Approximate Controllability of Continuity Equation of Transformers

Daniel Owusu Adu, *Member, IEEE*, and Bahman Gharesifard, *Senior Member, IEEE*

***Abstract*—Building on the recent work by Geshkovski et al. (2023) which provides an interacting particle system interpretation of Transformers with a continuous-time evolution, we study the controllability attributes of the corresponding continuity equation across the landscape of probability space curves. In particular, we consider the parameters of the Transformer's continuous-time evolution as control inputs. We prove that given an absolutely continuous probability measure and a non-local Lipschitz velocity field that satisfy a continuity equation, there exist control inputs such that the measure and the non-local velocity field of the Transformer's continuous-time evolution approximate them, respectively, in the p-Wasserstein and L^p-sense, where $1 \leq p < \infty$.**

***Index Terms*—Distributed parameter systems, machine learning, neural networks.**

I. Introduction

REMARKABLY, the recent work [1] shows that Transformers, which needless to say have revolutionized machine learning and beyond [2], [3], [4], [5], can be viewed as interacting particle systems. To be more precise, in [1], the authors considered the dynamics

$$\dot{x}_i(t) = \mathbf{P}_{x_i(t)}\left(\sum_{j=1}^{n} A_j\big(t, x_i(t), x_j(t)\big)V(t)x_j(t)\right). \tag{1}$$

Here:

1) $\mathbf{P}_{x_i} : \mathcal{S}^{d-1} \to \mathrm{T}_{x_i}\mathcal{S}^{d-1}$, defined as

$$\mathbf{P}_{x_i} y = y - \langle x_i, y\rangle x_i \tag{2}$$

represents the projection of $y \in \mathcal{S}^{d-1} \subset \mathbb{R}^d$ the unit sphere onto the tangent space $\mathrm{T}_{x_i}\mathcal{S}^{d-1}$. This projection map ensures that the relative positions of neighbouring states influence the dynamics of each state.

2) The self-attention $A_j(t, x_i(t), x_j(t))$ captures the importance or attention given by the i-th state to the j-th state relative to sequence of states $(x_i(t))_{i\in[n]} := (x_1(t)\ldots, x_n(t)) \in (\mathbb{R}^d)^n$, where $[n] := \{1, 2, \ldots, n\} \subset \mathbb{Z}$, at time t. Specifically,

$$A_j\big(t, x_i(t), x_j(t)\big) := \frac{e^{\langle Q(t)x_i(t), M(t)x_j(t)\rangle}}{\sum_{l=1}^{n} e^{\langle Q(t)x_i(t), M(t)x_l(t)\rangle}},$$

where $Q(t) \in \mathbb{R}^{p\times d}$ and $M(t) \in \mathbb{R}^{p\times d}$ determine the influence of the neighboring states.

3) The matrix $V(t) \in \mathbb{R}^{d\times d}$ scales the dot products in the self-attention mechanism, determining the strength of interactions among the states.

The analysis in [1] shows that due to the layer normalization property, the states evolve over the unit sphere $(x_i(t))_{i\in[n]} \in (\mathcal{S}^{d-1})^n$, for $t \geq 0$. Moreover, they showed that (1) inherits some clustering property almost surely, under some suitable condition.

Related to our work is the continuity equation

$$\partial_t \mu_t(x) + \nabla \cdot \left(\mathbf{P}_x\left(\int e^{\beta\langle x,y\rangle} y d\mu_t(y)\right)\mu_t(x)\right) = 0, \tag{3}$$

on $\mathbb{R}_{\geq 0} \times \mathcal{S}^{d-1}$, with initial distribution $\mu_0 \in \mathcal{P}(\mathcal{S}^{d-1})$, where $\beta > 0$ is fixed and determined by the magnitude of Q and M. In [1, Lemma 3.5] it is shown that (3) is a Wasserstein gradient flow of some energy function on $\mathcal{P}(\mathcal{S}^{d-1})$. The interacting particle systems interpretation of Transformers is also investigated in [6], where the clustering properties of the model is investigated, and it is shown that the absence of layer normalization leads to system instability.

This letter adopts a control-theoretic perspective. In particular, we consider

$$\partial_t \mu_{it}(x_i) + \nabla \cdot (F[\mu_i, W, A_i, V, b](t, x_i)\mu_{it}(x_i)) = 0, \tag{4}$$

with initial distributions $(\mu_{i0})_{i\in[n]} \in (\mathcal{P}(\mathcal{S}^{d-1}))^n$, where

$$F[\mu_i, W, A_i, V, b](t, x_i) := \mathbf{P}_{x_i}\Bigg(W(t)$$
$$\sigma\left(\sum_{k=1}^{2}\int_{\mathcal{S}^{d-1}} A_{i,k}(t, x_i, y)V_k(t)y d\mu_{it}(y) + b(t)\right)\Bigg) \tag{5}$$

is the trainable vector field and $A_i = (A_{i,1}, A_{i,2})$, where

$$A_{i,k}(t, x_i, y) = \frac{e^{\langle Q_k(t)x_i, M_k(t)y\rangle}}{\int_{\mathcal{S}^{d-1}} e^{\langle Q_k(t)x, M_k(t)y\rangle} d\mu_{it}(y)}, \tag{6}$$

Manuscript received 8 March 2024; revised 4 May 2024; accepted 23 May 2024. Date of publication 27 May 2024; date of current version 13 June 2024. This work was supported in part by the U.S. Department of Commerce under Grant BS123456. Recommended by Senior Editor P. Tesi. *(Corresponding author: Daniel Owusu Adu.)*

Daniel Owusu Adu is with the Faculty of Mathematics, University of Georgia, Athens, GA 30605 USA (e-mail: daniel.adu@uga.edu).

Bahman Gharesifard is with the Department of Electrical and Computer Engineering, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: gharesifard@ucla.edu).

