The emergence of clusters in self-attention dynamics

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, Philippe Rigollet

▶ **To cite this version:**

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, Philippe Rigollet. The emergence of clusters in self-attention dynamics. 2024. hal-04092937v2

HAL Id: hal-04092937

https://hal.science/hal-04092937v2

Preprint submitted on 20 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Open licence - etalab

arXiv:2305.05465v6 [cs.LG] 12 Feb 2024

THE EMERGENCE OF CLUSTERS IN SELF-ATTENTION DYNAMICS

BORJAN GESHKOVSKI, CYRIL LETROUIT, YURY POLYANSKIY,
AND PHILIPPE RIGOLLET

Abstract. Viewing Transformers as interacting particle systems, we describe the geometry of learned representations when the weights are not time dependent. We show that particles, representing tokens, tend to cluster toward particular limiting objects as time tends to infinity. Cluster locations are determined by the initial tokens, confirming context-awareness of representations learned by Transformers. Using techniques from dynamical systems and partial differential equations, we show that the type of limiting object that emerges depends on the spectrum of the value matrix. Additionally, in the one-dimensional case we prove that the self-attention matrix converges to a low-rank Boolean matrix. The combination of these results mathematically confirms the empirical observation made by Vaswani et al. [VSP+17] that *leaders* appear in a sequence of tokens when processed by Transformers.

Contents

Part 1. Introduction and main results

1. Introduction

The introduction of Transformers in 2017 [VSP+17] marked a turning point in the AI revolution, powering breakthroughs in natural language modeling and computer vision. With remarkable empirical success, Transformers enable large language models to compute very powerful representations using the self-attention

mechanism. Yet, little is known about the geometric structure of these representations. As the size of these models grows at an astonishing rate, the need to understand their inner workings is becoming a pressing scientific challenge. In this work, we make a first step in this direction by describing the geometry of learned representations.

To provide a transparent presentation of our findings, we take a leaf out of the literature on continuous-time dynamics such as neural ordinary differential equations (ODEs) [CRBD18, Wei17, HR17]. By viewing layers as a time variable, this formalism has emerged as a flexible mathematical framework to implement and study ResNets [HZRS16a] as particular discrete-time versions of a parametrized dynamics of the form

$$\dot{x}(t) = f_\theta(x(t)), \qquad t \in [0, T].$$

Here θ is the trained parameter of a neural network and f_θ is characterized by the precise architecture of the ResNet[1]. In turn, an input (e.g., an image) $x(0) \in \mathbb{R}^d$ is mapped to its representation $x(T)$.

Unlike neural ODEs and ResNets, the representation map of Transformers is not solely a function of an individual input $x(0) \in \mathbb{R}^d$ but rather of a set/sequence $(x_1(0), \dots, x_n(0))$ of $n \geqslant 1$ d-dimensional tokens. These tokens then evolve in time by interacting with each other per the self-attention mechanism. Namely, following [SABP22], we view tokens as particles, and the transformer dynamics as an interacting particle system of the form

$$\dot{x}_i(t) = \sum_{j=1}^{n} P_{ij}(t) V x_j(t), \qquad t \in [0, +\infty), \tag{1.1}$$

for any $i \in [n]$, where $P_{ij}(t)$ are the entries of a $n \times n$ stochastic matrix $P(t)$, given by

$$P_{ij}(t) := \frac{e^{\langle Q x_i(t), K x_j(t) \rangle}}{\sum_{\ell=1}^{n} e^{\langle Q x_i(t), K x_\ell(t) \rangle}}, \qquad (i, j) \in [n]^2. \tag{1.2}$$

Here the matrices Q (Query), K (Key), and V (Value) are learned from data. Note that Q, K need not be square. The $n \times n$ matrix $P(t)$ is called *self-attention matrix*. The wording *attention* stems precisely from the fact that $P_{ij}(t)$ captures the attention given by token i to token j relatively to all tokens $\ell \in [n]$. The matrices Q and K in (1.2) warp the geometry of the input tokens, so that a trained attention matrix contains weights which indicate semantic relations between words. Such conclusions have been drawn in the context of language processing tasks in [VSP+17, Figures 3-5].

Our goal is to showcase the fact that self-attention, which itself is the core novelty of Transformers, entails a clustering effect. To that end, we focus on the pure self-attention dynamics described in (1.1). In particular, we do not model variations such as multiple heads, feed-forward layers, and layer normalization that are typically adjoined to self-attention dynamics of (1.1). However, on this last point, we note that our theoretical findings indicate that without any normalization, the dynamics (1.1) can diverge in some (or even all) directions over time. We leave these additional questions for future research; see Section 12.

[1] A classical choice is $\theta = (W, A, b) \in \mathbb{R}^{d \times d} \times \mathbb{R}^{d \times d} \times \mathbb{R}^d$ and $f_\theta(x) = W\sigma(Ax + b)$ where σ is an elementwise nonlinearity such as the ReLU ([HZRS16b]).

1.1. **Organization of the paper and summary of contributions.** The goal of this paper is to characterize clustered representations of a *trained* Transformer by studying the asymptotic behavior of a sequence of tokens $(x_1(t), \dots, x_n(t))$ as they evolve through the layers of a transformer architecture using the dynamics (1.1). In this setup, a Transformer is completely described by the weight matrices (Q, K, V) obtained during training. Note that we assume that these three matrices are *time-independent*. While this assumption is motivated by mathematical convenience, it is worth noting that such weight-sharing scenarios are in fact used in practice—see, e.g., ALBERT [LCG+20]—as they drastically reduce the number of parameters of a network.

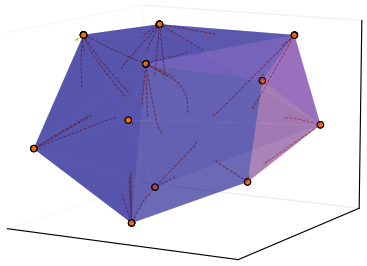


Figure 1. For $V = I_3$ tokens cluster toward the vertices of a convex polytope (Theorem 3.1).

With parameters (Q, K, V) fixed, tokens are subject to collective dynamics that we call *transformer dynamics*. While these dynamics are reminiscent of existing models for opinion dynamics and flocking, they present they own mathematical challenges requiring ad-hoc tools to study their asymptotic behavior.

The main conclusion of our analysis is that the set of tokens $\{x_1(t), \dots, x_n(t)\}$, appropriately rescaled, tends to a *clustered configuration* as $t \to \infty$. Our theoretical findings justify the empirical observation made in [VSP+17] that *leaders* appear in a sequence of tokens when processed by Transformers. We now list our main contributions.

(i) As a warm-up to the geometric characterization of the limits of sequences of tokens, we show in **Section 2** that when $d = 1$ and $V > 0$, the self-attention matrix $P(t)$ converges to a low-rank matrix with entries 0 and 1 as $t \to +\infty$ thus revealing the emergence of a small number of leaders that drive the transformer dynamics. The restriction $d = 1$ follows from technical considerations, and some pathological phenomena may occur in higher dimensions (see Remark 7.9). The proof may be found in **Section 7**. But numerical experiments (as well as past empirical work) indicate that the result may extend to higher dimensions for almost all initial sequences of tokens.

(ii) In **Section 3** we first focus on the case $V = I_d$ as a natural canonical choice that enables us to establish some of the main tools of the paper. We introduce a time re-scaling reminiscent of the layer normalization heuristics to alleviate the possible divergence of tokens. We show that along this scale the tokens converge to the boundary of a convex polytope. For almost all initial sequences they even converge to the vertices of the polytope, the number of which is significantly smaller than n. This elucidates the clustering phenomenon. (See Fig. 1.) When $V = -I_d$, all tokens following the dynamics (1.1) collapse to 0. The proofs are given in **Section 8**.

(iii) We build on these results and in **Section 4** consider the case wherein V is only assumed to have a simple and positive leading eigenvalue. This setting is much closer to reality and corresponds to actual learned matrices V (see Figure 10). We show that along the particular timescale, tokens cluster toward one of at most three

hyperplanes which are determined by the corresponding eigenvector. The proof is given in **Section 9**.

(iv) In **Section 5** we complete the results of Sections 3 and 4 by addressing the case where the leading eigenvalue has multiplicity. This results in clustering toward the vertices of a convex polytope in some directions, and a linear subspace in the others. The proof is provided in **Section 10**.

(v) We also prove the global existence and uniqueness of solutions of all dynamics considered in this work (including the mean field limit). We refer the reader to **Section 6** for more details.

We also observed numerically that our conclusions extend to more compound architectures (see Conjecture 4.3, **Section 12** and **Section 11**).

Value	**Key and Query**	**Limit geometry**	**Reference**
$V = I_d$	$Q^\top K > 0$	vertices of convex polytope	Theorem 3.1
$\lambda_1(V) > 0$, simple	$\langle Q\varphi_1, K\varphi_1\rangle > 0$	union of 3 parallel hyperplanes	Theorem 4.2
V paranormal	$Q^\top K > 0$	polytope $\times$ subspaces	Theorem 5.2
$V = -I_d$	$Q^\top K = I_d$	single cluster at origin*	Theorem 8.5

Table 1. Summary of the clustering results of this work. *All results except for the case $V = -I_d$ hold for the time-scaled dynamics (3.1).

Remark 1.1 (Discrete time)**.** *While we focus on the idealized setting of self-attention dynamics in continuous-time, this is solely done for convenience and all of our methods are straightforwardly applicable to the discrete-time setting. (See also Remark 3.4.) The discrete-time analog of* (1.1) *with time-step* $\Delta t > 0$ *(equal to* 1 *in practice) is simply the forward Euler iteration*

$$x_i((k+1)\Delta t) = x_i(k\Delta t) + \Delta t \sum_{j=1}^{n} \left(\frac{e^{\langle Qx_i(k\Delta t), Kx_j(k\Delta t)\rangle}}{\sum_{\ell=1}^{n} e^{\langle Qx_i(k\Delta t), Kx_\ell(k\Delta t)\rangle}} \right) Vx_j(k\Delta t), \quad (1.3)$$

for $k \in \mathbb{N}$.

1.2. **Notation.** We denote by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ the Euclidean dot product and norm respectively, and we use the shorthand $[n] := \{1, \dots, n\}$. For any matrix $M \in \mathbb{R}^{d\times d}$, we order its eigenvalues (repeated according to multiplicity) by decreasing order of modulus: $|\lambda_1(M)| \geqslant \dots \geqslant |\lambda_d(M)|$. We denote by $\|M\|_{\mathrm{op}}$ the ℓ^2—operator norm of the matrix M, equal to the largest singular value of M. Given a set $S \subset \mathbb{R}^d$, we define the distance of a point $x \in \mathbb{R}^d$ to S as $\mathrm{dist}(x, S) := \inf_{s\in S} \|x - s\|$, and by $\mathrm{conv}(S)$ the convex hull of S.

1.3. **Related work.** Our study and results build on several different lines of work, and we draw some parallels in what follows.

1.3.1. *Analysis of attention-based models.* Given the widespread use of Transformers in natural language processing, there has been a surge of interest in understanding the function and significance of attention layers within these models. In [YBR+20], the authors show that when treated as discrete-time systems with additional dense layers and multiple heads appended to the core attention mechanism,

Transformers exhibit the universal approximation property. In [LLH+20], the authors present, to the best of our knowledge, the first interacting particle systems perspective on Transformers. They then leverage the similarities between Transformers (with an additional feed-forward layer compared to (1.1)) and convection-diffusion equations to slightly improve the performance of Transformers by employing a Strang-Marchuk splitting scheme for time discretization. In [SABP22], the authors interpret system (1.1) as the characteristics of a continuity equation. Drawing on the similarities between (1.1) and Sinkhorn iterations, they propose a novel architecture dubbed *Sinkformer*, which possesses the desirable property of being a Wasserstein gradient flow.

1.3.2. *Quadratic complexity of Transformers.* The major computational challenge of Transformers is their high computational complexity, particularly when processing long sequences. Transformers require quadratic time and space complexity to process sequences, because each self-attention layer contains n^2 products of the form $\langle Qx_i, Kx_j\rangle$ (for $i,j \in [n]$). The empirical observation that the self-attention matrix P is close to a low rank matrix—see [LWLQ22, Section 4.4] for references—is cited as the inspiration behind *Linformers* [WLK+20] and the fine-tuning algorithm LoRA [HysW+22]. For both approaches, the low-rank structure is imposed rather than extracted from P itself. Other methods called *sparse attention* and *block attention* have been proposed to reduce the quadratic complexity—see [WLK+20, Section 2.2] for references. In the spirit of these works, a foreshadowing of the clustering mechanism was invoked in [VKF20], where queries are clustered into groups, again in view of reducing the quadratic complexity of self-attention. We point out that [DCL21] previously demonstrated that without skip connections, the dynamics trivializes and all tokens quickly lump together into a single tight cluster. Our work, in contrast, shows that in the presence of skip connections a rich cluster structure emerges.

Compared to the usual BERT, ALBERT [LCG+20] uses parameter-sharing across layers, meaning that the weight matrices Q, K, V in (1.1)-(1.2) do not depend on time, as in the present paper. This does not reduce the theoretical $O(n^2)$ complexity of the original Transformer, but, quoting [LCG+20], it "significantly reduce[s] the number of parameters for BERT without seriously hurting performance, thus improving parameter-efficiency. An ALBERT configuration similar to BERT-large has 18x fewer parameters and can be trained about 1.7x faster. The parameter reduction techniques also act as a form of regularization that stabilizes the training and helps with generalization".

1.3.3. *Neural collapse.* Our results and conclusions bear a resemblance to some geometric aspects of neural collapse for classification tasks [PHD20]. A key geometric aspect of neural collapse is the observation that, during the training of deep neural networks, the representation of different classes in the later layers of the network tends to form a tight cluster around the vertices of a simplex. The emergence of a simplex structure in the representation space provides insights into how the neural network organizes and separates the different classes.

1.3.4. *Clustering in interacting particle systems.* The transformer dynamics (1.1) have a strong connection to the vast literature on nonlinear systems arising in the modeling of opinion dynamics and flocking phenomena. In addition to the classical Kuramoto model describing synchronization/clustering of oscillators [Kur75,

ABV+05], the model which is most similar to (1.1) is the Krause model [Kra00]

$$\dot{x}_i(t) = \sum_{j=1}^{n} a_{ij}(x_j(t) - x_i(t)), \qquad a_{ij} = \frac{\phi(\|x_i - x_j\|^2)}{\sum_{k=1}^{n} \phi(\|x_i - x_k\|^2)}.$$

which is non-symmetric in general ($a_{ij} \neq a_{ji}$), much like (1.1). When ϕ is compactly supported, it has been shown in [JM14] that the particles $x_i(t)$ assemble in several clusters as $t \to +\infty$. Other models of opinion dynamics and flocking have been proposed and studied, among which the Vicsek model [VCBJ+95], the Hegselmann-Krause model [HK02] and the Cucker-Smale model [CS07]. These models may also exhibit a clustering behavior under various assumptions (see [MT14, CHH+16, HKPZ19] and the references therein). The transformer dynamics are also closely related to the dynamics employed in mean-shift clustering [Che95], and this work indirectly sheds some light on its theoretical properties.

The analysis of transformer dynamics presents unique mathematical challenges that cannot be addressed using the tools developed for these more primitive models. In particular, our work demonstrates how different choices for the parameters lead to remarkably diverse clustering patterns. Much more remains to be discovered and this work is a first attempt a rigorous mathematical analysis of these synthetic dynamics.

Acknowledgments. We thank Pierre Ablin, Léonard Boussioux, Enric Boix Adsera, Gabriel Peyré, Yair Shenfeld and Emmanuel Trélat for helpful discussions. C.L. was supported by the Simons Foundation Grant 601948, DJ. P.R. is supported by NSF grants IIS-1838071, DMS-2022448, and CCF-2106377. Y.P. is supported in part by the MIT-IBM Watson AI Lab.

2. Asymptotic low-rankness of the self-attention matrix

As mentioned in Section 1.3, numerical experiments in [WLK+20] show that the self-attention matrix P, defined in (1.2), has an almost low-rank structure. This observation has then been leveraged to reduce the quadratic complexity in the sequence length n which is inherent to Transformers, resulting in a non-negligible decrease in the cost of training.

As a warm-up to deriving complete geometric representations of the dynamics, our first result shows, in the simple $1d$ case that $P(t)$ indeed converges exponentially fast toward a matrix which is typically both Boolean and low-rank (see Fig. 3). Although there are clear obstructions to a rigorous extension of this result to higher dimensions (Remark 7.9), numerical experiments appear to show that this result holds in greater generality, for almost all initial sequences (Section 11).

To set this up, we introduce the set $\mathcal{P}$ of $n \times n$ matrices having the form illustrated in Fig. 2, where the asterisks denote arbitrary non-negative real numbers which add up to 1. The row of asterisks may actually be any row between the first and the last one.

Theorem 2.1 (Self-attention matrix converges to a low-rank Boolean matrix)**.** *Let $d = 1$. Suppose that the scalars (Q, K, V) satisfy $V > 0$ and $QK > 0$. For any initial sequence of pairwise distinct tokens $(x_1(0), \dots, x_n(0)) \in \mathbb{R}^n$, there exists some $P^* \in \mathcal{P}$ such that the self-attention matrix $P(t)$ defined in* (1.2) *converges to P^* as $t \to +\infty$.*

The proof may be found in Section 7. The rate of convergence toward P^* is in fact doubly exponential in t for coefficients outside the row of asterisks in Fig. 2. The proof the theorem also reveals that for almost all initial sequences of pairwise distinct tokens, P^* is actually of rank 1 or 2, i.e., the row of asterisks is equal to either $e_1 = (1, 0, \dots, 0)$ or $e_n = (0, \dots, 0, 1)$.

The interpretation of Theorem 2.1 is that in the $1d$ case, at most three tokens *capture the attention* of all tokens except at most one. Typically, these *leading* tokens are those carrying the largest amount of information. This is also illustrated in Fig. 4. Since the tokens x_i here evolve on $\mathbb{R}$, the right-most and left-most ones (which typically tend toward $\pm\infty$) capture the attention of all the others.

$$P_\sigma \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 0 \\ * & * & \dots & * \\ 0 & \dots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 1 \end{bmatrix}$$

Figure 2. Elements in $\mathscr{P}$, where $P_{\sigma_i} \in \mathbb{R}^{n\times n}$ are some permutation matrices, and asterisks denote arbitrary non-negative reals which add to 1.

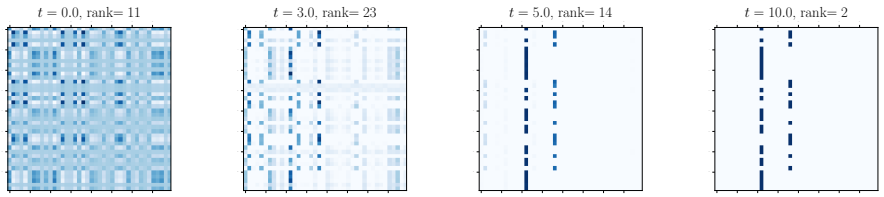


Figure 3. An illustration of the asymptotics of $P(t)$ entailed by Theorem 2.1 for $n = 40$ tokens, with $Q = K = 1$ and $V = 1$. (See Section 11 for details on computing.) Increasing n has no effect on this behavior of $P(t)$—see Fig. 11.

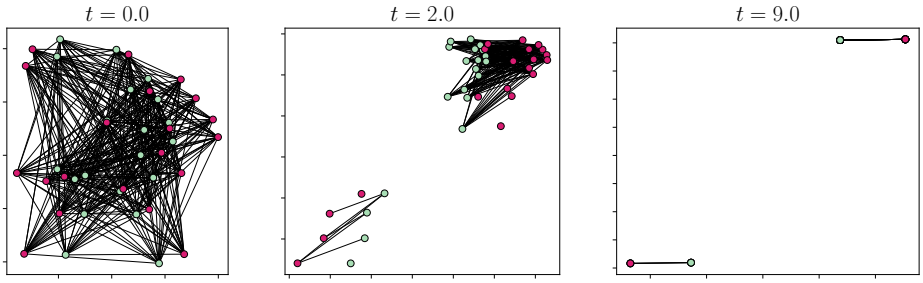


Figure 4. The clouds $\{Kx_i(t)\}_{i\in[20]}$ (green) and $\{Qx_j(t)\}_{j\in[20]}$ (purple) for $d = 2$ where pairwise points of clouds are connected by a line of width equal to $P_{ij}(t)$. Here $V \succ 0$ and $Q \succ 0$ are random matrices and $K = I_2$. The creation of clusters is reflected by the rank $\leqslant 2$ structure of the self-attention matrix $P(t)$. This interaction echoes findings illustrated in the original paper [VSP+17]—for instance, Figures 3-5 therein.

3. Clustering toward vertices of convex polytopes

In the rest of the paper, we seek to taxonomize various *clustering* results for the solutions to (3.1) when $t \to +\infty$, depending the sign and the multiplicity of the eigenvalues of V. We begin by focusing on what may appear to be the most natural[2] case $V = I_d$, as is also done in [SABP22]. In fact, we demonstrate (theoretically and numerically) later on, clustering is a generic phenomenon which holds under much less restrictive assumptions.

The transformer dynamics considered in (1.1) does not contain a layer normalization mechanism typically encountered in practice [VSP+17]. In absence of such a device, tokens may diverge to infinity as in Theorem 2.1. In fact, the norm of the tokens $x_i(t)$ typically diverges exponentially toward $+\infty$ for any d: this is expected, by analogy with the non-trivial solutions to $\dot{y}(t) = y(t)$.

To remedy this situation, we take inspiration from the solution $y(t) = e^{tV}y(0)$ to $\dot{y}(t) = Vy(t)$. Namely, for any $i \in [n]$ we consider the *rescaled* tokens

$$z_i(t) := e^{-tV}x_i(t),$$

which solve

$$\dot{z}_i(t) = \sum_{j=1}^{n} \left(\frac{e^{\langle Qe^{tV}z_i(t), Ke^{tV}z_j(t)\rangle}}{\sum_{k=1}^{n} e^{\langle Qe^{tV}z_i(t), Ke^{tV}z_k(t)\rangle}} \right) V(z_j(t) - z_i(t)), \qquad t \in [0, +\infty). \quad (3.1)$$

The initial condition remains the same: $x_i(0) = z_i(0)$ for any $i \in [n]$. More importantly, the coefficients of the self-attention matrix for the rescaled tokens $z_i(t)$ are the same as those for the original tokens $x_i(t)$. Whence, the conclusion of Theorem 2.1 also applies to the dynamics (3.1). We see this rescaling of tokens as a mathematically justified surrogate for the layer normalization.

The appearance of the exponential factor within the self-attention kernel facilitates the analysis of (3.1) compared to (1.1), and it is in fact instrumental in the proofs of all results that follow. Each result on the rescaled tokens $z_i(t)$ then gives information on the dynamics of the original tokens $x_i(t)$ by virtue of the relation $x_i(t) = e^{tV}z_i(t)$.

We are now able to state the main result of this section on the case $V = I_d$. The following theorem shows that the tokens $z_i(t)$ evolving per dynamics (3.1) converge to the boundary of a convex polytope as $t \to +\infty$. We present here a simplified but weaker version of our result for convenience, and refer the reader to Theorem 8.1 for a complete statement.

Theorem 3.1 (Convergence to points on the boundary of a convex polytope)**.** *Suppose $V = I_d$ and $Q^\top K \succ 0$. Then, for any initial sequence of tokens $\{z_i(0)\}_{i\in[n]} \subset \mathbb{R}^d$, there exists a convex polytope $\mathcal{K} \subset \mathbb{R}^d$ such that for any $i \in [n]$, $z_i(t)$ converges either to 0 or to some point on $\partial\mathcal{K}$ as $t \to +\infty$.*

The convex polytope $\mathcal{K}$ is completely determined by the initial sequence of tokens, and $Q^\top K$ (refer to Claim 1). Numerical experiments (e.g. Fig. 5) also lead us to claim that for almost all initial sequences of tokens, one should expect convergence of $z_i(t)$ ($i \in [n]$) toward some vertex of $\mathcal{K}$. (Furthermore, the number of vertices of $\mathcal{K}$ is often found to be significantly smaller than n.) It may however

[2]Note that the case $V = -I_d$ may appear equally natural. For such a choice of V, we show in Section 8.2 that the dynamics converge to a single cluster located at the origin. Multiplicative constants preserving the sign, i.e., $V = \pm cI_d, c > 0$ trivially yield the same conclusions.

happen that for initial sequences taken in some null set (not seen when tokens are drawn at random) some tokens converge to other points of the boundary $\partial\mathcal{K}$, namely in the interior of facets. On the other hand, for generic choices of initial sequences, we do not see a way to predict $\mathcal{K}$ explicitly besides running the full dynamics.

