A MATHEMATICAL PERSPECTIVE ON TRANSFORMERS

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

Abstract. A Transformer is a neural network architecture that plays a central role in the inner workings of large language models such as the ChatGPT chatbot. We set out a mathematical framework for analyzing Transformers based on their interpretation as interacting particle systems, which reveals that clusters emerge in long time. Our study explores the underlying theory and offers new perspectives for mathematicians as well as computer scientists.

Contents

1. Outline

The introduction of *Transformers* in 2017 by Vaswani et al. [VSP+17] marked a significant milestone in development of neural network architectures. Central to this contribution is the *self-attention mechanism*, a novelty which distinguishes Transformers from traditional architectures, and which contributes substantially to their superior practical performance. In fact, this innovation has been a key

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

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 a continuous-time dynamical system called neural ordinary differential equations (neural ODEs) [CRBD18, E17, HR17]. In this context, an input $x(0)$, 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].

In this article 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 indicate that the limiting distribution is actually a point mass, leaving no room for diversity or randomness, which is at odds with practical observations. To explain this phenomenon, we suggest the existence of a long metastable epoch during which the particles are assembled in a small number of clusters. They stay in this clustered metastable state for a long time until they relax to a point mass asymptotically.

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 allows one to draw concrete connections to established topics in mathematics, including nonlinear transport equations, Wasserstein gradient flows, opinion formation 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 consists in viewing the discrete layer indices as a continuous time variable. This abstraction is not new and parallels one employed in classical architectures such as

ResNets [CRBD18, E17, HR17]. This model focuses exclusively on two key components of the Transformers architecture: *self-attention* and *layer-normalization*. Layer-normalization essentially constrains particles to evolve on the unit sphere $\mathbb{S}^{d-1}$, whereas self-attention is the 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). 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.

Part 2: Clustering. In this part we establish new mathematical results that indicate clustering of tokens in the large time limit. Our main result, Theorem 4.1, indicates that in high dimension $d \geqslant n$, a set of n particles randomly initialized on $\mathbb{S}^{d-1}$ will cluster to a single point as $t \to +\infty$. We complement this result with a 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 (see Section 4). We also obtain a clustering result without assuming that the dimension d is large, in another asymptotic regime motivated by the statistical physics' nature of the system (Section 5).

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. In this context, we also propose evoke the BBGKY hierarchy for the two-point correlation function (Section 8). 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 interaction particle systems that allow for parameter tuning of the Transformers architectures, a key feature of practical implementations.

part: modeling

Part 1. Modeling

We begin our discussion by presenting the mathematical model for a Transformer (Section 2). While we focus on a simplified version that includes the self-attention mechanism as well as layer normalization, but excludes additional feed-forward layers commonly used in practice; See Section 2.3.2. This leads leading 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 distribution of particles that we study in Section 3.

sec: ips

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 residual neural network. The specifics of the Transformer model are presented in Section 2.2.

sec: resnets

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,

{eq: resnet}

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

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 of the ResNet given the i-th input, namely $x_i(L) \in \mathbb{R}^d$, is projected to $\mathbb{R}^m$ via a trained transformation as to match the label $f(x^{(i)})$ according to the user-specified objective. One can also devise generalizations of (2.1), for instance in which matrix-vector multiplications are replaced by discrete convolutions. 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

{eq: neural.ode}

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

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.

s:interacting

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.

Practical implementations of Transformers make use of *layer normalization* [BKH16], which amounts to an element-wise standardization of every particle at every layer. This effectively constrains particles to evolve on the unit sphere $\mathbb{S}^{d-1}$.

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

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

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

$$\mathbf{P}_x y = y - \langle x, y\rangle x$$

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

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

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[1], which, in light of the apparent statistical physics appearance of the system, can be seen as an inverse temperature. 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*

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

which is the nonlinear coupling mechanism in the interacting particle system. The $n \times n$ stochastic matrix $A(t)$ (rows are probability vectors) 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 gives 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 $(A_{ij}(\cdot))_{j\in[n]}$ ($i \in [n]$) 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]). While understanding this geometry is an important practical question, we focus on understanding the self-attention mechanism in the simplest possible geometry: the Euclidean geometry. In fact, 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, resulting in the dynamics

{eq: transformerSd}
$$\dot{x}_i(t) = \mathbf{P}_{x_i(t)} \left(\frac{1}{Z_{\beta,i}(t)} \sum_{j=1}^{n} e^{\beta\langle x_i(t), x_j(t)\rangle} x_j(t) \right) \tag{2.6}$$

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

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

[1] 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.

The dynamics (2.6) have a strong resemblance to the vast literature on nonlinear systems arising in the modeling of opinion dynamics and flocking phenomena. 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 dynamics 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 nonlinearly as in (2.6).

The appearance of clusters in Transformers is actually corroborated by numerical experiments with pre-trained models (see Figure 1 for instance). While we focus on a much simplified model, numerical evidence shows that the clustering phenomenon looks qualitatively the same in the cases $Q = K = V = I_d$ and generic random (Q, K, V) (see Figures 2 and 4 for instance). We defer the interested reader directly to Section 4; here, we continue the presentation on the modeling of different mechanisms appearing in the Transformer architecture.

Remark 2.1 (Positional encoding)**.** *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 $x_i(0) \in \mathbb{R}^d$ is not simply equal to a word embedding $w_i \in \mathbb{R}^d$, but contains an additional positional encoding, which allows tokens to also carry their position the input prompt.*

There are various 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

$$(p_i)_{2k} = \sin\left(\frac{k}{M^{\frac{2k}{d}}}\right), \qquad (p_i)_{2k+1} = \cos\left(\frac{k}{M^{\frac{2k}{d}}}\right)$$

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 of both encodings: $x_i(0) = w_i + p_i$. Subsequent studies simply use either a random positional encoding[2] (i.e., p_i is just some random vector) or a trained transformation. The addition of both codes can also be replaced by concatenation. (See [LWLQ22, XZ23] for details.) We do not cover the specifics of the positional encoding in this paper.

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*

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

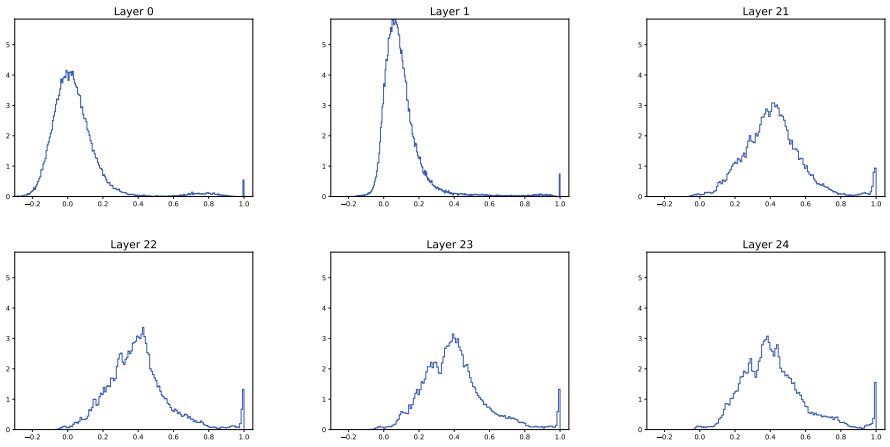

Figure 1. Histogram of $\{\langle x_i(t), x_j(t)\rangle\}_{(i,j)\in[n]^2, i\neq j}$ at different layers t in the context of the pre-trained ALBERT XLarge v2 model ([LCG+20] and https://huggingface.co/albert-xlarge-v2), which has constant parameter matrices. Here we randomly selected a single prompt, 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). If the number of layers is increased, up to 48 say, the clustering is further enhanced (bottom).