Digital Object Identifier 10.1109/LCSYS.2024.3406255

2475-1456 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: UCLA Library. Downloaded on June 06,2025 at 21:31:21 UTC from IEEE Xplore. Restrictions apply.

with $k \in \{1, 2\}$ and $i \in [n]$. Here, $W(t) \in \mathbb{R}^{d\times d}$ and $b(t) \in \mathbb{R}^d$ are the weight matrix and bias vector for the neural network, respectively, and σ is the activation function. The specific σ will be made explicit later. Note that if μ_i has density f_i, then $F[\mu_i, W, A_i, V, b](t, x_i)$ in (5) is not uniquely determined by the value of $f_i(x_i)$ but rather the value of f_i on the whole sphere $\mathcal{S}^{d-1}$. Therefore, F in (5) is non-local. If μ_i in (5) is an empirical measure, then

$$F[\mu_i, W, A_i, V, b](t, x_i) \coloneqq \mathbf{P}_{x_i}\left(W(t)\sigma\left(\sum_{k=1}^{2}\sum_{j=1}^{n} A_{i,k}\big(t, x_i, x_j\big)V_k(t)x_j + b(t)\right)\right)$$

is the complete feed-forward layer (see [1, Sec. 2.3.2]). Note that since the trainable parameters $W(t), V_k(t) \in \mathbb{R}^{d\times d}, Q_k(t), M_k(t) \in \mathbb{R}^{p\times d}$ and $b(t) \in \mathbb{R}^d$ in (5) are independent of $i \in [n]$, the system in (4)–(5) is the i-th component of a sequence. From this point on, we sometimes use the shorthand notations

$$F \coloneqq (F_i)_{i\in[n]} \quad \text{where} \quad F_i \coloneqq F[\mu_i, W, A_i, V, b]. \tag{7}$$

The problem under study in this letter is the following:

Problem 1.1: Given $(\chi, \nu) \coloneqq (\chi_i, \nu_i)_{i\in[n]}$, a sequence of Lipschtiz velocity fields and absolutely continuous probability measures, where each pair (ν_i, χ_i) satisfies the continuity equation of the form

$$\partial_t \nu_{it} + \nabla \cdot (\chi_i[\nu_{it}]\nu_{it}) = 0, \tag{8}$$

does there exist control inputs $W(t), V_k(t) \in \mathbb{R}^{d\times d}, Q_k(t), M_k(t) \in \mathbb{R}^{p\times d}$ and $b(t) \in \mathbb{R}^d$ such that the sequence $(F, \mu) \coloneqq (F_i, \mu_i)_{i\in[n]}$, where each (F_i, μ_i) in (4), approximate (χ, ν) in some proper sense?

The main objective of this letter is to provide an answer under appropriate regularity assumptions. Before we state this result, it is important to point out the technical difference between our work and that of Neural ODEs, for instance, in [7], [8], [9], [10], [11]. Beyond the fact that in the Neural ODE settings, one is often concerned with approximating one state, whereas here we deal with sequences, which somehow mimic ensemble control settings [12], complication arises from the fact that the velocity fields in (5) and (8) are *non-local* [13], [14], [15], [16]. Nevertheless, we show that $(F, \mu) = (F_i, \mu_i)_{i\in[n]}$, where each (F_i, μ_i) in (4), approximate $(\chi, \nu) \coloneqq (\chi_i, \nu_i)_{i\in[n]}$, where each pair (χ_i, ν_i) in (8) in some proper sense. As a result of the inherent non-local velocity fields F and χ, the approximation of the measures ν is in p-Wasserstein sense and that of the velocity fields χ is in L^p-sense, where $1 \leq p < \infty$. For simplicity, we only deal with the case of 1-Wasserstein sense. We assert that akin to the significant impact Neural ODEs have had on the understanding of performance and training of neural networks, it is reasonable to anticipate that the novel transformer model will assume a comparable role.

This letter is organized as follows; in Section II, we state some Preliminary result on the existence of a general continuity equation with a non-local velocity field. We state the main result in Section III and follow with the proof in Section IV.

II. Preliminaries on Transport PDEs With Non-Local Velocities

This section provides some mathematical background that will be used throughout this letter. The following notations will be needed: let $L^\infty(\mathbb{R}^d; \mathbb{R}^d)$ denote the space of essentially bounded functions from $\mathbb{R}^d$ to $\mathbb{R}^d$, $L^\infty_{\mathrm{loc}}(\mathbb{R}; \mathbb{R})$ be the space of locally essentially bounded functions from $\mathbb{R}$ to $\mathbb{R}$ and $C(\mathbb{R}^d; \mathbb{R}^d)$ be the space of continuous functions from $\mathbb{R}^d$ to $\mathbb{R}^d$. Let $\mathcal{P}(\mathbb{R}^d)$ and $\mathcal{P}_{ac}(\mathbb{R}^d)$ be the set of probability and absolutely continuous probability measures on $\mathbb{R}^d$, respectively. We define the metric $\mathcal{W}_1$ on $\mathcal{P}(\mathbb{R}^d)$ as

$$\mathcal{W}_1(\mu_0, \mu_f) \coloneqq \sup\left\{\int_{\mathbb{R}^d} f d(\mu_0 - \mu_f) : f \in C^\infty\left(\mathbb{R}^d\right) \cap \mathrm{Lip}_1\left(\mathbb{R}^d\right)\right\}, \tag{9}$$

where $\mathrm{Lip}_1(\mathbb{R}^d)$ is the set of Lipschitz functions on $\mathbb{R}^d$ with Lipschitz constant less than 1. It is well-known, see for instance [17, Th. 7.12], that the topology induced by $\mathcal{W}_1$ on $\mathcal{P}(\mathbb{R}^d)$ coincides[1] with weak convergence on $\mathcal{P}(\mathbb{R}^d)$. Suppose that $T : \mathbb{R}^d \to \mathbb{R}^d$ is a measurable map, and $\mu_0, \mu_f \in \mathcal{P}(\mathbb{R}^d)$. Then the pushforward of μ_0 ala T is denoted by $T_\# \mu_0 = \mu_f$ and given by

$$\int_{\mathbb{R}^d} g(T(x)) d\mu_0(x) = \int_{\mathbb{R}^d} g(y) d\mu_0\left(T^{-1}(y)\right) = \int_{\mathbb{R}^d} g(y) d\mu_f \tag{10}$$

where $g \in L^1(\mathbb{R}^d, \mu_f)$ is any integrable function and as usual $L^1(\mathbb{R}^d, \mu_f)$ is the space of μ_f-integrable functions defined on $\mathbb{R}^d$.

