Measure-to-measure interpolation using Transformers

Borjan Geshkovski, Philippe Rigollet, Domènec Ruiz-Balet

Measure-to-measure interpolation using Transformers

Borjan Geshkovski
Inria & Sorbonne Université

Philippe Rigollet
MIT

Domènec Ruiz-Balet
Université Paris Dauphine

November 8, 2024

Abstract

Transformers are deep neural network architectures that underpin the recent successes of large language models. Unlike more classical architectures that can be viewed as point-to-point maps, a Transformer acts as a measure-to-measure map implemented as specific interacting particle system on the unit sphere: the input is the empirical measure of tokens in a prompt and its evolution is governed by the continuity equation. In fact, Transformers are not limited to empirical measures and can in principle process any input measure. As the nature of data processed by Transformers is expanding rapidly, it is important to investigate their expressive power as maps from an arbitrary measure to another arbitrary measure. To that end, we provide an explicit choice of parameters that allows a single Transformer to match N arbitrary input measures to N arbitrary target measures, under the minimal assumption that every pair of input-target measures can be matched by some transport map.

Keywords. Transformers, optimal transport, mean-field, continuity equation, clustering, controllability, universal approximation.

AMS classification. 41A25, 68T07, 37C10.

Contents

1 Introduction

Transformers, introduced in 2017 with the groundbreaking paper [VSP+17], are the neural network architectures behind the recent successes of large language models. They owe their impressive results to the way they process data: inputs are length-n sequences of d-dimensional vectors called *tokens* (representing words, or patches of an image, for example), which are processed over several layers

of parametrized nonlinearities. Unlike conventional neural networks however, all tokens are coupled and mixed at every layer via the so-called *self-attention mechanism*.

To make this discussion transparent we take a leaf out of several recent works [SABP22, VBC20, GLPR23] which view Transformers as a flow maps on $\mathcal{P}(\mathbb{S}^{d-1})$—the space of probability measures over the unit sphere $\mathbb{S}^{d-1}$—realized by an interacting particle system: viewing each token as a particle, given an initial sequence of particles $(x_1(0), \dots, x_n(0)) \in (\mathbb{S}^{d-1})^n$, one considers

$$\dot{x}_i(t) = v[\mu(t)](t, x_i(t)) \qquad \text{for } t \in [0, T], \tag{1.1}$$

for all $i \in [n]$; here $\mu(t) = \frac{1}{n}\sum_{j=1}^n \delta_{x_j(t)}$ denotes the empirical measure. The vector field

$$v[\mu](t,x) = \mathsf{P}_x^{\perp}\Big(\boldsymbol{V}(t)\mathscr{A}_{\boldsymbol{B}}[\mu](t,x) + \boldsymbol{W}(t)\big(\boldsymbol{U}(t)x + b(t)\big)_+\Big) \tag{1.2}$$

depends on the empirical measure through self-attention

$$\mathscr{A}_{\boldsymbol{B}}[\mu](t,x) := \frac{\displaystyle\int e^{\langle \boldsymbol{B}(t)x, x'\rangle} x' \mu(\mathrm{d}x')}{\displaystyle\int e^{\langle \boldsymbol{B}(t)x, \zeta\rangle}\mu(\mathrm{d}\zeta)}. \tag{1.3}$$

The parameters $\boldsymbol{V}(t)$, $\boldsymbol{B}(t)$, $\boldsymbol{W}(t)$, $\boldsymbol{U}(t)$, which are all $d \times d$ matrices, and $b(t)$, which is a d-dimensional vector, are to be used to steer the flow to one's liking. The vector field $v[\mu(t)](t,\cdot)$ is a combination of the self-attention mechanism $\mathscr{A}_{\boldsymbol{B}}[\mu(t)](t,\cdot)$ and a *perceptron* at every layer t, ultimately projected onto $\mathsf{T}_x\mathbb{S}^{d-1}$ by virtue of the orthogonal projector $\mathsf{P}_x^{\perp} := I_d - xx^{\top}$, referred to as *layer normalization*. Practical implementations of Transformers are discrete-time versions, of course, and (1.1) originates from a Lie-Trotter splitting scheme—see **Appendix A** for details.

Since (1.1) only truly depends on the empirical measure, one can naturally turn to the *continuity equation* which governs its evolution. We can thus equivalently see the Transformer as the solution map of the Cauchy problem

$$\begin{cases} \partial_t \mu(t) + \operatorname{div}\Big(\mu(t)\, v[\mu(t)]\Big) = 0 & \text{on } [0,T]\times\mathbb{S}^{d-1} \\ \mu(0) = \mu_0 & \text{on } \mathbb{S}^{d-1}. \end{cases} \tag{1.4}$$

Here $-\operatorname{div}$ denotes the adjoint of the spherical gradient ∇. As the number n of particles can be large—orders of magnitude vary in different implementations, likely due to compute—in this paper we focus on (1.4), which makes sense for arbitrary measures, and encompasses (1.1) in the particular setting of empirical measures.

Transformers are used to solve learning tasks such as *next-token prediction*, wherein one seeks to map an ensemble of given input sequences of n tokens onto a corresponding ensemble of next tokens. In this case, the output measure encodes

the probability distribution of the next token. Motivated by further ubiquitous tasks including masked language prediction, sentiment analysis, and image classification, and taking an approximation/control theory perspective, in this paper we consider the canonical learning problem in which we are given data consisting of $N \gg 1$ pairs of input and output probability distributions

$$(\mu_0^i, \mu_1^i) \in \mathscr{P}(\mathbb{S}^{d-1}) \times \mathscr{P}(\mathbb{S}^{d-1}) \qquad \text{for } i \in [N], \tag{$\mathscr{D}$}$$

and we seek to match them through the solution map of (1.4). In the context of the applications evoked above, one always works with discrete measures, with the targets being a single point mass, but we consider a more general setting in what follows. This is an *ensemble transportation* or *controllability* problem, since we seek to accomplish this matching of measures by means of the flow of (1.4) for a *single* parameter or control $\theta = (\boldsymbol{W}(t), \boldsymbol{V}(t), \boldsymbol{B}(t), \boldsymbol{U}(t), b(t))_{t\in[0,T]}$.

In the discrete-time setting, and focusing solely on mapping sequences to sequences, the problem is first solved in [YBR+20] by using $\boldsymbol{B} = \beta\widetilde{\boldsymbol{B}}$ and $\beta = +\infty$ (a formal limit), as well as additional bias vectors within the inner products of the self-attention mechanism, but without employing layer normalization. Further work has focused on seeing whether one can do matching solely using self-attention, namely, without the perceptron component or layer normalization—results in this direction include [ADTK23, KZLD22]. See [CCP23, JL23, EGKZ22, JLLW23, WW24, PTB24, SP24] for further results.

In the continuous time and/or arbitrary measure setting, much less is known—we are aware of [AG24, AL24, FdHP24]. In [AL24], still in the context of empirical measures, the authors focus on self-attention dynamics only ($\boldsymbol{W} \equiv 0$) and prove that, *generically*, two vector fields in the class of permutation-equivariant vector fields suffice to match two ensembles of empirical measures with the same number of atoms. Their study is inspired by a flurry of works on matching one cloud of points to another using the flow of (1.2) with $\boldsymbol{V} \equiv 0$ (known as *neural ODEs*), where tools from geometric control theory can be useful [AS20, AS22, Sca23, EGBO22, TG22]. With the exception of [TG22, EGBO22], none of these papers actually state the specific vector fields that can be used, and none of them are constructive. On another hand, [AG24] address the setting of absolutely continuous measures, but use a slightly different vector field compared to (1.2). Finally, [FdHP24] address the discrete-time system and arbitrary measures, but use a slightly different model motivated by *in-context learning* [GTLV22] and approximate a map $\mathscr{P}(\Omega) \times \Omega \to \Omega$ over compact subsets $\Omega \subset \mathbb{R}^d$—the proof is based on a clever application of the Stone-Weierstrass theorem.

None of the above papers use layer normalization; moreover, the parameters used are not explicit due to the non-constructive strategy, and there are therefore no bounds on the number of switches. To address these pitfalls, we take inspiration from concurrent works on neural ODEs [LLS22, RBZ23, CLLS23] in which the parameters are fully explicit and piecewise constant by construction. Our goal is to focus on the most general case while constructing parameters that

leverage salient properties of all mechanisms involved in (1.2)—the prime example being the dynamic *emergence of clusters* proven in [GLPR24, GLPR23], which has been empirically observed and referred to as *token uniformity*, *oversmoothing* [CZC+22, RZZD23, GWDW23, WAWJ24, WAW+24, DBK24, SWJS24], or *rank collapse* [DCL21, FZH+22, NAB+22, JDB23, ZMZ+23, ZLL+23, NLL+24, BHK24, CNQG24] in the literature. In fact, we solely use the long-time behavior of (1.4) with explicit, well-chosen parameters throughout, and as such, our strategy also leads to a deeper understanding of the inner workings of all mechanisms in (1.2).

1.1 Main results

Set

$$\Theta := (\mathscr{M}_{d\times d}(\mathbb{R}))^4 \times \mathbb{R}^d.$$

We recall that for any $T > 0$ and $\theta = (\boldsymbol{V}, \boldsymbol{B}, \boldsymbol{W}, \boldsymbol{U}, b) \in L^\infty((0,T);\Theta)$, the Cauchy problem (1.4) is well-posed, in the sense that for every $\mu_0 \in \mathscr{P}(\mathbb{S}^{d-1})$ there exists a unique weak solution $\mu \in \mathscr{C}^0([0,T];\mathscr{P}(\mathbb{S}^{d-1}))$. This in turn yields a continuous[1] and invertible flow (or solution) map

$$\Phi_\theta^t : \mathscr{P}(\mathbb{S}^{d-1}) \to \mathscr{P}(\mathbb{S}^{d-1}),$$

for $t \in [0,T]$, with

$$\Phi_\theta^t(\mu_0) = \mu(t),$$

which we often use later on to simplify the presentation. These results follow from classical arguments using the Lipschitz properties of the underlying vector field—see [GLPR24, §6], [PT22] for details.

Henceforth, for simplicity, assume[2] that $\mu_0^i \not\equiv \mu_0^j$ and $\mu_1^i \not\equiv \mu_1^j$ for $i \neq j$.

When practically training a transformer, the target measures are simply a point mass on the next (or hidden) token. It turns out that this setup leads to a simpler construction so we start with this first theorem.

Theorem 1.1. *Suppose $d \geqslant 3$. Consider data ($\mathcal{D}$) such that*

1. *There exists $w_0 \in \mathbb{S}^{d-1}$ such that*

$$w_0 \notin \bigcup_{i\in[N]} \operatorname{supp}(\mu_0^i). \tag{1.5}$$

2. *For any $i \in [N]$, we have $\mu_1^i = \delta_{x^i}$.*

[1] with respect to the weak convergence on $\mathscr{P}(\mathbb{S}^{d-1})$, which is metrized by the W_2 distance (1.9).

[2] The assumption $\mu_1^i \not\equiv \mu_1^j$ for $i \neq j$ (as well as (1.5), and more generally (1.6)) can be removed at the cost of additional technicalities—see **Appendix B**. $\mu_0^i \not\equiv \mu_0^j$ for $i \neq j$ cannot be removed since (1.4) is well posed.

Then for any $T > 0$ and $\varepsilon > 0$, there exists $\theta \in L^\infty((0,T);\Theta)$ such that for any $i \in [N]$, the unique solution $\mu^i \in \mathscr{C}^0([0,T];\mathscr{P}(\mathbb{S}^{d-1}))$ to (1.4) with data μ_0^i and parameters θ satisfies

$$\mathsf{W}_2\left(\mu^i(T), \mu_1^i\right) \leqslant \varepsilon.$$

Moreover, θ can be chosen piecewise constant, with $O(d \cdot N)$ switches, and

$$\|\theta\|_{L^\infty((0,T);\Theta)} = O\left(\frac{d \cdot N}{T} + \log\frac{1}{\varepsilon}\right).$$

The fact that the parameters θ can be chosen to be piecewise-constant-in-time leads to a direct link with the discrete-time network used in practice: the number of switches provides a lower bound on the number of layers. Our estimates are in all likelihood sub-optimal (principally due to our inability to simultaneously use both components of the vector field in (1.2), as seen in **Section 1.2**) and we believe that there is great margin for improvement. The reader is referred to **Section 1.4.3** and **Section 6** for further comments on this particular aspect.

Theorem 1.1 follows as a corollary of the proof of the following general result.

Theorem 1.2. *Suppose $d \geqslant 3$. Consider data ($\mathscr{D}$) such that*

1. *There exist $w_0, w_1 \in \mathbb{S}^{d-1}$ such that*

$$w_0 \notin \bigcup_{i\in[N]} \operatorname{supp}(\mu_0^i) \quad \text{and} \quad w_1 \notin \bigcup_{i\in[N]} \operatorname{supp}(\mu_1^i). \tag{1.6}$$

2. *For any $i \in [N]$, there exists $\mathsf{T}^i \in L^2(\mathbb{S}^{d-1};\mathbb{S}^{d-1})$ such that $\mathsf{T}^i_\# \mu_0^i = \mu_1^i$.*

Then for any $T > 0$ and $\varepsilon > 0$, there exists $\theta \in L^\infty((0,T);\Theta)$ such that for any $i \in [N]$, the unique solution $\mu^i \in \mathscr{C}^0([0,T];\mathscr{P}(\mathbb{S}^{d-1}))$ to (1.4) with data μ_0^i and parameters θ satisfies

$$\mathsf{W}_2\left(\mu^i(T), \mu_1^i\right) \leqslant \varepsilon.$$

Moreover, θ can be chosen piecewise constant.

Here $\mathsf{T}_\#\mu(A) = \mu(\mathsf{T}^{-1}(A))$ for $A \subset \mathbb{S}^{d-1}$ is the image measure. The number of switches of the control θ can be estimated by using structural properties of the measures—we postpone a discussion thereon to **Section 1.4.3**.

1.2 Overview of the proof

We sketch the proof of Theorem 1.2. The solution map $\Phi^T_{\text{fin}} : \mathscr{P}(\mathbb{S}^{d-1}) \to \mathscr{P}(\mathbb{S}^{d-1})$ is constructed as[3]

$$\Phi^T_{\text{fin}} := \left(\Phi^{\frac{T}{3}}_{\theta_3}\right)^{-1} \circ \Phi^{\frac{2T}{3}}_{\theta_2} \circ \Phi^{\frac{T}{3}}_{\theta_1},$$

where

[3] The philosophy is reminiscent to the proof of the Chow-Rashevskii theorem using iterated Lie brackets for the controllability of driftless systems [Cor07, §3.3].

1. $\Phi^t_{\theta_1} : \mathcal{P}(\mathbb{S}^{d-1}) \to \mathcal{P}(\mathbb{S}^{d-1})$ is the solution map of (1.4) on $[0, \frac{T}{3}]$, generated by piecewise constant parameters θ_1 , having $O(d \cdot N)$ switches, as to disentangle the supports of the input measures (the use of the attention component is *necessary* for this step). After this step, the supports of the measures are disjoint:

$$\operatorname{supp}\left(\Phi^{\frac{T}{3}}_{\theta_1}(\mu^i_0)\right) \cap \operatorname{supp}\left(\Phi^{\frac{T}{3}}_{\theta_1}(\mu^j_0)\right) = \varnothing \quad \text{whenever } i \neq j. \tag{1.7}$$

This is done in Proposition 3.1 in **Section 3**. The clue lies in following the insights of [GLPR23], which entail clustering of every individual measure to a single point mass in long time in the special regime $\boldsymbol{B} = \beta I_d$ with $\beta \geqslant 0$ and $\boldsymbol{V} = I_d$. Should the limit point masses corresponding to every input measure be located at different positions, the disentanglement property (1.7) would readily follow by taking the time horizon T large enough. Unfortunately, characterizing the location of the limit point mass for general measures is an open problem. We instead consider a curated choice of $\boldsymbol{V}$ to facilitate locating the limiting cluster for every measure, which we now sketch. Consider $N = 2$ (the general case is argued by induction; see Lemma 3.3) and suppose that $\mathbb{E}_{\mu^1_0}[z]$ and $\mathbb{E}_{\mu^2_0}[z]$ are not colinear (this assumption is not needed, as seen in Lemma 3.4). We can take $\boldsymbol{B} \equiv 0$ (we provide an alternative proof when $\boldsymbol{B} \not\equiv 0$ in **Appendix D**) and

$$\boldsymbol{V}(t) := \sum_{k=1}^{d-1} \alpha_k \alpha_k^\top 1_{[T_k, T_{k+1}]}(t),$$

where $\{\alpha_k\}$ is an orthonormal basis of $(\operatorname{span} \mathbb{E}_{\mu^1_0}[z])^\perp$. Then there is some index ℓ such that $\langle \mathbb{E}_{\mu^1_0}[z], \alpha_\ell \rangle = 0$ and $\langle \mathbb{E}_{\mu^2_0}[z], \alpha_\ell \rangle \neq 0$. Consequently the quantity $t \mapsto \langle \mathbb{E}_{\mu^i(t)}[z], \alpha_\ell \rangle$ remains constant when $i = 1$, and does not change sign when $i = 2$. After an elementary computation one can then see that any $x(t) \in \operatorname{supp}(\mu^2(t))$ converges to $\pm\alpha_\ell$ in long time, whereas $\mu^1(t) = \mu^1_0$ throughout. One can always rescale time so that the above holds at an arbitrary prescribed horizon, at the cost of increasing the norm of the parameters.

2. In the same vein, $\Phi^t_{\theta_3} : \mathcal{P}(\mathbb{S}^{d-1}) \to \mathcal{P}(\mathbb{S}^{d-1})$ is the solution map of (1.4) on $[\frac{2T}{3}, T]$, generated by piecewise constant parameters θ_3, as to disentangle the supports of the target measures:

$$\operatorname{supp}\left(\Phi^{\frac{T}{3}}_{\theta_3}(\mu^i_1)\right) \cap \operatorname{supp}\left(\Phi^{\frac{T}{3}}_{\theta_3}(\mu^j_1)\right) = \varnothing \quad \text{whenever } i \neq j.$$

Inverting $\Phi^t_{\theta_3}$ simply corresponds to running time backwards from T to $\frac{2T}{3}$.

3. $\Phi^t_{\theta_2} : \mathcal{P}(\mathbb{S}^{d-1}) \to \mathcal{P}(\mathbb{S}^{d-1})$ is the solution map of (1.4) on $[\frac{T}{3}, \frac{2T}{3}]$, generated by piecewise constant parameters θ_2, alternating between $\boldsymbol{V} \equiv 0$ (namely,

using solely the perceptron component) and $\boldsymbol{W} \equiv 0$, $\boldsymbol{V} \equiv I_d$, which approximately matches the ensembles of disentangled input and target measures:

$$\mathsf{W}_2\left(\left(\Phi_{\theta_2}^{\frac{2T}{3}} \circ \Phi_{\theta_1}^{\frac{T}{3}}\right)(\mu_0^i), \Phi_{\theta_3}^{\frac{T}{3}}(\mu_1^i)\right) \leqslant \varepsilon$$

for all $i \in [N]$. This map can be constructed in three different ways depending on the nature of the target measures. If the target measures are point masses (Theorem 1.1), one simply clusters the disentangled input measures to point masses using Proposition 2.1 in **Section 2** ($\boldsymbol{W} \equiv 0$, $\boldsymbol{V} \equiv I_d$) up to time $\frac{T}{2}$ say, and then matches the resulting point masses to the targets using Proposition 4.1 in **Section 4** ($\boldsymbol{V} \equiv 0$) at time $\frac{2T}{3}$. This idea is then generalized to targets that are empirical measures with $M \geqslant 2$ atoms in **Section 5.1** (see the restricted case). The case of general, non-atomic target measures is significantly more involved. The construction is done in Lemma 5.4 in **Section 5** and the main idea is as follows. It can readily be seen (see Lemma 5.1) that the transport maps T^i are propagated by the flow maps constructed in the two previous steps, in the sense that there exists some integrable map $\Psi : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ with $\Psi\big|_{\mathrm{supp}(\Phi_{\theta_1}^{\frac{T}{3}}(\mu_0^i))} = \Psi^i$ and

$$\Psi^i_{\#}\left(\Phi_{\theta_1}^{\frac{T}{3}}(\mu_0^i)\right) = \Phi_{\theta_3}^{\frac{T}{3}}(\mu_1^i).$$

Since we construct $\Phi_{\theta_2}^t$ without using the nonlinear part of (1.4), we can identify $\Phi_{\theta_2}^t$ with a Lipschitz-continuous and invertible map from $\mathbb{S}^{d-1}$ to $\mathbb{S}^{d-1}$, which we also denote $\Phi_{\theta_2}^t$. Using standard arguments from optimal transport (Lemma 5.2), we find

$$\mathsf{W}_2\left(\left(\Phi_{\theta_2}^{\frac{2T}{3}}\right)_{\#}\left(\Phi_{\theta_1}^{\frac{T}{3}}(\mu_0^i)\right), \Phi_{\theta_3}^{\frac{T}{3}}(\mu_1^i)\right) \lesssim \left\|\Phi_{\theta_2}^{\frac{2T}{3}} - \Psi\right\|_{L^2(\mu)},$$

where

$$\mu = \sum_{i=1}^N \Phi_{\theta_1}^{\frac{T}{3}}(\mu_0^i).$$

The final result therefore boils down to approximating maps in $L^2(\mathbb{S}^{d-1}, \mu)$. This is technically involved due to the fact that μ can have both diffuse and atomic parts—both elements are treated using the clustering and matching constructions presented in **Section 2** and **Section 4** respectively.

Matching general ensembles of measures *cannot* be done with a single *linear* continuity equation, as is done in the Benamou-Brenier reformulation of optimal transport for instance [BB00], namely (1.4) in which the vector field v does not depend on $\mu(t)$. Indeed, take for instance $\mu_0^1, \mu_0^2 \in \mathcal{P}_{\mathrm{ac}}(\mathbb{S}^{d-1})$ such that

$$\mathrm{supp}(\mu_0^1) \cap \mathrm{supp}(\mu_0^2) \neq \varnothing,$$

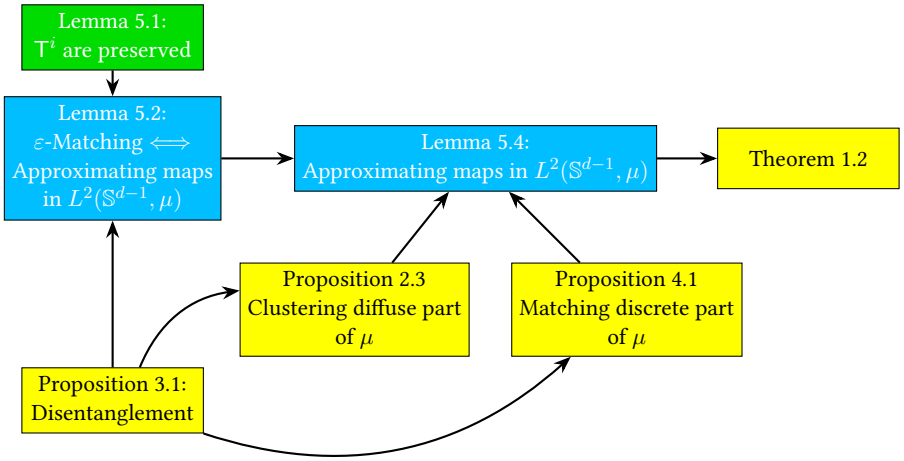

Figure 1: High-level overview of the proof of Theorem 1.2.

and similarly $\mu_1^1, \mu_1^2 \in \mathscr{P}_{\mathrm{ac}}(\mathbb{S}^{d-1})$ such that

$$\operatorname{supp}(\mu_1^1) \cap \operatorname{supp}(\mu_1^2) = \varnothing. \tag{1.8}$$

Then there cannot exist a single-valued $\mathsf{T} : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ such that $\mathsf{T}_{\#}\mu_0^1 = \mu_1^1$ and $\mathsf{T}_{\#}\mu_0^2 = \mu_1^2$, since there would have to exist $x \in \operatorname{supp}(\mu_0^1) \cap \operatorname{supp}(\mu_0^2)$ for which $\mathsf{T}(x)$ would have to take two different values due to (1.8). This elementary counterexample is the starting point of our strategy, as the self-attention mechanism $\mathscr{A}_{\boldsymbol{B}}[\mu]$ provides a nonlinear dependence[4] of the solution map to (1.4) with respect to μ, which we use precisely to disentangle overlapping measures. In this regard, Theorem 1.2 is an ensemble controllability result for a *nonlinear* continuity equation, thus extending existing results on the controllability of the *linear* continuity equation—see [Bro08, KL09, AC09, AL09, Rag24, CGP16, DMR19].

1.3 Outline

The remainder of the paper is organized as follows. We comment on assumptions and extensions of Theorem 1.2 in **Section 1.4**. In **Section 2**, we provide explicit parameters that yield long-time clustering (i.e., convergence to discrete measures). **Section 3** presents how initial measures with overlapping support can be disentangled over time using clustering. **Section 4** addresses the matching problem of clouds of points, which is used after clustering and disentanglement. The proofs of Theorem 1.2 and Theorem 1.1 can be found in **Section 5**. We discuss some interesting questions regarding the number of switches needed for disentanglement in **Section 6**.

[4] One can draw parallels with the failure of the Kalman rank condition [Son13, Cor07] for the ensemble controllability of linear systems in finite dimensions.

1.4 Discussion

1.4.1 On our assumptions

- The requirement $d \geqslant 3$ in Theorem 1.2 stems from matching disentangled measures. In $d = 2$, the problem becomes intrinsically one-dimensional, and in such a setting, the order of particles is preserved. This obstruction impedes the conclusion of the disentanglement. To carry through our strategy, one needs to use self-attention to order the input measures and disentangle them.

- When the targets are more general than point masses—as in Theorem 1.2—we operate under the assumption that there exists a transport map between every pair (μ_0^i, μ_1^i). This is again satisfied for many cases of interest; for instance, whenever μ_0^i is absolutely continuous with respect to the Lebesgue measure on $\mathbb{S}^{d-1}$ (per the celebrated theorem(s) of Brenier-McCann [Bre91, McC01]), or whenever μ_0^i and μ_1^i are empirical measures with μ_0^i having as many atoms as μ_1^i. This assumption is also minimal. Indeed, since $v[\mu(t)](t, \cdot)$ is Lipschitz, any solution to (1.4) can be written as the image measure of the initial data by means of some map. This presents a natural impediment to matching a measure constituted by a single atom to a measure with two atoms.

- If the input μ_0^i and output μ_1^i are empirical measures having n and m atoms respectively, with $n > m$ and $\frac{n}{m} \notin \mathbb{N}$, then there does not exist a transport map between μ_0^i and μ_1^i. Consequently, for the same reason as in the previous point, it is not possible to find a solution to (1.4) that approximates μ_1^i to any desired level of accuracy. However, if to each atom of μ_1^i one assigns $\lfloor n/m \rfloor$ atoms of μ_0^i, then one can construct a map T^i such that $\mathsf{W}_2(\mathsf{T}^i_{\#}\mu_0^i, \mu_1^i) = O(m/n)$. As a result, we could use the flow map of (1.4) to approximate all N target measures to error $O(\varepsilon + m/n)$.

1.4.2 On exact matching

One can raise the natural question if it is possible to have *exact* matching. i.e. $\varepsilon = 0$, in Theorem 1.2. We provide some comments:

- We can exactly match N empirical input measures to N empirical target measures as long as they have the same number of atoms. This follows as a corollary of the proof of Theorem 1.2, since no quantization is required in Lemma 5.4.

- Since $v[\mu(t)](t, \cdot)$ is Lipschitz, we cannot do exact transportation of an absolutely continuous measure to a discrete one even when $N = 1$. Similarly, we cannot match a single input measure with connected support to a target measure whose support has multiple connected components.

Remark 1.3 (Beyond W_2)**.** *We opted for approximation in the Wasserstein distance because working with distances is convenient. The result could be adapted to encompass the Kullback-Leibler divergence* (KL)*, which is the natural candidate in view*

of applications. (Note that this would be a stronger approximation result by virtue of [BV05], which only requires the second argument in the KL *to have a Gaussian moment, guaranteed in our case by working on* $\mathbb{S}^{d-1}$*.) To achieve this, after the step involving the disentanglement of supports, we can match the disentangled measures approximately in* TV *instead of* W_2 *by following a similar approach to the one developed in [RBZ24], and then apply a reverse Pinsker inequality [Ver14, Theorem 7] (see also [SV16]). However, to do so, one needs that the measures are mutually absolutely continuous and to have a bounded likelihood ratio. This approach would avoid quantizing the (disentangled) target measure and thus clustering the (disentangled) inputs into atoms.*

1.4.3 On the number of parameter switches

Our proof roughly yields

$$\#\text{switches} = \#\text{switches}_{\text{disentanglement}} + \#\text{switches}_{\text{clustering}} + \#\text{switches}_{\text{matching}}$$

for piecewise constant parameters. Our current best estimate, if all measures have pairwise overlapping support, is

$$\#\text{switches}_{\text{disentanglement}} = O(d \cdot N).$$

(See **Section 6** for an extended discussion thereon.) While all the parameters involved in the construction of the final map Φ_{fin}^T yielding Theorem 1.2 are explicit, the precise estimate of #switches in the most general case is difficult to determine, due to the clustering step. We discuss three concrete cases.

1. In Theorem 1.1, once the input measures are rendered disentangled, we use a single constant parameter that results in clustering over time (as per [GLPR23]), collapsing each input measure to a point mass. Thus,

   $$\#\text{switches}_{\text{clustering}} = 0.$$

   The resulting ensemble of point masses can then be matched to the targets using the perceptron component (adapting ideas from [RBZ23, LLS22]) with

   $$\#\text{switches}_{\text{matching}} = O(N).$$

   So the switches in Theorem 1.1 arise primarily from the disentangling step. In particular, if the source measures have disjoint support, then the number of switches becomes $O(N)$ which is dimension independent.

2. Consider Theorem 1.2, with targets that are empirical measures having $m \geqslant 2$ atoms, and inputs that are all absolutely continuous. As seen in Section 5.1 (restricted case), the proof can be significantly simplified in this case. Indeed, one can avoid a direct application of Lemma 5.4 and Proposition 2.3 by instead

iteratively applying Lemma C.3 to each input measure. Namely, after the disentanglement step (with a sufficiently large separation), we partition the support of each disentangled input measure into m pieces, which we can cluster to a point with a single constant parameter per piece using Lemma C.3. All in all,

$$\#\text{switches}_{\text{clustering}} = O(m \cdot N).$$

The resulting clustered measures can then be matched to the empirical target measures by Proposition 4.1 at the cost of

$$\#\text{switches}_{\text{matching}} = O(m \cdot N).$$

All in all,

$$\#\text{switches} = O\big((m+d)N\big).$$

3. Suppose both the inputs and targets are empirical measures, with n and m atoms, respectively. When it comes to $\#\text{switches}_{\text{clustering}}$, if $n \gg m$ or m is a divisor of n, one can use m balls per measure in Proposition 2.3 instead of packing, which would lead to

$$\#\text{switches}_{\text{clustering}} = O(m \cdot N).$$

To achieve this, one can combine Proposition 3.1 with the clustering of measures to a point mass as in Proposition 2.1, much like what is done in the Section 5.1 (restricted case). By virtue of the latter, we can partition the support of each measure using large balls that are not necessarily contained within the support. All in all,

$$\#\text{switches} = O\big((m+d)N\big).$$

To put things into context: in the discrete-time setting of [YBR+20], while the number of layers is seemingly independent of the number of sequences N, it is exponential in the dimension d.

In the case of general target measures as in Theorem 1.2, $\#\text{switches}_{\text{clustering}}$ is exponential in the dimension d due to the use of a specific strategy which leverages packing numbers—see Remark 2.4.

1.4.4 On generalities

We comment on even greater generality in the choice of the Transformer architecture, which typically varies slightly from implementation to implementation. (See **Appendix A** for further details.)

- **Increasing the width.** Many of the actions used throughout our proofs are performed using the perceptron component with constant parameter matrices $\boldsymbol{W}$ and $\boldsymbol{U}$ that are of rank 1. One could, of course, consider using rectangular matrices instead, thus increasing the *width* of the network, which could, in turn, reduce the number of switches ($\approx$ depth).