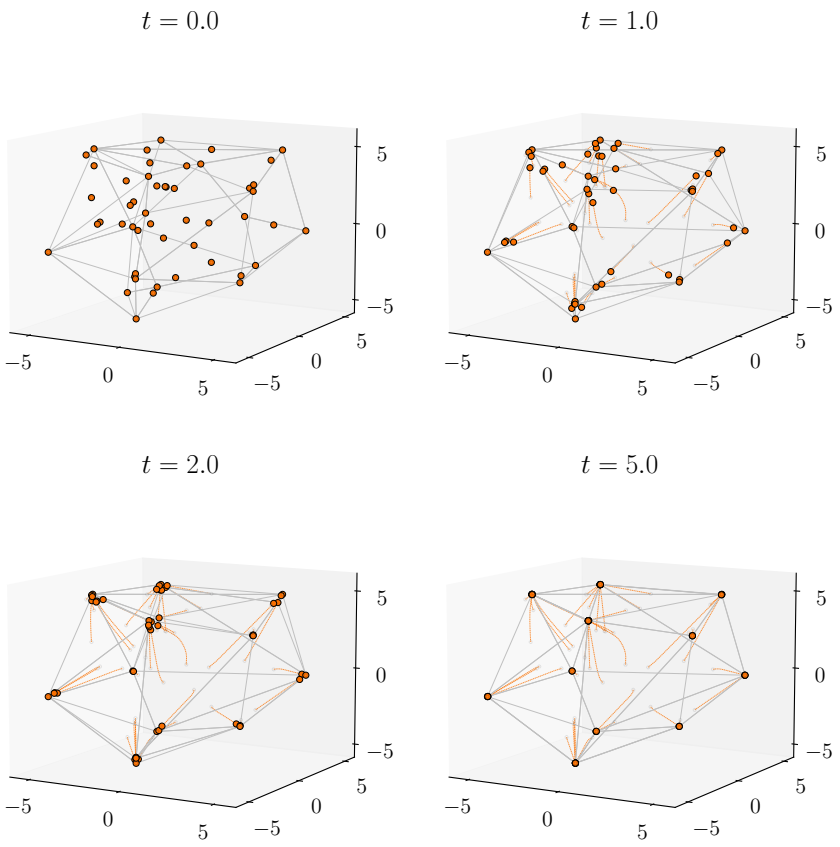


Figure 5. A toy example illustrating Theorem 3.1 with $n = 40$ tokens in $\mathbb{R}^3$. Here $Q = K = I_3$. The tokens converge to one of the vertices (*leaders*) of the limiting convex polytope.

Recall that the points $x_i(t) = e^t z_i(t)$ when $V = I_d$ follow the original dynamics (1.1). Akin to Theorem 2.1, this result also shows the emergence of a set of *leaders* (given by the vertices of $\mathcal{K}$) attracting all tokens as t grows. It has been experimentally observed (first in [VSP+17]) that in trained Transformers, tokens focus their attention on local leaders in a way that seems to reproduce the syntactic and semantic structure of sentences.

The proof of Theorem 3.1 is postponed to Section 8, and amounts to a couple of effects entailed by the dynamics. First of all, the convex hull of the particles is shrinking over time (Proposition 8.2). This is due to the fact that the distance of the particle nearest to any half-space (not containing the particles) increases with

time. On the other hand, the convex hull ought not collapse since particles which have not concentrated near the boundary of the limiting polytope will continue to increase in magnitude until they themselves reach this boundary (Step 2 in the proof). This occurs due to the time-rescaling.

Remark 3.2. *Assuming $Q^\top K \succ 0$ does not seem to be essential for our conclusions; instead, it guides the direction of the proof. To emphasize the broader validity of our conclusion beyond this specific assumption, we conducted additional experiments (refer to Section 12.1) which suggest that Theorem 3.1 (as well as Theorems 4.2 and 5.2 stated below) holds in more generality.*

Remark 3.3 (Rate of convergence)**.** *Although Theorem 3.1 (as well as Theorems 4.2 and 5.2 stated below) does not specify a rate of convergence toward $\partial\mathcal{K}$, we expect (and observe through numerics) that convergence happens very quickly—after few layers, most tokens are already clustered. What "few layers" means here necessarily depends on the typical modulus of the initial tokens, since the dynamics* (1.1) *is not invariant under multiplication of all initial conditions by a fixed real number.*

Remark 3.4 (Discrete time)**.** *As alluded to in Remark 1.1, all our results extend to the discrete-time Transformers* (1.3)*. Indeed, just as in the continuous-time case, there is a natural rescaled dynamics, which is the discrete analogue of* (3.1)*: if we set $R = I_d + V\Delta t$, and assume that R is invertible (which is the case for sufficiently small Δt), then $z_i(k\Delta t) = R^{-k}x_i(k\Delta t) := z_i^{[k]}$ satisfies*

$$z_i^{[k+1]} = z_i^{[k]} + \Delta t \sum_{j=1}^{n} \left(\frac{e^{\langle QR^k z_i^{[k]}, KR^k z_j^{[k]} \rangle}}{\sum_{\ell=1}^{n} e^{\langle QR^k z_i^{[k]}, KR^k z_\ell^{[k]} \rangle}} \right) R^{-1} V \left(z_j^{[k]} - z_i^{[k]} \right), \qquad k \in \mathbb{N}.$$

The proofs of Theorems 2.1, 8.5, 3.1, 4.2, and 5.2 carry through with straightforward modifications.

Let us provide some comments on the proof of Theorem 3.1 in the discrete-time setting, for the sake of completeness. First of all, Proposition 8.2 holds intuitively because for all integers $i \in [n]$ and $k \geqslant 1$,

$$z_i^{[k+1]} = \frac{1}{1+\Delta t} \left(z_i^{[k]} + \Delta t \sum_{j=1}^{n} P_{ij}^{[k]} z_j^{[k]} \right) \in \operatorname{conv} \left(\left\{ z_j^{[k]} \right\}_{j \in [n]} \right).$$

We then define the candidate set of limit points as in (8.6)*, and Claim 1 holds without any change in the statement or in the proof. Then, just as in Steps 2 and 3 in the proof of 8.1, we can first show that if $z_i^{[k]}$ is not already near some point in the candidate limit set, it will keep moving toward the boundary of the convex polytope. Finally, we can prove that tokens cannot circulate indefinitely between different points on the boundary. The combination of these arguments would establish the convergence of each token toward some point in the set given by* (8.6)*.*

4. Clustering toward hyperplanes

While being a natural example to consider, value matrices found empirically are much more general than $V = I_d$, which we considered in the previous section. We now turn our attention to a significantly more general setting of value matrices, which we formalize as follows.

Definition 4.1. *We call (Q, K, V) a* good triple *if the two following conditions are satisfied:*

- *the eigenvalue of* V *with largest modulus is real, positive, and simple; namely,*
$$\lambda_1(V) > |\lambda_2(V)| \geqslant \ldots \geqslant |\lambda_d(V)|.$$
- $\langle Q\varphi_1, K\varphi_1\rangle > 0$ *for any* $\varphi_1 \in \mathbb{R}^d$ *lying on the line* $\ker(V - \lambda_1(V)\mathrm{Id})$.

The second condition simply states that the quadratic form $\langle Q\cdot, K\cdot\rangle$ is positive definite along the eigenspace associated to the leading eigenvalue of V. Note also that if all entries of V are positive, the first condition is automatically satisfied by virtue of the Perron-Frobenius theorem. In fact, this assumption is generic. On the one hand, it is satisfied by some pre-trained value matrices for ALBERT (Figure 10). On the other hand, numerical experiments indicate that a constant fraction (about 14%) of matrices from the real Ginibre ensemble in dimension $d = 128$—this proportion is known to vanish as $d \to \infty$, albeit very slowly [RS14].

Our clustering result in the setting of good triples can be summarized as follows: the coordinate $\langle z_i(t), \frac{\varphi_1}{\|\varphi_1\|}\rangle$ of any token $z_i(t)$ along the eigenspace spanned by φ_1 converges, as $t \to +\infty$, toward one among possibly 3 real scalars. Consequently, all the tokens $z_i(t)$ converge toward one among at most three parallel hyperplanes; see Fig. 6 for an illustration.

Theorem 4.2 (Convergence toward $\leqslant 3$ hyperplanes)**.** *Assume that* (Q, K, V) *is a good triple in the sense of Definition 4.1. Then, for any initial sequence of tokens* $\{z_i(0)\}_{i\in[n]} \subset \mathbb{R}^d$*, there exist at most three parallel hyperplanes in* $\mathbb{R}^d$ *such that for any* $i \in [n]$*, the distance of the solution* $z_i(t)$ *to* (3.1) *to one of these hyperplanes converges to* 0 *as* $t \to +\infty$*.*

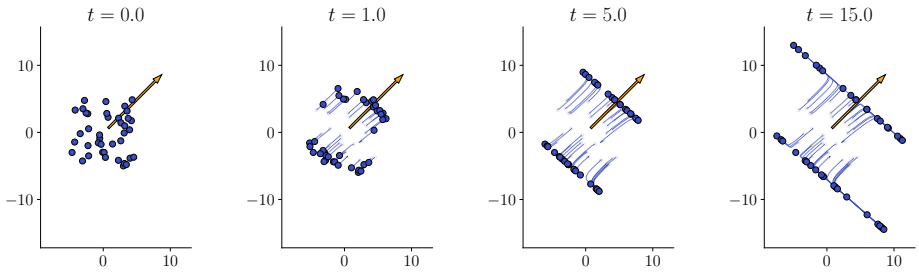


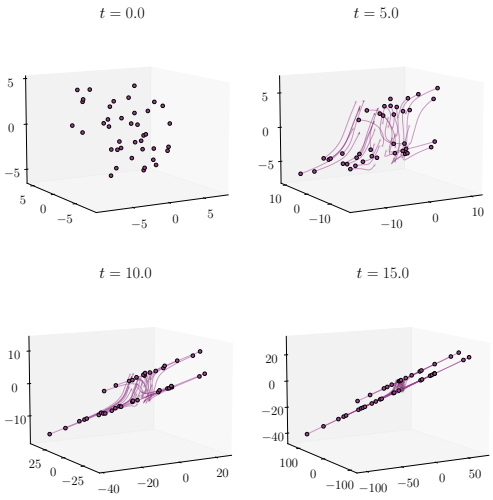
Figure 6. Illustrating Theorem 4.2 with $n = 40$ tokens in $\mathbb{R}^2$. Here $Q = K = I_2$, V is a random symmetric matrix with eigenvalues $\{1.35, -0.07\}$, and $\varphi_1 = (0.76, 0.65)$. The components of the tokens in the direction of φ_1 (orange arrow) cluster over time. (See Figures 13–14 for examples in $\mathbb{R}^3$.) We also observe that tokens typically cluster toward only two hyperplanes—a third one (passing through the origin) may appear for non-generic initial sequences. The hyperplanes are perpendicular to φ_1 since V is diagonalizable.

The proof may be found in Section 9. The important role played by $\lambda_1(V)$ in the dynamics may be seen in (3.1): the component of $z_i(t)$ along φ_1 determines the size of $e^{tV}z_i(t)$ in the exponent appearing in (3.1). The tokens $z_j(t)$ attracting other tokens $z_i(t)$ are those for which this component along φ_1 is largest in modulus. This attraction process forms the clusters. These *leaders*, as in all our results, have been empirically observed to be the ones carrying the largest amount of information in the sentence (see Supplementary material in [VSP+17]).

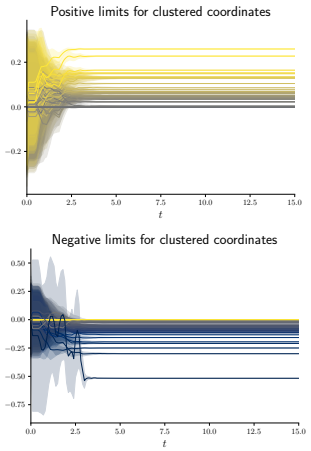
Furthermore, Theorem 4.2 can also be interpreted in more classical machine learning terms. On the one hand, it can be seen as an instance of *K-flats clustering* [BM00, Vid11]—points in the input sequence are clustered, based on their intrinsic similarity, to at most 3 "flats" of dimension $d-1$. On the other hand, it ensures that for a good triple (Q, K, V), (3.1) generates a *linearly separable* representation of tokens.

Beyond a single direction? Numerical experiments (e.g., Fig. 7) indicate that a similar phenomenon emerges for more complex V. We formulate following conjecture which is a natural generalization of Theorem 4.2.

Conjecture 4.3 (Codimension conjecture)**.** *Let $k \geqslant 1$ be the number of eigenvalues of V with positive real part. Then there exist at most three parallel Euclidean subspaces of $\mathbb{R}^d$ of codimension k such that for any $i \in [n]$, the distance of $z_i(t)$ to one of these subspaces converges to 0 as $t \to +\infty$.*



(a) Conjecture 4.3: low-dimensional case.



(b) Conjecture 4.3: high-dimensional case.

Figure 7. (a) $n = 40$, $d = 3$ and $Q = K = I_3$ with V a random matrix with eigenvalues $\{1.96, -0.22, 0.25\}$. The $k = 2$ positive eigenvalues of V generate attraction between the tokens and even convergence in the corresponding eigenspaces–this explains the codimension k statement. The negative eigenvalue generates a repulsive effect between the tokens, and we see a divergence along two lines (note the different scales between the four figures). (b) $n = 256$, $d = 128$, with (Q, K, V) fixed random matrices and V symmetric. For each coordinate j corresponding to a positive eigenvalue, the variance of the set $\{\varphi_j^*(z_i(t)) \colon i \in [n]\}$ (shaded area) tends to 0 with t, while the mean (solid lines) converges to one among two real scalars: one positive (top figure), one negative (bottom) figure. Coordinates corresponding to negative eigenvalues diverge (Fig. 15).

5. A mix of hyperplanes and polytopes

We now turn our attention to an even more general version of Theorem 4.2, which does not require the leading eigenvalue of V to be simple. The resulting theorem can be viewed as a combination of Theorem 4.2 and Theorem 3.1. Specifically, we assume that V behaves as the identity when acting on the eigenspace of the leading eigenvalue. This property is automatically satisfied if V is normal—so that its eigenvectors form an orthonormal basis—so we call such a V *paranormal.*

Definition 5.1. *We call* (Q, K, V) *a* good triple with multiplicity *if the following conditions hold:*

(i) $Q^\top K$ *is positive definite:* $Q^\top K > 0$*;*

(ii) V *is paranormal: there exist two linear subspaces* $\mathcal{F}, \mathcal{G} \subset \mathbb{R}^d$ *which are invariant under* V*, and such that* $\mathcal{F} \oplus \mathcal{G} = \mathbb{R}^d$*,* $V_{|\mathcal{F}} = \lambda \mathrm{Id}$ *for* $\lambda > 0$*, and* $\rho(V_{|\mathcal{G}}) < \lambda$*, where* $\rho(\cdot)$ *denotes the spectral radius (the maximal modulus of eigenvalues).*

An example of such a V is used for Fig. 8. We may now state our main result in the setting of good triples with multiplicity. The proof may be found in Section 10.

Theorem 5.2 (Clustering for λ_1 with multiplicity)**.** *Suppose that* (Q, K, V) *is a good triple with multiplicity in the sense of Definition 5.1. Then, for any initial sequence* $\{z_i(0)\}_{i\in[n]} \subset \mathbb{R}^d$*, there exists a bounded convex polytope* $\mathcal{K} \subset \mathcal{F}$ *such that setting* $\mathscr{H} := (\partial\mathcal{K} \cup \{0\}) \times \mathcal{G}$*, for any* $i \in [n]$*, we have* $\operatorname{dist}(z_i(t), \mathscr{H}) \to 0$ *as* $t \to +\infty$*.*

Part 2. Proofs

6. Well-posedness

We collect several facts regarding the global-in-time existence and uniqueness of solutions to all systems under consideration. Throughout the remainder of the paper, we use the terminology "tokens" and "particles" interchangeably.

To prove these results, we leverage the underlying continuity equation (see (6.1)). For the sake of future use, we prove a more general well-posedness result for the continuity equation than what is needed in this paper.

6.1. **Notation.** We denote by $\mathcal{P}_c(\mathbb{R}^d)$ the set of compactly supported probability measures on $\mathbb{R}^d$, and by $\mathcal{P}_2(\mathbb{R}^d)$ the set of probability measures μ on $\mathbb{R}^d$ having finite second moment: $\int_{\mathbb{R}^d} \|x\|^2 \,\mathrm{d}\mu(x) < +\infty$. Let $C^0(\mathbb{R}; \mathcal{P}_c(\mathbb{R}^d))$ denote the Banach space of continuous curves $\mathbb{R} \ni t \mapsto \mu(t) \in \mathcal{P}_c(\mathbb{R}^d)$. Here $\mathcal{P}_c(\mathbb{R}^d)$ is endowed with the weak topology, which coincides with the topology induced by the Wasserstein distance W_p for any $p \in [1, +\infty)$.

As seen below, for compactness purposes regarding solutions to the continuity equation, we consider an additional property on the support of such curves, summarized by the following definition.

Definition 6.1 (Equi-compactly supported curves)**.** *The set* $C^0_{\mathrm{co}}(\mathbb{R}; \mathcal{P}_c(\mathbb{R}^d))$ *consists of all elements* $\mu \in C^0(\mathbb{R}; \mathcal{P}_c(\mathbb{R}^d))$ *such that for any* $t_0, t_1 \in \mathbb{R}$*, there exists a compact subset* $\mathcal{K} \subset \mathbb{R}^d$ *such that* $\operatorname{supp}(\mu(t)) \subset \mathcal{K}$ *for any* $t \in [t_0, t_1]$*.*

We emphasise that there exist elements in $C^0(\mathbb{R}; \mathcal{P}_c(\mathbb{R}^d))$ which do not satisfy this property with regard to their support—e.g., $\mu(t) = (1 - e^{-\frac{1}{t^2}})\delta_0 + e^{-\frac{1}{t^2}}\delta_{\frac{1}{t}}$.

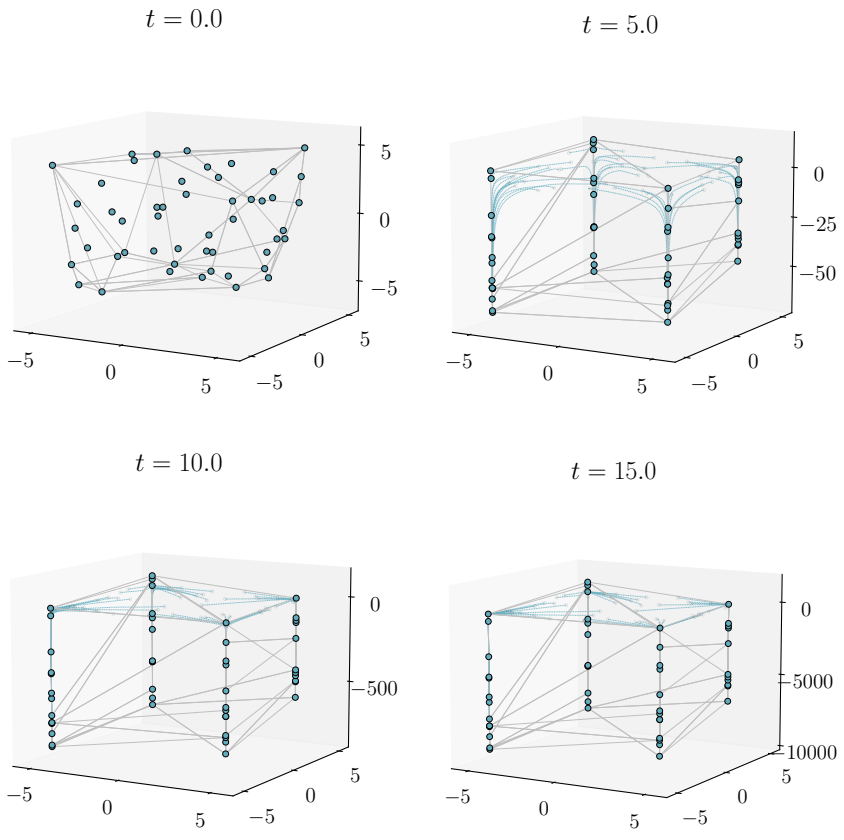


Figure 8. Illustrating Theorem 5.2 with $n = 40$ tokens in $\mathbb{R}^3$. As before, $Q = K = I_d$, and we take $V = \mathrm{diag}(1, 1, -\frac{1}{2})$. A convex polytope $\mathcal{K}$ emerges before time 5, toward which two coordinates of the tokens cluster, and persists throughout the evolution, while the tokens diverge along the coordinate corresponding to the eigenvalue $-\frac{1}{2}$ (note the different scales between the four figures).

6.2. **Well-posedness of the ODEs.** For any initial datum, i.e. a sequence of n points in $\mathbb{R}^d$, the dynamics (1.1) is well-posed, in the sense that it admits a unique solution defined for all times.

Proposition 6.2. *For any initial datum $\mathbf{X}_0 = (x_1^0, \dots, x_n^0) \in (\mathbb{R}^d)^n$, there exists a unique Lipschitz continuous function $\mathbb{R} \ni t \mapsto \mathbf{X}(t) = (x_1(t), \dots, x_n(t))$ such that $x_i(\cdot)$ solves (1.1) and satisfies $x_i(0) = x_i^0$ for any $i \in [n]$.*

We postpone the proof which is seen as a corollary of the well-posedness for the corresponding continuity equation. It follows that the equation (3.1) is also well-posed:

Proposition 6.3. *For any initial datum $\mathbf{Z}_0 = (z_1^0, \dots, z_n^0) \in (\mathbb{R}^d)^n$, there exists a unique Lipschitz continuous function $\mathbb{R} \ni t \mapsto \mathbf{Z}(t) = (z_1(t), \dots, z_n(t))$ such that $z_i(\cdot)$ solves (3.1) and satisfies $z_i(0) = z_i^0$ for any $i \in [n]$.*

Proof of Proposition 6.3. Since the equations (1.1) and (3.1) are related by the change of variables $x_i(t) = e^{tV} z_i(t)$, Proposition 6.3 is an immediate consequence of Proposition 6.2. □

6.3. **The continuity equation.** To prove Proposition 6.2, we show a more general result concerning global existence and uniqueness of solutions to the corresponding continuity equation[3]

$$\begin{cases} \partial_t \mu + \operatorname{div}(\mathcal{X}[\mu]\mu) = 0 & \text{in } (0,+\infty)\times\mathbb{R}^d \\ \mu_{|t=0} = \mu_0 & \text{in } \mathbb{R}^d, \end{cases} \tag{6.1}$$

when $\mathcal{X}[\mu]$ is the *attention kernel*

$$\mathcal{X}[\mu](x) := \frac{\displaystyle\int_{\mathbb{R}^d} e^{\langle Qx, Ky\rangle} Vy \,\mathrm{d}\mu(y)}{\displaystyle\int_{\mathbb{R}^d} e^{\langle Qx, Ky\rangle} \,\mathrm{d}\mu(y)}. \tag{6.2}$$

We will make use of the following notion of solution.

Definition 6.4. *Fix $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d)$. We say that $t \mapsto \mu(t) =: \mu_t$ is a solution to the Cauchy problem (6.1) if $\mu \in C^0_{\mathrm{co}}(\mathbb{R}, \mathcal{P}_c(\mathbb{R}^d))$, the function*

$$\mathbb{R} \ni t \mapsto \int_{\mathbb{R}^d} g(x) \,\mathrm{d}\mu_t(x)$$

is absolutely continuous for every $g \in C^\infty_c(\mathbb{R}^d)$, and

$$\int_{\mathbb{R}^d} g(x) \,\mathrm{d}\mu_t(x) = \int_{\mathbb{R}^d} g(x) \,\mathrm{d}\mu_0(x) + \int_0^t \int_{\mathbb{R}^d} \langle \nabla g(x), \mathcal{X}[\mu_t](x)\rangle \,\mathrm{d}\mu_s(x)\,\mathrm{d}s$$

holds for almost every $t \in \mathbb{R}$.

We will make use of the following lemma regarding (6.2).