We collect preliminary results of general non-local transport PDE [13], [14], [15], [16]. We say that a given pair (ϑ, ς) satisfies (8) if the function

$$t \mapsto \int_{\mathbb{R}^d} f(t, x) d\varsigma_t$$

is absolutely continuous for every $f \in C^\infty_c(\mathbb{R}_{\geq 0} \times \mathbb{R}^d; \mathbb{R})$ and

$$\int_{\mathbb{R}^d} f(t, x) d\varsigma_t(x) dx - \int_{\mathbb{R}^d} f(0, x) d\varsigma_0 = \int_0^t \int_{\mathbb{R}^d} (\partial_t f(s, x) + \nabla f(s, x) \cdot \vartheta[\varsigma_s](s, x)) d\varsigma_s(x) ds$$

holds, for almost every $t \in \mathbb{R}_{\geq 0}$. To guarantee the existence of a solution to (8), following from [13], [14], [15], [16], the following assumptions are needed:

Assumption 1: There exists positive functions $L_1, L_2, K \in L^\infty_{\mathrm{loc}}(\mathbb{R}_{\geq 0}; \mathbb{R})$ such that

$$\vartheta : C\left(\mathbb{R}_{\geq 0}; \mathcal{P}(\mathbb{R}^d)\right) \to C\left(\mathbb{R}_{\geq 0} \times \mathbb{R}^d; \mathbb{R}^d\right) \cap L^\infty\left(\mathbb{R}_{\geq 0} \times \mathbb{R}^d; \mathbb{R}^d\right)$$

satisfies:

1) the inequality

$$\|\vartheta[\varsigma_t](t, x) - \vartheta[\varsigma_t](t, y)\|_{\mathbb{R}^d} \leq L_1(t)\|x - y\|_{\mathbb{R}^d}, \tag{11}$$

for all $\varsigma_t \in \mathcal{P}(\mathbb{R}^d)$ and $x, y \in \mathbb{R}^d$ and $t \in \mathbb{R}_{\geq 0}$.

[1]The distance $\mathcal{W}_1$ metrizes the weak convergence of measures only if the measure has finite first moment. This condition is satisfied whenever the measures have compact support.

Authorized licensed use limited to: UCLA Library. Downloaded on June 06,2025 at 21:31:21 UTC from IEEE Xplore. Restrictions apply.

2) the inequality

$$\|\vartheta[\varsigma_t](t,x)\|_{\mathbb{R}^d} \le L_2(t)\big(1+\|x\|_{\mathbb{R}^d}\big), \tag{12}$$

for all $\varsigma_t \in \mathcal{P}(\mathbb{R}^d)$ and $x, y \in \mathbb{R}^d$ and $t \in \mathbb{R}_{\ge 0}$.

3) the inequality

$$\|\vartheta[\varsigma_0]-\vartheta[\varsigma_1]\|_{L^\infty(\mathbb{R}_{\ge 0};C^0(\mathbb{R}^d))} \le K(t)\mathcal{W}_1(\varsigma_0,\varsigma_1) \tag{13}$$

for all $\varsigma_0, \varsigma_1 \in \mathcal{P}(\mathbb{R}^d)$ and $t \in \mathbb{R}_{\ge 0}$.

It is well known (see for instance [16, Th. 2.3] and reference therein) that if ϑ satisfies Assumption 1, there exists a unique solution $\varsigma \in C(\mathbb{R}_{\ge 0}; \mathcal{P}_{ac}(\mathbb{R}^d))$ to (8), whenever $\varsigma_0 \in \mathcal{P}_{ac}(\mathbb{R}^d)$.

III. Main Results

This section presents the key findings of this letter. To this end, we fix the activation function $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ to be the rectified linear function (ReLU) applied to each component of a vector $x_i = (x_{i1}, \dots, x_{i,d}) \in \mathbb{R}^d$:

$$\sigma(x_i) = \mathrm{ReLU}(x_i) \coloneqq (\mathrm{ReLU}(x_{i1}), \dots, \mathrm{ReLU}(x_{id})), \tag{14}$$

where $\mathrm{ReLU}(x_{ij}) \coloneqq \max\{x_{ij}, 0\}, \quad j \in [d]$. Under the activation function, [18] has laid the groundwork by establishing that Transformers characterized as

$$\begin{aligned}\mathcal{T} \coloneqq &\big\{(x_i)_{i\in[n]} \mapsto (F[\mu_i, W, A_i, V, b](x_i))_{i\in[n]} \\ &\coloneqq (F[\mu_i, W, A_i, V, b](x_i+\omega_i))_{i\in[n]} : W, V_k \in \mathbb{R}^{d\times d}, \\ &\quad Q_k, M_k \in \mathbb{R}^{p\times d} \text{ and } b, \omega_i \in \mathbb{R}^d, \text{where } k \in \{1,2\}\big\}, \end{aligned} \tag{15}$$

where μ_i is a given empirical measure, serves as universal approximators of continuous sequence-to-sequence functions defined on a compact domain $f \in C((\mathcal{S}^{d-1})^n; (\mathbb{R}^d)^n)$, where

$$\begin{aligned}&C\Big((\mathcal{S}^{d-1})^n; (\mathbb{R}^d)^n\Big) \coloneqq \\ &\quad\Big\{f : \Big(\mathcal{S}^{d-1}\Big)^n \to \Big(\mathbb{R}^d\Big)^n : f \text{ is continuous functions}\Big\}.\end{aligned}$$

In particular, we later use the following universal approximation result, presented as [18, Th. 3].

Theorem 1: Let $1 \le p < \infty$ and $\epsilon > 0$. Then for any $f \in C((\mathcal{S}^{d-1})^n; (\mathbb{R}^d)^n)$, there exists $g \in \mathcal{T}$ such that

$$\begin{aligned}d(f,g) \coloneqq \Bigg(\int_{(\mathcal{S}^{d-1})^n} &\|f\big((x_i)_{i\in[n]}\big) \\ &- g\big((x_i)_{i\in[n]}\big)\|^p dx_1 \dots dx_n\Bigg)^{\frac{1}{p}} \le \epsilon.\end{aligned} \tag{16}$$

We state here that the universal approximation result also holds when $k \ge 3$ in (15), see [18]. We now state the main result of this letter.

Theorem 2: Consider $(\chi, \nu) \coloneqq (\chi_i, \nu_i)_{i\in[n]}$, where for each $i \in [n]$, we have that χ_i satisfies Assumption 1 and $\nu_i \in C(\mathbb{R}_{\ge 0}; \mathcal{P}_{ac}(\mathcal{S}^{d-1}))$ is the corresponding unique solution to (8) with initial measure $\nu_{i0} \in \mathcal{P}_{ac}(\mathcal{S}^{d-1})$. Then, for any $\epsilon > 0$, there exist a final time $t_f > 0$ and piecewise constant in time control inputs

$$W, V_k \in L^\infty\Big([0,t_f]; \mathbb{R}^{d\times d}\Big) \quad Q_k, M_k \in L^\infty\Big([0,t_f]; \mathbb{R}^{p\times d}\Big),$$

where $k \in \{1,2\}$ and $b, \omega \in L^\infty([0,t_f]; \mathbb{R}^d)$ such that the pair of sequence $(F,\mu) \coloneqq (F_i, \mu_i)_{i\in[n]}$, where each (F_i, μ_i) is characterized in (4)–(5), approximates (χ, ν) in the following sense:

$$d(F(t,\cdot), \chi(t,\cdot)) \le \epsilon,$$

where $\chi(t,(x_i)_{i\in[n]}) \coloneqq (\chi_i[\nu_{ti}](t,x_i))_{i\in[n]}$, for all $t \in [0,t_f]$ and

$$\sup_{t\in[0,t_f]} W_1(\mu_{it}, \nu_{it}) \le \epsilon,$$

for all $i \in [n]$.