- **Multi-head attention.** The computation of token similarities through self-attention is typically parallelized in practice throughout several *heads*, lending way to *multi-head self-attention*. This boils down to replacing the term

$$\boldsymbol{V}(t)\mathscr{A}_{\boldsymbol{B}}[\mu(t)](t,x)$$

in (1.2) by

$$\sum_{h=1}^{H} \boldsymbol{V}_h(t)\mathscr{A}_{\boldsymbol{B}_h}[\mu(t)](t,x).$$

We now dispose of $H \geqslant 1$ parameters $(\boldsymbol{V}_h(t), \boldsymbol{B}_h(t))_{h\in[H]}$ at every time t. We do not know how to exploit multiple heads in our proofs, although some theoretical insights thereon can be found in the proofs in [CL24].

- **Discrete time.** The continuous-time formulation gives rise to an equation that is *reversible in time*, which we often use in our construction. In particular, the task of disentangling supports becomes equivalent to the task of entangling supports, which is not the case in the non-reversible scenario. Our results are expected to hold for an appropriate discretization of (1.4) with a sufficiently small time step.

- **Beyond the ReLU perceptron.** All of our results remain unchanged if one replaces $(\cdot)_+$ (the ReLU) by any other Lipschitz nonlinearity that equals the ReLU near the origin. It is likely that the proofs can be generalized to even encompass the hyperbolic tangent. The relevant property that the nonlinear activation function ought to satisfy is to ensure that the resulting flow (when $\boldsymbol{V} \equiv 0$) leaves any spherical cap of choice invariant.

1.5 Notation and basic definitions

Unless stated otherwise, all integrals are taken over $\mathbb{S}^{d-1}$, and all ∇ denote the spherical gradient. We use $[n] := \{1,\dots,n\}$, and $f(x) \lesssim g(x)$ if there exists a finite positive constant C such that $f(x) \leqslant Cg(x)$. We write $f(x) \lesssim_S g(x)$ if the resulting constant depends on S. We denote by $d_g(x,y)$ the geodesic distance between $x, y \in \mathbb{S}^{d-1}$, which, as a reminder, is the great-circle distance $d_g(x,y) = \arccos(\langle x, y\rangle)$. For $A \subset \mathbb{S}^{d-1}$, we denote by $\operatorname{conv} A$ the convex hull of A in $\mathbb{R}^d$, and by $\operatorname{conv}_g A$ the geodesic convex hull of A in $\mathbb{S}^{d-1}$ (the smallest geodesically convex set containing A). Unless otherwise specified, all open balls are considered as subsets of $\mathbb{S}^{d-1}$ and are taken with respect to d_g. Recall the Wasserstein-p distance for $p \geqslant 1$:

$$\mathsf{W}_p^p(\mu,\nu) := \inf_{\pi\in\mathcal{C}(\mu,\nu)} \int d_g(x,y)^p \pi(\mathrm{d}x, \mathrm{d}y), \tag{1.9}$$

where $\mathcal{C}(\mu,\nu)$ denotes all couplings between μ and ν (see [Vil09] for details). We set $\mathsf{W}_\infty(\mu,\nu) := \lim_{p\to+\infty} \mathsf{W}_p(\mu,\nu)$. A geodesic is a smooth curve with zero

acceleration, namely, a smooth $\gamma : I \to \mathbb{S}^{d-1}$ with $\mathbf{P}^{\perp}_{\gamma(t)}(\ddot{\gamma}(t)) = 0$ for $t \in I$, where I is an open interval. Geodesics on $\mathbb{S}^{d-1}$ lie on great circles, specifically, there exists $v \in \mathbb{S}^{d-1}$ such that

$$\gamma(t) \subset \{x \in \mathbb{S}^{d-1} : \langle x, v\rangle = 0\}\,, \quad \forall\, t \in I\,.$$

Given two points $x, y \in \mathbb{S}^{d-1}$, a geodesic $\gamma : [a, b] \to \mathbb{S}^{d-1}$ is called a minimal or minimizing geodesic between x and y if γ is of speed 1 (i.e., $\|\dot{\gamma}(t)\| = 1$), satisfies $\gamma(a) = x$ and $\gamma(b) = y$, and $d_g(x, y) = b - a$. If x and y are not antipodal, the minimizing geodesic between them is unique and is simply the segment between them on the great circle on which they both lie. The geodesic open ball of radius $R > 0$ centered at $x \in \mathbb{S}^{d-1}$ is defined as

$$B(x, R) := \left\{y \in \mathbb{S}^{d-1} : d_g(x, y) < R\right\}. \tag{1.10}$$

Unless stated otherwise, by open (resp. closed) ball we understand a *geodesic* open (resp. closed) ball.

Acknowledgments

B.G. acknowledges financial support from the French government managed by the National Agency for Research under the France 2030 program, with the reference "ANR-23-PEIA-0004". P.R. was supported by NSF grants DMS-2022448, CCF-2106377, and a gift from Apple. D. Ruiz-Balet was funded by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/T024429/1.

2 Clustering of the input data

We begin by investigating how the input measures can be clustered using (1.4), in the sense that they are in the vicinity of discrete measures with few atoms.

In **Section 2.1**, we cover the special case of clustering to a single atom, while the case of general discrete measures is discussed in **Section 2.2**. The results of this section are used in **Section 3**, and they are also a key step in our final matching strategy.

2.1 Clustering to a single point mass

The following is essentially an adaptation of the so-called "cone collapse" argument presented in [GLPR23, Lemma 6.4].

Proposition 2.1. *Suppose $\mu_0^i \in \mathcal{P}(\mathbb{S}^{d-1})$ for $i \in [N]$, are such that*

$$\operatorname{supp}(\mu_0^i) \text{ is contained in an open hemisphere.}$$

Fix $\boldsymbol{B} \in \mathscr{M}_{d\times d}(\mathbb{R})$. Then, for any $i \in [N]$, there exists $z^i \in \operatorname{conv}_g \operatorname{supp}(\mu_0^i)$ such that the following holds. For any $\varepsilon > 0$, there exists $T > 0$ such that, the unique solution μ^i to

$$\begin{cases} \partial_t \mu^i(t) + \operatorname{div}\left(\mathbf{P}_x^{\perp} \mathscr{A}_{\boldsymbol{B}}[\mu^i(t)]\mu^i(t)\right) = 0 & \text{on } [0,T]\times \mathbb{S}^{d-1} \\ \mu^i(0) = \mu_0^i & \text{on } \mathbb{S}^{d-1} \end{cases} \tag{2.1}$$

satisfies

$$\mathsf{W}_\infty\left(\mu^i(T), \delta_{z^i}\right) \leqslant \varepsilon.$$

Moreover if

$$\operatorname{supp}(\mu_0^i) \cap \operatorname{supp}(\mu_0^j) = \varnothing$$

for all $i \neq j \in [N]$ then $z^i \neq z^j$ for all $i \neq j \in [N]$.

Proof of Proposition 2.1. It suffices to show the result for a single measure. We therefore drop indices to lighten the notation. Suppose[5]

$$\sigma_d(\operatorname{conv}_g \operatorname{supp}(\mu_0)) > 0.$$

Since $\boldsymbol{V} = I_d$, and $\operatorname{supp}(\mu_0)$ being contained in a half-sphere, for all $x \in \operatorname{supp}(\mu_0)$ we have

$$\|\mathscr{A}_{\boldsymbol{B}}[\mu_0](x)\| > 0, \tag{2.2}$$

and

$$\gamma(x) := \frac{\mathscr{A}_{\boldsymbol{B}}[\mu_0](x)}{\|\mathscr{A}_{\boldsymbol{B}}[\mu_0](x)\|} \in \operatorname{int}\left(\operatorname{conv}_g \operatorname{supp}(\mu_0)\right). \tag{2.3}$$

Now fix $\tau > 0$ and $x_0 \in \partial \operatorname{conv}_g \operatorname{supp}(\mu_0)$, and consider x solving

$$\begin{cases} \dot{x}(t) = \mathbf{P}^{\perp}_{x(t)}\left(\mathscr{A}_{\boldsymbol{B}}[\mu(t)](x(t))\right) & t \in [0,\tau] \\ x(0) = x_0. \end{cases} \tag{2.4}$$

We Taylor expand:

$$\langle x(\tau), \gamma(x_0)\rangle = \langle x_0, \gamma(x_0)\rangle + \tau \left\langle \mathbf{P}^{\perp}_{x_0}(\mathscr{A}_{\boldsymbol{B}}[\mu_0](x_0)), \gamma(x_0)\right\rangle + O(\tau^2).$$

Using $\mathscr{A}_{\boldsymbol{B}}[\mu_0](x_0) = \mathbf{P}^{\perp}_{x_0}(\mathscr{A}_{\boldsymbol{B}}[\mu_0](x_0)) + \langle \mathscr{A}_{\boldsymbol{B}}[\mu_0](x_0), x_0\rangle\, x_0$ as well as (2.2), we deduce

$$\tau \left\langle \mathbf{P}^{\perp}_{x_0}(\mathscr{A}_{\boldsymbol{B}}[\mu_0](x_0)), \gamma(x_0)\right\rangle = \tau \frac{\|\mathbf{P}^{\perp}_{x_0}(\mathscr{A}_{\boldsymbol{B}}[\mu_0](x_0))\|}{\|\mathscr{A}_{\boldsymbol{B}}[\mu_0](x_0)\|} > 0.$$

Therefore $\langle x(\tau), \gamma(x_0)\rangle > \langle x_0, \gamma(x_0)\rangle$ for τ small enough. In view of (2.3), we deduce

$$x(\tau) \in \operatorname{int}(\operatorname{conv}_g \operatorname{supp}(\mu_0)).$$

[5] σ_d henceforth denotes the normalized spherical Lebesgue measure, i.e. the uniform measure.

This entails that $t \mapsto \operatorname{conv}_g \operatorname{supp}(\mu(t))$ is strictly non-increasing (in the sense of set inclusion) for short times, i.e.

$$\operatorname{conv}_g \operatorname{supp}(\mu(\tau)) \subset \operatorname{conv}_g \operatorname{supp}(\mu_0) \tag{2.5}$$

for τ small enough. By the Lipschitz character of (2.4), we also gather that

$$\sigma_d(\operatorname{conv}_g \operatorname{supp}(\mu(t))) > 0$$

for all finite $t \geqslant 0$. Therefore we can bootstrap the argument for obtaining (2.5) to deduce that $t \mapsto \operatorname{conv}_g \operatorname{supp}(\mu(t))$ is strictly non-increasing over $\mathbb{R}_{\geqslant 0}$, and we also see that

$$\lim_{t \to +\infty} \sigma_d(\operatorname{conv}_g \operatorname{supp}(\mu(t))) = 0. \tag{2.6}$$

We use (2.6) and induction over the dimension d to conclude. Indeed note that in $d = 2$, (2.6) and the Portmanteau theorem yield

$$\lim_{t \to +\infty} \mathsf{W}_\infty(\mu(t), \delta_{x^*}) = 0 \tag{2.7}$$

for some $x^* \in \mathbb{S}^1$ depending only on μ_0. Now suppose $d = 3$. By (2.6) and the Portmanteau theorem, for every $\varepsilon_1 > 0$ there exist $T_{\varepsilon_1} > 0$ and $\mu^0_{\varepsilon_1} \in \mathscr{P}(\mathbb{S}^{d-1})$ such that $\operatorname{supp}(\mu^0_{\varepsilon_1}) \subset \mathbb{S}^1 \cap \{x \colon x_2 > 0\}$ and

$$\mathsf{W}_\infty\left(\mu(T_{\varepsilon_1}), \mu^0_{\varepsilon_1}\right) \leqslant \varepsilon_1. \tag{2.8}$$

By virtue of (2.7), for every $\varepsilon_2 > 0$ there exist $T_{\varepsilon_2} > 0$ and $x^*_{\varepsilon_1}$ (depending on ε_1 through the initial measure μ_{ε_1}) such that the solution μ_{ε_1} to

$$\begin{cases} \partial_t \mu_{\varepsilon_1} + \operatorname{div}\left(\mathbf{P}^\perp_x \mathscr{A}_{\boldsymbol{B}}[\mu_{\varepsilon_1}(t)]\mu_{\varepsilon_1}\right) = 0 & \text{on } [T_{\varepsilon_1}, T_{\varepsilon_1} + T_{\varepsilon_2}] \times \mathbb{S}^{d-1} \\ \mu_{\varepsilon_1}(T_{\varepsilon_1}) = \mu^0_{\varepsilon_1} & \text{on } \mathbb{S}^{d-1} \end{cases}$$

satisfies

$$\mathsf{W}_\infty\left(\mu_{\varepsilon_1}(T_{\varepsilon_2} + T_{\varepsilon_1}), \delta_{x^*_{\varepsilon_1}}\right) \leqslant \varepsilon_2. \tag{2.9}$$

Combining (2.8), (2.9) with the triangle inequality and Lemma C.1, we have

$$\begin{aligned} \mathsf{W}_\infty\left(\mu(T_{\varepsilon_2} + T_{\varepsilon_1}), \delta_{x^*_{\varepsilon_1}}\right) &\leqslant \mathsf{W}_\infty\left(\mu(T_{\varepsilon_2} + T_{\varepsilon_1}), \mu_{\varepsilon_1}(T_{\varepsilon_2} + T_{\varepsilon_1})\right) \\ &\quad + \mathsf{W}_\infty\left(\mu_{\varepsilon_1}(T_{\varepsilon_2} + T_{\varepsilon_1}), \delta_{x^*_{\varepsilon_1}}\right) \\ &\leqslant O\left(e^{e^{T_{\varepsilon_2}}}\right) \mathsf{W}_\infty\left(\mu(T_{\varepsilon_1}), \mu_{\varepsilon_1}\right) + \varepsilon_2 \\ &\leqslant O\left(e^{e^{T_{\varepsilon_2}}}\right) \varepsilon_1 + \varepsilon_2, \end{aligned}$$

where the implicit constants are independent of $\varepsilon_1, \varepsilon_2$. By compactness, there exists $x^* \in \mathbb{S}^1$ such that for any sequence $\varepsilon_{1,n} \to 0$ as $n \to +\infty$, there is a subsequence $\{\varepsilon_{1,n_k}\}_{k \in \mathbb{N}}$ such that $x_{\varepsilon_{1,n_k}} \to x^*$ as $k \to +\infty$. We relabel this

subsequence as $\varepsilon_{1,n}$ and the associated sequence of times by $\{T_{\varepsilon_{1,n}}\}_{n\in\mathbb{N}}$. Now note that

$$\mathsf{W}_\infty\left(\mu(T_{\varepsilon_2}+T_{\varepsilon_{1,n}}),\delta_{x^*}\right)\leqslant O\left(e^{e^{T_{\varepsilon_2}}}\right)\varepsilon_{1,n}+\varepsilon_2+d_g\left(x^*_{\varepsilon_{1,n}},x^*\right). \tag{2.10}$$

Moreover, there exists some $n(\varepsilon)\geqslant 1$ such that

$$d_g\left(x^*_{\varepsilon_{1,n}},x^*\right)\leqslant\frac{\varepsilon}{3}$$

for all $n\geqslant n(\varepsilon)$. We set $\varepsilon_2=\frac{\varepsilon}{3}$ and we choose $n\geqslant n(\varepsilon)$ large enough so that

$$O\left(e^{e^{T_{\varepsilon_2}}}\right)\varepsilon_{1,n}\leqslant\frac{\varepsilon}{3}.$$

Coming back to (2.10), we find

$$\mathsf{W}_\infty\left(\mu\left(T_{\varepsilon_{1,n}}+T_{\frac{\varepsilon}{3}}\right),\delta_{x^*}\right)\leqslant\varepsilon$$

for $n\geqslant n(\varepsilon)$. Therefore, we found a sequence of positive times $\{T_n\}_{n\in\mathbb{N}}$ with $T_n=T_{\varepsilon_{1,n}}+T_{\frac{\varepsilon}{3}}$ for which $\mu_n=\mu(T_n)$ converges to δ_{x^*} in W_∞. Thanks to (2.5), the convergence does not depend on the sequence of times. Indeed, for every $t\geqslant T_n$, we have

$$\begin{aligned}\mathsf{W}_\infty(\mu(t),\delta_{x^*})\leqslant\max_{y\in\mathrm{conv}_g\mathrm{supp}(\mu(t))}d_g(y,x^*)&\leqslant\max_{y\in\mathrm{conv}_g\mathrm{supp}(\mu_n)}d_g(y,x^*)\\&\leqslant\max_{y\in B(x^*,\mathsf{W}_\infty(\mu_n,\delta_{x^*}))}d_g(y,x^*)\\&=2\mathsf{W}_\infty(\mu_n,\delta_{x^*})\\&\leqslant 2\varepsilon\end{aligned}$$

where $B_n:=B(x^*,\mathsf{W}_\infty(\mu_n,\delta_{x^*}))$ is the ball on $\mathbb{S}^{d-1}$ centered at x^* with radius $\mathsf{W}_\infty(\mu_n,\delta_{x^*})$. As B_n is geodesically convex, and since $\mathrm{supp}(\mu_n)\subset B_n$, one has $\mathrm{conv}_g\mathrm{supp}(\mu_n)\subset B_n$.

The cases $d>3$ follow by repeating the above argument. □

We can in fact even show the following quantitative estimate.

Proposition 2.2. *Consider the setting of Proposition 2.1, for a fixed index i (which we remove), and $\boldsymbol{B}\equiv 0$. Let $x_0\in\mathrm{conv}_g\,\mathrm{supp}(\mu_0)$ be the limit of $\mu(t)$. Let $\varepsilon>0$ and set*

$$T_\varepsilon:=\inf\left\{t\geqslant 0\colon \mathsf{W}_2(\mu(t),\delta_{x_0})\leqslant\varepsilon\right\}.$$

Then

$$T_\varepsilon=O\left(\log\frac{1}{\varepsilon}\right).$$

The proof can be found in **Appendix C.2**.

2.2 Clustering to discrete measures

The following result ensures that an ensemble of measures with disjoint supports can be clustered, up to arbitrary precision, to finitely many atoms within their own support, all by means of the same flow map.

Proposition 2.3. *Suppose $\mu_0^i \in \mathcal{P}(\mathbb{S}^{d-1})$, $i \in [N]$, have no atoms, and satisfy*

$$\operatorname{conv}_g \operatorname{supp}(\mu_0^i) \cap \operatorname{conv}_g \operatorname{supp}(\mu_0^j) = \varnothing$$

for $i \neq j$. Fix $M \geqslant 1$, and for any $i \in [N]$ consider

$$\mu_1^i := \sum_{k=1}^{M} \alpha_k^i \delta_{x_k^i}$$

where $x_k^i \in \operatorname{conv}_g \operatorname{supp}(\mu_0^i)$, with $x_k^i = x_{k'}^j$ if and only if $(k,i) = (k',j)$, and where $\alpha_k^i \geqslant 0$ with $\sum_{k=1}^M \alpha_k^i = 1$. Then for any $T > 0$ and $\varepsilon > 0$ there exist piecewise constant $(\boldsymbol{W}, \boldsymbol{V}, b) : [0,T] \to \mathscr{M}_{d\times d}(\mathbb{R})^2 \times \mathbb{R}^d)$ such that for $i \in [N]$, the corresponding unique solution $\mu^i \in \mathscr{C}^0([0,T]; \mathcal{P}(\mathbb{S}^{d-1}))$ to (C.20) with data μ_0^i and these parameters satisfies

$$\operatorname{conv}_g \operatorname{supp}\left(\mu^i(T)\right) \cap \operatorname{conv}_g \operatorname{supp}\left(\mu^j(T)\right) = \varnothing$$

for $i \neq j$, and

$$\mathsf{W}_2\left(\mu^i(T), \mu_1^i\right) \leqslant \varepsilon.$$

The number of switches in $(\boldsymbol{W}, \boldsymbol{V}, b)$ can also be accounted for—see Remark 2.4.

Proof of Proposition 2.3. We split

$$[0,T] = \bigcup_{i\in[N]} [T_{i-1}, T_i],$$

where $0 = T_0 < T_1 < \ldots < T_N = T$ are to be determined later on. We look to apply Lemma C.2 separately within each interval, thus, dealing with one measure at a time. Namely, the parameters take the form

$$(\boldsymbol{W}, \boldsymbol{U}, b)(t) = \sum_{i=1}^{N} (\boldsymbol{W}_i, \boldsymbol{U}_i, b_i)(t) 1_{[T_{i-1}, T_i)}(t),$$

where $(\boldsymbol{W}_i, \boldsymbol{U}_i, b_i)$ are, roughly speaking, piecewise constant parameters stemming from a repeated application of Lemma C.2. We critically use (C.21) to ensure that when we act on the i-th measure in $[T_{i-1}, T_i]$, all the other measures remain invariant, so

$$\mu^i(T_{i-1}) = \mu_0^i. \tag{2.11}$$

Therefore, we take $i \in [N]$ to be arbitrary. We proceed in three steps.

Step 1. Partitioning each support into M pieces

Let $\mathscr{C}^i := \mathrm{supp}(\mu_0^i)$, and consider a partition $\{\mathscr{C}_k^i\}_{k\in[M]}$ of $\mathscr{C}^i$ consisting of pairwise disjoint sets with connected interiors and satisfying

$$\mu_0^i(\mathscr{C}_k^i) := \alpha_k^i.$$

Namely

$$\mathscr{C}^i = \bigcup_{k\in[M]} \mathscr{C}_k^i$$

with $\mathscr{C}_k^i \cap \mathscr{C}_{k'}^i = \varnothing$ if $k \neq k'$ (see Figure 2).

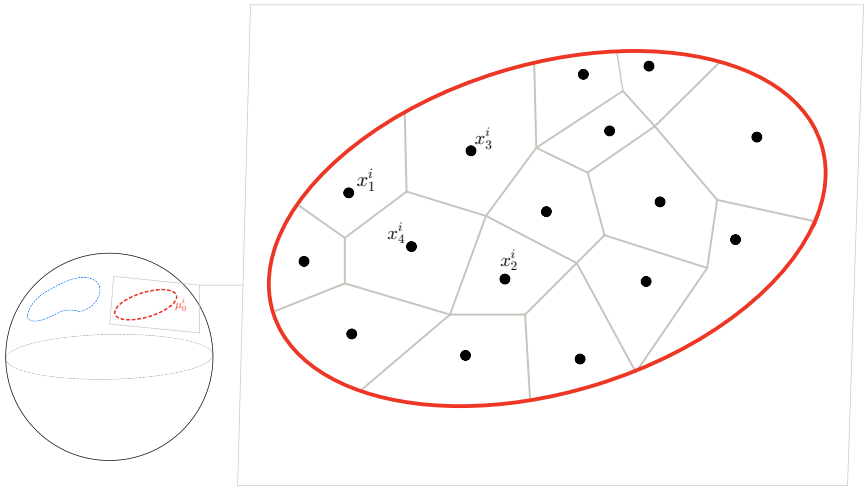

Figure 2: Partitioning $\mathscr{C}^i := \mathrm{supp}(\mu_0^i)$ into M pieces with connected interiors.

Step 2. Packing each part $\mathscr{C}_k^i$ with balls

We henceforth fix an arbitrary $i \in [N]$ and $k \in [M]$. Let $\delta > 0$ to be fixed and determined later on. Consider a packing of $\mathscr{C}_k^i$ consisting of $\mathsf{N}_k^i(\delta) \geqslant 1$ disjoint open balls[6]

$$B(z_{n,i,k}, R_{n,i,k}) \subset \mathscr{C}_k^i \qquad \text{for } n \in [\mathsf{N}_k^i(\delta)],$$

with $z_{n,i,k} \in \mathbb{S}^{d-1}$ and $R_{n,i,k} > 0$, such that

$$\mu_0^i\left(\bigcup_{n\in[\mathsf{N}_k^i(\delta)]} B(z_{n,i,k}, R_{n,i,k})\right) = \alpha_k^i - \delta. \tag{2.12}$$

We now define a target ball contained in $\mathscr{C}_k^i$ to which we aim to send the mass contained in the packing (2.12). Fix the anchor point $x_k^i \in \mathrm{int}\,(\mathscr{C}_k^i)$, and let $\eta > 0$

[6] Recall that all balls are considered as subsets of the sphere, so taken with respect to the geodesic distance, as in (1.10).

be arbitrary and to be determined later on (the same for all indices $(i,k) \in [N] \times [M]$), but also small enough so that

$$\mathscr{B}_k^i := B(x_k^i, \eta) \subset \operatorname{int}\left(\mathscr{C}_k^i\right).$$

We also choose the target ball $\mathscr{B}_k^i$ to satisfy $\mathscr{B}_k^i \subset B(z_{n,i,k}, R_{n,i,k})$ for some $n \in [\mathsf{N}_k^i(\delta)]$ (see Figure 3).

Step 3. Sending most of the mass to $\mathscr{B}_k^i$

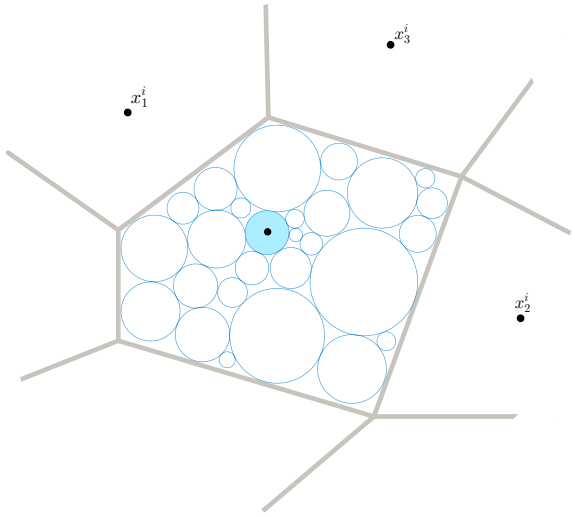

Figure 3: Step 2: packing the piece $\mathscr{C}_k^i$ of the partition of $\mathscr{C}^i = \operatorname{supp}(\mu_0^i)$ with balls whose union has mass $\mu_0^i(\mathscr{C}_k^i) - \delta$. A single anchorpoint x_k^i lies in this piece. The goal of Step 3 is to repeatedly use Lemma C.2 to transfer the mass of each ball to the one highlighted in blue.

As $\operatorname{int}\left(\mathscr{C}_k^i\right)$ is connected and thus path-connected (both are equivalent for open sets in our setup), for every $n \in [\mathsf{N}_k^i(\delta)]$ we can find a sequence of open balls $\{\mathscr{B}_{\ell,n}\}_{0 \leqslant \ell \leqslant L_{k,n}^i} \subset \mathscr{C}_k^i$ satisfying

$$\begin{aligned} &\mathscr{B}_{0,n} = B(z_{n,i,k}, R_{n,i,k}), \\ &\mathscr{B}_{\ell,n} \cap \mathscr{B}_{\ell+1,n} \neq \varnothing, \\ &\mathscr{B}_{L_n,n} = \mathscr{B}_k^i, \\ &\mathscr{B}_{\ell,n} \cap \mathscr{B}_{\ell',n} = \varnothing \qquad \text{if } |\ell' - \ell| \geqslant 2. \end{aligned} \tag{2.13}$$

Set $L_k^i := \max_{n \in [\mathsf{N}_k^i(\delta)]} L_n$ and fix an arbitrary $\bar{\varepsilon} > 0$ to be determined later on. We apply Lemma C.2 for each piece $k \in [M]$ and $n \in [\mathsf{N}_k^i(\delta)]$—recalling (2.11)—to find piecewise constant $(\boldsymbol{W}_i, \boldsymbol{U}_i, b_i) : [T_{i-1}, T_i] \to \mathscr{M}_{d\times d}(\mathbb{R})^2 \times \mathbb{R}^d$ with at most

$$K \cdot \max_{k \in [M]} \mathsf{N}_k^i(\delta) \cdot L_k^i$$

switches, such that

$$\begin{aligned}\mu^i(T_i, \mathscr{B}_k^i) &\geqslant (1-\overline{\varepsilon})^{L_k^i} \mu_0^i \left(\bigcup_{n \in [\mathsf{N}_k^i(\delta)]} \bigcup_{\ell=0}^{L_n} \mathscr{B}_{\ell,n} \right) \\ &\geqslant (1-\overline{\varepsilon})^{L_k^i} \mu_0^i \left(\bigcup_{n \in [\mathsf{N}_k^i(\delta)]} B(z_{n,i,k}, R_{n,i,k}) \right) \\ &\overset{(2.12)}{=} (1-\overline{\varepsilon})^{L_k^i} \left(\alpha_k^i - \delta \right). \end{aligned} \tag{2.14}$$

Moreover, $\mu^i(T_i) = \Phi^{T_i}_{\#} \mu_0^i$, and $\Phi^{T_i}(x) = x$ for all $x \notin \operatorname{supp}(\mu_0^i)$ because of (C.21). Using Kantorovich-Rubinstein duality,

$$\begin{aligned}\mathsf{W}_1(\mu^i(T_i), \mu_1^i) &= \sup_{\operatorname{Lip}(\phi) \leqslant 1} \left| \int \phi (\mu^i(T_i) - \mu_1^i) \right| \\ &= \sup_{\operatorname{Lip}(\phi) \leqslant 1} \left| \sum_{k=1}^{M} \int_{\mathscr{B}_k^i} \phi(\mu^i(T_i) - \mu_1^i) + \int_{\mathbb{S}^{d-1} \setminus (\bigcup_{k \in [M]} \mathscr{B}_k^i)} \phi(\mu^i(T_i) - \mu_1^i) \right|.\end{aligned}$$

Note that without loss of generality we can maximize over all $\phi \in W^{1,\infty}(\mathbb{S}^{d-1})$ with $\operatorname{Lip}(\phi) \leqslant 1$ and of average 0. Such functions have an $L^\infty(\mathbb{S}^{d-1})$–norm bounded by the length of any geodesic, namely 2π.[7] Going term by term in the identity above, using (2.14) and the definition of $\mathscr{B}_k^i$ we find

$$\begin{aligned}\int_{\mathscr{B}_k^i} \phi(\mu^i(T_i) - \mu_1^i) &= \int_{\mathscr{B}_k^i} \phi \mu^i(T_i) - \alpha_k^i \phi(x_k^i) \\ &= \int_{\mathscr{B}_k^i} \phi \mu^i(T_i) - (\alpha_k^i - \overline{\delta}) \phi(x_k^i) - \overline{\delta} \phi(x_k^i) \\ &= \int_{\mathscr{B}_k^i} (\phi(x) - \phi(x_k^i)) \mu^i(T_i) - \overline{\delta} \phi(x_k^i),\end{aligned}$$

where $\overline{\delta} := \alpha_k^i - \mu^i(T_i, \mathscr{B}_k^i) > 0$. By virtue of (2.13) and (C.21) we also gather that $\mu(T_i, \mathscr{C}_k^i) = \mu_0(\mathscr{C}_k^i) = \alpha_k^i$, and therefore

$$\alpha_k^i = \mu^i(T_i, \mathscr{C}_k^i) \geqslant \mu^i(T_i, \mathscr{B}_k^i).$$

Owing to (2.14), we find

$$\overline{\delta} \leqslant \alpha_k^i - (1-\overline{\varepsilon})^{L_k^i} (\alpha_k - \delta),$$

which clearly goes to 0 as $\overline{\varepsilon}$ and δ go to 0. Therefore

$$\left| \int_{\mathscr{B}_k^i} \phi(\mu^i(T_i) - \mu_1^i) \right| \leqslant \eta \mu^i(T_i, \mathscr{B}_k^i) + \overline{\delta} \|\phi\|_{L^\infty(\mathbb{S}^{d-1})},$$

[7] We can write $\phi = \psi + \int \phi$ (vertical shift), and clearly $\|\psi\|_{\operatorname{Lip}(\mathbb{S}^{d-1})} \leqslant 1$. Since ψ has average 0 and is continuous, $\psi(x) = 0$ for some $x \in \mathbb{S}^{d-1}$. Whence $|\psi(x)| \leqslant d_g(x, 0)$ for $x \in \mathbb{S}^{d-1}$.

which tends to 0 as δ, $\bar{\varepsilon}$ and η tend to zero. On the other hand, thanks to (2.14),

$$\begin{aligned}\left|\int_{\mathbb{S}^{d-1}\setminus(\bigcup_{k\in[M]}\mathcal{B}_k^i)}\phi(\mu^i(T_i)-\mu_1^i)\right| &\leqslant 2\pi\mu^i\left(T_i,\mathbb{S}^{d-1}\setminus\bigcup_{k\in[M]}\mathcal{B}_k^i\right)\\ &\leqslant 2\pi\left|1-(1-\bar{\varepsilon})^{\max_{k\in[M]}L_k^i}\sum_{k=1}^{M}(\alpha_k^i-\delta)\right|\\ &\leqslant 2\pi\left|1-(1-\bar{\varepsilon})^{\max_{k\in[M]}L_k^i}(1-M\delta)\right|,\end{aligned}$$

which also tends to 0 as $\bar{\varepsilon}$ and δ tend to 0. Therefore, we can choose $\bar{\varepsilon}, \delta$ and η small enough so that

$$\mathsf{W}_1\left(\mu^i(T_i),\mu_1^i\right)\leqslant\varepsilon.$$

We can conclude since all Wasserstein distances are equivalent on $\mathbb{S}^{d-1}$. □

Remark 2.4. *Looking at the proof, we deduce that* $(\boldsymbol{W},\boldsymbol{U},b)$ *have at most*

$$N\cdot M\cdot\max_{(i,k)\in[N]\times[M]}\mathsf{N}_k^i(\delta)\cdot\max_{n\in[\mathsf{N}_k^i(\delta)]}L_{k,n}^i$$

switches, where $\mathsf{N}_k^i(\delta)$ *and* $L_{k,n}^i$ *are defined in Step 2 and Step 3 respectively.*

3 Disentangling supports

In this section we show that flows generated by Transformers can disentangle measures with overlapping supports, in the sense that if

$$\operatorname{supp}(\mu_0^i)\cap\operatorname{supp}(\mu_0^j)\neq\varnothing,$$

for all $i\neq j\in[N]$, then we can find parameters θ so that the corresponding solution to (1.4) at time $T>0$ satisfies

$$\operatorname{supp}(\mu^i(T))\cap\operatorname{supp}(\mu^j(T))=\varnothing$$

for all $i\neq j\in[N]$. The fact that the vector field governing the continuity equation (1.4) is nonlinear as a function of $\mu(t)$ is essential in this endeavor. We provide two different proofs: when $\boldsymbol{B}\equiv 0$ (which is a generalization of the celebrated *Kuramoto model* to spheres) just below, and $\boldsymbol{B}\not\equiv 0$ in **Appendix D**.