Lemma 6.5. *For any $R > 0$ there exists a constant $C_1(R) > 0$ such that for any $\mu, \nu \in \mathcal{P}_c(\mathbb{R}^d)$ with support in $B(0,R)$,*

$$\|\mathcal{X}[\mu]\|_{L^\infty(\mathbb{R}^d;\mathbb{R}^d)} \leqslant \|V\|_{\mathrm{op}} R, \tag{6.3}$$

$$\|\nabla_x \mathcal{X}[\mu]\|_{L^\infty(\mathbb{R}^d;\mathbb{R}^{d\times d})} \leqslant 2\|Q^\top K\|_{\mathrm{op}} \|V\|_{\mathrm{op}} R^2 \tag{6.4}$$

$$\|\mathcal{X}[\mu](\cdot) - \mathcal{X}[\nu](\cdot)\|_{L^\infty(B(0,R);\mathbb{R}^d)} \leqslant C_1(R) W_2(\mu,\nu). \tag{6.5}$$

Proof. We henceforth set $\mathsf{G}(x,y) := e^{\langle Qx, Ky\rangle}$. To show (6.3), since $\mathsf{G} > 0$ we see that for any $x \in \mathbb{R}^d$,

$$\|\mathcal{X}[\mu](x)\| \leqslant \|V\|_{\mathrm{op}} \frac{\displaystyle\int_{B(0,R)} \mathsf{G}(x,y)\|y\| \,\mathrm{d}\mu(y)}{\displaystyle\int_{B(0,R)} \mathsf{G}(x,y) \,\mathrm{d}\mu(y)} \leqslant \|V\|_{\mathrm{op}} R.$$

[3]which can be seen as a mean-field limit, and is sometimes also referred to as a *Vlasov equation.*

We now show (6.4). Note that $\nabla_x \mathsf{G}(x,y) = Q^\top K y \mathsf{G}(x,y)$, thus, arguing as above, we find

$$\begin{aligned}\|\nabla_x \mathcal{X}[\mu](x)\| &\leqslant \frac{\displaystyle\int_{B(0,R)} \|\nabla_x \mathsf{G}(x,y)\| \|Vy\| \,\mathrm{d}\mu(y)}{\displaystyle\int_{B(0,R)} \mathsf{G}(x,y)\,\mathrm{d}\mu(y)} \\ &\qquad + \|V\|_{\mathrm{op}} \frac{\displaystyle\int_{B(0,R)} \mathsf{G}(x,y)\|y\|\,\mathrm{d}\mu(y)}{\displaystyle\int_{B(0,R)} \mathsf{G}(x,y)\,\mathrm{d}\mu(y)} \frac{\displaystyle\int_{B(0,R)} \|\nabla_x \mathsf{G}(x,y)\|\,\mathrm{d}\mu(y)}{\displaystyle\int_{B(0,R)} \mathsf{G}(x,y)\,\mathrm{d}\mu(y)} \\ &\leqslant 2\|Q^\top K\|_{\mathrm{op}} \|V\|_{\mathrm{op}} R^2.\end{aligned}$$

We finally prove (6.5). Using the fact that

$$\int_{\mathbb{R}^d} \mathsf{G}(x,y)\,\mathrm{d}\mu(y) \geqslant \left(\inf_{(x,y)\in B(0,R)^2} \mathsf{G}(x,y)\right) \mu(B(0,R)),$$

–with an analogous bound for ν–, we see that it suffices to bound

$$\left| \int_{\mathbb{R}^d} \mathsf{G}(x,y) V y \,\mathrm{d}\mu(y) \int_{\mathbb{R}^d} \mathsf{G}(x,y)\,\mathrm{d}\nu(y) - \int_{\mathbb{R}^d} \mathsf{G}(x,y) V y \,\mathrm{d}\nu(y) \int_{\mathbb{R}^d} \mathsf{G}(x,y)\,\mathrm{d}\mu(y) \right|$$

from above. We rewrite this difference by making $\mu - \nu$ appear artificially, and we then use the triangle inequality along with the fact that both $\int_{\mathbb{R}^d} \mathsf{G}(x,y) V y \,\mathrm{d}\mu(y)$ and $\int_{\mathbb{R}^d} \mathsf{G}(x,y)\,\mathrm{d}\mu(y)$ are bounded from above (by $e^{\|Q^\top K\|_{\mathrm{op}} R^2} \max(1, \|V\|_{\mathrm{op}} R)$). We thus end up with the task of bounding from above the absolute values of

$$\int_{\mathbb{R}^d} \mathsf{G}(x,y)(\,\mathrm{d}\nu - \,\mathrm{d}\mu)(y) \qquad \text{and} \qquad \int_{\mathbb{R}^d} \mathsf{G}(x,y) V y (\,\mathrm{d}\nu - \,\mathrm{d}\mu)(y). \tag{6.6}$$

For the first integral, from the Kantorovich-Rubinstein duality we deduce

$$\left| \int_{\mathbb{R}^d} \mathsf{G}(x,y)(\,\mathrm{d}\nu - \,\mathrm{d}\mu)(y) \right| \leqslant \|\mathsf{G}(x,\cdot)\|_{C^{0,1}(B(0,R))} W_1(\mu,\nu). \tag{6.7}$$

We now recall the following inequality relating Wasserstein distances of different orders: for any $p \geqslant 1$ and any bounded set B, for all Radon measures μ, ν supported in B,

$$W_1(\mu,\nu) \leqslant W_p(\mu,\nu) \leqslant \mathrm{diam}(B)^{1-\frac{1}{p}} W_1(\mu,\nu)^{1/p}. \tag{6.8}$$

Using (6.8) and the fact that the Lipschitz constant $\|\mathsf{G}(x,\cdot)\|_{C^{0,1}(B(0,R))}$ is uniformly bounded for $\|x\| \leqslant R$ by some $C_R > 0$ in (6.7), we end up with

$$\left| \int_{\mathbb{R}^d} \mathsf{G}(x,y)(\,\mathrm{d}\nu - \,\mathrm{d}\mu)(y) \right| \leqslant C_R W_2(\mu,\nu).$$

The same chain of inequalities applies to the second integral in (6.6) (with the additional multiplier $\|V\|_{\mathrm{op}} R$), which finally leads us to (6.5). □

The following existence and uniqueness result is adapted from [PRT15, Theorem 2.3]. In fact, the result holds true for any vector field $\mathcal{X}[\mu]$ on $\mathbb{R}^d$ satisfying conditions analog to those entailed by Lemma 6.5.

Proposition 6.6. *For any initial condition $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d)$, the Cauchy problem* (6.1) *admits a unique solution $\mu \in C^0_{\mathrm{co}}(\mathbb{R}; \mathcal{P}_c(\mathbb{R}^d))$ in the sense of Definition 6.4.*

Furthermore, we have the following stability estimate for solutions: for any $R > 0$ and $T > 0$, there exists a constant $C(T, R) > 0$ such that for any $\mu_0, \nu_0 \in \mathcal{P}_c(\mathbb{R}^d)$ with support in $B(0, R)$,

$$W_2(\mu(t), \nu(t)) \leqslant e^{C(T,R)t} W_2(\mu_0, \nu_0) \tag{6.9}$$

for any $t \in [0, T]$, where $\mu(t)$ and $\nu(t)$ solve (6.1) *with initial conditions μ_0 and ν_0 respectively.*

Results of this nature can be found in the literature—see for instance [PRT15]. They are however not sufficient for our purposes. We wrote Proposition 6.6 in the W_2 setting instead of the usual W_1 setting (used for instance for the classical *Dobrushin estimate* [Dob79, Gol13]) because it allows to extend the results of [WHL19] without difficulty from classical ResNets to self-attention dynamics. We recall that the goal of [WHL19] is to import classical (mean-field) optimal control tools such as the Pontryagin maximum principle and the analysis of Hamilton-Jacobi-Bellman equations to deep learning, and relies heavily on W_2 estimates (e.g., in [WHL19, Section 4]).

Proof of Proposition 6.6. To ease reading, we split the proof in three parts.

Part 1: Existence. Fix an arbitrary $T > 0$. For $k \geqslant 1$, set

$$\tau_k := \frac{T}{2^k}.$$

We define a sequence of curves $\mu^k : [0, T] \to \mathcal{P}_c(\mathbb{R}^d)$ by the following scheme[4]:

(i) $\mu^k(0) := \mu_0$;

(ii) $\mu^k(\ell\tau_k + t) := \left(\Phi^t_{\mathcal{X}[\mu^k(\ell\tau_k)]}\right)_{\#} \mu^k(\ell\tau_k)$ for $\ell \in \{0, \dots, 2^k - 1\}$ and $t \in (0, \tau_k]$,

where for any $x \in \mathbb{R}^d$, $\Phi^t_{\mathcal{X}[\mu^k(\ell\tau_k)]}(x)$ is the unique solution to the Cauchy problem

$$\begin{cases} \dot{y}(t) = \mathcal{X}[\mu^k(\ell\tau_k)](y(t)) & \text{on } [0, \tau_k] \\ y(0) = x. \end{cases}$$

(The above problem indeed has a unique solution for any $x \in \mathbb{R}^d$ by virtue of the Cauchy-Lipschitz theorem, using (6.4).) By construction, $\mu^k \in C^0([0, T]; \mathcal{P}_c(\mathbb{R}^d))$ for any $k \geqslant 1$.

We begin by showing that there exists a radius $R = R(T) > 0$ independent of k such that $\mathrm{supp}(\mu^k(t)) \subset B(0, R)$ for any $k \geqslant 1$ and $t \in [0, T]$. To this end, for any $t \in [0, T]$ and $k \geqslant 1$, let $R_k(t) > 0$ denote the smallest positive radius[5] such that $\mathrm{supp}(\mu^k(t)) \subset B(0, R_k(t))$. We will first look to show that

$$\mathrm{supp}(\mu^k(\ell\tau_k + t)) \subset B(0, R_k(\ell\tau_k) + t\|V\|_{\mathrm{op}} R_k(\ell\tau_k)). \tag{6.10}$$

Let $x \in \mathrm{supp}(\mu^k(\ell\tau_k + t))$, thus $\mu^k(\ell\tau_k + t)(B(x, \varepsilon)) > 0$ for any $\varepsilon > 0$. By the change of variables formula, we find that

$$\int_{(\Phi^t_{\mathcal{X}[\mu^k(\ell\tau_k)]})^{-1}(B(x,\varepsilon))} \mathrm{d}\mu^k(\ell\tau_k)(z) > 0.$$

[4] In other words we "freeze" the vector field $\mathcal{X}$ on each interval of the form $[\ell\tau_k, (\ell+1)\tau_k)$, and during this time interval, we follow the flow generated by this vector field starting from $\mu^k(\ell\tau_k)$.

[5] This radius always exists, since $\mu^k(t)$ is compactly supported.

Consequently $(\Phi^t_{\mathcal{X}[\mu^k(\ell\tau_k)]})^{-1}(B(x,\varepsilon))\cap\mathrm{supp}(\mu^k(\ell\tau_k))\neq\varnothing$, and let z be an element lying in this set. From the Duhamel formula, we gather that

$$\Phi^t_{\mathcal{X}[\mu^k(\ell\tau_k)]}(z) =: y(t) = z + \int_0^t \mathcal{X}[\mu^k(\ell\tau_k)](y(s))\,\mathrm{d}s.$$

Since $z \in (\Phi^t_{\mathcal{X}[\mu^k(\ell\tau_k)]})^{-1}(B(x,\varepsilon))$, we find that

$$\left\| z + \int_0^t \mathcal{X}[\mu^k(\ell\tau_k)](y(s))\,\mathrm{d}s - x \right\| \leqslant \varepsilon.$$

Using the triangle inequality, (6.3), and since $z \in \mathrm{supp}(\mu^k(\ell\tau_k))$ implies $z \in B(0, R_k(\ell\tau_k))$, we deduce that

$$\|x\| \leqslant \varepsilon + t\|V\|_{\mathrm{op}} R_k(\ell\tau_k) + R_k(\ell\tau_k).$$

Since $\varepsilon > 0$ is arbitrary, this inequality yields (6.10). We now use (6.10) to prove the original claim. Using the definition of the radius $R_k(t)$, we evaluate (6.10) at $t = \tau_k$ and find

$$R_k((\ell+1)\tau_k) \leqslant (1 + \|V\|_{\mathrm{op}}\tau_k) R_k(\ell\tau_k).$$

By induction, we deduce that

$$R_k(\ell\tau_k) \leqslant (1 + \|V\|_{\mathrm{op}}\tau_k)^\ell R_k(0),$$

whence

$$R_k(\ell\tau_k) \leqslant \left(1 + \|V\|_{\mathrm{op}}\frac{T}{2^k}\right)^{2^k} R_k(0) < e^{\|V\|_{\mathrm{op}}T} R_0,$$

where $R_0 > 0$ denotes the smallest positive radius such that $\mathrm{supp}(\mu_0) \subset B(0, R_0)$. Since the above bound is independent of k, the claim follows, yielding the desired radius $R = R(T) > 0$ bounding the support of every element in the sequence. In turn, we also deduce that $\mu^k \in C^0_{\mathrm{co}}(\mathbb{R}; \mathcal{P}_c(\mathbb{R}^d))$ for any $k \geqslant 1$.

Using the above fact, along with (6.3) and the definition of $\mu^k(\ell\tau_k + t)$, we find that

$$W_2\left(\mu^k(\ell\tau_k + t), \mu^k(\ell\tau_k)\right) \leqslant \|V\|_{\mathrm{op}} R t$$

for any $\ell \in \{0, \dots, 2^k - 1\}$, $t \in (0, \tau_k]$ and $k \geqslant 1$. Gluing these inequalities (for different ℓ and t) with the triangle inequality yields

$$W_2\left(\mu^k(t), \mu^k(s)\right) \leqslant \|V\|_{\mathrm{op}} R|t - s|$$

for any $t \in [0, T]$. Since $\mu^k(0) = \mu_0$ for any $k \geqslant 1$, and since $\mathcal{P}_2(\mathbb{R}^d)$ is the completion of $\mathcal{P}_c$ for the Wasserstein distance W_2, the Arzelà-Ascoli theorem implies the existence of a subsequence uniformly converging to some $\mu^* \in C^0([0,T]; \mathcal{P}_2(\mathbb{R}^d))$. Since for any $t \in [0,T]$ the curves $\mu^k(t)$ have their support enclosed in $B(0,R)$ for any $k \geqslant 1$, we even deduce that $\mu^* \in C^0_{\mathrm{co}}(\mathbb{R}, \mathcal{P}_c(\mathbb{R}^d))$. Note moreover that $\mu^*(0) = \mu_0$ and that

$$W_2(\mu^*(t), \mu^*(s)) \leqslant \|V\|_{\mathrm{op}} R|t - s|$$

for any $t, s \in [0, T]$.

The fact that μ^* is a solution of (6.1) follows exactly from the same computations as in [PRT15, p. 4711-4712], starting from (A.2) therein. We do not reproduce here this argument since the computations are the same word for word. The fact that for any $T > 0$ we have $\sup_{t\in[0,T]} W_1(\mu^*(t), \mu^k(t)) \to 0$ as $k \to +\infty$, which is instrumental in [PRT15, p. 4711-4712], follows in our case from the left-hand-side of (6.8).

Part 2: Uniqueness. Regarding uniqueness, we proceed as follows. We first recall the following estimate from [PR16, Proposition 4]. Let $p \geqslant 1$, let $t \geqslant 0$, let $v, w \in C^{0,1} \cap L^\infty([0,t] \times \mathbb{R}^d; \mathbb{R}^d)$ (both with Lipschitz constant $L > 0$, say), and let $\mu, \nu \in \mathcal{P}_c(\mathbb{R}^d)$. Then

$$W_p\left((\Phi_v^t)_\# \mu, (\Phi_w^t)_\# \nu\right) \leqslant e^{\frac{p+1}{p} Lt} W_p(\mu, \nu) + \frac{e^{\frac{Lt}{p}}(e^{Lt} - 1)}{L} \|v - w\|_{L^\infty([0,t]\times\mathbb{R}^d;\mathbb{R}^d)}. \tag{6.11}$$

Now assume that there are two solutions μ and ν of (6.1), with a spatial support that is locally bounded in time, and having the same initial condition. Define $v(t,x) := \mathcal{X}[\mu(t)](x)$ and $w(t,x) := \mathcal{X}[\nu(t)](x)$. Also set

$$t_0 := \inf\{t \geqslant 0 \colon W_2(\mu(t), \nu(t)) \neq 0\},$$

and assume that $t_0 \neq +\infty$. Fix $T > t_0$ and take $R > 0$ such that μ_t and ν_t are supported in $B(0,R)$ for any $t \in [0,T]$. Using (6.11) with $p = 2$, and setting $C_2(R) := 2\|Q^\top K\|_{\mathrm{op}} \|V\|_{\mathrm{op}} R^2$ in (6.4), we find

$$\begin{aligned} W_2(\mu(t_0+s), \nu(t_0+s)) &\leqslant e^{2C_2(R)s} W_2(\mu(t_0), \nu(t_0)) \\ &\quad + e^{C_2(R)s} \frac{e^{C_2(R)s} - 1}{C_2(R)} \sup_{\tau \in [t_0, t_0+s]} \|v(\tau,\cdot) - w(\tau,\cdot)\|_{L^\infty(\mathbb{R}^d)}. \end{aligned}$$

Choose $s > 0$ sufficiently small so that $e^{C_2(R)s} - 1 \leqslant 2C_2(R)s$. Then, by virtue of (6.5) and the fact that $W_2(\mu(t_0), \nu(t_0)) = 0$, we deduce

$$W_2(\mu(t_0+s), \nu(t_0+s)) \leqslant 2se^{C_2(R)s} \sup_{\tau \in [t_0, t_0+s]} W_2(\mu(\tau), \nu(\tau)). \tag{6.12}$$

We choose $s' > 0$ satisfying both $e^{C_2(R)s'} - 1 \leqslant 2C_2(R)s'$ and $2s'e^{C_2(R)s'} < 1$. Applying (6.12) to every $s \in [0, s']$ we obtain

$$\begin{aligned} \sup_{s \in [0,s']} W_2(\mu(t_0+s), \nu(t_0+s)) &\leqslant 2s'e^{C_2(R)s'} \sup_{\tau \in [t_0, t_0+s']} W_2(\mu(\tau), \nu(\tau)) \\ &< \sup_{s \in [0,s']} W_2(\mu(t_0+s), \nu(t_0+s)), \end{aligned}$$

which is a contradiction. Therefore $\mu(t) \equiv \nu(t)$ for any $t \geqslant 0$, which proves uniqueness, as desired.

Part 3: Stability. We do not detail the proof of estimate (6.9), which is very similar to the proof of (2.3) in Theorem 2.3 of [PRT15]: it follows from (6.11) with $p = 2$, and the argument after (A.7) in [PRT15], with W_2 instead of W_1. See also [PR13, Theorem 3]. □

We conclude this section with the proof of Proposition 6.2, which follows as a corollary of the above derivations.

Proof of Proposition 6.2. We first show existence. We apply Proposition 6.6 with $\mu_0 := \frac{1}{n}\sum_{j=1}^n \delta_{x_i^0}$, which in turn yields a solution $\mu(t)$ to (6.1). Following the proof of Proposition 6.6, we also know that this solution satisfies $\mu(t) = (\Phi^t_{\mathcal{X}[\mu(t)]})_\# \mu_0$ for any $t \in \mathbb{R}$, and the vector field $\mathcal{X}[\mu(t)]$ satisfies the assumptions of the Cauchy-Lipschitz theorem. In particular, $\mu(t)$ is of the form $\mu(t) = \frac{1}{n}\sum_{j=1}^n \delta_{x_i(t)}$ for some Lipschitz curves $\mathbb{R} \ni t \mapsto x_i(t)$, for $i \in [n]$. Then $t \mapsto \mu(t) = \frac{1}{n}\sum_{j=1}^n \delta_{x_i(t)}$ is a solution to the Cauchy problem (6.1)-(6.2) in the sense of Definition 6.4.

Secondly, we show uniqueness. Suppose that $\mathbf{X}(t) = (x_1(t), \dots, x_n(t))$ and $\mathbf{X}^*(t)$ are two Lipschitz solutions to (1.1), with the same initial conditions. Then for a.e. $t \geqslant 0$, using the equation (1.1) and the fact that the attention matrix coefficients $P_{ij}(t)$ defined in (1.2) belong to $[0,1]$, we obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\max_{i\in[n]}\|x_i(t)\|^2 \leqslant \|V\|_{\mathrm{op}}\max_{i\in[n]}\|x_i(t)\|^2$$

(and analogously for $x_i^*(t)$). Using Grönwall's inequality, we deduce the existence of two constants $c_1, c_2 > 0$ such that for any $t > 0$ and for any $i \in [n]$, $\|x_i(t)\|$ and $\|x_i^*(t)\|$ are bounded from above by $c_1 e^{c_2 t}$. It then follows that the empirical measures $\mu(\cdot) = \frac{1}{n}\sum_{j=1}^n \delta_{x_i(\cdot)}$ and $\mu^*(\cdot) = \frac{1}{n}\sum_{j=1}^n \delta_{x_i^*(\cdot)}$ belong to $C^0_{\mathrm{co}}(\mathbb{R}, \mathcal{P}_c(\mathbb{R}^d))$. Moreover, they satisfy $\mu(t) = (\Phi^t_{\mathcal{X}[\mu(t)]})_{\#}\mu_0$ and $\mu^*(t) = (\Phi^t_{\mathcal{X}[\mu^*(t)]})_{\#}\mu_0$ and are thus solutions to (6.1). Using the uniqueness result of Proposition 6.6, we obtain that $\mu = \mu^*$ which concludes the proof. □

7. Proof of Theorem 2.1

Throughout this section we focus on the following dynamics:

$$\dot{x}_i(t) = \sum_{j=1}^n \left(\frac{e^{\langle x_i(t), x_j(t)\rangle}}{\sum_{k=1}^n e^{\langle x_i(t), x_k(t)\rangle}}\right) x_j(t). \tag{7.1}$$

Note that for $d = 1$, the dot products in (7.1) are just multiplications of scalars.

We begin with the following observation, which holds for any $d \geqslant 1$.

Lemma 7.1. *For any $x_1, \dots, x_n \in \mathbb{R}^d$, the function $f : \mathbb{R}^d \to \mathbb{R}$ defined by*

$$f : x \mapsto \log\left(\sum_{j=1}^n e^{\langle x, x_j\rangle}\right) \tag{7.2}$$

is convex.

Proof. Using the elementary inequality $(a+b) \geqslant 2(ab)^{\frac{1}{2}}$ for any $a, b \geqslant 0$, we have

$$\begin{aligned}
\exp(f(x) + f(y)) &= \left(\sum_{j=1}^n \exp(\langle x, x_j\rangle)\right)\left(\sum_{j=1}^n \exp(\langle y, x_j\rangle)\right) \\
&= \frac{1}{2}\sum_{j=1}^n\sum_{k=1}^n \Big[\exp\left(\langle x, x_j\rangle + \langle y, x_k\rangle\right) + \exp(\langle x, x_k\rangle + \langle y, x_j\rangle)\Big] \qquad (7.3)\\
&\geqslant \sum_{j=1}^n\sum_{k=1}^n \exp\left(\left\langle \frac{x+y}{2}, x_j + x_k\right\rangle\right) \qquad (7.4)\\
&= \exp\left(2f\left(\frac{x+y}{2}\right)\right).
\end{aligned}$$

Taking the log on both sides yields the statement. □

The following lemma also holds for any $d \geqslant 1$.

Lemma 7.2. *Let $\mathbb{R} \ni t \mapsto \{x_i(t)\}_{i\in[n]}$ be a solution to (7.1). Then for any $i, j \in [n]$, the map $\mathbb{R} \ni t \mapsto \|x_i(t) - x_j(t)\|$ is non-decreasing.*

Proof. The dynamics (7.1) can be equivalently written as

$$\dot{x}_i(t) = \nabla f(x_i(t))$$

where f is as in (7.2). By convexity of f (Lemma 7.1),

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|x_i(t) - x_j(t)\|^2 &= \langle \dot{x}_i(t) - \dot{x}_j(t), x_i(t) - x_j(t)\rangle \\ &= \langle \nabla f(x_i(t)) - \nabla f(x_j(t)), x_i(t) - x_j(t)\rangle \geqslant 0,\end{aligned}$$

as desired. □

We now present the proof of Theorem 2.1, which assumes $d = 1$. We recall that in the statement, V is a positive scalar, but by reparametrizing time we may assume that $V = 1$, so the $1d$ dynamics under consideration is really given by (7.1). Also, to ease notations we focus on $QK = 1$, but the proof adapts straightforwardly to the setting $QK > 0$ assumed in the statement of Theorem 2.1.

As seen in Section 7.1, it is not difficult to prove the convergence of the coefficients $P_{ij}(t)$ of the attention matrix for indices $i \in [n]$ for which $x_i(t)$ becomes unbounded as $t \to +\infty$. This is the case for at least $n-1$ of the particles $x_i(t)$ (Lemma 7.6). But should one particle $x_i(t)$ remain bounded, proving the convergence of $P_{ij}(t)$ for $j \in [n]$ is slightly tedious (Section 7.2). Since $d = 1$, up to relabeling, we can order the initial collection of particles (which, we recall, are assumed distinct):

$$x_1(0) < \ldots < x_n(0).$$

We set

$$c := \min_{i\in[n-1]} |x_{i+1}(0) - x_i(0)|. \tag{7.5}$$

According to Lemma 7.2, we have $|x_i(t) - x_j(t)| \geqslant c$ for any $i \neq j$ and any $t \geqslant 0$. In particular, particles never "collide".

7.1. **Results about unbounded particles.** In this section we gather several results concerning the indices i corresponding to particles $x_i(t)$ which are not uniformly bounded in time. In particular, in Lemma 7.4 we show that for such indices i, $P_{ij}(t)$ converges toward 0 or 1 for any $j \in [n]$.

Lemma 7.3. *Let $A > 0$ denote the unique positive real number satisfying $A^2 = n^2 \exp(-A^2)$. If $x_n(t_0) > A$ for some time $t_0 \geqslant 0$, then there exists $c_1 > 0$ such that $x_n(t) \geqslant c_1 e^t$ for any sufficiently large $t > 0$. Similarly, if $x_1(t_0) < -A$ for some $t_0 \geqslant 0$, then $x_1(t) \leqslant -c_1 e^t$ for any sufficiently large $t > 0$.*

Proof. The two cases are symmetric since the evolution (7.1) commutes with the involution of $(\mathbb{R}^d)^n$ given by $(x_1, \ldots, x_n) \mapsto (-x_1, \ldots, -x_n)$. We thus focus on the case $x_n(t_0) > A$.

If $x_n(t) \geqslant 0$ for some $t \geqslant 0$, then

$$\dot{x}_n(t) = \sum_{j=1}^{n} \left(\frac{e^{x_n(t)(x_j(t)-x_n(t))}}{\sum_{k=1}^{n} e^{x_n(t)(x_k(t)-x_n(t))}} \right) x_j(t) \tag{7.6}$$

$$\geqslant \frac{x_n(t)}{1+(n-1)e^{-cx_n(t)}} + \sum_{\{j\in[n]\colon\, x_j(t)<0\}} e^{x_n(t)(x_j(t)-x_n(t))} x_j(t) \tag{7.7}$$

$$\geqslant \frac{x_n(t)}{1+(n-1)e^{-cx_n(t)}} - n\frac{e^{-x_n(t)^2}}{x_n(t)} \tag{7.8}$$

$$\geqslant \frac{x_n(t)}{n} - n\frac{e^{-x_n(t)^2}}{x_n(t)}. \tag{7.9}$$

We provide some detail on the above sequence of inequalities. First of all, to pass from (7.6) to (7.7), we use

$$e^{x_n(t)(x_k(t)-x_n(t))} \leqslant e^{-cx_n(t)}$$

for $j = n$ and for any $k \in [n]$ (which holds by virtue of (7.5)), combined with the fact that

$$\sum_{k=1}^{n} e^{x_n(t)(x_k(t)-x_n(t))} \geqslant 1$$

for all indices j such that $x_j(t) < 0$. To pass from (7.7) to (7.8), we use $e^{x_n(t)z} z \geqslant -\frac{1}{x_n(t)}$, which holds for any $z \leqslant 0$.