The theorem ensures that there exist appropriate control inputs such that (F,μ) characterized in (4)–(5) can approximate a given sequence of velocity fields and measures that solves (8).

IV. Proofs of Main Result

This section is devoted to presenting the proof of Theorem 2. The proof is organized into two subsections: firstly, we provide some properties on a generic sequence $\{(\chi_i^m, \nu_i^m)\}_{m\ge 1}$, where χ_i^m is piecewise constant in time vector field that approximate the solutions (χ_i, ν_i) to the continuity equations (8) in Proposition 2. Next, we prove in Corollary 1 that the set of control inputs W, V_k, Q_k, M_k, where $k \in \{1,2\}$ and b, ω can be selected in such a way as to ensure that the Transformer in (4) generates a sequence (F^N, μ^N), where F^N is in (15) that have the required properties in the former subsection. This core idea is inspired from [8]; however, in our derivations, the technical details differ as we deal with nonlocal vector fields and approximating sequence-to-sequence functions. For that part, we utilize Theorem 1 and modified techniques in [13], [14], [15], [16], [19].

A. Approximating the Solutions to (8)

We start with a stepping stone result on the support of the solutions to (8).

Proposition 1: Let $i \in [n]$ and consider the continuity equation (8) where each χ_i satisfies Assumption 1 and $\nu_{i0} \in \mathcal{P}_{ac}(\mathcal{S}^{d-1})$. There exists a final time $t_f > 0$ such that the solution $\nu_i \in C(\mathbb{R}_{\ge 0}; \mathcal{P}_{ac}(\mathbb{R}^d))$ to (8), satisfies

$$\mathrm{supp}(\nu_{it}) \subset \mathcal{S}^{d-1}, \tag{17}$$

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

Proof: Suppose that χ_i in (8) satisfies Assumption 1 and $\nu_{i0} \in \mathcal{P}_{ac}(\mathcal{S}^{d-1})$. Following from [16, Th. 2.3] the unique solution $\nu_i \in C(\mathbb{R}_{\ge 0}; \mathcal{P}_{ac}(\mathbb{R}^d))$ is characterized as

$$\nu_{it} = \Phi_i(t,\cdot)_\# \nu_{i0}, \tag{18}$$

where $\Phi_i(t,\cdot)$ is the diffeomorphic flow on $\mathbb{R}^d$ that satisfies

$$\partial_t \Phi_i(t,x) = \chi_i[\nu_{it}](t, \Phi_i(t,x)) \quad \text{and} \quad \Phi_i(0,x) = x, \tag{19}$$

on $\mathbb{R}_+ \times \mathbb{R}^d$, where $i \in [n]$. We proceed to show that there exists t_f such that (19) admits a unique solution on $[0,t_f]$. To this end, let $T > 0$ be fixed. From (12), since $\|\chi_i[\nu_{it}](t,x)\|_{\mathbb{R}^d} \le 2\bar{L}_{i2}$ holds for all $(t,x) \in [0,T] \times \mathcal{S}^{d-1}$, where $\bar{L}_{i2} = \mathrm{essup}_{t\in[0,T]} L_{i2}(t)$. If $t_f \le \min_{i\in[n]} \frac{1}{2\bar{L}_{i2}}$, then from [20, Ch. 1, Th. 1], for any $x \in \mathcal{S}^{d-1}$, we have that (19) admits a solution on $[0,t_f]$. For uniqueness, since the Lipschitz

Authorized licensed use limited to: UCLA Library. Downloaded on June 06,2025 at 21:31:21 UTC from IEEE Xplore. Restrictions apply.

condition (11) holds, for any $(t,x),(t,y)\in[0,T]\times\mathcal{S}^{d-1}$, from [20, Ch. 1, Th. 2] we have that the unique solution satisfies $\Phi_i(t,\text{supp}(\nu_{i0}))\subset\mathcal{S}^{d-1}$, where $t\in[0,t_f]$.

We now show (17). Let $z\in\text{supp}(\nu_{it})$. Then, from (18) and (10), since $\nu_{it}(\mathcal{B}_z(\epsilon))=\nu_{i0}(\Phi_i^{-1}(t,\mathcal{B}_z(\epsilon)))>0$ holds, for all $\epsilon>0$, where $\mathcal{B}_z(\epsilon)\subset\mathbb{R}^d$ is the ball with radius ϵ centred at $z\in\text{supp}(\nu_{it})$, we have that $\Phi_i^{-1}(t,\mathcal{B}_z(\epsilon))\subset\text{supp}(\nu_{i0})$ holds, for all $\epsilon>0$. Since $\Phi(t,\cdot)$ is a diffeomorphic flow on $\mathbb{R}^d$ and $t\mapsto\Phi(t,\cdot)$ is continuous map, we have that $\mathcal{B}_z(\epsilon)\subset\Phi_i(t,\text{supp}(\nu_{i0}))$, for all $\epsilon>0$. Since $\Phi_i(t,\text{supp}(\nu_{i0}))\subset\mathcal{S}^{d-1}$, holds, for all $t\in[0,t_f]$, we have that $\mathcal{B}_z(\epsilon)\subset\mathcal{S}^{d-1}$, for all $\epsilon>0$. This completes the proof. ∎

Next, we provide an approximation result for (8).