Set

$$\mathbb{Q}_1^{d-1}:=\mathbb{S}^{d-1}\cap(\mathbb{R}_{>0})^d.$$

We now prove that for disentangling the supports it suffices to consider $\boldsymbol{B}\equiv 0$. We consider (1.4) with

$$v[\mu](t,x)=\mathsf{P}_x^{\perp}\left(\boldsymbol{V}(t)\mathbb{E}_{\mu(t)}[z]+\boldsymbol{W}(t)(\boldsymbol{U}(t)x+b(t))_+\right).\tag{3.1}$$

The following is the main result of this section.

Proposition 3.1 (*A Separation*). *Let $T > 0$ and $\mu_0^i \in \mathcal{P}(\mathbb{Q}_1^{d-1})$, $i \in [N]$, be given. There exists $\theta \in L^\infty((0,T);\Theta)$ such that*

$$\operatorname{conv}_g \operatorname{supp}(\mu^i(T)) \cap \operatorname{conv}_g \operatorname{supp}(\mu^j(T)) = \varnothing$$

for all $i \neq j \in [N]$, where $\mu^i \in \mathcal{C}^0([0,T];\mathcal{P}(\mathbb{S}^{d-1}))$ denotes the unique solution to (1.4)–(3.1) corresponding to μ_0^i and θ.

Furthermore, we can take θ to be piecewise constant, having $O(d \cdot N)$ switches.

We defer the proof to **Section 3.3**. Proposition 3.1 entails the existence of a continuous solution map

$$\Phi_\theta^T : \mathcal{P}(\mathbb{S}^{d-1}) \to \mathcal{P}(\mathbb{S}^{d-1})$$

which satisfies

$$\operatorname{conv}_g \operatorname{supp}\left(\Phi_\theta^T(\mu_0^i)\right) \cap \operatorname{conv}_g \operatorname{supp}\left(\Phi_\theta^T(\mu_0^j)\right) = \varnothing$$

for all $i \neq j \in [N]$. This is of course totally equivalent to what is stated in Proposition 3.1, but in subsequent arguments, referring directly to the flow map Φ_θ^T instead of the parameters θ significantly eases the presentation, and we choose to do so.

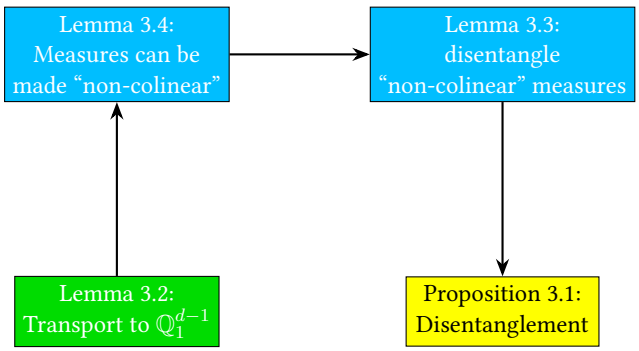

Figure 4: High-level overview of the proof of Proposition 3.1.

3.1 Transportation to $\mathbb{Q}_1^{d-1}$

By virtue of the following lemma, working with initial measures supported on $\mathbb{Q}_1^{d-1}$ is without loss of generality.

Lemma 3.2. *Let $T > 0$. Suppose that $\mu_0^i \in \mathcal{P}(\mathbb{S}^{d-1})$, for $i \in [N]$, are such that*

$$\bigcup_{i \in [N]} \operatorname{supp}(\mu_0^i) \subset \mathbb{S}^{d-1}.$$

Then there exists $\theta = (\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{B}, \boldsymbol{V}, b) \in L^\infty((0,T);\Theta)$ such that

$$\operatorname{supp}(\mu^i(T)) \subset \mathbb{Q}_1^{d-1}$$

for all $i \in [N]$ where $\mu^i \in \mathscr{C}^0([0,T];\mathscr{P}(\mathbb{S}^{d-1}))$ denotes the unique solution to (1.4) corresponding to μ_0^i and θ.

Furthermore, one can take $\boldsymbol{V} \equiv \boldsymbol{B} \equiv \boldsymbol{U} \equiv 0, b \equiv \mathbf{1}$, and $\boldsymbol{W}$ piecewise constant, having at most one switch, and satisfying

$$\|\boldsymbol{W}\|_{L^\infty((0,T);\mathscr{M}_{d\times d}(\mathbb{R}))} \leqslant \frac{C}{T},$$

where $C > 0$ depends only on the supports of the measures μ_0^i, $i \in [N]$.

Proof of Lemma 3.2. Set $\boldsymbol{V} \equiv \boldsymbol{B} \equiv \boldsymbol{U} \equiv 0$ and $b \equiv \mathbf{1}$.

We begin by noting that by assumption, there exists some $\omega \in \mathbb{S}^{d-1}$ with $\omega \notin \operatorname{supp}(\mu_0^i)$ for all $i \in [N]$. Let $T_0 \in (0,T)$ be chosen later on. We select

$$\boldsymbol{W}(t) \equiv \boldsymbol{W}_1 \qquad \text{for } t \in [0,T_0]$$

where $\boldsymbol{W}_1$ is any $d \times d$ matrix such that

$$\boldsymbol{W}_1 \mathbf{1} = -\omega.$$

For this choice of parameters, the characteristics of (1.4) read

$$\begin{cases} \dot{x}(t) = \mathbf{P}^{\perp}_{x(t)}(-\omega) & \text{in } [0,T_0] \\ x(0) = x_0. \end{cases} \tag{3.2}$$

For any $x_0 \in \mathbb{S}^{d-1} \setminus \{\omega\}$, we observe that the solution to (3.2) converges to $-\omega$ in long time. Indeed,

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle x(t),\omega\rangle = -1 + \langle x(t),\omega\rangle^2 < 0 \tag{3.3}$$

whenever $x(t) \in \mathbb{S}^{d-1}\setminus\{\pm\omega\}$. The solution to (3.2) defines a Lipschitz-continuous flow map

$$\Phi^t \colon \mathbb{S}^{d-1} \to \mathbb{S}^{d-1},$$

which is independent of any measure, and which allows to write the solution $\mu^i(t,\cdot)$ to (1.4) on $[0,T_0]$ as

$$\mu^i(t) = \Phi^t_{\#}\mu_0^i$$

for $i \in [N]$. Since ω does not belong to the support of any of the measures μ_0^i, by virtue of (3.3), we can choose $T_0 > 0$, depending on

$$\min_{x \in \bigcup_{i\in[N]} \operatorname{supp}(\mu_0^i)} d_g(x,\omega)$$

so that

$$\max_{x\in\operatorname{supp}(\mu^i(T_0))} d_g(x,-\omega) \leqslant \frac{\pi}{8}$$

for all $i \in [N]$.

Let $\alpha \in \mathbb{Q}_1^{d-1}$ be such that

$$d_g(-\omega, -\alpha) > \frac{\pi}{8}.$$

As a consequence, $-\alpha$ does not belong to the support of $\mu^i(T_0)$ for any $i \in [N]$. Now we choose

$$\boldsymbol{W}(t) \equiv \boldsymbol{W}_2 \qquad \text{for } t \in [T_0, T]$$

where $\boldsymbol{W}_2$ is any $d \times d$ matrix such that

$$\boldsymbol{W}_2 \mathbf{1} = \alpha.$$

The characteristics of (1.4) in $[T_0, T]$ then read

$$\begin{cases} \dot{x}(t) = \mathbf{P}^{\perp}_{x(t)}(\alpha) & \text{in } [T_0, T] \\ x(T_0) = x_0. \end{cases}$$

This differential equation defines a Lipschitz-continuous flow map

$$\Psi^{t-T_0} : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$$

with the same properties as in Step 1 (just replacing $-\omega$ by α), i.e., we can choose $T > 0$ large enough so that

$$\mu^i(T) = \left(\Psi^{T-T_0} \circ \Phi^{T_0}\right)_{\#} \mu_0^i$$

satisfies

$$\operatorname{supp}(\mu^i(T)) \subset \mathbb{Q}_1^{d-1}$$

for all $i \in [N]$, as desired. The bound on $\boldsymbol{W}$ follows by rescaling time. □

3.2 A pair of lemmas

The proof of Proposition 3.1 is based on the following lemmas.

Lemma 3.3. *Let $\mu_0^i \in \mathcal{P}(\mathbb{Q}_1^{d-1})$, $i \in [N]$, be given such that*

$$\mathbb{E}_{\mu_0^i}[x] \text{ is not colinear with } \mathbb{E}_{\mu_0^j}[x] \qquad \text{for } i \neq j.$$

Fix $j \in [N]$. Then for any $T > 0$ and $\varepsilon > 0$, and for any $\nu_0 \in \mathcal{P}(\mathbb{Q}_1^{d-1})$ such that $\mathbb{E}_{\nu_0}[x]$ is colinear with $\mathbb{E}_{\mu_0^N}[x]$, there exists $\theta \in L^\infty((0,T);\Theta)$ such that

$$\operatorname{supp}(\nu(T)) \cup \operatorname{supp}(\mu^j(T)) \subset B\left(\frac{\mathbb{E}_{\mu_0^j}[z]}{\left\|\mathbb{E}_{\mu_0^j}[z]\right\|}, \varepsilon\right),$$

and

$$\mu^i(T) = \mu_0^i$$

for $i \neq j \in [N]$, where $\mu^i, \nu \in \mathcal{C}^0([0,T]; \mathcal{P}(\mathbb{S}^{d-1}))$ denote the unique solution to (1.4)–(3.1) corresponding to data μ_0^i, ν_0, and the parameters θ.

Furthermore, one can take θ piecewise constant, having $O(d \cdot N)$ switches.

We postpone the proof to **Appendix C.4**.

Lemma 3.4. *Let $T > 0$ and let $\mu_0, \nu_0 \in \mathcal{P}(\mathbb{Q}_1^{d-1})$ be two different measures such that*

$$\mathbb{E}_{\mu_0}[x] = \gamma_1 \mathbb{E}_{\nu_0}[x]$$

for some $\gamma_1 \in (0, 1]$.

1. *If $\gamma_1 = 1$, then, setting $\boldsymbol{V} \equiv 0$, there exist $\boldsymbol{W}, \boldsymbol{U} \in \mathcal{M}_{d\times d}(\mathbb{R})$ and $b \in \mathbb{R}^d$ such that the unique solutions μ, ν to (1.4)–(3.1) corresponding to μ_0, ν_0 and these parameters, satisfy*

$$\mathbb{E}_{\mu(T)}[x] \neq \mathbb{E}_{\nu(T)}[x].$$

   Moreover the Lipschitz-continuous and invertible flow map $\Phi^T : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ induced by the characteristics of (1.4)–(3.1) with these parameters satisfies

$$\Phi^T(x) = x \qquad \text{for } x \in \mathbb{S}^{d-1} \setminus (\operatorname{supp}(\mu_0) \cup \operatorname{supp}(\nu_0)). \tag{3.4}$$

2. *If $\gamma \neq 1$, then, setting $\boldsymbol{B} \equiv 0$, there exist $(\boldsymbol{V}, \boldsymbol{W}, \boldsymbol{U}) \in L^\infty((0,T); \mathcal{M}_{d\times d}(\mathbb{R})^3)$ and $b \in L^\infty((0,T);\mathbb{R}^d)$, piecewise constant with at most 2 switches, such that the unique solutions μ, ν to (1.4)–(3.1) corresponding to data μ_0, ν_0 and these parameters satisfy*

$$\mathbb{E}_{\mu(T)}[x] \neq \gamma_2 \mathbb{E}_{\nu(T)}[x]$$

   for all $\gamma_2 \in \mathbb{R}$. In fact,

$$\boldsymbol{V}(t) = I_d 1_{(0,T_*)}(t) \qquad \boldsymbol{W}(t) = \boldsymbol{W} 1_{(T_*,T)}(t)$$
$$\boldsymbol{U}(t) = \boldsymbol{U} 1_{(T_*,T)}(t) \qquad b(t) = b 1_{(T_*,T)}(t)$$

   for some $T_* \in (0,T)$ and $\boldsymbol{W}, \boldsymbol{U} \in \mathcal{M}_{d\times d}(\mathbb{R})$, $b \in \mathbb{R}^d$. Moreover the Lipschitz-continuous and invertible flow map $\Phi^T : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ induced by the characteristics

$$\begin{cases} \dot{x}(t) = \mathbf{P}_x^\perp(\boldsymbol{W}(t)(\boldsymbol{U}(t)x + b(t))_+) & \text{in } [0,T] \\ x(0) = x \in \mathbb{S}^{d-1}, \end{cases}$$

   satisfies

$$\Phi^T(x) = x \qquad \text{for all } x \notin \operatorname{conv}_g \operatorname{supp}(\mu_0) \cup \operatorname{conv}_g \operatorname{supp}(\nu_0).$$

We postpone the proof to **Appendix C.5**.

3.3 Proof of Proposition 3.1

Proof of Proposition 3.1. We argue by induction over N. The base case $N = 1$ is trivially satisfied. Assume that

$$\operatorname{conv}_g \operatorname{supp}(\mu_0^i) \cap \operatorname{conv}_g \operatorname{supp}(\mu_0^j) = \varnothing \qquad \text{for } i \neq j \in [N-1], \tag{3.5}$$

and let $\mu_0^N \in \mathcal{P}(\mathbb{Q}_1^{d-1})$ be arbitrary. We prove that there exist parameters θ as in the statement, such that the solution μ^i to (1.4) satisfies

$$\operatorname{supp}(\mu^i(T)) \cap \operatorname{supp}(\mu^j(T)) = \varnothing \qquad \text{for } i \neq j \in [N].$$

Since $\operatorname{supp}(\mu_0^i) \subset \mathbb{Q}_1^{d-1}$, (3.5) implies that

$$\mathbb{E}_{\mu_0^i}[x] \text{ is not colinear with } \mathbb{E}_{\mu_0^j}[x] \qquad \text{for } i \neq j \in [N-1].$$

Now if $\mathbb{E}_{\mu_0^N}[x]$ is not colinear with $\mathbb{E}_{\mu_0^i}[x]$ for all $i \in [N-1]$, one can conclude by a simple application of Lemma 3.3. On another hand, as a consequence of (3.5), $\mathbb{E}_{\mu_0^N}[x]$ is colinear with $\mathbb{E}_{\mu_0^i}[x]$ for at most one $i \in [N-1]$. Suppose that this is the case, and without loss of generality, we label this index $i = N-1$. We now proceed as follows.

1. In $[0, T/4]$, we apply Lemma 3.3, with $\varepsilon > 0$ small enough, to guarantee the existence of piecewise constant $\theta_1 \in L^\infty((0, T/4); \Theta)$ having $O(d \cdot N)$ switches, such that the solution to (1.4) satisfies

$$\begin{aligned} \operatorname{supp}\left(\mu^j\left(\frac{T}{4}\right)\right) \cap \operatorname{supp}\left(\mu^N\left(\frac{T}{4}\right)\right) &= \varnothing \\ \operatorname{supp}\left(\mu^j\left(\frac{T}{4}\right)\right) \cap \operatorname{supp}\left(\mu^{N-1}\left(\frac{T}{4}\right)\right) &= \varnothing \end{aligned} \tag{3.6}$$

   for all $j \in [N-2]$.

2. In $[T/4, T/2]$, we apply the first part of Lemma 3.4 to find constant θ_2 such that

$$\mathbb{E}_{\mu^{N-1}\left(\frac{T}{2}\right)}[x] \neq \mathbb{E}_{\mu^N\left(\frac{T}{2}\right)}[x],$$

   whereas, thanks to (3.4) and the Lipschitz character of the ODE,

$$\operatorname{supp}\left(\mu^j\left(\frac{T}{2}\right)\right) \cap \operatorname{supp}\left(\mu^{N-1}\left(\frac{T}{2}\right)\right) = \varnothing$$

   for all $j \in [N-2]$.

3. In $[T/2, 3T/4]$, we apply the second part of Lemma 3.4 to $\mu^{N-1}(T/2)$ and $\mu^N(T/2)$ so that there are some piecewise constant $\theta_3 \in L^\infty((T/2, 3T/4); \Theta)$ such that

$$\mathbb{E}_{\mu^N\left(\frac{3T}{4}\right)}[x] \text{ is not colinear with } \mathbb{E}_{\mu^{N-1}\left(\frac{3T}{4}\right)}[x].$$

   Furthermore, owing to (3.6), and noting that $\boldsymbol{V} = I_d$ in Lemma 3.4, along with (2.5), we also have

$$\operatorname{supp}\left(\mu^i\left(\frac{3T}{4}\right)\right) \cap \operatorname{supp}\left(\mu^j\left(\frac{3T}{4}\right)\right) = \varnothing$$

   for all $i \neq j \in [N-1]$, and for all $i \in [N-2]$ and $j = N$.

4. The assumption of Lemma 3.3 is now fulfilled by all N measures, so by picking $\varepsilon > 0$ small enough and applying Lemma 3.3 once again, this time in $[3T/4, T]$, the conclusion follows. □

4 Matching discrete measures

The goal of this section is to prove the following result.

Proposition 4.1. *Suppose $d \geqslant 3$. Consider*

$$\left(x_0^i, y^i\right) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \qquad \text{for } i \in [M], \tag{$\mathscr{D}$}$$

with $x_0^i \neq x_0^j$ and $y^i \neq y^j$ for $i \neq j$, and suppose that for any $i \in [M]$, there exist $\gamma_i \in \mathbb{S}^{d-1}$ and $\varepsilon_i > 0$ such that

$$\left\langle \gamma_i, x_0^i - y^i \right\rangle = 0 \qquad \text{and} \qquad x_0^j \notin H_{\varepsilon_i}^{\gamma_i}$$

for $j \neq i \in [N]$, where

$$H_{\varepsilon_i}^{\gamma_i} := \left\{ x \in \mathbb{S}^{d-1} \colon |\langle x, \gamma_i \rangle| \leqslant \varepsilon_i \right\}.$$

Then for any $T > 0$, there exists $\theta = (\boldsymbol{W}, \boldsymbol{U}, b) \in L^\infty((0,T); \mathscr{M}_{d\times d}(\mathbb{R})^2 \times \mathbb{R}^d)$, piecewise constant having at most $6M$ switches, such that for any $i \in [M]$, the solution $x^i(\cdot) \in \mathscr{C}^0([0,T]; \mathbb{S}^{d-1})$ to

$$\begin{cases} \dot{x}^i(t) = \mathbf{P}_x^\perp \left(\boldsymbol{W}(t) \Big(\boldsymbol{U}(t) x^i(t) + b(t) \Big)_+ \right) & \text{in } [0,T] \\ x^i(0) = x_0^i, \end{cases} \tag{4.1}$$

satisfies

$$x^i(T) = y^i.$$

Moreover there exists some $C > 0$, not depending on $\mathscr{D}$ nor T, such that

$$\|\theta\|_{L^\infty((0,T);\Theta)} \leqslant \frac{C \cdot M}{T \min\limits_{i \in [M]} \varepsilon_i}.$$

The proof of Proposition 4.1 follows directly from the following result, combined with a straightforward induction argument.

Proposition 4.2. *Suppose $d \geqslant 3$. Consider*

$$\left(x_0^i, y^i\right) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \qquad \text{for } i \in [M], \tag{$\mathscr{D}$}$$

with $x_0^i \neq x_0^j$ and $y^i \neq y^j$ for $i \neq j$, with $x_0^i = y^i$ for $i \in [M-1]$, and suppose that there exist $\gamma \in \mathbb{S}^{d-1}$ and $\varepsilon > 0$ such that

$$\left\langle \gamma, x_0^M - y^M \right\rangle = 0 \qquad \text{and} \qquad x_0^i \notin H_\varepsilon^\gamma$$

for all $i \in [M-1]$. *Then, for every* $T > 0$, *there exist piecewise constant parameters* $\theta = (\boldsymbol{W}, \boldsymbol{U}, b) \in L^\infty((0,T); \mathscr{M}_{d\times d}(\mathbb{R})^2 \times \mathbb{R}^d)$, *all having at most* 6 *switches, such that for any* $i \in [M]$, *the solution* $x^i(\cdot) \in \mathscr{C}^0([0,T]; \mathbb{S}^{d-1})$ *to* (4.1) *satisfies*

$$x^i(T) = y^i.$$

Moreover, there exists some constant $C > 0$, *not depending on* $\mathscr{D}$ *and* T, *such that*

$$\|\theta\|_{L^\infty((0,T);\Theta)} \leqslant \frac{C}{T \cdot \varepsilon}.$$

Proof of Proposition 4.2. The final parameters take the form

$$\begin{aligned}
\boldsymbol{W}(t) &= \sum_{j=1}^{6} \boldsymbol{W}_j 1_{\left[\frac{(j-1)T}{6}, \frac{jT}{6}\right]}(t), \\
\boldsymbol{U}(t) &= \sum_{j=1}^{6} \boldsymbol{U}_j 1_{\left[\frac{(j-1)T}{6}, \frac{jT}{6}\right]}(t), \\
b(t) &= \sum_{j=1}^{6} b_j 1_{\left[\frac{(j-1)T}{6}, \frac{jT}{6}\right]}(t),
\end{aligned}$$

where

$$\begin{aligned}
&\boldsymbol{W}_5 = -\boldsymbol{W}_1 && \boldsymbol{U}_5 = \boldsymbol{U}_1 && b_5 = b_1, \\
&\boldsymbol{W}_6 = -\boldsymbol{W}_2 && \boldsymbol{U}_6 = \boldsymbol{U}_2 && b_6 = b_2, \\
& && \boldsymbol{U}_3 = \boldsymbol{U}_4 && b_3 = b_4.
\end{aligned}$$

Throughout, the time $T > 0$ is adjusted later by rescaling the norm of the parameters.

Step 1. The anchor points

In this step we find three anchor points which serve to build the parameters in what follows. Since $\langle \gamma, x_0^M - y^M \rangle = 0$, we can find some $\omega \in \mathbb{S}^{d-1}$ such that

$$\langle \gamma, \omega \rangle = 0, \tag{4.2}$$

as well as

$$d_g\left(\omega, x_0^M\right) \geqslant \frac{\pi}{2}, \qquad \text{and} \qquad d_g\left(\omega, y^M\right) \geqslant \frac{\pi}{2}. \tag{4.3}$$

Because of (4.2), we consider the point ω_+ lying on the minimizing geodesic between ω and γ, satisfying

$$d_g(\omega_+, \omega) = \frac{\pi}{8}.$$

Similarly, we consider the point ω_- lying on the minimizing geodesic between ω and $-\gamma$, satisfying

$$d_g(\omega_-, \omega) = \frac{\pi}{8}.$$

We have

$$\begin{aligned}
d_g\left(\omega_+, x_0^M\right) &\geqslant d_g\left(\omega, x_0^M\right) - d_g\left(\omega, \omega_+\right) = \frac{3\pi}{8}, \\
d_g\left(\omega_+, y^M\right) &\geqslant \frac{3\pi}{8}, \\
d_g\left(\omega_-, x_0^M\right) &\geqslant \frac{3\pi}{8}, \\
d_g\left(\omega_-, y^M\right) &\geqslant \frac{3\pi}{8}.
\end{aligned}$$

As a consequence, the hyperplane

$$\left\{x \in \mathbb{S}^{d-1} \colon \langle \omega, x\rangle = \cos\left(\frac{\pi}{8} + \tau\right)\right\}$$

is a separating hyperplane for the ball $B\left(\omega, \frac{\pi}{8} + \tau\right)$ and the points x_0^M and y^M for every $\tau \in (0, \frac{3\pi}{8})$; namely

$$\left\langle \omega, x_0^M \right\rangle - \cos\left(\frac{\pi}{8} + \tau\right) = \cos d_g\left(\omega, x_0^M\right) - \cos\left(\frac{\pi}{8} + \tau\right) < 0,$$

where the inequality is by virtue of (4.3). Analogous computations hold for y^M, whereas

$$\langle \omega, x\rangle - \cos\left(\frac{\pi}{8} + \tau\right) > 0$$

for all $x \in B(\omega, \frac{\pi}{8} + \frac{1}{2}\tau)$ and $\tau \in (0, \frac{3\pi}{8})$ (see Figure 5).

Step 2. Isolating x_0^M and y^M

Let

$$\epsilon := \min\left\{\varepsilon, \frac{\pi}{4}\right\}.$$

Consider

$$\boldsymbol{U}_1 = \gamma \mathbf{1}^\top \qquad \text{and} \qquad b_1 = -\frac{\epsilon}{2}\mathbf{1}.$$

Then

$$\left(\boldsymbol{U}_1 x + b_1\right)_+ = \left(\langle \gamma, x\rangle - \frac{\epsilon}{2}\right)_+ \mathbf{1}.$$

Choose any $\boldsymbol{W}_1$ so that

$$\boldsymbol{W}_1 \mathbf{1} = \omega_+.$$

Define

$$\mathcal{S}_+ := \left\{x \in \mathbb{S}^{d-1} : \langle \gamma, x\rangle \geqslant \epsilon\right\}.$$

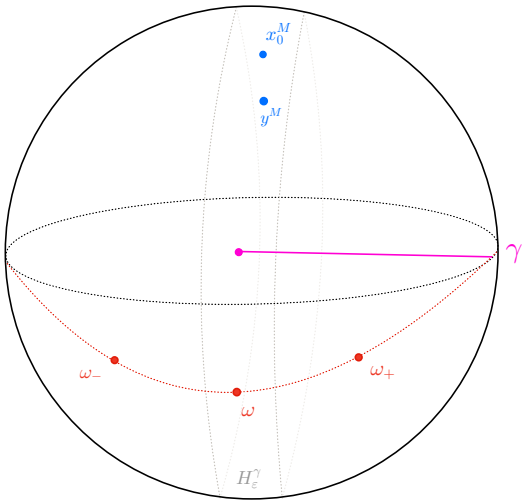

Figure 5: The geometric configuration of Step 1.

Obviously $\omega_+ \in \mathcal{S}_+$. Observe that the trajectories of the ODE

$$\dot{x}(t) = \left(\langle \gamma, x(t)\rangle - \frac{\epsilon}{2}\right)_+ \mathbf{P}^{\perp}_{x(t)}(\omega_+) \qquad \text{for } t \geqslant 0, \tag{4.4}$$

follow the Riemannian gradient flow of the distance between ω_+ and x in $\mathcal{S}_+$. Indeed,

$$\nabla_1 d_g(x, \omega_+) = -\frac{\mathbf{P}^{\perp}_x(\omega_+)}{\sqrt{1 - \langle x, \omega_+\rangle^2}}.$$

Then, setting $f(x) = \left(\langle \gamma, x\rangle - \frac{\epsilon}{2}\right)_+$, we have

$$\begin{aligned}\dot{x}(t) &= -f(x(t))\sqrt{1 - \langle x(t), \omega_+\rangle^2}\,\nabla_1 d_g(x(t), \omega_+) \\ &= -\overline{f}(x(t))\,\nabla_1 d_g(x(t), \omega_+).\end{aligned}$$

Since $\overline{f}$ is a nonnegative scalar function, by appropriately reparameterizing time, we conclude that $x(t)$ follows the desired gradient flow. In turn, the trajectory $x(t)$ of (4.4) starting from any $x_0 \in \mathcal{S}_+$ always lies on the minimal geodesic from x_0 to $\omega_+ \in \mathcal{S}_+$. Since $\mathcal{S}_+$ is geodesically convex, we gather that $x(t) \in \mathcal{S}_+$ for all $t \geqslant 0$. Then, notice that

$$\overline{f}(x) = 0 \qquad \Longleftrightarrow \qquad x = \omega_+ \quad \text{or} \quad x \in \left\{ y \in \mathbb{S}^{d-1} : \langle \gamma, y\rangle \leqslant \frac{\epsilon}{2} \right\}.$$

Thus, unless $x(t) = \omega_+$, $\overline{f}$ is uniformly bounded from below on $\mathcal{S}_+$, and since

$$\nabla_1 d_g(x, \omega_+) = 0 \qquad \Longleftrightarrow \qquad x = \pm\omega_+,$$

we can conclude that $x(t) \to \omega_+$ as $t \to +\infty$ for any $x_0 \in \mathcal{S}_+$ by applying the LaSalle invariance principle [LaS60].

For any $x_0 \in \mathbb{S}^{d-1}$, set

$$T_{\frac{\pi}{16}}(x_0) := \inf \left\{ t \geqslant 0 : \quad x(t) \in B\left(\omega_+, \frac{\pi}{16}\right) \right\},$$

where $x(\cdot)$ is the solution to the Cauchy problem for (4.4) with data x_0. Since $\|\gamma\| = 1$, $\|\boldsymbol{W}_1\|_{\mathrm{op}} \leqslant 1$ and $\|b_1\| \leqslant \frac{\epsilon}{2}\sqrt{d}$, bounding the L^∞–norm of the parameters comes from bounding $T_{\frac{\pi}{16}}(x_0)$ uniformly over $x_0 \in \mathcal{S}_+$ and rescaling time. For every $x_0 \in B\left(\omega_+, \frac{\pi}{16}\right)$ we see that $T_{\frac{\pi}{16}}(x_0)$ is trivially 0, whereas for $x_0 \in \mathcal{S}_+ \setminus B\left(\omega_+, \frac{\pi}{16}\right)$ one has

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\langle x(t), \omega_+ \rangle &= \left(\langle \gamma, x(t) \rangle - \frac{\epsilon}{2} \right)_+ \left(1 - \langle \omega_+, x \rangle^2 \right) \\ &\geqslant \frac{\epsilon}{2} \left(1 - \langle \omega_+, x \rangle^2 \right) \\ &\geqslant \frac{\epsilon}{2} \left(1 - \cos^2\left(\frac{\pi}{16}\right) \right). \end{aligned} \tag{4.5}$$

Hence $T_{\frac{\pi}{16}}(x_0) = O(1/\epsilon)$ for all $x_0 \in \mathcal{S}_+$.

