BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 00, Number 0, Pages 000–000
S 0273-0979(XX)0000-0

A MATHEMATICAL PERSPECTIVE ON TRANSFORMERS

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

Abstract. Transformers play a central role in the inner workings of large language models. We develop a mathematical framework for analyzing Transformers based on their interpretation as interacting particle systems, with a particular emphasis on long-time clustering behavior. Our study explores the underlying theory and offers new perspectives for mathematicians as well as computer scientists.

Contents

1. Outline 2

Part 1. Modeling 4
2. Interacting particle system 4
3. Measure to measure flow map 9

Part 2. Clustering 17
4. A single cluster for small β 18
5. A single cluster for large β 20
6. The high-dimensional case 20

Part 3. Further questions 27
7. Dynamics on the circle 28
8. BBGKY hierarchy 31
9. General matrices 32
10. Approximation, control, training 36
Acknowledgments 37

Appendix 37
Appendix A. Proof of Theorem 4.1 37
Appendix B. Proof of Theorem 5.1 39
Appendix C. Proof of Theorem 4.3 42
Appendix D. Proof of Theorem 6.9 43
References 47

2020 *Mathematics Subject Classification.* Primary: 34D05, 34D06, 35Q83; Secondary: 52C17.
Key words and phrases. Transformers, self-attention, interacting particle systems, clustering, gradient flows.

©0000 (copyright holder)

1. Outline

The introduction of *Transformers* in 2017 by Vaswani et al. [VSP+17] marked a significant milestone in the development of neural network architectures. Central to this contribution is *self-attention*, a novel mechanism which distinguishes Transformers from traditional architectures, and which plays a substantial role in their superior practical performance. In fact, this innovation has been a key catalyst for the progress of artificial intelligence in areas such as computer vision and natural language processing, notably with the emergence of large language models. As a result, understanding the mechanisms by which Transformers, and especially self-attention, process data is a crucial yet largely uncharted research area.

A common characteristic of deep neural networks (DNNs) is their compositional nature: data is processed sequentially, layer by layer, resulting in a discrete-time dynamical system (we refer the reader to the textbook [GBC16] for a general introduction). This perspective has been successfully employed to model *residual neural networks*—see Section 2.1 for more details—as continuous-time dynamical systems called neural ordinary differential equations (neural ODEs) [CRBD18, E17, HR17]. In this context, an input $x(0) \in \mathbb{R}^d$, say an image, is evolving according to a given time-varying velocity field as $\dot{x}(t) = v_t(x(t))$ over some time interval $(0, T)$. As such, a DNN can be seen as a flow map $x(0) \mapsto x(T)$ from $\mathbb{R}^d$ to $\mathbb{R}^d$. Even within the restricted class of velocity fields $\{v_t\}_{t \geqslant 0}$ imposed by classical DNN architectures, such flow maps enjoy strong approximation properties as exemplified by a long line of work on these questions [LJ18, ZGUA20, LLS22, TG22, RBZ23, CLLS23].

Following [SABP22] and [VBC20], we observe that Transformers are in fact flow maps on $\mathcal{P}(\mathbb{R}^d)$, the space of probability measures over $\mathbb{R}^d$. To realize this flow map from measures to measures, Transformers evolve a *mean-field interacting particle system*. More specifically, every particle (called a *token* in this context) follows the flow of a vector field which depends on the empirical measure of all particles. In turn, the *continuity equation* governs the evolution of the empirical measure of particles, whose long-time behavior is of crucial interest. In this regard, our main observation is that particles tend to cluster under these dynamics. This phenomenon is of particular relevance in learning tasks such as *next-token prediction*, wherein one seeks to map a given input sequence (i.e., a sentence) of n tokens (i.e., words) onto a given next token. In this case, the output measure encodes the probability distribution of the next token, and its clustering indicates a small number of possible outcomes. Our results on a simplified but insightful toy model indicate that the limiting distribution is actually a point mass, leaving no room for diversity or randomness, which is at odds with practical observations. This apparent paradox is resolved by the existence of a long-time metastable state. As can be seen from Figures 3 and 5, the Transformer flow appears to possess two different time-scales: in a first phase, tokens quickly form a few clusters, while in a second (much slower) phase, through the process of pairwise merging of clusters, all tokens finally collapse to a single point. This appears to corroborate behavior observed empirically in trained Transformer models, which goes under the names *token uniformity*, *oversmoothing* [CZC+22, RZZD23, GWDW23, WAWJ24, WAW+24, DBK24, SWJS24], or *rank-collapse* [DCL21, FZH+22, NAB+22, JDB23, ZMZ+23, ZLL+23, NLL+24, BHK24, CNQG24]; see also Figure 1.

The goal of this manuscript is twofold. On the one hand, we aim to provide a general and accessible framework to study Transformers from a mathematical perspective. In particular, the structure of these interacting particle systems enables concrete connections to established mathematical topics, such as nonlinear transport equations, Wasserstein gradient flows, collective behavior models, and optimal configurations of points on spheres, among others. On the other hand, we describe several promising research directions with a particular focus on the long-time clustering phenomenon. The main results we present are new, and we also provide what we believe are interesting open problems throughout the paper.

The rest of the paper is arranged in three parts.

Part 1: Modeling. We define an idealized model of the Transformer architecture that captures two of the main characteristics of transformers: *self-attention* and *layer-normalization*. Following a perspective put forward in classical architectures such as ResNets [CRBD18, E17, HR17], we view the successive layers of a neural network as time discretizations of a dynamical system of interacting particles. Layer-normalization effectively constrains particles to evolve on the unit sphere $\mathbb{S}^{d-1}$, whereas self-attention is the particular nonlinear coupling of the particles done through the empirical measure (Section 2). In turn, the empirical measure evolves according to the continuity partial differential equation (Section 3). Even after significant simplifications, this toy model retains macroscopic characteristics of trained Transformers, namely clustering. We also introduce a simpler surrogate model for self-attention which has the convenient property of being a Wasserstein gradient flow [AGS05] for an energy functional that is well-studied in the context of optimal configurations of points on the sphere and sheds a complementary light of the source of clustering.

Part 2: Clustering. In this part we recall existing and establish new mathematical results that indicate clustering of tokens in the long-time limit. (See Figure 2 for a summary.) Our first results (Theorem 4.3 in Section 4 and Theorem 5.1 in Section 5) are in extreme regimes of a temperature parameter β^{-1} that appears in the equation. We then move to the high-dimensional case in Section 6, where we begin by recalling Theorem 6.1—a result of [MTG17], recently revisited in [CRMB24]—which entails long-time clustering at any temperature when $d \geqslant 3$. We provide an exponential rate of convergence in Theorem 6.3 when $d \geqslant n$—here n denotes the number of particles—. We complement this result with an even more precise characterization of the rate of contraction of particles into a cluster. Namely, we describe the histogram of all inter-particle distances, and the time at which all particles are already nearly clustered (Theorem 6.9).

Part 3: Further questions. We propose potential avenues for future research, largely in the form of open questions substantiated by numerical observations. We first focus on the case $d = 2$ (Section 7) and elicit a link to Kuramoto oscillators. We briefly show in Section 9.1 how a simple and natural modification of our model leads to non-trivial questions related to optimal configurations on the sphere. The remaining sections explore interacting particle systems that allow for parameter tuning of the Transformers architectures, a key feature of practical implementations.

Part 1. Modeling

We begin this part by presenting the mathematical model for a Transformer in Section 2. Throughout the paper we focus on a simplified version that includes the self-attention mechanism as well as layer normalization (Section 2.2), but excludes additional feed-forward layers commonly used in practice (Section 2.3). This nonetheless leads to a highly nonlinear mean-field interacting particle system. In turn, this system implements, via the continuity equation, a flow map from initial to terminal distributions of particles that we present in Section 3.

2. Interacting particle system

Before writing down the Transformer model, we first provide a brief preliminary discussion to clarify our methodological choice of treating the discrete layer indices in the model as a continuous time variable in Section 2.1, echoing previous work on ResNets. The specifics of the toy Transformer model are presented in Section 2.2, and a complete model is presented in Section 2.3.

2.1. Residual neural networks.

One of the standard paradigms in machine learning is that of supervised learning, where one aims to approximate an unknown function $f : \mathbb{R}^d \to \mathbb{R}^m$, from data, $\mathcal{D} = \{x^{(i)}, f(x^{(i)})\}_{i\in[N]}$ say. This is typically done by choosing one among an arsenal of possible parametric models, whose parameters are then fit to the data by means of minimizing some user-specified cost. With the advent of graphical processing units (GPUs) in the realm of computer vision [KSH12], large neural networks have become computationally accessible, resulting in their popularity as one such parametric model.

Within the class of neural networks, *residual neural networks* (ResNets for short) have become a staple DNN architecture since their introduction in [HZRS16]. In their most basic form, ResNets approximate a function f at $x \in \mathbb{R}^d$ through a sequence of affine transformations, a component-wise nonlinearity, and skip connections. Put in formulae,

$$\begin{cases} x(k+1) = x(k) + w(k)\sigma(a(k)x(k) + b(k)) & \text{for } k \in \{0, \dots, L-1\} \\ x(0) = x\,. \end{cases} \tag{2.1}$$

Here σ is a Lipschitz function applied component-wise to the input vector, while $\theta(\cdot) = (w(\cdot), a(\cdot), b(\cdot)) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d} \times \mathbb{R}^d$ are trainable parameters. We say that (2.1) has $L \geqslant 1$ hidden layers (or $L+1$ layers, or is of depth L). The output $x_i(L)$ serves as a *representation* of the input $x^{(i)}$ that is then fed into a last layer that corresponds to a classical learning task such as linear or logistic regression in order to predict the label $f(x^{(i)})$. One can also devise generalizations of (2.1), for instance in which matrix-vector multiplications are replaced by discrete convolutions in order to reflect other common architectures such as convolutional neural networks [GBC16, Chapter 9]. The key element that all these models share is that they all have *skip-connections*, namely, the previous step $x_i(k)$ appears explicitly in the iteration for the next one.

One upside of (2.1), which is the one of interest to our narrative, is that the layer index k can naturally be interpreted as a time variable, motivating the continuous-time analogue

$$\begin{cases} \dot{x}(t) = w(t)\sigma(a(t)x(t) + b(t)) & \text{for } t \in (0, T) \\ x(0) = x. \end{cases} \tag{2.2}$$

These are dubbed *neural ordinary differential equations* (neural ODEs). Since their introduction in [CRBD18, E17, HR17], neural ODEs have emerged as a flexible mathematical framework to implement and study ResNets. We use this abstraction here for convenience, as dynamical systems are more naturally defined in continuous time. Although this approach is helpful for exposition, we emphasize that all the results presented can also be derived in discrete time using the same techniques.

2.2. **The interacting particle system.** Unlike ResNets, which operate on a single input vector $x(0) \in \mathbb{R}^d$ at a time, Transformers operate on a sequence of vectors of length n, namely, $(x_i(0))_{i\in[n]} \in (\mathbb{R}^d)^n$. This perspective is rooted in natural language processing, where each vector represents a word, and the entire sequence a sentence or a paragraph. In particular, it allows to process words together with their context. A sequence element $x_i(0) \in \mathbb{R}^d$ is called a *token*, and the entire sequence $(x_i(0))_{i\in[n]}$ a *prompt*. We use the words "token" and "particle" interchangeably.

All practical implementations of Transformers make use of *layer normalization* [BKH16], most commonly in the form of *root mean square (RMS) normalization* [ZS19]. (See lines 105–116 in [Mis24] for instance.) RMS normalization takes the sequence of tokens output after each layer, divides each token by its Euclidean norm[1] (plus a small parameter to avoid a possible division by zero), and multiplies the result by a trained diagonal matrix. This process aims to ensure that tokens do not diverge, thus avoiding rounding errors and overflow. The result is an evolution on a time-varying axis-aligned ellipsoid. To simplify the presentation and obtain insight and precise results, we assume that the trained diagonal matrix is equal to the idenitity, so that we work on the unit sphere $\mathbb{S}^{d-1}$ throughout. This simplification is justified empirically in the trained ALBERT XLarge v2 model described in Figure 1, wherein this diagonal matrix is constant over all layers, with entries of mean value equal to 0.44 and standard deviation equal to 0.008. Furthermore, current text embeddings provided by OpenAI, namely `text-embedding-3-small` and `text-embedding-3-large`, return norm-one embedding vectors. While we can only speculate as to the actual implementation of these models, this is an indication that layer normalization could be as simple as the one used in our toy model.

A Transformer is then a flow map on $(\mathbb{S}^{d-1})^n$: the input sequence $(x_i(0))_{i\in[n]} \in (\mathbb{S}^{d-1})^n$ is an initial condition which is evolved through the dynamics

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\frac{1}{Z_{\beta,i}(t)} \sum_{j=1}^{n} e^{\beta \langle Q(t)x_i(t), K(t)x_j(t)\rangle} V(t)x_j(t) \right) \tag{2.3}$$

for all $i \in [n]$ and $t \geqslant 0$. (We refer the reader to (2.8) and Section 2.3 for the full model.) Here and henceforth

$$\mathbf{P}^{\perp}_{x} y = y - \langle x, y \rangle x$$

[1]The original form instead consisted in an entry-wise standardization of every token, namely subtracting the mean of all tokens, then dividing by the standard deviation.

denotes the projection of $y \in \mathbb{R}^d$ onto $\mathrm{T}_x\mathbb{S}^{d-1}$. The *partition function* $Z_{\beta,i}(t) > 0$ reads

$$Z_{\beta,i}(t) = \sum_{k=1}^{n} e^{\beta\langle Q(t)x_i(t), K(t)x_k(t)\rangle}. \tag{2.4}$$

where $(Q(\cdot), K(\cdot), V(\cdot))$ (standing for Query, Key, and Value) are parameter matrices learned from data, and $\beta > 0$ a fixed number intrinsic to the model[2], which, can be seen as an inverse temperature using terminology from statistical physics. Note that $Q(\cdot), K(\cdot)$ need not be square.

The interacting particle system (2.3)–(2.4), a simplified version of which was first written down in [LLH+20, DGCC21, SABP22], importantly contains the true novelty that Transformers carry with regard to other models: the *self-attention mechanism*

$$A_{ij}(t) := \frac{e^{\beta\langle Q(t)x_i(t), K(t)x_j(t)\rangle}}{Z_{\beta,i}(t)}, \qquad (i,j) \in [n]^2, \tag{2.5}$$

which is the nonlinear coupling mechanism in the interacting particle system. The $n \times n$ stochastic matrix $A(t)$ (rows are probability vectors) is called the *self-attention matrix*. The wording *attention* stems from the fact that $A_{ij}(t)$ captures the attention given by particle i to particle j relatively to all particles $\ell \in [n]$. In particular, a particle pays attention to its neighbors where neighborhoods are dictated by the matrices $Q(t)$ and $K(t)$ in (2.5).

It has been observed numerically that the probability vectors $(A_{ij}(\cdot))_{j\in[n]}$ (for $i \in [n]$) in a trained self-attention matrix exhibit behavior related to the syntactic and semantic structure of sentences in natural language processing tasks (see [VSP+17, Figures 3-5]). To illustrate our conclusions as pedagogically as possible, throughout the paper we focus on a simplified scenario wherein the parameter matrices (Q, K, V) are constant, and even all equal to the identity unless stated otherwise, resulting in the dynamics

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\frac{1}{Z_{\beta,i}(t)} \sum_{j=1}^{n} e^{\beta\langle x_i(t), x_j(t)\rangle} x_j(t) \right) \tag{SA}$$

for $i \in [n]$ and $t \geqslant 0$ and, as before

$$Z_{\beta,i}(t) = \sum_{k=1}^{n} e^{\beta\langle x_i(t), x_k(t)\rangle}. \tag{2.6}$$

The appearance of clusters in Transformers is actually corroborated by numerical experiments with trained models (see Figure 1). While we focus on a much simplified model, numerical evidence shows that the clustering phenomenon looks qualitatively the same in the cases $Q = K = V = \lambda I_d$, $\lambda > 0$, and generic random (Q, K, V) (see Figures 3 and 5 for instance). We refer the interested reader directly to Sections 4, 5, 6; here, we continue the presentation on the modeling of different mechanisms appearing in the Transformer architecture.

---

[2]In practical implementations the inner products are multiplied by $d^{-\frac{1}{2}}$, which along with the typical magnitude of Q, K leads to the appearance of β.

[3]ALBERT XLarge v2 contains all the mechanisms described in this text, namely, is a system of the form (2.8) (or rather the discretization thereof) with 12 or 24 layers. The sequence length n is of the order of 512 or 1024, and the tokens evolve in $\mathbb{R}^{4096}$. The dynamics are therefore high-dimensional, lending weight to assumptions made later on (Section 6).

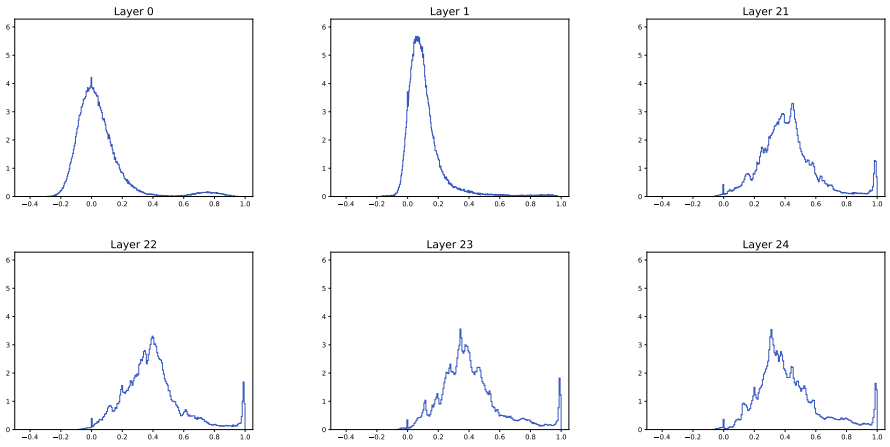
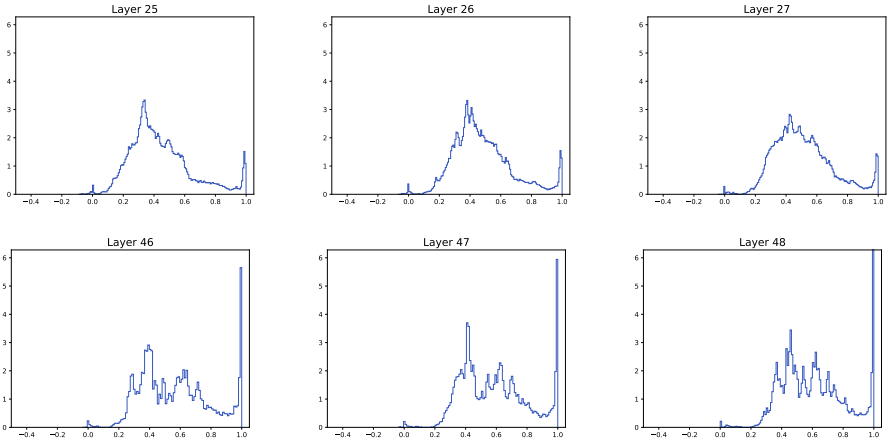


Figure 1. Histogram of $\{\langle x_i(t), x_j(t)\rangle\}_{(i,j)\in[n]^2, i\neq j}$ at different layers t in the context of the trained ALBERT XLarge v2 model ([LCG+20] and `https://huggingface.co/albert-xlarge-v2`)[3], which has constant parameter matrices. Here we randomly selected a single prompt, which in this context is a paragraph from a random Wikipedia entry, and then generate the histogram of the pairwise inner products. We see the progressive emergence of clusters all the way to the 24th (and last) hidden layer (top), as evidenced by the growing mass at 1. If the number of layers is increased, up to 48 say, the clustering is further enhanced (bottom).



Remark 2.1 (Collective behavior)**.** *The dynamics* (SA) *have a strong resemblance to the vast literature on nonlinear systems arising in the modeling of collective behavior. In addition to the connection to the classical Kuramoto model describing synchronization of oscillators* [Kur75, ABV+05] *(made evident in Section 7.2), Transformers are perhaps most similar to 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 (2.3)*. 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 related models include those of Vicsek* [VCBJ+95]*, Hegselmann-Krause* [HK02] *and Cucker-Smale* [CS07]*. All these models exhibit a clustering behavior under various assumptions (see* [MT14, Tad23] *and the references therein). Yet, none of the opinion dynamics models discussed above contain parameters appearing within the nonlinearity as in* (SA)*, whilst set on the sphere.*

Remark 2.2 (Permutation equivariance)**.** *A function $f : (\mathbb{S}^{d-1})^n \to (\mathbb{S}^{d-1})^n$ is permutation equivariant if $f(\pi X) = \pi(f_1(X), \dots, f_n(X))$ for any $X \in (\mathbb{R}^d)^n$ and for any permutation $\pi \in \mathbf{S}_n$ of n elements. Otherwise put, if we permute the input X, then the output $f(X)$ is permuted in the same way. Given $t > 0$, the Transformer* (SA)*, mapping $(x_i(0))_{i\in[n]} \mapsto (x_i(t))_{i\in[n]}$, is permutation-equivariant on $(\mathbb{S}^{d-1})^n$.*

2.3. **Toward the complete Transformer.** There are a couple of additional mechanisms used in practical implementations that we do not explicitly address or use in this study. The mathematical analysis of these mechanisms remains open.

2.3.1. *Multi-headed attention.* Practical implementations spread out the computation of the self-attention mechanism at every t through a sequence of *heads*, leading to the so-called *multi-headed self attention*. This consists in considering the following modification to (SA):

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\sum_{h=1}^{H} \sum_{j=1}^{n} \frac{e^{\beta \langle Q_h(t)x_i(t), K_h(t)x_j(t)\rangle}}{Z_{\beta,i,h}(t)} V_h(t)x_j(t) \right) \tag{2.7}$$

where $Z_{\beta,i,h}(t)$ is defined as in (2.4) for the matrices $Q_h(t)$ and $K_h(t)$. The integer $H \geqslant 1$ is called the number of heads[4].

The introduction of multiple heads also allows for drawing some interesting parallels with the literature on feed-forward neural networks, such as ResNets (2.1). Considerable effort has been expended to understand 2-layer neural networks with width tending to $+\infty$; more precisely, consider (2.1) with $L = 1$, $w \in \mathbb{R}^{d\times\ell}$, $a \in \mathbb{R}^{\ell\times d}$, and $\ell \to +\infty$. The infinite-width limit for Transformers is in fact very natural, as it is realized by stacking an arbitrary large number of heads: $H \to +\infty$. Hence, the same questions as for 1-hidden layer neural networks may be asked: for instance the question of universal approximation, in the vein of [Cyb89, Bar93].