Proposition 2: Suppose, for each $i\in[n]$, we have that χ_i satisfies Assumption 1 and $\nu_i\in C([0,t_f];\mathcal{P}_{ac}(\mathcal{S}^{d-1}))$ is the corresponding solution in (8) with initial distribution $\nu_{i0}\in\mathcal{P}_{ac}(\mathcal{S}^{d-1})$. Then, for each $i\in[n]$, there exists $\{(\chi_i^m,\nu_i^m)\}_{m\geq1}$ such that each χ_i^m is piecewise constant in time and (χ_i^m,ν_i^m) solves (8) with initial distribution $\nu_{i0}\in\mathcal{P}_{ac}(\mathcal{S}^{d-1})$ and $\chi_i^m: C([0,t_f];\mathcal{P}_{ac}(\mathcal{S}^{d-1}))\to C([0,t_f]\times\mathcal{S}^{d-1};\mathcal{S}^{d-1})\cap L^\infty([0,t_f]\times\mathcal{S}^{d-1};\mathcal{S}^{d-1})$ uniformly converges to χ_i and $\nu_i^m\in C([0,t_f];\mathcal{P}_{ac}(\mathcal{S}^{d-1}))$ weakly converges to ν_i.

Proof: Let $0:=t_0^m<t_1^m<\cdots<t_{2^m-1}^m:=t_f$, where $t_\ell^m=\ell 2^{-m}t_f$ be a regular partition of $[0,t_f]$ into 2^m subintervals. Given (χ,ν), let

$$\chi_i^m[\nu_{it}]:=\chi_i\Big[\nu_{it_l^m}\Big], \tag{20}$$

where $t\in[t_\ell^m,t_{\ell+1}^m]$, then χ_i^m is piecewise constant in time velocity field on $C([0,t_f];\mathcal{P}_{ac}(\mathcal{S}^{d-1}))$. Since χ_i satisfies Assumption 1, we have that χ_i^m satisfies

$$\begin{aligned}\big\|\chi_i^m[\nu_{it}](t,x)-\chi_i^m[\nu_{it}](t,y)\big\|_{\mathbb{R}^d}\leq L_{i1}\|x-y\|_{\mathbb{R}^d},\quad\text{and}\\ \big\|\chi_i^m[\nu_{it}](t,x)\big\|_{\mathbb{R}^d}\leq 2L_{i2},\end{aligned} \tag{21}$$

where $L_{i\alpha}=\text{esssup}_{t\in[0,t_f]}L_{i\alpha}(t)$, with $\alpha\in\{1,2\}$, hold for all $x,y\in\mathcal{S}^{d-1}$. Therefore, the unique solution ν_i^m to

$$\partial_t\nu_i+\nabla\cdot\big(\chi_i^m[\nu_i]\nu_i\big)=0\quad\text{and}\quad\nu_{i0}^m=\nu_{i0}$$

is characterized by

$$\begin{aligned}\nu_{it}^m&=\Phi_i^m(t,\cdot)_\#\nu_{it_l^m},\\ \partial_t\Phi_i^m(t,x)&=\chi_i\Big[\nu_{it_l^m}\Big]\big(t_\ell^m,\Phi_i^m(t,x)\big),\end{aligned} \tag{22}$$

and $\Phi_i^m(t_l^m,x):=\Phi_i(t_l^m,x)$, for all $t\in[t_\ell^m,t_{\ell+1}^m]$, where Φ_i is as given in (18). We show that, for each $i\in[n]$, the pair (χ_i^m,ν_i^m), characterized in (20) and (22), respectively, converges to (χ_i,ν_i) in some sense.

We show that for any $\epsilon>0$, there exists $N_0\in\mathbb{N}$ such that for any $m\geq N_0$, we have

$$\big\|\chi_i^m-\chi_i\big\|\leq\epsilon,$$

where

$$\big\|\chi_i^m-\chi_i\big\|:=\sup_{\nu\in C\left(\left[t_\ell^m,t_{\ell+1}^m\right];\mathcal{P}(\mathcal{S}^{d-1})\right)}\big\|\chi_i^m[\nu_i]-\chi_i[\nu_i]\big\|_{L^\infty} \tag{23}$$

and $L^\infty:=L^\infty([t_\ell^m,t_{\ell+1}^m]\times\mathcal{S}^{d-1};\mathcal{S}^{d-1})$. To this end, using (20) and since

$$\big\|\chi_i^m[\nu_{it}]-\chi_i[\nu_{it}]\big\|_{L^\infty}=\Big\|\chi_i\Big[\nu_{it_l^m}\Big]-\chi_i[\nu_{it}]\Big\|_{L^\infty},$$

for all $t\in[t_\ell^m,t_{\ell+1}^m]$, using (13), we have that

$$\begin{aligned}\big\|\chi_i^m[\nu_{it}]-\chi_i[\nu_{it}]\big\|_{L^\infty}&\leq K_i\mathcal{W}_2\Big(\nu_{it_l^m},\nu_{it}\Big)\\ &\leq K_ia_i\frac{t_f}{2^m},\end{aligned} \tag{24}$$

for all $\nu_i\in C([t_\ell^m,t_{\ell+1}^m];\mathcal{P}(\mathcal{S}^{d-1}))$, where the second inequality follows from the fact that the solution curve $t\mapsto\nu_{it}$ in (8) is a_i-Lipschitz continuous, for some positive constant $a_i\in\mathbb{R}$ (see for instance [15, Proposition 5]). Therefore, we conclude that χ_i^m uniformly converges to χ_i.

To show that the sequence $(\nu_i^m)_{m\geq1}$ weakly converges to ν_i, we first show that the family of functions $t\mapsto\nu_{it}^m$ is equi-continuous and equi-bounded. To this end, equi-bounded follows from the fact that from Proposition 1, we have that $\text{supp}(\nu_{it}),\text{supp}(\nu_{it}^m)\subset\mathcal{S}^{d-1}$, for all $t\in[0,t_f]$ and $m\geq1$. For equi-continuous, from (20), since $\|\chi_i^m\|\leq2L_{i2}$, following from [15, Proposition 1] we have that

$$\mathcal{W}_1\Big(\nu_{it}^m,\nu_{it_l^m}^m\Big)\leq2L_{i2}\big|t-t_\ell^m\big|. \tag{25}$$

Therefore, the family of functions $t\mapsto\nu_{it}^m$ is equi-Lipschitz and hence equi-continuous. Hence, by Arzela-Ascoli Theorem, we have that the sequence $(\nu_i^m)_{m\geq1}$ admits a subsequence $(\nu_i^{m_r})_{r\geq1}$ that weakly converges to $\tilde{\nu}_i$. We proceed to show that the limit measure $\tilde{\nu}_i$ satisfies

$$\int_0^{t_f}\int_{\mathcal{S}^{d-1}}\big(\partial_tf(t,x)+\nabla f(t,x)\cdot\chi_i\big[\tilde{\nu}_{it}\big](t,x)\big)d\tilde{\nu}_{it}(x)dt=0, \tag{26}$$

for every $f\in C^\infty([0,t_f]\times\mathcal{S}^{d-1};\mathbb{R})$, where $f(t_f,x)=0$ for all $x\in\mathcal{S}^{d-1}$. We prove this by showing that the following statements hold:

1)
$$\lim_{r\to\infty}\int_0^{t_f}\int_{\mathcal{S}^{d-1}}\partial_tf(t,x)d\big(\tilde{\nu}_{it}(x)-\nu_{it}^{m_r}(x)\big)dt=0.$$

2)
$$\begin{aligned}\lim_{r\to\infty}\int_0^{t_f}\int_{\mathcal{S}^{d-1}}\nabla f(t,x)\cdot(\chi_i\big[\tilde{\nu}_{it}\big](t,x)-\\ \chi_i^{m_r}\big[\nu_{it}^{m_r}\big](t,x))d\tilde{\nu}_{it}(x)dt=0.\end{aligned}$$

3)
$$\begin{aligned}\lim_{r\to\infty}\int_0^{t_f}\int_{\mathcal{S}^{d-1}}\nabla f(t,x)\cdot\chi_i^{m_r}\big[\nu_{it}^{m_r}\big](t,x)\\ d\big(\tilde{\nu}_{it}(x)-\nu_{it}^{m_r}(x)\big)dt=0.\end{aligned}$$

For Statement 1: using (9) and since

$$\begin{aligned}\left|\int_0^{t_f}\int_{\mathcal{S}^{d-1}}\partial_tf(t,x)d\big(\tilde{\nu}_{it}(x)-\nu_{it}^{m_r}(x)\big)dt\right|\leq\\ \max_{(t,x)\in[0,t_f]\times\mathcal{S}^{d-1}}\|\partial_tf(t,x)\|t_f\sup_{t\in[0,t_f]}\mathcal{W}_1\big(\tilde{\nu}_{it},\nu_{it}^{m_r}\big)\end{aligned}$$

and $\nu_{it}^{m_r}$ weakly converges to $\tilde{\nu}_{it}$, uniformly in $t\in[0,t_f]$, we conclude that Statement 1 holds.

For Statement 2, since

$$\left|\int_0^{t_f}\int_{\mathcal{S}^{d-1}}\nabla f(t,x)\cdot(\chi_i\big[\tilde{\nu}_{it}\big](t,x)-\right.$$

Authorized licensed use limited to: UCLA Library. Downloaded on June 06,2025 at 21:31:21 UTC from IEEE Xplore. Restrictions apply.

$$\chi_i^{m_r}[\nu_{it}^{m_r}](t,x))d\tilde{\nu}_{it}(x)dt\Big| \le \max_{(t,x)\in[0,t_f]\times\mathcal{S}^{d-1}} \|\partial_t f(t,x)\| \sum_{l=0}^{2^m-1}\int_{t_l}^{t_{l+1}}\int_{\mathcal{S}^{d-1}} \big\|\chi_i[\tilde{\nu}_{it}](t,x)-\chi_i^{m_r}[\nu_{it}^{m_r}](t,x)\big\|d\tilde{\nu}_{it}(x)dt,$$

from (13), we have that

$$\begin{aligned}\Big\|\chi_i[\tilde{\nu}_{it}]-\chi_i\Big[\nu_{it_l^{m_r}}^{m_r}\Big]\Big\| &\le K\mathcal{W}_1\Big(\tilde{\nu}_{it},\nu_{it_l^{m_r}}^{m}\Big)\\ &\le K\Big(\mathcal{W}_1(\tilde{\nu}_{it},\nu_{it}^{m_r})+\mathcal{W}_1(\nu_{it}^{m_r},\nu_{it_l^{m}}^{m_r})\Big)\\ &\le K\Big(\mathcal{W}_1\big(\tilde{\nu}_{it},\nu_{it}^{m_r}\big)+2L_{i2}\frac{t_f}{2^{m_r}}\Big),\end{aligned}$$

where the last term in the later inequality follows from (25). Since $\nu_{it}^{m_r}$ weakly converges to $\tilde{\nu}_{it}$, we have that Statement 2 holds.

For Statement 3, from (12), we have that

$$\Big|\int_0^{t_f}\int_{\mathcal{S}^{d-1}}\nabla f(t,x)\cdot\chi_i^{m_r}[\nu_{it}^{m_r}](t,x)d\big(\tilde{\nu}_{it}(x)-\nu_{it}^{m_r}(x)\big)dt\Big| \le \max_{(t,x)\in[0,t_f]\times\mathcal{S}^{d-1}}\|\partial_t f(t,x)\|2L_{i2}t_f\mathcal{W}_1\Big(\tilde{\nu}_{it},\nu_{it_l^{m_r}}^{m_r}\Big).$$

Using similar arguments in Statement 2, we conclude that Statement 3 holds. This completes the proof. ∎

B. Using Transformers to Approximate the Solutions to a Continuity Equation

We are now ready to present our main proof. We first state a corollary of the universal approximation result which we later combine with the observation made in the previous section to prove Theorem 2.

Corollary 1: Given $(\chi,\nu):=(\chi_i,\nu_i)_{i\in[n]}$, where

$$\chi\big(t,(x_i)_{i\in[n]}\big):=(\chi_i[\nu_{it}](t,x_i))_{i\in[n]}. \tag{27}$$

Suppose that χ_i satisfies Assumption 1 and $\nu_i \in C([0,t_f];\mathcal{P}_{ac}(\mathcal{S}^{d-1}))$ is the solution in (8) with initial distribution $\nu_{i0}\in\mathcal{P}_{ac}(\mathcal{S}^{d-1})$. Furthermore, suppose that the time component of χ_i remain constant on $[t_\ell^N,t_{\ell+1}^N]$, where $\ell\in\{0,\dots,2^N-1\}$ with $t_\ell^N:=\ell 2^{-N}t_f$. Then, there exist sequence of functions $F^N(t,\cdot)\in\mathcal{T}$, where $t\mapsto F^N(t,\cdot)$ is constant on the interval $[t_\ell^N,t_{\ell+1}^N]$, such that

$$\lim_{N\to\infty} d\big(F^N(t,\cdot),\chi(t,\cdot)\big)=0$$

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

Proof: Given (χ,ν), consider (27). Then, by assumption and from (11), since $(t,(x_i)_{i\in[n]}))\mapsto\chi(t,(x_i)_{i\in[n]})$ is constant on $[t_\ell^N,t_{\ell+1}^N]$ and Lipschitz over $(\mathcal{S}^{d-1})^n$, we have that $\chi\in C((\mathcal{S}^{d-1})^n;(\mathbb{R}^d)^n)$ on $[t_\ell^N,t_{\ell+1}^N]$. Therefore, from Theorem 1 given an empirical measure μ_i, we have that, for every $\epsilon>0$, there exists $F^N(t,\cdot)\in\mathcal{T}$, piecewise constant in time on $[0,t_f]$ such that $d(F^N(t,\cdot),\chi(t,\cdot))\le\epsilon$ holds, for all $t\in[t_\ell^N,t_{\ell+1}^N]$. This completes the proof. ∎

We are now ready to provide the proof of Theorem 2.