Finally, by following the same arguments leading to (4.5), beyond some large enough time, and for every $x_0 \in \mathcal{S}_+$, we can apply the Hartman-Grobman theorem [Har60, Har63, Shu13]: the behavior near the critical point ω_+ is governed[8] by the linearized system

$$\begin{cases} \dot{y}(t) = -\left(\langle \gamma, \omega_+ \rangle - \frac{\epsilon}{2} \right) y(t) & \text{in } \mathbb{R}_{\geqslant 0} \\ y(0) = y_0 \in \mathsf{T}_{\omega_+}\mathbb{S}^{d-1}, \end{cases}$$

which is exponentially stable. Thus, by the Hartman-Grobman theorem, for all $x_0 \in \mathcal{S}_+$,

$$d_g(x(t), \omega_+) \leqslant K e^{-\lambda t} \qquad \text{for all } t \geqslant 0, \tag{4.6}$$

and for some $\lambda > 0$ and $K \geqslant 1$ which depend on x_0, ϵ and γ only.

Similarly, consider

$$\boldsymbol{U}_2 = -\gamma \mathbf{1}^\top \qquad \text{and} \qquad b_2 = -\frac{\epsilon}{2}\mathbf{1}.$$

Then,

$$(\boldsymbol{U}_2 x + b_2)_+ = \left(\langle -\gamma, x \rangle - \frac{\epsilon}{2} \right)_+ \mathbf{1}.$$

Choose any $\boldsymbol{W}_2$ so that

$$\boldsymbol{W}_2 \mathbf{1} = \omega_-.$$

[8] Note that the critical point ω_+ is hyperbolic since we are working in $\mathsf{T}_{\omega_+}\mathbb{S}^{d-1}$. On $\mathbb{R}^d$, there is a zero eigenvalue associated to the radial direction.

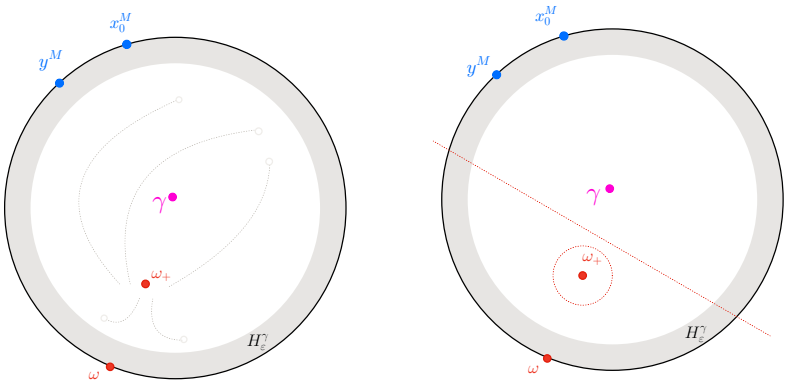

Figure 6: (*Left*) All points in one spherical cap converge to ω_+. (*Right*) All points aside the M-th one are in a neighbourhood of ω_+ or ω_-. Consequently there is a separating hyperplane between x_0^M and y^M (dashed).

Define

$$\mathcal{S}_- := \left\{ x \in \mathbb{S}^{d-1} : \langle -\gamma, x \rangle \geqslant \epsilon \right\}.$$

After reasoning similarly for $\mathcal{S}_-$ as for $\mathcal{S}_+$, and by rescaling time so that

$$\| \boldsymbol{W}_1 \|_{\text{op}} = O\left(\frac{1}{T \cdot \epsilon}\right), \qquad \| \boldsymbol{W}_2 \|_{\text{op}} = O\left(\frac{1}{T \cdot \epsilon}\right),$$

we deduce that for any $T > 0$ there exists $\theta_1 \in L^\infty((0, T/3); \mathscr{M}_{d\times d}(\mathbb{R})^2 \times \mathbb{R}^d)$, piecewise constant having two switches, such that the associated flow map of (4.1) is a Lipschitz-continuous and invertible map that satisfies

$$\begin{aligned} &\Phi_{\theta_1}^{\frac{T}{3}}(x_0^i) \in B\left(\omega, \frac{3\pi}{16}\right) \\ &\Phi_{\theta_1}^{\frac{T}{3}}(x) = x \qquad \text{if } x \in H_\epsilon^\gamma. \end{aligned} \tag{4.7}$$

Step 3. Steering x_0^M to y^M

By virtue of (4.7), the hyperplane

$$\left\{ x \in \mathbb{S}^{d-1} \colon \langle \omega, x \rangle = \cos\left(\frac{3\pi}{16}\right) \right\}$$

is a separating hyperplane: it separates x_0^N and y^N from $\Phi_{\theta_1}^{\frac{T}{3}}(x_0^i)$ for $i \in [M-1]$. Consider

$$\boldsymbol{U}_3 = -\omega \mathbf{1}^\top, \qquad \text{and} \qquad b_3 = \cos\left(\frac{3\pi}{16}\right) \mathbf{1}.$$

With this choice, we have

$$
\begin{aligned}
(\boldsymbol{U}_3 x + b_3)_+ &= 0 && \text{for } x \in B\left(\omega, \frac{3\pi}{16}\right), \\
(\boldsymbol{U}_3 x + b_3)_+ &= \underbrace{\left(\langle -\omega, x\rangle + \cos\left(\frac{3\pi}{16}\right)\right)}_{>0} \mathbf{1} && \text{for } x \in \mathbb{S}^{d-1} \setminus B\left(\omega, \frac{3\pi}{16}\right).
\end{aligned}
$$

Take two points $z_1, z_2 \in \mathbb{S}^{d-1} \setminus B\left(\omega, \frac{3\pi}{16}\right)$ such that

1. $\{c(s)\}_{s\in[0,1]}$ is a geodesic satisfying $c(0) = x_0^M$ and $c(1) = z_2$;
2. $c(1 - s_0) = y^M$ for some $s_0 \in (0, 1)$;
3. $c(s_{z_1}) = z_1$ for some $s_{z_1} \in (0, 1 - s_0)$;
4. $\{c(s)\}_{s\in[0,1]} \subset \mathbb{S}^{d-1} \setminus B\left(\omega, \frac{3\pi}{16}\right)$;
5. $d_g(z_1, x_0^M) \leqslant \kappa\pi$ and $d_g(z_1, z_2) \leqslant \kappa\pi$ for some[9] $\kappa < 1$.

Consider any $d \times d$ matrix $\boldsymbol{W}_3$ such that

$$
\boldsymbol{W}_3 \mathbf{1} = z_1.
$$

The Cauchy problem (4.1) with these parameters, for the $i = M$-th particle, reads

$$
\begin{cases}
\dot{x}(t) = \left(\langle -\omega, x(t)\rangle + \cos\left(\frac{3\pi}{16}\right)\right)_+ \mathbf{P}^{\perp}_{x(t)}(z_1) & \text{on } \mathbb{R}_{\geqslant 0} \\
x(0) = x_0^M.
\end{cases}
\tag{4.8}
$$

Since $d_g(x_0^M, z_1) \leqslant \kappa\pi$, and since the minimizing geodesic between x_0^M and z_1 is contained in $\{c(s)\}_{s\in[0,1]} \subset B(\omega, \frac{3\pi}{16})$, we gather that there exists some large enough time $\tau > 0$ such that

$$
d_g\left(x(\tau), z_2\right) \leqslant d_g\left(x(\tau), z_1\right) + d_g\left(z_1, z_2\right) \lesssim e^{-\lambda\tau} + \kappa\pi \leqslant \kappa_2\pi \tag{4.9}
$$

for some $\kappa_2 < 1$ and $\lambda > 0$. This comes from the long-time convergence of (4.8) to z_1, which can be shown by following the same arguments as for (4.6), replacing ω_+ by z_1. For any $d \times d$ matrix $\boldsymbol{W}_4$ such that

$$
\boldsymbol{W}_4 \mathbf{1} = z_2,
$$

the Cauchy problem (4.1), for the $i = M$-th particle, reads

$$
\begin{cases}
\dot{x}(t) = \left(\langle -\omega, x(t)\rangle + \cos\left(\frac{3\pi}{16}\right)\right)_+ \mathbf{P}^{\perp}_{x(t)}(z_2) & \text{for } t \geqslant \tau \\
x(\tau) = x(\tau)
\end{cases}
\tag{4.10}
$$

[9] can be chosen as such because $\{c(s)\}_{s\in[0,1]} \subset \mathbb{S}^{d-1} \setminus B\left(\omega, \frac{3\pi}{16}\right)$—indeed, take $\kappa = \frac{29}{32}$.

where $x(\tau)$ is the solution of (4.8) at $t = \tau$. Since $d_g(z_1, z_2) \leqslant \kappa\pi$, y^M lies on the minimizing geodesic between $x(\tau)$ and z_2. All the while, thanks to (4.9), taking T even larger than before, we deduce that that the solution to (4.10) satisfies

$$x(T) = y^M.$$

Therefore, as in the previous step, we deduce that for any $T > 0$ there exists some $\theta_2 \in L^\infty((T/3, 2T/3); \mathscr{M}_{d\times d}(\mathbb{R})^2 \times \mathbb{R}^d)$, piecewise constant having two switches, such that the associated flow map of (4.1) is a Lipschitz-continuous and invertible map that satisfies

$$\begin{aligned} \Phi_{\theta_2}^{\frac{2T}{3}}(x) &= x \qquad \text{if } x \in B\left(\omega, \frac{3\pi}{16}\right), \\ \Phi_{\theta_2}^{\frac{2T}{3}}(x_0^M) &= y^M, \end{aligned}$$

and

$$\left(\Phi_{\theta_2}^{\frac{2T}{3}} \circ \Phi_{\theta_1}^{\frac{T}{3}}\right)(x_0^i) = \Phi_{\theta_1}^{\frac{T}{3}}(x_0^i) \in B\left(\omega, \frac{3\pi}{16}\right) \qquad \text{for } i \in [M-1]$$

as well as

$$\left(\Phi_{\theta_2}^{\frac{2T}{3}} \circ \Phi_{\theta_1}^{\frac{T}{3}}\right)(x_0^M) = \Phi_{\theta_2}^{\frac{2T}{3}}(x_0^M) = y^M.$$

Step 4. Bringing $x^i(T)$ back to y^i

We conclude by applying the inverse of the flow map $\Phi_{\theta_1}^{\frac{T}{3}}$: defining

$$\Phi_{\text{fin}}^T := \left(\Phi_{\theta_1}^{\frac{T}{3}}\right)^{-1} \circ \Phi_{\theta_2}^{\frac{2T}{3}} \circ \Phi_{\theta_1}^{\frac{T}{3}},$$

we have $\Phi_{\text{fin}}^T(x_0^i) = y^i$ for all $i \in [M]$, as desired. □

Remark 4.3. *Proposition 2.3 yields a flow map that clusters the support of the input measure, which in turn allows to reduce a universal approximation property of maps in $L^p(\mathbb{S}^{d-1}; \mathbb{S}^{d-1})$ to interpolation via flow maps proved in Proposition 4.1. Indeed it suffices to consider a simple function $\varphi : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ defined as*

$$\varphi(x) = \sum_{i=1}^N y_i 1_{\Omega_i}(x)$$

with $y_i \in \mathbb{S}^{d-1}$. Universal approximation in $L^p(\mathbb{S}^{d-1}; \mathbb{S}^{d-1})$, $p < +\infty$, is equivalent to the W_p-approximate interpolation of

$$\mathrm{d}\mu_0^i(x) = 1_{\Omega_i}(x)\,\mathrm{d}x, \quad \mu_1^i = |\Omega_i|\delta_{y_i}$$

for $i \in [N]$. Note that, by construction, the supports of μ_0^i (and of μ_1^i) are pairwise disjoint. Thus the attention component of the vector field is not needed to perform this task. This is generalized in the next section.

5 Proofs of the main results

Our overarching goal is to construct the solution map $\Phi^T_{\text{fin}} : \mathscr{P}(\mathbb{S}^{d-1}) \to \mathscr{P}(\mathbb{S}^{d-1})$ of the form

$$\Phi^T_{\text{fin}} := \left(\Phi^{\frac{T}{3}}_{\theta_3}\right)^{-1} \circ \Phi^{\frac{T}{3}}_{\theta_2} \circ \Phi^{\frac{T}{3}}_{\theta_1},$$

where

$$\Phi^{\frac{T}{3}}_{\theta_1} =: \Phi_1$$

and

$$\Phi^{\frac{T}{3}}_{\theta_3} =: \Phi_3$$

are the flow maps given by Proposition 3.1, so that the measures $\Phi_1(\mu^i_0)$ (and $\Phi_3(\mu^i_1)$), for $i \in [N]$, have pairwise disentangled supports. The map

$$\Phi^{\frac{T}{3}}_{\theta_2} =: \Phi_2$$

is constructed in this section (see Figure 1 for a schematic overview of the entire proof). The main clue lies in the following three lemmas.

Lemma 5.1 (Propagating transport maps)**.** *Suppose that for every $i \in [N]$ there exists $\mathsf{T}^i \in L^2(\mathbb{S}^{d-1}; \mathbb{S}^{d-1})$ such that*

$$\mathsf{T}^i_{\#}\mu^i_0 = \mu^i_1. \tag{5.1}$$

Consider the flow map $\Phi_1 : \mathscr{P}(\mathbb{S}^{d-1}) \to \mathscr{P}(\mathbb{S}^{d-1})$ (respectively Φ_3) given by Proposition 3.1 with data $\mu^i_0 \in \mathscr{P}(\mathbb{S}^{d-1})$ (respectively $\mu^i_1 \in \mathscr{P}(\mathbb{S}^{d-1})$) for $i \in [N]$. Then, there exists a Lipschitz-continuous and invertible map $\Psi : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ such that

$$\Psi\big|_{\text{supp}\left(\Phi_1(\mu^i_0)\right)} = \Psi^i \tag{5.2}$$

for any $i \in [N]$, where $\Psi^i : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ is another Lipschitz-continuous and invertible map that satisfies

$$\left(\Psi^i \circ \mathsf{T}^i_{\Phi_1}\right)_{\#} \mu^i_0 = \left(\Psi \circ \mathsf{T}^i_{\Phi_1}\right)_{\#} \mu^i_0 = \left(\mathsf{T}^i_{\Phi_3}\right)_{\#} \mu^i_1 = \Phi_3(\mu^i_1)$$

for some Lipschitz-continuous and invertible maps $\mathsf{T}^i_{\Phi_1}, \mathsf{T}^i_{\Phi_3} : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$.

The proof can be found in **Appendix C.6**.

Lemma 5.2. *Suppose $\mu \in \mathscr{P}(\mathbb{S}^{d-1})$ and $\mathsf{T}^1, \mathsf{T}^2 : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ measurable, with T^1 bijective. Then*

$$\mathsf{W}_2\left(\mathsf{T}^1_{\#}\mu, \mathsf{T}^2_{\#}\mu\right) \lesssim \left\|\mathsf{T}^1 - \mathsf{T}^2\right\|_{L^2(\mu)}. \tag{5.3}$$

The proof is elementary, but brief, thus we provide it for completeness.

Proof of Lemma 5.2. Since T^1 is bijective, there is a measurable $\Psi : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ such that

$$\Psi(\mathsf{T}^1(x)) = \mathsf{T}^2(x) \qquad \text{for all } x \in \mathbb{S}^{d-1}.$$

Then

$$\mathsf{W}_2^2\left(\mathsf{T}^1_{\#}\mu, \mathsf{T}^2_{\#}\mu\right) \lesssim \int \|x - \Psi(x)\|^2 \left(\mathsf{T}^1_{\#}\mu\right)(\mathrm{d}x) = \left\|\mathsf{T}^1 - \mathsf{T}^2\right\|^2_{L^2(\mu)}. \qquad \square$$

Remark 5.3. *When μ is absolutely continuous with respect to the Lebesgue measure, and T^1 and T^2 are the optimal transport maps between μ and ν_1, and μ and ν_2 respectively, the upper bound in (5.3) is know as the* linearized optimal transport distance *(see [DM23, JCP23] and the references therein).*

Finally,

Lemma 5.4. *Suppose $\varepsilon > 0$ and $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$. For every $\Psi \in L^2(\mathbb{S}^{d-1}; \mathbb{S}^{d-1})$, there exists a Lipschitz-continuous and invertible map $\Psi_\varepsilon : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ induced by the solution map of (C.20), namely*

$$\Phi^T_{\theta_\varepsilon}(\mu) = (\Psi_\varepsilon)_{\#}\mu$$

for some piecewise constant $\theta_\varepsilon \in L^\infty((0,T);\Theta)$ with finitely many switches, such that

$$\|\Psi - \Psi_\varepsilon\|_{L^2(\mu)} \leqslant \varepsilon.$$

The proof of Lemma 5.4 is involved, so we postpone it to **Appendix C.7**.

5.1 Proof of Theorem 1.2

We provide *two* proofs: we first provide the proof in full generality, followed by a simpler proof that doesn't rely on Lemma 5.4, under stronger structural assumptions on the input and target measures.

Proof of Theorem 1.2 (general case). We split the proof in three steps.

Step 1. Disentanglement

We begin by rendering the supports of the initial measures $(\mu_0^i)_{i\in[N]}$ (respectively, the target measures $(\mu_1^i)_{i\in[N]}$) pairwise disjoint by virtue of applying Proposition 3.1 to (1.4) with data μ_0^i at time $t = 0$ (respectively μ_1^i at time $t = 2T/3$) for any $i \in [N]$. This entails the existence of two parameterized flow maps

$$\Phi^t_{\theta_1} : \mathcal{P}(\mathbb{S}^{d-1}) \mapsto \mathcal{P}(\mathbb{S}^{d-1})$$

for $t \in [0, T/3]$, and

$$\Phi^t_{\theta_3} : \mathcal{P}(\mathbb{S}^{d-1}) \mapsto \mathcal{P}(\mathbb{S}^{d-1})$$

for $t \in [2T/3, T]$, induced by (1.4), which are such that

$$\operatorname{supp}\left(\Phi_{\theta_1}^{\frac{T}{3}}(\mu_0^i)\right) \cap \operatorname{supp}\left(\Phi_{\theta_1}^{\frac{T}{3}}(\mu_0^j)\right) = \varnothing \qquad \text{if } i \neq j,$$

and

$$\operatorname{supp}\left(\Phi_{\theta_3}^{T}(\mu_1^i)\right) \cap \operatorname{supp}\left(\Phi_{\theta_3}^{T}(\mu_1^j)\right) = \varnothing \qquad \text{if } i \neq j.$$

Since (1.4) is well-posed and time-reversible, we further gather that there exists some constant $C = C(T, \theta_3) > 0$ such that

$$\mathsf{W}_2\left(\left(\Phi_{\theta_3}^{T}\right)^{-1}(\mu), \left(\Phi_{\theta_3}^{T}\right)^{-1}(\nu)\right) \leqslant C \cdot \mathsf{W}_2(\mu, \nu) \tag{5.4}$$

holds for any $\mu, \nu \in \mathcal{P}(\mathbb{S}^{d-1})$.

Step 2. Matching

By virtue of Lemma 5.1, there exists a Lipschitz-continuous and invertible map $\Psi : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ such that

$$\Psi\Big|_{\operatorname{supp}\left(\Phi_{\theta_1}^{\frac{T}{3}}(\mu_0^i)\right)} = \Psi^i,$$

for $i \in [N]$ where $\Psi^i : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ satisfies

$$\Psi^i_{\#}\left(\Phi_{\theta_1}^{\frac{T}{3}}(\mu_0^i)\right) = \Phi_{\theta_3}^{\frac{T}{3}}(\mu_1^i). \tag{5.5}$$

We consider

$$\mu = \sum_{i=1}^{N} \Phi_{\theta_1}^{\frac{T}{3}}(\mu_0^i),$$

and use Lemma 5.4 to find a flow map $\Psi_\varepsilon : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ such that

$$\left\| \Psi^i - \Psi_\varepsilon\Big|_{\operatorname{supp}\left(\Phi_{\theta_1}^{\frac{T}{3}}(\mu_0^i)\right)} \right\|_{L^2\left(\Phi_{\theta_1}^{\frac{T}{3}}(\mu_0^i)\right)} \leqslant \|\Psi - \Psi_\varepsilon\|_{L^2(\mu)} \leqslant \frac{\varepsilon}{C} \tag{5.6}$$

for $i \in [N]$. In fact by virtue of Lemma 5.4 there exists a parameterized flow map

$$\Phi_{\theta_2}^{t} : \mathcal{P}(\mathbb{S}^{d-1}) \to \mathcal{P}(\mathbb{S}^{d-1})$$

for $t \in [T/3, 2T/3]$, induced by (1.4) and defined, for $\nu \in \mathcal{P}(\mathbb{S}^{d-1})$, as

$$\Phi_{\theta_2}^{\frac{2T}{3}}(\nu) = (\Psi_\varepsilon)_{\#}\nu,$$

which by virtue of (5.5), (5.6) and Lemma 5.2, satisfies

$$\mathsf{W}_2\left(\left(\Phi_{\theta_2}^{\frac{2T}{3}} \circ \Phi_{\theta_1}^{\frac{T}{3}}\right)(\mu_0^i), \Phi_{\theta_3}^{T}(\mu_1^i)\right) \leqslant \frac{\varepsilon}{C} \tag{5.7}$$

for all $i \in [N]$.

Step 3. Continuity

We now apply the inverse of $\Phi^T_{\theta_3}$ to conclude that the map

$$\Phi^T_{\text{fin}} := \left(\Phi^T_{\theta_3}\right)^{-1} \circ \Phi^{\frac{2T}{3}}_{\theta_2} \circ \Phi^{\frac{T}{3}}_{\theta_1}$$

satisfies

$$\begin{aligned}
\mathsf{W}_2\left(\Phi^T_{\text{fin}}(\mu^i_0), \mu^i_1\right) &= \mathsf{W}_2\left(\Phi^T_{\text{fin}}(\mu^i_0), \left(\left(\Phi^T_{\theta_3}\right)^{-1} \circ \Phi^T_{\theta_3}\right)(\mu^i_1)\right) \\
&\overset{(5.4)}{\leqslant} C \cdot \mathsf{W}_2\left(\left(\Phi^{\frac{2T}{3}}_{\theta_2} \circ \Phi^{\frac{T}{3}}_{\theta_1}\right)(\mu^i_0), \Phi^T_{\theta_3}(\mu^i_1)\right) \\
&\overset{(5.7)}{\leqslant} \varepsilon,
\end{aligned}$$

for all $i \in [N]$, as desired. □

We now provide a different proof under the assumption that the input measures are absolutely continuous, and the targets are empirical measures with M atoms. The advantage of this proof is that it provides an explicit estimate on the number of parameter switches.

Proof of Theorem 1.2 (restricted case). We assume that the target measures μ^i_1 are all empirical measures with $M \geqslant 2$ atoms:

$$\mu^i_1 = \frac{1}{M} \sum_{m=1}^{M} \delta_{y^i_m},$$

for some $y^i_m \in \mathbb{S}^{d-1}$. The input measures μ^i_0 are assumed to be absolutely continuous with respect to the normalized Lebesgue measure, in addition to satisfying (1.6). Under these assumptions, the following proof is very similar to that of Theorem 1.1—it avoids the packing step of Proposition 2.3, and avoids a direct application of Lemma 5.4, steps where the number of switches are hard to track. We split the proof in three steps.

Step 1. Disentanglement

As before, we first disentangle the measures using Proposition 3.1. Furthermore since the vector field in (1.4) is Lipschitz, absolute continuity of all measures is preserved over time, and thus we find flow maps

$$\Phi^t_{\theta_1} : \mathscr{P}_{\text{ac}}(\mathbb{S}^{d-1}) \mapsto \mathscr{P}_{\text{ac}}(\mathbb{S}^{d-1})$$

for $t \in [0, T/5]$, and

$$\Phi^t_{\theta_5} : \mathscr{P}(\mathbb{S}^{d-1}) \mapsto \mathscr{P}(\mathbb{S}^{d-1})$$

for $t \in [4T/5, T]$, induced by the characteristics of (1.4) and piecewise constant parameters having $O(d \cdot N)$ switches, which satisfy

$$\operatorname{conv}_g \operatorname{supp}\left(\Phi_{\theta_1}^{\frac{T}{5}}(\mu_0^i)\right) \cap \operatorname{conv}_g \operatorname{supp}\left(\Phi_{\theta_1}^{\frac{T}{5}}(\mu_0^j)\right) = \varnothing \qquad \text{if } i \neq j,$$

and

$$\operatorname{conv}_g \operatorname{supp}\left(\Phi_{\theta_5}^{T}(\mu_1^i)\right) \cap \operatorname{conv}_g \operatorname{supp}\left(\Phi_{\theta_5}^{T}(\mu_1^j)\right) = \varnothing \qquad \text{if } i \neq j.$$

We label the disentangled targets as

$$\Phi_{\theta_5}^T(\mu_1^i) = \frac{1}{M} \sum_{m=1}^{M} \delta_{\widetilde{y}_m^i}. \tag{5.8}$$

Step 2. Clustering

Let $\varepsilon_1 > 0$ be arbitrary and to be chosen later on. We first employ Proposition 2.1 to cluster the disentangled input measures: there exists a flow map

$$\Phi_{\theta_2}^t : \mathcal{P}_{\mathrm{ac}}(\mathbb{S}^{d-1}) \mapsto \mathcal{P}_{\mathrm{ac}}(\mathbb{S}^{d-1})$$

for $t \in [T/5, 2T/5]$, which satisfies

$$\operatorname{diam}\left(\operatorname{supp}\left(\left(\Phi_{\theta_2}^{\frac{2T}{5}} \circ \Phi_{\theta_2}^{\frac{T}{5}}\right)(\mu_0^i)\right)\right) \leqslant \varepsilon_1 \tag{5.9}$$

for all $i \in [N]$. Instead of using Lemma 5.4 to approximate arbitrary transport maps as done in Step 2 in the previous proof, we rather use Lemma C.3 recursively to reduce the problem to an ensemble matching of points. As a consequence of Step 1 and (5.9), there exists some $\kappa > 0$ such that

$$\inf_{\substack{x \in \operatorname{conv}_g \operatorname{supp}(\nu^i), \\ y \in \operatorname{conv}_g \operatorname{supp}(\nu^j), \\ i \neq j}} d_g(x, y) \geqslant 2\kappa, \tag{5.10}$$

where we set

$$\nu^i := \left(\Phi_{\theta_2}^{\frac{2T}{5}} \circ \Phi_{\theta_2}^{\frac{T}{5}}\right)(\mu_0^i).$$

We use the following.

Claim 1. *There exists some small enough $\varepsilon_1 > 0$ such that for all $i \in [N]$, the measures ν^i in (5.10) are such that there exist balls $B(x_m^i, r^i)$, for $m \in [M]$, satisfying*

1.

$$\nu^i\left(B(x_m^i, r^i) \setminus B(x_{m-1}^i, r^i)\right) = \frac{1}{M} \qquad \text{for } 2 \leqslant m \leqslant M,$$
$$\nu^i\left(B(x_1^i, r^i)\right) = \frac{1}{M}.$$

2. *For any* $m \in [M-1]$ *there exists* $z_m^i \in B(x_m^i, r^i)$ *such that*

$$z_m^i \notin B(x_{m'}^i, r^i) \qquad \text{for } m' \geqslant m+1.$$

3. *For* $j \neq i \in [N]$,

$$\nu^i \left(B(x_m^j, r^j) \right) = 0 \qquad \text{for all } m \in [M]. \tag{5.11}$$

We postpone the proof of Claim 1 to after the present one (see also Figure 7). Fix an arbitrary $i \in [N]$. Applying Lemma C.3 M times successively using the balls stemming from Claim 1 and z_m^i in place of ω, we obtain M Lipschitz-continuous invertible flow maps $\psi_m^i : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ of (4.1) corresponding to constant parameters, such that, setting

$$\Psi^i := \psi_M^i \circ \psi_{M-1}^i \circ \cdots \circ \psi_1^i,$$

because of (5.11), we have

$$\Psi^i_{\#} \nu^j = \nu^j \qquad \text{for } j \neq i \in [N], \tag{5.12}$$

as well as

$$\mathsf{W}_2 \left(\Psi^i_{\#} \nu^i, \alpha^i \right) \leqslant \varepsilon \tag{5.13}$$

where

$$\alpha^i = \frac{1}{M} \sum_{m=1}^{M} \delta_{z_m^i}.$$

Due to (5.12) and (5.13), the map

$$\Psi := \Psi^N \circ \Psi^{N-1} \circ \cdots \circ \Psi^1$$

is a flow map of (C.20) induced by parameters having $O(M \cdot N)$ switches, and satisfying

$$\mathsf{W}_2 \left(\Psi_{\#} \nu^i, \alpha^i \right) \leqslant \varepsilon$$

for all $i \in [N]$. All in all, the flow map

$$\Phi_{\theta_3}^{\frac{3T}{5}} : \mathcal{P}(\mathbb{S}^{d-1}) \mapsto \mathcal{P}(\mathbb{S}^{d-1})$$
$$\Phi_{\theta_3}^{\frac{3T}{5}}(\mu) = \Psi_{\#} \mu$$

is such that

$$\mathsf{W}_2 \left(\left(\Phi_{\theta_3}^{\frac{3T}{5}} \circ \Phi_{\theta_3}^{\frac{2T}{5}} \circ \Phi_{\theta_3}^{\frac{T}{5}} \right) \mu_0^i, \alpha^i \right) \leqslant \varepsilon \tag{5.14}$$

for all $i \in [N]$.

Step 3. Matching

We apply[10] Proposition 4.1 to

$$\left(z_m^i, \widetilde{y}_m^i\right) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \qquad \text{for } (i,m) \in [N] \times [M], \tag{$\mathfrak{D}$}$$

with $\widetilde{y}_m^i$ as in (5.8). This yields a flow map $\phi : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ of (4.1) induced by piecewise constant parameters with $O(M \cdot N)$ switches satisfying

$$\phi(z_m^i) = \widetilde{y}_m^i$$

for all $(i,m) \in [N] \times [M]$. Define

$$\Phi_{\theta_4}^{\frac{4T}{5}}(\mu) := \phi_{\#}\mu.$$

Using the triangle inequality, the definition of α^i in Step 2, and the continuity of the solution to (4.1) with respect the initial conditions and (5.14), we find

$$\mathsf{W}_2\left((\phi \circ \Psi)_{\#}\nu^i, \Phi_{\theta_5}^T(\mu_1^i)\right) \lesssim_{M,N} \varepsilon$$

for all $i \in [N]$ where the implicit constant is independent of ε. This yields

$$\mathsf{W}_2\left(\left(\Phi_{\theta_4}^{\frac{4T}{5}} \circ \Phi_{\theta_3}^{\frac{3T}{5}} \circ \Phi_{\theta_2}^{\frac{2T}{5}} \circ \Phi_{\theta_1}^{\frac{T}{5}}\right)(\mu_0^i), \Phi_{\theta_5}^T(\mu_1^i)\right) \lesssim_{N,M} \varepsilon$$

for all $i \in [N]$. The conclusion follows by applying the inverse of $\Phi_{\theta_5}^T$ as in the previous proof. Finally, pasting the parameters used in all of the steps above, the resulting number of switches is $O((d+M) \cdot N)$. □

Proof of Claim 1. We fix $i \in [N]$. Due to (5.9), we have

$$\operatorname{supp}(\nu^i) \subset B\left(x_M^i, C\varepsilon_1\right) \tag{5.15}$$

for some $x_M^i \in \operatorname{conv}_g \operatorname{supp}(\nu^i)$ and $C > 0$. Take ε_1 small enough so that

$$\kappa \geqslant 4C\varepsilon_1. \tag{5.16}$$

Take

$$x_1^i \in \partial B\left(x_M^i, \frac{\kappa}{2}\right).$$

Consider the minimizing geodesic $\gamma : [0,1] \to \mathbb{S}^{d-1}$ between x_1^i and x_M^i, and the function

$$\begin{aligned} f : [0,1] \times [0,\pi] &\mapsto [0,1] \\ (s,r) &\mapsto f(s,r) = \nu^i\left(B(\gamma(s), r)\right). \end{aligned}$$

Since ν^i is absolutely continuous, we have

[10] Should the assumptions of Proposition 4.1 not hold, we consider a slight perturbation of the target measures ($\mathsf{W}_2(\mu_1^i, \tilde{\mu}_1^i) \leqslant \varepsilon$).

