This article has been accepted for publication in IEEE Control Systems Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LCSYS.2022.3182284

Neural ODE Control for Trajectory Approximation of Continuity Equation

Karthik Elamvazhuthi, Bahman Gharesifard, *Member, IEEE*, Andrea L. Bertozzi, *Member, IEEE*, and Stanley Osher

***Abstract*— We consider the controllability problem for the continuity equation, corresponding to neural ordinary differential equations (ODEs), which describes how a probability measure is pushedforward by the flow. We show that the controlled continuity equation has very strong controllability properties. Particularly, a given solution of the continuity equation corresponding to a bounded Lipschitz vector field defines a trajectory on the set of probability measures. For this trajectory, we show that there exist piecewise constant training weights for a neural ODE such that the solution of the continuity equation corresponding to the neural ODE is arbitrarily close to it. As a corollary to this result, we establish that the continuity equation of the neural ODE is approximately controllable on the set of compactly supported probability measures that are absolutely continuous with respect to the Lebesgue measure.**

***Index Terms*— Distributed parameter systems, Machine learning, Neural networks.**

I. INTRODUCTION

In recent years, there has been a considerable amount of work on deep neural networks, due to the flexibility they provide for training purposes. Continuum limits of such neural networks has lead to a wide literature on the so-called Neural ordinary differential equations (ODEs) [4], [13], [25], with a major upshot being that tools from dynamical systems and control theory can be used to understand and develop methods to train and synthesize neural networks. This includes extensions to stochastic settings [23], and higher order dynamical variants [21].

Control-theoretic tools have recently been utilized to address questions related to training neural networks. In [22], the authors use differential geometric techniques to establish the controllability properties of the underlying neural ODE and leverage that to obtain uniform approximation results. Controllability properties of neural ODEs on the group of diffeomorphisms has also been investigated in [1]. In addition to this, optimal control theory has been leveraged to train neural networks, and this direction has been an original motivation for neural ODEs, see [4], and also [3], [15], [18].

An interesting property that is the focal part of this paper is that data classification and universal approximation capabilities of neural ODEs can be related to problems arising in optimal transportation theory [2], where one aims to take a given probability density to another. Such transport problems also naturally arise for density estimation problems in machine learning, such as normalizing flows [16]. This insight has been used to leverage tools from optimal transportation theory [2] to find numerically efficient methods to train neural ODEs [12]. The resulting transport problems can be analyzed in terms of a controlled continuity equation, which describes how a probability density evolves under the action of flow of a differential equation. This motivates us to study the approximation capabilities of neural ODEs for density estimation, by studying the control properties of the corresponding continuity equation for which the vector-field is given by that of a neural ODE. Most closely related to our work, the authors in [20] establish approximate controllability of the underlying controlled continuity equation on the space of probability measures. Particularly, given a initial and final measure, it is shown that there exist weight parameters of the neural ODE that can be chosen such that the final condition of the continuity equation is arbitrarily close to the target probability measure in the Wasserstein-1 distance.

Statement of Contributions: We study the approximation capabilities of the continuity equation corresponding to a neural ODE. We show one can construct a sequence of control inputs such that the solutions of the continuity equation corresponding to the neural ODE uniformly converge to the solution of the continuity equation corresponding to any given Lipschitz vector field. This controllability property of the system is challenging to attain due to the low number of control parameters at hand, attributed to the limited width of the neural network. In general, it is not possible to select the control parameters in a way that the approximating vector fields are *strongly* converging to the original one. The key idea behind our result is the observation that one can instead construct admissible vector fields that are weakly converging to the original one, and more importantly, that this is sufficient to achieve uniform convergence of the curves on the set of probability measures.

First submission on March 21, 2022. This work was supported by National Science Foundation (NSF) award DMS-1952339 and AFOSR grants FA9550-18-1-0167 and FA9550-18-1-0502

Karthik Elamvazhuthi is with the Department of Mechanical Engineering, University of California, Riverside, CA 92521, USA (e-mail: kelamvazhuthu@engr.ucr.edu)

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

Andrea L. Bertozzi and Stanley Osher are with the Department of Mathematics, University of California, Los Angeles, CA 90095, USA (e-mail: bertozzi@math.ucla.edu, sjo@math.ucla.edu

This article has been accepted for publication in IEEE Control Systems Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LCSYS.2022.3182284

II. Problem Formulation and Motivation

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a given *activation function*. We define the map $\Sigma : \mathbb{R}^d \to \mathbb{R}^d$ by

$$\Sigma(x) = [\sigma(x_1), ..., \sigma(x_d)]^T.$$

An example of the class of activation functions that we consider is *sigmoidal functions* with globally bounded derivatives. An activation function σ is said to be *sigmoidal* if its range lies in $[0, 1]$,

$$\lim_{x\to-\infty} \sigma(x) = 0 \text{ and } \lim_{x\to\infty} \sigma(x) = 1.$$