For any $B > A$, we clearly have

$$\frac{B}{n} - n\frac{e^{-B^2}}{B} > 0.$$

We then deduce from (7.8) and the fact that $x_n(t_0) > A$ that $x_n(t) \to +\infty$ as $t \to +\infty$. Moreover due to the fact that the expression in (7.9) is bounded from below by $\frac{x_n(t)}{2n}$ whenever $x_n(t)$ is sufficiently large, we deduce that

$$x_n(t) \geqslant c_0 e^{\frac{t}{2n}}$$

for any sufficiently large $t > 0$.

Coming back to (7.8), we find that for sufficiently large $t > 0$,

$$\dot{x}_n(t) \geqslant x_n(t) \left(\frac{1}{1+(n-1)e^{-cc_0 e^{\frac{t}{2n}}}} - e^{-c_0^2 e^{\frac{t}{n}}} \right).$$

This implies that

$$\frac{\mathrm{d}}{\mathrm{d}t} \log(x_n(t)) \geqslant 1 - O\left(e^{-\frac{t}{3n}}\right),$$

whence

$$\log(x_n(t)) \geqslant t + O(1)$$

for sufficiently large $t > 0$, as desired. □

Here and in what follows, δ_{jk} denotes the Kronecker symbol.

Lemma 7.4. *If $i \in [n]$ is such that $x_i(t)$ is not uniformly bounded with respect to $t > 0$, then $x_i(t)$ converges to either $-\infty$ or $+\infty$ as $t \to +\infty$. Moreover,*

(1) if $x_i(t) \to +\infty$, then for any $j \in [n]$, $P_{ij}(t)$ converges to δ_{nj} as $t \to +\infty$, with doubly exponential rate.

(2) if $x_i(t) \to -\infty$, then for any $j \in [n]$, $P_{ij}(t)$ converges to δ_{1j} as $t \to +\infty$, with doubly exponential rate.

Proof. We assume that $x_i(t)$ is not uniformly bounded with respect to $t > 0$. Without loss of generality, we assume that there exists a sequence of positive times $\{t_k\}_{k=1}^{+\infty}$ with $t_k \to +\infty$ such that $x_i(t_k) \to +\infty$. Necessarily, $x_n(t_k) \to +\infty$. We notice that if $x_i(t) > 0$ for some $t \geqslant 0$, then, arguing as in (7.6)–(7.7)–(7.8), we have

$$\dot{x}_i(t) = \sum_{j=1}^{n} \left(\frac{e^{x_i(t)(x_j(t)-x_n(t))}}{\sum_{k=1}^{n} e^{x_i(t)(x_k(t)-x_n(t))}} \right) x_j(t) \geqslant \frac{x_n(t)}{n} - \frac{n}{x_i(t)} e^{-x_i(t)x_n(t)}. \quad (7.10)$$

For sufficiently large integers $k \geqslant 1$, from (7.10) we get $\dot{x}_i(t_k) > 0$ and $\dot{x}_n(t_k) > 0$. But as x_i and x_n increase, the lower bound in (7.10) becomes larger. It follows that

$$\dot{x}_i(t) \geqslant \frac{x_n(t)}{2n} \geqslant \frac{x_i(t)}{2n}$$

for sufficiently large t, implying that $x_i(t) \to +\infty$ with exponential rate as $t \to +\infty$.

We now prove point 1. regarding $P(t)$. We assume that $x_i(t) \to +\infty$ as $t \to +\infty$. In this case, for $j \neq n$ (namely $j \in [n-1]$),

$$P_{ij}(t) = \frac{e^{x_i(t)x_j(t)}}{\sum_{k=1}^{n} e^{x_i(t)x_k(t)}} \leqslant e^{x_i(t)(x_j(t)-x_n(t))} \leqslant e^{-cx_i(t)},$$

thus $P_{ij}(t)$ converges to 0 as $t \to +\infty$ (with doubly exponential rate). Consequently, we also deduce that

$$P_{in}(t) = 1 - \sum_{j=1}^{n-1} P_{ij}(t)$$

converges to 1, also with doubly exponential rate, as $t \to +\infty$.

The case where $x_i(t) \to -\infty$ is symmetric. This concludes the proof. □

Our last result is useful in the next section.

Lemma 7.5. *For any $i \in [n]$ such that $x_i(t)$ is not uniformly bounded with respect to $t > 0$, there exists some $\gamma_i \in \mathbb{R}$, $\gamma_i \neq 0$ such that $x_i(t) = \gamma_i e^t + o(e^t)$ as $t \to +\infty$.*

Proof. Without loss of generality we assume that $x_i(t) \to +\infty$ as $t \to +\infty$. For $j \neq n$, we find

$$P_{ij}(t) = \frac{e^{x_i(t)x_j(t)}}{\sum_{k=1}^{n} e^{x_i(t)x_k(t)}} = \frac{e^{x_i(t)(x_j(t)-x_n(t))}}{\sum_{k=1}^{n} e^{x_i(t)(x_k(t)-x_n(t))}} \leqslant e^{-cx_i(t)}.$$

Consequently,

$$P_{in}(t) \geqslant 1 - ne^{-cx_i(t)}.$$

Therefore, using Lemma 7.3 and the fact that $x_i(t) \geqslant b_i e^{\frac{t}{2n}}$ for some $b_i > 0$ (thanks to (7.10)), we gather that

$$\begin{aligned} \dot{x}_i(t) &\geqslant \left(1 - ne^{-cx_i(t)}\right) x_n(t) - ne^{-cx_i(t)} c_1 e^t \\ &\geqslant \left(1 - ne^{-cb_i e^{\frac{t}{2n}}}\right) x_n(t) - ne^{-cb_i e^{\frac{t}{2n}}} c_1 e^t \end{aligned} \quad (7.11)$$

for some $c_1 > 0$ independent of t. We also notice that due to (7.1), $\dot{x}_i(t) \leqslant x_n(t)$. Using (7.11), firstly for $i = n$, together with the trivial upper bound $x_n(t) \leqslant Ce^t$ for some $C > 0$ independent of t (immediately seen from (7.1)), we obtain

$$\dot{x}_n(t) = x_n(t)\left(1 + o\left(e^{-cb_i e^{\frac{t}{3n}}}\right)\right)$$

as $t \to +\infty$, which yields

$$x_n(t) = \gamma_n e^t + o(e^t)$$

for some $\gamma_n > 0$. Now using (7.11) for the index i, we gather that

$$\dot{x}_i(t) = x_n(t) + o\left(e^{-cb_i e^{\frac{t}{3n}}}\right),$$

and so we deduce that

$$x_i(t) = \gamma_n e^t + o(e^t).$$

Similarly, if $x_i(t) \to -\infty$, then $x_i(t) = \gamma_1 e^t + o(e^t)$. This proves Lemma 7.5 (and shows that $\gamma_i \in \{\gamma_1, \gamma_n\}$). □

7.2. **Results about bounded particles.** In this section we collect results concerning particles which remain uniformly bounded in time. The following lemma entails that there can be at most one particle with this property.

Lemma 7.6. *Consider*

$$\mathscr{B} := \Big\{ i \in [n] \colon x_i(\cdot) \in L^\infty([0, +\infty)) \Big\}.$$

Then $\#\mathscr{B} \in \{0, 1\}$.

Proof. We first prove that either $x_1(t) \to -\infty$ or $x_n(t) \to +\infty$ as $t \to +\infty$. By contradiction, if this is not the case, then by Lemma 7.3, $(x_1(t), \dots, x_n(t)) \in [-A, A]^n$ for any $t \geqslant 0$. We denote by $\mathscr{I}$ the set of configurations $(x_1^*, \dots, x_n^*) \in [-A, A]^n$ such that $|x_i^* - x_j^*| \geqslant |x_i(0) - x_j(0)| > 0$ for any distinct $i, j \in [n]$. For any $\mathbf{X}^* = (x_1^*, \dots, x_n^*) \in \mathscr{I}$, the function f defined in (7.2) (with anchor points given by $\mathbf{X}^*$) is strictly convex—the equality in the inequality between (7.3) and (7.4) is never achieved. Therefore, the proof of Lemma 7.2 shows that if $\mathbf{X}^*$ is seen as an initial datum for the dynamics (7.1), then

$$v(\mathbf{X}^*) := \frac{\mathrm{d}}{\mathrm{d}t}_{|t=0} |x_1^*(t) - x_n^*(t)| > 0.$$

Since $\mathscr{I}$ is compact, $v_0 := \inf_{\mathbf{X}^* \in \mathscr{I}} v(\mathbf{X}^*) > 0$. Hence, $t \mapsto |x_1(t) - x_n(t)|$ grows at least linearly, which is a contradiction.

We may therefore assume without loss of generality that $x_1(t) \to -\infty$ as $t \to +\infty$. We prove that $x_n(t)$ converges to either $-\infty$, or 0, or $+\infty$, as $t \to +\infty$. We assume in the sequel that $x_n(t)$ does not converge to $-\infty$ or 0. For any $i \in [n]$, if there exists $\varepsilon > 0$ and a sequence of positive times $\{s_k\}_{k=1}^{+\infty}$ tending to $+\infty$ such that $x_i(s_k) \leqslant -\varepsilon$, then it follows from (7.10) that $x_i(t) \to -\infty$. Therefore, by our assumptions, we have $\liminf_{t \to +\infty} x_n(t) \geqslant 0$. Also, since $x_n(t) \nrightarrow 0$, there exists $\varepsilon > 0$ and a sequence of positive times $\{t_k\}_{k=1}^{+\infty}$ tending to $+\infty$ such that $x_n(t_k) \geqslant \varepsilon$ for any integer $k \geqslant 1$. For any $t \geqslant 0$ such that $x_n(t) \geqslant \varepsilon$, we introduce the set of indices

$$N(t) = \{i \in [n] \colon x_i(t) < 0\},$$

and we write

$$\dot{x}_n(t) \geqslant \frac{e^{x_n(t)^2} x_n(t)}{\sum_{k=1}^{n} e^{x_n(t)x_k(t)}} + \frac{\sum_{j\in N(t)} e^{x_j(t)x_n(t)} x_j(t)}{\sum_{k=1}^{n} e^{x_n(t)x_k(t)}} \geqslant \frac{\varepsilon}{n} + \frac{1}{e^{\varepsilon^2}} \sum_{j\in N(t)} e^{\varepsilon x_j(t)} x_j(t). \quad (7.12)$$

According to Lemma 7.4, any point $x_i(t)$ which takes negative values for arbitrarily large times and does not converge to $-\infty$ has to converge to 0. Therefore, the second term in the lowermost bound in (7.12) is lower bounded by $-\frac{\varepsilon}{2n}$ for sufficiently large t. All in all, we gather that $\dot{x}_n(t) \geqslant \frac{\varepsilon}{2n}$ and $x_n(t)$ converges to $+\infty$ as $t \to +\infty$. If it converges to 0, then necessarily $x_{n-1}(t) \to -\infty$ by combining Lemma 7.2 with Lemma 7.4. This proves Lemma 7.6 in this case.

From now on we assume that $x_n(t) \to +\infty$. Using (7.10) we see that if there exists $\varepsilon > 0$ such that $x_i(t) > \varepsilon$ for an unbounded sequence of times t, then $x_i(t) \to +\infty$. The same is true symmetrically when $x_i(t) < -\varepsilon$ for an unbounded sequence of times t. Thus if $i \in \mathscr{B}$, necessarily $x_i(t) \to 0$. By Lemma 7.2 this can be true for at most one index i, which concludes the proof of Lemma 7.6. □

If $\mathscr{B} = \varnothing$, Theorem 2.1 follows from Lemma 7.4. From now on, we assume that $\#\mathscr{B} = 1$, and we denote by $i_0 \in [n]$ its unique element. We distinguish two cases: either $i_0 \in \{1, n\}$ (Lemma 7.7), or $i_0 \notin \{1, n\}$ (Lemma 7.8).

Lemma 7.7. *If $x_n(t)$ is bounded as $t \to +\infty$, then $P_{nn}(t) \to 1$, and $P_{nj}(t) \to 0$ for any $j \in [n-1]$, as $t \to +\infty$. Similarly, if $x_1(t)$ is bounded as $t \to +\infty$, then $P_{11}(t) \to 1$, and $P_{1j}(t) \to 0$ for any $j \in [n-1]$, as $t \to +\infty$.*

Proof. The two cases ($x_n(\cdot)$ bounded or $x_1(\cdot)$ bounded) are symmetric since the evolution (7.1) commutes with the involution of $(\mathbb{R}^d)^n$ given by $(x_1, \dots, x_n) \mapsto (-x_1, \dots, -x_n)$. Whence, we only address the first one: we assume that $x_n(t)$ is bounded as $t \to +\infty$. We first notice that all particles $x_j(t)$ for $j \in [n-1]$ tend to $-\infty$ as $t \to +\infty$ due to Lemma 7.6. We now prove the following properties:

(1) $x_n(t) > 0$ for any sufficiently large t;

(2) $x_n(t) \to 0$ as $t \to +\infty$;

(3) for any $j \in [n-1]$, $P_{nj}(t) \to 0$ as $t \to +\infty$.

To prove point (1), we notice that for sufficiently large t, $x_i(t) \leqslant 0$ for any $i \in [n-1]$. If in addition $x_n(t) \leqslant 0$, then due to (7.1), all $x_i(t)$ ($i \in [n]$) remain negative and due to (7.1), $x_n(t) \to -\infty$ as $t \to +\infty$, which is a contradiction.

For point (2), we fix $\varepsilon > 0$, and set

$$\mathsf{T}_\varepsilon^+ := \{t \geqslant 0 \colon x_n(t) \geqslant \varepsilon\}.$$

We prove that if T_ε^+ is unbounded, then $x_n(t) \to +\infty$ as $t \to +\infty$, which is a contradiction. As a consequence, T_ε^+ is bounded for any $\varepsilon > 0$, which implies (in conjunction with point 1.) that $x_n(t) \to 0$ as $t \to +\infty$. So let us assume that T_ε^+ is unbounded. We notice that for any $\delta > 0$, if $t \in \mathsf{T}_\varepsilon^+$ is sufficiently large then

$$\left| e^{x_n(t)x_j(t)} x_j(t) \right| \leqslant \delta$$

for any $j \in [n-1]$ since $x_j(t) \to +\infty$ as $t \to +\infty$. Therefore,

$$\sum_{j=1}^{n} e^{x_n(t)x_j(t)} x_j(t) \geqslant e^{\varepsilon^2}\varepsilon - (n-1)\delta \geqslant 0,$$

where we took $\delta > 0$ sufficiently small for the last inequality to hold. Consequently,

$$\dot{x}_n(t) = \frac{\sum_{j=1}^{n} e^{x_n(t)x_j(t)} x_j(t)}{\sum_{j=1}^{n} e^{x_n(t)x_j(t)}} \geqslant \frac{e^{x_n(t)^2} x_n(t) - (n-1)\delta}{e^{x_n(t)^2} + n - 1}.$$

It is not difficult to see that this implies that $x_n(t) \to +\infty$ as $t \to +\infty$, which is a contradiction.

For point (3), we first notice that for any $j \neq n$, since $x_j(t) \to -\infty$,

$$\dot{x}_j(t) = \sum_{k=1}^{n} \left(\frac{e^{x_j(t)(x_k(t)-x_n(t))}}{\sum_{\ell=1}^{n} e^{x_j(t)(x_\ell(t)-x_n(t))}} \right) x_k(t) \leqslant \frac{x_1(t)}{n} + \frac{n}{\varepsilon} e^{-x_j(t)x_n(t)}.$$

Using Lemma 7.3, we deduce the existence of some $c_2 > 0$ such that

$$x_j(t) \leqslant -c_2 e^t$$

for any sufficiently large $t > 0$. We now prove that for any $j \neq n$,

$$x_j(t)x_n(t) - x_n(t)^2 \underset{t \to +\infty}{\longrightarrow} -\infty. \tag{7.13}$$

Due to the ordering of the particles, it is enough to prove (7.13) for $j = n-1$. Fix $j = n-1$ and $\kappa > 0$, and assume that

$$x_n(t)x_j(t) \geqslant x_n(t)^2 - \kappa$$

for some $t \geqslant 0$. Then, using the fact that

$$x_n(t)x_j(t) \geqslant x_n(t)x_k(t)$$

for any $k \in [n-2]$, we get

$$P_{nj}(t) \geqslant \frac{e^{x_j(t)x_n(t)}}{e^{x_n(t)^2} + (n-1)e^{x_n(t)x_j(t)}} \geqslant \varepsilon,$$

where $\varepsilon = \frac{1}{n+e^\kappa}$. We obtain

$$\dot{x}_n(t) \leqslant P_{nn}(t)x_n(t) + P_{nj}(t)x_j(t) \leqslant x_n(t) + \varepsilon x_j(t),$$

hence

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\big(x_n(t)(x_n(t) - x_j(t))\big) &= \dot{x}_n(t)(2x_n(t) - x_j(t)) - x_n(t)\dot{x}_j(t) \\
&\leqslant (x_n(t) + \varepsilon x_j(t))(2x_n(t) - x_j(t)) - x_n(t)\dot{x}_j(t) \\
&= -\varepsilon x_j(t)^2 + x_n(t)(2\varepsilon x_j(t) + 2x_n(t) - x_j(t) - \dot{x}_j(t)) \\
&\leqslant -\varepsilon x_j(t)^2 + x_n(t)(2x_n(t) - 2x_1(t)),
\end{aligned} \tag{7.14}$$

where in the last line we used the fact that $\dot{x}_j(t) \geqslant x_1(t)$, which is due to (7.1), and that $x_1(t) < x_j(t)$, which is due to the ordering of the particles. Since $x_j(t) \leqslant -c_2 e^t$ and $x_1(t) \geqslant -c_1 e^t$, the upper bound in (7.14) is negative if t is large enough. We therefore conclude that for any fixed κ, if there exist unbounded times t such that $x_n(t)x_j(t) \geqslant x_n(t)^2 - \kappa$, then $x_n(t)x_j(t) \geqslant x_n(t)^2 - \kappa$ for any t large enough. But

this is excluded since $x_n(t) > 0$ and $x_j(t) \to -\infty$ as $t \to +\infty$. This concludes the proof of (7.13), and the lemma follows by plugging this information into the definition of $P_{nj}(t)$. □

Lemma 7.8. *If $i_0 \notin \{1, n\}$ and $x_{i_0}(t)$ remains uniformly bounded in t, then for any $j \in [n-1]$, there exists some $\alpha_j \in [0,1]$ such that $P_{i_0 j}(t) \to \alpha_j$ as $t \to +\infty$.*

Proof. Assume that $i_0 \notin \{1, n\}$. Then $x_1(t) \to -\infty$ and $x_n(t) \to +\infty$ as $t \to +\infty$. Also, $x_{i_0}(t) \to 0$ due to (7.10). We write $x_{i_0}(t) = y_{i_0}(t)e^{-t}$. Since $\gamma_n > 0$ and $\gamma_1 < 0$, we notice that the function

$$g : \theta \mapsto \frac{\displaystyle\sum_{i \in [n]\setminus\{i_0\}} e^{\gamma_i \theta}\gamma_i}{1 + \displaystyle\sum_{i \in [n]\setminus\{i_0\}} e^{\gamma_i \theta}}$$

takes value $-\infty$ at $-\infty$, and $+\infty$ at $+\infty$, and has a positive derivative. Thus, it takes the value 0 exactly once, and we denote this point by θ_0. We prove that $y_{i_0}(t) \to \theta_0$ as $t \to +\infty$. We observe that

$$e^{x_{i_0}(t)^2} = 1 + o(1).$$

Using Lemma 7.5 we have

$$\begin{aligned}
\dot{y}_{i_0}(t) &= e^t \dot{x}_{i_0}(t) - y_{i_0}(t) \\
&= (P_{i_0 i_0}(t) - 1) y_{i_0}(t) \\
&\quad + e^{2t} \sum_{j \in [n]\setminus\{i_0\}} \left(\frac{e^{y_{i_0}(t)(\gamma_j + o(1))}}{1 + o(1) + \displaystyle\sum_{k \in [n]\setminus\{i_0\}} e^{y_{i_0}(t)(\gamma_k + o(1))}} \right) (\gamma_j + o(1)).
\end{aligned}$$

We recognize that the sum in the above expression is roughly equal to $g(y_{i_0})$. If the latter is not close to 0 for large times, then $\dot{y}_{i_0}(t)$ necessarily have a huge magnitude due to the e^{2t} factor, leading to a contradiction. Fix $\varepsilon > 0$. If $y_{i_0}(t) > \theta_0 + \varepsilon$ for some large time $t > 0$, then, noticing that

$$|y_{i_0}(t)| = e^t |x_{i_0}(t)| = o(e^t), \tag{7.15}$$

we get

$$\dot{y}_{i_0}(t) = o(e^t) + e^{2t}\Big(g\big(y_{i_0}(t) + o(y_{i_0}(t))\big)\Big).$$

But $g(y_{i_0}(t)) \geqslant \delta = \delta(\varepsilon)$, and hence

$$\dot{y}_{i_0}(s) \geqslant \frac{\delta}{2} e^{2s}$$

for any larger time $s \geqslant t$, which contradicts (7.15). We get a similar contradiction if $y_{i_0}(t) < \theta_0 - \varepsilon$ for large enough t. This concludes the proof that $y_{i_0}(t) \to \theta_0$. As a consequence, $x_{i_0}(t)x_i(t) \to \theta_0 \gamma_i$ for any $i \neq i_0$, and we deduce Lemma 7.8. □

7.3. **Concluding the proof of Theorem 2.1.**

Proof of Theorem 2.1. By Lemma 7.6, there is at most one index $i_0 \in [n]$ for which the particle $x_{i_0}(t)$ remains bounded for any $t > 0$. In turn, for any $i \in [n]\backslash\{i_0\}$, we may invoke Lemma 7.4 which entails that $P_{ij}(t)$ converges to either δ_{1j} or δ_{nj} as $t \to +\infty$ (with doubly exponential rate). And by ordering of the particles, for indices $i_1 \leqslant i_2$ different from i_0, and $P_{i_1 j}(t) \to \delta_{nj}$ then necessarily $P_{i_2 j}(t) \to \delta_{nj}$ as well. Consequently, all but at most one row of $P(t)$ converge to either $e_1 = (1, 0, \dots, 0)$ or $e_n = (0, \dots, 0, 1)$ as $t \to +\infty$. For the i_0-th row, we may invoke either Lemma 7.7 or Lemma 7.8. The former applies if $i_0 \in \{1, n\}$, and entails that the i_0-th row of $P(t)$ converges either to e_1 or e_n, while the latter applies if $i_0 \notin \{1, n\}$, and entails that the i_0-th row of $P(t)$ converges to some vector $\alpha \in \mathbb{R}^d$ with non-negative entries. Finally, since the i_0-th row of $P(t)$ has entries which sum up to 1, then so does α. These conclusions lead us to a final limit matrix P^* which has precisely the form indicated in Fig. 2 (namely, $P^* \in \mathcal{P}$), as desired. □

Remark 7.9 (Higher dimensions)**.** *The extension of Theorem 2.1 to $d \geqslant 2$ is not straightforward due to rare pathological situations. For example, suppose $d = 2$, $n = 2$, and the initial configuration $x_1(0) = (1, \varepsilon)$ and $x_2(0) = (1, -\varepsilon)$. One can check that $x_i(t) \to (1, 0)$ as $t \to +\infty$, for $i = 1, 2$, which means that a single cluster appears. However, the self-attention matrix converges toward the identity (which has rank 2). Therefore, it is not true in full generality that the rank of the limiting self-attention matrix is equal to the number of clusters as $t \to +\infty$, although we believe that the result is true for almost all initial conditions.*

8. Proofs of Theorems 3.1 and 8.5

In this section, we focus on proving the result in the case

$$V = I_d.$$

We also provide a full picture of the behavior of the dynamics in the case $V = -I_d$ in Section 8.2.

8.1. **Clustering towards vertices of convex polytopes: Theorem 3.1.** In this section, we prove Theorem 8.1—namely, we show that particles $\{z_i(t)\}_{i\in[n]}$ following the rescaled dynamics

$$\dot{z}_i(t) = \sum_{j=1}^{n} \left(\frac{e^{e^{2t}\langle Az_i(t), Az_j(t)\rangle}}{\sum_{k=1}^{n} e^{e^{2t}\langle Az_i(t), Az_k(t)\rangle}} \right) (z_j(t) - z_i(t)) \tag{8.1}$$

converge, as $t \to \infty$, toward points lying on the boundary of a particular convex polytope. In (8.1) we made use of the shorthand notation

$$A := \left(Q^\top K\right)^{\frac{1}{2}}. \tag{8.2}$$

The precise statement is the following:

Theorem 8.1. *Suppose $V = I_d$ and $Q^\top K \succ 0$. Then, for any initial datum $\{z_i(0)\}_{i\in[n]} \subset \mathbb{R}^d$, the solution to (8.1) is such that its convex hull $\operatorname{conv}\left(\{z_i(t)\}_{i\in[n]}\right)$ converges to some convex polytope $\mathcal{K} \subset \mathbb{R}^d$ as $t \to +\infty$. Furthermore, let $\mathcal{V} = \{v_1, \dots, v_m\}$ ($m \leqslant n$) denote the set of vertices of $\mathcal{K}$, and consider*

$$\mathcal{S} := \left\{ x \in \mathcal{K} \colon \|Ax\|^2 = \max_{j\in[m]} \langle Ax, Av_j\rangle \right\},$$

with A defined in (8.2). *Then $\mathcal{S}$ has finite cardinality, and $\mathcal{V} \subset \mathcal{S} \subset \partial\mathcal{K} \cup \{0\}$. Finally, for any $i \in [n]$ there exists a point $\bar{z} \in \mathcal{S}$ such that $z_i(t) \to \bar{z}$ as $t \to +\infty$. In particular, $z_i(t)$ converges either to some point on the boundary of $\mathcal{K}$, or to* 0.