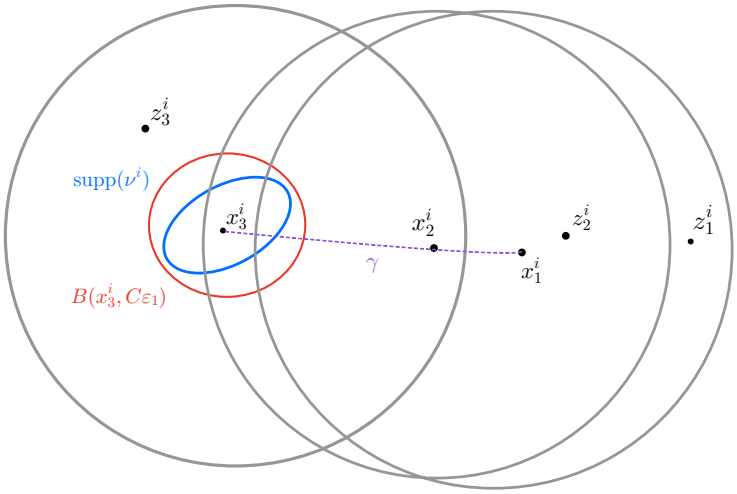

Figure 7: High-level overview of Claim 1: M balls partition the support of the absolutely continuous measure ν^i into M pieces of mass $\frac{1}{M}$. Here $M = 3$.

1. $f \in \mathscr{C}^0([0,1] \times [0,\pi]; [0,1])$.
2. $f(1,r) = 1$ for all $r \geqslant C\varepsilon_1$
3. $f(0,r) = 0$ for all $r \leqslant \kappa/2 - C\varepsilon_1$
4. $f(0,r) = 1$ for all $r \geqslant 1/2 + C\varepsilon_1$.

By continuity, there exists $r^i \in (\kappa/2 - C\varepsilon_1, \kappa/2 + C\varepsilon_1)$ such that

$$f(0, r^i) = \frac{1}{M}.$$

Furthermore, due to (5.15) and (5.10) we also have

$$\nu^j(B(\gamma(s), r^i)) = 0 \qquad \text{for } s \in [0,1],\ j \neq i \in [N].$$

Finally, note that $f(\cdot, r^i)$ is continuous and monotonically increasing provided $r^i \geqslant C\varepsilon_1$, which is guaranteed by (5.16). We can thus pick $\{s_m^i\}_{m=2}^{M-1} \subset (0,1)$ such that

$$f(s_m^i, r^i) = \frac{m}{M}.$$

Hence, the desired balls are

$$B\left(x_m^i = \gamma(s_m^i), r^i\right), \qquad \text{with } (s_1^i, s_M^i) = (0,1).$$

Finally, since for fixed i, all balls have the same radius, the existence of z_m^i is straightforward. □

5.2 Proof of Theorem 1.1

Proof of Theorem 1.1. The proof follows the same ideas as that of Theorem 1.2, but is significantly simpler since some steps can be omitted completely. Indeed, we can consider

$$\Phi^T_{\mathrm{fin}} := \Phi^T_{\theta_3} \circ \Phi^{\frac{2T}{3}}_{\theta_2} \circ \Phi^{\frac{T}{3}}_{\theta_1},$$

where

- $\Phi^t_{\theta_1} : \mathscr{P}(\mathbb{S}^{d-1}) \to \mathscr{P}(\mathbb{S}^{d-1})$ for $t \in [0, T/3]$ is the flow map provided by Proposition 3.1;
- $\Phi^t_{\theta_2} : \mathscr{P}(\mathbb{S}^{d-1}) \to \mathscr{P}(\mathbb{S}^{d-1})$ for $t \in [T/3, 2T/3]$ is the flow map provided by Proposition 2.1, which can be applied since $\Phi^{\frac{T}{3}}(\mu^i_0)$ are pairwise disjoint and supported in a single hemisphere for all $i \in [N]$;
- $\Phi^t_{\theta_3}$ for $t \in [2T/3, T]$ is the flow map provided by Proposition 4.1.

To conclude, we demonstrate how to derive the bound on the norm of the parameters θ. Recall that by rescaling time, bounding the final time horizon is equivalent to bounding the L^∞–norm of θ.

1. In the proof of Proposition 3.1, after tracking dependencies one sees that

   $$\|\theta_1\|_{L^\infty((0,T);\Theta)} \lesssim \frac{d \cdot N}{T}$$

   where the implicit constant depends only on the supports of the initial measures.

2. Once the measures are disentangled, we further cluster them before using Proposition 4.1. We thus quantify the convergence in Proposition 2.1, when $\boldsymbol{B} \equiv 0$. Specifically, by virtue of Proposition 2.2, we deduce that for every $i \in [N]$,

   $$\mathsf{W}_2\left(\mu^i(T_\delta), \delta_{x^i_0}\right) \leqslant \delta$$

   with $T_\delta = O(\log 1/\delta)$, which implies

   $$\|\theta_2\|_{L^\infty((0,T);\Theta)} \lesssim \log \frac{1}{\delta}.$$

3. Finally, we apply Proposition 4.1 for the ensemble of atoms x^i_0: since all measures are δ-close to $\delta_{x^i_0}$, we have

   $$\mathsf{W}_2\left(\mu^i(T), \delta_{y^i}\right) \leqslant e^{O(1)\cdot N \cdot T}\delta,$$

   at a cost

   $$\|\theta_3\|_{L^\infty((0,T);\Theta)} \lesssim \frac{N}{T}.$$

All in all,

$$\mathsf{W}_2\left(\mu^i(T), \delta_{y^i}\right) \leqslant \varepsilon,$$

with

$$\|\theta\|_{L^\infty((0,T);\Theta)} = O\left(\frac{d\cdot N}{T} + \log\frac{1}{\varepsilon}\right). \qquad \square$$

6 Complexity of disentanglement

6.1 Number of switches

In view of what precedes, we know that the Transformer can be used to disentangle the supports of N probability measures on $\mathbb{S}^{d-1}$ by using parameters with $O(d\cdot N)$ switches. The proof thereof relies on separating one probability measure from all the others in a successive manner, leading to the linear dependence in N. This dependence is very likely sub-optimal, and could be improved upon having a precise characterization of the ω-limit set for

$$\partial_t\mu(t) + \operatorname{div}\Big(\mathbf{P}^\perp_x(\mathscr{A}_{\beta\boldsymbol{B}}[\mu(t)](x))\mu(t)\Big) = 0 \qquad \text{in } \mathbb{R}_{\geqslant 0}\times\mathbb{S}^{d-1}$$

for $\boldsymbol{B}\in\mathscr{M}_{d\times d}(\mathbb{R})$ and $\beta\geqslant 0$. Specifically, if given $\mu_0^i\in\mathscr{P}(\mathbb{S}^{d-1})$, $i\in[N]$, we were to know that there exists a matrix $\boldsymbol{B}\in\mathscr{M}_{d\times d}(\mathbb{R})$ such that for all $i\in[N]$, there exists $z_i\in\mathbb{S}^{d-1}$, with $z_i\neq z_j$, $i\neq j$, such that

$$\lim_{T\to+\infty}\mathsf{W}_\infty\left(\mu^i(T),\delta_{z_i}\right) = 0,$$

then there would exist a time $T>0$ for which the measures are disentangled with *a single constant parameter*. However, we believe that characterizing z_i from μ_0^i is far from straightforward in general.

Example 6.1. *We provide an example of a small class of probability measures that can be simultaneously disentangled with a single constant parameter. Let $d=2$ and consider $\mu_0^i(\,\mathrm{d}x) = \frac{1}{|\mathscr{B}_i|}1_{\mathscr{B}_i}\,\mathrm{d}x$, with $\mathscr{B}_i$ being connected, $\mathscr{B}_i\neq\mathbb{S}^1$, and the barycenters of all $\mathscr{B}_i$ being different. When $\boldsymbol{B}=0$ the barycenter is preserved along the flow by symmetry. This implies that each measure clusters to the barycenter of $\mathscr{B}_i$.*

Furthermore, one has to note that in the construction we presented, there is a bottleneck (in terms of number of switches) in Proposition 2.3, which solely uses the perceptron component of the vector field. Although Proposition 2.3 can be parallelized by adding width to the perceptron, one could inquire if nonlinear effects of the self-attention mechanism could be of help in obtaining the same property with even less switches—avenues in this direction include using the *dynamic metastability* property shown in [GKPR24] (see also [BPA24]).

6.2 Fastest disentanglement

One can view the number of switches in the parameters as a natural measure of complexity for achieving a particular task (we focus on disentanglement for simplicity). However there are other relevant notions of complexity that ought to be kept track of, such as the *effective depth* of the architecture, namely the time horizon $T > 0$ needed for achieving disentanglement with $\boldsymbol{V}$ and $\boldsymbol{W}$ rescaled so that the maximal velocity of the characteristics of (1.4) is bounded by 1. Indeed, we can always rescale the dynamics so that any property can be achieved in the time we want, simply because we can absorb the change of time-scale in the parameters. Bounding the speed of the particles gives us a judicious way of comparing different choices of parameters, which could help in understanding self-attention across different parameters.

We now propose an example which demonstrates that the shortest time horizon T in which disentanglement occurs is necessarily achieved by using self-attention with $\boldsymbol{B} \not\equiv 0$.

Suppose $\mu_0^1, \mu_0^2 \in \mathcal{P}(\mathbb{S}^1 \cap (\mathbb{R}_{\geqslant 0})^2)$[11] such that $\operatorname{supp}(\mu_0^i)$ is connected for $i = 1, 2$ and

$$\operatorname{supp}(\mu_0^1) \cap \operatorname{supp}(\mu_0^2) \neq \varnothing.$$

Set

$$x^{i+} := \operatorname*{arg\,max}_{x \in \operatorname{supp}(\mu_0^i)} \langle x, e_2 \rangle, \qquad x^{i-} := \operatorname*{arg\,min}_{x \in \operatorname{supp}(\mu_0^i)} \langle x, e_2 \rangle,$$

and assume that $x^{1+} > x^{2+}$ and $x^{1-} > x^{2-}$. Consider

$$\begin{cases} \partial_t \mu^1 + \operatorname{div}\left(\mathbf{P}_x^{\perp}(x^{1+})\mu^1\right) = 0 \\ \mu^1(0) = \mu_0^1, \end{cases} \qquad \begin{cases} \partial_t \mu^2 + \operatorname{div}\left(\mathbf{P}_x^{\perp}(x^{2-})\mu^2\right) = 0 \\ \mu^2(0) = \mu_0^2. \end{cases} \tag{6.1}$$

The vector fields in (6.1) can be achieved by considering the self-attention mechanism with a *hardmax* nonlinearity. Indeed, pick

$$\boldsymbol{B} = \alpha\alpha^\top$$

where $\alpha \in \mathbb{S}^1 \cap (\mathbb{R}_{\leqslant 0})^2$ satisfies

$$\operatorname*{arg\,max}_{y \in \operatorname{supp}(\mu_0^1)} \langle \boldsymbol{B}x, y \rangle = \operatorname*{arg\,max}_{y \in \operatorname{supp}(\mu_0^1)} \langle \alpha, y \rangle = x^{1+},$$

and similarly

$$\operatorname*{arg\,max}_{y \in \operatorname{supp}(\mu_0^2)} \langle \boldsymbol{B}x, y \rangle = x^{2-}.$$

One such vector is

$$\alpha = -\frac{x^{1+} + x^{2-}}{2\|x^{1+} + x^{2-}\|}.$$

[11] Similar examples can be constructed in higher dimensions with appropriate assumptions on the supports.

The respective solutions to (6.1) disentangle the supports of μ_0^1 and μ_0^2 faster than those for the equation considered with $\mathscr{A}_{\boldsymbol{B}}[\mu(t)]$ for any other constant $\boldsymbol{B}$. Indeed, the measures are disentangled in time T if the trajectories of

$$\begin{cases} \dot{x}^-(t) = \mathbf{P}^{\perp}_{x^-(t)}\left(\mathscr{A}_{\boldsymbol{B}}[\mu^1(t)](x^-(t))\right) \\ x^-(0) = x^{1-}, \end{cases} \qquad \begin{cases} \dot{x}^+(t) = \mathbf{P}^{\perp}_{x^+(t)}\left(\mathscr{A}_{\boldsymbol{B}}[\mu^2(t)](x^+(t))\right) \\ x^+(0) = x^{2+}, \end{cases}$$

satisfy

$$\langle x^+(T), e_2\rangle > \langle x^-(T), e_2\rangle.$$

Since $\mathscr{A}_{\boldsymbol{B}}[\mu^2(t)](x)$ points inward the convex hull of $\mathrm{supp}(\mu^2(t))$ for any $\boldsymbol{B}$, one has

$$\left\|\mathbf{P}^{\perp}_{x^+(t)}\left(\mathscr{A}_{\boldsymbol{B}}[\mu^2(t)](x^+(t))\right)\right\| \leqslant \left\|\mathbf{P}^{\perp}_{x^+(t)}(x^{2-})\right\|,$$

and similarly,

$$\left\|\mathbf{P}^{\perp}_{x^-(t)}\left(\mathscr{A}_{\boldsymbol{B}}[\mu^1(t)](x^-(t))\right)\right\| \leqslant \left\|\mathbf{P}^{\perp}_{x^-(t)}(x^{1+})\right\|.$$

Therefore, for any $\boldsymbol{B}$, disentanglement will necessarily be achieved in a greater time than the time needed using (6.1).

In that regard, it is natural to define the *hardmax* dynamics. Consider the Cauchy problem

$$\begin{cases} \partial_t \mu(t) + \mathrm{div}\left(\mathbf{P}^{\perp}_{x}\left(\int y \mu_{\boldsymbol{B},x}(t,\,\mathrm{d}y)\right)\mu(t)\right) = 0 & \text{on } \mathbb{R}_{\geqslant 0}\times\mathbb{S}^{d-1} \\ \mu(0) = \mu_0 & \text{on } \mathbb{S}^{d-1}, \end{cases} \tag{6.2}$$

where $\mu_{\boldsymbol{B},x}(\,\mathrm{d}y)$ denotes the conditional probability at the $\arg\max$ of the map $\mathrm{supp}(\mu) \ni y \to \langle \boldsymbol{B}x, y\rangle$, namely

$$\mu_{\boldsymbol{B},x}(\,\mathrm{d}y) := \mu\left(\mathrm{d}y \,\middle|\, y \in \underset{x'\in\mathrm{supp}(\mu)}{\arg\max}\,\langle \boldsymbol{B}x, x'\rangle\right).$$

The vector field in (6.2) is not continuous in x in general, and as such, uniqueness of solutions may not be expected. The example presented above is a particular case in which the $\arg\max$ is constant for all x, and therefore uniqueness is not an issue.

The *hardmax* vector field defined above is therefore a natural alternative to self-attention. Yet, even-though for $\mu \in \mathscr{P}_{\mathrm{ac}}(\mathbb{S}^{d-1})$ and $\boldsymbol{B} \in \mathscr{M}_{d\times d}(\mathbb{R})$, one has

$$\lim_{\beta\to+\infty} \mathscr{A}_{\beta\boldsymbol{B}}[\mu](x) = \underset{y\in\mathrm{supp}(\mu)}{\arg\max}\,\langle \boldsymbol{B}x, y\rangle,$$

the following question is open.

Problem 1. *Given $\boldsymbol{B} \in \mathscr{M}_{d\times d}(\mathbb{R})$ and $\mu_0 \in \mathscr{P}(\mathbb{S}^{d-1})$, does the solution to*

$$\begin{cases} \partial_t \mu^\beta(t) + \mathrm{div}\left(\mathbf{P}^{\perp}_{x}\left(\mathscr{A}_{\beta\boldsymbol{B}}\left[\mu^\beta(t)\right]\right)\mu^\beta(t)\right) = 0 & \text{on } \mathbb{R}_{\geqslant 0}\times\mathbb{S}^{d-1} \\ \mu^\beta(0) = \mu_0 & \text{on } \mathbb{S}^{d-1}, \end{cases}$$

converge, in an appropriate sense, to an appropriate solution to (6.2), as $\beta \to +\infty$?

This is a *singular perturbation* limit.

All in all, one can speculate that the fastest disentanglement problem among vector fields consisting of parametrized attention mechanisms or hardmax ones, comes precisely from the latter, namely (6.2). It would however be of interest to fully understand the following problem.

Problem 2 (Fastest disentanglement). *Given $\mu_0^i \in \mathcal{P}(\mathbb{S}^{d-1})$, $i \in [N]$,*

$$\text{minimize} \quad T$$

among all $\boldsymbol{B} \in L^\infty((0,T); \mathscr{M}_{d\times d}(\mathbb{R}))$, subject to

$$\operatorname{supp}(\mu^i(T)) \cap \operatorname{supp}(\mu^j(T)) = \varnothing \qquad \text{if } i \neq j,$$

and

$$\begin{cases} \partial_t \mu^i(t) + \operatorname{div}\left(\mathbf{P}_x^\perp\left(\mathscr{A}_{\boldsymbol{B}}[\mu^i(t)]\right)\mu^i(t)\right) = 0 & \text{on } [0,T]\times\mathbb{S}^{d-1} \\ \mu^i(0) = \mu_0^i & \text{on } \mathbb{S}^{d-1}. \end{cases}$$

Remark 6.2 (Minimum-exit-time problem). *It is worth noting that the fastest disentanglement problem has a link in spirit with the* minimum-exit-time problem *(motion with constant velocity). Indeed, for disentanglement to hold, every particle in the support of one measure needs to exit the supports of the other measures. Typically, the minimum-exit-time problem has as input a bounded domain $\Omega \subset \mathbb{R}^d$ and $x_0 \in \Omega$, and one solves*

$$\text{minimize} \quad T$$

among all $\alpha \in L^\infty((0,T);\mathbb{R}^d)$ with $\|\alpha(t)\| = 1$ for $t \in [0,T]$, subject to

$$\begin{cases} \dot{x}(t) = \alpha(t) & \text{in } [0,T], \\ x(0) = x_0, \\ x(T) \in \partial\Omega. \end{cases}$$

We recall ([BD97][12]) that the solution to this problem is $\alpha^*(t) = \nabla u(x(t))$, where u solves the eikonal equation

$$\begin{cases} \|\nabla u\| = 1 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

A Deriving the model

A.1 From the code to a model

Our starting point in writing down a mathematical model for a Transformer is the following code snippet:

[12] The interested reader is also referred to [KS09] for an exquisite introduction.

```
123 def block(x, scope, *, past, hparams):
124     with tf.variable_scope(scope):
125         nx = x.shape[-1].value
126         a, present = attn(norm(x, 'ln_1'), 'attn', nx, past=past, hparams=hparams)
127         x = x + a
128         m = mlp(norm(x, 'ln_2'), 'mlp', nx*4, hparams=hparams)
129         x = x + m
130         return x, present
```

Figure 8: Lines $123 - 130$ in the source code of OpenAI's GPT-2, freely available online at https://github.com/openai/gpt-2/blob/master/src/model.py.

Figure 8 shows a single layer of a practical implementation of a Transformer. Combining this code with transparent models written in [LLH+20, SABP22], we write the full architecture in mathematical symbols. Our model is not exactly the same as the one used in GPT-2, as it differs in two ways:

- We focus on *encoder* models, namely those in which self-attention is defined as in (1.3). Models such as GPT are *decoder* models: they are causal as they use a masked self-attention mechanism. These models are not addressed herein. It is worth noting that the equation is remarkably similar, and the probability flow interpretation persists—see [CAP24].
- We use *post-layer normalization*, instead of *pre-layer normalization* as done in Figure 8, which consists in simply permuting the order of the `norm` and `attn`, and `norm` and `mlp` operations. This is a minor modeling detail.

Our setting is however exactly that used in the celebrated BERT model [DCLT19] (see [PH22, Algorithm 9] as well) and vision Transformers [DBK+21].

We proceed in writing the full Transformer architecture. Suppose we are given a sequence of initial particles (tokens)

$$x^0 = (x_1^0, \dots, x_n^0) := (x_1(0), \dots, x_n(0)) \in (\mathbb{S}^{d-1})^n.$$

At layer $t \geqslant 0$, given parameters $\boldsymbol{V}^t, \boldsymbol{B}^t, \boldsymbol{W}^t, \boldsymbol{U}^t, b^t$ the Transformer processes

the token x_i^t via[13][14]

$$y_i^t = x_i^t + \triangle t\ \boldsymbol{V}^t \sum_{j=1}^{n} \frac{e^{\langle \boldsymbol{B}^t x_i^t, x_j^t \rangle}}{\sum_{k=1}^{n} e^{\langle \boldsymbol{B}^t x_i^t, x_k^t \rangle}} x_j^t,$$

$$x_i^{t+1} = \frac{y_i^t + \triangle t\ \boldsymbol{W}^t \left(\boldsymbol{U}^t y_i^t + b^t\right)_+}{\left\| y_i^t + \triangle t\ \boldsymbol{W}^t \left(\boldsymbol{U}^t y_i^t + b^t\right)_+ \right\|}.$$

A.2 The differential equation

Although classical, for the sake of clarity and completeness, we briefly sketch the derivation of the continuous-time model from the discrete-time scheme written above. Setting

$$\mathscr{A}_{\boldsymbol{B}^t}[\mu](x_i^t) = \sum_{j=1}^{n} \frac{e^{\langle \boldsymbol{B}^t x_i^t, x_j^t \rangle}}{\sum_{k=1}^{n} e^{\langle \boldsymbol{B}^t x_i^t, x_k^t \rangle}} x_j^t,$$

and

$$f_{\theta^t}^{\triangle t}(x_i^t) := \boldsymbol{V}^t \mathscr{A}_{\boldsymbol{B}^t}[\mu](x_i^t) + \boldsymbol{W}^t \left(\boldsymbol{U}^t \left(x_i^t + \triangle t\ \boldsymbol{V}^t \mathscr{A}_{\boldsymbol{B}^t}[\mu](x_i^t)\right) + b^t\right)_+,$$

we can rewrite the above scheme as

$$x_i^{t+1} = \frac{x_i^t + \triangle t\, f_{\theta^t}^{\triangle t}(x_i^t)}{\left\| x_i^t + \triangle t\, f_{\theta^t}^{\triangle t}(x_i^t) \right\|}.$$

We Taylor-expand the denominator as

$$\left\| x_i^t + \triangle t\, f_{\theta^t}^{\triangle t}(x_i^t) \right\| = \left\| x_i^t \right\| + \triangle t \left\langle x_i^t, f_{\theta^t}^{\triangle t=0}(x_i^t) \right\rangle + O\left((\triangle t)^2\right),$$

thus

$$\frac{1}{\left\| x_i^t + f_{\theta^t}^{\triangle t}(x_i^t) \right\|} = \frac{1}{\|x_i^t\|} - \triangle t \left\langle \frac{x_i^t}{\|x_i^t\|}, f_{\theta^t}^{\triangle t=0}(x_i^t) \right\rangle + O\left((\triangle t)^2\right).$$

[13] Strictly speaking, $\triangle t = 1$ in practical implementations. We can rescale the multiplicative parameters $\boldsymbol{V}^t$, $\boldsymbol{W}^t$ to recover the time-step $\triangle t$ which we make small to derive a differential equation. One ought to be wary about the practical validity of such approximations—see [SAP22, MWSB24].

[14] In truth, looking closely at the code, one sees that even in the equation for y_i^t, one ought to divide the right-hand side by its Euclidean norm. We choose not to do this so as to derive a "cleaner" equation. Since this normalization is parametrized as well, we can choose the parameters so that it doesn't appear anyway.

Consequently,

$$x_i^{t+1} = \left(1 - \triangle t \left\langle x_i^t, f_{\theta^t}^{\triangle t=0}(x_i^t)\right\rangle + O\left((\triangle t)^2\right)\right) \cdot \left(x_i^t + \triangle t \left\langle x_i^t, f_{\theta^t}^{\triangle t=0}(x_i^t)\right\rangle + O\left((\triangle t)^2\right)\right),$$

and expanding the product yields

$$x_i^{t+1} = x_i^t + \triangle t\, \mathbf{P}_{x_i^t}^{\perp}\left(f_{\theta^t}^{\triangle t=0}(x_i^t)\right) + O\left((\triangle t)^2\right).$$

Letting $\triangle t \to 0$ we find the desired equation.

B On condition (1.6)

Lemma B.1. *Suppose $\nu_0^i \in \mathcal{P}(\mathbb{S}^{d-1})$, $i \in [N]$. Assume that for every $\eta > 0$ there exists a Lipschitz-continuous and invertible $\Phi_\eta : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ such that*

$$\left(\Phi_{\eta\#}\nu_0^i\right)(\mathbb{Q}_1^{d-1}) = 1 - \eta \tag{B.1}$$

for all $i \in [N]$. Then for all $i \in [N]$ there exists $\mu_0^i \in \mathcal{P}(\mathbb{Q}_1^{d-1})$ and a universal numerical constant $C > 0$ such that

$$\mathsf{W}_2\left(\Phi_{\eta\#}\nu_0^i, \mu_0^i\right) \leqslant C\eta.$$

Proof of Lemma B.1. For every $A \subset \mathbb{S}^{d-1}$ consider

$$\mu_0^i(A) = \nu_0^i(A \cap \mathbb{Q}_1^{d-1}) + \nu_0^i\left(\mathbb{S}^{d-1} \setminus \mathbb{Q}_1^{d-1}\right)\delta_{x_0^i}(A)$$

with $x_0^i \in \mathbb{Q}_1^{d-1}$. □

Lemma 3.2 provides a map Φ_η that ensures (B.1). By virtue of Lemma B.1, we can extend Theorem 1.2 to the setting of measures whose support fill $\mathbb{S}^{d-1}$, namely, the assumption of having a point $\omega \notin \bigcup_i \mathrm{supp}(\mu_0^i)$ can be removed. The result then follows by a continuity argument: we apply Theorem 1.2 to the measures μ_0^i provided by Lemma B.1 to obtain

$$\mathsf{W}_2\left(\mu^i(T), \nu^i(T)\right) \lesssim_T \mathsf{W}_2\left(\mu_0^i, \Phi_\eta\nu_0^i\right) \lesssim_T \eta,$$

where $\mu^i(t)$ is the solution to (1.4) given by Theorem 1.2 with initial data μ_0^i, and $\nu(t)$ is the solution to (1.4) with initial data $(\Phi_\eta)_\#\nu_0^i$. On the other hand, we can simply approximate the targets μ_1^i by measures that directly satisfy (1.6).

C Technical proofs

C.1 W_∞-stability

Lemma C.1. *Suppose $\mu_0, \nu_0 \in \mathcal{P}(\mathbb{S}^{d-1})$, $\boldsymbol{B} \in \mathcal{M}_{d\times d}(\mathbb{R})$. Let μ and ν denote the unique solutions to (2.1) on $\mathbb{R}_{\geqslant 0}\times\mathbb{S}^{d-1}$ corresponding to data μ_0 and ν_0 respectively. Then there exist a couple of constants $M_0 \geqslant 1$ and $M_1 > 0$ depending only on $\boldsymbol{B}$ such that*

$$\mathsf{W}_\infty(\mu(t), \nu(t)) \leqslant M_0 e^{e^{M_1 t}}\, \mathsf{W}_\infty(\mu_0, \nu_0)$$

for all $t \geqslant 0$.

Proof of Lemma C.1. Fix $p \geqslant 1$. Let $\gamma_0 \in \mathcal{P}(\mathbb{S}^{d-1}\times\mathbb{S}^{d-1})$ be the optimal transport plan between μ_0 and ν_0 given by solving

$$\mathsf{W}_p(\mu_0, \nu_0)^p = \inf_{\gamma\in\mathcal{C}(\mu_0,\nu_0)} \iint d_g(x,y)^p \gamma(\mathrm{d}x, \mathrm{d}y).$$

Consider $\Lambda^t : \mathbb{S}^{d-1}\times\mathbb{S}^{d-1} \to \mathbb{S}^{d-1}\times\mathbb{S}^{d-1}$ defined as

$$\Lambda^t(x,y) = \left(\Phi^t(x), \Psi^t(y)\right),$$

where $\Phi^t : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ and $\Psi^t : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ are the flow maps induced by (2.1) with data μ_0 and ν_0. Namely, setting $v[\nu](x) = \mathbf{P}_x^\perp \mathcal{A}_{\boldsymbol{B}}[\nu](x)$, Φ^t solves

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}\tau}\Phi^\tau(x) = v[\mu(\tau)](\Phi^\tau(x)) & \tau\in[0,t] \\ \Phi^0(x) = x, \end{cases}$$

and Ψ^t solves

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}\tau}\Psi^\tau(x) = v[\nu(\tau)](\Psi^\tau(x)) & \tau\in[0,t] \\ \Psi^0(x) = x. \end{cases}$$

We have $\mu(t) = \Phi^t_\#\mu_0$ and $\nu(t) = \Psi^t_\#\nu_0$, and $\Lambda^t_\#\gamma$ is a transport plan between $\mu(t)$ and $\nu(t)$. Denote by $\Pi^x : \mathbb{S}^{d-1}\times\mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ (resp. Π^y) the projection operator onto the first (resp. second) variable. Note that

$$\left(\Pi^x \circ \Lambda^t\right)(x,y) = \Phi^t(x) = \left(\Phi^t \circ \Pi^x\right)(x,y).$$

Therefore

$$\Pi^x_\#\left(\Lambda^t_\#\gamma_0\right) = \left(\Pi^x\circ\Lambda^t\right)_\#\gamma_0 = \left(\Phi^t\circ\Pi^x\right)_\#\gamma_0 = \Phi^t_\#\mu_0.$$

(Analogous computations follow for $\Pi^y_\#(\Lambda^t_\#\gamma_0)$.) Therefore,

$$\begin{aligned}\mathsf{W}_p(\mu(t),\nu(t)) &\leqslant \left(\iint d_g(x,y)^p \Lambda^t_\#\gamma_0(\mathrm{d}x, \mathrm{d}y)\right)^{\frac{1}{p}} \\ &= \left(\iint d_g(\Phi^t(x),\Psi^t(y))^p \gamma_0(\mathrm{d}x,\mathrm{d}y)\right)^{\frac{1}{p}}.\end{aligned}$$

We thus have

$$\mathsf{W}_p(\mu(t),\nu(t)) \leqslant \left(\iint d_g(\Phi^t(x),\Phi^t(y))^p \gamma_0(\,\mathrm{d}x,\,\mathrm{d}y)\right)^{\frac{1}{p}} + \left(\iint d_g(\Phi^t(y),\Psi^t(y))^p \gamma_0(\,\mathrm{d}x,\,\mathrm{d}y)\right)^{\frac{1}{p}}. \tag{C.1}$$

We analyze the integrals above separately, starting with the first one, since $v[\mu](\cdot)$ is Lipschitz, by the Grönwall Lemma, the flow map Φ^t is Lipschitz as well:

$$\iint d_g(\Phi^t(x),\Phi^t(y))^p \gamma_0(\,\mathrm{d}x,\,\mathrm{d}y) \leqslant K_1^p e^{pK_0 t} \iint d_g(x,y)^p \gamma_0(\,\mathrm{d}x,\,\mathrm{d}y) \tag{C.2}$$

where $K_1 \geqslant 1$ and $K_0 > 0$ depend only on $\boldsymbol{B}$. For the second integral, setting $f(t,y) = \|\Phi^t(y) - \Psi^t(y)\|$, since both flows are actually $\mathscr{C}^\infty$ on $\mathbb{R}_{\geqslant 0} \times \mathbb{S}^{d-1}$, for a.e. t and y we can differentiate to find

$$\begin{aligned}
\partial_t f(t,y) &\leqslant \left\| v[\mu(t)](\Phi^t(y)) - v[\nu(t)](\Psi^t(y)) \right\| \\
&\leqslant \left\| v[\mu(t)](\Phi^t(y)) - v[\nu(t)](\Phi^t(y)) \right\| + \left\| v[\nu(t)](\Phi^t(y)) - v[\nu(t)](\Psi^t(y)) \right\| \\
&\leqslant \left\| v[\mu(t)] - v[\nu(t)] \right\|_{L^\infty(\mathbb{S}^{d-1})} + C_0 d_g(\Phi^t(y),\Psi^t(y)) \\
&\leqslant C_1 \mathsf{W}_p(\mu(t),\nu(t)) + C_0 f(t,y),
\end{aligned}$$

where $C_1, C_0 \geqslant 0$ depend only on $\boldsymbol{B}$ as shown in [GLPR24, CAP24]. Applying Grönwall's inequality, and noting that $f(0,\cdot) \equiv 0$, we obtain

$$f(t,y) \leqslant \int_0^t e^{C_1(t-s)} \mathsf{W}_p(\mu(s),\nu(s))\,\mathrm{d}s.$$