Proof of Theorem 2: Given $(\chi,\nu):=(\chi_i,\nu_i)_{i\in[n]}$, where for each $i\in[n]$, we have that χ_i satisfies Assumption 1 and $\nu_i\in C(\mathbb{R}_{\ge0};\mathcal{P}_{ac}(\mathcal{S}^{d-1}))$ is the corresponding unique solution to (8) with initial measure $\nu_{i0}\in\mathcal{P}_{ac}(\mathcal{S}^{d-1})$. From Proposition 2, we have that for each $i\in[n]$, there exists $\{(\chi_i^m,\nu_i^m)\}_{m\ge1}$, where the pair (χ_i^m,ν_i^m) is characterized in (20) and (22), respectively, such that

$$\lim_{m\to\infty}\|\chi_i^m-\chi_i\|=0,\quad\text{and}\quad\lim_{m\to\infty}\mathcal{W}_1\big(\nu_{it}^m,\nu_{it}\big)=0, \tag{28}$$

for all $t\in[0,t_f]$. Here $\|\cdot\|$ is defined in (23) and, for any $m\ge1$, we have that χ_i^m is constant in time on $[t_\ell^m,t_{\ell+1}^m]$ with $t_\ell^m=\ell 2^{-m}t_f$. Let $m\ge1$ be fixed. Then, from Corollary 1, given an empirical measure μ_i, there exists a sequence $\{F^N\}_{N\ge1}\subset\mathcal{T}$ in (15) such that

$$\lim_{N\to\infty} d\big(F^N(t,\cdot),\chi^m(t,\cdot)\big)=0 \tag{29}$$

holds, for all $t\in[0,t_f]$. Furthermore, from (7), we have that

$$\Big\{F^N:=\big(F_i^N\big)_{i\in[n]}\Big\}_{N\ge1}$$

admits the following characterization: for a given empirical measure μ_i, there exists piecewise time control inputs $W^N,V_k^N\in L^\infty([0,t_f];\mathbb{R}^{d\times d})$, $Q_k^N,M_k^N\in L^\infty([0,t_f];\mathbb{R}^{p\times d})$ and $b^N,\omega_i^N\in L^\infty([0,t_f];\mathbb{R}^d)$, where $k\in\{1,2\}$ such that for $N\ge1$, we have that

$$F_i^N(t,x_i):=\mathbf{P}_{x_i}\left(W^N(t)\sigma\left(\sum_{k=1}^{2}\sum_{j=1}^{n}A_{i,k}^N\big(t,x_i,x_j\big)V_k^N(t)x_j+b^N(t)\right)\right).$$

We proceed to generate the corresponding measures μ_i^N. To this end, we consider the piecewise constant in time vector field

$$\mathrm{F}_i^N(t,x_i):=\mathbf{P}_{x_i}\left(W^N(t)\sigma\left(\sum_{k=1}^{2}\int_{\mathcal{S}^{d-1}}A_{i,k}^N(t,x_i,y)V_k^N(t)yd\mu_{it}(y)+b^N(t)\right)\right), \tag{30}$$

and show that, for a fixed N and $i\in[n]$, we have that $\mathrm{F}_i^N(t,x_i)$ above satisfies Assumption 1.

To this end, since the projection map $\mathbf{P}$ in (2) is uniformly bounded and Lipschitz on $\mathcal{S}^{d-1}$, it is enough to show that

$$\mathrm{G}_i^N(t,x_i):=W^N(t)\sigma\left(\sum_{k=1}^{2}\int_{\mathcal{S}^{d-1}}A_{i,k}^N(t,x_i,y)V_k^N(t)yd\mu_{it}(y)+b^N(t)\right)$$

satisfies Assumption 1. For Statement 1 in Assumption 1, from (6) since

$$\nabla_{x_i}\mathrm{G}_i^N(t,x_i)=W^N(t)\sigma\left(\sum_{k=1}^{2}\int_{\mathcal{S}^{d-1}}\nabla_{x_i}A_{i,k}^N(t,x_i,y)V_k^N(t)yd\mu_{it}(y)\right)$$

for every $i\in[n]$ and $m\ge1$, we have that

$$\big\|\nabla_{x_i}\mathrm{G}_i^N\big\|_{L^\infty}\le 2\big\|W^N\big\|_{L^\infty}\left(\sum_{k=1}^{2}\big\|M_k^N\big\|_{L^\infty}\big\|Q_k^N\big\|_{L^\infty}\big\|V_k^N\big\|_{L^\infty}\right).$$

Therefore, we conclude that $x_i\mapsto\mathrm{G}_i^N(t,x_i)$ is Lipschitz.

Authorized licensed use limited to: UCLA Library. Downloaded on June 06,2025 at 21:31:21 UTC from IEEE Xplore. Restrictions apply.

For uniform boundedness, since

$$\left\| W^N(t)\sigma\left(\sum_{k=1}^{2}\int_{\mathcal{S}^{d-1}} A^N_{i,k}(t,x_i,y)V^N_k(t)y\,d\mu_{it}(y) + b^N(t)\right)\right\| \le \left\|W^N\right\|_{L^\infty}\left(\sum_{k=1}^{2}\left\|V^N_k\right\|_{L^\infty}+\left\|b^N\right\|_{L^\infty}\right)$$

we have that

$$\left\|\mathrm{G}^N_i\right\|_{L^\infty} \le \left\|W^N\right\|_{L^\infty}\left(\sum_{k=1}^{2}\left\|V^N_k\right\|_{L^\infty}+\left\|b^N\right\|_{L^\infty}\right).$$

This concludes that $\mu_i \mapsto \mathrm{G}^N_i$ is uniformly bounded.

