Journal of Machine Learning **ISSN**: 2790-2048(e), 2790-203X(p)

Variational Formulations of ODE-Net as a Mean-Field Optimal Control Problem and Existence Results

Noboru Isobe [*1] and Mizuho Okumura [† 2]

[1]Graduate School of Mathematical Sciences, The University of Tokyo, Tokyo, Japan.
[2]Graduate School of Science, Tohoku University, Sendai, Japan.

Abstract. This paper presents a mathematical analysis of ODE-Net, a continuum model of deep neural networks (DNNs). In recent years, machine learning researchers have introduced ideas of replacing the deep structure of DNNs with ODEs as a continuum limit. These studies regard the "learning" of ODE-Net as the minimization of a "loss" constrained by a parametric ODE. Although the existence of a minimizer for this minimization problem needs to be assumed, only a few studies have investigated the existence analytically in detail. In the present paper, the existence of a minimizer is discussed based on a formulation of ODE-Net as a measure-theoretic mean-field optimal control problem. The existence result is proved when a neural network describing a vector field of ODE-Net is linear with respect to learnable parameters. The proof employs the measure-theoretic formulation combined with the direct method of calculus of variations. Secondly, an idealized minimization problem is proposed to remove the above linearity assumption. Such a problem is inspired by a kinetic regularization associated with the Benamou-Brenier formula and universal approximation theorems for neural networks.

Keywords:
Deep learning,
ResNet,
ODE-Net,
Benamou-Brenier formula,
Mean-field game.

Article Info.:
Volume: 3
Number: 4
Pages: 413 - 444
Date: December/2024
doi.org/10.4208/jml.231210

Article History:
Received: 10/12/2023
Accepted: 11/09/2024

Communicated by:
Jiequn Han

1 Introduction

Deep neural networks, or deep learning, now constitute a core of artificial intelligence technology, but their theoretical inner mechanisms have yet to be explored. In particular, there have been few theoretical contributions regarding "learning" DNNs, despite practical demands for them, where "learning" is, broadly speaking, to minimize the so-called "loss" by optimizing a parameter θ of DNNs.

Our research aims to establish a well-posed mathematical formulation of the learning. To achieve this aim, some researchers have brought languages of dynamical systems and differential equations into DNNs, for example, in [22, 27, 54]. In short, one can regard a continuum limit of DNNs in their depth as an ODE. Many researchers have attempted to dissect DNNs through some ODEs, designated as ODE-Net throughout the paper. For more information on these attempts, see the survey in Section 2. Based on this survey, well-posednesses, such as the existence of a minimizer of loss, have not yet been fully explored in the context of these studies.

*Corresponding author. nobo0409@g.ecc.u-tokyo.ac.jp
†okumura.mizuho.p3@gmail.com

Accordingly, our goal in this paper is to prove the existence of a minimizer for learning ODE-Net, formulated as a regularized minimization problem constrained by a continuity equation.

1.1 Target problems and main results

First of all, we are going to study the existence of a minimizer of the following kinetic-regularized minimization problem.

Problem 1.1 (Kinetic Regularized Learning Problem Constrained by ODE-Net)**.** Let $\lambda \geq 0$ and $\epsilon > 0$ be constants, let $\mathcal{Y}$ be a subset of $\mathbb{R}^d$ and let $v\colon \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ and $\ell\colon \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ be continuous. Let $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d \times \mathcal{Y})$ be a given training data. Set

$$J(\mu,\theta) := \int_{\mathbb{R}^d\times\mathcal{Y}} \ell \mathrm{d}\mu_T + \int_0^T \int_{\mathbb{R}^d\times\mathcal{Y}} \left(\frac{\lambda}{2}|v(x,\theta_t)|^2 + \frac{\epsilon}{2}|\theta_t|^2\right) \mathrm{d}\mu_t(x,y)\mathrm{d}t \tag{1.1}$$

for $\mu \in C([0,T];(\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y}), W_2))$ and $\theta \in L^2(0,T;\mathbb{R}^m)$. Note that $v(\bullet,\theta) \in L^2(\mathrm{d}\mu)$ is a vector field on $\mathbb{R}^d$ for $\mu \in \mathcal{P}_c(\mathbb{R}^d)$ and $\theta \in \mathbb{R}^m$. The learning problem constrained by ODE-Net is posed as the following constrained minimization problem:

$$\begin{aligned} &\inf\left\{J(\mu,\theta) \;\middle|\; \mu \in C\big([0,T];\big(\mathcal{P}_2(\mathbb{R}^d\times\mathcal{Y}),W_2\big)\big),\ \theta \in L^2(0,T;\mathbb{R}^m)\right\} \\ &\text{subject to} \\ &\begin{cases} \partial_t \mu_t + \mathrm{div}_x\left(\mu_t(x,y)v(x,\theta_t)\right) = 0, \quad (x,y) \in \mathbb{R}^d\times\mathcal{Y}, \quad t \in (0,T), \\ \mu_t|_{t=0} = \mu_0, \end{cases} \end{aligned} \tag{1.2}$$