8.1.1. *The convex hull is shrinking.* To prove Theorem 8.1, we begin with the following illustrative result.

Proposition 8.2. *Suppose $V = I_d$ and $Q^\top K \succ 0$. Then the solution $\{z_i(\cdot)\}_{i\in[n]}$ to* (8.1) *is such that $t \mapsto \mathrm{conv}(\{z_i(t)\}_{i\in[n]})$ is non-increasing in the sense of set-inclusion.*

Proof of Proposition 8.2. Fix $t > 0$ and let $H \subset \mathbb{R}^d$ be a closed half-space which does not contain any of the points $z_i(t)$. We define the map

$$\alpha : s \mapsto \min_{i\in[n]} \mathrm{dist}(z_i(s), H)$$

for $s \geqslant 0$. We claim that

$$\alpha \text{ is non-decreasing on } [t, +\infty). \tag{8.3}$$

Before proving (8.3), let us show how to conclude the proof of Proposition 8.2 using this claim. It follows from (8.3) that if $\mathrm{conv}(\{z_i(t)\}_{i\in[n]}) \cap H = \varnothing$, then $\mathrm{conv}(\{z_i(t')\}_{i\in[n]}) \cap H = \varnothing$ for any $t' \geqslant t$. Writing the convex set $\mathrm{conv}(\{z_i(t)\}_{i\in[n]})$ as

$$\mathrm{conv}(\{z_i(t)\}_{i\in[n]}) = \bigcap_{\substack{H' \text{ open half-space} \\ \mathrm{conv}(\{z_i(t)\}_{i\in[n]}) \subset H'}} H' = \bigcap_{\substack{H \text{ closed half-space} \\ \mathrm{conv}(\{z_i(t)\}_{i\in[n]}) \cap H = \varnothing}} \mathbb{R}^d \backslash H,$$

we get that $\mathrm{conv}(\{z_i(t')\}_{i\in[n]}) \subset \mathrm{conv}(\{z_i(t)\}_{i\in[n]})$ for any $t' \geqslant t$.

We now turn to the proof of the claim (8.3). Denoting by $\mathbf{n}$ the unit outer normal to H and by proj_H the orthogonal projection onto the closed set H, we have

$$\mathrm{dist}(x, H) = \langle x - \mathrm{proj}_H(x), \mathbf{n}\rangle.$$

If $t \mapsto x(t)$ is a differentiable curve, writing $\dot{x}(t) = \langle \dot{x}(t), \mathbf{n}\rangle \mathbf{n} + v(t)$ where $v(t) \in H$ we have $\frac{\mathrm{d}}{\mathrm{d}t}(\mathrm{proj}_H(x(t))) = v(t)$, whence

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{dist}(x(t), H) = \langle \dot{x}(t), \mathbf{n}\rangle. \tag{8.4}$$

Let $T > t$ denote the infimum of the times for which one of the points $z_i(t)$ lies in H. Now fix $s \in [t, T)$, and denote by $M(s)$ the set of indices $i \in [n]$ such that $\mathrm{dist}(z_i(s), H)$ is minimal. For $h \to 0$, we have

$$\begin{aligned}
\alpha(s+h) &= \min_{i\in M(s)} \mathrm{dist}(z_i(s+h), H) \\
&= \min_{i\in M(s)} \left(\mathrm{dist}(z_i(s), H) + h\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{dist}(z_i(s), H) + o(h) \right) \\
&= \alpha(s) + h \left(\min_{i\in M(s)} \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{dist}(z_i(s), H) \right) + o(h).
\end{aligned}$$

Consequently,

$$\frac{\mathrm{d}\alpha}{\mathrm{d}t}(s) = \min_{i\in M(s)} \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{dist}(z_i(s), H).$$

Moreover, for any $i \in M(s)$, one has

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{dist}(z_i(s), H) \overset{(8.4)}{=} \langle \dot{z}_i(s), \mathbf{n}\rangle = \sum_{j=1}^{n} P_{ij}(s)\langle z_j(s) - z_i(s), \mathbf{n}\rangle \geqslant 0,$$

where the last inequality comes from the fact that each term in the sum is non-negative, since $i \in M(s)$. This proves (8.3) (and, as a byproduct, that $T = +\infty$). □

The following fact immediately ensues.

Corollary 8.3. *For any $i \in [n]$ and $t \geqslant 0$, $z_i(t) \in \mathrm{conv}(\{z_i(0)\}_{i\in[n]})$. In particular, $z_i(\cdot)$ is uniformly bounded in time.*

8.1.2. *Proof of Theorem 8.1.*

Proof of Theorem 8.1. As a consequence of Proposition 8.2, the set $\mathrm{conv}(\{z_i(t)\}_{i\in[n]})$ converges as $t \to +\infty$ toward some convex polytope $\mathcal{K}$. In the remainder of the proof, we look to show that the particles $z_i(t)$ can in fact converge only to some well-distinguished points lying on the boundary of this polytope.

Step 1. The candidate set of limit points. We denote by $\mathcal{V} = \{v_1, \dots, v_m\}$ the set of vertices of $\mathcal{K}$. Writing any $x \in \mathcal{K}$ as a convex combination of these vertices: $x = \sum_{j=1}^m \alpha_j v_j$ for some weights $\alpha_j \geqslant 0$ with $\sum_{j=1}^m \alpha_j = 1$, we gather that

$$\|Ax\|^2 = \left\langle Ax, \sum_{j=1}^{m} \alpha_j Av_j \right\rangle = \sum_{j=1}^{m} \alpha_j \langle Ax, Av_j\rangle \leqslant \max_{j\in[m]} \langle Ax, Av_j\rangle. \tag{8.5}$$

Let $\mathcal{S} \subset \mathcal{K}$ denote the set of points $w \in \mathcal{K}$ such that

$$\|Aw\|^2 = \max_{j\in[m]} \langle Aw, Av_j\rangle. \tag{8.6}$$

The following holds—we postpone the proof to after that of the theorem.

Claim 1. *$\mathcal{V} \subset \mathcal{S}$. Moreover, if $0 \in \mathcal{K}$, then $0 \in \mathcal{S}$. Finally, $\mathcal{S} \subset \partial\mathcal{K} \cup \{0\}$, and $\mathcal{S}$ has finite cardinality.*

Now, for $\delta > 0$, we define the set $\mathcal{S}_\delta$ of points in $\mathcal{K}$ at distance at most δ from $\mathcal{S}$:

$$\mathcal{S}_\delta := \{x \in \mathcal{K} \colon \mathrm{dist}(x, \mathcal{S}) \leqslant \delta\}.$$

Since $\mathcal{S}$ is finite, there exists a sufficiently small $\delta_0 > 0$ such that for any $\delta \leqslant \delta_0$, the set $\mathcal{S}_\delta$ has $M := \#\mathcal{S}$ connected components, with any two of these connected components being separated by a distance of at least δ_0. Our goal is to prove that for any $i \in [n]$, and for sufficiently large t, the particle $z_i(t)$ remains in one of these connected components. In the sequel, we fix $i \in [n]$.

Step 2. $z_i(t)$ must grow if it is not already in $\mathcal{S}_\delta$. We now prove that there exists some $\gamma = \gamma(\mathcal{K}) > 0$ (depending only on the geometry of $\mathcal{K}$) such that for any $\delta \in (0, \delta_0]$, there exists $T(\delta) > 0$ such that if $t \geqslant T(\delta)$ and $z_i(t) \notin \mathcal{S}_\delta$, then

$$\frac{\mathrm{d}}{\mathrm{d}t}\|Az_i(t)\|^2 \geqslant \gamma\delta. \tag{8.7}$$

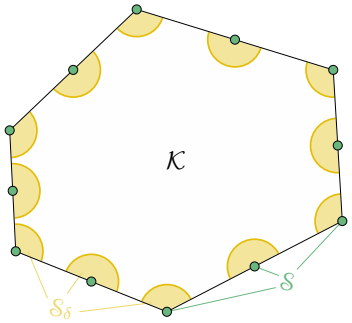


Figure 9. An example configuration of the sets $\mathcal{S}$ and $\mathcal{S}_\delta$ in $\mathbb{R}^2$. The set $\mathcal{S}$ consists of all green nodes along the boundary of $\partial\mathcal{K}$, while $\mathcal{S}_\delta$ is the union of all yellow "hemispheres". The latter are pairwise disjoint and are the connected components of $\mathcal{S}_\delta$, which we denote by $\mathcal{C}_k$, for $k \in [M]$.

To this end, we observe that

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|Az_i(t)\|^2 &= \langle A\dot{z}_i(t), Az_i(t)\rangle \\ &= \sum_{j=1}^n \left(\frac{e^{\langle Az_i(t), Az_j(t)\rangle e^{2t}}}{\sum_{k=1}^n e^{\langle Az_i(t), Az_k(t)\rangle e^{2t}}}\right)\langle A(z_j(t)-z_i(t)), Az_i(t)\rangle \\ &= \sum_{j=1}^n \underbrace{\left(\frac{e^{a_j(t)e^{2t}}}{\sum_{k=1}^n e^{a_k(t)e^{2t}}}\right)a_j(t)}_{:=b_j(t)} \end{aligned} \tag{8.8}$$

where we have set

$$a_j(t) := \langle A(z_j(t)-z_i(t)), Az_i(t)\rangle.$$

(To obtain the last equality in (8.8), divide both the numerator and the denominator by $e^{\|Az_i(t)\|^2 e^{2t}}$.) The following holds.

Claim 2. *There exists some constant $\gamma' = \gamma'(\mathcal{K}) > 0$ depending only on the geometry of $\mathcal{K}$ such that the following holds. Fix $\delta \in (0, \delta_0]$. There exists $T'(\delta) > 0$ such that if $t \geqslant T'(\delta)$ and $z_i(t) \notin \mathcal{S}_\delta$, then there exists $j \in [n]$ such that $a_j(t) \geqslant \gamma'\delta$.*

We postpone the proof of this claim to after that of the theorem. We seek to use this claim in obtaining a lower bound of $b_j(t)$ for any j, whenever δ is small enough and t is large enough. Since by Corollary 8.3, for any $j \in [n]$, $t \mapsto z_j(t)$ is uniformly bounded on $[0, +\infty)$, we gather that $a_j(\cdot) \in L^\infty(0, +\infty)$. So, we may set

$$\kappa := \max_{j\in[n]} \sup_{t\geqslant 0} |a_j(t)|.$$

Let $t \geqslant 0$ be fixed. We define

$$B(t) := \{j \in [n] \colon a_j(t) \geqslant 0\}.$$

We pick an index $j_0(t)$ maximizing $a_j(t)$, namely

$$j_0(t) \in \operatorname{argmax}_{j\in[n]} a_j(t).$$

Observe that $j_0(t) \in B(t)$ since $a_{j_0(t)}(t) \geqslant a_i(t) = 0$. Clearly

$$b_j(t) \geqslant 0 \qquad \text{for all } j \in B(t). \tag{8.9}$$

In fact, we also have

$$b_{j_0(t)}(t) \geqslant \frac{a_{j_0(t)}(t)}{n}. \tag{8.10}$$

Now suppose that $j \notin B(t)$; since $a_j(t) \geqslant -\kappa$, and

$$\frac{e^{a_j(t)e^{2t}}}{\sum_{k=1}^n e^{a_k(t)e^{2t}}} \leqslant \frac{1}{\sum_{k=1}^n e^{a_k(t)e^{2t}}} \leqslant e^{-a_{j_0(t)}e^{2t}},$$

we gather that

$$b_j(t) \geqslant -\kappa e^{-a_{j_0(t)}(t)e^{2t}} \qquad \text{for all } j \in [n]\backslash B(t). \tag{8.11}$$

Using (8.9), (8.10) and (8.11) in (8.8), we find

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|Az_i(t)\|^2 \geqslant \frac{a_{j_0(t)}(t)}{n} - \kappa n e^{-a_{j_0(t)}(t)e^{2t}}.$$

The above inequality along with Claim 2 lead us to deduce that there exists $T(\delta) > 0$ (possibly larger than $T'(\delta)$) such that (8.7) holds whenever $t \geqslant T(\delta)$, with $\gamma = \frac{\gamma'}{2n}$, as desired.

Step 3. $z_i(t)$ cannot circulate indefinitely between the connected components of $\mathcal{S}_\delta$. Since $z_i \in L^\infty([0,+\infty))$ by Corollary 8.3, from (8.1) we gather that $\dot{z}_i \in L^\infty([0,+\infty))$ as well. And since any two connected components of $\mathcal{S}_{\delta_0}$ are separated by a distance at least δ_0, we deduce that it takes a time at least

$$T_0 := \frac{\delta_0}{\|\dot{z}_i\|_{L^\infty([0,+\infty))}}$$

for z_i to go from one connected component of $\mathcal{S}_{\delta_0}$ to another one. Fix $\delta \in (0, \delta_0)$ such that

$$\delta < \frac{T_0\gamma\delta_0}{8R\|A\|_{\mathrm{op}}}, \tag{8.12}$$

where $R := \max_{j\in[n]} \|z_j\|_{L^\infty([0,+\infty))}$. Denote by

$$\mathscr{C}_1, \dots, \mathscr{C}_M$$

the connected components of $\mathcal{S}_\delta$, each of which being the intersection of $\mathcal{K}$ with a Euclidean ball of radius δ centered at some point of $\mathcal{S}$ (see Fig. 9). For any $k \in [M]$,

$$\sup_{x\in\mathscr{C}_k} \|Ax\|^2 - \inf_{x\in\mathscr{C}_k} \|Ax\|^2 \leqslant 4R\|A\|_{\mathrm{op}}\delta. \tag{8.13}$$

We introduce the following binary relation on $[M]$:

$$k \succ \ell \iff \inf_{x\in\mathscr{C}_k} \|Ax\|^2 > \sup_{x\in\mathscr{C}_\ell} \|Ax\|^2,$$

which is transitive. The underlying idea is the following: if t is sufficiently large, and if z_i starts from some connected component $\mathscr{C}_\ell$, then the only connected components $\mathscr{C}_k$ which z_i is able to visit later on are those for which $k \succ \ell$. This travel of z_i

has to stop after some time since $[M]$ is finite, $\succ$ is transitive, and for any ℓ, the relation $\ell \succ \ell$ does not hold.

Let $T = T(\delta)$ be as in Step 2. Suppose that $t_2 > t_1 \geqslant T$ and $k_1, k_2 \in [M]$ are distinct and such that $z_i(t_1) \in \mathscr{C}_{k_1}$, $z_i(t_2) \in \mathscr{C}_{k_2}$ and $z_i(t) \notin \mathcal{S}_\delta$ for any $t \in (t_1, t_2)$. Per Step 2 (more specifically, (8.7)),

$$\|Az_i(t_2)\|^2 \geqslant \|Az_i(t_1)\|^2 + T_0\gamma\delta_0.$$

Therefore using (8.13) twice and since δ is chosen as in (8.12), we gather that

$$\begin{aligned}\inf_{x\in\mathscr{C}_{k_2}} \|Ax\|^2 \geqslant \|Az_i(t_2)\|^2 - 4R\|A\|_{\mathrm{op}}\delta &\geqslant \|Az_i(t_1)\|^2 + T_0\gamma\delta_0 - 4R\|A\|_{\mathrm{op}}\delta \\ &\geqslant \inf_{x\in\mathscr{C}_{k_1}} \|Ax\|^2 + T_0\gamma\delta_0 - 4R\|A\|_{\mathrm{op}}\delta \\ &\geqslant \sup_{x\in\mathscr{C}_{k_1}} \|Ax\|^2 + T_0\gamma\delta_0 - 8R\|A\|_{\mathrm{op}}\delta \\ &> \sup_{x\in\mathscr{C}_{k_1}} \|Ax\|^2.\end{aligned} \tag{8.14}$$

Whence $k_2 \succ k_1$. We therefore deduce that there exist some $T' \geqslant T$ and $k \in [M]$ such that $z_i(t) \notin \mathcal{S}_\delta \backslash \mathscr{C}_k$ for any $t \geqslant T'$.

Step 4. Conclusion. To conclude, it remains to be shown that $z_i(t)$ stays in $\mathscr{C}_k$ for t large enough. For this, in addition to (8.12), we impose

$$\delta^{\frac{1}{4}} < \frac{\gamma T_0}{8R\|A\|_{\mathrm{op}}\delta_0}. \tag{8.15}$$

For $r > 0$, we denote by $\mathscr{C}_k^r$ the intersection of $\mathcal{K}$ with the closed Euclidean ball of radius δ^r having the same center as $\mathscr{C}_k$. In particular, $\mathscr{C}_k^1 = \mathscr{C}_k$. If, after time T', z_i travels from $\mathscr{C}_k$ to the complement of $\mathscr{C}_k^{\frac{1}{4}}$, it spends a time at least

$$\frac{(\delta^{\frac{1}{4}} - \delta^{\frac{1}{2}})}{\|\dot{z}_i\|_{L^\infty([0,+\infty))}}$$

in $\mathscr{C}_k^{\frac{1}{4}} \backslash \mathscr{C}_k^{\frac{1}{2}}$. Per Step 2 (used with $\delta^{\frac{1}{2}}$), $\|Az_i\|^2$ has to increase by at least

$$\frac{\gamma\delta^{\frac{1}{2}}\left(\delta^{\frac{1}{4}} - \delta\right)}{\|\dot{z}_i\|_{L^\infty([0,+\infty))}} \geqslant \frac{\gamma\delta^{\frac{3}{4}}}{2\|\dot{z}_i\|_{L^\infty([0,+\infty))}} > 4R\|A\|_{\mathrm{op}}\delta \tag{8.16}$$

during this travel (the last inequality in (8.16) stems from (8.15)). This implies that z_i cannot reenter $\mathscr{C}_k$ after having reached the boundary of $\mathscr{C}_k^{\frac{1}{4}}$, due to (8.13). Thus $z_i(t) \notin \mathcal{S}_\delta$ for any sufficiently large t, which is impossible due to Step 2 and the uniform boundedness of $t \mapsto \|Az_i(t)\|$. Hence, for sufficiently large t, $z_i(t) \in \mathscr{C}_k^{\frac{1}{4}}$. Since δ may be chosen arbitrarily small, this concludes the proof of Theorem 8.1. □

8.1.3. *Proving Claims 1 and 2.* We now address the proofs of the two claims which were instrumental in what precedes (along with a sketch of the proof of $\mathcal{V} \subset \mathcal{S}$, as implied).

Proof of Claim 1. The fact that $0 \in \mathcal{S}$ if $0 \in \mathcal{K}$ is immediate. We now show that $\mathcal{S}$ is finite and $\mathcal{S} \subset \partial\mathcal{K} \cup \{0\}$. Let $w \in \mathcal{S}\backslash\{0\}$. As

$$w = \sum_{j=1}^{m} \alpha_j v_j$$

for some $\alpha_j \geqslant 0$ with $\sum_{j=1}^m \alpha_j = 1$, and since (8.6) holds by definition, it follows that $\alpha_j = 0$ for any j not attaining the maximum in (8.6). Let $\mathsf{I} \subset [m]$ denote the set of all such indices. We have

$$w = \sum_{j \in \mathsf{I}} \alpha_j v_j$$

with $\|Aw\|^2 = \langle Aw, Av_j \rangle$ for any $j \in \mathsf{I}$. Whence w is the orthogonal projection onto $\operatorname{span}\{v_j\}_{j \in \mathsf{I}}$ with respect to $\langle A\cdot, A\cdot \rangle$. This yields $\mathcal{S} \subset \partial\mathcal{K}$. Moreover, since for each subset $\mathsf{I} \subset [m]$ there exists a unique such projection w, $\mathcal{S}$ is finite. □

Sketch of proof of $\mathcal{V} \subset \mathcal{S}$. We notice that for any $i \in [n]$ and for t large enough, we have

$$\dot{z}_i(t) = \sum_{j=1}^n \left(\frac{e^{e^{2t}\langle Az_i(t), Az_j(t) \rangle}}{\sum_{k=1}^n e^{e^{2t}\langle Az_i(t), Az_k(t) \rangle}} \right) (z_j(t) - z_i(t)) \tag{8.17}$$

$$\approx \sum_{j \in M_i(t)} \left(\frac{e^{e^{2t}\langle Az_i(t), Az_j(t) \rangle}}{\sum_{k=1}^n e^{e^{2t}\langle Az_i(t), Az_k(t) \rangle}} \right) (z_j(t) - z_i(t)), \tag{8.18}$$

where $M_i(t)$ is the subset of $[n]$ containing all indices j such that

$$\max_{k \in [n]} \langle Az_i(t), Az_k(t) \rangle - \langle Az_i(t), Az_j(t) \rangle \leqslant e^{-t}$$

(all other terms in the sum (8.17) are negligible). Due to the convergence of $\operatorname{conv}(\{z_i(t)\}_{i \in [n]})$ toward $\mathcal{K}$, we also know that for t large enough,

- all the points $z_i(t)$ are contained in a small neighborhood of $\mathcal{K}$,
- near any element of $\mathcal{V}$, there exists some particle $z_i(t)$.

Assume, for the sake of contradiction, that there exists a vertex $v_j \in \mathcal{V}$ such that $v_j \notin \mathcal{S}$. Set $\mathcal{C} := \operatorname{conv}(\{v_i\}_{i \in [m] \setminus \{j\}})$. In particular, $\operatorname{dist}(v_j, \mathcal{C}) > 0$ since v_j is a vertex of $\mathcal{K}$. If $\mathsf{I} \subset [n]$ denotes the set of indices i such that $z_i(t)$ lies near v_j, then $M_i(t) \cap \mathsf{I} = \varnothing$ for any $i \in \mathsf{I}$, since $v_j \notin \mathcal{S}$. For $i \in \mathsf{I}$, using (8.18), we find that $\operatorname{dist}(z_i(t), \mathcal{C})$ decays as $t \to +\infty$ as long as $i \notin M_i(t)$—indeed, (8.18) implies that $z_i(t)$ is attracted by $\mathcal{C}$. This implies that $v_j \notin \operatorname{conv}(\{z_k(t')\}_{k \in [n]})$ for t' large enough. This is a contradiction since $\mathcal{K} \subset \operatorname{conv}(\{z_k(t)\}_{k \in [n]})$ for any $t \geqslant 0$ according to Proposition 8.2. □

Proof of Claim 2. To simplify the notation, we only prove Claim 2 when $A = I_d$. Assume that $t \geqslant 0$ and that $z_i(t) \notin \mathcal{S}_\delta$.

First case. Firstly, we prove the claim in the case where $z_i(t) \notin \mathcal{S}_{\delta_0}$. For this, we notice that the function

$$f : x \mapsto \max_{j \in [n]} \langle v_j, x \rangle - \|x\|^2$$

is continuous, and by definition of $\mathcal{S}$, f is strictly positive on the compact set $\mathcal{K} \setminus \operatorname{Int}(\mathcal{S}_{\delta_0})$ (the complement in $\mathcal{K}$ of the interior of $\mathcal{S}_{\delta_0}$). Hence $f(x) \geqslant c'$ in this set for some constant $c' > 0$. Setting

$$\mathcal{K}_\varepsilon := \{x \in \mathbb{R}^d \colon \operatorname{dist}(x, \mathcal{K}) \leqslant \varepsilon\},$$

by continuity we find that $f(x) \geqslant c'/2$ for $x \in \mathcal{K}_\varepsilon \setminus \operatorname{Int}(\mathcal{S}_{\delta_0})$ and for sufficiently small $\varepsilon > 0$ (fixed in the sequel). For sufficiently large t, we have $z_i(t) \in \mathcal{K}_\varepsilon$ for any

$i \in [n]$, thus

$$\max_{j\in[n]}\langle z_i(t), z_j(t) - z_i(t)\rangle \geqslant \max_{j\in[m]}\langle z_i(t), v_j - z_i(t)\rangle \geqslant \frac{c'}{2}.$$

Since c' is independent of δ, we deduce the claim in this case (notice that it suffices to prove the claim for sufficiently small δ).