Lastly, we show that $\mu_i \mapsto \mathrm{G}^N_i$ is Lipschitz. From (14), since for $x_i = (x_{i1},\dots,x_{id}) \in \mathcal{S}^{d-1}$ we have that $\mathrm{ReLU}(x_{ij})$ is Lipschitz and

$$\left\|\sum_{k=1}^{2}\int_{\mathcal{S}^{d-1}} A^N_{i,k}(t,x_i,y)V^N_k(t)y\,d(\mu_{it1}(y)-\mu_{it2}(y))\right\| \le \left(\sum_{k=1}^{2}\left\|A^N_{i,k}(t)\right\|_{L^\infty}\left\|V^N_k\right\|_{L^\infty}\right)\mathcal{W}_1(\mu_{it1},\mu_{it2}).$$

we conclude that $\mu_i \mapsto \mathrm{G}^N_i$ is Lipschitz. Therefore, $\mathrm{F}^N_i(t,x_i)$ satisfies Assumption 1. From [16, Th. 2.3], we have that there exists a unique solution μ^N_i to

$$\partial_t \mu_{it}(x_i) + \nabla\cdot\left(\mathrm{F}^N_i(t,x_i)\mu_{it}(x_i)\right) = 0, \tag{31}$$

with initial distribution $\mu_{i0} \in (\mathcal{P}_{ac}(\mathcal{S}^{d-1}))^n$, where $i \in [n]$. Furthermore, from (29), for a fixed $m \ge 1$, by Arzela-Ascoli Theorem, we have that

$$\lim_{N\to\infty} W_1\left(\mu^N_{it}, \nu^m_{it}\right) = 0 \tag{32}$$

uniformly in $t \in [0,t_f]$.

Since $W_1(\mu^N_{it},\nu_{it}) \le W_1(\mu^N_{it},\nu^m_{it}) + W_1(\nu^m_{it},\nu_{it})$, for all $m \ge 1$, from (32), we have that $0 \le \lim_{N\to\infty} W_1(\mu^N_{it},\nu_{it}) \le W_1(\nu^m_{it},\nu_{it})$, for all $m \ge 1$, uniformly in $t \in [0,t_f]$. From (28), we have that $\lim_{N\to\infty} W_1(\mu^N_{it},\nu_{it}) = 0$, uniformly in $t \in [0,t_f]$. Therefore, for every $\epsilon > 0$, there exists $N_0 \in \mathbb{N}$ such that for $N \ge N_0$, we have that $\sup_{t\in[0,t_f]} W_1(\mu^N_{it},\nu_{it}) \le \epsilon$, holds, for all $i \in [n]$. Similarly, since $d(F^N(t,\cdot),\chi(t,\cdot)) \le d(F^N(t,\cdot),\chi^m(t,\cdot)) + d(\chi^m(t,\cdot),\chi(t,\cdot))$, from (29), we conclude that given (χ,ν), there exists sequence of piecewise constant control inputs $W^N, V^N_k \in L^\infty([0,t_f];\mathbb{R}^{d\times d})$, Q^N_k, $M^N_k \in L^\infty([0,t_f];\mathbb{R}^{p\times d})$ and $b^N, \omega^N_i \in L^\infty([0,t_f];\mathbb{R}^d)$ such that $\lim_{N\to\infty} d(F^N(t,\cdot),\chi(t,\cdot)) = 0$, for all $t \in [0,t_f]$. This finishes the proof. ■

REFERENCES

[1] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, "A mathematical perspective on transformers," 2023, *arXiv:2312.10794*.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[3] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[4] M. Chen et al., "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.

[5] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–15.

[6] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, "The emergence of clusters in self-attention dynamics," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–55.

[7] D. Ruiz-Balet and E. Zuazua, "Neural ODE control for classification, approximation, and transport," *SIAM Rev.*, vol. 65, no. 3, pp. 735–773, 2023.

[8] K. Elamvazhuthi, B. Gharesifard, A. L. Bertozzi, and S. Osher, "Neural ODE control for trajectory approximation of continuity equation," *IEEE Control Syst. Lett.*, vol. 6, pp. 3152–3157, 2022.

[9] P. Tabuada and B. Gharesifard, "Universal approximation power of deep residual neural networks via nonlinear control theory," 2020, *arXiv:2007.06007*.

[10] A. Agrachev and A. Sarychev, "Control on the manifolds of mappings with a view to the deep learning," *J. Dyn. Control Syst.*, vol. 28, no. 4, pp. 989–1008, 2022.

[11] C. Cuchiero, M. Larsson, and J. Teichmann, "Deep neural networks, generic universal interpolation, and controlled ODEs," *SIAM J. Math. Data Sci.*, vol. 2, no. 3, pp. 901–919, 2020.

[12] J.-S. Li and N. Khaneja, "Control of inhomogeneous quantum ensembles," *Phys. Rev. A*, vol. 73, no. 3, 2006, Art. no. 30302.

[13] L. Ambrosio and W. Gangbo, "Hamiltonian ODEs in the wasserstein space of probability measures," *Commun. Pure Appl. Math.*, vol. 61, no. 1, pp. 18–53, 2008.

[14] H. Dong, "Well-posedness for a continuity equation with non-local velocity," *J. Funct. Anal.*, vol. 255, no. 11, pp. 3070–3097, 2008.

[15] B. Piccoli and F. Rossi, "Continuity equation with non-local velocity in Wasserstein spaces: Convergence of numerical schemes," *Acta Applicandae Mathematicae*, vol. 124, pp. 73–105, Apr. 2013.

[16] B. Piccoli, F. Rossi, and E. Trélat, "Control to flocking of the kinetic Cucker–Smale model," *SIAM J. Math. Anal.*, vol. 47, no. 6, pp. 4685–4719, 2015.

[17] C. Villani, *Optimal Transport: Old and New*, vol. 338. Berlin, Germany: Springer, 2009.

[18] C. Yun, S. Bhojanapalli, A. S. Rawat, S. J. Reddi, and S. Kumar, "Are transformers universal approximators of sequence-to-sequence functions?" 2019, *arXiv:1912.10077*.

[19] H. O. Fattorini, *Infinite Dimensional Optimization and Control Theory*, vol. 54. Cambridge, U.K.: Cambridge Univ., 1999.

[20] A. F. Filippov, *Differential Equations With Discontinuous Righthand Sides: Control Systems*, vol. 18. Springer Science & Business Media, 2013.

Authorized licensed use limited to: UCLA Library. Downloaded on June 06,2025 at 21:31:21 UTC from IEEE Xplore. Restrictions apply.