Therefore,

$$\iint d_g(\Phi^t(y),\Psi^t(y))^p \gamma_0(\,\mathrm{d}x,\,\mathrm{d}y) \leqslant C_2^p \left(\int_0^t e^{C_1(t-s)} \mathsf{W}_p(\mu(s),\nu(s))\,\mathrm{d}s\right)^p. \tag{C.3}$$

where $C_2 > 0$ is such $d_g(z_1,z_2) \leqslant C_2\|z_1 - z_2\|$ for all $z_1, z_2 \in \mathbb{S}^{d-1}$, and thus, independent of p. Gathering (C.2) and (C.3) in (C.1), we find

$$\mathsf{W}_p(\mu(t),\nu(t)) \leqslant K_1 e^{K_0 t} \mathsf{W}_p(\mu_0,\nu_0) + C_2 \int_0^t e^{C_1(t-s)} \mathsf{W}_p(\mu(s),\nu(s))\,\mathrm{d}s.$$

Employing the integral form of Grönwall's inequality, we end up with

$$\mathsf{W}_p(\mu(t),\nu(t)) \leqslant M_0 e^{e^{M_1 t}} \mathsf{W}_p(\mu_0,\nu_0)$$

for $M_0 \geqslant 1$ and $M_1 > 0$ depending only on $\boldsymbol{B}$. Since none of the constants depend on p, we can let $p \to +\infty$ to conclude. □

C.2 Proof of Proposition 2.2

Proof of Proposition 2.2. Note that $\|\mathbb{E}_{\mu(t)}[x]\| > 0$ since $\operatorname{supp}(\mu_0)$ is contained in a hemisphere. Set

$$\gamma(t) := \frac{\mathbb{E}_{\mu(t)}[x]}{\|\mathbb{E}_{\mu(t)}[x]\|}.$$

and consider

$$\mathsf{V}[\mu(t)](t) := \int \|x - \gamma(t)\|^2 \mu(t, \mathrm{d}x).$$

We first show that $\mathsf{V}[\mu(t)](t)$ decays exponentially fast. Integration by parts yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{V}[\mu(t)](t) = 2\|\mathbb{E}_{\mu(t)}[x]\| \left(-1 + \int \langle x, \gamma(t)\rangle^2 \mu(t, \mathrm{d}x)\right) \\ + 2\int \langle x - \gamma(t), -\dot\gamma(t)\rangle \mu(t, \mathrm{d}x). \end{aligned} \tag{C.4}$$

Owing to Proposition 2.1, for any $\varepsilon_0 > 0$ there is some $t_0 > 0$ such that for all $t \geqslant t_0$, we have

$$d_g(x, y) \leqslant \varepsilon_0 \qquad \text{for all } x, y \in \operatorname{supp}(\mu(t_0)). \tag{C.5}$$

We treat the terms in (C.4) separately, starting with

$$\begin{aligned}-1 + \int \langle x, \gamma(t)\rangle^2 \mu(t, \mathrm{d}x) &= -1 + \int \cos^2\left(d_g(x, \gamma)\right) \mu(t, \mathrm{d}x) \\ &= -1 + \int \left(1 - \frac{1}{2} d_g(x, \gamma)^2 + O\left(\varepsilon_0^4\right)\right)^2 \mu(t, \mathrm{d}x) \\ &= -\int d_g(x, \gamma)^2 \mu(t, \mathrm{d}x) + O\left(\varepsilon_0^4\right). \end{aligned} \tag{C.6}$$

On the other hand,

$$\begin{aligned}\dot\gamma(t) &= \int \mathsf{P}_x^\perp(\gamma(t))\mu(t, \mathrm{d}x) - \frac{\mathbb{E}_{\mu(t)}[x]}{\|\mathbb{E}_{\mu(t)}[x]\|^2} \left\langle \frac{\mathbb{E}_{\mu(t)}[x]}{\|\mathbb{E}_{\mu(t)}[x]\|}, \int y \partial_t \mu(t, \mathrm{d}y) \right\rangle \\ &= \int \mathsf{P}_x^\perp(\gamma(t))\mu(t, \mathrm{d}x) - \gamma(t)\left(1 - \int \langle \gamma(t), x\rangle^2 \mu(t, \mathrm{d}x)\right). \end{aligned} \tag{C.7}$$

Plugging (C.7) into the second term in (C.4), we find

$$\begin{aligned}\int \langle y - \gamma(t), \dot\gamma(t)\rangle \mu(t, \mathrm{d}y) &= \int \langle y, \dot\gamma(t)\rangle \mu(t, \mathrm{d}y) - \frac{1}{2}\int \frac{\mathrm{d}}{\mathrm{d}t}\langle \gamma(t), \gamma(t)\rangle \mu(t, \mathrm{d}y) \\ &= \int \langle y, \dot\gamma(t)\rangle \mu(t, \mathrm{d}y). \end{aligned} \tag{C.8}$$

Expanding the inner product in (C.8), we obtain

$$\begin{aligned}
\langle y, \dot{\gamma}(t)\rangle &= \int \left\langle y, \mathbf{P}_x^{\perp}\gamma(t)\right\rangle \mu(t,\, \mathrm{d}x) - \left(1 - \int \langle \gamma(t), x\rangle^2 \mu(t,\, \mathrm{d}x)\right) \langle y, \gamma(t)\rangle \\
&= \int \left(\langle y, \gamma(t)\rangle - \langle \gamma(t), x\rangle\langle x, y\rangle\right) \mu(t,\, \mathrm{d}x) \\
&\qquad\qquad - \left(1 - \int \langle \gamma(t), x\rangle^2 \mu(t,\, \mathrm{d}x)\right) \langle y, \gamma(t)\rangle \\
&= -\int \langle \gamma(t), x\rangle\langle x, y\rangle \mu(t,\, \mathrm{d}x) + \left(\int \langle \gamma(t), x\rangle^2 \mu(t,\, \mathrm{d}x)\right) \langle y, \gamma(t)\rangle \\
&= -\int \cos(d_g(\gamma(t), x)) \cos(d_g(x, y)) \mu(t,\, \mathrm{d}x) \\
&\qquad + \left(\int \cos^2(d_g(\gamma(t), x)) \mu(t,\, \mathrm{d}x)\right) \cos(d_g(y, \gamma(t))). \qquad \text{(C.9)}
\end{aligned}$$

In view of (C.5), we Taylor expand to find

$$\begin{aligned}
&\int \cos(d_g(\gamma(t), x)) \cos(d_g(x, y)) \mu(t,\, \mathrm{d}x) \\
&= \int \left(1 - \frac{d_g(\gamma(t), x)^2}{2} + O\left(\varepsilon_0^4\right)\right) \left(1 - \frac{d_g(y, x)^2}{2} + O\left(\varepsilon_0^4\right)\right) \mu(t,\, \mathrm{d}x) \\
&= \int \left(1 - \frac{d_g(\gamma(t), x)^2}{2} - \frac{d_g(y, x)^2}{2}\right) \mu(t,\, \mathrm{d}x) + O\left(\varepsilon_0^4\right), \qquad \text{(C.10)}
\end{aligned}$$

and

$$\begin{aligned}
&\left(\int \cos^2(d_g(\gamma(t), x)) \mu(t,\, \mathrm{d}x)\right) \cos(d_g(y, \gamma(t))) \\
&= \left(\int \left(1 - d_g(\gamma(t), x)^2 + O\left(\varepsilon_0^4\right)\right) \mu(t,\, \mathrm{d}x)\right) \left(1 - \frac{d_g(y, \gamma(t))^2}{2} + O\left(\varepsilon_0^4\right)\right) \\
&= \left(\int (1 - d_g(\gamma(t), x)^2) \mu(t,\, \mathrm{d}x)\right) - \frac{d_g(y, \gamma(t))^2}{2} + O\left(\varepsilon_0^4\right). \qquad \text{(C.11)}
\end{aligned}$$

Combining (C.10) and (C.11) in (C.9) we obtain

$$\begin{aligned}
\langle y, \dot{\gamma}(t)\rangle &= \left(\int (1 - d_g(\gamma(t), x)^2) \mu(t,\, \mathrm{d}x)\right) - \frac{d_g(y, \gamma(t))^2}{2} \\
&\qquad - \int \left(1 - \frac{d_g(\gamma(t), x)^2}{2} - \frac{d_g(y, x)^2}{2}\right) \mu(t,\, \mathrm{d}x) + O\left(\varepsilon_0^4\right) \\
&= -\int d_g(\gamma(t), x)^2 \mu(t,\, \mathrm{d}x) - \frac{d_g(y, \gamma(t))^2}{2} \\
&\qquad + \frac{1}{2} \int \left(d_g(\gamma(t), x)^2 + d_g(y, x)^2\right) \mu(t,\, \mathrm{d}x) + O\left(\varepsilon_0^4\right).
\end{aligned}$$

Going back to (C.8), using (C.6) we gather that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{V}[\mu(t)](t) &= 2\|\mathbb{E}_{\mu(t)}[x]\|\left(-\int d_g(x,\gamma(t))^2\mu(t,\,\mathrm{d}x)\right) - 2\int\langle y,\dot\gamma(t)\rangle\mu(t,\,\mathrm{d}y)\\ &= -2\int d_g(x,\gamma(t))^2\mu(t,\,\mathrm{d}x) + O\left(\varepsilon_0^3\right)\\ &\quad -\int d_g(\gamma(t),x)^2\mu(t,\,\mathrm{d}x) - \iint d_g(y,x)^2\mu(t,\,\mathrm{d}x)\mu(t,\,\mathrm{d}y)\\ &\quad +2\int d_g(\gamma(t),x)^2\mu(t,\,\mathrm{d}x) + \int d_g(y,\gamma(t))^2\mu(t,\,\mathrm{d}y),\end{aligned} \tag{C.12}$$

where we used (C.5) to ensure $\|\mathbb{E}_{\mu(t)}[x]\| = 1-O(\varepsilon_0)$. Combining terms in (C.12),

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{V}[\mu(t)](t) &= \int d_g(y,\gamma(t))^2\mu(t,\,\mathrm{d}y) + O\left(\varepsilon_0^3\right)\\ &\quad -\int d_g(\gamma(t),x)^2\mu(t,\,\mathrm{d}x) - \iint d_g(y,x)^2\mu(t,\,\mathrm{d}x)\mu(t,\,\mathrm{d}y)\\ &= -\iint d_g(y,x)^2\mu(t,\,\mathrm{d}x)\mu(t,\,\mathrm{d}y) + O\left(\varepsilon_0^3\right).\end{aligned}$$

Taking ε_0 small enough,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{V}[\mu(t)](t) \leqslant -c\iint\|y-x\|^2\mu(t,\,\mathrm{d}x)\mu(t,\,\mathrm{d}y) \tag{C.13}$$

for some $c>0$ and all $t\geqslant t_0$. Fixing t, we consider the change of variables

$$y=\gamma(t)+z, \qquad \mathsf{T}(y)=y-\gamma(t),$$

and (C.13) rewrites as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{V}[\mu(t)](t) \leqslant -c\iint\|\gamma(t)+z-x\|^2\mu(t,\,\mathrm{d}x)\nu(t,\,\mathrm{d}z) \tag{C.14}$$

where $\nu(t)=\mathsf{T}_{\#}\mu(t)$. Expanding in (C.14) we find

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{V}[\mu(t)](t) &\leqslant -c\iint\left(\|\gamma(t)-x\|^2+\|z\|^2+2\langle z,\gamma(t)-x\rangle\right)\mu(t,\,\mathrm{d}x)\nu(t,\,\mathrm{d}z)\\ &= -c\int\|\gamma(t)-x\|^2\mu(t,\,\mathrm{d}x) - c\int\|z\|^2\nu(\,\mathrm{d}z)\\ &\qquad\qquad -2c\iint\langle z,\gamma(t)-x\rangle\mu(t,\,\mathrm{d}x)\nu(t,\,\mathrm{d}z)\\ &\leqslant -c\mathsf{V}[\mu(t)](t)+O\left(\varepsilon_0^3\right)\\ &\leqslant -c_0\mathsf{V}[\mu(t)](t)\end{aligned} \tag{C.15}$$

for some $c_0>0$ and all $t\geqslant t_0$, by choosing ε_0 small enough. By virtue of Grönwall's lemma,

$$\mathsf{V}[\mu(t)](t)\leqslant c_1e^{-c_0t} \tag{C.16}$$

for some constant $c_1 \geqslant 1$ and for all $t \geqslant t_0$. We now use (C.16) to conclude. We make use of

$$\begin{aligned} \mathsf{W}_2^2\left(\mu(t), \delta_{\gamma(t)}\right) &\lesssim \int \|x-\gamma(t)\|^2 \, \mu(t, \, \mathrm{d}x) \\ &= \int_{\{\|x-\gamma(t)\|^2 \leqslant \alpha\}} \|x-\gamma(t)\|^2 \, \mu(t, \, \mathrm{d}x) \\ &\quad + \int_{\{\|x-\gamma(t)\|^2 > \alpha\}} \|x-\gamma(t)\|^2 \, \mu(t, \, \mathrm{d}x), \end{aligned} \tag{C.17}$$

for $\alpha > 0$ to be determined later on, where we recall the integrals are taken over $\mathbb{S}^{d-1}$, and the first inequality follows by equivalence of norms and by considering the (optimal) transport plan which corresponds to $\mathsf{T}(x) = \gamma(t)$. We only need to estimate the second term in (C.17). By Markov's inequality and (C.16), we find

$$\mu\left(t, \left\{x \in \mathbb{S}^{d-1} : \|x-\gamma(t)\|^2 > \alpha\right\}\right) \leqslant \frac{\mathsf{V}[\mu(t)](t)}{\alpha} \leqslant \frac{c_1 e^{-2c_0 t}}{\alpha}.$$

Picking $\alpha = c_1 e^{-c_0 t}$, we deduce

$$\mu\left(t, \left\{x \in \mathbb{S}^{d-1} : \|x-\gamma(t)\|^2 > c_1 e^{-c_0 t}\right\}\right) \leqslant e^{-c_0 t}.$$

Coming back to (C.17), we find

$$\mathsf{W}_2^2\left(\mu(t), \delta_{\gamma(t)}\right) \lesssim e^{-c_0 t} + 2\pi \cdot \mu\left(t, \{\|x-\gamma(t)\|^2 > \alpha\}\right) \lesssim e^{-c_0 t}$$

for all $t \geqslant 0$. To conclude the proof, it suffices to show that

$$d_g(x_0, \gamma(t)) \lesssim e^{-c_2 t}$$

for some constant $c_2 > 0$ and for all $t \geqslant 0$. Similar computations to (C.9) give

$$\begin{aligned} \dot{\gamma}(t) &= \int \mathbf{P}_x^{\perp}(\gamma(t)) \mu(t, \, \mathrm{d}x) - \gamma(t)\left(1 - \int \langle \gamma(t), x\rangle^2 \mu(t, \, \mathrm{d}x)\right) \\ &= -\int \cos(d_g(x, \gamma(t))) x \mu(t, \, \mathrm{d}x) + \gamma(t) \int \cos^2(d_g(x, \gamma(t))) \mu(t, \, \mathrm{d}x) \\ &= -\int \cos(d_g(x, \gamma(t))) x \mu(t, \, \mathrm{d}x) + \gamma(t) \int \cos^2(d_g(x, \gamma(t))) \mu(t, \, \mathrm{d}x). \end{aligned}$$

We Taylor expand in the above identity to find

$$\begin{aligned}\dot\gamma(t) &= -\int \left(1-\frac{d_g(x,\gamma(t))^2}{2}\right) x\mu(t,\mathrm{d}x)\\ &\qquad\qquad +\gamma(t)\int (1-d_g(x,\gamma(t))^2)\mu(t,\mathrm{d}x) + O\left(\varepsilon_0^4\right)\\ &= -\int \left(1-\frac{d_g(x,\gamma(t))^2}{2}\right)(x-\gamma(t)+\gamma(t))\mu(t,\mathrm{d}x)\\ &\qquad\qquad +\gamma(t)\int (1-d_g(x,\gamma(t))^2)\mu(t,\mathrm{d}x) + O\left(\varepsilon_0^4\right)\\ &= -\int \left(1-\frac{d_g(x,\gamma(t))^2}{2}\right)(x-\gamma(t))\mu(t,\mathrm{d}x)\\ &\qquad\qquad -\frac{\gamma(t)}{2}\int d_g(x,\gamma(t))^2\mu(t,\mathrm{d}x) + O\left(\varepsilon_0^4\right).\end{aligned}$$

By Cauchy-Schwarz and (C.16), we deduce

$$\begin{aligned}&\left|\int \left(1-\frac{d_g(x,\gamma(t))^2}{2}\right)(x-\gamma(t))\mu(t,\mathrm{d}x)\right|\\ &\quad\leqslant \left(\int \left(1-\frac{d_g(x,\gamma(t))^2}{2}\right)^2\mu(t,\mathrm{d}x)\right)^{\frac12}\left(\int \|x-\gamma(t)\|^2\mu(t,\mathrm{d}x)\right)^{\frac12}\\ &\quad\leqslant \sqrt{\mathsf{V}[\mu(t)](t)} \lesssim e^{-\frac{c_0}{2}t}. \end{aligned}\tag{C.18}$$

Similarly,

$$\left|\gamma(t)\int d_g(x,\gamma(t))^2\mu(t,\mathrm{d}x)\right| \lesssim \int \|x-\gamma(t)\|^2\mu(t,\mathrm{d}x) \lesssim e^{-c_0 t}.\tag{C.19}$$

Combining (C.18) and (C.19), we deduce

$$\lim_{t\to+\infty}\gamma(t)=x_0 \qquad \text{and} \qquad |\dot\gamma(t)| \lesssim e^{-\frac{c_0}{2}t}$$

for all $t \geqslant t_0$. We conclude the proof by observing that

$$\begin{aligned}1-\langle x_0,\gamma(t)\rangle &= \langle x_0,\gamma(+\infty)\rangle - \langle x_0,\gamma(t)\rangle\\ &= \int_t^{+\infty}\frac{\mathrm{d}}{\mathrm{d}s}\langle x_0,\gamma(s)\rangle\,\mathrm{d}s \lesssim e^{-\frac{c_0}{2}t}.\end{aligned}$$

□

C.3 Transporting mass through overlapping balls

Lemma C.2. *Consider $K+1$ open balls $\mathscr{B}_K,\dots,\mathscr{B}_1,\mathscr{B}_0 \subset \mathbb{S}^{d-1}$ satisfying*

$$\begin{aligned}\mathscr{B}_k \cap \mathscr{B}_{k-1} &\neq \varnothing \qquad &&\text{for } k\in[K]\\ \mathscr{B}_k \cap \mathscr{B}_{k'} &= \varnothing \qquad &&\text{if } |k-k'|\geqslant 2.\end{aligned}$$

Then for any $T > 0$ and $\varepsilon > 0$, there exist $(\boldsymbol{W}, \boldsymbol{V}, b) : [0,T] \to \mathscr{M}_{d\times d}(\mathbb{R})^2 \times \mathbb{R}^d$, piecewise constant having at most K switches, such that for any $\mu_0 \in \mathscr{P}(\mathbb{S}^{d-1})$, the corresponding unique solution $\mu \in \mathscr{C}^0([0,T];\mathscr{P}(\mathbb{S}^{d-1}))$ to

$$\begin{cases} \partial_t \mu(t) + \operatorname{div}\left(\mathbf{P}_x^{\perp}\left(\boldsymbol{W}(t)\Big(\boldsymbol{U}(t)x + b(t)\Big)_+\right)\mu(t)\right) = 0 & \text{on } [0,T]\times\mathbb{S}^{d-1} \\ \mu(0) = \mu_0 & \text{on } \mathbb{S}^{d-1} \end{cases} \tag{C.20}$$

satisfies

$$\mu(T, \mathscr{B}_K) \geqslant (1-\varepsilon)^K \mu_0(\mathscr{B}_0).$$

Moreover, $\mu(T) = \Phi^T_{\#}\mu_0$ for a Lipschitz-continuous, invertible map $\Phi^t : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ which satisfies

$$\Phi^t(x) = x \qquad \text{for } x \notin \bigcup_{k=0}^{K} \mathscr{B}_k \tag{C.21}$$

for all $t \in [0,T]$.

We now focus on proving Lemma C.2, itself relying on the following lemma.

Lemma C.3. *Consider two open balls $\mathscr{B}_0, \mathscr{B}_1 \subset \mathbb{S}^{d-1}$ such that $\mathscr{B}_0 \cap \mathscr{B}_1 \neq \varnothing$. For any $\varepsilon > 0$ and $T > 0$, there exist $\boldsymbol{W}, \boldsymbol{U} \in \mathscr{M}_{d\times d}(\mathbb{R})$ and $b \in \mathbb{R}^d$ such that for any $\mu_0 \in \mathscr{P}(\mathbb{S}^{d-1})$, the unique solution $\mu \in \mathscr{C}^0([0,T];\mathscr{P}(\mathbb{S}^{d-1}))$ to (C.20) satisfies*

$$\mu(T, \mathscr{B}_0 \cap \mathscr{B}_1) \geqslant (1-\varepsilon)\mu_0(\mathscr{B}_0). \tag{C.22}$$

Moreover $\mu(T) = \Phi^T_{\#}\mu_0$ where the Lipschitz-continuous and invertible flow map $\Phi^t : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ of (4.1) satisfies

$$(\Phi^t)_{|\mathbb{S}^{d-1}\setminus\mathscr{B}_0} \equiv \mathrm{Id} \qquad \text{for } t \in [0,T].$$

Furthermore, for an arbitrary $\omega \in \operatorname{int}(\mathscr{B}_0)$, we can also choose $\boldsymbol{W}$, $\boldsymbol{V}$ and b, so that the solution to (C.20) satisfies:

$$\mathsf{W}_2(\mu(T), \alpha) \leqslant \varepsilon$$

where

$$\alpha(A) = \mu_0(\mathscr{B}_0)\delta_\omega(A) + \mu_0(A \setminus \mathscr{B}_0),$$

for any Borel set $A \subset \mathbb{S}^{d-1}$.

Proof of Lemma C.3. As done in previous proofs, we can take all time horizons to be as large as desired throughout by rescaling the norm of the parameters. Let $z \in \mathbb{S}^{d-1}$ denote the center and $R > 0$ the radius of $\mathscr{B}_0$. Take an arbitrary $\omega \in \operatorname{int}(\mathscr{B}_0 \cap \mathscr{B}_1)$. We consider

$$\begin{aligned} \boldsymbol{U} &= -\mathbf{1}z^\top, \\ b &= \cos(R)\mathbf{1}, \end{aligned}$$

as well as any $\boldsymbol{W} \in \mathscr{M}_{d\times d}(\mathbb{R})$ such that

$$\boldsymbol{W}\mathbf{1} = \omega.$$

Then

$$\boldsymbol{W}\,(\,\boldsymbol{U}x + b)_+ = \Big(-\cos d_g(z,x) + \cos(R)\Big)_+ \omega,$$

and note that

$$\Big(-\cos d_g(z,x) + \cos(R)\Big)_+ > 0 \qquad \Longleftrightarrow \qquad x \in \mathscr{B}_0. \tag{C.23}$$

Now observe that

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle x(t),\omega\rangle = \Big(-\cos d_g(z,x(t)) + \cos(R)\Big)_+ \Big(1 - \langle x(t),\omega\rangle^2\Big), \tag{C.24}$$

which is positive whenever $x(t) \in \mathscr{B}_0 \setminus \{\omega\}$. We claim that this implies the existence of a time $T_\varepsilon > 0$ for which

$$\mu(T_\varepsilon, \mathscr{B}_0 \cap \mathscr{B}_1) \geqslant (1-\varepsilon)\mu_0(\mathscr{B}_0). \tag{C.25}$$

To prove this claim, let $\delta > 0$ be fixed and to be determined later on. Because of (C.24), there exists some $T_\delta > 0$ such that

$$\Phi^{T_\delta}(x) \in \mathscr{B}_0 \cap \mathscr{B}_1 \qquad \text{for } x \in B(z, R-\delta), \tag{C.26}$$

where $\Phi^{T_\delta} : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ is the flow map of (4.1). Since $\mu(T_\delta) = \Phi^{T_\delta}_{\#}\mu_0$, we have

$$\mu(T_\delta, \mathscr{B}_0 \cap \mathscr{B}_1) = \mu_0\Big((\Phi^{T_\delta})^{-1}(\mathscr{B}_0 \cap \mathscr{B}_1)\Big) \overset{(C.26)}{\geqslant} \mu_0(B(z, R-\delta)). \tag{C.27}$$

Taking $\delta > 0$ small enough so that $\mu_0(B(z,R-\delta)) \geqslant (1-\varepsilon)\mu_0(\mathscr{B}_0)$ yields claim (C.25). We conclude that (C.22) holds by rescaling time. Finally, by virtue of (C.23), the flow map Φ^t is such that $\Phi^t(x) = x$ for $x \in \mathbb{S}^{d-1} \setminus \mathscr{B}_0$ and $t \in [0,T]$.

As for the second part of the statement, take $\mathscr{B}_1 = B(\omega,\eta) \subset \mathscr{B}_0$ with $\eta > 0$ to be determined later on. Owing to (C.27), we can argue in the same fashion as in the proof of Proposition 2.3. We have

$$\begin{aligned}
\mathsf{W}_1(\mu(T_\delta),\alpha) &= \sup_{\mathrm{Lip}(\phi)\leqslant 1} \left| \int \phi(\mu(T_\delta) - \alpha) \right| \\
&= \sup_{\mathrm{Lip}(\phi)\leqslant 1} \left| \int_{\mathscr{B}_0} \phi(\mu(T_\delta) - \alpha) + \int_{\mathbb{S}^{d-1}\setminus\mathscr{B}_0} \phi(\mu(T_\delta) - \alpha) \right|.
\end{aligned}$$

Let $\bar{\varepsilon} > 0$ be arbitrary and to be chosen small enough later. Using (C.27)—with $\bar{\varepsilon}$ instead of ε—and the definition of $\mathscr{B}_1$, we find

$$\begin{aligned}
\left| \int_{\mathscr{B}_0\setminus\mathscr{B}_1} \phi\,(\mu(T_\delta) - \alpha) + \int_{\mathscr{B}_1} \phi\,(\mu(T_\delta) - \alpha) \right| &\leqslant \left| \int_{\mathscr{B}_0\setminus\mathscr{B}_1} \phi\,(\mu(T_\delta) - \alpha) \right| \\
&\quad + \left| \int_{\mathscr{B}_1} \phi\mu(T_\delta) - \mu(T_\delta,\mathscr{B}_1)\phi(\omega) - (\mu_0(\mathscr{B}_0) - \mu(T_\delta,\mathscr{B}_1))\phi(\omega) \right| \\
&\leqslant \|\nabla\phi\|_{L^\infty(\mathbb{S}^{d-1})} \cdot \eta \cdot \bar{\varepsilon} \cdot \mu_0(\mathscr{B}_0) + \eta + \bar{\varepsilon}\cdot\mu_0(\mathscr{B}_0),
\end{aligned}$$

which tends to 0 as $\bar{\varepsilon}$ and η tend to zero. On the other hand,

$$\left| \int_{\mathbb{S}^{d-1} \setminus \mathscr{B}_0} \phi(\mu(T_\delta) - \alpha) \right| = 0$$

by construction. We choose $\bar{\varepsilon}$ and η small enough in a way that

$$\mathsf{W}_1(\mu(T_\delta), \alpha) \leqslant \varepsilon. \qquad \square$$

We finally provide the brief proof of Lemma C.2:

Proof of Lemma C.2. We write

$$[0, T) = \bigcup_{k \in [M]} [t_{k-1}, t_k)$$

where $t_k = \frac{kT}{K}$, and proceed by backward induction:

$$\begin{aligned} \mu(T, \mathscr{B}_K) &= \mu(T, \mathscr{B}_K \setminus \mathscr{B}_{K-1}) + \mu(T, \mathscr{B}_K \cap \mathscr{B}_{K-1}) \\ &\geqslant \mu(t_{K-1}, \mathscr{B}_K \setminus \mathscr{B}_{K-1}) + (1 - \varepsilon)\mu(t_{K-1}, \mathscr{B}_{K-1}), \end{aligned}$$

where the last inequality follows from Lemma C.3. Using $\mathscr{B}_k \cap \mathscr{B}_{k'} = \varnothing$ whenever $|k - k'| \geqslant 2$, we arrive to

$$\mu(T, \mathscr{B}_k) \geqslant \sum_{k=1}^{K} (1 - \varepsilon)^{K-k} \mu_0(\mathscr{B}_k \setminus \mathscr{B}_{k-1}) + (1 - \varepsilon)^K \mu_0(\mathscr{B}_0),$$

whereupon the conclusion follows.

$\square$

C.4 Proof of Lemma 3.3

Proof of Lemma 3.3. Without loss of generality suppose that $j = N$ in the statement. We proceed by induction over the number of measures N. Assume that for all $i, j \in [N - 1]$ we have

$$\operatorname{supp}\left(\mu_0^i\right) \cap \operatorname{supp}\left(\mu_0^j\right) = \varnothing.$$

Let us prove that we can find a solution map $\Phi_{\text{fin}} : \mathscr{P}(\mathbb{S}^{d-1}) \to \mathscr{P}(\mathbb{S}^{d-1})$ of (1.4) such that

$$\operatorname{supp}\left(\Phi_{\text{fin}}(\mu_0^i)\right) \cap \operatorname{supp}\left(\Phi_{\text{fin}}(\mu_0^j)\right) = \varnothing$$

for all $i \neq j \in [N - 1]$ and

$$\begin{aligned} \operatorname{supp}\left(\Phi_{\text{fin}}(\mu_0^i)\right) \cap \operatorname{supp}\left(\Phi_{\text{fin}}(\mu_0^N)\right) &= \varnothing, \\ \operatorname{supp}\left(\Phi_{\text{fin}}(\mu_0^i)\right) \cap \operatorname{supp}\left(\Phi_{\text{fin}}(\nu_0)\right) &= \varnothing \end{aligned}$$

for all $i \in [N - 1]$. As done in previous proofs, we can take all time horizons to be as large as desired throughout by rescaling the norm of the parameters.