where $\mathcal{P}_c(\mathbb{R}^d \times \mathcal{Y})$ denotes the set of regular and Borel probability measures compactly supported on $\mathbb{R}^d \times \mathcal{Y}$, $(\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y}), W_2)$ denotes the (L^2-)Wasserstein space defined in Section 3.2, $C([0,T];(\mathcal{P}(\mathbb{R}^d \times \mathcal{Y}), W_2))$ denotes the set of curves which is continuous with respect to the Wasserstein topology (see also Definition 3.1), and

$$\mu \in C\big([0,T];\big(\mathcal{P}_2(\mathbb{R}^d\times\mathcal{Y}),W_2\big)\big)$$

is supposed to solve the Eq. (1.2) in the distributional sense of Definition 3.2.

Remark 1.1. In Problem 1.1, the ODE-Net corresponds to the continuity equation (1.2) with a parameter θ_t, and the learning to the minimization of a functional J with respect to a parameter θ_t and a solution μ_t to ODE (1.2).

The first term in (1.1) measures the so-called loss. The second term in (1.1) is called a "kinetic regularization" in [25] because it represents the kinetic energy when $v(\bullet,\theta)$ ($\theta \in \mathbb{R}^m$) is regarded as a velocity field on $\mathbb{R}^d$. By letting this kinetic energy be as small as possible, we could control the velocity field so that the support of the solution μ_t to (1.2) does not change wildly. The third term is often called an L^2-regularization, which is familiar with the well-known Ridge regression.

In order to prove existence of a minimizer for Problem 1.1, we shall impose the following assumptions on $\mathcal{Y}$, ℓ and v.

Assumption 1.1. The label set $\mathcal{Y} \subset \mathbb{R}^d$ is compact, and the loss function $\ell\colon \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ is a continuous function of 2-growth, see also Definition 3.1.

In addition, following the previous works on ODE-Net [7, 48, 49, 52, 61], we impose the assumption below that the neural network $v(x,\theta)$ is linear with respect to θ, but not necessarily linear with respect to x.

Assumption 1.2. The neural network v in (1.2) is linear with respect to θ, i.e. the parameter θ is a $d \times p$ matrix and v satisfies

$$v(x,\theta) = \theta f(x), \tag{1.3}$$

where $f\colon \mathbb{R}^d \to \mathbb{R}^p$ is a Lipschitz continuous function.

Assumption 1.2 is not a serious restriction. In fact, [58, Theorem 1] shows that for a neural network v that is nonlinear with respect to θ, there exists another neural network that is linear with respect to θ and can approximate the solution μ of ODE-Net (1.2). Thus, Assumption 1.2 is not so restrictive in discussing the existence of the minimizer for Problem 1.1. Rather, Assumption 1.2 can address neural networks unbounded with respect to parameters θ, which commonly appear in modern DNNs. In contrast, previous theoretical works often assume a bounded neural network, details of which will be given in Section 2 below.

Under these assumptions, we obtain one of our main results in the present paper.

Theorem 1.1 (Existence of a Minimizer)**.** *Under Assumptions* 1.1 *and* 1.2*, there exists a minimizer* $(\mu,\theta) \in C([0,T];(\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y}),W_2)) \times L^2(0,T;\mathbb{R}^m)$ *for Problem* 1.1*.*

It should be noted that by virtue of this theorem, one can assume that the deep learning model as in Problem 1.1 with Assumptions 1.1 and 1.2 is well-defined so that we can pursue the mathematical analysis of the learning of ODE-Nets. We also remark here that the uniqueness of such minimizers cannot be generally expected since the problem is overdetermined with a large degree of freedom in θ. We will also mention the uniqueness in Remark 4.3 below.

We note that Assumption 1.2 does not hold for all neural networks. For example, two-layer ReLU networks $v(x,\theta) = A(Bx)_+$, $\theta = (A,B)$, $A,B \in \mathbb{R}^{d\times d}$, are not linear with respect to θ. This network is quite commonly used, not only in ODE-Net but also in the so-called ResNet, as illustrated in [28, Fig. 2].