2.3.2. *Feed-forward layers.* The complete Transformer dynamics combines all of the above mechanisms with a feed-forward layer; in the discrete-time context, this is actually done by using a Lie-Trotter splitting scheme for

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\sum_{h=1}^{H} \sum_{j=1}^{n} \frac{e^{\beta \langle Q_h(t)x_i(t), K_h(t)x_j(t)\rangle}}{Z_{\beta,i,h}(t)} V_h(t)x_j(t) + w(t)\sigma(a(t)x_i(t) + b(t)) \right), \tag{2.8}$$

[4]In practical implementations, H is a divisor of d, and the query and key matrices $Q_h(t)$ and $K_h(t)$ are $\frac{d}{H} \times d$ rectangular. This allows for further parallelization of computations and increased expressiveness. For mathematical purposes, we focus on working with arbitrary integers H, and square weight matrices Q_h and K_h.

where $w(t), a(t), b(t)$ and σ are all as in (2.2). The interested reader is referred to [LLH+20, PH22] for all the details[5]. The feed-forward layers (convolutional layers can alternatively be considered) are of critical importance in applications and drive the existing results on approximation properties of Transformers [YBR+19]. Nevertheless, the analysis of this model is beyond the scope of our current methods.

3. Measure to measure flow map

An important aspect of Transformers is that they are not hard-wired to take into account the order of the input sequence, contrary to other architectures used for natural language processing such as recurrent neural networks. In these applications, each token $x_i(0) \in \mathbb{R}^d$ contains not only a word embedding $w_i \in \mathbb{R}^d$, but also an additional *positional encoding* (we postpone a discussion to Remark 3.2) which allows tokens to also carry their position in the input sequence. Therefore, an input sequence is perfectly encoded as a *set* of tokens $\{x_1(0), \dots, x_n(0)\}$, or equivalently as the empirical measure of its constituent tokens $\frac{1}{n}\sum_{i=1}^n \delta_{x_i(0)}$. Recall that the output of a Transformer is also a probability measure, namely $\frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$, albeit one that captures the likelihood of the next token. As a result, one can view Transformers as flow maps between probability measures[6] on $\mathbb{S}^{d-1}$. To describe this flow map, we appeal to the continuity equation, which governs precisely the evolution of the empirical measure of particles subject to dynamics. This perspective is already present in [SABP22], the only modification here being that we add the projection on the sphere arising from layer normalization.

After introducing the continuity equation in Section 3.1, we show that a particular interaction energy functional, which is maximized at any point mass, increases along solutions thereof in Section 3.2. Motivated by this monotonicity property, in Section 3.3 we propose an illustrative modified model which has the nice property of being a Wasserstein gradient flow for this energy. Finally, in Section 3.4, we demonstrate that the original equation presented in Section 3.1 is itself a gradient flow for the same energy, upon changing the metric underlying the definition of the gradient.

3.1. The continuity equation.

The vector field driving the evolution of a single particle in (SA) clearly depends on all n particles. In fact, one can equivalently rewrite the dynamics as

$$\dot{x}_i(t) = \mathcal{X}[\mu(t)](x_i(t)) \tag{3.1}$$

for all $i \in [n]$ and $t \geqslant 0$, where

$$\mu(t, \cdot) = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}(\cdot)$$

is the empirical measure, while the vector field $\mathcal{X}[\mu] : \mathbb{S}^{d-1} \to \mathrm{T}\mathbb{S}^{d-1}$ reads

$$\mathcal{X}[\mu](x) = \mathbf{P}_x^{\perp}\left(\frac{1}{Z_{\beta,\mu}(x)}\int e^{\beta\langle x,y\rangle} y \,\mathrm{d}\mu(y)\right) \tag{3.2}$$

[5] and lines 123–130 in [Ope24] for some relevant source code.

[6] See [DBPC19, VBC20, ZB21] for further related work on neural networks acting on probability measures.

with

$$Z_{\beta,\mu}(x) = \int e^{\beta\langle x,y\rangle}\,\mathrm{d}\mu(y). \tag{3.3}$$

In other words, (SA) is a *mean-field interacting particle system*. The evolution of $\mu(t)$ is governed by the continuity equation[7]

$$\begin{cases} \partial_t \mu + \operatorname{div}(\mathcal{X}[\mu]\mu) = 0 & \text{on } \mathbb{R}_{\geqslant 0} \times \mathbb{S}^{d-1} \\ \mu_{|t=0} = \mu(0) & \text{on } \mathbb{S}^{d-1} \end{cases} \tag{3.4}$$

satisfied in the sense of distributions.

Remark 3.1 (Well-posedness)**.** *Global existence of weak, measure-valued solutions to* (3.4) *for arbitrary initial conditions* $\mu(0) \in \mathcal{P}(\mathbb{S}^{d-1})$ *follows by arguing exactly as in* [GLPR24, Lemma A.3]. *Here and henceforth,* $\mathcal{P}(\mathbb{S}^{d-1})$ *stands for the set of Borel probability measures on* $\mathbb{S}^{d-1}$.

Remark 3.2 (Positional encoding)**.** *For the sake of completeness, in this brief segue we discuss a few ways to perform positional encoding. The original one, proposed in* [VSP+17]*, proceeds as follows. Consider a sequence* $(w_i)_{i\in[n]} \in (\mathbb{R}^d)^n$ *of word embeddings. Then the positional encoding* $p_i \in \mathbb{R}^d$ *of the* i*-th word embedding is defined as* $(p_i)_{2k} = \sin(\frac{i}{M^{2k/d}})$ *and* $(p_i)_{2k+1} = \cos(\frac{i}{M^{2k/d}})$ *for* $k \in [d/2 - 1]$*, and* $M > 0$ *is a user-defined scalar equal to* 10^4 *in* [VSP+17]*. The* i*-th token is then defined as the addition:* $x_i(0) = w_i + p_i$*. Subsequent works simply use either a random*[8] *positional encoding (i.e.,* p_i *is just some random vector) or a trained transformation. The addition can also be replaced with a concatenation* $x_i(0) = [w_i; p_i]$*. (See* [LWLQ22, XZ23] *for details.)*

Remark 3.3 (Mean field limit)**.** *Although the analysis in this paper is focused on the flow of the empirical measure, one can also consider* (3.4) *for arbitrary initial probability measures* $\mu(0) \in \mathcal{P}(\mathbb{S}^{d-1})$*. Both views can be linked through a mean-field limit-type result, which can be shown by making use of the Lipschitz nature of the vector field* $\mathcal{X}[\mu]$*. The argument is classical and dates back at least to the work of Dobrushin* [Dob79]*. Consider an initial empirical measure* $\mu_n(0) = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(0)}$*, and suppose that the points* $x_i(0)$ *are such that* $\lim_{n\to+\infty} W_1(\mu_n(0), \mu(0)) = 0$ *for some probability measure* $\mu(0) \in \mathcal{P}(\mathbb{S}^{d-1})$*. (Here* W_1 *denotes 1-Wasserstein distance—see* [Vil09] *for definitions—.) Consider the solutions* $\mu_n(t)$ *and* $\mu(t)$ *to* (3.4) *with initial data* $\mu_n(0)$ *and* $\mu(0)$ *respectively. Dobrushin's argument is then centered around the estimate*

$$W_1(\mu_n(t), \mu(t)) \leqslant e^{O(1)|t|} W_1(\mu_n(0), \mu(0))$$

for any $t \in \mathbb{R}$*, which in the case of* (3.4) *can be shown without much difficulty (see* [Vil01, Chapter 4, Section 1] *or* [Gol16, Section 1.4.2]*). This elementary mean-field limit result has a couple of caveats. First, the time-dependence is exponential. Second, if one assumes that the points* $x_i(0)$ *are sampled i.i.d. according to* μ_0*, then* $W_1(\mu_n(0), \mu(0))$ *converges to zero at rate* $n^{-\frac{1}{d-1}}$ [Dud69, BLG14]*, which deteriorates quickly when* d *grows. Dimension-free convergence has been established in*

[7]Unless stated otherwise, ∇ and div henceforth stand for the spherical gradient and divergence respectively, and all integrals are taken over $\mathbb{S}^{d-1}$.

[8]This rationale supports the assumption that initial tokens are drawn at random, which we make use of later on.

some cases, for instance by replacing the Wasserstein distance with a more careful choice of metric as in [HHL23, Lac23] *or more generally in* [SFG$^+$12]. *Similarly, the exponential time-dependence might also be improved, as recent works in the context of flows governed by Riesz/Coulomb singular kernels, with diffusion, can attest* [RS23, GLBM24] *(see* [LLF23] *for a result in the smooth kernel case). We do not address this question in further detail here. For more references on this well-established topic, the reader is referred to* [Vil01, Gol16, Ser20] *and the references therein.*

3.2. **The interaction energy.** One can naturally ask whether the evolution in (3.4) admits some quantities which are monotonic when evaluated along the flow. As it turns out, the *interaction energy*

$$\mathsf{E}_\beta[\mu] = \frac{1}{2\beta} \iint e^{\beta\langle x,x'\rangle} \,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x') \tag{3.5}$$

is one such quantity. Indeed,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}_\beta[\mu(t)] &= \iint \beta^{-1} e^{\beta\langle x,x'\rangle} \,\mathrm{d}\partial_t\mu(t,x)\,\mathrm{d}\mu(t,x') \\
&= \int \mathcal{X}[\mu(t)](x) \cdot \int \nabla\left(\beta^{-1} e^{\beta\langle x,x'\rangle}\right) \mathrm{d}\mu(t,x')\,\mathrm{d}\mu(t,x) \\
&= \int \left\|\mathcal{X}[\mu(t)](x)\right\|^2 Z_{\beta,\mu(t)}(x)\,\mathrm{d}\mu(t,x) \qquad (3.6)
\end{aligned}$$

for any $t \geqslant 0$ by using integration by parts. Recalling the definition of $Z_{\beta,\mu}(x)$ in (3.3), we see that $e^{-\beta} \leqslant Z_{\beta,\mu}(x) \leqslant e^{\beta}$ for all $x \in \mathbb{S}^{d-1}$. The identity (3.6) therefore indicates that E_β increases along trajectories of (3.4). (Similarly, should $V = -I_d$, the energy E_β would decrease along trajectories.) This begs the question of characterizing the global minima and maxima of E_β, which is the goal of the following result.

Proposition 3.4. *Let $\beta > 0$ and $d \geqslant 2$. The unique global minimizer of E_β over $\mathcal{P}(\mathbb{S}^{d-1})$ is the uniform measure*[9] *σ_d. Any global maximizer of E_β over $\mathcal{P}(\mathbb{S}^{d-1})$ is a Dirac mass δ_{x^*} centered at some point $x^* \in \mathbb{S}^{d-1}$.*

This result lends credence to our nomenclature of the case $V = I_d$ as *attractive*, and $V = -I_d$ as *repulsive*. The reader should be wary however that in this result we are minimizing or maximizing E_β among *all* probability measures on $\mathbb{S}^{d-1}$. Should one focus solely on discrete measures, many global minima appear—these are discussed in Section 9.1—. This is one point where the particle dynamics and the mean-field flow deviate. We now provide a brief proof of Proposition 3.4 (see [Tan17] for a different approach).

Proof of Proposition 3.4. The fact that any global maximizer is a Dirac mass is easy to see. We proceed with proving the rest of the statement. Let $f(t) = e^{\beta t}$. The interaction energy then reads

$$\mathsf{E}_\beta[\mu] = \frac{1}{2} \iint f(\langle x,x'\rangle) \,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x').$$

[9]That is, the Lebesgue measure on $\mathbb{S}^{d-1}$, normalized to be a probability measure.

The proof relies on an ultraspherical (or Gegenbauer) polynomial expansion of $f(t)$:

$$f(t) = \sum_{k=0}^{+\infty} \widehat{f}(k;\lambda) \frac{k+\lambda}{\lambda} C_k^\lambda(t)$$

for $t \in [-1,1]$, where $\lambda = \frac{d-2}{2}$, C_k^λ are Gegenbauer polynomials, and

$$\widehat{f}(k;\lambda) = \frac{\Gamma(\lambda+1)}{\Gamma(\lambda+\frac{1}{2})\Gamma(\frac{1}{2})} \frac{1}{C_k^\lambda(1)} \int_{-1}^{1} f(t) C_k^\lambda(t)(1-t^2)^{\lambda-\frac{1}{2}} \,\mathrm{d}t$$

where $C_k^\lambda(1) > 0$ (see [DX13, Section 1.2]). According to [BD19, Proposition 2.2], a necessary and sufficient condition for Proposition 3.4 to hold is to ensure that $\widehat{f}(k;\lambda) > 0$ for all $k \geqslant 1$. To show this, we use the Rodrigues formula [Sze39, 4.1.72]

$$C_k^\lambda(t) = \frac{(-1)^k 2^k}{k!} \frac{\Gamma(k+\lambda)\Gamma(k+2\lambda)}{\Gamma(\lambda)\Gamma(2k+2\lambda)} (1-t^2)^{-(\lambda-\frac{1}{2})} \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^k (1-t^2)^{k+\lambda-\frac{1}{2}},$$

and the fact that $C_k^\lambda(-t) = (-1)^k C_k^\lambda(t)$ for $t \in [-1,1]$, which in combination with integration by parts yield

$$\int_{-1}^{1} t^\ell C_k^\lambda(t)(1-t^2)^{\lambda-\frac{1}{2}} \,\mathrm{d}t \begin{cases} > 0 & \text{if } \ell \geqslant k \text{ and } \ell - k \text{ is even} \\ = 0 & \text{otherwise .} \end{cases}$$

We conclude by using the power series expansion of f. □

3.3. **A Wasserstein gradient flow proxy.** In view of (3.6), one could hope to see the continuity equation (3.4) as the *Wasserstein gradient flow* of E_β, or possibly some other functional (see the seminal papers [Ott01, JKO98], and [AGS05, Vil09] for a complete treatment). The long-time asymptotics of the PDE can then be analyzed by studying convexity properties of the underlying functional, by analogy with gradient flows in the Euclidean case.

For (3.4) to be the Wasserstein gradient flow of E_β, the vector field $\mathcal{X}[\mu]$ defined in (3.2) ought to be the gradient of the first variation $\delta\mathsf{E}_\beta$ of E_β. However, notice that $\mathcal{X}[\mu]$ is a logarithmic derivative:

$$\mathcal{X}[\mu](x) = \nabla \log \int \beta^{-1} e^{\beta\langle x,y\rangle} \,\mathrm{d}\mu(y). \tag{3.7}$$

(This observation goes beyond $Q = K = I_d$ and $V = \pm I_d$ insofar as $Q^\top K = K^\top Q = \pm V$; see [SABP22, Assumption 1]. Because of the lack of symmetry, it has been shown in [SABP22] that (3.7) is not the gradient of the first variation of a functional.

To overcome this limitation on $\mathbb{R}^d$, thus without layer normalization, [SABP22] propose two ways to "symmetrize" (3.4) that both lead to a Wasserstein gradient flow; see [SABP22, Proposition 2]. We focus here on the simplest one which consists in removing the logarithm in (3.7), or equivalently to removing the denominator in (3.2). This is one point where working on the unit sphere is useful: otherwise, the equation on $\mathbb{R}^d$ without layer normalization (as considered in [SABP22]) is ill-posed for general choices of matrices V, due to the fact that the magnitude of the vector field $\mathcal{X}[\mu]$ grows exponentially with the size of the support of μ. On the contrary, on $\mathbb{S}^{d-1}$ the resulting equation is perfectly well-posed.

Remark 3.5 (Doubly stochastic kernel)**.** *Considering the Transformer dynamics on $\mathbb{R}^d$, thus without layer normalization, the authors in* [SABP22] *propose an alternative symmetric model: they replace the self-attention (stochastic) matrix by a doubly stochastic one, generated from the Sinkhorn iteration. This leads to a Wasserstein gradient flow, whereby the resulting attention mechanism is implicitly expressed as a limit of Sinkhorn iterations. Understanding the emergence of clusters for this model is an interesting but possibly challenging question.*

In view of the above discussion, we are inclined to propose the surrogate model

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\frac{1}{n} \sum_{j=1}^{n} e^{\beta \langle x_i(t), x_j(t) \rangle} x_j(t) \right), \tag{USA}$$

which is obtained by replacing the partition function $Z_{\beta,i}(t)$ by n. As a matter of fact, (USA) presents a remarkably similar qualitative behavior—all of the results we show in this paper are essentially the same for both dynamics—.

The continuity equation corresponding to (USA), namely

$$\begin{cases} \partial_t \mu(t,x) + \operatorname{div} \left(\mathbf{P}^{\perp}_{x} \left(\int e^{\beta \langle x, x' \rangle} x' \, \mathrm{d}\mu(t,x') \right) \mu(t,x) \right) = 0 \\ \mu_{|t=0} = \mu_0 \end{cases} \tag{3.8}$$

for $(t,x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{S}^{d-1}$, can now be seen as a Wasserstein gradient flow for the interaction energy E_β defined in (3.5).

Lemma 3.6. *Consider the interaction energy $\mathsf{E}_\beta : \mathcal{P}(\mathbb{S}^{d-1}) \to \mathbb{R}_{\geqslant 0}$ defined in* (3.5)*. Then the vector field*

$$\mathcal{X}[\mu](x) = \mathbf{P}^{\perp}_{x} \left(\int e^{\beta \langle x, x' \rangle} x' \, \mathrm{d}\mu(x') \right)$$

satisfies

$$\mathcal{X}[\mu](x) = \nabla \delta \mathsf{E}_\beta[\mu](x) \tag{3.9}$$

for any $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ and $x \in \mathbb{S}^{d-1}$, where $\delta \mathsf{E}_\beta[\mu]$ denotes the first variation of E_β.

We omit the proof which follows from standard Otto calculus [Ott01], [Vil09, Chapter 15], [CNWR24, Chapter 5]. We can actually write (3.9) more succinctly by recalling the definition of the convolution of two functions on $\mathbb{S}^{d-1}$ [DX13, Chapter 2]: for any $g \in L^1(\mathbb{S}^{d-1})$ and $f : [-1,1] \to \mathbb{R}$ such that $t \mapsto (1-t^2)^{\frac{d-3}{2}} f(t)$ is integrable,

$$(f * g)(x) = \int f(\langle x, y \rangle) g(y) \, \mathrm{d}\sigma_d(y).$$

This definition has a natural extension to the convolution of a function f (with the above integrability) and a measure $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$. We can hence rewrite

$$\mathsf{E}_\beta[\mu] = \frac{1}{2} \int (\mathsf{G}_\beta * \mu)(x) \, \mathrm{d}\mu(x)$$

where $[-1,1] \ni \mathsf{G}_\beta(t) = \beta^{-1} e^{\beta t}$, and so

$$\mathcal{X}[\mu](x) = \nabla (\mathsf{G}_\beta * \mu)(x).$$

Thus, (3.8) takes the equivalent form

$$\begin{cases} \partial_t \mu(t,x) + \operatorname{div}\Big(\nabla\big(\mathsf{G}_\beta * \mu(t,\cdot)\big)(x)\mu(t,x)\Big) = 0 & \text{for } (t,x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{S}^{d-1} \\ \mu_{|t=0} = \mu_0 & \text{for } x \in \mathbb{S}^{d-1}. \end{cases} \tag{3.10}$$

The considerations above lead us to the following Lyapunov identity.

Lemma 3.7. *The solution $\mu \in C^0(\mathbb{R}_{\geqslant 0}; \mathcal{P}(\mathbb{S}^{d-1}))$ to* (3.8) *satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}_\beta[\mu(t)] = \int \Big\|\nabla\Big(\mathsf{G}_\beta * \mu(t,\cdot)\Big)(x)\Big\|^2 \,\mathrm{d}\mu(t,x)$$

for $t \geqslant 0$.

Interestingly, (3.10) is an *aggregation* equation, versions of which have been studied in great depth in the literature. For instance, clustering in the spirit of an asymptotic collapse to a single Dirac measure located at the center of mass of the initial density $\mu(0,\cdot)$ has been shown for aggregation equations with singular kernels in [BCM08, BLR11, CDF$^+$11], motivated by the Patlak-Keller-Segel model of chemotaxis. Here, one caveat (and subsequently, novelty) is that (3.10) is set on $\mathbb{S}^{d-1}$ which makes the analysis developed in these references difficult to adapt or replicate.

Remark 3.8 (Particle version)**.** *Let us briefly sketch the particle version of the Wasserstein gradient flow* (3.8)*. When $\mu(t) = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$, the interaction energy* (3.5) *takes the form*

$$\mathsf{E}_\beta(X) = \frac{1}{2\beta n^2}\sum_{i=1}^n \sum_{j=1}^n e^{\beta\langle x_i, x_j\rangle}$$

where $X = (x_1, \dots, x_n) \in (\mathbb{S}^{d-1})^n$. Denoting by ∇_X the gradient associated to the standard Riemannian metric on $(\mathbb{S}^{d-1})^n$, we get the dynamics

$$\dot{X}(t) = n\nabla_X \mathsf{E}_\beta(X(t)). \tag{3.11}$$

Indeed, the gradient on $(\mathbb{S}^{d-1})^n$ is simply $\nabla = (\partial_1, \dots, \partial_n)$ where ∂_i is the gradient in $\mathbb{S}^{d-1}$ acting on the i-th copy in $(\mathbb{S}^{d-1})^n$. Therefore

$$\partial_i \mathsf{E}_\beta(X(t)) = \frac{1}{\beta n^2}\sum_{j=1}^n \mathbf{P}^\perp_{x_i(t)}\Big(e^{\beta\langle x_i(t), x_j(t)\rangle}\beta x_j(t)\Big) = \frac{1}{n}\dot{x}_i(t)$$

which yields (3.11)*.*

Note that (SA) *also corresponds to a gradient flow of the same interaction energy albeit with respect to a Riemannian metric on the sphere different from the standard one (for $n = 2$ the two are conformally equivalent). We provide more detail in the following section.*

3.4. (SA) **is a gradient flow for a modified metric.** We will now briefly demonstrate that for a particular choice of parameters (Q, K, V), the true dynamics (SA) can be seen as a gradient flow for E_β upon a modification of the metric on the tangent space of $(\mathbb{S}^{d-1})^n$. This will facilitate qualitative analysis later on by using standard tools from dynamical systems. This insight can then be extrapolated to the corresponding continuity equation (3.4) as well, as seen in Section 3.4.2.