One such sigmoidal function is

$$\sigma(x) = \frac{1}{1+e^{-x}}. \tag{1}$$

Another important example of an activation function is the *Rectified Linear Unit (ReLU)* function defined by

$$\sigma(x) = \begin{cases} x & x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

We consider the neural ODE given by

$$\dot{x}(t) = A(t)\Sigma(W(t)x + \theta(t)), \tag{3}$$

where $A : [0, T] \to \mathbb{R}^{d\times d}$, $W : [0, T] \to \mathbb{R}^d$ and $\theta : [0, T] \to \mathbb{R}^d$ are the control inputs or weights for the neural network. See [4], [13], [20], [22] for a discussion on the relation between the above ODE and deep residual neural networks.

Suppose that the initial condition $x(0)$ of (3) is chosen at random from a distribution with a probability density function ρ_0. The uncertainty in the state $x(t)$ is determined by the time-dependent probability density $\rho(t)$ which evolves according to the continuity equation,

$$\frac{\partial\rho}{\partial t} + \nabla \cdot \Big(\big(A(t)\Sigma(W(t)x + \theta(t))\big)\rho\Big) = 0, \qquad \rho(0) = \rho_0. \tag{4}$$

In density estimation problems such as the ones considered in [4], [12], the goal is to construct weight functions (or control inputs) $A(\cdot), W(\cdot), \theta(\cdot)$ so that the endpoint of the solution $\rho(T)$ of (4) is approximately equal to an unknown probability distribution ρ^f, using available samples of ρ^f. From a control-theoretic point of view, it is natural to ask for which class of target distributions ρ^f, solutions of (4) can be controlled to ρ^f within final time T. This problem has been recently considered in [20], for the special case of ReLU activation functions and $d \geq 2$, where it has been shown that $\rho(T)$ can be made arbitrarily close to any given compactly supported ρ^f in the *Wasserstein* 1*-metric*. The purpose of this short paper is to consider the more general *trajectory approximation* problem stated below.

Problem II.1. *Given a curve on the set of probability densities $t \mapsto \tilde{\rho}(t)$, can we construct control inputs $A(\cdot), W(\cdot), \theta(\cdot)$ such that the solution of* (4) *is arbitrary close to $\tilde{\rho}(t)$ in a suitable sense, for all $t \in [0, T]$?*

We answer this problem affirmatively in Theorem IV.1 for the case when the curve $t \mapsto \tilde{\rho}(t)$ is the pushforward of the flow of a uniformly Lipschitz bounded vector field. Then we show that the controllability result in [20] can be derived as a Corollary to our main result, for general activation functions that satisfy Assumption III.2. Few motivations for considering the more general trajectory approximation problem include interpolation of data lying on the set of probability measures [5], identifying dynamical systems from population data [26], and control of large swarms [9].

III. Notation and Preliminaries

In this section, we define some notation that will be used throughout the paper. We refer the readers to [2] for more details. Let $\mathcal{P}_2(\mathbb{R}^d)$ denote the set of Borel probability measures on $\mathbb{R}^d$ with finite second moment: $\int_\Omega |x|^2 d\mu(x) < \infty$. For a given Borel map $T : \mathbb{R}^d \to \mathbb{R}^d$ we will denote by $T_\#$ the corresponding pushforward map, which maps any measure μ to a measure $T_\#\mu$, where $T_\#\mu$ is the measure defined by

$$(T_\#\mu)(B) = \mu(T^{-1}(B)), \tag{5}$$

for all Borel measurable sets $B \subseteq \mathbb{R}^d$. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, we denote the set of transport plans from μ to ν by

$$\Gamma(\mu, \nu) := \{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) | \pi^1_\#\gamma = \mu, \pi^2_\#\gamma = \nu\}, \tag{6}$$

where $\pi^i : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ are the projections on to the ith coordinates, respectively. We will define the $2-$Wasserstein distance between two probability measures μ, ν as the following

$$W_2(\mu, \nu) = \min_{\gamma\in\Gamma(\mu,\nu)} \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} |x - y|^2 d\gamma(x, y) \right)^{1/2}. \tag{7}$$