In order to provide existence results for these cases as well, we shall consider an ideal or relaxed version of Problem 1.1. To this end, we shall employ the universal approximation theorem by Cybenko [18] or the Kolmogorov-Arnol'd representation theorem shown by [5,33,55], they insist that neural networks v can approximate or represent arbitrary vector fields. Those theorems inspire that the ODE-Net is no longer parametrized by θ, i.e. the ODE-Net is just driven by a family of vector fields $(v_t)_{t\in[0,T]}$. From this perspective, our ideal setting for the learning reads:

Problem 1.2 (Ideal Learning Problem)**.** Let $\lambda > 0$ be a strictly positive constant, let $\ell\colon \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ be continuous, and let $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d \times \mathcal{Y})$ be a given input data. Set

$$\widehat{J}(\mu, v) := \int_{\mathbb{R}^d \times \mathcal{Y}} \ell \mathrm{d}\mu_T + \int_0^T \int_{\mathbb{R}^d \times \mathcal{Y}} \frac{\lambda}{2} |v(x,t)|^2 \mathrm{d}\mu_t(x,y)\mathrm{d}t \tag{1.4}$$

for $\mu \in C([0,T]; (\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y}), W_2))$ and $v \in L^2(\mathrm{d}\mu_t \mathrm{d}t)$, where $v \in L^2(\mathrm{d}\mu_t \mathrm{d}t)$ means that the squared integral of $v(x,t)$ in the measure $\mathrm{d}\mu_t(x)$ over $\mathbb{R}^d$ is integrable in time t over $[0,T]$. Then an ideal learning problem constrained by ODE-Net is posed as the following constrained minimization problem:

$$\begin{aligned} &\inf \left\{ \widehat{J}(\mu, v) \;\middle|\; \mu \in C\big([0,T]; \big(\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y}), W_2\big)\big),\ v \in L^2(\mathrm{d}\mu_t \mathrm{d}t) \right\} \\ &\text{subject to} \\ &\begin{cases} \partial_t \mu_t + \mathrm{div}_x(v_t \mu_t) = 0 \quad \text{(in the sense of Definition 3.2)}, \\ \mu_t|_{t=0} = \mu_0. \end{cases} \end{aligned} \tag{1.5}$$

In contrast to Problem 1.1, where the parameter θ is a variable to the functional J, the vector field v itself is a variable to the functional $\widehat{J}$ in Problem 1.2. For this idealized problem containing a broader class of vector fields, we also establish the existence of a minimizer as in the following theorem.

Theorem 1.2 (Existence of a Minimizer)**.** *Under Assumption* 1.1*, there exists a minimizer* $\mu \in C([0,T]; (\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y}), W_2))$ *and* $v \in L^2(\mathrm{d}\mu_t \mathrm{d}t)$ *for Problem* 1.2*.*

We have been discussing the well-posedness of "learning" of ODE-Net by formulating it via a mean-field optimal control problem, in the sense that we have to control trajectories in the space of probability measures μ_t. Through the above discussion, it is suggested that such a learning framework successfully gives a mathematical way to analyze the learning processes of DNNs. Our main results obtained in this analysis are interesting from the viewpoint of the calculus of variations in that minimizers exist for nonlinear optimal control problems such as Problems 1.1 and 1.2. Moreover, the proofs of our theorems will ensure that every minimizing sequence contains a convergent subsequence in a suitable topology, leading to the well-posedness of sequential minimization algorithms such as gradient descent (GD).

1.2 Contributions of the paper

The present paper contributes to establishing the existence of a minimizer under situations where the regularization parameter λ is not necessarily large in Theorem 1.1. This situation can be addressed because, in contrast to the paper [10], we use an argument that does not rely on a strong convexity of J to prove the existence of a minimizer. In addition, this theorem can apply to unbounded and non-differentiable neural networks $v(x,\theta)$, which are important targets in applications.

As a comparison, a key to our convergence results is to obtain the existence of a minimizer of both μ and θ under reasonable assumptions. Bonnet *et al.* [10] required strong convexity of J, or a sufficiently large parameter λ, in order to obtain strong compactness. In addition, Thorpe and Gennip [57] and Esteve *et al.* [24] obtained existence results under an H^1-regularization of θ_t, and Herty *et al.* [29] under boundedness for the Lipschitz

constant of $\theta\colon [0,T] \to \mathbb{R}^m$, broadly speaking, both of them are assuming that the "differentials" of the parameters θ_t in time are controlled. While these studies are novel in that they do not impose assumptions on regularization parameters such as λ, the assumptions of the continuity or differentiability on parameters $\theta\colon [0,T] \to \mathbb{R}^m$ should be relaxed or removed because the functions that ODE-Net can approximate are limited and the expected value of ℓ cannot be sufficiently small. Furthermore, in the field of ensemble optimal control, an existence result in the L^2-setting using ODE similar to (2.4) is proved by Scagliotti [53, Theorem 3.2]. Pogodaev [45] proved the existence of optimal control of the continuity equation with parameters θ relaxed to Young measures on a bounded domain. As a corollary, the existence of optimal parameters follows if the neural network satisfies certain convexity conditions, but the continuity of an optimal curve μ^* with respect to t is not clear.