3.4.1. *The case of particles.* We suppose that

$$Q^\top K \text{ is symmetric}, \qquad V = Q^\top K.$$

We define a new metric on $(\mathbb{S}^{d-1})^n$ as follows. Let $X = (x_1, \dots, x_n) \in (\mathbb{S}^{d-1})^n$. Consider the inner product on $\mathrm{T}_X(\mathbb{S}^{d-1})^n$ given by

$$\langle (a_1, \dots, a_n), (b_1, \dots, b_n) \rangle_X = \sum_{i=1}^n Z_{\beta,i}(X) \langle a_i, b_i \rangle, \tag{3.12}$$

where $a_i, b_i \in \mathrm{T}_{x_i}\mathbb{S}^{d-1}$, and

$$Z_{\beta,i}(X) = \sum_{j=1}^n e^{\beta \langle Vx_i, x_j \rangle}.$$

Set

$$\mathsf{E}_\beta(X) = \frac{1}{2\beta} \sum_{i=1}^n \sum_{j=1}^n e^{\beta \langle Vx_i, x_j \rangle}.$$

We now show that the dynamics (2.3) can be equivalently written as

$$\dot{X}(t) = \nabla \mathsf{E}_\beta(X(t)),$$

where the gradient ∇ is computed with respect to the metric (3.12) on $(\mathbb{S}^{d-1})^n$. To this end, we ought to show that for all vector fields Y on $(\mathbb{S}^{d-1})^n$ and for all $X \in (\mathbb{S}^{d-1})^n$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \mathsf{E}_\beta \left(\Phi_Y^t(X) \right) = \langle Y(X), B(X) \rangle_X \tag{3.13}$$

holds, where Φ_Y^t is the flow associated to the vector field Y, whereas $B = (B_1, \dots, B_n)$ with

$$B_i = \mathbf{P}_{x_i}^\perp \left(\frac{1}{Z_{\beta,i}(X)} \sum_{j=1}^n e^{\beta \langle Vx_i, x_j \rangle} Vx_j \right) \in \mathrm{T}_{x_i}\mathbb{S}^{d-1}.$$

By linearity, it is sufficient to show (3.13) for vector fields Y of the form

$$Y(X) = (Ax_1, 0, \dots, 0) \in \mathrm{T}_X(\mathbb{S}^{d-1})^n$$

where A is an arbitrary non-zero skew-symmetric matrix. Clearly

$$\Phi_Y^t(X) = (e^{tA}x_1, x_2, \dots, x_n). \tag{3.14}$$

One first computes

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \mathsf{E}_\beta \left(\Phi_Y^t(X) \right) = \sum_{j=1}^n e^{\beta \langle Vx_1, x_j \rangle} \langle Ax_1, Vx_j \rangle.$$

Now observe that $\langle Ax_1, y \rangle = \langle Ax_1, z \rangle$ for all skew-symmetric matrices A if and only if $x_1(y^\top - z^\top)$ is a symmetric matrix. Since $\mathbf{P}_{x_1}^\perp = I_d - x_1 x_1^\top$, we see that

$$\sum_{j=1}^n e^{\beta \langle Vx_1, x_j \rangle} \langle Ax_1, Vx_j \rangle = \langle Y(X), B(X) \rangle_X,$$

as desired.

3.4.2. *The case of measures.* The above insight, which consists in reweighing the metric with respect to which the gradient is being taken, can be formally adapted to Wasserstein setting for (3.4)—which we recall, is not a gradient flow for the standard Wasserstein gradient of E_β—. But note that (3.4) writes as

$$\partial_t \mu(t) + \operatorname{div}\left(\frac{\nabla \delta \mathsf{E}_\beta[\mu(t)]}{\delta \mathsf{E}_\beta[\mu(t)]}\mu(t)\right) = 0, \tag{3.15}$$

with $\delta \mathsf{E}_\beta[\mu](x) = \int e^{\beta\langle x,y\rangle}\,\mathrm{d}\mu(y)$. To avoid technicalities we henceforth only focus on the absolutely continuous case. For a fixed $\mu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{S}^{d-1})$, as in the well-known formal Riemannian reinterpretation of the Wasserstein space using Otto calculus [Ott01], [CNWR24, Chapter 5], we consider

$$\mathrm{T}_\mu \mathcal{P}_{\mathrm{ac}}(\mathbb{S}^{d-1}) = \overline{\{\nabla\psi\colon \psi \in C^\infty(\mathbb{S}^{d-1})\}}^{L^2(\mu)},$$

which, rather than endowing with the standard formal metric tensor given by the $\dot{H}^1$–inner product

$$\langle \nabla\psi_1, \nabla\psi_2\rangle_\mu := \int \langle \nabla\psi_1(x), \nabla\psi_2(x)\rangle\,\mathrm{d}\mu(x),$$

we instead endow with

$$\langle \nabla\psi_1, \nabla\psi_2\rangle_{\mu,\mathsf{E}_\beta} := \int \langle \nabla\psi_1(x), \nabla\psi_2(x)\rangle\,\delta\mathsf{E}_\beta[\mu](x)\,\mathrm{d}\mu(x).$$

Continuing the Riemannian geometry analogy, through this metric tensor we can define a distance between $\mu_0, \mu_1 \in \mathcal{P}_{\mathrm{ac}}(\mathbb{S}^{d-1})$ by solving the variational problem

$$\inf_{(\mu(t),v(t))_{t\in[0,1]}} \left\{ \int_0^1 \|v(t)\|^2_{\mu(t),\mathsf{E}_\beta}\,\mathrm{d}t\colon (\mu, v) \text{ satisfy (3.16)}, \mu(0) = \mu_0, \mu(1) = \mu_1 \right\},$$

where

$$\partial_t \mu(t,x) + \operatorname{div}(v(t,x)\mu(t,x)) = 0 \qquad \text{on } [0,1]\times\mathbb{S}^{d-1}. \tag{3.16}$$

This variational problem is a generalization of the celebrated Benamou-Brenier formula [BB00], the value of which we dub $W^2_{2,\mathsf{E}_\beta}(\mu_0,\mu_1)$: it is a weighed Wasserstein distance. For a curve of measures $(\mu(t))_{t\geqslant 0}$ with tangent vectors $(v(t))_{t\geqslant 0}$ (meaning they solve (3.16)), the Wasserstein gradient $\mathbb{W}_{\mathsf{E}_\beta}\mathsf{E}_\beta$ induced by this geometric setup is then the element of $\mathrm{T}_{\mu(t)}\mathcal{P}_{\mathrm{ac}}(\mathbb{S}^{d-1})$ such that

$$\partial_t \mathsf{E}_\beta[\mu(t)] = \langle \mathbb{W}_{\mathsf{E}_\beta}\mathsf{E}_\beta[\mu(t)], v(t)\rangle_{\mu(t),\mathsf{E}_\beta}.$$

We can now demonstrate that the vector field driving (3.15) is the Wasserstein gradient of E_β corresponding to this geometric setup. Indeed as in [CNWR24, Definition 5.9] we first have

$$\partial_t \mathsf{E}_\beta[\mu(t)] = \int \delta\mathsf{E}_\beta[\mu(t)]\,\mathrm{d}\partial_t\mu(t).$$

We then find

$$\int \delta\mathsf{E}_\beta[\mu(t)]\,\mathrm{d}\partial_t\mu(t) = \int \langle \nabla\delta\mathsf{E}_\beta[\mu(t)], v(t)\rangle\,\mathrm{d}\mu(t) = \left\langle \frac{\nabla\delta\mathsf{E}_\beta[\mu(t)]}{\delta\mathsf{E}_\beta[\mu(t)]}, v(t)\right\rangle_{\mu(t),\mathsf{E}_\beta},$$

as desired. The literature studying weighed Wasserstein distances such as the one above is rather scarce, but a relevant reference is [Li21].

Part 2. Clustering

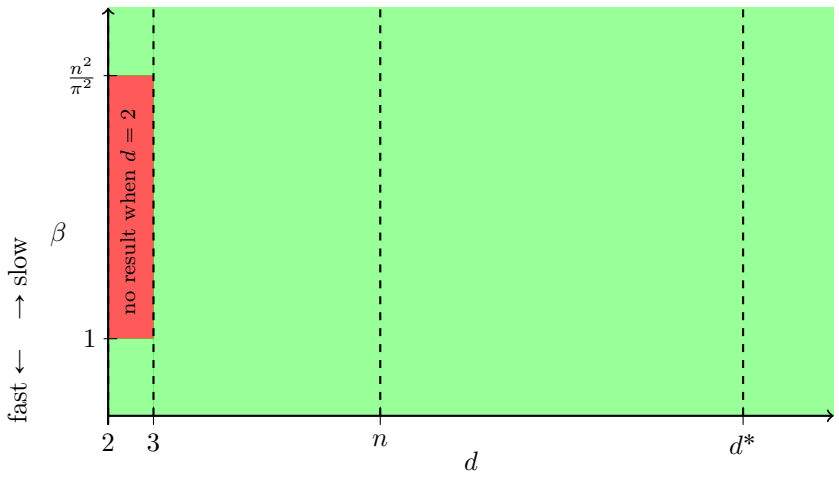


Figure 2. Green zones indicate regimes where convergence to a single cluster as $t \to +\infty$ can be proven. Here $n \geqslant 2$ is fixed. When d is larger than specific thresholds, the long-time asymptotics can be chiseled out in finer detail. Convergence is slow when $\beta \gg 1$ (relative to the size of d, n), as even the exponential decay constant when $d \geqslant n$ is of the form $\lambda = O(e^{-\beta})$ and thus degenerates. One rather expects dynamic metastability. Section 4 addresses the case where β is small and Section 5 where it's large, whereas Section 6 covers the high-dimensional case at arbitrary β.

As alluded to in the introductory discussion, clustering is of particular relevance in tasks such as sentiment analysis, masked language modeling, summarization, and so on. Therein, the output measure encodes the probability distribution of the missing tokens for instance, and its clustering indicates a small number of possible outcomes. In Sections 4, 5, and 6, we show several results (mostly summarized in Figure 2) which indicate that the limiting distribution is a point mass. While it may appear that this leaves no room for diversity or randomness, which is at odds with practical observations, these results hold for the specific choice of parameter matrices, and apply in possibly very long-time horizons. Numerical experiments indicate a more subtle picture for different parameters—for instance, there is an appearance of a long metastable phase during which the particles coalesce in a small number of clusters, which seems consistent with behavior in trained models (Figure 1 in Section 6.3)—. We are not able to theoretically explain this behavior as of now.

Ultimately, the appearance of clusters is somewhat natural since the Transformer dynamics are a weighted average of all particles, with the weights being hard-wired to perform a fast selection of particles most similar to the i-th particle being queried. This causes the emergence of leaders which attract all particles in their vicinity. In the natural language processing interpretation, where particles represent tokens, this further elucidates the wording *attention* as the mechanism of inter-token

attraction, and the amplitude of the inner product between tokens can be seen as a measure of their *semantic similarity*.

4. A single cluster for small β

As seen in Figure 2, the dimension d and inverse temperature β appear to play a key role in the clustering results we obtain. In this section and in Section 5, we begin by focusing on extreme choices of β whilst d, n are fixed. We first focus on the case $\beta = 0$ (Section 4.1), before moving to the case $\beta \ll 1$ by a perturbation argument (Section 4.2). We cover the case where β is sufficiently large, but finite, in Section 5. The case $\beta = +\infty$ is of little interest since all particles are fixed by the evolution.

4.1. The case $\beta = 0$.

For $\beta = 0$, both (SA) and (USA) read as

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\frac{1}{n} \sum_{j=1}^{n} x_j(t) \right), \qquad t \geqslant 0. \tag{4.1}$$

The following result shows that generically over the initial points, a single cluster emerges. It complements a known convergence result ([FL19, Theorem 2]) for (4.1). In [FL19, Theorem 2], the authors show convergence to an antipodal configuration, in the sense that $n-1$ particles converge to some $x^* \in \mathbb{S}^{d-1}$, with the last particle converging to $-x^*$. Moreover, once convergence is shown to hold, it holds with an exponential rate. Mimicking the proof strategy of [BCM15, Theorem 2.2] and [HKR18, Theorem 3.2], we sharpen this result by showing that the appearance of an antipodal particle is non-generic over the choice of initial conditions.

Theorem 4.1. *Let $d, n \geqslant 2$. For Lebesgue almost any initial sequence $(x_i(0))_{i\in[n]} \in (\mathbb{S}^{d-1})^n$, there exists some point $x^* \in \mathbb{S}^{d-1}$ such that the unique solution $(x_i(\cdot))_{i\in[n]} \in C^0(\mathbb{R}_{\geqslant 0}; (\mathbb{S}^{d-1})^n)$ to the corresponding Cauchy problem for* (4.1) *satisfies*

$$\lim_{t\to+\infty} x_i(t) = x^*$$

for any $i \in [n]$.

This is also referred to as convergence toward *consensus* in collective behavior models. We refer the interested reader to Appendix A for the proof, which relies on a gradient flow reinterpretation of (4.1), much like the proofs of several subsequent theorems. We provide some comments in Section 5.

4.2. The case $\beta \ll 1$.

Theorem 4.1 has some implications for small but positive β, something which is already seen in Figure 3 and Figure 4. This is essentially due to the fact that, formally,

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\frac{1}{n} \sum_{j=1}^{n} x_j(t) \right) + O(\beta)$$

for $\beta \ll 1$. So, during a time $\ll \beta^{-1}$, the particles do not feel the influence of the remainder $O(\beta)$ and behave as in the regime $\beta = 0$. This motivates

Theorem 4.2. *Fix $d, n \geqslant 2$. For $\beta \geqslant 0$, let $\mathcal{S}_\beta \subset (\mathbb{S}^{d-1})^n$ be the subset consisting of all initial sequences for which the associated solution to the Cauchy problem for* (SA) *(or* (USA)*) converges to one cluster as $t \to +\infty$. Then*

$$\lim_{\beta\to 0} \mathbb{P}(\mathcal{S}_\beta) = 1.$$

More generally, if Q and K are arbitrary $d \times d$ matrices, then the same result also holds for the Cauchy problem for (2.3) *with $V = I_d$ (or the natural analogue of* (USA) *with these parameters).*

Proof. We focus on the dynamics (SA), but the proof is in fact identical in the case of (USA).

For $\alpha \in [0,1)$, we say that a set formed from n points $z_1, \dots, z_n \in (\mathbb{S}^{d-1})^n$ is α–*clustered* if for any $i, j \in [n]$, $\langle z_i, z_j \rangle > \alpha$ holds. Observe that if $\{z_1, \dots, z_n\}$ is α–clustered for some $\alpha \geqslant 0$, then the solution to the Cauchy problem for (SA) (for arbitrary $\beta \geqslant 0$) with this sequence as initial condition converges to a single cluster, since $w = z_1$ satisfies the assumption in Lemma 6.4.

Now, for any integer $m \geqslant 1$, we denote by $\mathcal{S}_0^m \subset \mathcal{S}_0$ the set of initial sequences $x_1(0), \dots, x_n(0)$ in $(\mathbb{S}^{d-1})^n$ for which the solution $(x_i^0(\cdot))_{i\in[n]}$ to the associated Cauchy problem for (4.1) is $\frac{3}{4}$–clustered at time $t = m$, namely

$$\langle x_i^0(m), x_j^0(m) \rangle > \frac{3}{4} \tag{4.2}$$

holds for all $i, j \in [n]$. We see that $\mathcal{S}_0^m$ is an open set for any integer $m \geqslant 1$. Moreover, $\mathcal{S}_0^m \subset \mathcal{S}_0^{m+1}$ according to the proof of Lemma 6.4, and $\bigcup_{m=1}^{+\infty} \mathcal{S}_0^m = \mathcal{S}_0$. This implies that

$$\lim_{m \to +\infty} \mathbb{P}(\mathcal{S}_0^m) = 1. \tag{4.3}$$

We now show that the solution to (SA) is near that of (4.1), starting from the same initial condition, when β is small. Using the Duhamel formula, we find

$$\begin{aligned}
x_i^\beta(t) - x_i^0(t) &= \int_0^t \sum_{j=1}^n \left(\frac{e^{\beta \langle Q x_i^\beta(s), K x_j^\beta(s) \rangle}}{\sum_{k=1}^n e^{\beta \langle Q x_i^\beta(s), K x_k^\beta(s) \rangle}} \right) \mathbf{P}^\perp_{x_i^\beta(s)}(x_j^\beta(s))\, \mathrm{d}s \\
&\quad - \int_0^t \frac{1}{n} \sum_{j=1}^n \mathbf{P}^\perp_{x_i^0(s)}(x_j^0(s))\, \mathrm{d}s \\
&= \int_0^t \sum_{j=1}^n \left(\frac{1}{n} + O\left(\frac{\beta}{n}\right) \right) \mathbf{P}^\perp_{x_i^\beta(s)}(x_j^\beta(s))\, \mathrm{d}s \\
&\quad - \int_0^t \frac{1}{n} \sum_{j=1}^n \mathbf{P}^\perp_{x_i^0(s)}(x_j^0(s))\, \mathrm{d}s,
\end{aligned}$$

where we used that all particles lie on $\mathbb{S}^{d-1}$ for all times. Employing Grönwall, we deduce

$$\left\| x_i^\beta(t) - x_i^0(t) \right\| \leqslant O(\beta) e^{3t} \tag{4.4}$$

for all $t \geqslant 0$, $\beta \geqslant 0$ and $i \in [n]$. Due to (4.4), there exists some $\beta_m > 0$ such that for any $\beta \in [0, \beta_m]$,

$$\left\| x_i^\beta(m) - x_i^0(m) \right\| \leqslant \frac{1}{8}. \tag{4.5}$$

For this to hold, we clearly need $\beta_m \to 0$ as $m \to +\infty$. Combining (4.2) and (4.5), we gather that for any initial condition in $\mathcal{S}_0^m$, the solution $(x_i^\beta(\cdot))_{i\in[n]}$ to the corresponding Cauchy problem for (SA) is $\frac{1}{2}$–clustered at time $t = m$, namely satisfies

$$\langle x_i^\beta(m), x_j^\beta(m) \rangle > \frac{1}{2}$$

for all $i, j \in [n]$ and $\beta \in [0, \beta_m]$. Thus $\mathcal{S}_0^m \subset \mathcal{S}_\beta$ for any $\beta \in [0, \beta_m]$ by virtue of Lemma 6.4, which together with (4.3) concludes the proof. □

In the specific case where $Q^\top K = I_d$, we can in fact significantly sharpen Theorem 4.2 by relying on the gradient flow structure evoked in Section 3.4. Namely, we can show the following.

Theorem 4.3. *Fix $d, n \geqslant 2$. There exists a numerical constant $C > 0$ such that whenever*

$$\beta \leqslant Cn^{-1}$$

the following holds. For Lebesgue almost any $(x_i(0))_{i\in[n]} \in (\mathbb{S}^{d-1})^n$, there exists $x^* \in \mathbb{S}^{d-1}$ such that the solution $(x_i(\cdot))_{i\in[n]} \in C^0(\mathbb{R}_{\geqslant 0}; (\mathbb{S}^{d-1})^n)$ to the corresponding Cauchy problem for (SA) *(resp. for* (USA)*) satisfies*

$$\lim_{t\to+\infty} x_i(t) = x^*$$

for all $i \in [n]$.

Moreover, when $d = 2$ one can take $\beta \leqslant 1$ and the same conclusion holds.

The proof can be found in Appendix C. As for Theorem 4.1, we briefly discuss the general strategy in Section 5 just below. The improvement to $\beta \leqslant 1$ when $d = 2$ is due to the recent paper [CRMB24].

5. A single cluster for large β

The chain of thought leading to the proof of Theorem 4.3 also leads to

Theorem 5.1. *Fix $d, n \geqslant 2$. There exists a constant $C = C(d) > 0$ depending only on d such that whenever*

$$\beta \geqslant Cn^2,$$

the conclusion of Theorem 4.3 holds for both (SA) *and* (USA).

We refer the interested reader to Appendix B for the proof, which, much like those of Theorem 4.1 and 4.3 relies on the gradient flow interpretation of the dynamics. We give a general outline thereof. Since all the intervening functions and metrics are real-analytic, the celebrated *Łojasiewicz theorem* [Loj63] implies that for any initial condition, the gradient flow converges to some critical point of E_β as $t \to +\infty$. The genericity over the initial configurations seen in the statements comes from the *center-stable manifold theorem* (Lemma A.1). The former ensures that for almost every initial configuration, the gradient flow does not converge to a strict saddle point of E_β (namely critical points where the Hessian has at least one positive eigenvalue). Whence, generically over the initial configurations, the gradient flow converges to a local maximum. One is then left analyzing the landscape of the underlying energy, with the goal of ensuring that all local maxima are necessarily global.

6. The high-dimensional case

We now elucidate the role that the dimension d plays in clustering results. To start, it turns out that the restrictions on β provided by Theorems 4.3 and 5.1 are specific to the two-dimensional case. The following result is shown in [MTG17, CRMB24].

Theorem 6.1 ([MTG17, CRMB24])**.** *Fix $n \geqslant 2$, $d \geqslant 3$ and $\beta \geqslant 0$. Then the conclusion of Theorem 4.3 holds for both* (SA) *and* (USA)*.*

In Section 6.1 we show that the convergence of Theorem 6.1 is exponentially fast when $d \geqslant n$ (although, with a decay constant that is exponentially small in β) and in Section 6.2 we describe the full dynamics when n is fixed and $d \to +\infty$. We discuss some numerical experiments and posit questions on the intermediate behavior of the dynamics when $\beta \gg 1$ in Section 6.3.

We were made aware of Theorem 6.1 after the first version of this manuscript. The downside of our original proof, which caused us to miss the full range of β, is that we focused on $d = 2$, where the natural perturbations to a critical point involve selecting which particles are moving and which are standing still. However, in $d > 2$, one can move all particles in the same direction (as is done in the proof in [CRMB24]). Using that there is a continuum of directions and only finitely many points, one can find some direction that brings all of them closer.