The set $C([0, T], \mathcal{P}_2(\mathbb{R}^d))$ will refer to the set of continuous curves $t \mapsto \mu_t$ in $\mathcal{P}_2(\mathbb{R}^d)$ with respect to the topology induced by the $2-$Wasserstein distance. We will say that a sequence $\{\mu^N\}_{N\in\mathbb{Z}_+}$ in $C([0, T]; \mathcal{P}_2(\mathbb{R}^d))$ converges to $\mu \in C([0, T]; \mathcal{P}_2(\mathbb{R}^d))$ if $\lim_{N\to\infty} \sup_{t\in[0,T]} W_2(\mu^N_t, \mu_t) = 0$. Given a vector-field $V : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$, we will consider solutions μ of the continuity equation,

$$\frac{\partial\mu}{\partial t} + \nabla \cdot \big(V_t(x)\mu\big) = 0, \qquad \mu(0) = \mu_0. \tag{8}$$

Naturally, we say that $\mu \in C([0, T], \mathcal{P}_2(\mathbb{R}^d))$ is a weak solution, or a *solution in the sense of distributions* [2, Section 8.1] of the continuity equation (8) if

$$\int_0^T \int_{\mathbb{R}^d} \left(\frac{\partial\phi(t, x)}{\partial t} + \nabla\phi(t, x) \cdot V_t(x) \right) d\mu_t(x)dt = -\int_{\mathbb{R}^d} \phi(0, x)d\mu_0(x), \tag{9}$$

for all compactly supported real-valued functions $\phi \in C^\infty([0, T) \times \mathbb{R}^d)$. We make the following assumption.

Assumption III.1. *The vector field $V : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$ is such that $t \mapsto V_t(x)$ is measurable for every $x \in \mathbb{R}^d$ and it is uniformly Lipschitz in x. That is, there exists $K > 0$ such that*

$$|V_t(x) - V_t(y)| \leq K|x - y|,$$

for all $x, y \in \mathbb{R}^d$ and all $t \in [0, T]$.

In addition to this, we will need some mild assumptions on the activation function $\sigma : \mathbb{R} \to \mathbb{R}$. For this purpose, let us define the set of functions

$$\mathcal{F} = \bigcup_{m \in \mathbb{Z}_+} \{\sum_{i=1}^{m} \alpha_i \sigma(w_i^T x + \theta_i) \mid \alpha_i \in \mathbb{R}, w_i \in \mathbb{R}^d, \theta_i \in \mathbb{R}\}.$$

Note that the set $\mathcal{F}$ is the set of arbitrarily wide single-hidden layer neural networks.

Assumption III.2. *We make the following assumptions:*

1) ***(Regularity)*** *The activation function σ is globally Lipschitz, that is, there exists $K > 0$ such that*

$$|\sigma(x) - \sigma(y)| \leq K|x - y|, \tag{10}$$

for all $x, y \in \mathbb{R}$.

2) ***(Density of superpositions)*** *The set of functions $\mathcal{F}$ is dense in $C(\mathbb{R}^d; \mathbb{R})$ in the uniform norm topology on compact sets. Particularly, given a function $f \in C(\mathbb{R}^d; \mathbb{R})$, for each compact set $\Omega \subset \mathbb{R}$ and $\delta > 0$, there exists a function $g \in \mathcal{F}$ such that*

$$\sup_{x \in \Omega} |f(x) - g(x)| < \delta.$$

It is well-known that the Logistic function (1) and the ReLU function (2) satisfy the density property, see [7], [17]. Given Assumption III.2, it is easy to see that the subset of vector-valued functions $\mathcal{F}_d$ defined by

$$\mathcal{F}_d = \bigcup_{m \in \mathbb{Z}_+} \{\sum_{i=1}^{m} A_i \Sigma(W_i x + \theta_i) \mid A_i, W_i \in \mathbb{R}^{d \times d}, \theta_i \in \mathbb{R}^d\},$$

is dense in $C(\mathbb{R}^d; \mathbb{R}^d)$ in the uniform norm topology on compact sets.

IV. Analysis

In this section, we perform our controllability analysis. We show that given a solution of the continuity equation (8), we can approximate the solution arbitrarily well using solutions of the equation (4).

Theorem IV.1. ***(Main Result)*** *Suppose that Assumptions III.2 and III.1 hold and $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ has compact support. Let μ be the weak solution of the continuity equation* (8) *corresponding to the vector field V. Additionally, suppose that V is uniformly bounded in space and time. Then for every $\epsilon > 0$, there exist piecewise constant control inputs $A^\epsilon(\cdot), W^\epsilon(\cdot)$ and $\theta^\epsilon(\cdot)$, such that the corresponding weak solutions μ^ϵ of* (4) *satisfy*

$$\sup_{t \in [0,T]} W_2(\mu_t^\epsilon, \mu_t) \leq \epsilon. \tag{11}$$

As a consequence, we obtain the following result which was established as [20, Theorem 5].