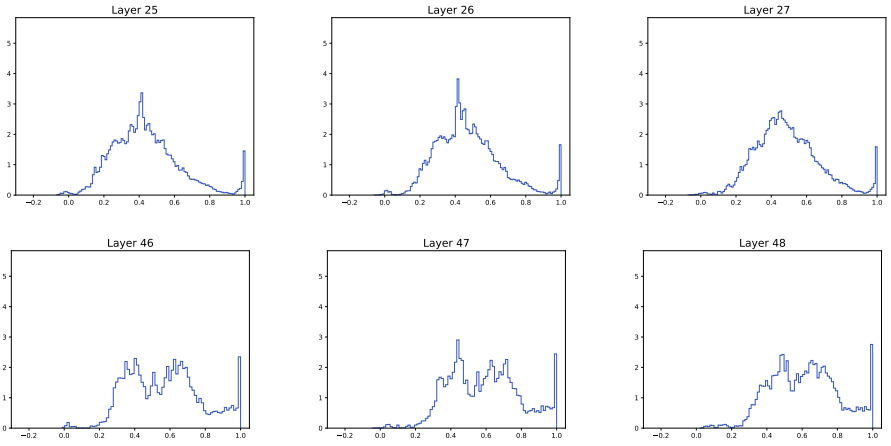

fig: albert

Transformer (2.6)*, mapping* $(x_i(0))_{i\in[n]} \in (\mathbb{S}^{d-1})^n$ *to* $(x_i(t))_{i\in[n]} \in (\mathbb{S}^{d-1})^n$*, is a permutation-equivariant function. It is unclear whether this feature is useful in common natural language processing applications, as the Transformer output yields a probability distribution from which samples are drawn. It has however found applications in the sciences (see* [BVE23]*).*

2.3. **Toward the complete Transformer.** There are a couple of additional mechanisms used in practical implementations that we do not cover 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 (2.6):

{eq: multihead}

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

where $Z_{\beta,i,h}(t)$ is defined as in (2.4) for the matrices Q_h and K_h. The integer $H \geqslant 1$ is called the number of heads[3].

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, in the vein of [Cyb89, Bar93],

Problem 1 (Approximation)**.** *Fix $d, n \geqslant 2$ and consider the 1-hidden layer spherical Transformer with multi-headed self attention $f_\theta^H : (\mathbb{S}^{d-1})^n \to (\mathbb{S}^{d-1})^n$ defined as*

$$f_\theta^H(x_1, \dots, x_n)_i = \mathbf{P}_{x_i} \left(\sum_{h=1}^{H} \sum_{j=1}^{n} \frac{e^{\beta \langle Q_h x_i, K_h x_j \rangle}}{Z_{\beta,i,h}} V_h x_j \right),$$

where $H \geqslant 1$ and $\theta = (Q_h, K_h, V_h)_{h \in [H]}$ are as for (2.7). Can one approximate, in some appropriate topology, any continuous and permutation-equivariant function $f : (\mathbb{S}^{d-1})^n \to (\mathbb{S}^{d-1})^n$ by means of some f_θ^H as $H \to +\infty$? The same question for the multi-headed Transformer without layer normalization: $f_\theta^H : (\mathbb{R}^d)^n \to (\mathbb{R}^d)^n$ defined as

$$f_\theta^H(x_1, \dots, x_n)_i = \sum_{h=1}^{H} \sum_{j=1}^{n} \frac{e^{\beta \langle Q_h x_i, K_h x_j \rangle}}{Z_{\beta,i,h}} V_h x_j,$$

is also open.

A universal approximation property of the above kind would then motivate studying the training dynamics of infinite-width (i.e., infinite number of heads) 1-hidden layer spherical Transformers, similar to what has been done for the neural network analog in recent years [CB18, MMN18, RVE22]. None of these questions has received a definitive answer for Transformers; see [YBR+19] for related work when the depth is taken to infinity.

sec:feedforward

2.3.2. *Feed-forward layers.* The complete Transformer dynamics combines all of the above mechanisms with a feed-forward layer. This amounts to considering dynamics of the form

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

[3]In practical implementations, H is a divisor of d, and the query and key matrices Q_h and K_h 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 $w(t), b(t)$ and σ are as in (2.2). These layers are critical 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, and we leave it open to further investigation.

3. Measure to measure flow map

s:WGF

The vector field driving the evolution of a single particle in (2.6)–(SA) clearly depends on all n particles. In fact, one can equivalently rewrite the dynamics as

{eq: mean.field.ips}
$$\dot{x}_i(t) = \mathcal{X}[\mu(t)](x_i(t)) \tag{3.1}$$

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

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

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

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

with

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

In other words, (2.6) is a *mean-field interacting particle system*. And due to the equivalence of the dynamics (3.1) satisfied by the particles, and the continuity equation[4]

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

satisfied (in the sense of distributions) by the empirical measure, we naturally see Transformers as flow maps between probability measures.

Remark 3.1. *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 identically as in* [GLPR23, Lemma A.3]. *Here and henceforth,* $\mathcal{P}(\mathbb{S}^{d-1})$ *stands for the set of Borel probability measures on* $\mathbb{S}^{d-1}$.

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 works of Dobrushin in the late 1970s [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[5] $\lim_{n\to+\infty} W_1(\mu_n(0), \mu(0)) = 0$ for some probability measure $\mu(0) \in \mathcal{P}(\mathbb{S}^{d-1})$. Consider the solutions $\mu_n(t)$ and $\mu(t)$ to

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

[5] Here W_p denotes the Wasserstein distance of order p–see [Vil09] for definitions.

(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, Chapitre 4, Section 1] or [Gol16, Section 1.4.2]). This elementary mean-field limit result has a couple of caveats. First of all, the time-dependence is exponential. Second of all, if one assumes that the points $x_i(0)$ are sampled i.i.d. according to μ_0, then the convergence of the empirical measure to μ_0 suffers from the curse of dimensionality [BLG14, WB19]. There may be hope that a dimension-free limit can be obtained, for instance by a more careful choice of metric ([HHL23], see also [Lac23] for recent work in the diffusive case). 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, GBM21]. We do not address this question in further detail here. For a non-exhaustive list of references on this well-established topic, the reader is referred to [Vil01, Gol16, Ser20] and the references therein.

3.1. **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* (or *free*) *energy*

{eq: interaction.energy}

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

is one such quantity. Indeed,

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

{eq: dissipation.softmax}

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.

prop: existence.uniqueness.energy

Proposition 3.2. *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[6] 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 3.2 (see [Tan17] for a different approach).

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

Proof of Proposition 3.2. Let $f(t) = e^{\beta t}$. The free energy then reads

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

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 3.2 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. □

3.2. **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 descent in the Euclidean context.

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

{eq: logder}

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

(This observation goes beyond $Q = K = I_d$ and $V = \pm I_d$ as long as $Q^\top K = V$.) And 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.

One is then naturally led to look for ways to "symmetrize" (3.4). A first attempt would be to remove the logarithm in (3.7), which amounts to removing the denominator in (3.2). This is one point where working on the unit sphere is useful: should one work on $\mathbb{R}^d$, without layer normalization, the resulting vector field would grow exponentially with the size of the support of the measure, rendering a Cauchy-Lipschitz argument inapplicable. On the contrary, on $\mathbb{S}^{d-1}$ the resulting equation is perfectly well-posed.

Remark 3.3. *Considering the Transformer dynamics on $\mathbb{R}^d$, thus without layer normalization, the authors in* [SABP22] *propose an alternative symmetric yet non-trivial 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 (see* [SABP22, Proposition 2]*), but since the resulting kernel is a limit of iterations and is not explicit, it appears difficult to analyze at a first glance.*

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

{eq:spherenonsoftmax}

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

which is obtained by replacing the partition function $Z_{\beta,i}(t)$ by n. As a matter of fact, (USA) will present a remarkably similar qualitative behavior. The continuity equation corresponding to (USA), namely

{eq: pde.nosoftmaxZ}

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

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

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

satisfies

{e:XmuE}

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

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

We omit the proof. Here $\mathcal{P}_{\mathrm{ac}}(\mathbb{S}^{d-1})$ denotes the set of probability measures which are absolutely continuous with respect to the Lebesgue measure σ_d on $\mathbb{S}^{d-1}$. A derivation of the gradient flow formulation for discrete measures is provided in Remark 3.6, for which all of the conclusions discussed in this sectionalso hold.

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

{eq: aggregation.eq}

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

The considerations above lead us to the following Lyapunov identity.

lem: dissipation

Lemma 3.5. *The solution $\mu \in C^0(\mathbb{R}_{\geqslant 0}; \mathcal{P}_{\mathrm{ac}}(\mathbb{S}^{d-1}))$ 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.

rem: particle.gf

Remark 3.6. *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

{e:dynonX}

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

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

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

which yields (3.11).

part: clustering

Part 2. Clustering

As alluded to in the introductory discussion, clustering is of particular relevance in tasks such as next-token prediction. Therein, the output measure encodes the probability distribution of the next token, and its clustering indicates a small number of possible outcomes. In Sections 4 and 5, we prove several results 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 in particular asymptotic regimes and for the specific choice of parameter matrices, and apply in possibly very long time horizons. Numerical experiments indicate a more complicated 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 appears consistent with behavior

in pre-trained models (Figure 1). We are not able to theoretically explain this behavior as of now.

Ultimately, the appearance of clusters is somewhat natural, since the Transformer dynamics is 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*.

sec: single.cluster

4. A single cluster in high dimension

The clustering results we present in this section are restricted to the high-dimensional regime. We cover the case of arbitrary dimension d, when $\beta \approx 0$, in Section 5. Further avenues for tackling the low-dimensional case are given in Section 7 and Section 8, whereas the repulsive case $V = -I_d$ is commented in Section 9.1.