Remark 6.2 (Invariant measures)**.** *In the theory of dynamical systems, often the first question of interest is to find smooth invariant measures. It is clear that whenever conclusions of Theorem 4.3 hold (in particular, for all β, n when $d \geqslant 3$), neither* (SA) *nor* (USA) *may possess a smooth invariant measure.*

6.1. **Clustering at an exponential rate when $d \geqslant n$.** One can ask whether for almost every initial configuration, the convergence provided by all of the results above holds with some rate. The answer is affirmative—and the rate is in fact exponential—when the initial configuration lies in an open hemisphere.

Theorem 6.3. *Let $n \geqslant 1$ and $\beta > 0$. Suppose $d \geqslant n$. Consider the unique solution $(x_i(\cdot))_{i\in[n]} \in C^0(\mathbb{R}_{\geqslant 0}; (\mathbb{S}^{d-1})^n)$ to the Cauchy problem for* (SA) *or* (USA)*, corresponding to an initial sequence of points $(x_i(0))_{i\in[n]} \in (\mathbb{S}^{d-1})^n$ distributed uniformly at random. Then almost surely there exists $x^* \in \mathbb{S}^{d-1}$ and constants $C, \lambda > 0$ such that*

$$\|x_i(t) - x^*\| \leqslant Ce^{-\lambda t} \tag{6.1}$$

holds for all $i \in [n]$ and $t \geqslant 0$.

In fact, let Q and K be arbitrary $d \times d$ matrices. Then the same result also holds for the solution to the corresponding Cauchy problem for (2.3) *with $V = I_d$ (or the natural analogue of* (USA) *with these parameters).*

When $d \geqslant n$ and the points $(x_i(0))_{i\in[n]} \in (\mathbb{S}^{d-1})^n$ are distributed uniformly at random, with probability one there exists[10] $w \in \mathbb{S}^{d-1}$ such that $\langle w, x_i(0)\rangle > 0$ for any $i \in [n]$. In other words, all of the initial points lie in an open hemisphere almost surely. The proof of Theorem 6.3 thus follows as a direct corollary of the following result, which holds for any $n \geqslant 1$ and $d \geqslant 2$:

Lemma 6.4 (Cone collapse)**.** *Let $\beta > 0$ and let $(x_i(0))_{i\in[n]} \in (\mathbb{S}^{d-1})^n$ be such that there exists $w \in \mathbb{S}^{d-1}$ for which $\langle x_i(0), w\rangle > 0$ for any $i \in [n]$. Consider the unique solution $(x_i(\cdot))_{i\in[n]} \in C^0(\mathbb{R}_{\geqslant 0}; (\mathbb{S}^{d-1})^n)$ to the corresponding Cauchy problem for* (SA) *or* (USA)*. Then there exists $x^* \in \mathbb{S}^{d-1}$ and constants $C, \lambda > 0$ such that*

$$\|x_i(t) - x^*\| \leqslant Ce^{-\lambda t}$$

[10]This weak version of Wendel's theorem (Theorem 6.7) is easy to see directly.

holds for all $i \in [n]$ and $t \geqslant 0$.

In fact, let Q and K be arbitrary $d \times d$ matrices. Then the same result also holds for the solution to the corresponding Cauchy problem for (2.3) *with $V = I_d$ (or the natural analogue of* (USA) *with these parameters).*

Remark 6.5. *Lemma 6.4 implies that $\big\{(\bar{x}_i)_{i\in[n]} \in (\mathbb{S}^{d-1})^n \colon \bar{x}_1 = \ldots = \bar{x}_n\big\}$ is Lyapunov asymptotically stable as a set. In fact, it is exponentially stable.*

Lemma 6.4 is reminiscent of results on interacting particle systems on the sphere (see [CLP15, Theorem 3.7] for instance), and the literature on synchronization for the Kuramoto model on the circle ([ABK$^+$22, Lemma 2.8], [HR20, Theorem 3.1] and Section 7.2). We often make use of the following elementary lemma.

Lemma 6.6. *Let $f : \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ be a differentiable function such that*

$$\int_0^{+\infty} |f(t)| \, \mathrm{d}t + \sup_{t\in\mathbb{R}_{\geqslant 0}} \left|\dot{f}(t)\right| < +\infty.$$

Then $\lim_{t\to+\infty} f(t) = 0$.

The proof of Lemma 6.4 is an adaptation of [CLP15, Theorem 1]. We present it here for completeness.

Proof of Lemma 6.4. We focus on the case (USA), and set

$$a_{ij}(t) := n^{-1} e^{\beta\langle x_i(t), x_j(t)\rangle} > 0.$$

The proof for (SA) is identical, and one only needs to change the coefficients $a_{ij}(t)$ by $Z_{\beta,i}(t)^{-1} e^{\beta\langle x_i(t), x_j(t)\rangle}$ throughout. Also note that since we only make use of the positivity of the coefficients $a_{i,j}(t)$ throughout the proof, all arguments are readily generalizable to the case of arbitrary $d \times d$ matrices Q and K appearing in the inner products.

Step 1. Clustering. For $t \geqslant 0$, consider

$$i(t) \in \operatorname*{arg\,min}_{i\in[n]} \langle x_i(t), w\rangle.$$

Fix $t_0 \geqslant 0$. We have

$$\begin{aligned}
&\left(\frac{\mathrm{d}}{\mathrm{d}t}\langle x_{i(t_0)}(\cdot), w\rangle\right)\Big|_{t=t_0} \\
&= \sum_{j=1}^{n} a_{i(t_0)j}(t_0)\Big(\langle x_j(t_0), w\rangle - \langle x_{i(t_0)}(t_0), x_j(t_0)\rangle\langle x_{i(t_0)}(t_0), w\rangle\Big) \geqslant 0.
\end{aligned}$$

This implies that all points remain within the same open hemisphere at all times and the map

$$t \mapsto r(t) := \min_{i\in[n]} \langle x_i(t), w\rangle$$

is non-decreasing on $\mathbb{R}_{\geqslant 0}$. It is also bounded from above by 1. We may thus define $r_\infty := \lim_{t\to+\infty} r(t)$. Note that $r_\infty \geqslant r(0) > 0$ by assumption. By compactness, there exist a sequence of times $\{t_k\}_{k=1}^{+\infty}$ with $t_k \to +\infty$, and some $(\overline{x}_i)_{i\in[n]} \in (\mathbb{S}^{d-1})^n$ such that $\lim_{k\to+\infty} x_i(t_k) = \overline{x}_i$ for all $i \in [n]$. Using the definition of $r(t)$, we also find that

$$\langle \overline{x}_j, w\rangle \geqslant r_\infty$$

for all $j \in [n]$, and by continuity, there exists $i \in [n]$ such that $\langle \overline{x}_i, w\rangle = r_\infty$. Then

(6.2)
$$\lim_{k\to+\infty} \langle \dot{x}_i(t_k), w\rangle = \sum_{j=1}^n \overline{a}_{ij}(\langle w, \overline{x}_j\rangle - \langle \overline{x}_i, \overline{x}_j\rangle\langle \overline{x}_i, w\rangle) \geqslant r_\infty \sum_{j=1}^n \overline{a}_{ij}(1 - \langle \overline{x}_i, \overline{x}_j\rangle),$$

where we set $\overline{a}_{ij} := e^{\beta\langle \overline{x}_i, \overline{x}_j\rangle} > 0$. Notice that

$$\lim_{k\to+\infty} \int_{t_k}^{+\infty} \langle \dot{x}_i(s), w\rangle \,\mathrm{d}s = r_\infty - \lim_{k\to+\infty} \langle x_i(t_k), w\rangle = 0,$$

and by using the equation (USA) we also find that $|\langle \ddot{x}_i(t), w\rangle| = O(e^{2\beta})$ for any $t \geqslant 0$. Therefore by Lemma 6.6, the left-hand side of (6.2) is equal to 0, and consequently the right-hand side term as well. This implies that $\overline{x}_1 = \ldots = \overline{x}_n := x^*$. Repeating the argument by replacing w with x^*, we see that the extraction of a sequence $\{t_k\}_{k=1}^{+\infty}$ as above is not necessary, and therefore

$$\lim_{t\to+\infty} x_i(t) = x^* \tag{6.3}$$

for all $i \in [n]$.

Step 2. Exponential rate. We now improve upon (6.3). Set

$$\alpha(t) := \min_{i\in[n]} \langle x_i(t), x^*\rangle.$$

From (6.3) we gather that there exists some $t_0 > 0$ such that $\alpha(t) \geqslant \frac{1}{2}$ for all $t \geqslant t_0$. Also, in view of what precedes we know that x^* lies in the convex cone generated by the points $x_1(t), \ldots, x_n(t)$ for any $t > 0$. Thus, there exists some $\eta \in (0, 1]$ such that ηx^* is a convex combination of the points $x_1(t), \ldots, x_n(t)$, which implies that

$$x^* = \sum_{k=1}^n \theta_k(t) x_k(t), \qquad \text{for some} \qquad \sum_{k=1}^n \theta_k(t) \geqslant 1, \quad \theta_k(t) \geqslant 0 \quad \forall k \in [n]. \tag{6.4}$$

For any t, we denote by $i(t)$ an element of $\arg\min(\langle x_i(t), x^*\rangle)$ for which $\langle \dot{x}_i(t), x^*\rangle$ is smallest. It follows from a Taylor expansion of $\langle x_i(t+h), x^*\rangle$ for $h > 0$ and $i \in [n]$ that

$$\dot{\alpha}(t) = \langle \dot{x}_{i(t)}(t), x^*\rangle.$$

Therefore

$$\dot{\alpha}(t) = \langle \dot{x}_{i(t)}(t), x^*\rangle \geqslant \sum_{j=1}^n a_{i(t)j}(t)(1 - \langle x_{i(t)}(t), x_j(t)\rangle)\alpha(t) \tag{6.5}$$

On another hand,

$$\min_{j\in[n]} \langle x_{i(t)}(t), x_j(t)\rangle \leqslant \sum_{k=1}^n \theta_k(t)\langle x_{i(t)}(t), x_k(t)\rangle = \langle x_{i(t)}(t), x^*\rangle = \alpha(t). \tag{6.6}$$

Plugging (6.6) into (6.5) and using $a_{ij}(t) \geqslant n^{-1}e^{-2\beta}$ we get

$$\dot{\alpha}(t) \geqslant \frac{1}{2ne^{2\beta}}(1 - \alpha(t)) \tag{6.7}$$

for $t \geqslant t_0$. Applying the Grönwall inequality we get

$$1 - \alpha(t) \leqslant \frac{1}{2} e^{-\frac{1}{2ne^\beta}(t-t_0)} \tag{6.8}$$

for all $t \geqslant t_0$. The conclusion follows. □

In the case $d < n$, we can still apply Wendel's theorem (recalled below) together with Lemma 6.4 to obtain clustering to a single point with probability at least $p_{n,d}$ for some explicit $p_{n,d} \in (0,1)$.

Theorem 6.7 (Wendel, [Wen62])**.** *Let $d, n \geqslant 1$ be such that $d \leqslant n$. Let $x_1, \dots, x_n$ be n i.i.d. uniformly distributed points on $\mathbb{S}^{d-1}$. The probability that these points all lie in the same hemisphere is:*

$$\mathbb{P}\Big(\exists w \in \mathbb{S}^{d-1} : \langle x_i, w\rangle > 0 \quad \text{for all} \quad i \in [n]\Big) = 2^{-(n-1)} \sum_{k=0}^{d-1} \binom{n-1}{k}.$$

6.2. **More precise quantitative convergence.** When n is fixed and $d \to +\infty$, in addition to showing the formation of a cluster as in Theorem 6.3, it is possible to quantitatively describe the entire evolution of the particles with high probability. To motivate this, on the one hand we note that since the dynamics evolve on $\mathbb{S}^{d-1}$, inner products are representative of the distance between points, and clustering occurs if $\langle x_i(t), x_j(t)\rangle \to 1$ for any $(i,j) \in [n]^2$ as $t \to +\infty$. On the other hand, if $d \gg n$, n points in a generic initial sequence are *almost orthogonal* by concentration of measure [Ver18, Chapter 3], and we are thus able to compare their evolution with that of an initial sequence of truly *orthogonal* ones.

We begin by describing the case of exactly orthogonal initial particles, which is particularly simple as the dynamics are described by a single parameter.

Theorem 6.8. *Let $\beta \geqslant 0$, $d, n \geqslant 2$ be arbitrary. Consider an initial sequence $(x_i(0))_{i\in[n]} \in (\mathbb{S}^{d-1})^n$ of n pairwise orthogonal points: $\langle x_i(0), x_j(0)\rangle = 0$ for $i \neq j$, and let $(x_i(\cdot))_{i\in[n]} \in C^0(\mathbb{R}_{\geqslant 0}; (\mathbb{S}^{d-1})^n)$ denote the unique solution to the corresponding Cauchy problem for* (SA) *(resp. for* (USA)*). Then the angle $\angle(x_i(t), x_j(t))$ is the same for all distinct $i, j \in [n]$:*

$$\angle(x_i(t), x_j(t)) = \theta_\beta(t)$$

for $t \geqslant 0$ and some $\theta_\beta \in C^0(\mathbb{R}_{\geqslant 0}; \mathbb{T})$. Furthermore, for (SA)*, $\gamma_\beta(t) := \cos(\theta_\beta(t))$ satisfies*

$$\begin{cases} \dot\gamma_\beta(t) = \dfrac{2e^{\beta\gamma_\beta(t)}(1-\gamma_\beta(t))((n-1)\gamma_\beta(t)+1)}{e^\beta + (n-1)e^{\beta\gamma_\beta(t)}} & \text{for } t \geqslant 0 \\ \gamma_\beta(0) = 0\,, \end{cases} \tag{6.9}$$

and for (USA)*, we have*

$$\begin{cases} \dot\gamma_\beta(t) = \dfrac{2}{n} e^{\beta\gamma_\beta(t)}(1-\gamma_\beta(t))((n-1)\gamma_\beta(t)+1) & \text{for } t \geqslant 0 \\ \gamma_\beta(0) = 0\,. \end{cases} \tag{6.10}$$

Here and henceforth, $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$ denotes the one-dimensional torus. We provide a brief proof of Theorem 6.8 just below. The following result then shows that when $d \gg n$, $t \mapsto \gamma_\beta(t)$ is a valid approximation for $t \mapsto \langle x_i(t), x_j(t)\rangle$ for any distinct $i, j \in [n]$.

Theorem 6.9. *Fix $\beta \geqslant 0$ and $n \geqslant 2$. Then there exists some $d^*(n,\beta) \geqslant n$ such that for all $d \geqslant d^*(n,\beta)$, the following holds. Consider a sequence $(x_i(0))_{i\in[n]}$ of n i.i.d. uniformly distributed points on $\mathbb{S}^{d-1}$, and let $(x_i(\cdot))_{i\in[n]} \in C^0(\mathbb{R}_{\geqslant 0}; (\mathbb{S}^{d-1})^n)$ denote the unique solution to the corresponding Cauchy problem for* (SA)*. Then*

there exist $C = C(n,\beta) > 0$ and $\lambda = \lambda(n,\beta) > 0$, such that with probability at least $1 - 2n^2 d^{-1/64}$,

$$\left|\langle x_i(t), x_j(t)\rangle - \gamma_\beta(t)\right| \leqslant \min\left\{2 \cdot c(\beta)^{nt}\sqrt{\frac{\log d}{d}}, Ce^{-\lambda t}\right\} \tag{6.11}$$

holds for any $i \neq j$ and $t \geqslant 0$, where $c(\beta) = e^{10\max\{1,\beta\}}$, and γ_β is the unique solution to (6.9).

Since the proof is rather lengthy, we defer it to Appendix D. It relies on combining the stability of the flow with respect to the initial data (entailed by the Lipschitz nature of the vector field) with concentration of measure. An analogous statement also holds for (USA), and more details can be found in Remark D.1, whereas the explicit values of C and λ can be found in (D.15). The upper bound in (6.11) is of interest in regimes where d and/or t are sufficiently large as the error in (6.11) is trivially bounded by 2.

Proof of Theorem 6.8. We split the proof in two parts. We focus on proving the result for the dynamics (SA), since the same arguments readily apply to the dynamics (USA).

Part 1. The angle $\theta_\beta(t)$. We first show there exists $\theta \in C^0(\mathbb{R}_{\geqslant 0};\mathbb{T})$ such that $\theta(t) = \angle(x_i(t), x_j(t))$ for any distinct $(i,j) \in [n]^2$ and $t \geqslant 0$. Since the initial tokens are orthogonal (and thus $d \geqslant n$), we may consider an orthonormal basis $(e_1,\dots,e_d)$ of $\mathbb{R}^d$ such that $x_i(0) = e_i$ for $i \in [n]$. Let $\pi : [d] \to [d]$ be a permutation. By decomposing any $x \in \mathbb{S}^{d-1}$ in this basis, we define $P_\pi : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ as

$$P_\pi\left(\sum_{i=1}^n a_i e_i\right) = \sum_{i=1}^n a_i e_{\pi(i)}.$$

Setting $y_i(t) = P_\pi(x_i(t))$ for $i \in [n]$, we see that $y_i(t)$ solves (SA) with initial condition $y_i(0) = P_\pi(x_i(0))$. But $(x_{\pi(1)}(t),\dots,x_{\pi(n)}(t))$ is a solution of (SA) by permutation equivariance, and it has the same initial condition since $P_\pi(x_i(0)) = x_{\pi(i)}(0)$. Consequently, we deduce that $P_\pi(x_i(t)) = x_{\pi(i)}(t)$ for any $t \geqslant 0$ and any $i \in [d]$. Hence

$$\langle x_i(t), x_j(t)\rangle = \langle P_\pi(x_i(t)), P_\pi(x_j(t))\rangle = \langle x_{\pi(i)}(t), x_{\pi(j)}(t)\rangle$$

which concludes the proof.

Part 2. The curve $\gamma_\beta(t)$. By virtue of the orthogonality assumption we have $\gamma_\beta(0) = \cos(\theta_\beta(0)) = 0$. To prove that $\gamma_\beta(t)$ satisfies (6.9) for the case of (SA), recall that

$$\mathbf{P}^\perp_{x_i(t)}(x_j(t)) = x_j(t) - \langle x_i(t), x_j(t)\rangle x_i(t).$$

Then for $k \neq i$,

$$\begin{aligned}\dot\gamma_\beta(t) &= 2\langle \dot x_i(t), x_k(t)\rangle \\ &= 2\sum_{j=1}^n \left(\frac{e^{\beta\langle x_i(t), x_j(t)\rangle}}{\sum_{\ell=1}^n e^{\beta\langle x_i(t), x_\ell(t)\rangle}}\right)\left(\langle x_j(t), x_k(t)\rangle - \langle x_i(t), x_j(t)\rangle\langle x_i(t), x_k(t)\rangle\right).\end{aligned}$$

Since the denominator in the above expression is equal to $(n-1)e^{\beta\gamma_\beta(t)} + e^\beta$, we end up with

$$\begin{aligned}\dot\gamma_\beta(t) &= \frac{2e^{\beta\gamma_\beta(t)}}{(n-1)e^{\beta\gamma_\beta(t)}+e^\beta}\sum_{j=1}^n\Big(\langle x_j(t),x_k(t)\rangle - \langle x_i(t),x_j(t)\rangle\langle x_i(t),x_k(t)\rangle\Big)\\ &= \frac{2e^{\beta\gamma_\beta(t)}}{(n-1)e^{\beta\gamma_\beta(t)}+e^\beta}(1-\gamma_\beta(t)^2+(n-2)(\gamma_\beta(t)-\gamma_\beta(t)^2)),\end{aligned}$$

as desired. □

6.3. **Metastability and a phase transition.** An interesting byproduct of Theorem 6.8 and Theorem 6.9 is the fact that they provide an accurate approximation of the exact *phase transition curve* delimiting the clustering and non-clustering regimes, in terms of t and β. To be more precise, given an initial sequence $(x_i(0))_{i\in[n]} \in (\mathbb{S}^{d-1})^n$ of random points distributed independently according to the uniform distribution on $\mathbb{S}^{d-1}$, and for any fixed $0<\delta\ll 1$, we define the phase transition curve as the boundary

$$\Gamma_{d,\delta} = \partial\left\{t,\beta\geqslant 0\colon t = \operatorname*{arg\,inf}_{s\geqslant 0}\Big(\mathbb{P}(\langle x_1(s),x_2(s)\rangle\geqslant 1-\delta) = 1-2n^2d^{-\frac{1}{64}}\Big)\right\}$$

where $(x_i(\cdot))_{i\in[n]}$ denotes the solution to the corresponding Cauchy problem for (SA). (Here the choice of the first two particles instead of a random distinct pair is justified due to permutation equivariance.) Theorem 6.9 then gives the intuition that over compact subsets of $(\mathbb{R}_{\geqslant 0})^2$, $\Gamma_{d,\delta}$ should be well-approximated by

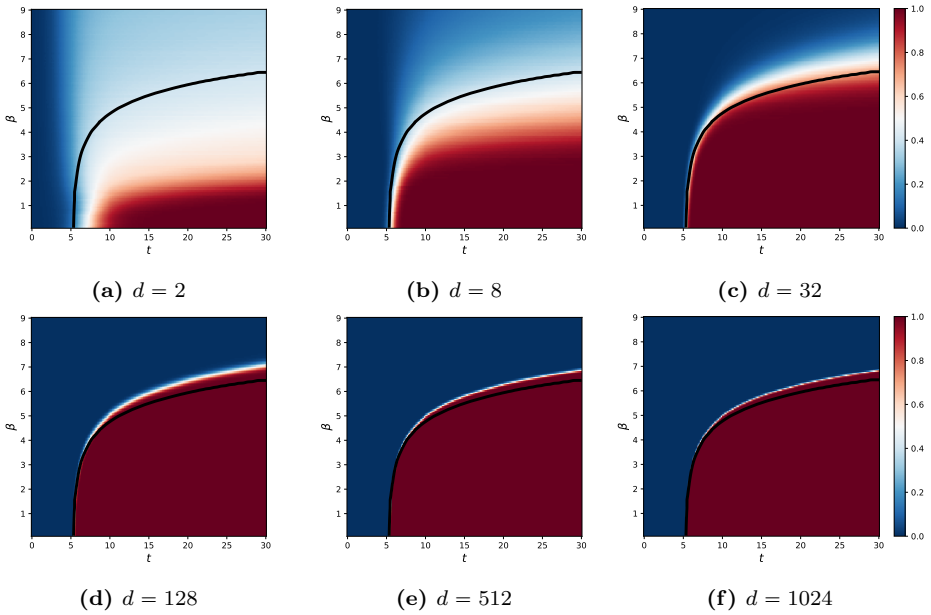
$$\Gamma_{\infty,\delta} = \Big\{t,\beta\geqslant 0\colon \gamma_\beta(t) = 1-\delta\Big\}. \tag{6.12}$$

This is clearly seen in Figure 3, along with the fact that the resolution of this approximation increases with $d\to+\infty$.