Corollary IV.2. ***(Approximate Controllability)*** *Suppose that Assumption III.2 holds and $\mu_0, \mu^f \in \mathcal{P}_2(\mathbb{R}^d)$ have compact supports, and are absolutely continuous with respect to the Lebesgue measure. Then for every $\epsilon > 0$, there exist piecewise constant control inputs $A^\epsilon(\cdot), W^\epsilon(\cdot)$ and $\theta^\epsilon(\cdot)$, such that the corresponding weak solutions μ^ϵ of the equation* (4)*, satisfy*

$$W_2(\mu_T^\epsilon, \mu^f) \leq \epsilon. \tag{12}$$

In order to prove the main result and its corollary, we will need some preliminary results. The idea behind the proof is that due to Assumption III.2, the convex closure of the set of admissible vector fields includes V. This is a well-known idea in theory of differential inclusions and relaxed controls [10]. These existing results are not directly applicable to (4). That being said, we adapt the arguments to prove the above results.

We first observe some regularity properties of the solution of the continuity equation (8), with respect to the time variable, which will be used later to invoke compactness of certain approximating sequences.

Lemma IV.3. *Suppose that $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ has compact support and V is a Borel measurable vector field for which $\mu \in C([0, T]; \mathcal{P}_2(\mathbb{R}^d))$ is a solution of the continuity equation. Suppose there exists $C > 0$ such that $|V_t(x)| \leq C$ for μ_t almost every $x \in \mathbb{R}^d$, for (Lebesgue) almost every $t \in (0, T)$. Then the curve μ is Lipschitz:*

$$W_2(\mu_t, \mu_s) \leq K|t - s|, \tag{13}$$

for all $0 < t \leq s < T$, where K is a positive constant that depends only on C.

Proof. From [2, Theorem 8.3.1], we know the curve μ is absolutely continuous in the sense of [2, Definition 1.1.1], since $\int_{\mathbb{R}^d} \|V_t(x)\|^2 d\mu_t(x)$ is essentially bounded over $(0, T)$. Moreover, from [2, Theorem 8.3.1], the *metric derivative* of μ defined by

$$|\mu'(t)| := \lim_{s \to t} \frac{W_2(\mu_t, \mu_s)}{|t - s|},$$

is essentially bounded by $\left(\int_{\mathbb{R}^d} \|V_t\|^2 d\mu_t(x)\right)^{1/2}$. Since, $|V_t| \leq C$ for μ_t almost every $\mathbb{R}^d$, for (Lebesgue) almost every $t \in (0, T)$, from [2, Theorems 1.1.2 and 8.3.1], we have that $W_2(\mu(t), \mu(s)) \leq \int_s^t |\mu'(\tau)| d\tau \leq K|t - s|$ for all $0 < t \leq s < T$. This concludes the result. □

Next, we observe some classical properties on the relation between solutions of the continuity equation (8) and an associated ODE. This result enables some control on the growth of the support of the solution of the continuity equation, due to the Caratheodory existence theorem for solutions of ODEs. This, again, will be used later to establish the compactness of certain sequence of curves on $\mathcal{P}_2(\mathbb{R}^d)$. In what follows, we denote by $B_c(0) := \{x \in \mathbb{R}^d; |x| \leq c\}$ the closed ball of radius $c > 0$ centered at the origin.

Proposition IV.4. *Suppose that Assumption III.1 holds and $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ has a compact support. If $r, R, C > 0$ are such that* supp $\mu_0 \subset B_r(0)$*, $|V_t(x)| < C$ for all $(t, x) \in [0, T] \times B_{R+r}(0)$ and $T < \frac{R+r}{C}$. Then there exists a unique solution μ to the continuity equation* (8)*. Additionally, the solution μ is given by $\mu_t = (X_t)_\# \mu_0$ for all $t \in [0, T]$, where $X : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$ is such that*

$$\frac{dX_t(x)}{dt} = V_t(X_t(x)); \quad X_0(x) = x.$$

Moreover, $\operatorname{supp}\ \mu_t \subset B_{R+r}(0)$ *for all* $t \in [0,T]$.

Proof. Due to Assumption III.1, for each $x \in B_r(0)$, there exists a unique local solution $y(t)$, of the differential equation

$$\frac{dy(t)}{dt} = V_t(y(t)); \quad y(0) = x.$$

From the assumption that there exist $r, R, C > 0$ such that $x \in B_r(0)$, $|V_t(x)| < C$ for all $(t,x) \in [0,T] \times B_{R+r}(0)$ and $T < \frac{R+r}{C}$, and Caratheodory's existence theorem on the existence of solutions to ODEs [11, Chapter 1, Theorem 1], we can conclude that the solution y of the above ODE is defined over the interval $[0,T]$ and $y(t) \in B_{R+r}(0)$ for all $t \in [0,T]$. Hence, for μ_0 every $x \in \mathbb{R}^d$, the solution of this ODE is well defined over the interval $[0,T]$ and the result then follows from [2, Lemma 8.1.6]. □

In the next proposition we prove the straightforward idea that given a vector-field we can approximate the solution of (8) using piecewise constant in time vector fields.