4.1. Clustering when $d \geqslant n$.

Our first result shows the emergence of a single cluster in high dimension and reads as follows.

thm: d.infty

Theorem 4.1. *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[7] for (2.6)–(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 with probability equal to one, there exists some $x^* \in \mathbb{S}^{d-1}$ such that*

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

for all $i \in [n]$.

This is referred to as convergence toward *consensus* in collective behavior models.

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[8] $w \in \mathbb{S}^{d-1}$ such that $\langle w, x_i(0)\rangle > 0$ for any $i \in [n]$. In other words, all of the initial points must lie in an open hemisphere. The proof of Theorem 4.1 thus follows as a direct corollary of

lem: hemisphere.clustering

Lemma 4.2. *Let $\beta > 0$. Let $(x_i(0))_{i\in[n]} \in (\mathbb{S}^{d-1})^n$ be such that there exists $w \in \mathbb{S}^{d-1}$ with $\langle x_i(0), w\rangle > 0$ for any $i \in [n]$. 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 (2.6)–(SA) or (USA). Then there exists $x^* \in \mathbb{S}^{d-1}$ such that*

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

for all $i \in [n]$.

[7] We refer to the initial value problem for the ODE as Cauchy problem (terminology typically reserved for PDEs) due to the equivalence with the PDE satisfied by the empirical measure.

[8] This instance of *Wendel's theorem* (Theorem 4.5) is easy to prove. Let us denote by $\mathcal{A}$ the event on which the points $x_1(0), \dots, x_n(0)$ are linearly independent. When $d \geqslant n$, $\mathcal{A}$ has probability 1. Assuming that $\mathcal{A}$ is realized, we denote by $X \in \mathbb{R}^{n\times d}$ the matrix whose i-th row is given by $x_i(0)^\top$. There exists a subset of indices $J \subset [d]$ of size n such that the $n \times n$ submatrix of X, formed by the columns of X with indices in J, is invertible. Pick $v \in \mathbb{R}^d \backslash \{0\}$ with coordinates $v_j = 0$ for $j \notin J$ such that $Xv = (1, \dots, 1)^\top \in \mathbb{R}^n$. Then $w = v/\|v\|$ satisfies the claim.

Remark 4.3. *Lemma 4.2 implies that* $\left\{(\bar{x}_i)_{i\in[n]} \in (\mathbb{S}^{d-1})^n \colon \bar{x}_1 = \ldots = \bar{x}_n\right\}$ *is Lyapunov asymptotically stable as a set (and actually, exponentially stable).*

Results like Lemma 4.2 are commonplace in the literature on interacting particle systems on the sphere–see for instance 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.

lem: ez.lemma

Lemma 4.4. *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 $\lim_{t\to+\infty} f(t) = 0$.

Since the proof of Lemma 4.2, which is an adaptation of [CLP15, Theorem 1], is rather succinct, we present it here.

Proof of Lemma 4.2. We focus on the case (USA) and set $a_{ij}(t) := e^{\beta\langle x_i(t), x_j(t)\rangle}$. (The proof for (2.6)–(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.) For $t \geqslant 0$, consider

$$i(t) \in \arg\min_{i\in[n]} \langle x_i(t), w\rangle.$$

Fix $t_0 \geqslant 0$. We have

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

with equality only if $x_1(t_0) = \ldots = x_n(t_0)$. Therefore 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

{eq: therighthandside}

$$\lim_{k\to+\infty} \langle \dot{x}_i(t_k), w\rangle = \sum_{j=1}^n \overline{a}_{ij}(\langle w, \overline{x}_j\rangle - \langle \overline{x}_i, \overline{x}_j\rangle\langle \overline{x}_i, w\rangle) \geqslant r_\infty \sum_{j=1}^n \overline{a}_{ij}(1 - \langle \overline{x}_i, \overline{x}_j\rangle), \tag{4.1}$$

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 4.4, the left-hand side term in (4.1) 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. The result follows. □

If we cease to assume that $d \geqslant n$ in Theorem 4.1, we can still apply Wendel's theorem (recalled below) together with Lemma 4.2 to obtain clustering to a single point with probability at least p_0 for some explicit $p_0 \in (0,1)$.

r:wendel

Theorem 4.5 (Wendel, [Wen62])**.** *Let $d, n \geqslant 1$. 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:*

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

4.2. **Precise convergence in high dimension.** In the regime where n is fixed and $d \to +\infty$, in addition to showing the formation of a cluster as in Theorem 4.1, it is possible to quantitatively describe the evolution of the particles with high probability. To motivate this, on 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* (due to concentration of measure), and we are thus able to compare their evolution with that of an initial sequence of truly *orthogonal* ones. For orthogonal initial particles, the dynamics is particularly simple, as described in Theorem 4.6 below.

thm: orthogonal

Theorem 4.6. *Let $\beta \geqslant 0$, $d, n \geqslant 2$ be arbitrary. Consider a 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* (2.6)–(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, $\gamma_\beta(t) := \cos(\theta_\beta(t))$ satisfies

eq: ybeta

$$\begin{cases} \dot\gamma_\beta(t) = \dfrac{2e^{\beta\gamma_\beta(t)}(1-\gamma_\beta(t))((n-1)\gamma_\beta(t)+1)}{e^\beta + (n-1)e^{\beta\gamma_\beta(t)}} & \text{for } t \in \mathbb{R}_{\geqslant 0} \\ \gamma_\beta(0) = 0 \end{cases} \tag{4.2}$$

(resp.

eq: ybetaUSA

$$\begin{cases} \dot\gamma_\beta(t) = \dfrac{2}{n} e^{\beta\gamma_\beta(t)}(1-\gamma_\beta(t))((n-1)\gamma_\beta(t)+1) & \text{for } t \in \mathbb{R}_{\geqslant 0} \\ \gamma_\beta(0) = 0 \end{cases} \tag{4.3}$$

for (USA)*.)*

Here and henceforth, $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$ denotes for the one-dimensional torus. We provide a brief proof of Theorem 4.6 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 $i, j \in [n]$.

thm: phase.transition.curve

Theorem 4.7. *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* (2.6)–(SA)*.*

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

{eq: upto-t}

$$\left|\langle x_i(t), x_j(t)\rangle - \gamma_\beta(t)\right| \leqslant \min\left\{2 \cdot C(\beta)^{nt}\sqrt{\frac{\log d}{d}}, \kappa e^{-\lambda t}\right\} \tag{4.4}$$

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 (4.2).

Since the proof is rather lengthy, we defer the reader to Section 6.1. It essentially 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 6.1, whereas the explicit values of κ and λ can be found in (6.16). The upper bound in (4.4) is of interest in regimes where d and/or t are sufficiently large–one can otherwise consider the trivial bound equal to 2.

Proof of Theorem 4.6. We split the proof in two parts. We focus on proving the result for the dynamics (2.6)–(SA), since the arguments adapt straightforwardly 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 (2.6) 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 (2.6) 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 (4.2) for the (2.6)–(SA) dynamics, we note that

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

Then for $k \neq i$,

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

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. □

4.3. **Metastability and a phase transition.** An interesting byproduct of Theorem 4.6 and Theorem 4.7 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

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

where $(x_i(\cdot))_{i\in[n]}$ denotes the solution to the corresponding Cauchy problem for (2.6)–(SA). (Here the choice of the first two particles instead of a random distinct pair is justified due to permutation equivariance.) Theorem 4.7 then gives the intuition that over compact subsets of $(\mathbb{R}_{\geqslant 0})^2$, $\Gamma_{d,\delta}$ should be well-approximated by

{eq: gamma.infty}

$$\Gamma_{\infty,\delta}=\Big\{(t,\beta)\in(\mathbb{R}_{\geqslant 0})^2\colon \gamma_\beta(t)=1-\delta\Big\}. \tag{4.5}$$

This is clearly seen in Figure 2[9], along with the fact that the resolution of this approximation increases with $d\to+\infty$.

Figure 2 appears to contain more information than what we may gather from Theorem 4.1, Theorem 4.6 and Theorem 4.7. In particular, we see the appearance of a zone (light shade of blue in Figure 2) of (t,β) for which the probability of particles being clustered is positive, but not close to one. This zone appears to shrink as $d\to+\infty$, and in low-dimension entails a sort of metastability (see Figure 3), which leads us to formulate the following question.

Problem 2. *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^*?*

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 are given in Figure 4.