Second case. Secondly, we prove the claim when $z_i(t) \in \mathcal{S}_{\delta_0}\backslash\mathcal{S}_\delta$. The proof mainly relies on the following result:

Lemma 8.4. *For any $w \in \mathcal{S}$, there exists $\beta > 0$ such that if[6] $x \in \mathcal{K} \cap B(w, \delta_0)$, then*

$$\max_{j\in[m]}\langle x, v_j - x\rangle \geqslant \beta\|x - w\|. \tag{8.19}$$

We postpone the proof of Lemma 8.4 and show how to conclude the proof of Claim 2. Fix $\delta > 0$. We set

$$\eta := \frac{\beta\delta}{6R}$$

where

$$R := \max_{j\in[n]} \|z_j\|_{L^\infty(\mathbb{R})}.$$

Since $\mathrm{conv}(\{z_j(t)\}_{j\in[n]})$ converges to $\mathcal{K}$ as $t \to +\infty$, there exists $T(\delta) > 0$ such that for any $t \geqslant T(\delta)$, if $z_i(t) \in B(w, \delta_0)\backslash B(w, \delta)$ for some $w \in \mathcal{S}$, then

$$\|z_i(t) - x\| \leqslant \eta$$

for some $x \in \mathcal{K} \cap (B(w, \delta_0)\backslash B(w, \delta))$. Therefore, using Lemma 8.4,

$$\begin{aligned}\max_{j\in[m]}\langle z_i(t), v_j - z_i(t)\rangle &\geqslant \max_{j\in[m]}\langle x, v_j - x\rangle - 3R\eta \\ &\geqslant \beta\delta - 3R\eta \\ &= \frac{\beta}{2}\delta.\end{aligned}$$

To summarize, we have found that for any $\delta > 0$ there exists $T(\delta) > 0$ such that if $t \geqslant T(\delta)$ and $z_i(t) \in \mathcal{S}_{\delta_0}\backslash\mathcal{S}_\delta$, then

$$\max_{j\in[m]}\langle z_i(t), v_j - z_i(t)\rangle \geqslant \frac{\beta}{2}\delta. \tag{8.20}$$

Combining (8.20) with

$$\max_{j\in[n]}\langle z_i(t), z_j(t) - z_i(t)\rangle \geqslant \max_{j\in[m]}\langle z_i(t), v_j - z_i(t)\rangle$$

concludes the proof of Claim 2 in this second case. □

Proof of Lemma 8.4. Let us first address the case where $w = 0$. Writing any $x \in \mathcal{K}\backslash\{0\}$ as a convex combination of the vertices: $x = \sum_{j=1}^m \alpha_j v_j$, we find

$$0 = \left\langle x, \sum_{j=1}^m \alpha_j(v_j - x)\right\rangle = \sum_{j=1}^m \alpha_j\langle x, v_j - x\rangle. \tag{8.21}$$

[6]Here, $B(y, r)$ denotes the closed ball with center $y \in \mathbb{R}^d$ and radius $r > 0$.

We can exclude having $\langle x, v_j - x\rangle = 0$ for all $j \in [m]$, as this would necessarily imply that $\|x\|^2 = 2\sum_{j=1}^m \alpha_j \langle x, v_j - x\rangle = 0$. We deduce from (8.21) that

$$\max_{j\in[m]} \langle x, v_j - x\rangle > 0$$

for any $x \in \mathcal{K}\backslash\{0\}$. Hence, it is sufficient to prove (8.19) for $\|x\|$ small enough. We notice that for any $x \in \mathcal{K}\backslash\{0\}$ written as above,

$$\|x\|^2 = \sum_{j=1}^m \alpha_j \langle v_j, x\rangle.$$

Hence $x \mapsto \max_{j\in[m]} \langle v_j, x\rangle$ is positive for $x \in \mathcal{K}\backslash\{0\}$. Since this function is continuous and homogeneous in x, we deduce the existence of $\beta > 0$ such that

$$\max_{j\in[m]} \langle v_j, x\rangle \geqslant 2\beta\|x\|$$

for any $x \in \mathcal{K}$. For $x \in \mathcal{K}$ with $\|x\|$ sufficiently small, we obtain (8.19).

We now assume that $w \in \mathcal{S}\backslash\{0\}$. We set

$$\mathsf{I}_w := \left\{j \in [n] \colon \|w\|^2 = \langle w, v_j\rangle\right\}$$

and

$$\mathcal{A} := \operatorname{span}\left(\left\{v_j - w \colon j \in \mathsf{I}_w\right\}\right),$$

which is orthogonal to w. We also introduce

$$\mathscr{R} := \left(\mathbb{R}w \oplus \mathcal{A}\right)^{\perp},$$

and we denote by $\pi_{\mathscr{R}}$ the orthogonal projection on $\mathscr{R}$. We claim that there exists some $\rho > 0$ such that for any $j \in [m]$, we have

$$\langle w - v_j, w\rangle \geqslant \rho\|\pi_{\mathscr{R}} v_j\|.$$

This follows from the observation that $[m]$ is finite, and that $\|\pi_{\mathscr{R}} v_j\| > 0$ implies $\langle w - v_j, w\rangle > 0$. Therefore, for any $x \in \mathcal{K}$, writing x as a convex combination of the vertices, namely $x = \sum_{j=1}^m \alpha_j v_j$, we find that

$$\rho\|\pi_{\mathscr{R}} x\| \leqslant \sum_{j=1}^m \alpha_j \|\pi_{\mathscr{R}} v_j\| \leqslant \sum_{j=1}^m \alpha_j \langle w - v_j, w\rangle = \langle w - x, w\rangle. \tag{8.22}$$

Fix $x \in \mathcal{K} \cap B(w, \delta_0)$. We write $x = w + \delta' u$ with $0 \leqslant \delta' \leqslant \delta_0$ and $\|u\| = 1$. Then we have the orthogonal decomposition

$$u = bw + a + r \tag{8.23}$$

where $a \in \mathcal{A}$, $r \in \mathscr{R}$ and $b \in \mathbb{R}$. Since a is a convex combination of the form

$$a = \sum_{j\in\mathsf{I}_w} \beta_j (v_j - w),$$

we have

$$\|a\|^2 = \sum_{j\in\mathsf{I}_w} \beta_j \langle v_j - w, a\rangle,$$

whence

$$\max_{j\in\mathsf{I}_w} \langle a, v_j - w\rangle \geqslant \|a\|^2.$$

We deduce that

$$\begin{aligned}\max_{j\in \mathsf{I}_w}\langle x, v_j - x\rangle &= \max_{j\in \mathsf{I}_w}\langle w + \delta' u, (v_j - w) - \delta' u\rangle \\ &= -\delta' b\|w\|^2 - \delta'^2 + \delta' \max_{j\in \mathsf{I}_w}\langle a, v_j - w\rangle \\ &\geqslant -\delta' b\|w\|^2 - \delta'^2 + \delta'\|a\|^2. \end{aligned} \tag{8.24}$$

Notice that $b \leqslant 0$ by combining (8.22) and (8.23). Since $\|u\| = 1$ and using (8.22) we have

$$1 = b^2 + \|a\|^2 + \|r\|^2 \leqslant \|a\|^2 + \kappa b^2 \leqslant \kappa(\|a\|^2 + b^2)$$

where $\kappa := 1+\rho^{-2}\|w\|^4$. We deduce that either $\|a\|^2 \geqslant (2\kappa)^{-1}$ or $-b = |b| \geqslant (2\kappa)^{-\frac{1}{2}}$. Plugging this knowledge in (8.24) and using the fact that $\|w\| > 0$, we finally deduce the existence of an $\alpha > 0$ (independent of $\delta > 0$ and $x \in \mathcal{K} \cap B(w, \delta_0)$) such that

$$\max_{j\in[m]}\langle x, v_j - x\rangle \geqslant \alpha\delta' - \delta'^2 = \alpha\|x - w\| - \|x - w\|^2.$$

This proves (8.19) when $\|x - w\| \leqslant \alpha/2$.

It thus remains to show that (8.19) holds for all $x \in \mathcal{K} \cap (B(w, \delta_0)\backslash B(w, \frac{\alpha}{2}))$. To this end, we notice that $x \mapsto \max_{j\in[m]}\langle x, v_j - x\rangle$ is continuous in the connected set $\mathcal{K} \cap (B(w, \delta_0)\backslash B(w, \frac{\alpha}{2}))$, non-negative according to (8.5), and it is nowhere 0 (by definition of $\mathcal{S}$). Therefore, it is strictly positive, and denote by $\alpha' > 0$ some lower bound. Then for $x \in \mathcal{K} \cap (B(w, \delta_0)\backslash B(w, \frac{\alpha}{2}))$, we have

$$\max_{j\in[m]}\langle x, v_j - x\rangle \geqslant \alpha' \geqslant \frac{\alpha'}{\delta_0}\|x - w\|.$$

This concludes the proof of Lemma 8.4. □

8.2. **A cluster at the origin.** We complete this section by addressing the case $V = -I_d$, for which the convergence of the solutions of (1.1) is the simplest, since a unique cluster forms at the origin. We also suppose that $Q^\top K = I_d$: in other words, we consider the dynamics

$$\dot{x}_i(t) = -\sum_{j=1}^{n}\left(\frac{e^{\langle x_i(t), x_j(t)\rangle}}{\sum_{k=1}^{n} e^{\langle x_i(t), x_k(t)\rangle}}\right) x_j(t), \qquad t \in [0, +\infty), \tag{8.25}$$

with a prescribed initial condition $\{x_i(0)\}_{i\in[n]} \subset \mathbb{R}^d$.

Theorem 8.5 (Convergence toward the origin)**.** *Suppose $V = -I_d$ and $Q^\top K = I_d$. Then, for any initial sequence of tokens $\{x_i(0)\}_{i\in[n]} \subset \mathbb{R}^d$, and for any $i \in [n]$, we have $\|x_i(t)\| \to 0$ as $t \to +\infty$.*

Remark 8.6. *In the setting of Theorem 8.5, the self-attention matrix $P(t)$ defined in* (1.2) *converges, as $t \to +\infty$, to the $n \times n$ matrix with all entries equal to $1/n$.*

8.2.1. *Proof of Theorem 8.5.* We begin by showing that for any $i \in [n]$, the solution to (8.25) is uniformly bounded for all $t > 0$. In the sequel, we fix an initial configuration $\{x_i(0)\}_{i\in[n]} \subset \mathbb{R}^d$.

Lemma 8.7. *The trajectories of* (8.25) *are uniformly bounded in time—namely, there exists $R > 0$ (depending solely on n and the initial configuration) such that the solution $x_i(\cdot)$ to* (8.25) *satisfies $\|x_i(t)\| \leqslant R$ for any $i \in [n]$ and $t \geqslant 0$.*

Proof of Lemma 8.7. We fix $i \in [n]$. For $t \geqslant 0$, we denote by $D_i(t)$ the set of points $x_k(t)$ such that $\langle x_i(t), x_k(t)\rangle \geqslant 0$. We also set

$$S_i(t) := \sum_{k \in D_i(t)} e^{\langle x_i(t), x_k(t)\rangle} \langle x_i(t), x_k(t)\rangle,$$

and

$$R_i(t) := \sum_{k=1}^{n} e^{\langle x_i(t), x_k(t)\rangle}.$$

Since $1 + x \leqslant e^x$ whence $e^{-x}x \leqslant 1$, we deduce that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|x_i(t)\|^2 = -\frac{\sum_{k=1}^{n} e^{\langle x_i(t), x_k(t)\rangle}\langle x_i(t), x_k(t)\rangle}{R_i(t)} \leqslant \frac{-S_i(t) + n}{R_i(t)}.$$

Now since $1 - x \leqslant e^{-x}$ whence $e^x \leqslant 1 + e^x x$, we find that $R_i(t) \leqslant n + S_i(t)$. Consequently, if we assume that $\|x_i(t)\|^2 \geqslant 2n$ then $S_i(t) \geqslant 2n$, and therefore

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|x_i(t)\|^2 \leqslant \frac{-S_i(t) + n}{n + S_i(t)} \leqslant -1.$$

This shows that $\|x_i(t)\| \leqslant \max\{\|x_i(0)\|, \sqrt{2n}\}$ for any $t \geqslant 0$, which concludes the proof. □

By virtue of Lemma 7.1, we are able to characterize the stationary configurations for the dynamics (8.25)—namely, the set of points $(\bar{x}_1, \dots, \bar{x}_n) \in (\mathbb{R}^d)^n$ satisfying

$$\sum_{j=1}^{n} \left(\frac{e^{\langle \bar{x}_i, \bar{x}_j\rangle}}{\sum_{k=1}^{n} e^{\langle \bar{x}_i, \bar{x}_k\rangle}} \right) \bar{x}_j = 0$$

for all $i \in [n]$.

Lemma 8.8. *The only stationary configuration for the dynamics* (8.25) *is* $\bar{x}_1 = \dots = \bar{x}_n = 0$.

Proof. Assume that $(\bar{x}_1, \dots, \bar{x}_n) \in (\mathbb{R}^d)^n$ is a stationary configuration for the dynamics (8.25). We consider $f : \mathbb{R}^d \to \mathbb{R}$ defined as

$$f : x \mapsto \log\left(\sum_{j=1}^{n} e^{\langle x, \bar{x}_j\rangle} \right).$$

Per Lemma 7.1, f is convex, whence

$$f(x) \geqslant f(\bar{x}_i) + \langle \nabla f(\bar{x}_i), x - \bar{x}_i\rangle$$

for $x \in \mathbb{R}^d$ and $i \in [n]$. Since $\nabla f(\bar{x}_i) = 0$ for any $i \in [n]$, we gather that $f(x) \geqslant f(\bar{x}_i)$, whence $\bar{x}_i$ is a global minimizer of f for any $i \in [n]$. By convexity, f is constant on $\mathrm{conv}(\{\bar{x}_i\}_{i \in [n]})$. Since f is analytic on the affine space E spanned by the points $\bar{x}_i$, $i \in [n]$, it is then constant on E as well. Now assume that not all of the points $\bar{x}_i$ are equal, and pick an index $i_0 \in [n]$ such that $\bar{x}_{i_0}$ is not equal to the projection of the origin onto E. Then there exists some $j_0 \in [n]$ such that $\langle \bar{x}_{i_0} - \bar{x}_{j_0}, \bar{x}_{i_0}\rangle \neq 0$. For any $s \in \mathbb{R}$, we set $P_s := \bar{x}_{j_0} + s(\bar{x}_{i_0} - \bar{x}_{j_0}) \in E$, and we notice that $f(P_s) \geqslant \langle P_s, \bar{x}_{i_0}\rangle$, where the lower bound tends to $+\infty$ either when $s \to +\infty$ or when $s \to -\infty$. This contradicts the fact that f is constant on E. We conclude that the $\bar{x}_i$ are all equal for $i \in [n]$. The only value they can then take is necessarily 0. □

Lemma 8.9. *The trajectories of* (8.25) *satisfy* $\int_0^{+\infty} \|\dot{x}_i(t)\|^2 \,\mathrm{d}t < +\infty$ *for any* $i \in [n]$.

Proof. The function

$$\mathscr{L} : t \mapsto \sum_{i=1}^{n} \sum_{j=1}^{n} e^{\langle x_i(t), x_j(t)\rangle}$$

is non-increasing, as demonstrated by the following simple computation:

$$\begin{aligned}\frac{\mathrm{d}\mathscr{L}(t)}{\mathrm{d}t} &= 2\sum_{i=1}^{n}\sum_{j=1}^{n} e^{\langle x_i(t),x_j(t)\rangle}\langle \dot{x}_i(t), x_j(t)\rangle = 2\sum_{i=1}^{n}\left\langle \dot{x}_i(t), \sum_{j=1}^{n} e^{\langle x_i(t),x_j(t)\rangle} x_j(t)\right\rangle \\ &= -2\sum_{i=1}^{n}\sum_{j=1}^{n} e^{\langle x_i(t),x_j(t)\rangle}\|\dot{x}_i(t)\|^2.\end{aligned}$$

Being non-negative, $\mathscr{L}(t)$ thus converges as $t \to +\infty$. Since $\langle x_i(t), x_j(t)\rangle \geqslant R$ for some (possibly negative) $R \in \mathbb{R}$ by virtue of Lemma 8.7, we deduce that

$$\int_0^{+\infty} \|\dot{x}_i(t)\|^2 \,\mathrm{d}t \leqslant e^{-R}\int_0^{+\infty} \sum_{i=1}^{n}\sum_{j=1}^{n} e^{\langle x_i(t),x_j(t)\rangle}\|\dot{x}_i(t)\|^2 \,\mathrm{d}t = e^{-R}(\mathscr{L}(0) - \lim_{t\to+\infty}\mathscr{L}(t)),$$

which concludes the proof. □

We are now able to conclude the proof of Theorem 8.5.

Proof of Theorem 8.5. We set $\mathbf{X}(t) := (x_1(t), \dots, x_n(t)) \in (\mathbb{R}^d)^n$. If $\mathbf{X}(t)$ does not converge to 0, the compactness provided by Lemma 8.7 implies that there is a sequence $\{t_k\}_{k=1}^{+\infty}$ with $t_k \to +\infty$, and $\mathbf{X}^* = (x_1^*, \dots, x_n^*) \in (\mathbb{R}^d)^n \backslash \{0\}$, such that $\mathbf{X}(t_k) \to \mathbf{X}^*$ as $k \to +\infty$. To conclude the proof, it suffices to show that $\mathbf{X}^*$ is a stationary configuration of the dynamics: this directly leads to a contradiction per Lemma 8.8. Therefore, assume that $\mathbf{X}^*$ is not a stationary configuration of the dynamics. We denote by $\mathbf{X}^*(t) = (x_1^*(t), \dots, x_n^*(t))$ the solution of (8.25) with initial condition $\mathbf{X}^*$. Then, there exists $i \in [n]$ such that $\dot{x}_i^*(0) \neq 0$. We set $\varepsilon = \|\dot{x}_i^*(0)\|$. We select $T_0 > 0$ (possibly small) such that $\|\dot{x}_i^*(t)\| \geqslant \varepsilon/2$ for $t \in [0, T_0]$. It follows from (6.9) (which is verified according to Corollary 6.6) that for any $\delta > 0$ there exists $k_0 \in \mathbb{N}$ such that $\|\mathbf{X}(t_k + t) - \mathbf{X}^*(t)\| \leqslant \delta$ for any $t \in [0, T_0]$ and any $k \geqslant k_0$. By (6.5) (which is verified according to Corollary 6.6), we obtain that $\|\dot{x}_i(t_k + t) - \dot{x}_i^0(t)\| \leqslant C\delta$ for $t \in [0, T_0]$ and any $k \geqslant k_0$. Choosing $\delta > 0$ sufficiently small, we obtain that $\|\dot{x}_i(t_k + t)\| \geqslant \varepsilon/4$ for $t \in [0, T_0]$ and any $k \geqslant k_0$. This contradicts Lemma 8.9. □

9. Proof of Theorem 4.2

To ensure clarity, we present the proof of Theorem 4.2 under the assumption that V is diagonalizable. However, this assumption is not necessary. In Remark 9.5, we explain how the proof can be modified to accommodate for non-diagonalizable V.

Let us therefore assume that V is diagonalizable. Let $(\varphi_1, \dots, \varphi_d)$ be an orthonormal basis of eigenvectors associated to eigenvalues $(\lambda_1, \dots, \lambda_d)$, ordered in a decreasing manner with respect to their modulus: $|\lambda_1| \geqslant \ldots \geqslant |\lambda_d|$. (Starting from this point and throughout, we use the symbol λ exclusively to denote the eigenvalues of V.) Except for $\lambda_1 \in \mathbb{R}$, all the other eigenvalues (and eigenvectors) may be complex. We denote by $(\varphi_1^*, \dots, \varphi_d^*)$ the dual basis of $(\varphi_1, \dots, \varphi_d)$.

9.1. **Some monotonicity properties and bounds.** To start, we present some general facts that are prove useful in all subsequent sub-cases.

Lemma 9.1. *Suppose $k \in [d]$ is such that $\lambda_k \geqslant 0$. Then $t \mapsto \max_{j\in[n]} \varphi_k^*(z_j(t))$ is a non-increasing and bounded function, and $t \mapsto \min_{j\in[n]} \varphi_k^*(z_j(t))$ is a non-decreasing and bounded function. In particular, $t \mapsto \varphi_k^*(z_i(t))$ is uniformly bounded as a function on $[0, +\infty)$ for any $i \in [n]$.*

Proof. For any $k \in [d]$ and any $t \geqslant 0$, set

$$\alpha_k(t) = \min_{j\in[n]} \varphi_k^*(z_j(t)), \qquad \beta_k(t) = \max_{j\in[n]} \varphi_k^*(z_j(t)).$$

Let $i \in [n]$ be an index such that $\alpha_k(t) = \varphi_k^*(z_i(t))$. Then we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\varphi_k^*(z_i(t)) &= \sum_{j=1}^n P_{ij}(t)\varphi_k^*\left(V(z_j(t) - z_i(t))\right) \\ &= \lambda_k \sum_{j=1}^n P_{ij}(t)(\varphi_k^*(z_j(t)) - \varphi_k^*(z_i(t))) \geqslant 0\end{aligned}$$

where the last inequality stems from the fact that $\lambda_k \geqslant 0$ and the choice of index i. This proves that $\alpha_k(\cdot)$ is non-decreasing, as desired. Arguing similarly, one finds that $\beta_k(\cdot)$ is non-increasing. As a consequence, $\alpha_k(0) \leqslant \alpha_k(t) \leqslant \beta_k(t) \leqslant \beta_k(0)$ for any $t \geqslant 0$, which shows that $\alpha_k(\cdot)$ and $\beta_k(\cdot)$ are bounded. □

Corollary 9.2. *If V only has real non-negative eigenvalues, then $z_i(\cdot) \in L^\infty([0, +\infty))$.*

Lemma 9.3. *Fix $k \in [d]$ and $i \in [n]$. Then there exists a constant $C > 0$ such that*

$$\left|\varphi_k^*\left(e^{tV} z_i(t)\right)\right| \leqslant Ce^{|\lambda_k|t}$$

holds for all $t \geqslant 0$.

Proof. We naturally make use of the equation for $x_i(t) := e^{tV} z_i(t)$. Fix $t \geqslant 0$. We have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left|\varphi_k^*(x_i(t))\right|^2 &= 2 \cdot \mathbf{Re}\left(\overline{\varphi_k^*(x_i(t))}\frac{\mathrm{d}}{\mathrm{d}t}\varphi_k^*(x_i(t))\right) \\ &= 2 \cdot \mathbf{Re}\left(\sum_{j=1}^n P_{ij}(t)\varphi_k^*(Vx_j(t))\overline{\varphi_k^*(x_i(t))}\right) \\ &= 2 \cdot \mathbf{Re}\left(\sum_{j=1}^n P_{ij}(t)\lambda_k\varphi_k^*(x_j(t))\overline{\varphi_k^*(x_i(t))}\right) \\ &\leqslant 2|\lambda_k| \max_{j\in[n]} \left|\varphi_k^*(x_j(t))\right|^2.\end{aligned}$$

Choosing $i \in [n]$ running over the set of indices such that $|\varphi_k^*(x_i(t))|$ is maximal, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} \max_{j\in[n]} |\varphi_k^*(x_j(t))|^2 \leqslant 2|\lambda_k| \max_{j\in[n]} |\varphi_k^*(x_j(t))|^2.$$

We conclude the proof by applying Grönwall's lemma. □

9.2. **Proof of Theorem 4.2.** We now prove Theorem 4.2. We again recall that λ_1 is simple and positive, and the eigenvalues of V are ordered in decreasing order of modulus: $\lambda_1 > |\lambda_2| \geqslant \ldots \geqslant |\lambda_d|$.

Proof of Theorem 4.2. We look to prove that for any $i \in [n]$, the component of $z_i(t)$ along the principal eigenvector φ_1, i.e. $\varphi_1^*(z_i(t))$, converges as $t \to +\infty$. We also show that there exists a set of at most 3 real numbers (depending on the initial datum $(z_1(0), \ldots, z_n(0))$) such that for any $i \in [n]$ the limit of $\varphi_1^*(z_i(t))$ belongs to this set. Theorem 4.2 directly follows from these facts.

Let $i \in [n]$ be fixed. Recall from Lemma 9.1 that $\varphi_1^*(z_i(t))$ is uniformly bounded for any $t \in [0, +\infty)$. We set

$$a := \lim_{t \to +\infty} \min_{j \in [n]} \varphi_1^*(z_j(t)), \qquad b := \lim_{t \to +\infty} \max_{j \in [n]} \varphi_1^*(z_j(t)). \tag{9.1}$$

(Note that by Lemma 9.1, $a \geqslant \min_{j \in [n]} \varphi_1^*(z_j(0))$ and $b \leqslant \max_{j \in [n]} \varphi_1^*(z_j(0))$.) For $c \in \{0, a, b\}$, we define the candidate limiting hyperplanes for $z_i(t)$:

$$H_c := \{x \in \mathbb{R}^d \colon \varphi_1^*(x) = c\}.$$

We show that $z_i(t)$ converges either to H_0, to H_a or to H_b. If $a = b = 0$, then according to (9.1) all particles converge to H_0 and there is nothing left to prove. We now distinguish two scenarios:

(i) either for any $\varepsilon > 0$, $|\varphi_1^*(z_i(t))| \leqslant \varepsilon$ for t large enough—in which case, we deduce that $z_i(t)$ converges toward H_0 as $t \to +\infty$—,

(ii) or $|\varphi_1^*(z_i(t_k))| > \varepsilon_0$ for some $\varepsilon_0 > 0$ and for some sequence of positive times $\{t_k\}_{k=1}^{+\infty}$ with $t_k \to +\infty$.