Step 1. Isolating μ_0^N and ν_0

Throughout this first step, $\boldsymbol{W} \equiv 0$. Consider $0 < T_0 < T_1 < \ldots < T_{d-1}$ to be chosen later on, and set

$$\boldsymbol{V}(t) = \sum_{k=1}^{d-1} \alpha_k \alpha_k^\top 1_{[T_{k-1},T_k]}(t)$$

with $\{\alpha_k\}_{k\in[d-1]}$ being an orthonormal basis of $\operatorname{span}\left(\left\{\mathbb{E}_{\mu_0^N}[z]\right\}\right)^\perp$, namely

$$\left\langle \mathbb{E}_{\mu_0^N}[x], \alpha_k \right\rangle = 0$$

for all $k \in [d-1]$. We proceed recursively, starting from $k = 1$. Observe that the solution to

$$\begin{cases} \partial_t \mu(t) + \operatorname{div}\left(\mathbf{P}_x^\perp\left(\langle \alpha_1, \mathbb{E}_{\mu(t)}[x]\rangle \alpha_1\right)\mu(t)\right) = 0 & \text{on } \mathbb{R}_{\geqslant 0} \times \mathbb{S}^{d-1}, \\ \mu(0) = \mu_0 & \text{on } \mathbb{S}^{d-1} \end{cases} \tag{C.28}$$

for $\mu_0 \in \mathscr{P}(\mathbb{Q}_1^{d-1})$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \mathbb{E}_{\mu(t)}[x], \alpha_1 \right\rangle = \left\langle \mathbb{E}_{\mu(t)}[x], \alpha_1 \right\rangle \left(1 - \int \langle x', \alpha_1\rangle^2 \mu(t,\, \mathrm{d}x')\right).$$

This implies

$$\left\langle \mathbb{E}_{\mu(t)}[x], \alpha_1 \right\rangle = \langle \mathbb{E}_{\mu_0}[x], \alpha_1\rangle \exp\left(t - \int_0^t \int \langle x', \alpha_1\rangle^2 \mu(s,\, \mathrm{d}x')\, \mathrm{d}s\right).$$

Therefore $\langle \mathbb{E}_{\mu(t)}[x], \alpha_1\rangle$ does not change sign along the trajectory $\mu(t)$, and also $\frac{\mathrm{d}}{\mathrm{d}t}\langle \mathbb{E}_{\mu(t)}[x], \alpha_1\rangle = 0$ whenever $\mathbb{E}_{\mu_0}[x]$ is orthogonal to α_1 or if $\mu(t) = \delta_{\pm\alpha_1}$. Hence, for any $x(t) \in \operatorname{supp}(\mu(t))$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle x(t), \alpha_1\rangle = \left\langle \mathbb{E}_{\mu(t)}[x], \alpha_1 \right\rangle \left(1 - \langle \alpha_1, x(t)\rangle^2\right)$$

which implies that

$$\lim_{t\to+\infty} x(t) = \pm\alpha_1$$

whenever $\langle \mathbb{E}_{\mu_0}[x], \alpha_1\rangle \neq 0$. Therefore, for every $\varepsilon_1 > 0$ we can take $T_1 > 0$ large enough so that

$$\operatorname{supp}(\mu(T_1)) \subset B(\alpha_1, \varepsilon_1) \cup B(-\alpha_1, \varepsilon_1)$$

whenever $\langle \mathbb{E}_{\mu_0}[x], \alpha_1\rangle \neq 0$. We can repeat the argument for every k to deduce

$$\operatorname{supp}(\mu(T_{d-1})) \subset \bigcup_{k\in[d-1]} B(\alpha_k, C_k\varepsilon_k) \cup B(-\alpha_k, C_k\varepsilon_k) \tag{C.29}$$

where $C_k > 0$ does not depend on ε_k, but does depend on ε_ℓ for $\ell > k$. We can choose all radii ε_k small enough so that

$$\bigcup_{k\in[d-1]} B(\alpha_k, C_k\varepsilon_k) \cup B(-\alpha_k, C_k\varepsilon_k) \subset \mathbb{S}^{d-1} \setminus \mathbb{Q}_1^{d-1}. \tag{C.30}$$

Whence we have constructed a map

$$\Psi_1 : \mathscr{P}(\mathbb{S}^{d-1}) \to \mathscr{P}(\mathbb{S}^{d-1}),$$

with $\Psi_1(\mu_0) = \mu(T_d)$ where μ denotes the solution to the Cauchy problem (C.28) with the choice of parameters specified at the very beginning. Since $\mathbb{E}_{\mu_0^i}[x]$ is not colinear with $\mathbb{E}_{\mu_0^N}[x]$, and thanks to (C.29) and (C.30),

$$\operatorname{supp} \Psi_1(\mu_0^j) \subset \mathbb{S}^{d-1} \setminus \mathbb{Q}_1^{d-1}$$

for $j \in [N-1]$, and

$$\Psi_1(\mu_0^N) = \mu_0^N, \qquad \Psi_1(\nu_0) = \nu_0.$$

Step 2. Clustering the supports of μ_0^N and ν_0

Let $a \in \mathbb{S}^{d-1}$ and $\underline{b} \in \mathbb{R}$ be such that

$$\begin{aligned} &\langle a, x\rangle + \underline{b} > 0 && \text{for } x \in \mathbb{Q}_1^{d-1} \\ &\langle a, x\rangle + \underline{b} < 0 && \text{for } x \in \bigcup_{k\in[d-1]} B(\alpha_k, C_k\varepsilon_k) \cup B(-\alpha_k, C_k\varepsilon_k). \end{aligned}$$

For instance, this can be ensured by taking $\{\varepsilon_k\}_{k\in[d-1]}$ small enough and setting

$$a = \frac{\mathbb{E}_{\mu_0^N}[x]}{\|\mathbb{E}_{\mu_0^N}[x]\|} \qquad \text{and} \qquad b = -\max_{k\in[d-1]} C_k\varepsilon_k.$$

Let $\delta > 0$ be arbitrary; in the interval (T_d, T_δ), for $T_\delta > 0$ to be determined later on, consider

$$\begin{aligned} \boldsymbol{W}(t) &= \boldsymbol{W}_2 \cdot 1_{[T_d,T_\delta]}(t) & \boldsymbol{U}(t) &= \boldsymbol{U} \cdot 1_{[T_d,T_\delta]}(t), \\ b(t) &= \underline{b}\mathbf{1} \cdot 1_{[T_d,T_\delta]}(t), & \boldsymbol{U} &\equiv \mathbf{1}a^\top, \end{aligned}$$

where $\boldsymbol{W}_2$ is any $d \times d$ matrix such that

$$\boldsymbol{W}_2\mathbf{1} = \mathbb{E}_{\mu_0^N}[x].$$

For this choice of parameters, the measures $\mu^i(T_d)$, for $i \in [N-1]$, are invariant by the corresponding flow map of (1.4). We can choose $T_\delta > 0$ large enough so that

$$\operatorname{supp}(\nu(T_\delta)) \cup \operatorname{supp}\left(\mu^N(T_\delta)\right) \subset B\left(\frac{\mathbb{E}_{\mu_0^N}[x]}{\|\mathbb{E}_{\mu_0^N}[x]\|}, \delta\right). \tag{C.31}$$

This follows by observing that

$$\lim_{t\to+\infty}\left\langle x(t), \frac{\mathbb{E}_{\mu_0^N}[x]}{\|\mathbb{E}_{\mu_0^N}[x]\|}\right\rangle = 1$$

for every $x_0 \in \mathrm{supp}(\mu_0^N)$, where $x(t)$ follows the characteristics of (1.4), by following the same arguments as for (C.24) in the proof of Lemma C.3, or (4.6) in the proof of Proposition 4.2. This construction yields a flow map

$$\Psi_2 : \mathscr{P}(\mathbb{S}^{d-1}) \to \mathscr{P}(\mathbb{S}^{d-1}),$$

with

$$\Psi_2(\mu_0) = \mu(T_\delta)$$

where μ denotes the solution to the Cauchy problem (1.4) on $[T_d, T_\delta]$ with the choice of parameters specified in this step, which satisfies

$$\Psi_2(\mu^j(T_d)) = \mu^j(T_d)$$

for $j \in [N-1]$, and $\Psi_2(\mu^N(T_d))$, $\Psi_2(\nu(T_d))$ satisfy (C.31).

Step 3. Flow reversal

We finally employ the inverse map Ψ_1^{-1}, and choose $\delta > 0$ small enough to obtain the result; namely setting $\Phi_{\text{fin}} := \Psi_1^{-1} \circ \Psi_2 \circ \Psi_1$, we have

$$\Phi_{\text{fin}}(\mu_0^i) = \mu_0^i$$

for $i \in [N-1]$, and

$$\mathrm{supp}(\Phi_{\text{fin}}(\nu_0)) \cup \mathrm{supp}(\Phi_{\text{fin}}(\mu_0^N)) \subset B\left(\frac{\mathbb{E}_{\mu_0^N}[x]}{\|\mathbb{E}_{\mu_0^N}[x]\|}, C_T\delta\right),$$

for some $C_T > 0$ depending on Ψ_1 but not on Ψ_2. Therefore, by choosing $\delta > 0$ small enough, we can conclude. □

C.5 Proof of Lemma 3.4

Proof of Lemma 3.4. We begin with the first part of the statement. As before, we can take all time horizons to be as large as desired throughout by rescaling the norm of the parameters.

Part 1.

There exists an open ball $\mathscr{B} \subset \mathrm{supp}(\mu_0) \cup \mathrm{supp}(\nu_0)$ such that

$$\mu_0(\mathscr{B}) \neq \nu_0(\mathscr{B}).$$

We now claim that there exists some $x^* \in \mathscr{B}$ such that

$$\mu_0(\mathscr{B})x^* + \int_{\mathbb{S}^{d-1}\setminus\mathscr{B}} x\mu_0(\,\mathrm{d}x) \neq \nu_0(\mathscr{B})x^* + \int_{\mathbb{S}^{d-1}\setminus\mathscr{B}} x\nu_0(\,\mathrm{d}x).$$

Indeed if this were to be false, then we'd have

$$x^* = \frac{1}{\mu_0(\mathscr{B}) - \nu_0(\mathscr{B})} \int_{\mathbb{S}^{d-1}\setminus\mathscr{B}} x(\nu_0(\,\mathrm{d}x) - \mu_0(\,\mathrm{d}x))$$

for all $x^* \in \mathscr{B}$, which cannot hold. Take $x^* \in \mathscr{B}$ as above. Let $a \in \mathbb{S}^{d-1}$ be the center of $\mathscr{B}$ and $R > 0$ its radius. Consider

$$\begin{aligned} \boldsymbol{U} &= -\mathbf{1}a^\top, \\ b &= R\mathbf{1}, \end{aligned} \tag{C.32}$$

and any $\boldsymbol{W} \in \mathscr{M}_{d\times d}(\mathbb{R})$ satisfying

$$\boldsymbol{W}\mathbf{1} = x^*. \tag{C.33}$$

Then, by Lemma C.3, for any $\varepsilon > 0$ we can take a large enough $T > 0$ such that the solution to

$$\begin{cases} \partial_t \mu(t) + \operatorname{div}\Big(\mathsf{P}_x^{\perp}\Big(\boldsymbol{W}(\boldsymbol{U}x + b)_+\Big)\mu(t)\Big) = 0 & \text{on } \mathbb{R}_{\geqslant 0} \times \mathbb{S}^{d-1} \\ \mu(0) = \mu_0 & \text{on } \mathbb{S}^{d-1} \end{cases}$$

satisfies

$$\mathsf{W}_2(\mu(T), \alpha) \leqslant \varepsilon$$

where

$$\alpha(A) = \mu_0(\mathscr{B})\delta_{x^*}(A \setminus \mathscr{B}) + \mu_0(A \setminus \mathscr{B})$$

for any Borel $A \subset \mathbb{S}^{d-1}$. Since the expectation of a measure is continuous with respect to the measure in the sense of the Wasserstein distance, it follows that there is a Lipschitz invertible flow map $\Phi : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ of (4.1) such that $\mathbb{E}_{\Phi_\# \mu_0}[x] \neq \mathbb{E}_{\Phi_\#\nu_0}[x]$. Furthermore, $\Phi(x) = x$ for $x \notin \mathscr{B}$ by construction.

Part 2.

The parameters take the form

$$\begin{aligned} \boldsymbol{V}(t) &= I_d 1_{[0,T_*]}(t) & \boldsymbol{W}(t) &= \boldsymbol{W} 1_{[T_*,T]}(t) \\ \boldsymbol{U}(t) &= \boldsymbol{U} 1_{[T_*,T]}(t) & b(t) &= b 1_{[T_*,T]}(t), \end{aligned}$$

for $T_* > 0$ and $T > T_*$ to be determined later on. Recall that $\boldsymbol{B} \equiv 0$. We first prove that if

$$\operatorname{supp}(\mu_0) \neq \operatorname{supp}(\nu_0) \tag{C.34}$$

is not satisfied, it ought to hold after some time. Indeed, suppose that (C.34) does not hold. Let $\tau > 0$ be arbitrary. For any $x_0 \in \partial \mathrm{conv}_g \mathrm{supp}(\mu_0)) \cap \mathrm{supp}(\mu_0)$ consider

$$\begin{cases} \dot{x}(t) = \mathbb{E}_{\mu(t)}[x] - \left\langle \mathbb{E}_{\mu(t)}[x], x(t) \right\rangle x(t) & \text{in } [0,\tau] \\ x(0) = x_0 \end{cases}$$

and

$$\begin{cases} \dot{y}(t) = \mathbb{E}_{\nu(t)}[x] - \left\langle \mathbb{E}_{\nu(t)}[x], y(t) \right\rangle y(t) & \text{in } [0,\tau] \\ y(0) = x_0. \end{cases}$$

Taylor-expanding within the Duhamel formula, for τ small enough, we find

$$x(\tau) = x_0 + \tau \left(\mathbb{E}_{\mu_0}[x] - \langle \mathbb{E}_{\mu_0}[x], x_0 \rangle \, x_0 \right) + O(\tau^2)$$

and

$$y(\tau) = x_0 + \frac{\tau}{\gamma_1} \left(\mathbb{E}_{\mu_0}[x] - \langle \mathbb{E}_{\mu_0}[x], x_0 \rangle \, x_0 \right) + O(\tau^2)$$

Then

$$\left\langle y(\tau) - x(\tau), \frac{\mathbb{E}_{\mu_0}[x]}{\|\mathbb{E}_{\mu_0}[x]\|} \right\rangle = \tau \left(\frac{1}{\gamma_1} - 1 \right) \left(\|\mathbb{E}_{\mu_0}[x]\| - \frac{\langle \mathbb{E}_{\mu_0}[x], x_0 \rangle^2}{\|\mathbb{E}_{\mu_0}[x]\|} \right) + O(\tau^2).$$

Suppose[15] $\sigma_d(\mathrm{conv}_g \, \mathrm{supp}(\mu_0)) > 0$. Since $x_0 \in \partial \mathrm{conv}_g \, \mathrm{supp}(\mu_0)$ as well as $\frac{\mathbb{E}_{\mu_0}[x]}{\|\mathbb{E}_{\mu_0}[x]\|} \in \mathrm{int}(\mathrm{conv}_g \, \mathrm{supp}(\mu_0))$,

$$\|\mathbb{E}_{\mu_0}[x]\| - \frac{\langle \mathbb{E}_{\mu_0}[x], x_0 \rangle^2}{\|\mathbb{E}_{\mu_0}[x]\|} \geqslant c$$

for some $c > 0$. Since $\gamma_1 \in (0,1)$ we gather that

$$\left\langle y(\tau) - x(\tau), \frac{\mathbb{E}_{\mu_0}[x]}{\|\mathbb{E}_{\mu_0}[x]\|} \right\rangle > c_1 \tau + O(\tau^2) > 0$$

for some $c_1 > 0$ and for τ small enough. Consequently for T_* small enough, we have $\mathrm{supp}(\nu(T_*)) \subset \mathrm{supp}(\mu(T_*))$ as well as $\mathrm{supp}(\mu(T_*)) \neq \mathrm{supp}(\nu(T_*))$. Therefore, there exist $\varepsilon > 0$ and an open ball $\mathscr{B}$ such that

$$\mathscr{B} \cap \mathrm{supp}(\nu(T_*)) \neq \varnothing, \quad \mathscr{B} \cap \mathrm{supp}(\mu(T_*)) = \varnothing \tag{C.35}$$

and

$$\mathscr{B} \subset \left\{ x \in \mathbb{S}^{d-1} : \inf_{y \in \mathrm{conv}_g(\mu(T_*))} d_g(x,y) \leqslant \varepsilon \right\}.$$

Let $a \in \mathbb{S}^{d-1}$ be the center of $\mathscr{B}$ and R be its radius. Now, in the interval (T_*, T) we take $\boldsymbol{V} \equiv 0$ and $\boldsymbol{W}, \boldsymbol{U} \in \mathscr{M}_{d\times d}(\mathbb{R})$ and $b \in \mathbb{R}^d$ as in (C.32)–(C.33) for some

[15] If $\sigma_d(\mathrm{conv}_g \, \mathrm{supp}(\mu_0)) = 0$, we can argue as in the proof of Proposition 2.1, reducing the dynamics to $\mathbb{S}^{d-2}$ (or a lower-dimensional sphere), where the same proof can be repeated.

$x^* \in \mathscr{B}$ to be determined later on. Because of (C.35), ν is invariant with respect to the flow generated by the parameters defined in (C.32) and (C.33). We change the coordinate system so that

$$\left(\int_{\mathbb{S}^{d-1}} x\nu(T_*)\right)_1 = \alpha, \quad \left(\int_{\mathbb{S}^{d-1}} x\nu(T_*)\right)_k = 0 \quad \text{for } k \geqslant 2.$$

Using the fact that $\mathscr{B}$ is open and (C.35), it is impossible that for every $x^* \in \mathscr{B}$,

$$\left(\int_{\mathbb{S}^{d-1}\setminus\mathscr{B}} x\,\mu(T_*)\right)_2 + \mu(T_*, \mathscr{B})(x^*)_2 = 0.$$

Consequently there exist $x^* \in \mathscr{B}$ for which

$$\int_{\mathbb{S}^{d-1}\setminus\mathscr{B}} x\,\mu(T_*) + \mu(T_*, \mathscr{B})x^* \qquad \text{and} \qquad \int_{\mathbb{S}^{d-1}} x\nu(T_*)$$

are not colinear. Therefore, letting T large enough, by the same arguments as in Lemma 3.4 and since $\mathscr{B} \subset \operatorname{conv}_g \operatorname{supp}(\mu_0) \cup \operatorname{conv}_g \operatorname{supp}(\nu_0)$, we can conclude. □

C.6 Proof of Lemma 5.1

Proof of Lemma 5.1. Since the vector field in (1.4) (or (3.1)) is Lipschitz, for every $i \in [N]$ there exist Lipschitz-continuous and invertible maps $\mathsf{T}^i_{\Phi_1} : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ and $\mathsf{T}^i_{\Phi_3} : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ such that

$$\Phi_1(\mu_0^i) = \left(\mathsf{T}^i_{\Phi_1}\right)_{\#} \mu_0^i,$$

and

$$\Phi_3(\mu_1^i) = \left(\mathsf{T}^i_{\Phi_3}\right)_{\#} \mu_1^i.$$

Then

$$\operatorname{supp}\left(\left(\mathsf{T}^i_{\Phi_1}\right)_{\#} \mu_0^i\right) \cap \operatorname{supp}\left(\left(\mathsf{T}^j_{\Phi_1}\right)_{\#} \mu_0^j\right) = \varnothing, \tag{C.36}$$

and

$$\operatorname{supp}\left(\left(\mathsf{T}^i_{\Phi_3}\right)_{\#} \mu_1^i\right) \cap \operatorname{supp}\left(\left(\mathsf{T}^j_{\Phi_3}\right)_{\#} \mu_1^j\right) = \varnothing$$

for $i \neq j \in [N]$. We wish to find an integrable map $\Psi^i : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ that satisfies

$$\left(\Psi^i \circ \mathsf{T}^i_{\Phi_1}\right)_{\#} \mu_0^i = \left(\mathsf{T}^i_{\Phi_3}\right)_{\#} \mu_1^i.$$

Using (5.1), since $\mathsf{T}^i_{\Phi_1}$ and $\mathsf{T}^i_{\Phi_1}$ are bijective, this is equivalent to

$$\left(\mathsf{T}^i_{\Phi_3}\right)^{-1} \circ \Psi^i \circ \mathsf{T}^i_{\Phi_1} = \mathsf{T}^i,$$

so

$$\Psi^i = \mathsf{T}^i_{\Phi_3} \circ \mathsf{T}^i \circ \left(\mathsf{T}^i_{\Phi_1}\right)^{-1}.$$

Due to (C.36), there also exists a Lipschitz-continuous map $\Psi : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ satisfying (5.2). □

C.7 Proof of Lemma 5.4

Proof of Lemma 5.4. Consider

$$\Psi_\varepsilon^\dagger(x) := \sum_{m=1}^{M(\varepsilon)} y_m^\varepsilon 1_{\Omega_m(\varepsilon)}(x), \tag{C.37}$$

where $\Omega_m(\varepsilon) \subset \mathbb{S}^{d-1}$ are connected and pairwise disjoint with

$$\bigcup_{m\in[M(\varepsilon)]} \Omega_m(\varepsilon) = \mathbb{S}^{d-1}, \tag{C.38}$$

whereas $y_m^\varepsilon \neq y_{m'}^\varepsilon$ when $m \neq m'$, and

$$\left\|\Psi_\varepsilon^\dagger - \Psi\right\|_{L^2(\mu)} \leqslant \frac{\varepsilon}{2}. \tag{C.39}$$

Our goal is then to approximate $\Psi_\varepsilon^\dagger$ by means of some flow map $\Psi_\varepsilon : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ of (4.1). To this end, we also approximate μ as

$$\left|\mu(\mathbb{S}^{d-1}) - \mu^\eta(\mathbb{S}^{d-1})\right| \leqslant \eta,$$

with μ^η curated so that we can apply Proposition 4.1 and Proposition 2.3 "more easily". Then,

$$\begin{aligned}\int \left\|\Psi_\varepsilon(x) - \sum_{m=1}^{M(\varepsilon)} y_m^\varepsilon 1_{\Omega_m}\right\|^2 \mu(\,\mathrm{d}x) &= \int \left\|\Psi_\varepsilon(x) - \sum_{m=1}^{M(\varepsilon)} y_m^\varepsilon 1_{\Omega_m}\right\|^2 \mu^\eta(\,\mathrm{d}x) \\ &\quad + \int \left\|\Psi_\varepsilon(x) - \Psi_\varepsilon^\dagger(x)\right\|^2 (\mu(\,\mathrm{d}x) - \mu^\eta(\,\mathrm{d}x)) \\ &\leqslant \int \left\|\Psi_\varepsilon(x) - \sum_{m=1}^{M(\varepsilon)} y_m^\varepsilon 1_{\Omega_m}\right\|^2 \mu^\eta(\,\mathrm{d}x) + 2\pi\eta.\end{aligned} \tag{C.40}$$

Step 1: Constructing μ^η

Fix $\eta > 0$. By the Lebesgue decomposition theorem, we split μ into purely atomic and diffusive parts:

$$\mu = \mu_{\text{pp}} + \mu_{\text{diff}},$$

with μ_{diff} having no atoms, and[16]

$$\mu_{\text{pp}} = \sum_{n=1}^{\infty} \mu(\{x_n\})\delta_{x_n}.$$

[16] Recall that the atomic part of any σ-finite measure is always a countable set.

Let $N(\eta) \geqslant 1$ be such that

$$\mu_{\mathrm{pp}}^{\eta} := \sum_{n=1}^{N(\eta)} \mu(\{x_n\})\delta_{x_n}$$

satisfies

$$\mu_{\mathrm{pp}}(A) - \mu_{\mathrm{pp}}^{\eta}(A) \leqslant \frac{\eta}{2}$$

for any Borel $A \subset \mathbb{S}^{d-1}$. Fix $\eta_1 > 0$ to be determined later on but such that for all $n \in [N(\eta)]$,

$$B(x_n, \eta_1) \cap B(x_m, \eta_1) = \varnothing \qquad \text{for } m \neq n \in [N(\eta)]. \tag{C.41}$$

Consider

$$\mu^{\eta} := \mu_{\mathrm{pp}}^{\eta} + \mu_{\mathrm{diff}}^{\eta}, \tag{C.42}$$

where[17]

$$\mu_{\mathrm{diff}}^{\eta}(A) := \mu_{\mathrm{diff}}\left(A \setminus \bigcup_{n\in[N(\eta)]} B(x_n, \eta_1)\right) \tag{C.43}$$

for any Borel $A \subset \mathbb{S}^{d-1}$. Furthermore, take $\eta_1 > 0$ small enough so that, in addition to (C.41),

$$\left|\mu(\mathbb{S}^{d-1}) - \mu^{\eta}(\mathbb{S}^{d-1})\right| \leqslant \eta.$$

Step 2: Toward a sufficient matching problem

We further decompose μ^{η} in several parts. For $m \in [M(\varepsilon)]$, consider

$$\mu_m(A) := \mu_{\mathrm{diff}}^{\eta}(A \cap \Omega_m) \tag{C.44}$$

for any Borel $A \subset \mathbb{S}^{d-1}$. Because of (C.43), (C.44) and (C.38), we have

$$\mu_{\mathrm{diff}}^{\eta}(A) = \sum_{m=1}^{M(\varepsilon)} \mu_m(A) \tag{C.45}$$

for any Borel $A \subset \mathbb{S}^{d-1}$. Therefore, thanks to (C.42) and (C.45), bounding (C.40) boils down to bounding

$$\begin{aligned}
&\int \left\| \Psi_\varepsilon(x) - \sum_{m=1}^{M(\varepsilon)} y_m^{\varepsilon} 1_{\Omega_m} \right\|^2 \mu^{\eta}(\,\mathrm{d}x) \\
&= \sum_{m=1}^{M(\varepsilon)} \int \left\|\Psi_\varepsilon(x) - y_m^{\varepsilon}\right\|^2 \mu_m + \sum_{n=1}^{N(\eta)} \mu(\{x_n\}) \left\|\Psi_\varepsilon(x_n) - \Psi_\varepsilon^{\dagger}(x_n)\right\|^2.
\end{aligned} \tag{C.46}$$

[17] In case $\mu_{\mathrm{pp}} = 0$, we consider an arbitrary point $x_1 \in \mathbb{S}^{d-1}$ and then define

$$\mu_{\mathrm{diff}}^{\eta}(A) := \mu_{\mathrm{diff}}\left(A \setminus B(x_1, \eta_1)\right).$$

For the second term in (C.46) we will employ exact matching via Proposition 4.1, whereas for the first, we first note that for any $\eta_3 > 0$, one has the trivial identity

$$\begin{aligned}
&\int \|\Psi_\varepsilon(x) - y_m^\varepsilon\|^2 \, \mu_m(\mathrm{d}x) \\
&= \mu_m(\mathbb{S}^{d-1}) \Bigg(\int_{(\Psi_\varepsilon)^{-1}(B(x_m,\eta_3))} \|\Psi_\varepsilon(x) - y_m^\varepsilon\|^2 \frac{\mu_m(\mathrm{d}x)}{\mu_m(\mathbb{S}^{d-1})} \\
&\qquad + \int_{(\Psi_\varepsilon)^{-1}(B(x_m,\eta_3))^c} \|\Psi_\varepsilon(x) - y_m^\varepsilon\|^2 \frac{\mu_m(\mathrm{d}x)}{\mu_m(\mathbb{S}^{d-1})} \Bigg).
\end{aligned} \tag{C.47}$$

We use the following.

Claim 2. *Suppose $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ and $x_0 \in \mathbb{S}^{d-1}$ satisfy*

$$\mathsf{W}_2(\mu, \delta_{x_0}) \leqslant \eta_2.$$

Then there exists some numerical constant $C > 0$ (not depending on η_2) such that

$$1 - \mu(B(x_0, \eta_3)) \leqslant C \cdot \frac{\eta_2}{\eta_3}$$

for all $\eta_3 > 0$.

Proof of Claim 1. By compactness of $\mathbb{S}^{d-1}$ and Kantorovich-Rubinstein duality,

$$\mathsf{W}_1(\mu, \delta_{x_0}) = \sup_{\mathrm{Lip}(g)\leqslant 1} \int g(\mu - \delta_{x_0}) \leqslant C \cdot \eta_2$$

for some numerical constant $C > 0$. Hence, for $g : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ defined as

$$g(x) = \begin{cases} 1 & x \in B(x, \eta_3) \\ 1 - \frac{1-\eta_3}{\eta_3} & x \in B(x,\eta_3) \cap B(x, 2\eta_3) \\ 0 & x \notin B(x, 2\eta_3), \end{cases}$$

we obtain

$$1 - \mu(B(x, \eta_3)) \leqslant C \cdot \frac{\eta_2}{\eta_3}.$$ □

From (C.47), if

$$\mathsf{W}_2(\mu_m, \delta_{y_m^\varepsilon}) \leqslant \eta_2$$

were to hold, by applying Claim 2 one would find

$$\int \|\Psi_\varepsilon(x) - y_m\|^2 \, \mu_m(\mathrm{d}x) \leqslant \mu_m(\mathbb{S}^{d-1}) \left(\eta_3^2 + 2\pi \cdot C \cdot \frac{\eta_2}{\eta_3} \right). \tag{C.48}$$

(C.46) and (C.48) naturally raise the following problem: find a flow map that matches

$$\begin{aligned}
\left(\mu_m, \mu_m(\mathbb{S}^{d-1})\delta_{y_m}\right) &\quad \text{for } m \in [M(\varepsilon)], \\
\left(\mu(\{x_n\})\delta_{x_n}, \mu(\{x_n\})\delta_{\Psi_\varepsilon^\dagger(x_n)}\right) &\quad \text{for } n \in [N(\eta)].
\end{aligned}$$

We aim for the matching to be exact for the discrete input measures (second line) and approximate in W_2 for the diffuse ones (first line).

Step 3: Constructing Ψ_ε through matching

We look to use Proposition 2.3 to cluster the diffuse input measures to a single atom, which paired with Proposition 2.1 for matching all atoms approximately, would lead to the desired conclusion. Specifically, we construct the approximation candidate $\Psi_\varepsilon : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ as

$$\Psi_\varepsilon := \Phi_3 \circ \Phi_2 \circ \Phi_1, \tag{C.49}$$

where

- $\Phi_1 : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ is the flow map induced by Proposition 4.1[18], which exactly matches $\mu(\{x_n\})\delta_{x_n}$ to $\mu(\{x_n\})\delta_{\Psi_\varepsilon^\dagger(x_n)}$, for all $n \in [N(\eta)]$;
- $\Phi_2 : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ is the flow map induced by Proposition 2.3 that concentrates μ_m near some atom inside $\operatorname{supp}(\Phi_{1\#}\mu_m)$, for all $m \in [M(\varepsilon)]$;
- $\Phi_3 : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ is the flow map induced by Proposition 4.1 that matches the atoms from the previous step to $\mu_m(\mathbb{S}^{d-1})\delta_{y_m^\varepsilon}$, for all $m \in [M(\varepsilon)]$.

We now make the construction of (C.49) precise, and with the help of (C.48), Proposition 2.3, and Proposition 4.1, we bound the right hand side in (C.46).

1. Thanks to Proposition 4.1 we have

   $$\Phi_{1\#}\mu_{\mathrm{pp}}^\eta = \Psi_{\varepsilon\#}^\dagger \mu_{\mathrm{pp}}^\eta.$$

   Exact matching ensures

   $$\sum_{n=1}^{N(\eta)} \mu(\{x_n\}) \left\| \Phi_1(x_n) - \Psi_\varepsilon^\dagger(x_n) \right\|^2 = 0. \tag{C.50}$$

2. We apply Proposition 2.3 to the measures $(\Phi_{1\#}\mu_m)_{m\in[M(\varepsilon)]}$ to deduce that, for all $m \in [M(\varepsilon)]$,

   $$\mathsf{W}_2\left(\left(\Phi_2 \circ \Phi_1\right)_{\#} \frac{\mu_m}{\mu_m(\mathbb{S}^{d-1})}, \delta_{x_m} \right) \leqslant \eta_2$$

   for some $x_m \in \operatorname{supp}(\Phi_{1\#}\mu_m)$ and for small enough $\eta_2 > 0$ to be determined later on. Note that when we apply Proposition 2.3 for each m in view of clustering $\Phi_{1\#}\mu_m$ to a discrete measure supported inside $\operatorname{supp}(\Phi_{1\#}\mu_m)$, the flow map stemming from Proposition 2.3 also satisfies—because of how Lemma C.2 is applied in the proof of Proposition 2.3—

   $$\Phi_2\Big|_{\left(\mathbb{S}^{d-1}\setminus\bigcup_{m\in[M(\varepsilon)]}\operatorname{supp}(\Phi_{1\#}\mu_m)\right)} \equiv \mathrm{Id}. \tag{C.51}$$

[18] Should the assumption in Proposition 4.1 not hold, one can always choose slightly different y_m^ε in (C.37) so that the approximation error is not altered and the assumption does hold.