Problem 3. *Can the conclusions of Theorem 4.6–Theorem 4.7 be generalized to the case of random matrices (Q,K,V)?*

[9]Figures 2–4 contain phase diagrams which in this context are heatmaps of the function $(t,\beta)\mapsto\mathbb{E}_{(x_1(0),\dots,x_n(0))\sim\sigma_d}\big[1_{\{\langle x_1(t),x_2(t)\rangle\geqslant 1-\delta\}}\big]$. By permutation equivariance, this is equal to the function $(t,\beta)\mapsto\mathbb{E}_{(x_1(0),\dots,x_n(0))\sim\sigma_d,i\neq j\text{ fixed}}\big[1_{\{\langle x_1(t),x_2(t)\rangle\geqslant 1-\delta\}}\big]$. 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. We take $\delta=10^{-3}$ throughout.

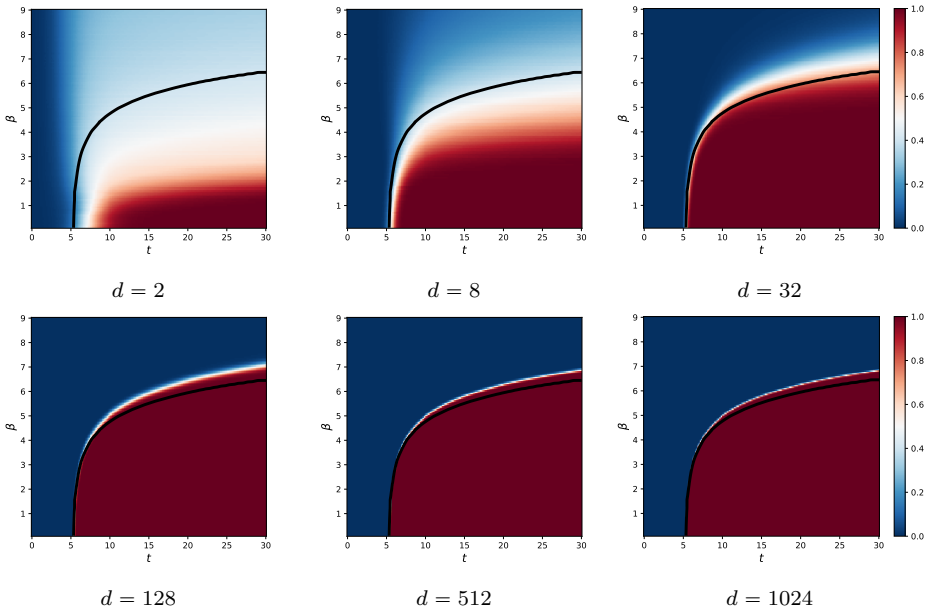

Figure 2. Plots of the probability that randomly initialized particles following (2.6)–(SA) cluster to a single point, in the sense that $\langle x_1(t), x_2(t)\rangle \geqslant 0.999$ (thus $\delta = 10^{-3}$) as a function of t and β. Here, $n = 32$, while d varies. We see that the curve $\Gamma_{\infty,\delta}$ defined in (4.5) approximates the actual phase transition with increasing accuracy as d grows, as implied by Theorem 4.7.

fig: phase.diag.Id

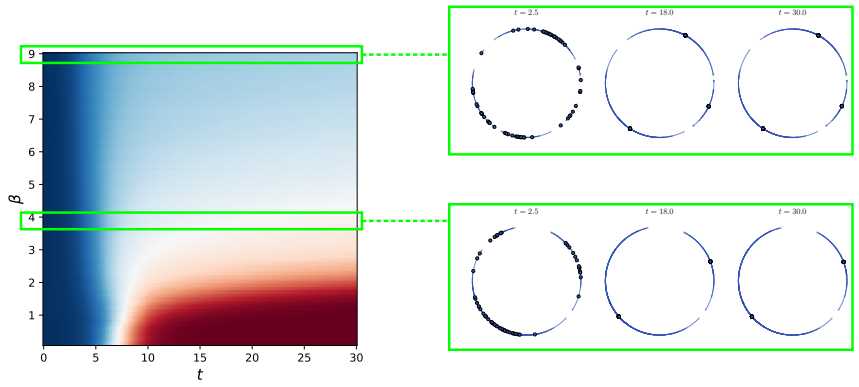

Figure 3. We zoom in on the phase diagram for the dynamics on the circle: $d = 2$. For $\beta = 4, 9$, we also display a trajectory of (2.6)–(SA) for a randomly drawn initial condition at times $t = 2.5, 18, 30$. We see that the particles settle at 2 clusters when $\beta = 4$ (bottom right) and 3 clusters when $\beta = 9$ (top right), for a duration of time. This reflects our metastability claim for large β in the low-dimensional case. The regime $\beta \ll 1$ (a single cluster emerges) is covered in Section 5.

fig: metastability

A similar metastable zone can also be seen when (Q, K, V) are random in Figure 4. 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]. At

github.com/borjanG/2023-revenge-of-the-fallen

the reader can find additional figures illustrating that our conclusions appear to hold in even more generality.

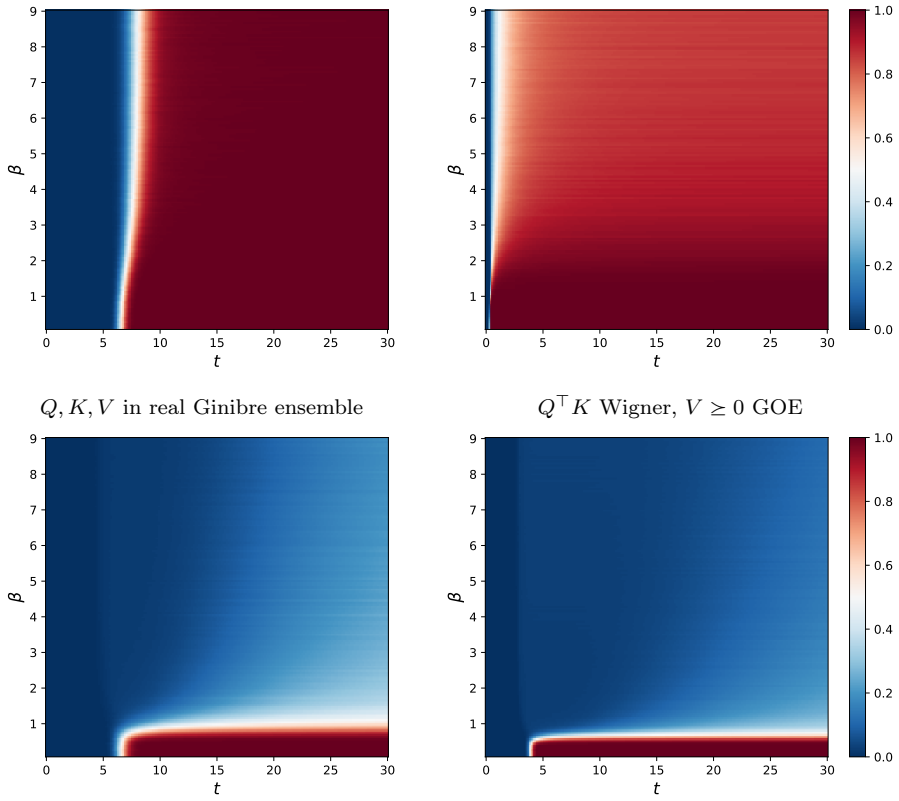

Figure 4. Phase diagrams for some choices of random matrices (Q, K, V); here $d = 128$, $n = 32$. Sharp phase transitions as well as metastable regions appear in all cases. In the "gradient flow" case $Q^\top K = V$, there is a remarkable resemblance to the well-understood case $Q^\top K = V = I_d$, which we are not able to explain for the moment.

fig: random.QKV

5. A single cluster for small β

sec: temperature

Our first attempt to remove the assumption $d \gg 1$ consists in looking at extreme choices of β. The case $\beta = +\infty$ is of little interest since all particles are fixed by the evolution. Therefore, we first focus on the case $\beta = 0$, before moving to the case $\beta \ll 1$ by a perturbation argument.

5.1. **The case $\beta = 0$.** For $\beta = 0$, both (2.6) and (USA) read as

{e:Snonres0}

$$\dot{x}_i(t) = \mathbf{P}_{x_i(t)}\left(\frac{1}{n}\sum_{j=1}^{n} x_j(t)\right), \qquad t \geqslant 0. \tag{5.1}$$

The following result shows that generically over the initial points, a single cluster emerges. Since the proof is rather lengthy, we refer the interested reader to Section 6.2.

p:beta0

Theorem 5.1. *Fix $d, n \geqslant 2$. 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* (5.1) *satisfies*

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

for all $i \in [n]$.