Since case (i) is straightforward, let us handle case (ii). Without loss of generality, we can extract a subsequence of times (which we do not relabel, for simplicity of notation) along which

$$\varphi_1^*(z_i(t_k)) > \varepsilon_0. \tag{9.2}$$

Let $\varepsilon \in (0, \varepsilon_0]$ be fixed and to be chosen later. We set

$$w_j(t) := \left\langle Q e^{tV} z_i(t), K e^{tV} z_j(t) \right\rangle,$$

so that

$$\frac{1}{\lambda_1} \frac{\mathrm{d}}{\mathrm{d}t} \varphi_1^*(z_i(t)) = \sum_{j=1}^n \frac{e^{w_j(t)}}{\sum_{k=1}^n e^{w_k(t)}} \left(\varphi_1^*(z_j(t)) - \varphi_1^*(z_i(t))\right). \tag{9.3}$$

We look to obtain a lower bound for the right-hand side in the above identity. Let us use the shorthand

$$c_{k\ell} := \langle Q\varphi_k, K\varphi_\ell \rangle$$

for $k, \ell \in [d]$. By assumption, $c_{11} > 0$. We have $\varphi_k^*(e^{tV} z_i(t)) = e^{t\lambda_k} \varphi_k^*(z_i(t))$ and the following spectral expansion holds:

$$e^{tV} z_i(t) = \sum_{k=1}^d e^{t\lambda_k} \varphi_k^*(z_i(t)) \varphi_k.$$

Using this fact, as well as Lemma 9.3, we gather that

$$\begin{aligned}\left|w_j(t) - c_{11}e^{2\lambda_1 t}\varphi_1^*(z_i(t))\varphi_1^*(z_j(t))\right| &= \left|\sum_{(k,\ell)\neq(1,1)} c_{k\ell}\varphi_k^*(e^{tV}z_i(t))\varphi_\ell^*\left(e^{tV}z_j(t)\right)\right| \\ &\leqslant \sum_{(k,\ell)\neq(1,1)} |c_{k\ell}|\left|\varphi_k^*\left(e^{tV}z_i(t)\right)\right|\left|\varphi_\ell^*\left(e^{tV}z_j(t)\right)\right| \\ &\leqslant C^2\|Q^\top K\|_{\mathrm{op}} \sum_{(k,\ell)\neq(1,1)} e^{(|\lambda_k|+|\lambda_\ell|)t} \\ &\leqslant \underbrace{C^2\|Q^\top K\|_{\mathrm{op}}(d-1)^2}_{=:C'}\, e^{(\lambda_1+|\lambda_2|)t} \qquad (9.4)\end{aligned}$$

holds for all $t \geqslant 0$ and $j \in [n]$. Now since $\lambda_1 > 0$, Lemma 9.1 implies that for any $t \geqslant 0$ there exists an index $i_0(t) \in [n]$ such that

$$\varphi_1^*(z_{i_0(t)}(t)) \geqslant b. \qquad (9.5)$$

With $j_0(t) \in \operatorname{argmax}_{j\in[n]} w_j(t)$, using (9.4) and (9.5) we see that

$$w_{j_0(t)}(t) \geqslant w_{i_0(t)}(t) \geqslant c_{11}\varphi_1^*(z_i(t))be^{2\lambda_1 t} - C'e^{(\lambda_1+|\lambda_2|)t}. \qquad (9.6)$$

Now for any t within the sequence $\{t_k\}_{k=1}^{+\infty}$, combining the first inequality in (9.6) with the fact that $c_{11} > 0$, (9.2) and (9.4), we deduce that

$$\varphi_1^*(z_{j_0(t)}(t)) - \varphi_1^*(z_{i_0(t)}(t)) \geqslant -\frac{2C'}{c_{11}\varepsilon}e^{-(\lambda_1-|\lambda_2|)t}. \qquad (9.7)$$

As $\lambda_1 > |\lambda_2|$, for t large enough, we find that we can lower bound the above expression by $-\frac{\varepsilon}{4}$. We now define the set of indices

$$N(t) := \{j \in [n] \colon \varphi_1^*(z_i(t)) - \varphi_1^*(z_j(t)) \geqslant 0\}.$$

Take t within the sequence $\{t_k\}_{k=1}^{+\infty}$ such that $\varphi_1^*(z_i(t)) \leqslant b - \varepsilon$ and large enough so that (9.7) is lower bounded by $-\frac{\varepsilon}{4}$ (if such a t does not exist, we immediately conclude that $\varphi_1^*(z_i(t)) \to b$ as $t \to +\infty$). Using (9.5) and the subsequent derivations, we deduce that

$$\varphi_1^*(z_{j_0(t)}(t)) - \varphi_1^*(z_i(t)) \geqslant \frac{3\varepsilon}{4},$$

and since $\varphi_1^*(z_j(t)) - \varphi_1^*(z_i(t)) \geqslant 0$ for $j \notin N(t)$, we expand in (9.3) to get

$$\frac{1}{\lambda_1}\frac{\mathrm{d}}{\mathrm{d}t}\varphi_1^*(z_i(t)) \geqslant \frac{e^{w_{j_0(t)}(t)}}{\sum_{k=1}^n e^{w_k(t)}}\frac{3\varepsilon}{4} + \sum_{j\in N(t)} \frac{e^{w_j(t)}}{\sum_{k=1}^n e^{w_k(t)}}\left(\varphi_1^*(z_j(t)) - \varphi_1^*(z_i(t))\right). \qquad (9.8)$$

On another hand, for $j \in N(t)$, we may use (9.4) to find

$$w_j(t) \leqslant c_{11}\varphi_1^*(z_i(t))^2 e^{2\lambda_1 t} + C'e^{(\lambda_1+|\lambda_2|)t}. \qquad (9.9)$$

We set

$$C_0 := \max_{j\in[n]} \varphi_1^*(z_j(0)) - \min_{j\in[n]} \varphi_1^*(z_j(0)).$$

Using the monotonicity properties from Lemma 9.1, as well as (9.9) in (9.8), we obtain

$$\frac{1}{\lambda_1}\frac{\mathrm{d}}{\mathrm{d}t}\varphi_1^*(z_i(t)) \geqslant \frac{3\varepsilon}{4n} - C_0 n \frac{\exp\Big(c_{11}\varphi_1^*(z_i(t))^2 e^{2\lambda_1 t} + C' e^{(\lambda_1+|\lambda_2|)t}\Big)}{\exp\Big(c_{11}\varphi_1^*(z_i(t)) b e^{2\lambda_1 t} - C' e^{(\lambda_1+|\lambda_2|)t}\Big)}.$$

Given our choice of t, we have $\varphi_1^*(z_i(t))^2 - b\varphi_1^*(z_i(t)) \leqslant -\varepsilon(b-\varepsilon)$, so, we conclude from the inequality just above that

$$\frac{1}{\lambda_1}\frac{\mathrm{d}}{\mathrm{d}t}\varphi_1^*(z_i(t)) \geqslant \frac{3\varepsilon}{4n} - C_0 n \exp\Big(-c_{11}\varepsilon(b-\varepsilon)e^{2\lambda_1 t} + 2C' e^{(\lambda_1+|\lambda_2|)t}\Big). \tag{9.10}$$

Since $\lambda_1 > |\lambda_2|$, it follows from (9.10) that there exists $T > 0$ such that for any t within the sequence $\{t_k\}_{k=1}^{+\infty}$ for which $t \geqslant T$ and $\varphi_1^*(z_i(t)) \in [\varepsilon, b-\varepsilon]$, there holds

$$\frac{\mathrm{d}}{\mathrm{d}t}\varphi_1^*(z_i(t)) \geqslant \frac{\lambda_1\varepsilon}{2n}.$$

This shows the existence of a larger time horizon $T' > T$ such that $\varphi_1^*(z_i(t)) \geqslant b-\varepsilon$ whenever $t \geqslant T'$. And since ε can be taken arbitrarily small, we deduce that $\varphi_1^*(z_i(t))$ converges toward b, namely that $z_i(t)$ converges toward H_b, as $t \to +\infty$.

Arguing in the same way as above, and assuming without loss of generality that $a < 0$, we may find that all indices $i \in [n]$ for which $\varphi_1^*(z_i(t_k)) \leqslant -\varepsilon_0$ for some $\varepsilon_0 > 0$ and some sequence $t_k \to +\infty$, the particle $z_i(t)$ converges toward H_a as $t \to +\infty$. This concludes the proof. □

9.3. **Remarks.**

Remark 9.4. *Theorem 4.2 establishes the convergence of $\varphi_1^*(z_i(t))$ for any $i \in [n]$ as $t \to +\infty$, but does not preclude the fact that $\|z_i(t)\|$ may diverge toward $+\infty$ (along the hyperplane) as $t \to +\infty$. This is indeed expected (and observed numerically—see Fig. 6) when V has some negative eigenvalues. We also note that when all the eigenvalues of V are non-negative, Corollary 9.2 shows that all the $z_i(t)$ remain bounded.*

Remark 9.5 (The case where V is not diagonalizable)**.** *If V is not assumed to be diagonalizable, Lemma 9.3 (or, at least the proof thereof) requires some modifications. Let $\delta := \lambda_1 - |\lambda_2| > 0$. Let $\varepsilon > 0$ be fixed and to be chosen later. We decompose V in Jordan blocks, and we consider*

$$\mathbb{C}^d = \bigoplus_{k=1}^{m} \mathscr{F}_k, \tag{9.11}$$

where $\mathscr{F}_k$ is the span of the Jordan chain corresponding to the k-th Jordan block. By a slight abuse of notation (solely for the purpose of this remark), we denote by λ_k the eigenvalue associated to the k-th Jordan block. We recall that we can choose a basis $(\varphi_{k,1}, \dots, \varphi_{k,j_k})$ of each $\mathscr{F}_k$ in a way that $V_{|\mathscr{F}_k}$ reads in this basis as[7]

$$\begin{bmatrix} \lambda_k & \varepsilon & & \\ & \ddots & \ddots & \\ & & \ddots & \varepsilon \\ & & & \lambda_k \end{bmatrix}. \tag{9.12}$$

[7]Recall that Jordan blocks are commonly written with a $+1$ in the superdiagonal. This can be replaced by any non-zero complex scalar as done here—see [HJ12, Chapter 3, Corollary 3.1.21].

We observe that if ε is chosen sufficiently small (depending only on δ), Lemma 9.3 may be replaced by the following estimate in each $\mathscr{F}_k$:

$$\exists C > 0,\ \forall t \geqslant 0,\ \forall i \in [n], \qquad \left\|\pi_{\mathscr{F}_k}\left(e^{tV} z_i(t)\right)\right\| \leqslant C e^{(|\lambda_k|+\delta)t}. \tag{9.13}$$

Here, $\pi_{\mathscr{F}_k}$ denotes the orthogonal projection onto $\mathscr{F}_k$. To prove estimate (9.13)*, we follow the proof of Lemma 9.3, with $\frac{\mathrm{d}}{\mathrm{d}t}\|\pi_{\mathscr{F}_k}(x_i(t))\|^2$ playing the role of $\frac{\mathrm{d}}{\mathrm{d}t}\,|\varphi_k^*(x_i(t))|^2$. The key observation is that combining* (9.11) *and* (9.12) *we obtain*

$$\|\pi_{\mathscr{F}_k}(V x_i(t))\| \leqslant (|\lambda_k| + \delta)\|\pi_{\mathscr{F}_k}(x_i(t))\|,$$

provided ε is chosen sufficiently small. Then (9.13) *follows as in Lemma 9.3.*

With (9.11) *at hand, the proof of Theorem 4.2 carries through, under the impactless modification that $Ce^{(\lambda_1+|\lambda_2|+\delta)t}$ replaces* (9.4) *(and subsequent estimates are modified in the same way).*

10. Proof of Theorem 5.2

In this section, we establish the proof for Theorem 5.2. Since the proof is essentially a combination of the proofs of Theorems 4.2 and 8.1, we may occasionally skip certain details and refer to the proofs of these two results. As done throughout this work, we set

$$A := (Q^\top K)^{\frac{1}{2}}.$$

We denote by $\pi_{\mathscr{F}} : \mathbb{R}^d \to \mathscr{F}$ the projection onto $\mathscr{F}$ parallel to $\mathscr{G}$, and by $\pi_{\mathscr{G}} : \mathbb{R}^d \to \mathscr{G}$ the projection onto $\mathscr{G}$ parallel to $\mathscr{F}$. The set $\pi_{\mathscr{F}}(\mathrm{conv}(\{z_i(t)\}_{i\in[n]}))$ is a convex subset of $\mathscr{F}$ which is non-increasing with respect to t (the proof of this fact is identical to that of Proposition 8.2). It therefore converges toward some convex polytope $\mathcal{K}$ as $t \to +\infty$.

Fix $i \in [n]$. We have

$$\begin{aligned}
\pi_{\mathscr{F}}(\dot{z}_i(t)) &= \sum_{j=1}^n \left(\frac{e^{\left\langle Ae^{tV} z_i(t), Ae^{tV} z_j(t)\right\rangle}}{\sum_{k=1}^n e^{\left\langle Ae^{tV} z_i(t), Ae^{tV} z_k(t)\right\rangle}} \right) \pi_{\mathscr{F}}(V(z_j(t) - z_i(t))) \\
&= \sum_{j=1}^n \left(\frac{e^{\left\langle Ae^{tV} z_i(t), Ae^{tV} (z_j(t) - z_i(t))\right\rangle}}{\sum_{k=1}^n e^{\left\langle Ae^{tV} z_i(t), Ae^{tV} (z_k(t) - z_i(t))\right\rangle}} \right) \pi_{\mathscr{F}}(V(z_j(t) - z_i(t))).
\end{aligned}$$

From this point on, we follow the proof of Theorem 8.1, and we solely highlight the changes compared to the original proof. Roughly speaking, this new proof amounts to adding projections $\pi_{\mathscr{F}}$ at several places. We denote by $\mathcal{S} \subset \mathscr{F}$ the set of points $w \in \mathcal{K}$ such that

$$\|\pi_{\mathscr{F}}(Aw)\|^2 = \max_{j\in[m]} \left\langle \pi_{\mathscr{F}}(Aw), \pi_{\mathscr{F}}(Av_j)\right\rangle.$$

The fact that $\mathcal{S} \subset \partial\mathcal{K}$ and that $\mathcal{S}$ has finite cardinality is proved precisely as Claim 1 (in the proof of Theorem 8.1), simply by replacing all occurrences of $A\cdot$ by $\pi_{\mathscr{F}}(A\cdot)$. Once again, $\mathcal{S}_\delta$ denotes the set of all points in $\mathcal{K}$ at distance $\leqslant \delta$ to some point of $\mathcal{S}$.

Step 2 in the proof of Theorem 8.1 (i.e., (8.7)) is replaced by the following statement:

Step 2': There exists a constant $\gamma = \gamma(\mathcal{K}) > 0$ (depending only on the geometry of $\mathcal{K}$) such that for any $\delta \in (0, \delta_0]$, there exists $T = T(\delta) > 0$ such that if $t \geqslant T$ and

$\pi_{\mathcal{F}}(z_i(t)) \notin \mathcal{S}_\delta$, then

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\pi_{\mathcal{F}}(Az_i(t))\|^2 \geqslant \gamma\delta.$$

We now proceed in proving this statement.

Proof of Step 2'. We set

$$a_j(t) := \langle \pi_{\mathcal{F}}(Az_i(t)), \pi_{\mathcal{F}}(A(z_j(t) - z_i(t))\rangle$$

and

$$r_j(t) := \left\langle Ae^{tV}z_i(t), Ae^{tV}(z_j(t) - z_i(t))\right\rangle - a_j(t)e^{2\lambda_1 t}.$$

We find

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\pi_{\mathcal{F}}(Az_i(t))\|^2 &= \langle \pi_{\mathcal{F}}(A\dot{z}_i(t)), \pi_{\mathcal{F}}(Az_i(t))\rangle \\
&= \sum_{j=1}^n \left(\frac{e^{\left\langle Ae^{tV}z_i(t), Ae^{tV}z_j(t)\right\rangle}}{\sum_{k=1}^n e^{\left\langle Ae^{tV}z_i(t), Ae^{tV}z_k(t)\right\rangle}}\right) \langle \pi_{\mathcal{F}}(A(z_j(t) - z_i(t))), \pi_{\mathcal{F}}(Az_i(t))\rangle \\
&= \sum_{j=1}^n \left(\frac{e^{\left\langle Ae^{tV}z_i(t), Ae^{tV}(z_j(t)-z_i(t))\right\rangle}}{\sum_{k=1}^n e^{\left\langle Ae^{tV}z_i(t), Ae^{tV}(z_k(t)-z_i(t))\right\rangle}}\right) \langle \pi_{\mathcal{F}}(A(z_j(t) - z_i(t))), \pi_{\mathcal{F}}(Az_i(t))\rangle \\
&= \sum_{j=1}^n \underbrace{\left(\frac{e^{a_j(t)e^{2\lambda_1 t} + r_j(t)}}{\sum_{k=1}^n e^{a_k(t)e^{2\lambda_1 t} + r_k(t)}}\right) a_j(t)}_{=:b_j(t)}. \qquad (10.1)
\end{aligned}$$

We now make use of the following adaptation of Claim 2.

Claim 3. *There exists some constant $\gamma' = \gamma'(\mathcal{K}) > 0$ depending only on the geometry of $\mathcal{K}$ such that the following holds. Fix $\delta \in (0, \delta_0]$. There exists $T = T(\delta) > 0$ such that if $t \geqslant T$ and $z_i(t) \notin \mathcal{S}_\delta \times \mathcal{G}$, then there exists $j \in [n]$ such that $a_j(t) \geqslant \gamma'\delta$.*

Compared to Step 2 in the proof of Theorem 8.1, we now have to estimate the coefficients $r_j(t)$. To this end, setting $y_j(t) := Ae^{tV}z_j(t)$ for $j \in [n]$, we notice that $r_j(t) = P_1(t) + P_2(t) + P_3(t)$ where

$$\begin{aligned}
P_1(t) &= \langle \pi_{\mathcal{F}}(y_i(t)), \pi_{\mathcal{G}}(y_j(t) - y_i(t))\rangle, \\
P_2(t) &= \langle \pi_{\mathcal{G}}(y_i(t)), \pi_{\mathcal{F}}(y_j(t) - y_i(t))\rangle, \\
P_3(t) &= \langle \pi_{\mathcal{G}}(y_i(t)), \pi_{\mathcal{G}}(y_j(t) - y_i(t))\rangle.
\end{aligned}$$

By virtue of Lemma 9.3 we have $|\pi_{\mathcal{F}}(y_j(t))| \leqslant Ce^{\lambda_1 t}$ and $|\pi_{\mathcal{G}}(y_j(t))| \leqslant Ce^{t|\lambda_2|}$ for any $t \geqslant 0$ (or $Ce^{t|\lambda_2|+\varepsilon}$ if $V_{|\mathcal{G}}$ is not diagonalizable—see Remark 9.5), hence

$$|r_j(t)| \leqslant Ce^{t(\lambda_1 + |\lambda_2|)}. \qquad (10.2)$$

Since $\pi_{\mathcal{F}}(z_j(t))$ is uniformly bounded in $t \in [0, +\infty)$ for any $j \in [n]$ due to Corollary 8.3, we get $a_j(\cdot) \in L^\infty(0, +\infty)$. So, we may set

$$\kappa := \max_{j\in[n]} \sup_{t\geqslant 0} |a_j(t)|.$$

Let $t \geqslant 0$. We define

$$B(t) := \left\{j \in [n] \colon a_j(t)e^{2\lambda_1 t} + r_j(t) \geqslant 0\right\}.$$

Let $j_0(t) \in \operatorname{argmax}_{j\in[n]}(a_j(t)e^{2\lambda_1 t} + r_j(t))$. Note that $j_0(t) \in B(t)$ since

$$a_{j_0}(t)e^{2\lambda_1 t} + r_{j_0}(t) \geqslant a_i(t)e^{2\lambda_1 t} + r_i(t) = 0.$$

We notice the following three properties:

- For $j = j_0(t)$, we have $b_{j_0(t)}(t) \geqslant \frac{a_{j_0(t)}(t)}{n}$ (recall the definition of b_j in (10.1));
- for any $j \in B(t)\backslash\{j_0\}$, we have $b_j(t) \geqslant 0$;
- for any $j \notin B(t)$, we have

$$b_j(t) \geqslant -\kappa \exp\left(-a_{j_0}(t)e^{2\lambda_1 t} + Ce^{(\lambda_1+|\lambda_2|)t}\right).$$

Indeed, using the fact that $j \in B(t)$ and (10.2), we find

$$\begin{aligned}\frac{\exp\left(a_j(t)e^{2\lambda_1 t} + r_j(t)\right)}{\sum_{k=1}^n \exp\left(a_k(t)e^{2\lambda_1 t} + r_k(t)\right)} &\leqslant \frac{1}{\sum_{k=1}^n \exp\left(a_k(t)e^{2\lambda_1 t} + r_k(t)\right)} \\ &\leqslant \frac{1}{\exp\left(a_{j_0}(t)e^{2\lambda_1 t} + r_{j_0}(t)\right)} \\ &\leqslant \exp\left(-a_{j_0}(t)e^{2\lambda_1 t} + Ce^{(\lambda_1+|\lambda_2|)t}\right).\end{aligned}$$

Making use of these properties in (10.1) yields the desired lower bound—indeed, if t is sufficiently large and $z_i(t) \notin \mathcal{S}_\delta \times \mathcal{G}$, we have $\{j \in [n] \colon a_j(t) \geqslant \gamma'\delta\} \neq \varnothing$ according to Claim 3, and so we deduce that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|Az_i(t)\|^2 \geqslant \frac{\gamma'\delta}{n} - \kappa n e^{-\gamma'\delta e^{2\lambda_1 t} + Ce^{(\lambda_1+|\lambda_2|)t}}.$$

Taking t possibly larger (and depending on δ), we obtain the result of Step 2'. □

Steps 3 and 4 in the proof of Theorem 8.1 are essentially unchanged—we replace all the occurrences of $\|A \cdot\|$ by $\|\pi_{\mathscr{F}}(A\cdot)\|$ (for instance in (8.13) and (8.14)). Although $\|Az_i(t)\|$ may not be uniformly bounded in t, it is important to note that $\|\pi_{\mathscr{F}}(Az_i(t))\|$ is uniformly bounded. Similarly, while $\dot{z}_i(t) \notin L^\infty([0,+\infty))$, we do have $\|\frac{\mathrm{d}}{\mathrm{d}t}\pi_{\mathscr{F}}(z_i(\cdot))\|_{L^\infty([0,+\infty))} < +\infty$. The sets $\mathcal{S}_\delta$, $\mathscr{C}_k$ and $\mathscr{C}_k^r$ are replaced by $\mathcal{S}_\delta \times \mathcal{G}$, $\mathscr{C}_k \times \mathcal{G}$ and $\mathscr{C}_k^r \times \mathcal{G}$ respectively. The conclusion is that $\|\pi_{\mathscr{F}}(Az_i(t))\|^2$ has to increase by at least

$$\frac{\gamma\delta^{\frac{1}{2}}(\delta^{\frac{1}{4}} - \delta)}{\|\dot{z}_i\|_{L^\infty([0,+\infty))}} \geqslant \frac{\delta^{\frac{3}{4}}}{2\|\dot{z}_i\|_{L^\infty([0,+\infty))}} > 4R\|A\|_{\mathrm{op}}\delta$$

during a travel from $\mathscr{C}_k \times \mathcal{G}$ to the complement of $\mathscr{C}_k^{\frac{1}{4}} \times \mathcal{G}$. As in the proof of Theorem 8.1 this implies that for any $i \in [n]$ there exists $s \in \mathcal{S}$ such that $z_i(t)$ remains at distance at most δ away from $\{s\} \times \mathcal{G}$. This being true for any $\delta > 0$, we obtain the desired result.

11. Numerical experiments

11.1. **Setup.** Unless indicated otherwise, all figures presented in this paper were generated by discretizing the underlying dynamics (either (1.1) or (3.1)) using a fourth order Runge-Kutta scheme with a step size of 0.1. All points in the initial sequence were drawn independently from the uniform distribution over the hypercube $[-5,5]^d$. Random matrices (e.g., Q, K, V) have entries drawn independently from the uniform distribution on $[-1,1]$. Codes and animated plots of all examples may be found online at

https://github.com/borjanG/2023-transformers.

We now present some experiments which motivate some conjectures and claims made in what precedes.

11.2. **Eigenvalues of ALBERT's value matrices.** In Figure 10 we illustrate the eigenvalues of the value matrices V_h for a couple of heads h in a pre-trained ALBERT model. We focus on ALBERT-xlarge-v2 available online at

https://huggingface.co/albert-xlarge-v2.

This version uses 16 heads, with sequences of length $n = 256$ and tokens of dimension $d = 128$. While not all value matrices V_h per head $h \in [16]$ satisfy the assumptions made in Section 4, we illustrate the eigenvalues of a couple of them which do.

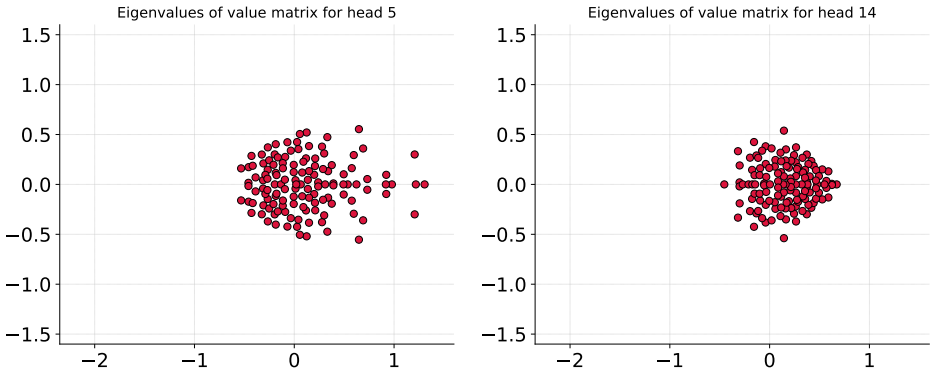


Figure 10. The eigenvalues of V_5 and V_{14} in the pre-trained ALBERT satisfy the eigenvalue assumption made in Definition 4.1. Furthermore, the second assumption made in Definition 4.1 is satisfied by (Q_5, K_5) and (Q_{14}, K_{14}) (the inner products evaluated along the eigenvector of norm 1 equal 1.3060 and 0.6719 respectively). In other words, the triples (Q_h, K_h, V_h) corresponding to heads $h = 5$ and $h = 14$ in ALBERT satisfy all the assumptions made in the statement of Theorem 4.2.