Proposition IV.5. *Suppose that* V *satisfies Assumption III.1, is uniformly bounded in space and time and* $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ *has a compact support. Then there exists a sequence* $\{V^N\}_{N\in\mathbb{Z}_+}$ *of piecewise constant in time vector fields* $V^N : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ *such that, the sequence of weak solutions* $\{\mu^N\}_{N\in\mathbb{Z}_+}$*, corresponding to these vector-fields, converges to the weak solution* μ *corresponding to the vector field* V.

Proof. Define V^N by

$$V_t^N(x) = \begin{cases} \frac{N}{T}\int_{\frac{(n-1)T}{N}}^{\frac{nT}{N}} V_\tau(x)d\tau; t \in [\frac{(n-1)T}{N}, \frac{nT}{N}) \\ \qquad\qquad \text{for } n = 1,..,N-1, \\ \frac{N}{T}\int_{\frac{(N-1)}{T}}^{T} V_\tau(x)d\tau; t \in [\frac{(N-1)T}{N}, T], \end{cases}$$

for all $x \in \mathbb{R}^d$. By Lemma A.1 $t \mapsto V_t^N(x)$ weakly converges to $t \mapsto V_t(x)$ in $L^1(0,1;\mathbb{R}^d)$ for every $x \in \mathbb{R}^d$. Moreover, it is easy to verify that the vector-fields V^N satisfy Assumption III.1 Let X^N be the flow corresponding to the vector fields V^N, for each $N \in \mathbb{Z}_+$. It follows [19, Lemma 2.8] that $\mu_t^N = (X_t^N)_{\#}\mu_0$ are converging to $\mu_t = (X_t)_{\#}\mu_0$ in the weak topology of measures, for each $t \in [0,T]$, as N tends to ∞. Invoking Proposition IV.4, that there exists a compact set Ω such that the supports of μ_t^N, μ_t are contained in Ω for all $t \in [0,T]$ and for all $N \in \mathbb{Z}_+$. Therefore, since convergence in the weak topology is equivalent to the convergence in the 2-Wasserstein distance for probability measures with compact support [24, Theorem 6.9], this implies that $\{\mu_t^N\}_{N\in\mathbb{Z}_+}$ converges to μ_t in $\mathcal{P}_2(\mathbb{R}^d)$, for each $t \in [0,T]$. Moreover, due to the uniform bound $|V_t^N(x)| \le C$, Lemma IV.3 implies that $\{\mu^N\}_{N\in\mathbb{Z}_+}$ are uniformly Lipschitz in the time variable and hence, invoking the Arzelà-Ascoli theorem, there exists a subsequence of $\{\mu^N\}_{N\in\mathbb{Z}_+}$ that is converging to $\tilde{\mu}$ in $C([0,T],\mathcal{P}_2(\mathbb{R}^d))$. But we know that $\{\mu_t^N\}_{N\in\mathbb{Z}_+}$ converges to μ_t in $\mathcal{P}_2(\mathbb{R}^d)$, for each $t \in [0,T]$. Therefore, it must be that $\tilde{\mu} = \mu$. This concludes the proof. □

Proposition IV.6. *Suppose that* $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ *has a compact support, and that* $V : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ *is uniformly bounded in space and time and satisfies Assumption III.1. Additionally, assume that the vector field is piecewise constant in time. Given Assumption III.2, for* $N \in \mathbb{Z}_+$*, there exist vector fields* Q^N *that are piecewise constant and such that* $Q_t^N \in \mathcal{F}_d$ *for all* $t \in [0,T]$*, and the sequence of weak solutions* $\{\mu^N\}_{N\in\mathbb{Z}_+}$*, corresponding to the vector-fields* $\{Q^N\}_{N\in\mathbb{Z}_+}$*, converges to the weak solution* μ *corresponding to the vector field* V.

Proof. Suppose that the support of μ_0 lies in $B_r(0)$ for some $r > 0$. Since the vector-field V is uniformly Lipschitz and bounded, the support of μ_t lies in $B_{R+r}(0)$ for all sufficiently large $R > 0$. Choose R such that $T < \frac{R+r}{C+\delta}$ for some $\delta > 0$ and the support of μ_t lies in $B_{R+r}(0)$ for all $t \in [0,T]$. Define $\Omega := B_{R+r}(0)$. By Assumption III.2, we can construct approximating vector-fields Q^N such that Q^N are piecewise constant in time, for $N \in \mathbb{Z}_+$, $Q_t^N \in \mathcal{F}_d$ for all $t \in [0,T]$, and $\{Q^N\}_{N\in\mathbb{Z}_+}$ strongly converges to V uniformly in time and space on compact sets:

$$\lim_{N\to\infty} \sup_{(t,x)\in[0,T]\times\Omega} \|V_t(x) - Q_t^N(x)\|_\infty = 0.$$

and $|Q^N(t,x)| < C + \delta$ for all $(t,x) \in [0,T] \times \Omega$ and all $N \in \mathbb{Z}_+$. We can conclude μ_t^N is contained in Ω for all $t \in [0,T]$ and all $N \in \mathbb{Z}_+$, due to Proposition IV.4. Due to the uniform bound on the velocity fields on Ω, Lemma IV.3 implies that $\{\mu^N\}_{N\in\mathbb{Z}_+}$ are uniformly Lipschitz in time. Therefore, there exists a subsequence of $\{\mu^N\}_{N\in\mathbb{Z}_+}$ that converges to a limit $\tilde{\mu}$ in $C([0,T];\mathcal{P}_2(\mathbb{R}^d))$. Next, we will verify that $\tilde{\mu}$ is the weak solution of the continuity equation (8) corresponding to the curve V. Let $\phi \in C^\infty([0,T) \times \mathbb{R}^d)$ be a compactly supported function. Since the supports of μ^N and $\tilde{\mu}$ are contained in the compact set Ω,

$$\int_0^T \int_{\mathbb{R}^d} \left(\frac{\partial\phi(t,x)}{\partial t} + \nabla\phi(t,x)\cdot Q_t^N(x)\right) d\mu_t^N(x)dt$$
$$- \int_0^T \int_{\mathbb{R}^d} \left(\frac{\partial\phi(t,x)}{\partial t} + \nabla\phi(t,x)\cdot V_t(x)\right) d\tilde{\mu}_t(x)dt =$$

$$\int_0^T \int_{\Omega} \left(\frac{\partial\phi(t,x)}{\partial t} + \nabla\phi(t,x)\cdot Q_t^N(x)\right) d\mu_t^N(x)dt$$
$$- \int_0^T \int_{\Omega} \left(\frac{\partial\phi(t,x)}{\partial t} + \nabla\phi(t,x)\cdot V_t(x)\right) d\tilde{\mu}_t(x)dt. \quad (14)$$

Since $\{Q^N\}_{N\in\mathbb{Z}_+}$ is uniformly converging to V on $[0,T] \times \Omega$, we can conclude that the terms $\left(\frac{\partial\phi}{\partial t} + \nabla\phi \cdot Q^N\right)$ are uniformly converging to $\left(\frac{\partial\phi}{\partial t} + \nabla\phi \cdot V\right)$ on $[0,T] \times \Omega$, as N tends to ∞. Moreover, the sequence $\{\mu^N\}_{N\in\mathbb{Z}_+}$ is converging to $\tilde{\mu}$ in $C([0,T];\mathcal{P}_2(\mathbb{R}^d))$. By an application of the Dominated Convergence Theorem, (14) converges to 0 as N tends to ∞. This implies that $\tilde{\mu}$ is the weak solution of the continuity equation (8) corresponding to the velocity field V since we conclude that

$$\int_0^T \int_{\mathbb{R}^d} \left(\frac{\partial\phi(t,x)}{\partial t} + \nabla\phi(t,x)\cdot V_t(x)\right) d\tilde{\mu}_t(x)dt$$
$$= -\int_{\mathbb{R}^d} \phi(0,x)d\mu_0(x),$$

for all compactly supported functions $\phi \in C^\infty([0,T)\times\mathbb{R}^d)$. The solution μ to the continuity equation (8) is unique. Therefore, $\tilde{\mu}=\mu$. □

Next, we show that given a vector field that is a superposition of functions of the form $\Sigma(W\cdot+\theta)$, we can construct an oscillating sequence of admissible vector fields that converge to the superposition weakly in time.

Lemma IV.7. *Let $A_i, W_i \in \mathbb{R}^{d\times d}, \theta_i \in \mathbb{R}^d$ be weight parameters for $i=1,...,m$. For each $N\in\mathbb{Z}_+$. Let Q^N be a $\frac{T}{N}$-periodic vector field defined by*

$$Q_{t+\frac{nT}{N}}(x) = mA_i\Sigma(W_ix+\theta_i),\ \ t\in[\frac{iT}{mN},\frac{(i+1)T}{mN}), \quad (15)$$

for all $n\in\{0,...,N-1\}$, $i\in\{0,1,...,m-1\}$ and $x\in\mathbb{R}^d$. Then, for each $x\in\mathbb{R}^d$, $t\mapsto Q^N_t(x)$ weakly converges to $\sum_{i=1}^m A_i\Sigma(W_ix+\theta_i)$ in $L^1(0,T;\mathbb{R}^d)$ for all $x\in\mathbb{R}^d$, as N tends to ∞.