5.2. **The case $\beta \ll 1$.** Theorem 5.1 has some implications for small but positive β, something which is already seen in Figure 2 and Figure 3. This is essentially due to the fact that, formally,

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

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 5.2. *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 solutions to* (2.6)–(SA) *(or* (USA)*) converge to one cluster as $t \to +\infty$. Then*

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

Proof. We focus on the dynamics (2.6)–(SA), but the proof is in fact identical in the case of (USA).

For $\alpha > 0$, 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 $i, j \in [n]$, there holds $\langle z_i, z_j\rangle > \alpha$. Observe that if $\{z_1, \dots, z_n\}$ is α–clustered for some $\alpha \geqslant 0$, then the solution to the Cauchy problem for (2.6)–(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 4.2.

Now, for any integer $n \geqslant 1$, we denote by $\mathcal{S}_0^n \subset \mathcal{S}_0$ 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 (5.1) is $\frac{3}{4}$–clustered at time $t = n$, namely

{eq: harry.potter}

$$\langle x_i^0(n), x_j^0(n)\rangle > \frac{3}{4} \tag{5.2}$$

holds for all $i, j \in [n]$. We see that $\mathcal{S}_0^n$ is an open set for any integer $n \geqslant 1$. Moreover, $\mathcal{S}_0^n \subset \mathcal{S}_0^{n+1}$ according to the proof of Lemma 4.2, and $\bigcup_{n=1}^{+\infty} \mathcal{S}_0^n = \mathcal{S}_0$. This implies that

{e:Ps0n}

$$\lim_{n\to+\infty} \mathbb{P}(\mathcal{S}_0^n) = 1. \tag{5.3}$$

We now show that the solution to (2.6)–(SA) is near that of (5.1), starting from the same initial condition, when β is small. Using the Duhamel formula, we find

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

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

{e:approxsphere}

$$\left\| x_i^\beta(t) - x_i^0(t) \right\| \leqslant O(\beta) e^{3t} \tag{5.4}$$

for all $t \geqslant 0$, $\beta \geqslant 0$ and $i \in [n]$. Due to (5.4), there exists some $\beta_n > 0$ such that for any $\beta \in [0, \beta_n]$,

{eq: youareawizardharry}

$$\left\| x_i^\beta(n) - x_i^0(n) \right\| \leqslant \frac{1}{8}. \tag{5.5}$$

For this to hold, we clearly need $\beta_n \to 0$ as $n \to +\infty$. Combining (5.2) and (5.5), we gather that for any initial condition in $\mathcal{S}_0^n$, the solution $(x_i^\beta(\cdot))_{i\in[n]}$ to the corresponding Cauchy problem for (2.6)–(SA) is $\frac{1}{2}$–clustered at time $t = n$, namely satisfies

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

for all $i, j \in [n]$ and $\beta \in [0, \beta_n]$. Thus $\mathcal{S}_0^n \subset \mathcal{S}_\beta$ for any $\beta \in [0, \beta_n]$ by virtue of Lemma 4.2, which together with (5.3) concludes the proof. □

One can naturally ask

o:sbeta

Problem 4. *Does $\mathbb{P}(\mathcal{S}_\beta) = 1$ hold for any $\beta \geqslant 0$?*

6. Proofs

app: proof.1

6.1. **Proof of Theorem 4.7.** We focus on the dynamics (2.6)–(SA), since the proof for (USA) follows from very similar computations.

Step 1. The flow map is Lipschitz. We begin by showing that spherical transformer 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 (2.6) 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

$$\begin{aligned} x_i(t) - y_i(t) &= x_i(0) - y_i(0) \\ &\quad + \int_0^t \sum_{j=1}^n \left(\frac{e^{\beta \langle x_i(s), x_j(s) \rangle}}{\sum_{k=1}^n e^{\beta \langle x_i(s), x_k(s) \rangle}} \right) \big(x_j(s) - \langle x_i(s), x_j(s) \rangle x_i(s) \big) \, \mathrm{d}s \\ &\quad - \int_0^t \sum_{j=1}^n \left(\frac{e^{\beta \langle y_i(s), y_j(s) \rangle}}{\sum_{k=1}^n e^{\beta \langle y_i(s), y_k(s) \rangle}} \right) \big(y_j(s) - \langle y_i(s), y_j(s) \rangle y_i(s) \big) \, \mathrm{d}s. \end{aligned} \tag{6.1}$$

{eq: lip.1}

We see that

$$\left\| \int_0^t \sum_{j=1}^n \left(\frac{e^{\beta \langle x_i(s), x_j(s) \rangle}}{\sum_{k=1}^n e^{\beta \langle x_i(s), x_k(s) \rangle}} \right) (x_j(s) - y_j(s)) \, \mathrm{d}s \right\| \leqslant \int_0^t \max_{j \in [n]} \|x_j(s) - y_j(s)\| \, \mathrm{d}s. \tag{6.2}$$

{eq: lip.2}

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

$$\begin{aligned} &\left\| \int_0^t \sum_{j=1}^n \left(\frac{e^{\beta \langle x_i(s), x_j(s) \rangle}}{\sum_{k=1}^n e^{\beta \langle x_i(s), x_k(s) \rangle}} - \frac{e^{\beta \langle y_i(s), y_j(s) \rangle}}{\sum_{k=1}^n e^{\beta \langle y_i(s), y_k(s) \rangle}} \right) y_j(s) \, \mathrm{d}s \right\| \\ &\qquad \leqslant \beta n^{\frac{1}{2}} \int_0^t \left(\sum_{j=1}^n \Big[\langle x_i(s), x_j(s) \rangle - \langle y_i(s), y_j(s) \rangle \Big]^2 \right)^{\frac{1}{2}} \mathrm{d}s \\ &\qquad \leqslant 2\beta n \int_0^t \max_{j \in [n]} \|x_j(s) - y_j(s)\| \, \mathrm{d}s. \end{aligned} \tag{6.3}$$

{eq: lip.3}

Using (6.2), (6.3) and arguing similarly for the remaining terms in (6.1), 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

$$\max_{j \in [n]} \|x_j(t) - y_j(t)\| \leqslant C(\beta)^{nt} \max_{j \in [n]} \|x_j(0) - y_j(0)\|, \tag{6.4}$$

{eq: stability.4ortho}

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 $i \in [n]$,

$$\|x_i(0) - y_i(0)\| \leqslant \sqrt{\frac{\log d}{d}}. \tag{6.5}$$

{eq: almost.ortho.vec}

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 $\mathbb{S}^{d-1}$,

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

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

$$\mathscr{A}_0 = \{(6.5) \text{ is satisfied for any } i \in [n]\}$$

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

$$\mathscr{A} = \mathscr{A}_0 \cap \{\text{Theorem 4.1 is satisfied}\}$$

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 (4.4). 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 (2.6) corresponding to the initial datum $(y_i(0))_{i\in[n]}$. A combination of (6.4) and (6.5) yields

{e:shortdist}

$$\|x_i(t) - y_i(t)\| \leqslant C(\beta)^{nt}\sqrt{\frac{\log d}{d}} \tag{6.6}$$

for any $i \in [n]$ and $t \geqslant 0$, under $\mathscr{A}$. Combining (6.6) with Theorem 4.6 we obtain

{e:ineqfirstpart}

$$\left|\langle x_i(t), x_j(t)\rangle - \gamma_\beta(t)\right| \leqslant 2C(\beta)^{nt}\sqrt{\frac{\log d}{d}} \tag{6.7}$$

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

We turn to the proof of the second part of (4.4). 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

{e:ybetacloseto1}

$$1 - \gamma_\beta(t) \leqslant \frac{1}{2}\exp\left(\frac{n^2 e^\beta}{2\left(n + e^{\frac{\beta}{2}}\right)} - \frac{nt}{n + e^{\frac{\beta}{2}}}\right) \tag{6.8}$$

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 (6.8). We now set $d^*(n, \beta) \geqslant n$ such that

{eq: d.large}

$$\frac{d}{\log d} \geqslant \frac{16C(\beta)^2}{\gamma_\beta(\frac{1}{n})^2} \tag{6.9}$$

holds for any $d \geqslant d^*(n, \beta)$. According to Lemma 4.2, 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

{e:productcloseto1}

$$1 - \alpha(t) \leqslant \exp\left(\frac{1 - \gamma_\beta\left(\frac{1}{n}\right)t}{2ne^{2\beta}}\right). \tag{6.10}$$

To this end, let us first prove that

{e:1/n} (6.11) $$\alpha\left(\frac{1}{n}\right) \geqslant \frac{1}{2}\gamma_\beta\left(\frac{1}{n}\right).$$

We saw in the proof of Lemma 4.2 that x^* lies in the convex cone generated by the points $x_1(t), \dots, x_n(t)$ for any $t > 0$. Thus, there exists some $\eta \in (0, 1]$ such that ηx^* is a convex combination of the points $x_1(t), \dots, x_n(t)$, which implies that