11.3. **Experiments related to Theorem 2.1.** We begin with the setup of Theorem 2.1, which we recall was proven to hold in the case $d = 1$. Herein we present a couple of examples (Figures 11 and 12) which elucidate the role that d and n appear to play in this fact.

Notably, as seen in Fig. 4, we believe that the conclusion of Theorem 2.1 could plausibly be extended to any $d > 1$, assuming $V \succ 0$.

11.4. **Illustrating Theorem 4.2 in $\mathbb{R}^3$.** To precisely illustrate the appearance of at most three hyperplanes in the setting of Theorem 4.2, we gave an example in $\mathbb{R}^2$. We expand on this and provide a couple of toy examples in $\mathbb{R}^3$ for the purpose of visualization (we recall that these are toy models, as Transformers in practice are high-dimensional), and namely focus in both examples on the case where the two latter eigenvalues are complex. In Fig. 14, we see the effect of having eigenvalues with a negative real part, and the complementary case is illustrated in Fig. 13.

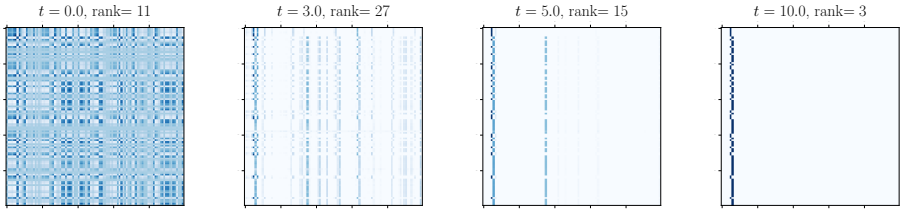


Figure 11. We expand on Fig. 3—for the same setup, consider $n = 100$. The sequence length n does not appear to influence the rank of $P(t)$, which is expected since the rank of P corresponds to the number of leaders.

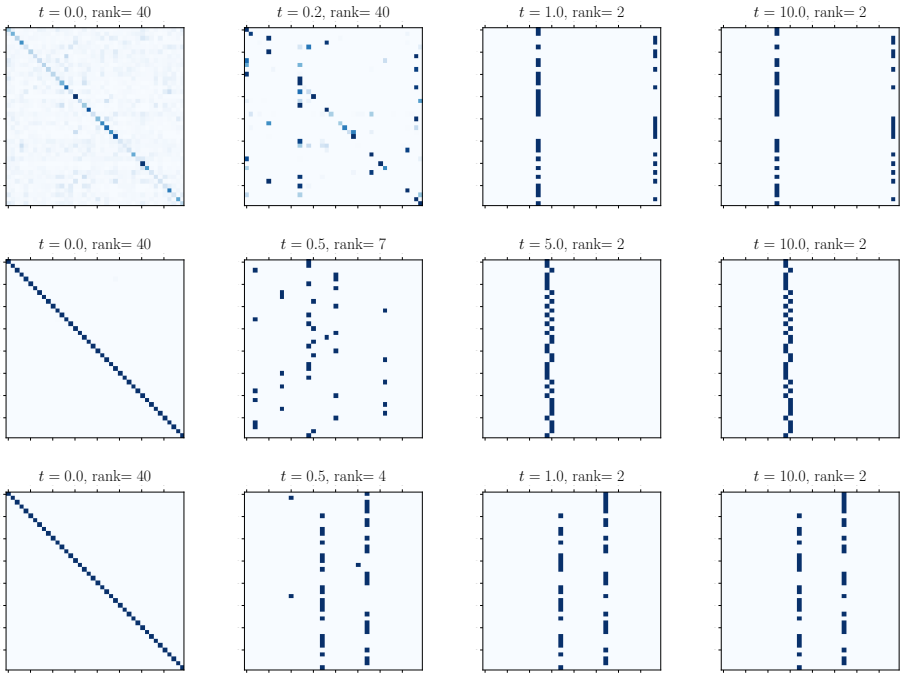


Figure 12. We consider $n = 40$, $Q = K = I_d$ and a random matrix $V \succ 0$ in dimensions $d = 10$ (first row), $d = 40$ (second row), and $d = 80$ (third row). The conclusion of Theorem 2.1 appears to transfer to the higher dimensional case, and this would actually follow from Conjecture 4.3 (should it hold).

11.5. **Complementing Figure 7.** In Figure 7, we illustrate the appearance of clustering in high-dimension (the ALBERT setup: $n = 256$ and $d = 128$) for generic random matrices (Q, K, V). The value matrix V in question has 65 positive eigenvalues, and we show the conjectured convergence of the 65 coordinates along the corresponding eigenvectors to one of possibly 3 (generically 2) real scalars. In Figure 15, we complement this illustration by showing the possible oscillatory and divergent behavior of the remaining coordinates.

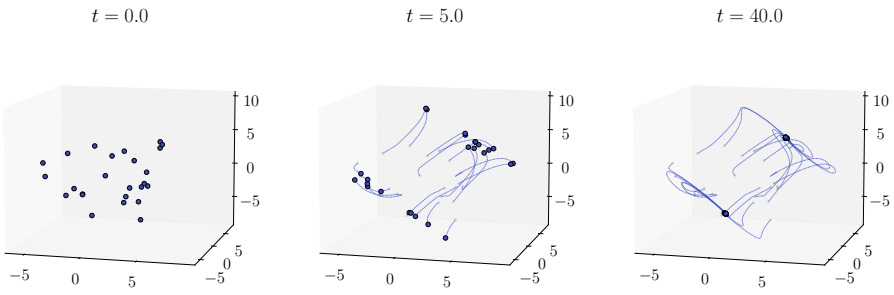


Figure 13. We consider $n = 25$, $Q = K = I_d$, and V a random matrix with positive entries and eigenvalues $\{1, 0.1 + 0.08i, 1 - 0.08i\}$. The pair of complex eigenvalues have a positive real part. We not only see convergence to one of two hyperplanes determined by the direction $\varphi_1 = (0.38, 0.8, 0.47)$, but in fact, the particles appear to collapse to two points. In other words, the "hyperplanes" are of codimension 3, which is in line with Conjecture 4.3.

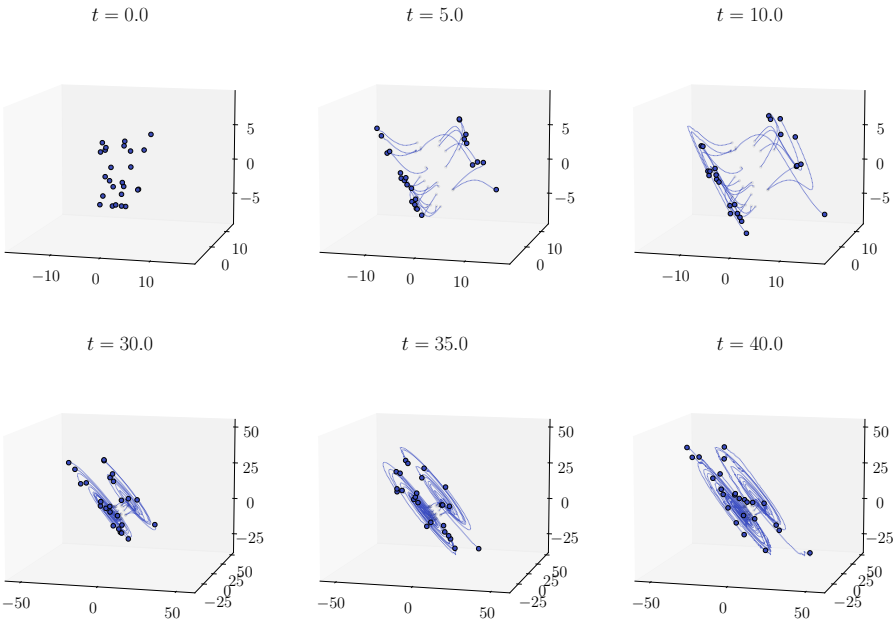


Figure 14. We consider $n = 25$, $Q = K = I_d$, and V a random matrix with positive entries and eigenvalues $\{1, -0.05 + 0.25i, -0.05 - 0.25i\}$. The pair of complex eigenvalues have a negative real part, which entails the rotation of the particles. We see that the particles rotate within a couple of 2-dimensional hyperplanes determined by $\varphi_1 = (-0.3, -0.8, -0.45)$, as implied by Theorem 4.2.

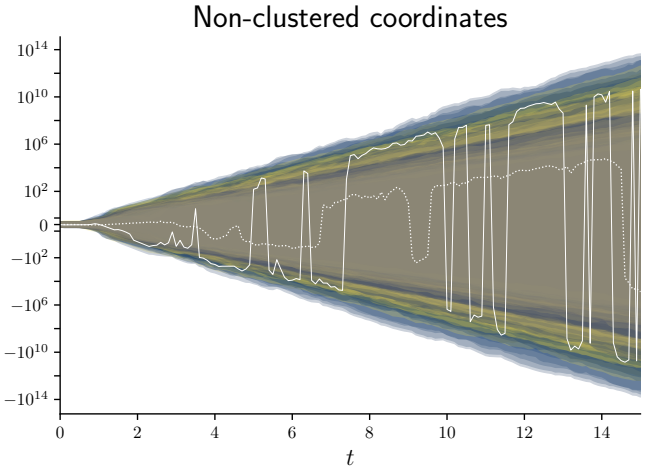


Figure 15. We complement Figure 7 and plot the variance of the set $\{\varphi_j^*(z_i(t))\colon i \in [n]\}$ of all coordinates j corresponding to negative eigenvalues of V. We also show the mean along tokens of a couple of coordinates (white lines). Coordinates diverge rapidly to $\pm\infty$ over time t; y-axis is in log scale.

Part 3. Discussion and open questions

12. Outlook

Several important directions regarding the mathematical theory of Transformers remain unexplored. An important extension of our work would amount to studying *multi-headed* Transformers—borrowing the notation from Remark 3.4, they amount to:

$$x_i^{[k+1]} = x_i^{[k]} + \Delta t \sum_{h=1}^{H} \sum_{j=1}^{n} \left(\frac{e^{\langle Q_h x_i^{[k]}, K_h x_j^{[k]} \rangle}}{\sum_{\ell=1}^{n} e^{\langle Q_h x_i^{[k]}, K_h x_\ell(k) \rangle}} \right) V_h x_j^{[k]}, \qquad k \in \mathbb{N}.$$

For each $h \in [H]$ (corresponding to a different *head*), the weight matrices Q_h, K_h, V_h are constant. Proofs regarding clustering or convergence of the self-attention matrix for such dynamics is an open problem. Preliminary numerical investigations seem to indicate that interesting clustering phenomena also occur in this context. A characterization or properties of optimal weights by invoking the optimal control correspondence in the spirit of [Wei17] is also an interesting avenue for future research.

We hereby list a couple of additional numerical experiments suggesting generalizations of our results, which we leave as open problems.

12.1. **Beyond $Q^\top K > 0$ in Theorems 3.1 and 5.2.** As seen throughout all the presented proofs, assumptions on the value matrix V are significantly more rigid than assumptions on the matrices Q and K. For instance, should the eigenvalue λ with the largest real part of V be negative, all rescaled tokens will diverge to infinity. Should λ be complex, we do not expect any clustering to occur (for the

rescaled tokens). Yet, none of the conclusions of Theorems 3.1 or 5.2 seem to change for generic choices of $Q^\top K$. This is illustrated in Figures 16 and 17 respectively.

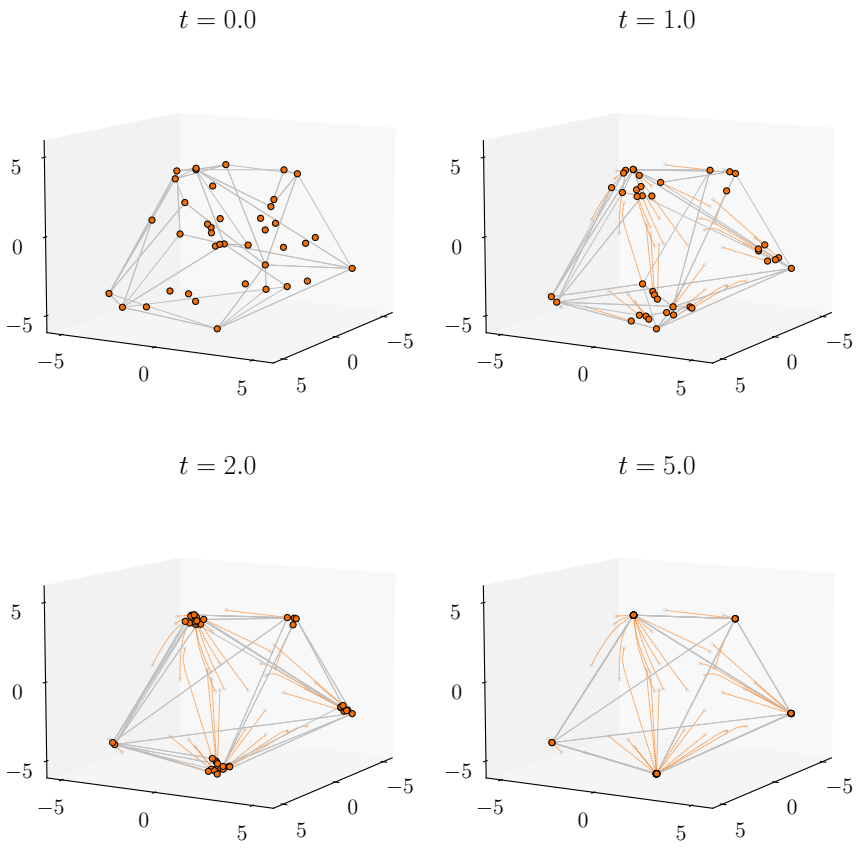


Figure 16. Here, $V = I_d$, while $Q^\top K$ violates the PSD assumption–it is a random matrix (with entries drawn from the uniform distribution on $[-1, 1]$). Nonetheless, the clustering pattern entailed by Theorem 3.1 persists.

12.2. **Beyond pure self-attention: adding a feed-forward layer.** Practical implementations of the Transformer architecture combine the self-attention mechanism with a feed-forward neural network. While extending the mathematical analysis from this paper to such a broader setting would be challenging, we can offer some numerical insights into the expected outcomes.

The feed-forward neural network which can be adjoined to the Transformer dynamics in one of two ways. The first way consists in running the pure self-attention dynamics up to time $t \leqslant T$ (or equivalently, for $O(T)$ layers), and then applying a pure feed-forward neural network to the concatenated vector of clustered features at time T. This amounts to seeing the feed-forward network as a map from $\mathbb{R}^{nd}$ to $\mathbb{R}^m$ (for some $m \geqslant 1$), which can be studied independently with existing theory.

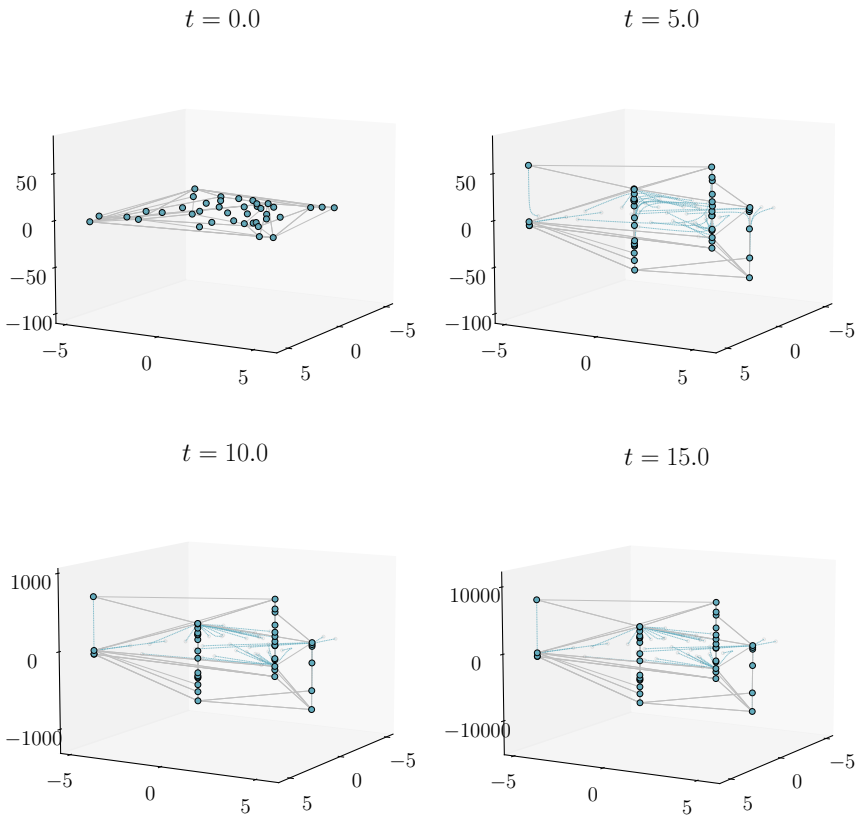


Figure 17. Here, V is paranormal, while $Q^\top K$ violates the PSD assumption–it is a random matrix (with entries drawn from the uniform distribution on $[-1, 1]$). Nonetheless, the clustering pattern entailed by Theorem 5.2 persists.

The second way consists in using both the self-attention and feed-forward mechanisms in parallel at every layer t. In this case, clustering in the exact sense of Theorems 3.1 and Theorems 5.2 would be difficult to anticipate since the weights of the feed-forward network play the role of a value matrix V (as they can be absorbed within V), and the conclusions of these theorems strongly depend on the identity-like structure.

In Figure 18, we focus on the second of the above-discussed examples, and illustrate a possible generalization of Theorem 4.2 to this setup. For simplicity, we focus on a 2-layer neural network: we apply a component-wise nonlinear activation function σ (either the ReLU or tanh) to the self-attention dynamics, and then multiply by a weight matrix $W \in \mathbb{R}^{d\times d}$. Namely, we consider

$$\dot{z}_i(t) = W\sigma\left(V\sum_{j=1}^{n}\left(\frac{e^{\langle Qe^{tV}z_i(t),Ke^{tV}z_j(t)\rangle}}{\sum_{k=1}^{n}e^{\langle Qe^{tV}z_i(t),Ke^{tV}z_k(t)\rangle}}\right)(z_j(t)-z_i(t))\right) \tag{12.1}$$

for $i \in [n]$ and $t \geqslant 0$. A bias vector $b \in \mathbb{R}^d$ (whether inside or outside the activation function) can also be included to allow for translations. The clustering property appears to persist, the pattern depending on the weight matrix W and on the activation function σ. We leave this problem open to further investigation.

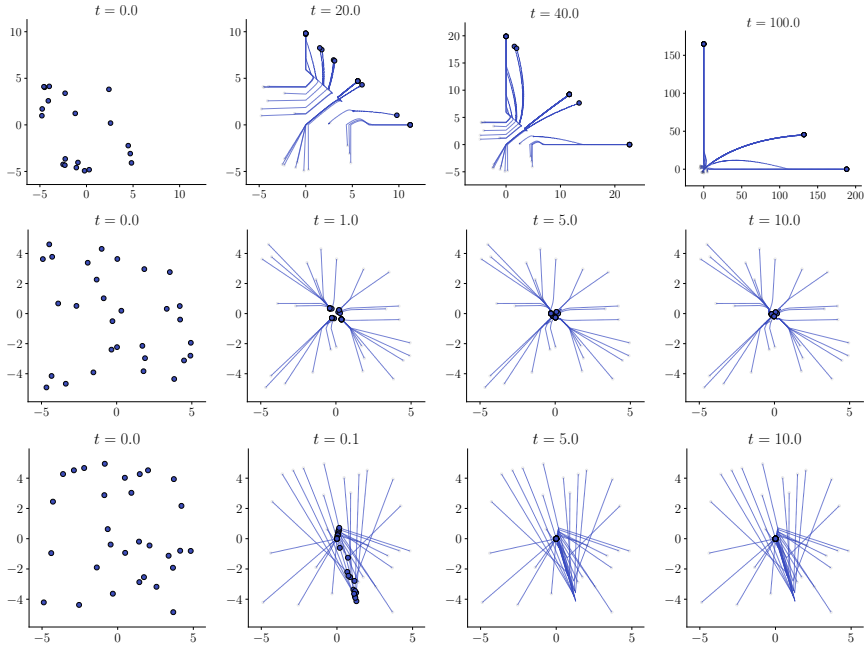


Figure 18. The setup of Theorem 4.2 with a 2-layer neural network appended to the dynamics (i.e., (12.1)). Top: $\sigma = \mathrm{ReLU}$ with $W = I_d$. Middle: $\sigma = \tanh$ with $W = I_d$. Bottom: $\sigma = \mathrm{ReLU}$ with W being a random matrix. In the first row, we see that the particles first evolve as to reach the upper right quadrant $(\mathbb{R}_{>0})^d$ (due to the ReLU). Once they reach it, every particle eventually follows one of three hyperplanes determined by the spectrum of V and the projection onto $(\mathbb{R}_{>0})^d$. In the other two cases, all particles appear to collapse to 0.

References

[ABV+05] Juan A Acebrón, Luis L Bonilla, Conrad J Pérez Vicente, Félix Ritort, and Renato Spigler. The Kuramoto model: A simple paradigm for synchronization phenomena. *Reviews of modern physics*, 77(1):137, 2005.

[BM00] Paul S Bradley and Olvi L Mangasarian. K-plane clustering. *Journal of Global optimization*, 16:23–32, 2000.

[Che95] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

[CHH+16] Junghee Cho, Seung-Yeal Ha, Feimin Huang, Chunyin Jin, and Dongnam Ko. Emergence of bi-cluster flocking for the Cucker–Smale model. *Mathematical Models and Methods in Applied Sciences*, 26(06):1191–1218, 2016.

[CRBD18] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018.

[CS07] Felipe Cucker and Steve Smale. Emergent behavior in flocks. *IEEE Transactions on Automatic Control*, 52(5):852–862, 2007.

[DCL21] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pages 2793–2803. PMLR, 2021.

[Dob79] Roland L'vovich Dobrushin. Vlasov equations. *Funktsional'nyi Analiz i ego Prilozheniya*, 13(2):48–58, 1979.

[Gol13] François Golse. Mean field kinetic equations. *Course of Polytechnique*, 2013.

[HJ12] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge University Press, 2012.

[HK02] Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence: models, analysis and simulation. *Journal of Artifical Societies and Social Simulation (JASSS)*, 5(3), 2002.

[HKPZ19] Seung-Yeal Ha, Jeongho Kim, Jinyeong Park, and Xiongtao Zhang. Complete cluster predictability of the Cucker–Smale flocking model on the real line. *Archive for Rational Mechanics and Analysis*, 231:319–365, 2019.

[HR17] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34(1), 2017.

[HysW+22] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[HZRS16a] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[HZRS16b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016.

[JM14] Pierre-Emmanuel Jabin and Sebastien Motsch. Clustering and asymptotic behavior in opinion formation. *Journal of Differential Equations*, 257(11):4165–4187, 2014.

[Kra00] Ulrich Krause. A discrete nonlinear and non-autonomous model of consensus. In *Communications in Difference Equations: Proceedings of the Fourth International Conference on Difference Equations*, page 227. CRC Press, 2000.

[Kur75] Yoshiki Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. In *International Symposium on Mathematical Problems in Theoretical Physics: January 23–29, 1975, Kyoto University, Kyoto/Japan*, pages 420–422. Springer, 1975.

[LCG+20] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*, 2020.

[LLH+20] Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. In *International Conference on Learning Representations*, 2020.

[LWLQ22] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, 3:111–132, 2022.

[MT14] Sebastien Motsch and Eitan Tadmor. Heterophilious dynamics enhances consensus. *SIAM Review*, 56(4):577–621, 2014.

[PHD20] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[PR13] Benedetto Piccoli and Francesco Rossi. Transport equation with nonlocal velocity in Wasserstein spaces: convergence of numerical schemes. *Acta Applicandae Mathematicae*, 124:73–105, 2013.

[PR16] Benedetto Piccoli and Francesco Rossi. On properties of the generalized Wasserstein distance. *Archive for Rational Mechanics and Analysis*, 222:1339–1365, 2016.

[PRT15] Benedetto Piccoli, Francesco Rossi, and Emmanuel Trélat. Control to flocking of the kinetic Cucker–Smale model. *SIAM Journal on Mathematical Analysis*, 47(6):4685–4719, 2015.

[RS14] Brian Rider and Christopher D. Sinclair. Extremal laws for the real Ginibre ensemble. *The Annals of Applied Probability*, 24(4):1621 – 1651, 2014.

[SABP22] Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pages 3515–3530. PMLR, 2022.

[VCBJ+95] Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet. Novel type of phase transition in a system of self-driven particles. *Physical Review Letters*, 75(6):1226, 1995.

[Vid11] René Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

[VKF20] Apoorv Vyas, Angelos Katharopoulos, and François Fleuret. Fast transformers with clustered attention. *Advances in Neural Information Processing Systems*, 33:21665–21674, 2020.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[Wei17] E Weinan. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 1(5):1–11, 2017.

[WHL19] E Weinan, Jiequn Han, and Qianxiao Li. A mean-field optimal control formulation of deep learning. *Research in Mathematical Sciences*, 6(1):10, 2019.

[WLK+20] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[YBR+20] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020.

Borjan Geshkovski
Department of Mathematics
MIT
Cambridge, MA
02139 USA
e-mail: borjan@mit.edu

Cyril Letrouit
Department of Mathematics
MIT
Cambridge, MA
02139 USA
e-mail: letrouit@mit.edu

Yury Polyanskiy
Department of EECS
MIT
Cambridge, MA
02139 USA
e-mail: yp@mit.edu

Philippe Rigollet
Department of Mathematics
MIT
Cambridge, MA
02139 USA
e-mail: rigollet@mit.edu