Proof. We note that $\frac{1}{T}\int_0^T Q^N_t(x) = \sum_{i=1}^m A_i\sigma(W_ix+\theta_i)$ for all $x\in\mathbb{R}^d$ and all $N\in\mathbb{Z}_+$. The weak convergence of $t\mapsto Q^N_t(x)$ to $\sum_{i=1}^m A_i\Sigma(W_ix+\theta_i)$ in $L^1(0,T;\mathbb{R}^d)$, for each $x\in\mathbb{R}^d$, as N tends to ∞, follows from [6, Theorem 8.2]. Note that the latter result is stated of functions that are p-integrable for $p>1$. However, since $\sum_{i=1}^m A_i\Sigma(W_ix+\theta_i)$ is essentially bounded, the result applies. □

Proposition IV.8. *Let $\mu_0\in\mathcal{P}_2(\mathbb{R}^d)$ have a compact support. Suppose Assumption III.2 holds. Let $Q:[0,T]\times\mathbb{R}^d\to\mathbb{R}^d$ be a piecewise constant in time vector field such that $Q_t\in\mathcal{F}_d$ for all $t\in[0,T]$. Then there exist vector fields $Q^N:[0,T]\times\mathbb{R}^d\to\mathbb{R}^d$ that are of the form of the right hand side of* (3) *for piecewise constant controls $A^N(\cdot),W^N(\cdot),\theta^N(\cdot)$ such that the sequence of solutions $\{\mu^N\}_{N\in\mathbb{Z}_+}$ of* (4) *for these choices of weights, converges to the solution μ, corresponding to the vector field Q, in $C([0,T];\mathcal{P}_2(\mathbb{R}^d))$.*

Proof. From Lemma IV.7 it follows that, for Q given, there exist weakly approximating admissible vector-fields Q^N, of the form in the right hand side of (3), by repeating the construction in (15) over the time intervals on which Q is constant and concatenating the approximating vector fields. Moreover, from the construction in Lemma IV.7, the map $t\mapsto Q^N_t(x)$ weakly converges to $t\mapsto Q_t(x)$, for each x, in $L^1(0,T;\mathbb{R}^d)$, as N tends to ∞. From [19, Lemma 2.8], it follows that $\{\mu^N_t\}_{N\in\mathbb{Z}_+}$ converges to μ_t in $\mathcal{P}_2(\mathbb{R}^d)$, for each $[0,T]$. From the construction of the weakly converging vector fields Q^N in Lemma IV.3, the vector fields Q^N are uniform bounded on compact sets and therefore, it follows that the curves μ^N are uniformly Lipschitz in time. As a result, there exists a subsequence of $\{\mu^N\}_{N\in\mathbb{Z}_+}$ converging in $C([0,T];\mathcal{P}_2(\mathbb{R}^d))$. But we have already established that $\{\mu^N_t\}_{N\in\mathbb{Z}_+}$ converges to μ_t in $\mathcal{P}_2(\mathbb{R}^d)$, for each $[0,T]$. Therefore, the convergence of $\{\mu^N_t\}_{N\in\mathbb{Z}_+}$ to μ_t must be uniform in the time variable, and hence $\{\mu^N\}_{N\in\mathbb{Z}_+}$ converges to μ in $C([0,T];\mathcal{P}_2(\mathbb{R}^d))$. □

Now, we are ready to prove our main result on approximate controllability of (4) about trajectories of (8).

Proof of Theorem IV.1. The result follows by applying Proposition (IV.5) to approximate V using a vector fields that are piecewise constant in time and, then using Proposition IV.8 to approximate the piecewise constant approximations using vector fields of the form in the right-hand side of (3). □

Finally, owing to an existing result on the approximate controllability of the continuity equation (8) proved in [8], we can establish approximate controllability of (4).

Proof of Corollary IV.2. According to [8, Proposition 3.1], it is known that, for every $\epsilon>0$, there exists a uniformly bounded vector field V satisfying Assumption III.1 such that the solution μ of (8) satisfies $W_2(\mu_T,\mu_f)\le\epsilon/2$. Then, due to Theorem IV.1, there exist piecewise constant control inputs $A^\epsilon(\cdot),W^\epsilon(\cdot)$ and $\theta^\epsilon(\cdot)$, such that the corresponding weak solutions μ^ϵ of the equation (4), satisfies

$$W_2(\mu^\epsilon_T,\mu_T)\le\epsilon/2. \quad (16)$$

Using the triangle inequality property of the W_2-distance, we can conclude that $W_2(\mu^\epsilon_T,\mu^f)\le\epsilon$. This concludes the proof. □

V. CONCLUSION

We demonstrated how neural ODEs can be used to approximate solutions of the continuity equation with a uniformly Lipschitz bounded vector field. Interesting future directions include extending the result to vector fields that are not Lipschitz, such as those arising from solution of the Benamou-Brenier formulation of optimal transport. Lastly, one could also consider similar approximation results for stochastic and higher order dynamical variants of neural ODEs.