{e:decompox*} (6.12) $$x^* = \sum_{k=1}^n \theta_k(t) x_k(t), \qquad \text{for some} \qquad \sum_{k=1}^n \theta_k(t) \geqslant 1, \quad \theta_k(t) \geqslant 0 \quad \forall k \in [n].$$

Taking the inner product of $x_i(\frac{1}{n})$ with the decomposition (6.12) at time $t = \frac{1}{n}$, we get

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

where the second inequality comes from (6.6) evaluated at time $t = \frac{1}{n}$, and the last inequality comes from (6.9). This is precisely (6.11). 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 4.2, we now find

{e:dotalpha} (6.13) $$\dot\alpha(t) = \langle \dot x_{i(t)}(t), x^*\rangle \geqslant \sum_{j=1}^n a_{i(t)j}(t)(1 - \langle x_{i(t)}(t), x_j(t)\rangle)\alpha(t)$$

for one of the indices $i(t) \in [n]$ achieving the minimum in the definition of $\alpha(t)$. Combining this with (6.11), we gather that $\alpha(t) \geqslant \alpha(\frac{1}{n})$ for $t \geqslant \frac{1}{n}$. But

{e:mineqalpha} (6.14) $$\min_{j\in[n]} \langle x_{i(t)}(t), x_j(t)\rangle \leqslant \sum_{k=1}^n \theta_k(t)\langle x_{i(t)}(t), x_k(t)\rangle = \langle x_{i(t)}(t), x^*\rangle = \alpha(t).$$

Plugging (6.14) into (6.13) and using $a_{ij}(t) \geqslant n^{-1}e^{-2\beta}$ we get

{e:diffineqalpha} (6.15) $$\dot\alpha(t) \geqslant \frac{1}{ne^{2\beta}}\alpha\left(\frac{1}{n}\right)(1 - \alpha(t))$$

for $t \geqslant \frac{1}{n}$. Integrating (6.15) from $\frac{1}{n}$ to t, we get (6.10). We therefore deduce from (6.10) 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 (6.8), we then get

{e:ineqsecondpart} (6.16)

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

Finally, combining (6.7) and (6.16) we obtain (4.4). □

rem: usa.d **Remark 6.1.** *An analogous statement to Theorem 4.7 holds for* (USA)*, where γ_β would rather be the unique solution to* (4.3)*. More concretely, Step 1 in the proof is only slightly changed–the constant one obtains in the analogue of* (4.4) *is rather*

$C(\beta)^{nt}$ with $C(\beta) = e^{10\beta e^{2\beta}}$. Step 2 remains unchanged. In Step 3, (6.8) *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.

app: proof.2

6.2. **Proof of Theorem 5.1.** We split the proof in four steps.

Step 1. A gradient flow. We first notice that the evolution (5.1) is a (continuous-time) gradient ascent for the energy $\mathsf{E}_0 : (\mathbb{S}^{d-1})^n \to \mathbb{R}$ defined as

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

Let

$$m(t) := \frac{1}{n} \sum_{j=1}^n x_j(t).$$

In the sequel, we assume that $m(0) \neq 0$ and that $x_1(0), \dots, x_n(0)$ are pairwise distinct, since this is true for Lebesgue almost any initial configuration. We notice that

$$\|m(t)\|^2 = \frac{1}{n} \mathsf{E}_0(x_1(t), \dots, x_n(t))$$

and therefore the map $t \mapsto \|m(t)\|^2$ is non-decreasing. Indeed, one can check that

{eq: mtfi} (6.17) $$\frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|m(t)\|^2 = \frac{1}{n} \sum_{i=1}^n f_i(t),$$

where

$$f_i(t) := \|m(t)\|^2 - \langle m(t), x_i(t) \rangle^2 \geqslant 0.$$

From (6.17) and the fact that $\|m(t)\| \leqslant 1$ for any $t \geqslant 0$, we also gather that

{eq: f_i.L1} (6.18) $$\int_0^{+\infty} f_i(t) \, \mathrm{d}t < +\infty$$

for all $i \in [n]$. Furthermore,

{eq: f_i.W1infty} (6.19) $$\left|\dot{f}_i(t)\right| = 2 \left|\langle \dot{m}(t), m(t) \rangle - \langle m(t), x_i(t) \rangle \left(\langle \dot{m}(t), x_i(t) \rangle + \langle m(t), \dot{x}_i(t) \rangle\right)\right| \leqslant C$$

for some $C > 0$ independent of $t > 0$ by virtue of the uniform boundedness of all of the involved quantities. Using (6.18) and (6.19) along with Lemma 4.4 we deduce that

$$\lim_{t \to +\infty} f_i(t) = 0 \qquad \text{for any } i \in [n].$$

Consequently,

$$\|m(t)\|^2 - \langle m(t), x_i(t) \rangle^2 \xrightarrow[t \to +\infty]{} 0 \qquad \text{for any } i \in [n].$$

Since $|\langle m(t), x_i(t) \rangle| \leqslant \|m(t)\|$, with equality only when $m(t)$ and $x_i(t)$ are aligned (using the fact that $\|m(t)\|$ is bounded from below), we deduce that $x_i(t)$ gets progressively aligned with $m(t)$ as $t \to +\infty$ for any $i \in [n]$. In other words, for any $i \in [n]$ there exists $\varepsilon_i \in \{\pm 1\}$ such that

{e:convepsi} (6.20) $$x_i(t) - \varepsilon_i \frac{m(t)}{\|m(t)\|} \xrightarrow[t \to +\infty]{} 0.$$

Step 2. All but one particles point in the same direction. We now show that $\varepsilon_i = -1$ for at most one index $i \in [n]$. To this end we argue by contradiction, and consider $i, j \in [n]$ with $i \neq j$ such that $\varepsilon_i = \varepsilon_j = -1$. Using the fact that

$$\dot{x}_i(t) = m(t) - \alpha_i(t) x_i(t)$$

where $\alpha_i(t) := \langle m(t), x_i(t) \rangle$, we get

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

where we used that $x_i(t), x_j(t) \in \mathbb{S}^{d-1}$. Consequently,

{eq: grow.apart}

$$\frac{\mathrm{d}}{\mathrm{d}t}\|x_i(t) - x_j(t)\|^2 = -\left(\alpha_i(t) + \alpha_j(t)\right)\|x_i(t) - x_j(t)\|^2. \tag{6.21}$$

Since both $x_i(t)$ and $x_j(t)$ approach $-\frac{m(t)}{\|m(t)\|}$ as t goes to infinity, we gather that $\alpha_i(t)$ and $\alpha_j(t)$ are both bounded above by a fixed negative number for sufficiently large t. Using this in (6.21) we deduce that $x_i(t)$ and $x_j(t)$ must move apart when t grows large (recalling that $x_i(t)$ and $x_j(t)$ are distinct at any time due to Cauchy uniqueness and the fact that $x_i(0) \neq x_j(0)$). This is a contradiction with (6.20) whence $\varepsilon_i = -1$ for at most one $i \in [n]$. (We will show in Step 5 that generically, $\varepsilon_i = 1$ for all $i \in [n]$.)

Step 3. The mean converges. From now on, we assume that $n \geqslant 3$ (for $n = 2$, the system is solved easily). Let us begin by considering the case in which $\varepsilon_i = 1$ for any $i \in [n]$. Define the convex cone generated by the particles:

$$\mathscr{K}(t) := \left\{ \sum_{i=1}^{n} \alpha_i x_i(t) \colon \alpha_i \geqslant 0 \text{ for all } i \in [n] \right\}.$$

The cone $\mathscr{K}(t)$ is polyhedral for any $t \geqslant 0$. We will show that $t \mapsto \mathscr{K}(t)$ is decreasing for large enough times t, unless $\mathscr{K}(t)$ is a point. If $x_i(t)$ lies on the boundary of $\mathscr{K}(t)$, then

{e:commentlesxibougent}

$$\dot{x}_i(t) = m(t) - \langle m(t), x_i(t) \rangle x_i(t) = \frac{1}{n}\sum_{j=1}^{n} \left(x_j(t) - \langle x_j(t), x_i(t) \rangle x_i(t)\right). \tag{6.22}$$

Assume that $\mathscr{K}(t)$ is not reduced to a point. If the $x_i(t)$ are close to each other, each of the n vectors in the right-hand side of (6.22) points inward $\mathscr{K}(t)$, therefore $\dot{x}_i(t)$ points inward too. As a consequence, $t \mapsto \mathscr{K}(t)$ is decreasing for large enough t. The only possible limit of $\mathscr{K}(t)$ is therefore a single point, which implies that $x_1(t), \dots, x_n(t)$ converge to the same limit as $t \to +\infty$.