Figure 3 appears to contain more information than what we may gather from Theorem 6.3, Theorem 6.8 and Theorem 6.9. In particular, for small d, we see the appearance of a zone (white/light blue in Figure 3) of parameters (t,β) for which the probability of particles being clustered is positive, but not close to one. A careful inspection of this region reveals that points are grouped in a finite number of clusters; see Figure 4. The presence of such a zone indicates the emergence of a long-time *metastable* state where points are clustered into several groups but eventually relax to a single cluster in long-time. This two-time-scale phenomenon is illustrated in Figure 4 and prompts us to formulate the following question.

Problem 1. *Do the dynamics enter a transient metastable state, in the sense that for $\beta\gg 1$, all particles stay in the vicinity of $m<n$ clusters for long periods of time, before they all collapse to the final cluster $\{x^*\}$?*

There have been important steps towards a systematic theory of metastability for gradient flows, with applications to nonlinear parabolic equations—typically reaction-diffusion equations such as the Allen-Cahn or Cahn-Hilliard equations [OR07, KO02]—. While these tools to not readily apply to the current setup, they form an important starting point to answer this question.



(a) $d = 2$ **(b)** $d = 8$ **(c)** $d = 32$

(d) $d = 128$ **(e)** $d = 512$ **(f)** $d = 1024$

Figure 3. Plots of the probability that randomly initialized particles following (SA) cluster to a single point as a function of t and β: we graph the function $(t, \beta) \mapsto \mathbb{P}_{(x_1(0),\dots,x_n(0))\sim\sigma_d}(\{\langle x_1(t), x_2(t)\rangle \geqslant 1-\delta\})$, which is equal to $(t, \beta) \mapsto \mathbb{P}_{(x_1(0),\dots,x_n(0))\sim\sigma_d, i\neq j \text{ fixed}}(\{\langle x_1(t), x_2(t)\rangle \geqslant 1-\delta\})$ by permutation equivariance. We compute this function by generating the average of the histogram of $\{\langle x_i(t), x_j(t)\rangle \geqslant 1-\delta \colon (i,j) \in [n]^2, i \neq j\}$ over 2^{10} different realizations of initial sequences. Here, $\delta = 10^{-3}$, $n = 32$, while d varies. We see that the curve $\Gamma_{\infty,\delta}$ defined in (6.12) approximates the actual phase transition with increasing accuracy as d grows, as implied by Theorem 6.9.

Finally, one may naturally ask whether the clustering and phase diagram conclusions persist when the parameter matrices (Q, K, V) are significantly more general: some illustrations[11] are given in Figure 5.

Problem 2. *Can the conclusions of Theorem 6.8–Theorem 6.9 be generalized to the case of random matrices (Q, K, V)?*

Part 3. Further questions

We conclude this manuscript by discussing several avenues of research that can lead to a finer understanding of the clustering phenomenon and generalizations of our results, and which, we believe, are of independent mathematical interest. Specifically,

- In Section 7, we zero in on the special case $d = 2$, where we make a link with the celebrated *Kuramoto model* when $\beta = 0$;

[11] See `github.com/borjanG/2023-transformers-rotf` for additional figures which indicate that this phenomenon appears to hold in even more generality.

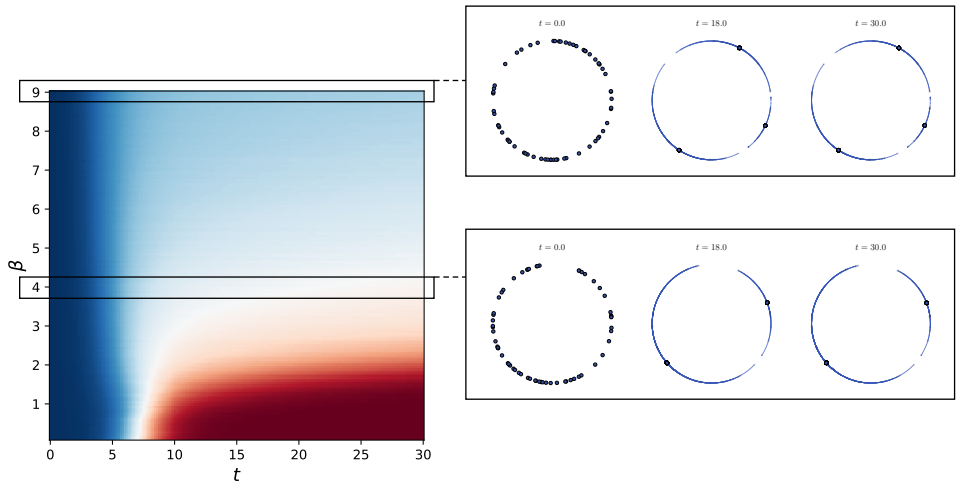


Figure 4. We zoom in on the phase diagram (Figure 3) for the dynamics on the circle: $d = 2$. For $\beta = 4, 9$, we also display a trajectory of (SA) for a randomly drawn initial condition at times $t = 2.5, 18, 30$. We see that the particles settle at 2 clusters when $\beta = 4$ (bottom right) and 3 clusters when $\beta = 9$ (top right), for a duration of time. This reflects our metastability claim for large β.

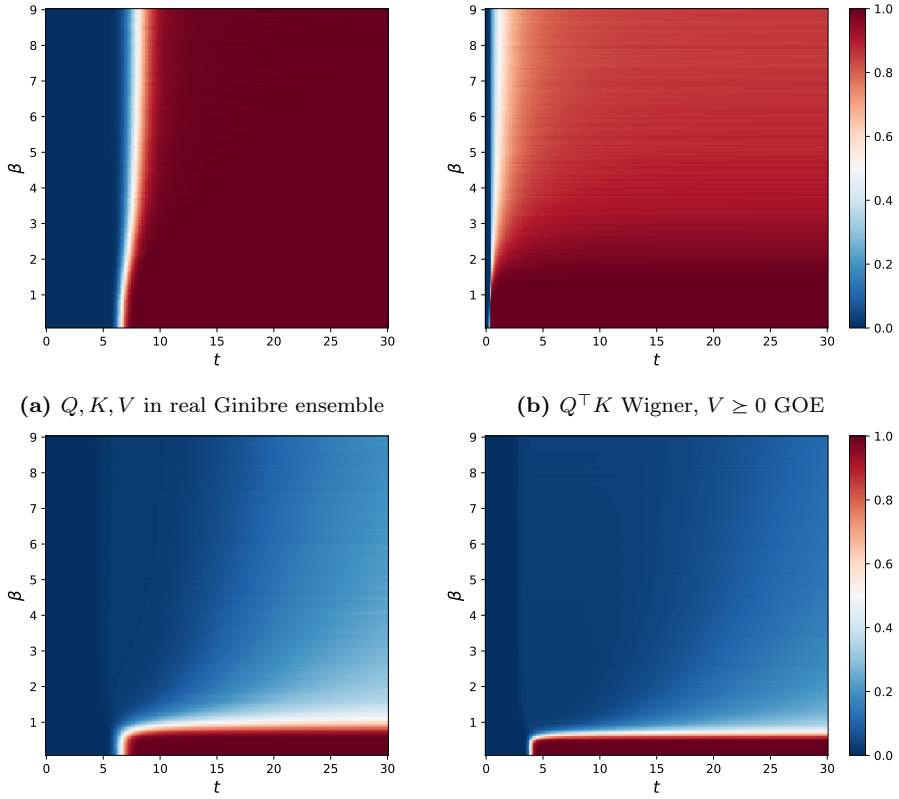
- In Section 8, we discuss an alternative approach for analyzing clustering, based on the so-called *BBGKY hierarchy* from statistical mechanics;
- In Section 9, we foray beyond the case $Q^\top K = V = I_d$ in a few directions. We start with Section 9.1 by relating the case $V = -I_d$ to *optimal configurations* on the sphere. We then discuss existing results in the absence of layer normalization in Section 9.2, which motivate the study of a related singular equation (Section 9.3) and a diffusive regularization (Section 9.4);
- Finally, in Section 10, we briefly discuss various ways focusing on the tuning of the parameter matrices themselves.

7. Dynamics on the circle

We study the dynamics (SA) and (USA) in the special case $d = 2$, namely on the unit circle $\mathbb{S}^1 \subset \mathbb{R}^2$. This model, parametrized by angles and related to the celebrated Kuramoto model, is of independent interest and deserves a complete mathematical analysis.

7.1. **Angular equations.** On the circle $\mathbb{S}^1$, all particles $x_i(t) \in \mathbb{S}^1$ are of course completely characterized by the angle $\theta_i(t) \in \mathbb{T}$: $x_i(t) = \cos(\theta_i(t))e_1 + \sin(\theta_i(t))e_2$ where $e_1 = (1, 0)$ and $e_2 = (0, 1) \in \mathbb{R}^2$. We focus on the dynamics (USA) for simplicity. For any $i \in [n]$ and $t \geqslant 0$, we may derive the equation satisfied by $\theta_i(t)$ from $\cos(\theta_i(t)) = \langle x_i(t), e_1 \rangle$: differentiating in t and plugging into (USA) we obtain

$$\dot\theta_i(t) = -\frac{n^{-1}}{\sin(\theta_i(t))}\left(\sum_{j=1}^n e^{\beta\langle x_i(t), x_j(t)\rangle}\Big[\langle x_j(t), e_1\rangle - \langle x_i(t), x_j(t)\rangle\langle x_i(t), e_1\rangle\Big]\right)$$



(a) Q, K, V in real Ginibre ensemble

(b) $Q^\top K$ Wigner, $V \geq 0$ GOE

(c) Q, K in real Ginibre ensemble, $V = Q^\top K$

(d) $Q^\top K$ Wigner, $V = Q^\top K$

Figure 5. Phase diagrams (see Figure 3 for explanations) for some choices of random matrices (Q, K, V); here $d = 128$, $n = 32$. Sharp phase transitions as well as metastable regions appear in all cases.

where we used the definition of the projection (if $\theta_i(t) = 0$ for some t, we differentiate the equality $\sin(\theta_i(t)) = \langle x_i(t), e_2 \rangle$ instead, which also leads to (7.1) in the end). Observing that

$$\langle x_i(t), x_j(t) \rangle = \cos(\theta_i(t) - \theta_j(t)),$$

we find

$$\dot{\theta}_i(t) = -\frac{n^{-1}}{\sin(\theta_i(t))} \left(\sum_{j=1}^{n} e^{\beta \cos(\theta_i(t) - \theta_j(t))} \Big[\cos(\theta_j(t)) - \cos(\theta_i(t) - \theta_j(t)) \cos(\theta_i(t)) \Big] \right).$$

Using elementary trigonometry, we conclude that

$$\dot{\theta}_i(t) = -\frac{1}{n} \sum_{j=1}^{n} e^{\beta \cos(\theta_i(t) - \theta_j(t))} \sin(\theta_i(t) - \theta_j(t)). \tag{7.1}$$

The case $\beta = 0$ is exactly the Kuramoto model recalled in Section 7.2. Suppose for the time being that $\beta > 0$. Defining the function $h_\beta : \mathbb{T} \to \mathbb{R}_{\geqslant 0}$ as

$$h_\beta(\theta) = e^{\beta \cos(\theta)},$$

we have effectively deduced that the empirical measure of the angles, $\nu(t) = \frac{1}{n}\sum_{j=1}^n \delta_{\theta_j(t)}$, which is a measure on the torus $\mathbb{T}$, is a solution to the continuity equation

$$\partial_t \nu(t) + \partial_\theta(\mathcal{X}[\nu(t)]\nu(t)) = 0, \qquad \text{on } \mathbb{R}_{\geqslant 0} \times \mathbb{T},$$

where

$$\mathcal{X}[\nu](\theta) = \frac{1}{\beta}\Big(h'_\beta * \nu\Big)(\theta).$$

When the particles $x_i(t)$ follow (SA), one readily checks that the same continuity equation is satisfied but rather with the field

$$\mathcal{X}[\nu](\theta) = \frac{1}{\beta}\left(\frac{h'_\beta * \nu}{h_\beta * \nu}\right)(\theta).$$

7.2. **The Kuramoto model.** As mentioned above, when $\beta = 0$, (7.1) is a particular case of the Kuramoto model [Kur75]:

$$\dot\theta_i(t) = \omega_i + \frac{K}{n}\sum_{j=1}^n \sin(\theta_j(t) - \theta_i(t)), \tag{7.2}$$

where $K > 0$ is a prescribed coupling constant, and $\omega_i \in \mathbb{T}$ are the intrinsic natural frequencies of the oscillators $\theta_i(t)$. It is known that for sufficiently small coupling strength K, the oscillators $\theta_i(t)$ in the Kuramoto model (7.2) do not synchronize in long-time. It is also known that when K exceeds some critical threshold value, a phase transition occurs, leading to the synchronization of a fraction of the oscillators. If K is chosen very large, there is total synchronization of the oscillators in long-time. For more on the mathematical aspects of the Kuramoto model, we refer the reader to the review papers [Str00, ABV$^+$05, HKPZ16] (see also [CCH$^+$14, Chi15, FGVG16, DFGV18, HR20, TSS20, ABK$^+$22, MP22] for a non-exhaustive list of other recent mathematical results on the subject).

When all the frequencies ω_i are equal to some given frequency, $\omega \in \mathbb{R}$ say, after a change of variable of the form $\theta_i(t) \leftarrow \theta_i(t) - \omega t$, the dynamics in (7.2) become the gradient flow

$$\dot\theta(t) = n\nabla \mathsf{F}(\theta)$$

where the energy $\mathsf{F} : \mathbb{T}^n \to \mathbb{R}_{\geqslant 0}$ reads

$$\mathsf{F}(\theta) = \frac{K}{2n^2}\sum_{i=1}^n\sum_{j=1}^n \cos(\theta_i - \theta_j). \tag{7.3}$$

The oscillators can be viewed as attempting to maximize this energy. The energy F is maximized when all the oscillators are synchronized, that is, $\theta_i = \theta^*$ for some $\theta^* \in \mathbb{T}$ and for all $i \in [n]$. As the dynamics follow a gradient system, the equilibrium states are the critical points of the energy, namely those satisfying $\nabla\mathsf{F}(\theta) = 0$. The local maxima of F correspond to equilibrium states θ that are physically achievable, since small perturbations thereof return the system back to θ.

Some authors consider a variant of the Kuramoto model where the oscillators are interacting according to the edges of a graph. In other words, the coefficients A_{ij} of the graph's adjacency matrix are inserted in the sum in (7.3) as weights, and the dynamics are then the corresponding gradient flow. A recent line of work culminating with [ABK$^+$22] has established that synchronization occurs with high probability for Erdős–Rényi graphs with parameter p, for every p right above the connectivity threshold.

Coming back to our dynamics (7.1), we notice that it can also be written as a gradient flow on $\mathbb{T}^n$:

$$\dot{\theta}(t) = n\nabla \mathsf{E}_\beta(\theta(t)),$$

for the interaction energy $\mathsf{E}_\beta : \mathbb{T}^n \to \mathbb{R}_{\geqslant 0}$ defined as

$$\mathsf{E}_\beta(\theta) = \frac{1}{2\beta n^2} \sum_{i=1}^n \sum_{j=1}^n e^{\beta \cos(\theta_i - \theta_j)}, \tag{7.4}$$

which is maximized when $\theta_i = \theta^*$ for some $\theta^* \in \mathbb{T}$ and for all $i \in [n]$. In the spirit of [LXB19], we suggest the following open problem—we recall that a critical point is called a *strict saddle point* of E_β if the Hessian of E_β at these points has at least one positive eigenvalue.x

Problem 3. *With the exception of the global maxima, are all critical points of* E_β *strict saddle points?*

The proofs of Theorems 4.3 and 5.1 already yield a positive answer to Problem 5 in the regimes $\beta \leqslant 1$ or $\beta \geqslant \frac{n^2}{\pi^2}$. The complementary regime remains open. By classical arguments, recalled in Appendix A, a positive answer to Problem 3 would imply that for all initial conditions except a set of measure zero, all $\theta_i(t)$ converge under the dynamics (7.1) to a common limit as $t \to +\infty$.

Extensions of the Kuramoto model of the form

$$\dot{\theta}_i(t) = \omega_i + \frac{K}{n} \sum_{j=1}^n h(\theta_j(t) - \theta_i(t)), \tag{7.5}$$

for a general non-linearity $h : \mathbb{T} \to \mathbb{R}$, which contains both (7.2) and our model (7.1) as particular cases, have already been studied in the physics literature. For instance, we refer the reader to [Dai92] (see also [ABV$^+$05, page 158]), where many heuristics are proposed to address the behavior of solutions to these dynamics. We are not aware of mathematical results for (7.1) besides Theorem 5.1. We nevertheless have some hope that handling the dynamics (7.1) is easier than dealing with (7.5) for a general h; for instance, we have

$$h_\beta(\theta) = e^{\beta \cos(\theta)} = \sum_{k \in \mathbb{Z}} I_k(\beta) e^{\mathrm{i}k\theta}$$

where $I_k(\beta)$ are the modified Bessel function of the first kind, whose properties have been extensively studied.

8. BBGKY hierarchy

For the sake of simplicity, we again focus on the dynamics on the circle $\mathbb{S}^1$, where recall that all particles are parametrized by angles (which we also refer to as particles). To carve out an even more complete understanding of the clustering phenomenon, it is natural to consider initial particles sampled i.i.d. from the uniform distribution on $\mathbb{S}^1$ and to study the time-evolution of the r-particle distribution $\rho_n^{(r)}(t, \theta_1, \dots, \theta_r)$, defined as the joint law of the particles $\theta_1(t), \dots, \theta_r(t)$. Otherwise put, it is the r-point marginal of the joint distribution $\rho^{(n)}(t, \cdot) \in \mathcal{P}(\mathbb{T}^n)$ of all n particles. Note that because of rotational invariance, $\rho^{(1)}(t, \cdot)$ is just the

uniform distribution equal to $\frac{1}{2\pi}$ for all $t \geqslant 0$. For $r = 2$, again by rotational invariance, there exists some $\psi(t,\cdot) : \mathbb{T} \to \mathbb{R}_{\geqslant 0}$ such that

$$\rho^{(2)}(t,\theta_1,\theta_2) = \frac{1}{2\pi}\psi(t,\theta_2-\theta_1).$$

Proving the clustering/synchronization of all $\theta_i(t)$ in long-time amounts to proving that $\psi(t,\cdot)$ converges to a Dirac mass centered at 0 as $t \to +\infty$. Using the fact that $\rho^{(n)}(t,\cdot)$ solves the Liouville equation, by following the method used to derive the BBGKY[12] hierarchy [GSRT13, Gol16], it is possible to show that $\psi(t,\cdot)$ satisfies

$$\begin{cases} \partial_t\psi(t,x) + \partial_x(v(t,x)\psi(t,x)) = 0 & \text{in } \mathbb{R}_{\geqslant 0}\times\mathbb{T} \\ \psi(0,x) = (2\pi)^{-1} & \text{in } \mathbb{T}, \end{cases} \tag{8.1}$$

where

$$v(t,x) = \frac{2}{\beta n}h'_\beta(x) - \frac{2(n-2)}{\beta n}g(t,x),$$

and

$$g(t,x) = \mathbb{E}\Big[-h'_\beta(\theta_3(t)) \,\Big|\, \theta_1(t) = 0,\ \theta_2(t) = x\Big].$$

Note that the equation (8.1) is not closed since $g(t,x)$ depends on the 3-point correlation function. This is typical in the BBGKY hierarchy, whereupon physical theory and experimental evidence is typically used to devise an ansatz for closing the system. For instance, the Boltzmann equation is derived from the BBGKY hierarchy by assuming the *molecular chaos hypothesis (Stosszahlansatz)* at the level of $r = 2$. We suggest to close (8.1) in a way that reflects the formation of clusters:

Problem 4. *Devise a realistic ansatz for $g(t,x)$ which allows to close equation* (8.1)*, and allows to prove the convergence of $\psi(t,\cdot)$ to a Dirac mass centered at* 0 *as $t \to +\infty$.*

The derivation of a BBGKY hierarchy when $d > 2$, as well as for (SA), are also problems which we believe merit further investigation.

9. General matrices

Figure 5 hints at the likelihood of the clustering phenomenon being significantly more general than just the case $Q = K = V = I_d$. However, extending our proofs to more general parameter matrices does not appear to be straightforward and is an open problem. Here we discuss some particular cases (without excluding other approaches).

9.1. **The repulsive case.** As seen from Lemma 3.7, in the repulsive case $V = -I_d$ the interaction energy E_β decreases along trajectories. Recall that the unique global minimum of E_β over $\mathcal{P}(\mathbb{S}^{d-1})$ is the uniform distribution (Proposition 3.4). In contrast, we explain in this section that many different configurations of n points may yield global minima for E_β when minimized over empirical measures with n atoms.

We thus focus on minimizing E_β over the set $\mathcal{P}_n(\mathbb{S}^{d-1})$ of empirical measures, namely sums of n Dirac masses. Rewriting E_β as

$$\mathsf{E}_\beta[\mu] = \frac{e^\beta}{2\beta}\iint e^{-\frac{\beta}{2}\|x-x'\|^2}\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x'),$$

[12]Bogoliubov–Born–Green–Kirkwood–Yvon.

it turns out that minimizing E_β over $\mathcal{P}_n(\mathbb{S}^{d-1})$ is precisely the problem of finding *optimal configurations of points* on $\mathbb{S}^{d-1}$, which has direct links to the sphere packing problem [CK07, CKM+22] and coding theory [DGS91]. For $\mu \in \mathcal{P}_n(\mathbb{S}^{d-1})$, we can equivalently rewrite E_β in terms of the set of support points $\mathscr{C} \subset \mathbb{S}^{d-1}$, $\#\mathscr{C} = n$:

$$\mathsf{E}_\beta[\mu] = \mathsf{H}_\beta[\mathscr{C}] = \frac{e^\beta}{2n^2\beta} \sum_{x,x' \in \mathscr{C}} e^{-\frac{\beta}{2}\|x-x'\|^2}.$$

In [CK07], Cohn and Kumar characterize the global minima $\mathscr{C}$ of H_β. To state their result, we need the following definition.