ACKNOWLEDGEMENTS

The authors thank Katy Craig and Levon Nurbekyan for many helpful discussions.

APPENDIX

Lemma A.1. *Let $f\in L^1(0,T;\mathbb{R}^n)$. Suppose that there exists a constant $C>0$ such that $|f(t)|<C$ for almost every $t\in(0,T)$. For each $N\in\mathbb{Z}_+$, consider $f^N\in L^1(0,T;\mathbb{R}^d)$ defined by*

$$f^N(t) = \frac{N}{T}\int_{\frac{(n-1)T}{N}}^{\frac{(n)T}{N}} f(\tau)d\tau;\ \ t\in[\frac{(n-1)T}{N},\frac{nT}{N}), \quad (17)$$

for $n=1,..,N$. Then the sequence $\{f^N\}_{N\in\mathbb{Z}_+}$ weakly converges to f in $L^1(0,T;\mathbb{R}^d)$ as $N\to\infty$.

Proof. By the Lebesgue differentiation theorem [14, Theorem 2.3.4], $\{f^N(t)\}_{N\in\mathbb{Z}_+}$ converges to $f(t)$ for almost every $t\in(0,T)$. Since $|f(t)|$ and $|f^N(t)|$ are bounded by C for almost every $t\in(0,T)$, it follows from the dominated convergence theorem that

$$\lim_{N\to\infty}\int_0^T |f(t)-f^N(t)|dt = 0.$$

Therefore, $\{f^N\}_{N\in\mathbb{Z}_+}$ converges to f in the strong topology in $L^1(0,T;\mathbb{R}^d)$ and hence, also in the weak topology. □

REFERENCES

[1] Andrei Agrachev and Andrey Sarychev. Control on the manifolds of mappings with a view to the deep learning. *Journal of Dynamical and Control Systems*, pages 1–20, 2021.

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[3] Martin Benning, Elena Celledoni, Matthias J Ehrhardt, Brynjulf Owren, and Carola-Bibiane Schönlieb. Deep learning as optimal control problems: Models and numerical methods. *Journal of Computational Dynamics*, 6(2):171, 2019.

[4] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *NeurIPS*, 2018.

[5] Yongxin Chen, Giovanni Conforti, and Tryphon T Georgiou. Measure-valued spline curves: An optimal transport viewpoint. *SIAM Journal on Mathematical Analysis*, 50(6):5947–5968, 2018.

[6] Michel Chipot. *Elliptic equations: an introductory course*. Springer Science & Business Media, 2009.

[7] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[8] 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.

[9] Karthik Elamvazhuthi and Spring Berman. Mean-field models in swarm robotics: A survey. *Bioinspiration & Biomimetics*, 15(1):015001, 2019.

[10] Hector O Fattorini. *Infinite dimensional optimization and control theory*, volume 54. Cambridge University Press, 1999.

[11] Aleksei Fedorovich Filippov. *Differential equations with discontinuous righthand sides: control systems*, volume 18. Springer Science & Business Media, 2013.

[12] Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam Oberman. How to train your neural ode: the world of jacobian and kinetic regularization. In *International Conference on Machine Learning*, pages 3154–3164. PMLR, 2020.

[13] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34(1):014004, 2017.

[14] Tuomas Hytönen, Jan Van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach spaces*, volume 12. Springer, 2016.

[15] Jean-François Jabir, David Šiška, and Łukasz Szpruch. Mean-field neural odes via relaxed optimal control. *arXiv preprint arXiv:1912.05475*, 2019.

[16] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

[17] Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

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

[19] Nikolay Pogodaev. Optimal control of continuity equations. *Nonlinear Differential Equations and Applications NoDEA*, 23(2):21, 2016.

[20] Domènec Ruiz-Balet and Enrique Zuazua. Neural ode control for classification, approximation and transport. *arXiv preprint arXiv:2104.05278*, 2021.

[21] Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Momentum residual neural networks. *arXiv preprint arXiv:2102.07870*, 2021.

[22] Paulo Tabuada and Bahman Gharesifard. Universal approximation power of deep residual neural networks via nonlinear control theory. In *International Conference on Learning Representations*, 2020.

[23] Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*, 2019.

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

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

[26] Caleb Weinreb, Samuel Wolock, Betsabeh K Tusi, Merav Socolovsky, and Allon M Klein. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences*, 115(10):E2467–E2476, 2018.