Now assume that there exists a single $i \in [n]$ such that $\varepsilon_i = -1$. Without loss of generality, assume $i = n$. Consider the convex cone generated by the first $n - 1$ particles

$$\mathscr{K}_1(t) = \left\{ \sum_{i=1}^{n-1} \alpha_i x_i(t) \colon \alpha_i \geqslant 0 \text{ for all } i \in [n-1] \right\}.$$

Let us prove that $m(t) \in \mathscr{K}_1(t)$ for t large enough. We fix $t_0 > 0$ large enough so that $x_1(t), \dots, x_{n-1}(t)$ and $-x_n(t)$ are close to $\frac{m(t)}{\|m(t)\|}$ for any $t \geqslant t_0$, which is possible according to (6.20). Fix $t \geqslant t_0$. For any $v \in \mathbb{S}^{d-1}$ for which $\langle v, x_i(t) \rangle > 0$

for any $i \in [n-1]$, we have that $\langle v, x_n(t)\rangle \leqslant 0$ according to Lemma 4.2 and the fact that $\varepsilon_n = -1$. This implies that $-x_n(t) \in \mathscr{K}_1(t)$. We now write

$$-x_n(t) = \sum_{j=1}^{n-1} \alpha_j(t) x_j(t)$$

with $\alpha_j(t) \geqslant 0$. We notice that since $1 \geqslant \langle -x_n(t), x_j(t)\rangle$ for any $j \in [n-1]$, we have $\alpha_j(t) \leqslant 1$. Consequently,

$$m(t) = \frac{1}{n}\sum_{j=1}^{n-1} (1-\alpha_j(t)) x_j(t) \in \mathscr{K}_1(t).$$

The inequality $\alpha_j(t) \leqslant 1$ is strict if $-x_n(t) \neq x_j(t)$. As a consequence, unless $\mathscr{K}_1(t)$ is a point for t large enough, we gather using the first equality in (6.22) that $t \mapsto \mathscr{K}_1(t)$ is decreasing for large enough t. The only possible limit of $\mathscr{K}_1(t)$ is a point, which shows that $x_1(t), \dots, x_{n-1}(t)$ converge to the same limit x^* as $t \to +\infty$, and that $\frac{m(t)}{\|m(t)\|}$ also converges to x^*. As a consequence, $x_n(t)$ converges to $-x^*$ as $t \to +\infty$.

Step 4. Antipodal configurations are strict saddle points. We call an *antipodal configuration* any tuple $(x_1, \dots, x_n) \in (\mathbb{S}^{d-1})^n$ for which there exists $\overline{x} \in \mathbb{S}^{d-1}$ such that for any $i \in [n]$, there exists $\varepsilon_i \in \{\pm 1\}$ such that $x_i = \varepsilon_i \overline{x}$, and $\varepsilon_i = -1$ for *exactly one index* i. We show that any antipodal configuration is a strict saddle point, i.e., a critical point of E_0 at which the Hessian has at least one strictly *positive* eigenvalue[10]. Up to relabelling of the points $x_1, \dots, x_n$, any antipodal configuration is of the form $X^0 = (\overline{x}, \dots, \overline{x}, -\overline{x})$ for some $\overline{x} \in \mathbb{S}^{d-1}$. We consider a smooth function $[0,1] \ni \epsilon \mapsto x(\epsilon) \in \mathbb{S}^{d-1}$ satisfying $x(0) = \overline{x}$ and

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}_{|\epsilon=0} x(\epsilon) = \gamma \in \mathrm{T}_{\overline{x}}\mathbb{S}^{d-1}.$$

We then define $X^\epsilon = (x(\epsilon), \dots, x(\epsilon), -\overline{x})$. We easily check that

$$\begin{aligned}\mathsf{E}_0(X^\epsilon) &= \mathsf{E}_0(X^0) + \frac{2(n-1)}{n}(1 - \cos(x(\epsilon) - \overline{x})) \\ &= \mathsf{E}_0(X^0) + \frac{(n-1)}{n}\epsilon^2 \|\gamma\|^2 + O(\epsilon^2),\end{aligned}$$

whence $\frac{\mathrm{d}^2}{\mathrm{d}\epsilon^2}\mathsf{E}_0$ evaluated at X^0 has at least one positive eigenvalue, and consequently this antipodal configuration is a strict saddle point.

To conclude we invoke existing results ([LSJR16, PP17]): it is known that the set of initial conditions for which gradient ascent (in our case given in Step 1) converges to points where the Hessian has a positive eigenvalue has zero Lebesgue measure. The result in [LSJR16, PP17] is stated for functions defined on $\mathbb{R}^n$, and in the context of discrete-time evolution with small positive step-size, but the proof can be adapted to the continuous-time setting, and for functions defined on $(\mathbb{S}^{d-1})^n$.[11]

[10] We look at positive eigenvalues because the dynamics is a (continuous-time) gradient *ascent*, and not gradient descent for which one would look at *negative* eigenvalues of the Hessian.

[11] The proof of [PP17, Theorem 3] needs the following modification: on the compact set $(\mathbb{S}^{d-1})^n$, since the derivatives of any order of our energy functional are bounded, the inverse image of each local stable-center disk $W^{\mathrm{sc}}_{\mathrm{loc}}(\mathbf{r}_m)$ through the diffeomorphism induced by the continuous-time gradient descent at time t, that is, $\bigcup_{t\geqslant 0} g^{-t}(W^{\mathrm{sc}}_{\mathrm{loc}}(\mathbf{r}_m))$, has zero Lebesgue measure due to

Together with Step 4, this proves that for Lebesgue-almost any initial configuration, the particles converge to 1 cluster as $t \to +\infty$. □

sec: beyond

Part 3. Further

We discuss several avenues of research which would lead to a finer understanding of the clustering phenomenon and generalizations of our results, and which we believe are of independent mathematical interest.

sec: circle

7. Dynamics on the circle

We discuss the dynamics (2.6) 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 completely characterized by the angle $\theta_i(t) \in \mathbb{T}$ for which $x_i(t) = \cos(\theta_i(t))e_1 + \sin(\theta_i(t))e_2$ 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 this equality with respect to time and plugging into (USA) we obtain

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

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

eq:onangles

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

The case $\beta = 0$ is exactly the Kuramoto model recalled in Section 7.2. Assume for a moment 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 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

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

---

the strict saddle-point assumption. Using a finite covering of the set of strict saddle points by small balls (in particular included in local charts of $(\mathbb{S}^{d-1})^n$), we deduce the result.

When the particles $x_i(t)$ follow (2.6), one readily checks that the same continuity equation is satisfied but rather with the field

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

s:kuramoto

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

{e:kuramoto}

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

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] 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) becomes a gradient flow

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

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

{e:energykuramoto}

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

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 is then the corresponding gradient flow. In [ABK+22] for instance, the authors prove 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

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

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.

Problem 5. *With the exception of the global maximum, are all critical points of E_β strict saddle points?*

Extensions of the Kuramoto model of the form

{e:variantkuramoto}

$$\dot{\theta}_i(t) = \omega_i + \frac{K}{n}\sum_{j=1}^{n} h(\theta_j(t) - \theta_i(t)), \tag{7.4}$$

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) if β is not close to 0. We nevertheless have some hope that handling the dynamics (7.1) is easier than dealing with (7.4) for a general h: for instance, the Fourier coefficients of the function $h_\beta(\theta) = e^{\beta\cos(\theta)}$ in question are completely explicit, and given by the modified Bessel function of the first kind, whose properties have been extensively studied.

sec: bbgky

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

{e:transport}

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

where

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

and

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

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

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 6. *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 $d > 2$, and for the original spherical Transformer, are also problems which we believe merit further investigation.

9. General matrices

Figure 4 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 a couple of particular cases (without excluding other approaches).

sec:repulsive

9.1. The repulsive case.

As seen from Lemma 3.5, 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 3.2). 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

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

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 $\mathscr{C} \subset \mathbb{S}^{d-1}$, $\#\mathscr{C} = n$:

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

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 defer the reader to [CK07] and the illustrative experimental paper [BBC+09] for further detail. 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 $\mathbf{P}_x$, leading to

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

for all $i \in [n]$ and $t \in \mathbb{R}_{\geqslant 0}$. In fact, in [GLPR23], 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* [GLPR23]. 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]

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

which are solutions to

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

for $i \in [n]$ and $t \in \mathbb{R}_{\geqslant 0}$, where

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

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 [GLPR23] 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.

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

[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.