Definition 9.1. *Let $n \geqslant 2$. A set of points $\mathscr{C} = \{x_1, \dots, x_n\} \subset \mathbb{S}^{d-1}$ is called a* spherical t-design *if*

$$\int p(x)\, \mathrm{d}\sigma_d(x) = \frac{1}{n} \sum_{i=1}^n p(x_i)$$

for all polynomials p of d variables, of total degree at most t. The set of points $\mathscr{C}$ is called a sharp configuration *if there are m distinct inner products between pairwise distinct points in $\mathscr{C}$, for some $m > 1$, and if it is a spherical $(2m-1)$-design.*

The following result is a special case of [CK07, Theorem 1.2].

Theorem 9.2 ([CK07])**.** *Let $n \geqslant 2$. Any global minimum of H_β among $\mathscr{C} \subset \mathbb{S}^{d-1}$, $\#\mathscr{C} = n$ is either a sharp configuration, or the vertices of a 600-cell*[13]*.*

The set of sharp configurations is not known for all regimes of n, d or m (the largest m such that the configuration is a spherical m-design). A list of known examples is provided in [CK07, Table 1]: it consists of vertices of full-dimensional polytopes (specifically, regular polytopes whose faces are simplices), or particular derivations of the E_8 root lattice in $\mathbb{R}^8$ and the Leech lattice in $\mathbb{R}^{24}$. We refer the reader to [CK07] and the illustrative experimental paper [BBC+09] for further detail. A complete picture of the long-time behavior of Transformers in the repulsive case remains open.

9.2. **Pure self-attention.** An alternative avenue for conducting such an analysis which has shown to be particularly fruitful consists in removing the projector $\mathbf{P}_x^\perp$, leading to

$$\dot{x}_i(t) = \frac{1}{Z_{\beta,i}(t)} \sum_{j=1}^n e^{\beta \langle Qx_i(t), Kx_j(t) \rangle} V x_j(t) \tag{9.1}$$

for all $i \in [n]$ and $t \in \mathbb{R}_{\geqslant 0}$. In fact, in [GLPR24] we analyze precisely these dynamics, and show different clustering results depending on the spectral properties of the matrix V. We briefly summarize our findings in what follows.

9.2.1. *A review of* [GLPR24]. For most choices of value matrices V, without rescaling time, most particles diverge to $\pm\infty$ and no particular pattern emerges. To make a very rough analogy, (9.1) "looks like" $\dot{x}_i(t) = Vx_i(t)$ (which amounts to having $P_{ij}(t) = \delta_{ij}$ instead of (2.5)), whose solutions are given by $x_i(t) = e^{tV}x_i(0)$. To discern the formation of clusters, we introduce the rescaling[14]

$$z_i(t) = e^{-tV}x_i(t), \tag{9.2}$$

[13] A 600-cell is a particular 4-dimensional convex polytope with $n = 120$ vertices.

[14] The rescaling (9.2) should be seen as a surrogate for layer normalization.

which are solutions to

$$\dot{z}_i(t) = \frac{1}{Z_{\beta,i}(t)} \sum_{j=1}^{n} e^{\beta \langle Q e^{tV} z_i(t), K e^{tV} z_j(t) \rangle} V(z_j(t) - z_i(t)) \tag{9.3}$$

for $i \in [n]$ and $t \geqslant 0$, where

$$Z_{\beta,i}(t) = \sum_{k=1}^{n} e^{\beta \langle Q e^{tV} z_i(t), K e^{tV} z_k(t) \rangle},$$

whereas the initial condition remains the same, namely $x_i(0) = z_i(0)$. It is crucial to notice that the coefficients $A_{ij}(t)$ (see (2.5)) of the self-attention matrix for the rescaled particles $z_i(t)$ are the same as those for the original particles $x_i(t)$. The weight $A_{ij}(t)$ indicates the strength of the attraction of $z_i(t)$ by $z_j(t)$. In [GLPR24] we show that the rescaled particles $z_i(t)$ cluster toward well-characterized geometric objects as $t \to +\infty$ for various choices of matrices (Q, K, V). Our results are summarized in Table 1 below, whose first two lines are discussed thereafter.

V	K and Q	Limit geometry	Result in [GLPR24]
$V = I_d$	$Q^\top K > 0$	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 4.1
V paranormal	$Q^\top K > 0$	polytope $\times$ subspaces	Theorem 5.1
$V = -I_d$	$Q^\top K = I_d$	single cluster at origin*	Theorem C.5

Table 1. Summary of the clustering results of [GLPR24]. *All results except for the case $V = -I_d$ hold for the time-scaled dynamics (9.3).

When $V = I_d$, outside from exceptional situations, all particles cluster to vertices of some convex polytope. Indeed, since the velocity $\dot{z}_i(t)$ is a convex combination of the attractions $z_j(t) - z_i(t)$, the convex hull $\mathcal{K}(t)$ of the $z_i(t)$ shrinks and thus converges to some convex polytope. The vertices of the latter attract all particles as $t \to +\infty$. When the eigenvalue with largest real part of V, denoted by $\lambda_1(V)$, is simple and positive, the rescaled particles $z_i(t)$ cluster on hyperplanes which are parallel to the direct sum of the eigenspaces of the remaining eigenvalues. Roughly speaking, the coordinates of the points $z_i(t)$ along the eigenvector of V corresponding to $\lambda_1(V)$ quickly dominate the matrix coefficients $P_{ij}(t)$ in (9.3) due to the factors $e^{tV} z_j(t)$. For more results and insights regarding clustering on $\mathbb{R}^d$, we refer the reader to [GLPR24]. We nonetheless leave the reader with the following general question:

Problem 5. *Is it possible to extend the clustering results of Table 1 to other cases of (Q, K, V)? What are the resulting limit shapes?*

9.3. **Singular dynamics.** We mention another intriguing question, whose answer would allow for a transparent geometric understanding of clustering for (9.3). Let (Q, K, V) be given $d \times d$ matrices. For $\beta > 0$, we consider the system of coupled ODEs

$$\dot{z}_i(t) = \frac{1}{Z_{\beta,i}(t)} \sum_{j=1}^{n} e^{\beta \langle Q z_i(t), K z_j(t) \rangle} V(z_j(t) - z_i(t)), \tag{9.4}$$

where once again

$$Z_{\beta,i}(t) = \sum_{k=1}^{n} e^{\beta\langle Qz_i(t), Kz_k(t)\rangle}.$$

For any $T > 0$, and any fixed initial condition $(z_i(0))_{i\in[n]} \in (\mathbb{R}^d)^n$, as $\beta \to +\infty$, we expect that the solution to (9.4) converges uniformly on $[0,T]$ to a solution of

$$\dot{z}_i(t) = \frac{1}{|C_i(t)|} \sum_{j\in C_i(t)} V(z_j(t) - z_i(t)) \tag{9.5}$$

where

$$C_i(t) = \Big\{ j \in [n] \colon \langle Qz_i(t), Kz_j(t)\rangle \geqslant \langle Qz_i(t), Kz_k(t)\rangle \quad \text{for all } k \in [n] \Big\}. \tag{9.6}$$

However, defining a notion of solution to (9.5)–(9.6) is not straightforward, as illustrated by the following example.

Example 9.3. *Suppose $d = 2$, $n = 3$. Let $Q = K = V = I_d$ and $z_1(0) = (1,1)$, $z_2(0) = (-1,1)$, $z_3(0) = (0,0)$. Consider the evolution of these particles through* (9.5)–(9.6)*. The points $z_1(t)$ and $z_2(t)$ do not move, because it is easily seen that $C_i(t) = \{i\}$ for $i \in \{1,2\}$. On the other hand, the point $z_3(t)$ can be chosen to solve either of three equations: $\dot{z}_3(t) = z_1(t) - z_3(t)$, or $\dot{z}_3(t) = z_2(t) - z_3(t)$, or even $\dot{z}_3(t) = \frac{1}{2}(z_1(t) + z_2(t)) - z_3(t)$. In any of these cases, both* (9.5) *and* (9.6) *remain satisfied for almost every $t \geqslant 0$.*

It is possible to prove the existence of solutions to (9.5)–(9.6) defined in the sense of Filippov[15]: for this, we can either use a time-discretization of (9.5)–(9.6), or use a convergence argument for solutions to (9.4) as $\beta \to +\infty$. Uniqueness however does not hold, as illustrated by Example 9.3. This naturally leads us to the following question:

Problem 6. *Is it possible to establish a selection principle (similar to viscosity or entropy solutions) for solutions to* (9.5)–(9.6) *which allows to restore uniqueness? In the affirmative, is it possible to revisit the clustering results of* [GLPR24] *and Problem 5 in the setting of* (9.5)–(9.6)*?*

9.4. **Diffusive regularization.** We believe that (9.5)–(9.6) is also an original model for collective behavior. There are some similarities in spirit with methods arising in *consensus based optimization* (CBO for short), [PTTM17, CJLZ21]. With CBO methods, one wishes to minimize a smooth and bounded, but otherwise arbitrary function $f : \mathbb{R}^d \to \mathbb{R}$ by making use of the Laplace method

$$\lim_{\beta\to+\infty} \left(-\frac{1}{\beta} \log \int_{\mathbb{R}^d} e^{-\beta f(x)} \, \mathrm{d}\rho(x) \right) = \inf_{x\in\mathrm{supp}(\rho)} f(x),$$

which holds for any fixed $\rho \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$. This is accomplished by considering a McKean-Vlasov particle system of the form

$$\mathrm{d}x_i(t) = -\lambda(x_i(t) - v_f) H^\epsilon(f(x_i(t)) - f(v[\mu_n(t)])) \, \mathrm{d}t + \sqrt{2}\sigma |x_i(t) - v[\mu_n(t)]| \, \mathrm{d}W_i(t)$$

for fixed $\beta > 0$, with drift parameter $\lambda > 0$ and noise parameter $\sigma \geqslant 0$; H^ϵ is a particular smoothed Heaviside function, and $\mu_n(t)$ is the empirical measure of the

[15]We thank Enrique Zuazua for this suggestion.

particles. The point $v[\mu] \in \mathbb{R}^d$ is a weighted average of the particles:

$$v[\mu] = \frac{1}{Z_{\beta,\mu}} \int_{\mathbb{R}^d} e^{-\beta f(x)} x \, \mathrm{d}\mu(x)$$

where $Z_{\beta,\mu} = \int_{\mathbb{R}^d} e^{-\beta f(x)} \, \mathrm{d}\mu(x)$. Morally speaking, particles which are near a minimum of f have a larger weight. The drift term is a gradient relaxation (for a quadratic potential) towards the current weighted average position of the batch of particles. The diffusion term is an exploration term whose strength is proportional to the distance of the particle from the current weighted average. Results of convergence to a global minimizer do exist, under various smallness assumptions on the initial distribution of the particles, and assumptions on the relative size of the coefficients. They rely on the analysis of the associated Fokker-Planck equation, see [CJLZ21, CD22], and also [FHPS21] for the analog on $\mathbb{S}^{d-1}$. We point out that similarities are mainly in spirit—these results and analysis are inapplicable to our setting because there is no analog for $f(x)$—. Nonetheless, they do raise the following interesting question:

Problem 7. *What can be said about the long-time limit of Transformers with a noise/diffusion term of strength $\sigma > 0$?*

The question is of interest for any of the Transformers models presented in what precedes.

10. Approximation, control, training

Understanding the *expressivity*, namely the ability of a neural network to reproduce any map in a given class (by tuning its parameters), is essential. Two closely related notions reflect the expressivity of neural networks: *interpolation*—the property of exactly matching arbitrarily many input and target samples—and *(universal) approximation*—the property of approximating input-target functional relationships in an appropriate topology—. We refer the reader to [CLLS23] for a primer on the relationship between these two notions in the contex of deep neural networks.

For discrete-time Transformers, universal approximation has been shown to hold in [YBR+19], making use of a variant of the architecture with translation parameters and letting the number of layers go to infinity; see also [ADTK23, JL23] and the review [JLLW23].

In the context of flow maps (from $\mathbb{R}^d$ to $\mathbb{R}^d$), it is now well understood that interpolation and approximation reflect the *controllability* properties of the system. The transfer of control theoretical techniques to the understanding of expressivity has borne fruit, both in terms of controllability results [AS22, CLT20, TG22, LLS22, RBZ23, VR23, CLLS23] and optimal control insights [LCT18, EGPZ20, GZ22, EYG23]. We refer the reader to [AL24, AG24, FdHP24] for the first universal approximation results for Transformers, viewed as measure-to-measure maps, using control theoretic tools.

Besides approximation, understanding the training dynamics of Transformers is another major challenge which we haven't covered herein. As it is impossible to reference all works on this burgeoning topic, we refer the interested reader to [TLTO23, ACDS23, DGTT24] and references therein.

Acknowledgments

We thank Pierre Ablin, Sébastien Bubeck, Gabriel Peyré, Matthew Rosenzweig, Sylvia Serfaty, Kimi Sun, and Rui Sun for discussions. We thank Nicolas Boumal for referring us to [MTG17, CRMB24] and for clarifying comments.

B.G. acknowledges financial support from the French government managed by the National Agency for Research under the France 2030 program, with the reference "ANR-23-PEIA-0004". The work of Y.P. was supported in part by the MIT-IBM Watson AI Lab and by the National Science Foundation under Grant No CCF-2131115. P.R. was supported by NSF grants DMS-2022448, CCF-2106377, and a gift from Apple.

Appendix

Appendix A. Proof of Theorem 4.1

The proof of Theorem 4.1 relies on standard arguments from dynamical systems, upon noticing that the evolution (4.1) is a (continuous-time) gradient ascent for the energy $\mathsf{E}_0 : (\mathbb{S}^{d-1})^n \to \mathbb{R}$ defined as

$$\mathsf{E}_0(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \langle x_i, x_j \rangle.$$

Since the dynamics are the gradient ascent of a real-analytic functional on the compact real-analytic manifold $(\mathbb{S}^{d-1})^n$, the celebrated Łojasiewicz theorem [Loj63], in the form given by [HKR18, Corollary 5.1]—which is valid in the context of general compact Riemannian manifolds—, implies that for any initial condition $X \in (\mathbb{S}^{d-1})^n$, the solution $\Phi^t(X) \in (\mathbb{S}^{d-1})^n$ converges to some critical point $X^* \in (\mathbb{S}^{d-1})^n$ of E_0 as $t \to +\infty$.

We recall that a *strict saddle point* of E_0 is a critical point of E_0 at which the Hessian of E_0 has at least one strictly positive eigenvalue. Theorem 4.1 then follows by combining the following couple of lemmas with the Łojasiewicz theorem.

Lemma A.1. *Let $\mathcal{M}$ be a compact Riemannian manifold and let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function. The set of initial conditions $X_0 \in \mathcal{M}$ for which the gradient ascent*

$$\begin{cases} \dot{X}(t) = \nabla f(X(t)) \\ X(0) = X_0 \end{cases} \tag{A.1}$$

converges to a strict saddle point of f is of volume zero.

Proof of Lemma A.1. Let us denote by $\Phi^t(X_0) := X(t)$, $t \geqslant 0$ the solution to (A.1). We denote by $\mathcal{S} \subset \mathcal{M}$ the set of strict saddle points of f, and by $\mathcal{A} \subset \mathcal{M}$ the set of initial conditions $X_0 \in \mathcal{M}$ for which $\Phi^t(X_0)$ converges to a strict saddle point of f as $t \to +\infty$. For any $y \in \mathcal{S}$, we denote by B_y a ball in which the local center-stable manifold $W^{\mathrm{sc}}_{\mathrm{loc}}(y)$ exists (see [Shu13], Theorem III.7 and Exercise III.3 for the adaptation to flows). Using compactness, we may write the union of these balls as a countable union $\bigcup_{k \in I} B_{y_k}$ (where I is countable and $y_k \in \mathcal{M}$ for $k \in I$). If $X_0 \in \mathcal{A}$, there exists some $t_0 \geqslant 0$ and $k \in I$ such that $\Phi^t(X_0) \in B_{y_k}$ for all $t \geqslant t_0$. From the center-stable manifold theorem ([Shu13], Theorem III.7 and Exercise III.3, where we note that the Jacobian of a gradient vector field coincides, at a critical point, with the Hessian of the corresponding function) we gather that $\Phi^t(X_0) \in W^{\mathrm{sc}}_{\mathrm{loc}}(y_k)$

for $t \geqslant t_0$, hence $X_0 \in \Phi^{-t}(W^{\mathrm{sc}}_{\mathrm{loc}}(y_k))$ for all $t \geqslant t_0$. The dimension of $W^{\mathrm{sc}}_{\mathrm{loc}}(y_k)$ is at most $\dim(\mathcal{M}) - 1$, thus it has zero volume. Since Φ^t is a diffeomorphism on a compact manifold, Φ^{-t} preserves null-sets and hence $\Phi^{-t}(W^{\mathrm{sc}}_{\mathrm{loc}}(y_k))$ has zero volume for all $t \geqslant 0$. Therefore $\mathscr{A}$, which satisfies

$$\mathscr{A} \subset \bigcup_{k \in I} \bigcup_{\ell \in \mathbb{N}} \Phi^{-\ell}(W^{\mathrm{sc}}_{\mathrm{loc}}(y_k))$$

has volume zero. □

Lemma A.2. *Any critical point $(x_1, \dots, x_n) \in (\mathbb{S}^{d-1})^n$ of E_0 which is not a global maximum, namely such that $x_1 = \dots = x_n$, is a strict saddle point. In particular, all local maxima are global.*

Proof of Lemma A.2. We extend the proof idea of [Tay12, Theorem 4.1] as follows. Let $(x_1, \dots, x_n) \in (\mathbb{S}^{d-1})^n$ be a critical point of E_0, and assume that the points x_i are not all equal to each other.

Step 1. We first prove that there exists a set of indices $\mathcal{S} \subset [n]$ such that

$$\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}^c} \langle x_i, x_j \rangle < 0. \tag{A.2}$$

To this end, define

$$m := \sum_{j=1}^{n} x_j,$$

and consider two cases. If $m \neq 0$, then we deduce from $\nabla \mathsf{E}_0(x_1, \dots, x_n) = 0$ that for any $j \in [n]$, x_j is collinear with m. Thus $x_j = \pm x_1$ for any $j \in [n]$. Setting

$$\mathcal{S} = \{j \in [n] \colon x_j = +x_1\},$$

we can see that (A.2) holds, unless $\mathcal{S} = [n]$ which has been excluded. Now suppose that $m = 0$. Then by expanding $\langle m, x_i \rangle = 0$, we find that for any $i \in [n]$

$$-1 = \sum_{j=2}^{n} \langle x_j, x_1 \rangle,$$

holds, which again implies (A.2) with $\mathcal{S} = \{1\}$.

Step 2. In this second step we look to deduce from (A.2) that $(x_1, \dots, x_n)$ is a strict saddle point. Consider an arbitrary non-zero skew-symmetric matrix B and define the perturbation

$$x_i(t) = \begin{cases} x_i & i \notin \mathcal{S} \\ e^{tB} x_i & i \in \mathcal{S}. \end{cases}$$

Set $\mathsf{E}_0(t) = \mathsf{E}_0(x_1(t), \dots, x_n(t))$. Note that we have

$$\mathsf{E}_0(t) = \text{const.} + \frac{2}{n} \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}^c} \langle x_i(t), x_j \rangle,$$

where we grouped time-independent terms into the constant (recall that e^{tB} is an orthogonal matrix, since skew-symmetric matrices are the Lie algebra of $\mathrm{SO}(d)$).

Thus

$$\mathsf{E}_0'(t) = \frac{2}{n} \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}^c} \langle \dot{x_i}(t), x_j \rangle$$
$$\mathsf{E}_0''(t) = \frac{2}{n} \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}^c} \langle \ddot{x_i}(t), x_j \rangle .$$

Since $(x_1, \dots, x_n)$ is a critical point of E_0, we have $\mathsf{E}_0'(0) = 0$. On the other hand, since $\ddot{x}_i(0) = B^2 x_i$ we have

$$\mathsf{E}_0''(0) = \frac{2}{n} \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}^c} \langle B^2 x_i, x_j \rangle . \tag{A.3}$$

We claim that given (A.2), there must exist some skew-symmetric matrix B such that $\mathsf{E}_0''(0) > 0$. Indeed, if d is even, then we just take B as the block-diagonal matrix with repeated block

$$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},$$

so that $B^2 = -I_d$. If d is odd, we can represent

$$-I_d = \frac{1}{d-1} \sum_{j=1}^{d} B_j^2, \tag{A.4}$$

where B_j is the same block-diagonal matrix, with the exception that the j-th block is a 1×1 zero-matrix. If each B_j were to yield $\mathsf{E}_0''(0) \leqslant 0$, then it would violate (A.2). Thus, $\mathsf{E}_0''(0) > 0$ for some well-chosen skew-symmetric B, which proves that $(x_1, \dots, x_n)$ is a strict saddle point. □

Appendix B. Proof of Theorem 5.1

Proof of Theorem 5.1. We leverage the gradient flow structure presented in Remark 3.8 and Section 3.4 (the manifold is compact, and the metric and functional are analytic), and use Lemma A.1 as in the proof of Theorem 4.1. Consequently, it suffices to show that, in the stated regime of β, the critical points of E_β which are not global maxima are strict saddle points, namely, that all local maxima are global. For simplicity we write the argument for (USA) and explain the extension to the case of (SA) in Remark B.1.

We begin by focusing on the case $d = 2$, and provide a brief argument which shows that the case of arbitrary $d \geqslant 2$ readily follows.