Then, by the continuity of the flow map Φ_1, and (C.43), we have

$$\operatorname{supp}\left(\Phi_{1\#}\mu^{\eta}_{\mathrm{pp}}\right) \subset \mathbb{S}^{d-1} \setminus \left(\bigcup_{m\in[M(\varepsilon)]} \operatorname{supp}(\Phi_{1\#}\mu_m) \right),$$

and from (C.51)

$$\left(\Phi_2 \circ \Phi_1\right)_{\#}\mu^{\eta}_{\mathrm{pp}} = \Psi^{\dagger}_{\varepsilon\#}\mu^{\eta}_{\mathrm{pp}}.$$

This means that, paired with (C.50), we also have

$$\sum_{n=1}^{N(\eta)} \mu(\{x_n\}) \left\| (\Phi_2 \circ \Phi_1)(x_n) - \Psi^{\dagger}_{\varepsilon}(x_n) \right\|^2 = 0.$$

3. We then apply Proposition 4.1 to find a flow map Φ_3 which matches the pairs $(x_m, y_m)_{m\in[M(\varepsilon)]}$, and leads us to deduce, by virtue of continuity with respect to the data of (C.20), that

$$\mathsf{W}_2\left(\left(\Phi_3 \circ \Phi_2 \circ \Phi_1\right)_{\#}\frac{\mu_m}{\mu_m(\mathbb{S}^{d-1})}, \delta_{y^{\varepsilon}_m}\right) \leqslant C_{M(\varepsilon)} \cdot \eta_2 \tag{C.52}$$

holds for some $C_{M(\varepsilon)} > 0$ independent of η. Moreover, after applying Φ_3, thanks to Proposition 4.1 (or Proposition 4.2), we have that the pure point part remains unaltered:

$$\left(\Phi_3 \circ \Phi_2 \circ \Phi_1\right)_{\#}\mu^{\eta}_{\mathrm{pp}} = \Psi^{\dagger}_{\varepsilon\#}\mu^{\eta}_{\mathrm{pp}}.$$

Hence,

$$\sum_{n=1}^{N(\eta)} \mu(\{x_n\}) \left\| \Psi_{\varepsilon}(x_n) - \Psi^{\dagger}_{\varepsilon}(x_n) \right\|^2 = 0.$$

Step 4: Putting everything together

Thanks to (C.48) and (C.52), for any $\varepsilon_1 > 0$ we can choose η_2 and η_3 small enough as to ensure

$$\int \|\Psi_{\varepsilon}(x) - y^{\varepsilon}_m\|^2 \mu_m(\,\mathrm{d}x) \leqslant \mu_m(\mathbb{S}^{d-1})\varepsilon_1.$$

Since $\sum_{m\in[M(\varepsilon)]} \mu_m(\mathbb{S}^{d-1}) \leqslant 1$ by construction,

$$\sum_{m=1}^{M(\varepsilon)} \int \|\Psi_{\varepsilon}(x) - y^{\varepsilon}_m\|^2 \mu_m(\,\mathrm{d}x) \leqslant \varepsilon_1.$$

Combining all the estimates, and choosing ε_1 and η small enough, we can deduce that

$$\left\|\Psi_{\varepsilon} - \Psi^{\dagger}_{\varepsilon}\right\|^2_{L^2(\mu)} = \int \left\| \Psi_{\varepsilon}(x) - \sum_{m=1}^{M(\varepsilon)} y^{\varepsilon}_m 1_{\Omega_m} \right\|^2 \mu(\,\mathrm{d}x) \leqslant \varepsilon_1 + 2\pi \cdot \eta \leqslant \frac{\varepsilon^2}{4},$$

which paired with (C.39) leads us to the conclusion. □

Remark C.4 (Number of switches)**.** *In the proof above,*

- Φ_1 *is induced by parameters with* $O(N(\eta))$ *switches, and therefore depends on the decay of* $(\mu(\{x_n\}))_{n\geqslant 1}$*. If there are finitely many atoms, then there is no dependence on* η*.*
- Φ_2 *is induced by parameters where the number of switches depends on the packing numbers of the supports of* $\Phi^1_{\#}\mu_m$*—see Remark 2.4.*
- Φ_3 *is induced by parameters with* $O(M(\varepsilon))$ *switches, where* $M(\varepsilon)$ *depends on the approximation of* T^i *by a simple function. It is worth noting that whenever a simple function with* M *components is considered as a transport map* T^i*, the resulting measure* $\mathsf{T}^i_{\#}\mu$ *consists of a combination of* M *atoms. As a consequence, if we have* N *targets each having* M *atoms, after disentangling and clustering them, the matching with* Φ_3 *can be done with* $O(M\cdot N)$ *switches.*

D Disentangling through continuous feedback

For purely demonstrative purposes, in this section we show that measures can also be disentangled by using self-attention with $\boldsymbol{B}\not\equiv 0$. The proof is rather technical and does not yield the most desirable estimates on the number of switches—in fact, we even take the control $\boldsymbol{B}(t)$ in continuous[19] feedback form, meaning it is not piecewise constant.

We begin with the following lemma, which provides a flow map that, roughly speaking, reduces the entire system to one defined on the circle.

Lemma D.1. *Let* $\mu_0^i\in\mathscr{P}(\mathbb{S}^{d-1})$*,* $i\in[N]$*, be such that*

$$\bigcup_{i\in[N]}\operatorname{supp}\left(\mu_0^i\right)\subset\mathbb{S}^{d-1}.$$

For any $i\in[N]$ *consider the marginal* $\nu^i\in\mathscr{P}(\mathbb{S}^1)$ *defined as*

$$\nu^i(x_1,x_2)=\int_{[0,\pi]^{d-2}}\mu_0^i(x_1,x_2,\mathrm{d}\phi_3,\dots,\mathrm{d}\phi_d),$$

where ϕ_k*,* $k\geqslant 3$*, correspond to angular hyper-spherical coordinates. Then for every* $\varepsilon>0$ *and* $T>0$*, there exists* $\theta=(\boldsymbol{V},\boldsymbol{B},\boldsymbol{W},\boldsymbol{U},b)\in L^\infty((0,T);\Theta)$ *such that for any* $i\in[N]$*, the solution* $\mu^i\in\mathscr{C}^0([0,T];\mathscr{P}(\mathbb{S}^{d-1}))$ *to* (1.4) *corresponding to the initial data* μ_0^i *and parameters* θ *satisfies*[20]

$$\mathsf{W}_2\left(\mu^i(T),\nu^i\otimes\delta_0^{\otimes(d-2)}\right)\leqslant\varepsilon.$$

Moreover, we can take $\boldsymbol{B}\equiv\boldsymbol{V}\equiv b\equiv 0$*, and* $\boldsymbol{W}$ *and* $\boldsymbol{U}$ *piecewise constant with at most* $2(d-2)$ *switches.*

[19] This is to contrast the piece-wise constant controls constructed in what precedes, which can also be interpreted as some sort of feedback, since at every switch we choose a constant control depending on the location of the particles.

[20] We use the standard shorthand $\delta_0^{\otimes m}(x)=\delta_0(x_1)\otimes\cdots\otimes\delta_0(x_m)$ for $x\in\mathbb{R}^m$.

Proof of Lemma D.1. The proof is done by induction—it is thus enough to prove that we can "collapse" one dimension/coordinate. We begin by the last coordinate. Consider

$$\mathscr{P}(\mathbb{S}^{d-2}) \ni \nu_k^i(x_1, x_2, \dots, x_k) := \int_0^\pi \mu_0^i(x_1, x_2, \dots, x_k, \mathrm{d}\phi_{k+1}, \dots \mathrm{d}\phi_d).$$

With this notation, $\nu^i = \nu_2^i$. Let $T_d > 0$ and $\varepsilon_d > 0$ to be chosen later on. Consider

$$\begin{aligned} \boldsymbol{U}_+(t) &= \mathbf{1}e_d^\top, \\ \boldsymbol{W}_+(t)\mathbf{1} &= -e_d \end{aligned}$$

for $t \in [0, T_d]$. Then on $[0, T_d]$, the characteristics of (1.4) become

$$\begin{cases} \dot{x}(t) = -\langle e_d, x(t)\rangle_+ \mathbf{P}^\perp_{x(t)} e_d & \text{in } [0, T_d] \\ x(0) = x_0. \end{cases} \tag{D.1}$$

One sees that for any $x_0 \in \mathbb{S}^{d-1} \setminus \{e_d\}$ with $\langle x_0, e_d\rangle > 0$, we have $\langle x(t), e_d\rangle \to 0$ as $t \to +\infty$. Denoting the flow of (D.1) as $\Phi_+^{\frac{t}{2}} : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$, and similarly, denoting by $\Phi_-^{\frac{t}{2}} : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ the flow map associated to

$$\begin{aligned} \boldsymbol{U}_-(t) &= -\mathbf{1}e_d^\top, \\ \boldsymbol{W}_-(t)\mathbf{1} &= e_d, \end{aligned}$$

we have that

$$\Phi_+^{\frac{t}{2}} \circ \Phi_-^{\frac{t}{2}} = \Phi_-^{\frac{t}{2}} \circ \Phi_+^{\frac{t}{2}} =: \Psi_d^t.$$

(The subscript d indicates the coordinate which we collapsing to the equator.) Since for any $i \in [N]$, μ_0^i has no atom at e_d nor at $-e_d$, we can choose T_d such that $\mu^i(T_d) = (\Psi_d^{T_d})_\# \mu_0^i$ satisfies

$$\mathsf{W}_2\left(\mu^i(T_d), \nu_{d-1}^i \otimes \delta_0\right) \leqslant \varepsilon_d$$

for $i \in [N]$.

Now assume heredity:

$$\mathsf{W}_2\left(\left(\Psi_4^{T_4} \circ \cdots \circ \Psi_d^{T_d}\right)_\# \mu_0^i, \nu_3^i \otimes \delta_0^{\otimes(d-3)}\right) \leqslant \varepsilon_4 \tag{D.2}$$

for an arbitrary $\varepsilon_4 > 0$, $i \in [N]$, and for some flow maps $\Psi_k^{T_k} : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ induced by the characteristics of (1.4). For $\varepsilon_3 > 0$ to be chosen later on, we apply the same reasoning as done above to find a flow map $\Psi_3^{T_3} : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ induced by the characteristics of (1.4) such that

$$\mathsf{W}_2\left(\left(\Psi_3^{T_3}\right)_\# \nu_3^i, \nu^i \otimes \delta_0^{\otimes(d-3)}\right) \leqslant \varepsilon_3. \tag{D.3}$$

Then using standard continuity estimates for the continuity equation, along with (D.3), (D.2) and the triangle inequality, we deduce

$$\begin{aligned}
&\mathsf{W}_2\left(\left(\Psi_3^{T_3}\circ\Psi_4^{T_4}\circ\cdots\circ\Psi_d^{T_d}\right)_{\#}\mu_0^i,\nu^i\otimes\delta_0^{\otimes(d-3)}\right)\\
&\qquad\leqslant \mathsf{W}_2\left(\left(\Psi_3^{T_3}\circ\Psi_4^{T_4}\circ\cdots\circ\Psi_d^{T_d}\right)_{\#}\mu_0^i,\left(\Psi_3^{T_3}\right)_{\#}\nu_3^i\right)\\
&\qquad\qquad+\mathsf{W}_2\left(\left(\Psi_3^{T_3}\right)_{\#}\nu_3^i,\nu^i\otimes\delta_0^{\otimes(d-2)}\right)\\
&\qquad\leqslant C(\varepsilon_3)\varepsilon_4+\varepsilon_3
\end{aligned}$$

for some $C(\varepsilon_3)>0$. We may choose ε_4 and ε_3 small enough to conclude. □

We now show that disentanglement can also be accomplished by taking $\boldsymbol{B}(t)$ in feedback form. The proof is based, roughly speaking, on the ability to determine the location of the cluster of one of the measures and ensuring that it is different from the limit cluster of the remaining measures. With that in hand, we can separate one measure from the rest and then proceed by induction.

Proposition D.2. *Let $T>0$. Consider $\mu_0^i\in\mathcal{P}(\mathbb{S}^{d-1})$, $i\in[N]$ satisfying*

$$\partial\mathrm{supp}(\nu_0^i)\cap\partial\mathrm{supp}(\nu_0^j)=\varnothing\qquad\text{for } i\neq j,$$

where, as before, $\nu_0^i\in\mathcal{P}(\mathbb{S}^1)$ denotes the marginal

$$\nu_0^i(x_1,x_2)=\int_{[0,\pi]^{d-2}}\mu_0^i(x_1,x_2,\,\mathrm{d}\phi_3,\ldots,\,\mathrm{d}\phi_d).$$

Then there exists $\theta\in L^\infty((0,T);\Theta)$ such that

$$\mathrm{supp}\left(\mu^i(T)\right)\cap\mathrm{supp}\left(\mu^j(T)\right)=\varnothing\qquad\text{for } i\neq j.$$

Proof of Proposition D.2. We proceed in several steps. Throughout, $\boldsymbol{V}\equiv I_d$ and $b\equiv 0$.

Step 1. Squashing to the equator; transporting to the first orthant

By virtue of Lemma D.1, for any $\varepsilon>0$ we can find a flow map $\Phi_\varepsilon:\mathbb{S}^{d-1}\to\mathbb{S}^{d-1}$ induced by the characteristics of (1.4) such that

$$\mathsf{W}_2\left(\Phi_{\varepsilon\#}\mu_0^i,\eta_0^i\otimes\delta_0^{\otimes(d-2)}\right)\leqslant\varepsilon,\tag{D.4}$$

where $\eta_0^i=\Psi_{\#}\nu_0^i\in\mathcal{P}(\mathbb{S}^1\cap(\mathbb{R}_{\geqslant 0})^2)$, with Ψ being the flow map given by Lemma 3.2. Since $\varepsilon>0$ can be chosen arbitrarily small, by virtue of (D.4) we can, without loss of generality, assume that μ_0^i are defined on $\mathbb{Q}_1^1:=\mathbb{S}^1\cap(\mathbb{R}_{\geqslant 0})^2$.

Step 2. Creating an atom at the argmax

We order (and relabel) the measures μ_0^i by decreasing order of the respective magnitude of

$$x^{i+} := \underset{x \in \mathrm{supp}(\mu_0^i)}{\arg\max} \langle x, e_2 \rangle.$$

Let $\eta > 0$ be chosen later on. We apply Lemma C.3 with $\mathscr{B}_0 = B(x^{+1}, \rho)$, where $\rho < d_g(x^{+1}, x^{+2})$, and $\mathscr{B}_1 = B(x^{+1}, \eta)$, and by choosing $\omega = x^{+1}$ in the proof, it follows that there exists some time $T_1(\eta) > 0$ such that

$$\mathsf{W}_2 \left(\mu^1(T_1(\eta)), \alpha \right) \leqslant \delta, \tag{D.5}$$

where

$$\alpha(A) = \mu_0^1(\mathscr{B}_0)\delta_{x^{1+}}(A) + \mu_0^1(A \setminus \mathscr{B}_0)$$

for any Borel $A \subset \mathbb{S}^1$. Furthermore, due to the choice of ρ, we have

$$\mu^i(T_1(\eta)) = \mu_0^i$$

for $2 \leqslant i \leqslant N$.

Step 3. A feedback to counter attention

Let $T_2(\eta) > T_1(\eta)$ be chosen later on. For $t \in [T_1(\eta), T_2(\eta)]$ we choose

$$\boldsymbol{U}(t) \equiv \mathbf{1}^\top a$$

where $a \in \mathbb{S}^1$ is such that

$$\begin{aligned} &\langle a, x^{1+} \rangle > 0, \\ &\langle a, x \rangle < 0 \qquad \text{for } x \in \bigcup_{j=2}^{N} \mathrm{supp}(\mu_0^j). \end{aligned}$$

(As such, the perceptron component of the vector field vanishes for $2 \leqslant i \leqslant N$.) We then define $\boldsymbol{W}(t)$ in feedback form:

$$\left(\left\langle a, x^{1+} \right\rangle \right)_+ \boldsymbol{W}(t)\mathbf{1} + \mathscr{A}_{\boldsymbol{B}} \left[\mu^1(t) \right] (x^{1+}) = x^{1+}. \tag{D.6}$$

(In this way, after applying $\mathsf{P}_x^\perp$, the atom located at x^{1+} remains invariant.) Equivalently,

$$\boldsymbol{W}(t)1 = \frac{1}{(\langle a, x^{1+} \rangle)_+} \left(x^{1+} - \mathscr{A}_{\boldsymbol{B}} \left[\mu^1(t) \right] (x^{1+}) \right).$$

Note that $\boldsymbol{W}(t)$ can be chosen to be a diagonal matrix. Since we now operate on $\mathbb{Q}_1^1$, and the vector field $x \mapsto \boldsymbol{W}(\boldsymbol{U}x)_+$ does not affect μ_0^i for $2 \leqslant i \leqslant N$, the solution to the Cauchy problem (1.4) emanating from μ_0^i converges weakly to a

point mass δ_{z^i} as $t \to +\infty$, for some z^i lying in $\operatorname{conv}_g \operatorname{supp}(\mu_0^i)$ (see Proposition 2.1, or the arguments of [GLPR23, Theorem 4.1]). On the other hand, thanks to (D.5) and the definition of the feedback $\boldsymbol{W}(t)$, we also have a way of identifying the position z^1 of the limit δ_{z^1} for the solution emanating from μ_0^1. Indeed, let us choose $\boldsymbol{B} \in \mathscr{M}_{2\times 2}(\mathbb{R})$ such that

$$\begin{aligned} \boldsymbol{B}x^{1+} &= \beta x^{1+} \\ \boldsymbol{B}x &= 0 \quad \text{for } x \in \mathbb{S}^1 \text{ such that } \langle x, x^{1+}\rangle = 0, \end{aligned}$$

where the eigenvalue $\beta > 0$ is to be chosen later on. Observe that if, for every $\varepsilon > 0$, there exists a $\beta > 0$ such that the solution to (1.4) for $\mu^1(t)$ in the interval $(0, +\infty)$ satisfies

$$\mathscr{A}_{\boldsymbol{B}}[\mu^1(t)](x) \in sB(x^{1+}, C\varepsilon) \tag{D.7}$$

for $s \geqslant c > 0$, then, for every $\delta > 0$ there exists T_* for which

$$\operatorname{supp}\left(\mu^1(T_*)\right) \subset B(x^{1+}, C\varepsilon + \delta)$$

We prove that the solution $\mu^1 \in \mathscr{C}^0([0,T];\mathscr{P}(\mathbb{S}^1))$, whenever μ_0^1 has an atom at x^{1+}, satisfies (D.7).

1. Consider
$$\mathscr{C} := \operatorname{conv}(\mathbb{Q}_1^1).$$
For every $\mu \in \mathscr{P}(\mathbb{Q}_1^1)$ and $x \in \operatorname{supp}(\mu)$ one has $\mathscr{A}_{\boldsymbol{B}}[\mu](x) \in \mathscr{C}$ since $\mathscr{A}_{\boldsymbol{B}}[\mu](x)$ is a weighted average of the elements of the support of μ. This implies that
$$\|\mathscr{A}_{\boldsymbol{B}}[\mu](x)\| \geqslant c \tag{D.8}$$
for some $c > 0$ and for every $x \in \operatorname{supp}(\mu)$.

2. Thanks to (D.8), it only remains to assess the direction in which $\mathscr{A}_{\boldsymbol{B}}[\mu](x)$ is pointing. If $\varepsilon > 0$ is fixed, and for every μ that has an atom at x^{1+}, we notice that we can decompose the integral in three parts
$$\begin{aligned} e^{-\beta}\int e^{\langle \boldsymbol{B}x, x'\rangle} x' \mu(\mathrm{d}x') = \mu(\{x^{1+}\})x^{1+} \\ + \int_{B(x^{1+},\varepsilon)\setminus\{x^{1+}\}} e^{\beta\langle x,x'\rangle\langle x^{1+},x'\rangle-\beta} x'\mu(\mathrm{d}x') \\ + \int_{\mathbb{S}^{d-1}\setminus B(x^{1+},\varepsilon)} e^{\beta\langle x,x'\rangle\langle x^{1+},x'\rangle-\beta} x'\mu(\mathrm{d}x'), \end{aligned}$$
The above identity is a sum of three vectors with
$$\begin{aligned} \|w\| &:= \left\| \int_{B(x^{1+},\varepsilon)\setminus\{x^{1+}\}} e^{\beta\langle x,x'\rangle\langle x^{1+},x'\rangle-\beta} x'\mu(\mathrm{d}x') \right\| \\ &\leqslant \mu\left(B(x^{1+},\varepsilon)\setminus\{x^{1+}\}\right), \end{aligned}$$

as well as

$$\|v\| := \left\| \int_{\mathbb{S}^{d-1}\setminus B(x^{1+},\varepsilon)} e^{\beta\langle x,x'\rangle\langle x^{1+},x'\rangle-\beta} x' \mu(\,\mathrm{d}x') \right\| \leqslant e^{-O(\beta\varepsilon^2)},$$

where the last bound stems from a Taylor expansion of $\langle x^{1+}, x'\rangle$, and the implicit constant is universal. Applying this to $\mu^1(t,\cdot)$ for any $t \in [T_1(\eta), T_2(\eta)]$, and choosing $\beta = \varepsilon^{-3}$ so that $|v|$ goes to zero as ε goes to zero, we deduce

$$\mu^1(t, \{x^{1+}\})x^{1+} + w + v \in sB(x^{1+}, C\varepsilon)$$

for some $C > 0$ independent of ε, t, and for some $s \in [c, 1]$. Since x^{1+} is invariant by (D.6), $\mu(t, x^{1+}) = \mu(0, x^{1+})$ and we deduce (D.7).

Step 4. Clustering and separation

Because of the choice of $\boldsymbol{U}$ and $\boldsymbol{V}$, for $i \in [N]$ we have

$$\operatorname{supp}\left(\mu^i(t)\right) \subset \operatorname{supp}\left(\mu_0^i\right)$$

for any $t \in [T_1(\eta), T_2(\eta)]$. Let $\varepsilon < \rho$ be small enough so that for any $x \in \operatorname{supp}(\mu^1(t))$, we have $\mathcal{A}_{\boldsymbol{B}}(x) \in sB(x^{1+}, d_g(x^{1+}, x^{2+}))$ where $s \in [c, 1]$. It follows that we can choose $T_2(\eta) > 0$ large enough so that $\operatorname{supp}(\mu^1(T_2(\eta))) \subset B(x^{1+}, d_g(x^{1+}, x^{2+}))$, and hence,

$$\operatorname{supp}\left(\mu^1(T_2(\eta))\right) \cap \operatorname{supp}\left(\mu^i(T_2(\eta))\right) = \varnothing$$

for $2 \leqslant i \leqslant N$.

Step 5. Rotation and induction

For convenience, we relabel the measures obtained from the previous step by resetting time: $\mu_0^i = \mu^i(T_2(\eta))$. In this last step, we set $\boldsymbol{V} \equiv 0$, to send the first measure counter-clockwise to $\mathbb{Q}_1^1$ so that it has disjoint support with the other measures, and so that $\langle x^{1+}, e_2\rangle$ is smaller than the infimum of $x \mapsto \langle x, e_2\rangle$ over the supports of all the other measures. This argument can then be repeated for every $i \in [N]$. To do so, let $T_1 > 0$ to be chosen later, and

$$\boldsymbol{U}(t) \equiv \mathbf{1}a^\top$$

for $t \in [0, T_1]$ with $a \in \mathbb{S}^1$ satisfying

$$\langle a, z\rangle = 0$$

for all $z \in \mathbb{S}^1$ with $\langle z - y, e_2\rangle > 0$ for all $y \in \bigcup_{j=2}^N \operatorname{supp}(\mu_0^j)$ and $\langle z - y, e_2\rangle < 0$ for all $y \in \operatorname{supp}(\mu_0^1)$. Define

$$\mathcal{S}_a := \left\{x \in \mathbb{S}^1 : \langle a, x\rangle > 0\right\},$$

take $\omega \in \mathcal{S}_a \cap \operatorname{int}(\mathbb{Q}_4^1)$, and choose

$$\boldsymbol{W}(t)\mathbf{1} \equiv \omega$$

for $t \in [0, T_1]$. (Here $\mathbb{Q}_4^1 := \mathbb{S}^1 \cap \{x \in \mathbb{R}^2 \colon x_1 > 0, x_2 < 0\}$.) We can choose $T_1 > 0$ large enough so that

$$\operatorname{supp}\left(\mu^1(T_1)\right) \subset \operatorname{int}\left(\mathbb{Q}_4^1\right),$$

while the rest of the measures remain invariant: $\mu^i(T_1) = \mu_0^i$ for $2 \leqslant i \leqslant N$. Now let $T_2 > T_1$ to be chosen later, and $\boldsymbol{U}(t) \equiv \mathbf{1}a_2^\top$ for $t \in [T_1, T_2]$, where a_2 is such that

$$\mathcal{S}_{a_2} \cap \operatorname{int}\left(\mathbb{Q}_1^1\right) \neq \varnothing, \qquad \omega \in \operatorname{int}(\mathcal{S}_{a_2}), \qquad \mathcal{S}_{a_2} \cap \bigcup_{j=2}^{N} \operatorname{supp}\left(\mu_0^j\right) = \varnothing.$$

Then fix $\omega_2 \in \operatorname{int}(\mathbb{Q}_1^1) \cap \mathcal{S}_{a_2}$ and proceed as before to find a $T_2 > 0$ large enough so that

$$\operatorname{supp}\left(\mu^1(T_2)\right) \subset \operatorname{int}\left(\mathbb{Q}_1^1\right) \cap \mathcal{S}_{a_2}$$

then, by this argument, the intersection between the support of the first measure and all the others is empty. Furthermore, notice that from Step 3 of this proof, we have

$$\operatorname{supp}\left(\mu^j(T)\right) \subset \operatorname{supp}\left(\mu_0^j\right) \qquad \text{for } j \in [N].$$

Consequently, if two measures μ_0^i and μ_0^j had disjoint support, the supports of $\mu^j(T)$ and $\mu^i(T)$ would remain disjoint for all $T > 0$. We can inductively repeat the whole argument simply by relabelling the measures as

$$\mu_0^i := \mu^{i+1}(T_2) \qquad \text{for } i \in [N-1],$$

and

$$\mu_0^N := \mu^1(T_2),$$

to conclude. □

References

[AC09] Andrei Agrachev and Marco Caponigro. Controllability on the group of diffeomorphisms. *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, 26(6):2503–2509, 2009.

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

[AG24] Daniel Owusu Adu and Bahman Gharesifard. Approximate Controllability of Continuity Equation of Transformers. *IEEE Control Systems Letters*, 2024.

[AL09] Andrei Agrachev and Paul Lee. Optimal transportation under nonholonomic constraints. *Transactions of the American Mathematical Society*, 361(11):6019–6047, 2009.

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

[AS20] Andrei Agrachev and Andrey Sarychev. Control in the spaces of ensembles of points. *SIAM Journal on Control and Optimization*, 58(3):1579–1596, 2020.

[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.

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

[BD97] Martino Bardi and Italo Capuzzo Dolcetta. *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*, volume 12. Springer, 1997.

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

[BPA24] Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. Emergence of meta-stable clustering in mean-field transformer models. *arXiv preprint arXiv:2410.23228*, 2024.

[Bre91] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

[Bro08] R Brockett. On the control of Liouville equations. In *Differential Equation and Topology: Abstracts of International Conference Dedicated to the Centennial Anniversary of Lev Semenovich Pontryagin. Lomonosov Moscow State University. Moscow*, page 7, 2008.

[BV05] François Bolley and Cédric Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales de la Faculté des sciences de Toulouse: Mathématiques*, 14(3):331–352, 2005.

[CAP24] Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *Forty-first International Conference on Machine Learning*, 2024.

[CCP23] David Chiang, Peter Cholak, and Anand Pillay. Tighter bounds on the expressivity of transformer encoders. In *International Conference on Machine Learning*, pages 5544–5562. PMLR, 2023.

[CGP16] Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Optimal transport over a linear dynamical system. *IEEE Transactions on Automatic Control*, 62(5):2137–2152, 2016.

[CL24] Sitan Chen and Yuanzhi Li. Provably learning a multi-head attention layer. *arXiv preprint arXiv:2402.04084*, 2024.

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

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

[Cor07] Jean-Michel Coron. *Control and nonlinearity*. Number 136 in Mathematical Surveys and Monographs. American Mathematical Soc., 2007.

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

[DBK+21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

[DBK24] Gbètondji JS Dovonon, Michael M Bronstein, and Matt J Kusner. Setting the Record Straight on Transformer Oversmoothing. *arXiv preprint arXiv:2401.04301*, 2024.

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

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[DM23] Alex Delalande and Quentin Merigot. Quantitative stability of optimal transport maps under variations of the target measure. *Duke Mathematical Journal*, 172(17):3321–3357, 2023.

[DMR19] Michel Duprez, Morgan Morancey, and Francesco Rossi. Approximate and exact controllability of the continuity equation with a localized vector field. *SIAM Journal on Control and Optimization*, 57(2):1284–1311, 2019.

[EGBO22] Karthik Elamvazhuthi, Bahman Gharesifard, Andrea L Bertozzi, and Stanley Osher. Neural ode control for trajectory approximation of continuity equation. *IEEE Control Systems Letters*, 6:3152–3157, 2022.

[EGKZ22] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022.

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

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

[GKPR24] Borjan Geshkovski, Hugo Koubbi, Yury Polyanskiy, and Philippe Rigollet. Dynamic metastability in the self-attention model. *arXiv preprint arXiv:2410.06833*, 2024.

[GLPR23] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *arXiv preprint arXiv:2312.10794*, 2023.

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

[GTLV22] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

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

[Har60] Philip Hartman. A lemma in the theory of structural stability of differential equations. *Proceedings of the American Mathematical Society*, 11(4):610–620, 1960.

[Har63] Philip Hartman. On the local linearization of differential equations. *Proceedings of the American Mathematical Society*, 14(4):568–573, 1963.

[JCP23] Yiheng Jiang, Sinho Chewi, and Aram-Alexandre Pooladian. Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space. *arXiv preprint arXiv:2312.02849*, 2023.

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

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

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

[KL09] Boris Khesin and Paul Lee. A nonholonomic Moser theorem and optimal transport. *Journal of Symplectic Geometry*, 7(4):381 – 414, 2009.

[KS09] Robert V Kohn and Sylvia Serfaty. Second-order PDE's and deterministic games. In *Proc. Internat. Congress Ind. Appl. Math.(ICIAM'07)*, pages 239–249, 2009.

[KZLD22] Anastasis Kratsios, Behnoosh Zamanlooy, Tianlin Liu, and Ivan Dokmanić. Universal approximation under constraints is possible with transformers. In *International Conference on Learning Representations*, 2022.

[LaS60] Joseph LaSalle. Some extensions of Liapunov's second method. *IRE Transactions on circuit theory*, 7(4):520–527, 1960.

[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 *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

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

[McC01] Robert J McCann. Polar factorization of maps on Riemannian manifolds. *Geometric & Functional Analysis GAFA*, 11(3):589–608, 2001.

[MWSB24] Pierre Marion, Yu-Han Wu, Michael Eli Sander, and Gérard Biau. Implicit regularization of deep residual networks towards neural ODEs. In *The Twelfth International Conference on Learning Representations*, 2024.

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

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

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

[PT22] Thierry Paul and Emmanuel Trélat. From microscopic to macroscopic scale equations: mean field, hydrodynamic and graph limits. *arXiv preprint arXiv:2209.08832*, 2022.

[PTB24] Aleksandar Petrov, Philip HS Torr, and Adel Bibi. Prompting a pretrained transformer can be a universal approximator. *arXiv preprint arXiv:2402.14753*, 2024.

[Rag24] Maxim Raginsky. Some Remarks on Controllability of the Liouville Equation. *arXiv preprint arXiv:2404.14683*, 2024.

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

[RBZ24] Domènec Ruiz-Balet and Enrique Zuazua. Control of neural transport for normalising flows. *Journal de Mathématiques Pures et Appliquées*, 181:58–90, 2024.

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

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

[SAP22] Michael Sander, Pierre Ablin, and Gabriel Peyré. Do residual neural networks discretize neural ordinary differential equations? *Advances in Neural Information Processing Systems*, 35:36520–36532, 2022.

[Sca23] Alessandro Scagliotti. Deep learning approximation of diffeomorphisms via linear-control systems. *Mathematical Control and Related Fields*, 13(3):1226–1257, 2023.

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

[Son13] Eduardo D Sontag. *Mathematical control theory: deterministic finite dimensional systems*, volume 6. Springer Science & Business Media, 2013.

[SP24] Michael E Sander and Gabriel Peyré. Towards understanding the universality of transformers for next-token prediction. *arXiv preprint arXiv:2410.03011*, 2024.

[SV16] Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

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

[TG22] Paulo Tabuada and Bahman Gharesifard. Universal approximation power of deep residual neural networks through the lens of control. *IEEE Transactions on Automatic Control*, 68(5):2715–2728, 2022.

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

[Ver14] Sergio Verdú. Total variation distance and the distribution of relative information. In *2014 information theory and applications workshop (ITA)*, pages 1–3. IEEE, 2014.

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

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

[WAW+24] Xinyi Wu, Amir Ajorlou, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the Role of Attention Masks and LayerNorm in Transformers. *arXiv preprint arXiv:2405.18781*, 2024.

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

[WW24] Mingze Wang and E Weinan. Understanding the expressive power and mechanisms of transformer for sequence modeling. *arXiv preprint arXiv:2402.00522*, 2024.

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

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

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

Borjan Geshkovski
Inria & Laboratoire Jacques-Louis Lions
Sorbonne Université
4 Place Jussieu
75005 Paris, France
e-mail: borjan.geshkovski@inria.fr

Philippe Rigollet
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Ave
Cambridge 02139 MA, United States
e-mail: rigollet@math.mit.edu

Domènec Ruiz-Balet
CEREMADE, UMR CNRS 7534
Université Paris-Dauphine, Université PSL
Pl. du Maréchal de Lattre de Tassigny
75016 Paris, France
e-mail: domenec.ruiz-i-balet@dauphine.psl.eu