V	K and Q	Limit geometry	Result in [GLPR23]
$V = I_d$	$Q^\top K > 0$	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:results

Table 1. Summary of the clustering results of [GLPR23]. *All results except for the case $V = -I_d$ hold for the time-scaled dynamics (9.3).

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 [GLPR23]. We nonetheless leave the reader with the following general question:

prob: Rd

Problem 7. *Is it possible to extend the clustering results of Table 1 to other cases of (Q, K, V)? What are the resulting limit shapes?*

9.2.2. *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

{e:exp}

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

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

{e:maineq}

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

where

{e:Ci(t)}

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

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

e:selection

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 8. *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* [GLPR23] *and Problem 7 in the setting of* (9.5)–(9.6)*?*

We believe that (9.5)–(9.6) is also an original model for opinion formation. 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 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 9. *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.

[15]We thank Enrique Zuazua for this suggestion.

10. Approximation and control

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. For discrete-time Transformers, universal approximation has been shown to hold in [YBR+19], making use of a variant of the architecture with translate parameters and letting the number of layers go to infinity. See also [ADTK23], and [JLLW23] for a review. 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, GZ22]. We are however not aware of control-theoretical results in which arbitrarily many input measures ought to be mapped to as many output measures, as would be the case for Transformers.

Acknowledgments

We thank Sébastien Bubeck, Matthew Rosenzweig, Kimi Sun, and Rui Sun for discussions.

References

abdalla2022expander [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.

acebron2005kuramoto [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.

alberti2023sumformer [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.

ambrosio2005gradient [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.

agrachev2020control [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.

barron1993universal [Bar93] Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

ballinger2009experimental [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.

blanchet2008infinite [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.

bilyk2019geodesic [BD19] Dmitriy Bilyk and Feng Dai. Geodesic distance Riesz energy on the sphere. *Transactions of the American mathematical Society*, 372(5):3141–3166, 2019.

ba2016layer [BKH16] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

boissard2014mean [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.

bertozzi2011lp [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.

boffi2023deep [BVE23] Nicholas M Boffi and Eric Vanden-Eijnden. Deep learning probability flows and entropy production rates in active matter. *arXiv preprint arXiv:2309.12991*, 2023.

chizat2018global [CB18] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems*, 31, 2018.

carrillo2014contractivity [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.

chaintron2022propagation [CD22] Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: a review of models, methods and applications. II. Applications. 2022.

carrillo2011global [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.

chiba2015proof [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.

carrillo2021consensus [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.

cohn2007universally [CK07] Henry Cohn and Abhinav Kumar. Universally optimal distribution of points on spheres. *Journal of the American Mathematical Society*, 20(1):99–148, 2007.

cohn2022universal [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.

cheng2023interpolation [CLLS23] Jingpu Cheng, Qianxiao Li, Ting Lin, and Zuowei Shen. Interpolation, approximation and controllability of deep neural networks. *arXiv preprint arXiv:2309.06015*, 2023.

caponigro2015nonlinear [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.

cuchiero2020deep [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.

chen2018neural [CRBD18] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018.

cucker2007emergent [CS07] Felipe Cucker and Steve Smale. Emergent behavior in flocks. *IEEE Transactions on Automatic Control*, 52(5):852–862, 2007.

cybenko1989approximation [Cyb89] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

daido1992order [Dai92] Hiroaki Daido. Order function and macroscopic mutual entrainment in uniformly coupled limit-cycle oscillators. *Progress of Theoretical Physics*, 88(6):1213–1218, 1992.

dietert2018landau [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.

dutta2021redesigning [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.

delsarte1991spherical [DGS91] Philippe Delsarte, Jean-Marie Goethals, and Johan Jacob Seidel. Spherical codes and designs. In *Geometry and Combinatorics*, pages 68–93. Elsevier, 1991.

dobrushin1979vlasov [Dob79] Roland L'vovich Dobrushin. Vlasov equations. *Funktsional'nyi Analiz i ego Prilozheniya*, 13(2):48–58, 1979.

dai2013approximation [DX13] Feng Dai and Yuan Xu. *Approximation theory and harmonic analysis on spheres and balls*. Springer, 2013.

weinan2017proposal [E17] Weinan E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 1(5):1–11, 2017.

fernandez2016landau [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.

fornasier2021consensus [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.

GooBenCou16 [GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016.

guillin2021uniform [GBM21] Arnaud Guillin, Pierre Le Bris, and Pierre Monmarché. Uniform in time propagation of chaos for the 2d vortex model and other singular stochastic systems. *arXiv preprint arXiv:2108.08675*, 2021.

geshkovski2023emergence [GLPR23] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *arXiv preprint arXiv:2305.05465*, 2023.

golse2016dynamics [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.

gallagher2013newton [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.

geshkovski2022turnpike [GZ22] Borjan Geshkovski and Enrique Zuazua. Turnpike in optimal control of pdes, resnets, and beyond. *Acta Numerica*, 31:135–263, 2022.

han2023class [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.

rainer2002opinion [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.

ha2016collective [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.

haber2017stable [HR17] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34(1), 2017.

ha2020asymptotic [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.

he2016identity [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.

jordan1998variational [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.

jiang2023brief [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.

jabin2014clustering [JM14] Pierre-Emmanuel Jabin and Sebastien Motsch. Clustering and asymptotic behavior in opinion formation. *Journal of Differential Equations*, 257(11):4165–4187, 2014.

kohn2002upper [KO02] Robert V Kohn and Felix Otto. Upper bounds on coarsening rates. *Communications in Mathematical Physics*, 229(3):375–395, 2002.

krause2000discrete [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.

krizhevsky2012imagenet [KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

kuramoto1975self [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.

lacker2023hierarchies [Lac23] Daniel Lacker. Hierarchies, entropy, and quantitative propagation of chaos for mean field diffusions. *Probability and Mathematical Physics*, 4(2):377–432, 2023.

lanalbert [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.

li2018maximum [LCT18] Qianxiao Li, Long Chen, and Cheng Tai. Maximum principle based algorithms for deep learning. *Journal of Machine Learning Research*, 18:1–29, 2018.

lin2018resnet [LJ18] Hongzhou Lin and Stefanie Jegelka. Resnet with one-neuron hidden layers is a universal approximator. *Advances in neural information processing systems*, 31, 2018.

lu2019understanding [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.

li2022deep [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.

lee2016gradient [LSJR16] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257. PMLR, 2016.

lin2022survey [LWLQ22] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, 2022.

ling2019landscape [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.

mei2018mean [MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

motsch2014heterophilious [MT14] Sebastien Motsch and Eitan Tadmor. Heterophilious dynamics enhances consensus. *SIAM Review*, 56(4):577–621, 2014.

otto2007slow [OR07] Felix Otto and Maria G Reznikoff. Slow motion of gradient flows. *Journal of Differential Equations*, 237(2):372–420, 2007.

otto2001geometry [Ott01] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.

panageas2017gradient [PP17] Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

pinnau2017consensus [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.

ruiz2023neural [RBZ23] Domenec Ruiz-Balet and Enrique Zuazua. Neural ode control for classification, approximation, and transport. *SIAM Review*, 65(3):735–773, 2023.

rosenzweig2023global [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.

rotskoff2022trainability [RVE22] Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022.

sander2022sinkformers [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.

serfaty2020mean [Ser20] Sylvia Serfaty. Mean field limit for Coulomb-type flows. *Duke Mathematical Journal*, 169(15), 2020.

strogatz2000kuramoto [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.

szeg1939orthogonal [Sze39] Gabor Szegö. *Orthogonal polynomials*, volume 23. American Mathematical Soc., 1939.

tadmor2023swarming [Tad23] Eitan Tadmor. Swarming: hydrodynamic alignment with pressure. *Bulletin of the American Mathematical Society*, 60(3):285–325, 2023.

shuotan2017 [Tan17] Yan Shuo Tan. Energy optimization for distributions on the sphere and improvement to the Welch bounds. *Electronic Communications in Probability*, 22(none):1 – 12, 2017.

tabuada2020universal [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.

townsend2020dense [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.

vicsek1995novel [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.

villani2001limite [Vil01] Cédric Villani. Limite de champ moyen. *Cours de DEA*, 2002:49, 2001.

villani2009optimal [Vil09] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

veeravalli2023nonlinear [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.

vaswani2017attention [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.

weed2019sharp [WB19] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

wendel1962problem [Wen62] James G Wendel. A problem in geometric probability. *Mathematica Scandinavica*, 11(1):109–111, 1962.

xiao2023introduction [XZ23] Tong Xiao and Jingbo Zhu. Introduction to transformers: an nlp perspective. *arXiv preprint arXiv:2311.17633*, 2023.

yun2019transformers [YBR+19] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019.

zhang2020approximation [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.

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