Let $(\theta_1, \dots, \theta_n) \in \mathbb{T}^n$ be a critical point such that all eigenvalues of the Hessian of E_β are non-positive. We intend to show that if β is sufficiently large, then necessarily $\theta_1 = \cdots = \theta_n$. To that end, note that the non-positivity of the Hessian of E_β implies in particular that for any subset of indices $\mathcal{S} \subset [n]$, we must have

$$\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} \partial_{\theta_i} \partial_{\theta_j} \mathsf{E}_\beta(\theta_1, \dots, \theta_n) \leqslant 0 . \tag{B.1}$$

Notice that for any $i, j \in [n]$,

$$\partial_{\theta_i} \mathsf{E}_\beta(\theta_1, \dots, \theta_n) = \frac{1}{n^2} \sum_{m \in [n] \setminus \{i\}} -\sin(\theta_i - \theta_m) e^{\beta \cos(\theta_i - \theta_m)}$$

and

$$\partial_{\theta_i}\partial_{\theta_j}\mathsf{E}_\beta(\theta_1,\dots,\theta_n) = \frac{1}{n^2}\cdot\begin{cases} g(\theta_i-\theta_j), & i\neq j \\ -\sum_{m\in[n]\setminus\{i\}} g(\theta_i-\theta_m), & i=j\,, \end{cases}$$

where we set $g(x) := (\cos(x) - \beta\sin^2(x))e^{\beta\cos(x)}$. Plugging this expression back into (B.1) and simplifying, we obtain

$$\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}^c} g(\theta_i-\theta_j) \geqslant 0\,. \tag{B.2}$$

Let us now define τ^*_β be the unique solution on $[0,\frac{\pi}{2})$ of the equation

$$\beta\sin^2(\tau) = \cos(\tau)\,.$$

Note that τ^*_β is a monotonically decreasing function of β, and in fact

$$\tau^*_\beta = \frac{1+o(1)}{\sqrt{\beta}}$$

as $\beta\to+\infty$. The importance of τ^*_β is in implying the following property of the function g: for any $\tau\notin[-\tau^*_\beta,\tau^*_\beta]$, we must have that $g(\tau)<0$ (see Figure 6). We arrive at the following conclusion: it must be that for any proper subset $\mathcal{S}\subset[n]$ there exists, by virtue of (B.2), some index $j\in\mathcal{S}^c$ such that

$$\inf_{k\in\mathcal{S}}|\theta_j-\theta_k| < \tau^*_\beta.$$

So now let us start with $\mathcal{S}=\{1\}$ and grow $\mathcal{S}$ inductively by adding those points θ_j at distance $<\tau^*_\beta$ from $\{\theta_k\colon k\in\mathcal{S}\}$ at each induction step. If β is large enough so that

$$(n-1)\tau^*_\beta < \frac{\pi}{2},$$

then in the process of adding points we have travelled a total arc-length $<\pi/2$ on each side of x_1. Thus it must be that the collection of points $\theta_1,\dots,\theta_n$ is strictly contained inside a half-circle of angular width $<\pi$. By Lemma 6.4 we know that there can be no critical points of E_β that are strictly inside some half-circle, unless that critical point is trivial: $\theta_1=\cdots=\theta_n$. This completes the proof when $d=2$.

We can show that the same conclusion holds for any dimension $d\geqslant 2$. The proof follows by arguing just as above, making instead use of the following generalization of (B.2): given a collection $x_1,\dots,x_n\in\mathbb{S}^{d-1}$ at which the Hessian of E_β is non-positive, we must have for any subset $\mathcal{S}\subset[n]$ that

$$\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}^c} g(\theta_{ij}) \geqslant 0, \tag{B.3}$$

where $g(\zeta) = e^{\beta\cos(\zeta)}((d-1)\cos(\zeta)-\beta\sin^2(\zeta))$ and $\theta_{ij}\in[0,\pi]$ is the geodesic distance between x_i and x_j, namely $\cos(\theta_{ij}) = \langle x_i, x_j\rangle$. We now show (B.3). By repeating the argument in Step 2 of the proof of Lemma A.2 , we see that for any skew-symmetric matrix B we must have

$$\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}^c} e^{\beta\langle x_i,x_j\rangle}\Big(\beta\langle Bx_i,x_j\rangle^2 + \langle B^2x_i,x_j\rangle\Big) \leqslant 0. \tag{B.4}$$

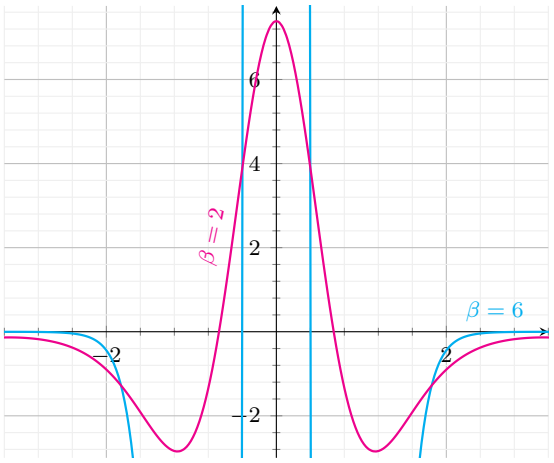


Figure 6. The function $\tau \mapsto g(\tau)$ for two values of β.

Now we take B to be random by generating $B_{ij} \overset{\text{i.i.d.}}{\sim} P$, $i < j$ and P being any zero-mean, unit-variance distribution. We set $B_{ji} = -B_{ij}$ and $B_{ii} = 0$. Then it is easy to check that

$$\mathbb{E}[B^2] = -(d-1)I_d$$

and

$$\mathbb{E}[\langle Bx_i, x_j\rangle^2] = 1 - \langle x_i, x_j\rangle^2 = \sin^2(\theta_{ij}).$$

Thus, taking the expectation over all such B in (B.4) yields (B.3). Mirroring the proof for $d = 2$, we define τ^*_β to be the unique solution on $[0, \frac{\pi}{2})$ of the equation $\beta \sin^2(\tau) = (d-1)\cos(\tau)$. We note that

$$\tau^*_\beta = \sqrt{\frac{(d-1) + o(1)}{\beta}}$$

for $\beta \to +\infty$. Repeating verbatim the argument for the case $d = 2$, we deduce the convergence to a single cluster whenever $\beta \gtrsim (d-1)n^2$. □

Remark B.1. *We comment on the extension of the above proof to the dynamics* (SA). *We recall that* (SA) *is a gradient flow, but for a different metric—see Section 3.4—and we show that the saddle point property is preserved across metrics. Our proof is an adaptation of a classical argument: the Hessian of a function at a critical point is a notion which does not depend on the choice of Riemannian metric.*

Let $x = (x_1, \dots, x_n) \in (\mathbb{S}^{d-1})^n$ be a critical point of E_β (this does not depend on the metric) such that not all x_i are equal to each other. Recall that for $f : (\mathbb{S}^{d-1})^n \to \mathbb{R}$, for any metric on $(\mathbb{S}^{d-1})^n$ (with associated Christoffel symbols Γ^k_{ij}) and any associated orthonormal basis $y_1, \dots, y_{(d-1)n}$, the Hessian of f reads

$$\text{Hess}(f) = \left(\frac{\partial^2 f}{\partial y_i \partial y_j} - \Gamma^k_{ij}\frac{\partial f}{\partial y_k}\right) dy_i \otimes dy_j. \tag{B.5}$$

Since we are evaluating the Hessian at a critical point x of E_β, the term carrying the Christoffel symbols Γ^k_{ij} vanishes. In the above argument, we saw that Hess E_β *evaluated at x, and written in an orthonormal basis for the canonical metric g on*

$(\mathbb{S}^{d-1})^n$, is not negative semi-definite. We denote this matrix by M_g; we know that there exists $v \in \mathrm{T}_x(\mathbb{S}^{d-1})^n$ such that $g(v,v) = 1$ and $v^\top M_g v > 0$. Let $\tilde{g}$ be another metric on $(\mathbb{S}^{d-1})^n$; we denote by $M_{\tilde{g}}$ the Hessian evaluated at x, and written in an orthonormal basis for $\tilde{g}$. Let $c : \mathbb{R}_{\geqslant 0} \to (\mathbb{S}^{d-1})^n$ be such that $c(0) = x$ and $\dot{c}(0) = v$. Since x is a critical point (for both metrics), a Taylor expansion to second order in the two orthonormal bases yields

$$\mathsf{E}_\beta(c(t)) = \mathsf{E}_\beta(c(0)) + \frac{1}{2}t^2 v^\top M_g v + O(t^3)$$

as well as

$$\mathsf{E}_\beta(c(t)) = \mathsf{E}_\beta(c(0)) + \frac{1}{2}t^2 \|v\|_{\tilde{g}}^{-2} v^\top M_{\tilde{g}} v + O(t^3)$$

thanks to (B.5). *Hence $v^\top M_{\tilde{g}} v > 0$. Specializing to $\tilde{g}$ being the metric of Section 3.4, with respect to which* (SA) *is a gradient flow for E_β, we conclude for* (SA).

Appendix C. Proof of Theorem 4.3

Proof of Theorem 4.3. We leverage the gradient flow structure and follow the same strategy as in the proof of Theorem 5.1 presented above. For simplicity we write the argument for (USA) and explain the extension to the case of (SA) in Remark C.1.

Consider

$$\mathsf{E}_\beta(x_1, \dots, x_n) = \frac{1}{2\beta} \sum_{i=1}^n \sum_{j=1}^n \left(e^{\beta \langle x_i, x_j \rangle} - 1 \right).$$

Note that this is only a slight deviation from the energy studied in Section 3.4: we solely subtracted a constant. Consequently the dynamics (USA) are also a gradient flow for this energy. The main interest of considering this modified energy is the observation that

$$\mathsf{E}_\beta(x_1, \dots, x_n) = \mathsf{E}_0(x_1, \dots, x_n) + \beta\, \mathsf{R}_\beta(x_1, \dots, x_n),$$

where R_β is smooth. Hence R_β has a bounded Hessian on $(\mathbb{S}^{d-1})^n$ uniformly with respect to β, and

$$\nabla \mathsf{E}_\beta = \nabla \mathsf{E}_0 + O(\beta), \qquad \operatorname{Hess} \mathsf{E}_\beta = \operatorname{Hess} \mathsf{E}_0 + O(\beta). \tag{C.1}$$

Observe that in the proof of Theorem 4.1, we actually showed that there exists $c > 0$ such that at any critical point $(x_1, \dots, x_n)$ of E_0 for which $x_i \neq x_j$ whenever $i \neq j$, at least one of the eigenvalues of the Hessian of E_0, λ say, satisfies $\lambda \geqslant c$. Indeed, in (A.2) the proof actually shows the existence of some $\mathcal{S} \subset [n]$ such that

$$\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}^c} \langle x_i, x_j \rangle \leqslant -1.$$

Then, (A.3), together with (A.4) for instance, yield

$$\mathsf{E}_0''(0) \geqslant \frac{2(d-1)}{dn} =: c \tag{C.2}$$

for one of the B_j.

Now suppose that there exists a positive sequence $\beta_k \to 0$ as well as $X_k \in (\mathbb{S}^{d-1})^n$ such that X_k is a critical point of E_{β_k} and all of the eigenvalues of $\operatorname{Hess} \mathsf{E}_{\beta_k}(X_k)$ are non-positive. Then by virtue of the continuity properties of E_β with respect to β in (C.1), we find that, up to extracting a subsequence, there is some limit point $\overline{X} = (x_1, \dots, x_n) \in (\mathbb{S}^{d-1})^n$ of X_k which is a critical point of E_0, and such that all of

the eigenvalues of $\operatorname{Hess} \mathsf{E}_0(\overline{X})$ are non-positive. Per Theorem 4.1, this implies that $x_1 = \ldots = x_n$. But then, for large enough k, X_k is also constituted of points which are all nearly equal, whence in the same hemisphere, and the only such critical point of E_β is that in which all points are equal (synchronized). This, combined with the continuity of the eigenvalues of $\operatorname{Hess} \mathsf{E}_\beta$ with respect to β and (C.2), proves that there exists some $c > 0$ independent of n such that whenever $\beta \leqslant c\, n^{-1}$, all critical points of E_β except synchronized ones are strict saddle points. □

Remark C.1. *We comment on the extension of the above proof to the dynamics* (SA)*. The point of contention is* (C.1)*, since the metric with respect to which the gradient and Hessian of* E_0 *are taken is not the same as that for* E_β*. Denote the modified metric defined in Section 3.4 by* g_β*, and the canonical metric by* g*. For any* $x \in (\mathbb{S}^{d-1})^n$ *and* $v \in \mathrm{T}_x(\mathbb{S}^{d-1})^n$ *we have*

$$\mathrm{D}\mathsf{E}_\beta(x)[v] = g_\beta(\nabla_{g_\beta} \mathsf{E}_\beta(x), v),$$

but also $\mathrm{D}\mathsf{E}_\beta(x)[v] = g(\nabla_g \mathsf{E}_\beta(x), v)$*. By virtue of the explicit form of* g_β *and* E_β *as well as* (C.1)*, we gather that*

$$g_\beta(\nabla_{g_\beta} \mathsf{E}_\beta(x), v) = g(\nabla_g \mathsf{E}_0(x), v) + O(\beta) \tag{C.3}$$

which implies that any sequence of critical points of E_β *converges to a critical point for* E_0*. Similarly, since* $\operatorname{Hess}_{g_\beta} \mathsf{E}_\beta(x)[v] = \mathrm{D}(\nabla_{g_\beta} \mathsf{E}_\beta(x))[v]$*, we find*

$$\operatorname{Hess}_{g_\beta} \mathsf{E}_\beta(x)[v] = \operatorname{Hess}_g \mathsf{E}_0(x)[v] + O(\beta). \tag{C.4}$$

We can then repeat the argument in the proof above by replacing (C.1) *by* (C.3) *and* (C.4)*.*

Appendix D. Proof of Theorem 6.9

Proof. We focus on the dynamics (SA), since the proof for (USA) follows from very similar computations.

Step 1. The flow map is Lipschitz. We begin by showing that the trajectories satisfy a Lipschitz property with respect to the initial data. To this end, let $(x_i(\cdot))_{i\in[n]} \in C^0(\mathbb{R}_{\geqslant 0}; (\mathbb{S}^{d-1})^n)$ and $(y_i(\cdot))_{i\in[n]} \in C^0(\mathbb{R}_{\geqslant 0}; (\mathbb{S}^{d-1})^n)$ be two solutions to the Cauchy problem for (SA) associated to data $(x_i(0))_{i\in[n]}$ and $(y_i(0))_{i\in[n]}$ respectively. For any $i \in [n]$ and $t \geqslant 0$, we have

$$\begin{aligned} x_i(t) - y_i(t) &= x_i(0) - y_i(0) \\ &\quad + \int_0^t \sum_{j=1}^n \left(\frac{e^{\beta\langle x_i(s), x_j(s)\rangle}}{\sum_{k=1}^n e^{\beta\langle x_i(s), x_k(s)\rangle}} \right) \big(x_j(s) - \langle x_i(s), x_j(s)\rangle x_i(s)\big)\, \mathrm{d}s \\ &\quad - \int_0^t \sum_{j=1}^n \left(\frac{e^{\beta\langle y_i(s), y_j(s)\rangle}}{\sum_{k=1}^n e^{\beta\langle y_i(s), y_k(s)\rangle}} \right) \big(y_j(s) - \langle y_i(s), y_j(s)\rangle y_i(s)\big)\, \mathrm{d}s. \end{aligned} \tag{D.1}$$

We see that

$$\left\| \int_0^t \sum_{j=1}^n \left(\frac{e^{\beta\langle x_i(s), x_j(s)\rangle}}{\sum_{k=1}^n e^{\beta\langle x_i(s), x_k(s)\rangle}} \right) (x_j(s) - y_j(s))\, \mathrm{d}s \right\| \leqslant \int_0^t \max_{j\in[n]} \|x_j(s) - y_j(s)\|\, \mathrm{d}s. \tag{D.2}$$

On another hand, since the softmax function with a parameter β is β–Lipschitz (with respect to the Euclidean norm), we also get

$$\begin{aligned}
&\left\| \int_0^t \sum_{j=1}^n \left(\frac{e^{\beta\langle x_i(s), x_j(s)\rangle}}{\sum_{k=1}^n e^{\beta\langle x_i(s), x_k(s)\rangle}} - \frac{e^{\beta\langle y_i(s), y_j(s)\rangle}}{\sum_{k=1}^n e^{\beta\langle y_i(s), y_k(s)\rangle}} \right) y_j(s)\, \mathrm{d}s \right\| \\
&\qquad \leqslant \beta n^{\frac{1}{2}} \int_0^t \left(\sum_{j=1}^n \Big[\langle x_i(s), x_j(s)\rangle - \langle y_i(s), y_j(s)\rangle \Big]^2 \right)^{\frac{1}{2}} \mathrm{d}s \\
&\qquad \leqslant 2\beta n \int_0^t \max_{j\in[n]} \|x_j(s) - y_j(s)\|\, \mathrm{d}s. \qquad \text{(D.3)}
\end{aligned}$$

Using (D.2), (D.3) and arguing similarly for the remaining terms in (D.1), we deduce that

$$\|x_i(t) - y_i(t)\| \leqslant \|x_i(0) - y_i(0)\| + 10\max\{1,\beta\} n \int_0^t \max_{j\in[n]} \|x_j(s) - y_j(s)\|\, \mathrm{d}s.$$

Maximizing over $i \in [n]$ and applying the Grönwall inequality yields

$$\max_{j\in[n]} \|x_j(t) - y_j(t)\| \leqslant c(\beta)^{nt} \max_{j\in[n]} \|x_j(0) - y_j(0)\|\,, \qquad \text{(D.4)}$$

for any $i \in [n]$ and $t \geqslant 0$.

Step 2. Almost orthogonality. Let $x_1(0), \dots, x_n(0) \in \mathbb{S}^{d-1}$ be the random i.i.d. initial points. We prove that with high probability, there exist n pairwise orthogonal points $y_1(0), \dots, y_n(0) \in \mathbb{S}^{d-1}$, such that for any $i \in [n]$,

$$\|x_i(0) - y_i(0)\| \leqslant \sqrt{\frac{\log d}{d}}. \qquad \text{(D.5)}$$

To this end, we take $y_1(0) = x_1(0)$ and then construct the other points $y_i(0)$ by induction. Assume that $y_1(0), \dots, y_i(0)$ are constructed for some $i \in [n]$, using only knowledge about the points $x_1(0), \dots, x_i(0)$. Then by Lévy's concentration of measure, since $x_{i+1}(0)$ is independent from $x_1(0), \dots, x_i(0)$ and uniformly distributed on $\mathbb{S}^{d-1}$,

$$\mathbb{P}\left(\left\{ \operatorname{dist}\left(x_{i+1}(0), \operatorname{span}\{y_1(0), \dots, y_i(0)\}^\perp\right) \leqslant \sqrt{\frac{\log d}{d}} \right\} \right) \geqslant 1 - 4id^{-1/64},$$

for some universal constants $c, C > 0$. Using the union bound, we gather that the event

$$\mathscr{A}_0 = \{\text{(D.5) is satisfied for any } i \in [n]\}$$

has probability at least $p_0 = 1 - 2n^2 d^{-1/64}$. We now consider the event

$$\mathscr{A} = \mathscr{A}_0 \cap \{\text{for some } C, \lambda > 0, \text{ (6.1) holds for any } i \in [n] \text{ and } t \geqslant 0\}$$

which, since $d \geqslant n$ and thus the second event has probability 1, also holds with probability at least $p_0 = 1 - 2n^2 d^{-1/64}$. For the remainder of the proof, we assume that $\mathscr{A}$ is satisfied.

Step 3. Proof of (6.11). Let $(y_i(\cdot))_{i\in[n]} \in C^0(\mathbb{R}_{\geqslant 0}; (\mathbb{S}^{d-1})^n)$ denote the unique solution to the Cauchy problem for (SA) corresponding to the initial datum $(y_i(0))_{i\in[n]}$. A combination of (D.4) and (D.5) yields

$$\|x_i(t) - y_i(t)\| \leqslant c(\beta)^{nt}\sqrt{\frac{\log d}{d}} \tag{D.6}$$

for any $i \in [n]$ and $t \geqslant 0$, under $\mathscr{A}$. Combining (D.6) with Theorem 6.8 we obtain

$$\left|\langle x_i(t), x_j(t)\rangle - \gamma_\beta(t)\right| \leqslant 2c(\beta)^{nt}\sqrt{\frac{\log d}{d}} \tag{D.7}$$

for any $i \neq j$ and $t \geqslant 0$, under $\mathscr{A}$.

We turn to the proof of the second part of (6.11). For this, we prove that for large times t, both $\gamma_\beta(t)$ and $\langle x_i(t), x_j(t)\rangle$ are necessarily close to 1. We first show that

$$1 - \gamma_\beta(t) \leqslant \frac{1}{2}\exp\left(\frac{n^2 e^\beta}{2\left(n + e^{\frac{\beta}{2}}\right)} - \frac{nt}{n + e^{\frac{\beta}{2}}}\right) \tag{D.8}$$

for any $t \geqslant 0$. To this end, we notice that $t \mapsto \gamma_\beta(t)$ is increasing and thus $\gamma_\beta(t) \geqslant 0$, as well as $\dot\gamma_\beta(t) \geqslant \frac{1}{ne^\beta}$ as long as $\gamma_\beta(t) \leqslant \frac{1}{2}$. Therefore,

$$\gamma_\beta\left(\frac{ne^\beta}{2}\right) \geqslant \frac{1}{2}.$$

We deduce that for $t \geqslant \frac{ne^\beta}{2}$,

$$\dot\gamma_\beta(t) \geqslant \frac{n(1-\gamma_\beta(t))}{n + e^{\frac{\beta}{2}}}.$$

Integrating this inequality from $\frac{ne^\beta}{2}$ to t, we obtain (D.8). We now set $d^*(n,\beta) \geqslant n$ such that

$$\frac{d}{\log d} \geqslant \frac{16c(\beta)^2}{\gamma_\beta(\frac{1}{n})^2} \tag{D.9}$$

holds for any $d \geqslant d^*(n,\beta)$. According to Lemma 6.4, since $\mathscr{A}$ is satisfied, there exists $x^* \in \mathbb{S}^{d-1}$ such that $x_i(t) \to x^*$ for any $i \in [n]$ as $t \to +\infty$. We set

$$\alpha(t) := \min_{i\in[n]}\langle x_i(t), x^*\rangle,$$

and prove that

$$1 - \alpha(t) \leqslant \exp\left(\frac{1 - \gamma_\beta\left(\frac{1}{n}\right)t}{2ne^{2\beta}}\right). \tag{D.10}$$

To this end, let us first prove that

$$\alpha\left(\frac{1}{n}\right) \geqslant \frac{1}{2}\gamma_\beta\left(\frac{1}{n}\right). \tag{D.11}$$

From Step 2 in the proof of Lemma 6.4, we gather that x^* lies in the convex cone generated by the points $x_1(t), \dots, x_n(t)$ for any $t > 0$, and so the decomposition

(6.4) holds. Taking the inner product of $x_i(\frac{1}{n})$ with the decomposition (6.4) at time $t = \frac{1}{n}$, we get

$$\alpha\left(\frac{1}{n}\right) \geqslant \min_{(i,j)\in[n]^2} \left\langle x_i\left(\frac{1}{n}\right), x_j\left(\frac{1}{n}\right)\right\rangle \geqslant \gamma_\beta\left(\frac{1}{n}\right) - 2c(\beta)\sqrt{\frac{\log(d)}{d}} \geqslant \frac{1}{2}\gamma_\beta\left(\frac{1}{n}\right),$$

where the second inequality comes from (D.6) evaluated at time $t = \frac{1}{n}$, and the last inequality comes from (D.9). This is precisely (D.11). Using the notation $a_{ij}(t) = Z_{\beta,i}(t)^{-1}e^{\beta\langle x_i(t),x_j(t)\rangle}$ as in the proof of Lemma 6.4, we now find

$$\dot\alpha(t) = \langle \dot x_{i(t)}(t), x^*\rangle \geqslant \sum_{j=1}^n a_{i(t)j}(t)(1 - \langle x_{i(t)}(t), x_j(t)\rangle)\alpha(t) \tag{D.12}$$

for one of the indices $i(t) \in [n]$ achieving the minimum in the definition of $\alpha(t)$. Combining this with (D.11), we gather that $\alpha(t) \geqslant \alpha(\frac{1}{n})$ for $t \geqslant \frac{1}{n}$. But

$$\min_{j\in[n]}\langle x_{i(t)}(t), x_j(t)\rangle \leqslant \sum_{k=1}^n \theta_k(t)\langle x_{i(t)}(t), x_k(t)\rangle = \langle x_{i(t)}(t), x^*\rangle = \alpha(t). \tag{D.13}$$

Plugging (D.13) into (D.12) and using $a_{ij}(t) \geqslant n^{-1}e^{-2\beta}$ we get

$$\dot\alpha(t) \geqslant \frac{1}{ne^{2\beta}}\alpha\left(\frac{1}{n}\right)(1 - \alpha(t)) \tag{D.14}$$

for $t \geqslant \frac{1}{n}$. Integrating (D.14) from $\frac{1}{n}$ to t, we get (D.10). We therefore deduce from (D.10) that

$$\langle x_i(t), x_j(t)\rangle \geqslant 1 - \exp\left(\frac{1 - \gamma_\beta(\frac{1}{n})t}{2ne^{2\beta}}\right)$$

holds for any distinct $i, j \in [n]$. Together with (D.8), we then get

$$\left|\langle x_i(t), x_j(t)\rangle - \gamma_\beta(t)\right| \leqslant \exp\left(\frac{1 - \gamma_\beta(\frac{1}{n})t}{2ne^{2\beta}}\right) + \frac{1}{2}\exp\left(\frac{n^2e^\beta}{2(n + e^{\frac{\beta}{2}})} - \frac{nt}{n + e^{\frac{\beta}{2}}}\right). \tag{D.15}$$

Finally, combining (D.7) and (D.15) we obtain (6.11). □

Remark D.1. *An analogous statement to Theorem 6.9 holds for* (USA)*, where γ_β would rather be the unique solution to* (6.10)*. More concretely, Step 1 in the proof is only slightly changed—the constant one obtains in the analogue of* (6.11) *is rather $c(\beta)^{nt}$ with $c(\beta) = e^{10\beta e^{2\beta}}$ —. Step 2 remains unchanged. In Step 3,* (D.8) *is replaced by $\gamma_\beta(\frac{n}{2}) \geqslant \frac{1}{2}$ and*

$$1 - \gamma_\beta(t) \leqslant \frac{1}{2}\exp\left(-e^{\frac{\beta}{2}}\left(t - \frac{n}{2}\right)\right).$$

The rest of the proof then remains essentially unchanged.

References

[ABK+22] Pedro Abdalla, Afonso S Bandeira, Martin Kassabov, Victor Souza, Steven H Strogatz, and Alex Townsend. Expander graphs are globally synchronising. *arXiv preprint arXiv:2210.12788*, 2022.

[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.

[ACDS23] Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36, 2023.

[ADTK23] Silas Alberti, Niclas Dern, Laura Thesing, and Gitta Kutyniok. Sumformer: Universal Approximation for Efficient Transformers. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pages 72–86. PMLR, 2023.

[AG24] Daniel Owusu Adu and Bahman Gharesifard. Approximate controllability of continuity equation of transformers. *IEEE Control Systems Letters*, 2024.

[AGS05] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[AL24] Andrei Agrachev and Cyril Letrouit. Generic controllability of equivariant systems and applications to particle systems and neural networks. *arXiv preprint arXiv:2404.08289*, 2024.

[AS22] Andrei Agrachev and Andrey Sarychev. Control on the manifolds of mappings with a view to the deep learning. *Journal of Dynamical and Control Systems*, 28(4):989–1008, 2022.

[Bar93] Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

[BB00] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[BBC+09] Brandon Ballinger, Grigoriy Blekherman, Henry Cohn, Noah Giansiracusa, Elizabeth Kelly, and Achill Schürmann. Experimental study of energy-minimizing point configurations on spheres. *Experimental Mathematics*, 18(3):257–283, 2009.

[BCM08] Adrien Blanchet, José A Carrillo, and Nader Masmoudi. Infinite time aggregation for the critical Patlak-Keller-Segel model in $\mathbb{R}^2$. *Communications on Pure and Applied Mathematics*, 61(10):1449–1481, 2008.

[BCM15] Dario Benedetto, Emanuele Caglioti, and Umberto Montemagno. On the complete phase synchronization for the Kuramoto model in the mean-field limit. *Communications in Mathematical Sciences*, 13(7):1775–1786, 2015.

[BD19] Dmitriy Bilyk and Feng Dai. Geodesic distance Riesz energy on the sphere. *Transactions of the American Mathematical Society*, 372(5):3141–3166, 2019.

[BHK24] Han Bao, Ryuichiro Hataya, and Ryo Karakida. Self-attention networks localize when qk-eigenspectrum concentrates. *arXiv preprint arXiv:2402.02098*, 2024.

[BKH16] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[BLG14] Emmanuel Boissard and Thibaut Le Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pages 539–563, 2014.

[BLR11] Andrea L Bertozzi, Thomas Laurent, and Jesús Rosado. L^p theory for the multidimensional aggregation equation. *Communications on Pure and Applied Mathematics*, 64(1):45–83, 2011.

[CCH+14] José A Carrillo, Young-Pil Choi, Seung-Yeal Ha, Moon-Jin Kang, and Yongduck Kim. Contractivity of transport distances for the kinetic Kuramoto equation. *Journal of Statistical Physics*, 156(2):395–415, 2014.

[CD22] Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: a review of models, methods and applications. II. Applications. 2022.

[CDF+11] J. A. Carrillo, M. DiFrancesco, A. Figalli, T. Laurent, and D. Slepčev. Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations. *Duke Mathematical Journal*, 156(2):229 – 271, 2011.

[Chi15] Hayato Chiba. A proof of the Kuramoto conjecture for a bifurcation structure of the infinite-dimensional Kuramoto model. *Ergodic Theory and Dynamical Systems*, 35(3):762–834, 2015.

[CJLZ21] José A Carrillo, Shi Jin, Lei Li, and Yuhua Zhu. A consensus-based global optimization method for high dimensional machine learning problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 27:S5, 2021.

[CK07] Henry Cohn and Abhinav Kumar. Universally optimal distribution of points on spheres. *Journal of the American Mathematical Society*, 20(1):99–148, 2007.

[CKM+22] Henry Cohn, Abhinav Kumar, Stephen Miller, Danylo Radchenko, and Maryna Viazovska. Universal optimality of the E_8 and Leech lattices and interpolation formulas. *Annals of Mathematics*, 196(3):983–1082, 2022.

[CLLS23] Jingpu Cheng, Qianxiao Li, Ting Lin, and Zuowei Shen. Interpolation, approximation and controllability of deep neural networks. *arXiv preprint arXiv:2309.06015*, 2023.

[CLP15] Marco Caponigro, Anna Chiara Lai, and Benedetto Piccoli. A nonlinear model of opinion formation on the sphere. *Discrete & Continuous Dynamical Systems-A*, 35(9):4241–4268, 2015.

[CLT20] Christa Cuchiero, Martin Larsson, and Josef Teichmann. Deep neural networks, generic universal interpolation, and controlled ODEs. *SIAM Journal on Mathematics of Data Science*, 2(3):901–919, 2020.

[CNQG24] Aditya Cowsik, Tamra Nebabu, Xiao-Liang Qi, and Surya Ganguli. Geometric dynamics of signal propagation predict trainability of transformers. *arXiv preprint arXiv:2403.02579*, 2024.

[CNWR24] Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 2024.

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

[CRMB24] Christopher Criscitiello, Quentin Rebjock, Andrew D. McRae, and Nicolas Boumal. Synchronization on circles and spheres with nonlinear interactions, 2024.

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

[Cyb89] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[CZC+22] Tianlong Chen, Zhenyu Zhang, Yu Cheng, Ahmed Awadallah, and Zhangyang Wang. The principle of diversity: Training stronger vision transformers calls for reducing all levels of redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12020–12030, 2022.

[Dai92] Hiroaki Daido. Order function and macroscopic mutual entrainment in uniformly coupled limit-cycle oscillators. *Progress of Theoretical Physics*, 88(6):1213–1218, 1992.

[DBK24] Gbètondji JS Dovonon, Michael M Bronstein, and Matt J Kusner. Setting the record straight on transformer oversmoothing. *arXiv preprint arXiv:2401.04301*, 2024.

[DBPC19] Gwendoline De Bie, Gabriel Peyré, and Marco Cuturi. Stochastic deep networks. In *International Conference on Machine Learning*, pages 1556–1565. PMLR, 2019.

[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.

[DFGV18] Helge Dietert, Bastien Fernandez, and David Gérard-Varet. Landau damping to partially locked states in the Kuramoto model. *Communications on Pure and Applied Mathematics*, 71(5):953–993, 2018.

[DGCC21] Subhabrata Dutta, Tanya Gautam, Soumen Chakrabarti, and Tanmoy Chakraborty. Redesigning the transformer architecture with insights from multi-particle dynamical systems. *Advances in Neural Information Processing Systems*, 34:5531–5544, 2021.

[DGS91] Philippe Delsarte, Jean-Marie Goethals, and Johan Jacob Seidel. Spherical codes and designs. In *Geometry and Combinatorics*, pages 68–93. Elsevier, 1991.

[DGTT24] Puneesh Deora, Rouzbeh Ghaderi, Hossein Taheri, and Christos Thrampoulidis. On the optimization and generalization of multi-head attention. *Transactions on Machine Learning Research*, 2024.

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

[Dud69] R. M. Dudley. The Speed of Mean Glivenko-Cantelli Convergence. *The Annals of Mathematical Statistics*, 40(1):40 – 50, 1969.

[DX13] Feng Dai and Yuan Xu. *Approximation theory and harmonic analysis on spheres and balls*. Springer, 2013.

[E17] Weinan E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 1(5):1–11, 2017.

[EGPZ20] Carlos Esteve, Borjan Geshkovski, Dario Pighin, and Enrique Zuazua. Large-time asymptotics in deep learning. *arXiv preprint arXiv:2008.02491*, 2020.

[EYG23] Carlos Esteve-Yagüe and Borjan Geshkovski. Sparsity in long-time control of neural ODEs. *Systems & Control Letters*, 172:105452, 2023.

[FdHP24] Takashi Furuya, Maarten V de Hoop, and Gabriel Peyré. Transformers are universal in-context learners. *arXiv preprint arXiv:2408.01367*, 2024.

[FGVG16] Bastien Fernandez, David Gérard-Varet, and Giambattista Giacomin. Landau damping in the Kuramoto model. In *Annales Henri Poincaré*, volume 17, pages 1793–1823. Springer, 2016.

[FHPS21] Massimo Fornasier, Hui Huang, Lorenzo Pareschi, and Philippe Sünnen. Consensus-based optimization on the sphere: Convergence to global minimizers and machine learning. *The Journal of Machine Learning Research*, 22(1):10722–10776, 2021.

[FL19] Amic Frouvelle and Jian-Guo Liu. Long-time dynamics for a simple aggregation equation on the sphere. In *Stochastic Dynamics Out of Equilibrium: Institut Henri Poincaré, Paris, France, 2017*, pages 457–479. Springer, 2019.

[FZH+22] Ruili Feng, Kecheng Zheng, Yukun Huang, Deli Zhao, Michael Jordan, and Zheng-Jun Zha. Rank diminishing in deep neural networks. *Advances in Neural Information Processing Systems*, 35:33054–33065, 2022.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016.

[GLBM24] Arnaud Guillin, Pierre Le Bris, and Pierre Monmarché. Uniform in time propagation of chaos for the 2D vortex model and other singular stochastic systems. *Journal of the European Mathematical Society*, 2024.

[GLPR24] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36, 2024.

[Gol16] François Golse. On the dynamics of large particle systems in the mean field limit. *Macroscopic and large scale phenomena: coarse graining, mean field limits and ergodicity*, pages 1–144, 2016.

[GSRT13] Isabelle Gallagher, Laure Saint-Raymond, and Benjamin Texier. *From Newton to Boltzmann: hard spheres and short-range potentials*. European Mathematical Society Zürich, Switzerland, 2013.

[GWDW23] Xiaojun Guo, Yifei Wang, Tianqi Du, and Yisen Wang. Contranorm: A contrastive learning perspective on oversmoothing and beyond. In *The Eleventh International Conference on Learning Representations*, 2023.

[GZ22] Borjan Geshkovski and Enrique Zuazua. Turnpike in optimal control of PDEs, ResNets, and beyond. *Acta Numerica*, 31:135–263, 2022.

[HHL23] Jiequn Han, Ruimeng Hu, and Jihao Long. A class of dimension-free metrics for the convergence of empirical measures. *Stochastic Processes and their Applications*, 164:242–287, 2023.

[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.

[HKPZ16] Seung-Yeal Ha, Dongnam Ko, Jinyeong Park, and Xiongtao Zhang. Collective synchronization of classical and quantum oscillators. *EMS Surveys in Mathematical Sciences*, 3(2):209–267, 2016.

[HKR18] Seung-Yeal Ha, Dongnam Ko, and Sang Woo Ryoo. On the relaxation dynamics of Lohe oscillators on some Riemannian manifolds. *Journal of Statistical Physics*, 172:1427–1478, 2018.

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

[HR20] Seung-Yeal Ha and Seung-Yeon Ryoo. Asymptotic phase-locking dynamics and critical coupling strength for the Kuramoto model. *Communications in Mathematical Physics*, 377(2):811–857, 2020.

[HZRS16] 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.

[JDB23] Amir Joudaki, Hadi Daneshmand, and Francis Bach. On the impact of activation and normalization in obtaining isometric embeddings at initialization. *Advances in Neural Information Processing Systems*, 36:39855–39875, 2023.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[JL23] Haotian Jiang and Qianxiao Li. Approximation theory of transformer networks for sequence modeling. *arXiv preprint arXiv:2305.18475*, 2023.

[JLLW23] Haotian Jiang, Qianxiao Li, Zhong Li, and Shida Wang. A brief survey on the approximation theory for sequence modelling. *arXiv preprint arXiv:2302.13752*, 2023.

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

[KO02] Robert V Kohn and Felix Otto. Upper bounds on coarsening rates. *Communications in Mathematical Physics*, 229(3):375–395, 2002.

[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.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[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.

[Lac23] Daniel Lacker. Hierarchies, entropy, and quantitative propagation of chaos for mean field diffusions. *Probability and Mathematical Physics*, 4(2):377–432, 2023.

[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.

[LCT18] Qianxiao Li, Long Chen, and Cheng Tai. Maximum principle based algorithms for deep learning. *Journal of Machine Learning Research*, 18:1–29, 2018.

[Li21] Wuchen Li. Hessian metric via transport information geometry. *Journal of Mathematical Physics*, 62(3), 03 2021.

[LJ18] Hongzhou Lin and Stefanie Jegelka. Resnet with one-neuron hidden layers is a universal approximator. *Advances in Neural Information Processing Systems*, 31, 2018.

[LLF23] Daniel Lacker and Luc Le Flem. Sharp uniform-in-time propagation of chaos. *Probability Theory and Related Fields*, pages 1–38, 2023.

[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.

[LLS22] Qianxiao Li, Ting Lin, and Zuowei Shen. Deep learning via dynamical systems: An approximation perspective. *Journal of the European Mathematical Society*, 25(5):1671–1709, 2022.

[Loj63] Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

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

[LXB19] Shuyang Ling, Ruitu Xu, and Afonso S Bandeira. On the landscape of synchronization networks: A perspective from nonconvex optimization. *SIAM Journal on Optimization*, 29(3):1879–1907, 2019.

[Mis24] MistralAI. `https://github.com/mistralai/mistral-finetune/blob/main/model/transformer.py`, 2024.

[MP22] Javier Morales and David Poyato. On the trend to global equilibrium for Kuramoto oscillators. *Annales de l'Institut Henri Poincaré C*, 40(3):631–716, 2022.

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

[MTG17] Johan Markdahl, Johan Thunberg, and Jorge Gonçalves. Almost global consensus on the n-sphere. *IEEE Transactions on Automatic Control*, 63(6):1664–1675, 2017.

[NAB+22] Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022.

[NLL+24] Lorenzo Noci, Chuning Li, Mufan Li, Bobby He, Thomas Hofmann, Chris J Maddison, and Dan Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. *Advances in Neural Information Processing Systems*, 36, 2024.

[Ope24] OpenAI. `https://github.com/openai/gpt-2/blob/master/src/model.py`, 2024.

[OR07] Felix Otto and Maria G Reznikoff. Slow motion of gradient flows. *Journal of Differential Equations*, 237(2):372–420, 2007.

[Ott01] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.

[PH22] Mary Phuong and Marcus Hutter. Formal algorithms for transformers. *arXiv preprint arXiv:2207.09238*, 2022.

[PTTM17] René Pinnau, Claudia Totzeck, Oliver Tse, and Stephan Martin. A consensus-based model for global optimization and its mean-field limit. *Mathematical Models and Methods in Applied Sciences*, 27(01):183–204, 2017.

[RBZ23] Domenec Ruiz-Balet and Enrique Zuazua. Neural ODE control for classification, approximation, and transport. *SIAM Review*, 65(3):735–773, 2023.

[RS23] Matthew Rosenzweig and Sylvia Serfaty. Global-in-time mean-field convergence for singular Riesz-type diffusive flows. *The Annals of Applied Probability*, 33(2):954–998, 2023.

[RZZD23] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2023.

[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.

[Ser20] Sylvia Serfaty. Mean field limit for Coulomb-type flows. *Duke Mathematical Journal*, 169(15), 2020.

[SFG+12] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550 – 1599, 2012.

[Shu13] Michael Shub. *Global stability of dynamical systems*. Springer Science & Business Media, 2013.

[Str00] Steven H Strogatz. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D: Nonlinear Phenomena*, 143(1-4):1–20, 2000.

[SWJS24] Michael Scholkemper, Xinyi Wu, Ali Jadbabaie, and Michael Schaub. Residual connections and normalization can provably prevent oversmoothing in gnns. *arXiv preprint arXiv:2406.02997*, 2024.

[Sze39] Gabor Szegö. *Orthogonal polynomials*, volume 23. American Mathematical Soc., 1939.

[Tad23] Eitan Tadmor. Swarming: hydrodynamic alignment with pressure. *Bulletin of the American Mathematical Society*, 60(3):285–325, 2023.

[Tan17] Yan Shuo Tan. Energy optimization for distributions on the sphere and improvement to the Welch bounds. *Electronic Communications in Probability*, 22:1 – 12, 2017.

[Tay12] Richard Taylor. There is no non-zero stable fixed point for dense networks in the homogeneous Kuramoto model. *Journal of Physics A: Mathematical and Theoretical*, 45(5):055102, 2012.

[TG22] Paulo Tabuada and Bahman Gharesifard. Universal approximation power of deep residual neural networks through the lens of control. *IEEE Transactions on Automatic Control*, 2022.

[TLTO23] Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *Advances in Neural Information Processing Systems*, 36, 2023.

[TSS20] Alex Townsend, Michael Stillman, and Steven H Strogatz. Dense networks that do not synchronize and sparse ones that do. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(8), 2020.

[VBC20] James Vuckovic, Aristide Baratin, and Remi Tachet des Combes. A mathematical theory of attention. *arXiv preprint arXiv:2007.02876*, 2020.

[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.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[Vil01] Cédric Villani. Limite de champ moyen. *Cours de DEA*, 2002:49, 2001.

[Vil09] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[VR23] Tanya Veeravalli and Maxim Raginsky. Nonlinear controllability and function representation by neural stochastic differential equations. In *Learning for Dynamics and Control Conference*, pages 838–850. PMLR, 2023.

[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.

[WAW+24] Xinyi Wu, Amir Ajorlou, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the role of attention masks and layernorm in transformers. *arXiv preprint arXiv:2405.18781*, 2024.

[WAWJ24] Xinyi Wu, Amir Ajorlou, Zihui Wu, and Ali Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[Wen62] James G Wendel. A problem in geometric probability. *Mathematica Scandinavica*, 11(1):109–111, 1962.

[XZ23] Tong Xiao and Jingbo Zhu. Introduction to Transformers: an NLP Perspective. *arXiv preprint arXiv:2311.17633*, 2023.

[YBR+19] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2019.

[ZB21] Aaron Zweig and Joan Bruna. A functional perspective on learning symmetric functions with neural networks. In *International Conference on Machine Learning*, pages 13023–13032. PMLR, 2021.

[ZGUA20] Han Zhang, Xi Gao, Jacob Unterman, and Tom Arodz. Approximation capabilities of neural ODEs and invertible residual networks. In *International Conference on Machine Learning*, pages 11086–11095. PMLR, 2020.

[ZLL+23] Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pages 40770–40803. PMLR, 2023.

[ZMZ+23] Haiteng Zhao, Shuming Ma, Dongdong Zhang, Zhi-Hong Deng, and Furu Wei. Are more layers beneficial to graph transformers? In *The Eleventh International Conference on Learning Representations*, 2023.

[ZS19] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Ave, 02139 Cambridge MA, USA
Email address: borjan@mit.edu

CNRS & Université Paris-Saclay, Laboratoire de mathématiques d'Orsay, 307 rue Michel Magat, Bâtiment 307, 91400 Orsay, France
Email address: cyril.letrouit@universite-paris-saclay.fr

Department of EECS, Massachusetts Institute of Technology, 77 Massachusetts Ave, 02139 Cambridge MA, USA
Email address: yp@mit.edu

Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Ave, 02139 Cambridge MA, USA
Email address: rigollet@math.mit.edu