Theorem 1.1 also provides one theoretical justification for the experimental algorithm in [25]. The authors developed an algorithm to approximate the minimizer of the Benamou-Brenier type problem, which is guaranteed to exist. However, the guarantee does not hold for the algorithm because the vector field v in the continuity equation is constrained by the neural network v_θ. This study supplements the existence of a minimizer, even in this case.

In Section 5, we combine the neural network property of universal approximation with the training of ODE-Net in Problem 1.2. This new combination makes it possible to obtain existence results (Theorem 1.2) without the linearity assumption (Assumption 1.2). It is also interesting that Problem 1.2 has a similar formulation to (variational) mean-field game (MFG) [9, 35, 50]. This similarity between deep learning and MFG has recently been pointed out by E *et al.* [23] and Ruthotto *et al.* [47]. Our results are expected to suggest a strong connection between MFG and ODE-Net. In fact, for the proof of Theorem 1.2, we will give an auxiliary theorem (Lemma 5.1) that is proved via the so-called Lagrange perspective for easy handling of the vector fields v (see also [50, Section 2.2.2]).

1.3 Organization of the paper

This paper is organized as follows. In Section 2, we will give a brief review of previous studies on ODE-Net. In the first half, we summarize the history of the development of ODE-Net, and in the second half, we review mathematical formulations of the learning of ODE-Net. In Section 3, we will provide preliminary facts on the convergence of probability measures and distributional solutions of the continuity equation, which will be used to set up and prove our main results. In Section 4, we will prove Theorem 1.1. By virtue of the regularization term in (1.1) and the Benamou-Brenier formula in Lemma 3.6, we will easily get the appropriate compactness of minimizing sequences. Hence, we can apply the direct method of the calculus of variations to reach the existence results. In Section 5, we will exhibit how Problem 1.2 is formulated through an idealization in a detailed manner, and then we prove our main result (Theorem 1.2). One cannot prove the theorem by simply applying the arguments used in Section 4. Instead, we show the theorem by the use of a supplementary problem (see Problem 5.1 and Lemma 5.1) based on the Lagrange perspective. Section 6 presents a summary of the paper and discusses some tasks to be

undertaken in future studies. In Appendices A and B, we show and review the existence results for problems given by Bonnet *et al.* [10] and Thorpe and Gennip [57]. These problems adopt different regularization terms from Problem 1.1. By comparing the proofs of Theorems 1.1, A.1 and B.1, one can observe differences in how compactness is obtained to minimizing sequences.

2 Background and related works

This section provides an overview of previous research on learning of ODE-Net. Section 2.1 reviews how ODE-Net has been proposed. Section 2.2 describes how the learning has been formulated and discussed.

2.1 Background to the development of ODE-Net

Before describing the history of the development of ODE-Net, we shall review a type of DNN called ResNet that led to the improvement of DNN's performance. ResNet was devised to facilitate optimization of DNN [28]. The simplest L-layer ResNet consists of the difference equation

$$\begin{aligned} &x_0 = g(x,\theta),\\ &x_{t+1} = x_t + v(x_t,\theta_t), \quad t = 0,\dots,L-1,\\ &y = h(x_L,\theta_L), \end{aligned} \tag{2.1}$$

where $x \in \mathbb{R}^d$ is an input data, $y \in \mathcal{Y}$ denotes a final output, and $g(\bullet,\theta)\colon \mathbb{R}^d \to \mathbb{R}^{d_0}$ and $h(\bullet,\theta_L)\colon \mathbb{R}^{d_L} \to \mathcal{Y} \subset \mathbb{R}^{d_\mathcal{Y}}$ are some linear maps with parameters $\theta \in \mathbb{R}^{d_0\times d}$ and $\theta_L \in \mathbb{R}^{d_\mathcal{Y}\times d_L}$ respectively. In addition, $v(\bullet,\theta_t)\colon \mathbb{R}^{d_t} \to \mathbb{R}^{d_{t+1}}$ is multiple compositions of some affine maps with θ_t, and nonlinear functions, called activation functions, such as rectified linear unit (ReLU) [42]. Out of various models of (Deep) Neural Networks, we shall refer to the above mapping $v(\bullet,\theta_t)$ associated with ResNet as a neural network simply in this paper. Experimentally, ResNet is known to perform better than other DNNs. In particular, deep ResNet, i.e. (2.1) with $L \gg 1$ outperforms other machine learning methods.

When ResNet is very deep, it is natural to observe ResNet (2.1) as the explicit Euler discretization of an ODE with unit step size. With the pioneering works [22, 27, 54], a trend started to analyze DNNs and develop algorithms by replacing "discrete" DNNs with "continuum" ODEs. For example, Haber *et al.* [27] employed the linear stability analysis in the theory of dynamical systems to stabilize ResNet, and Lorin *et al.* [37] utilized the parallel computing for differential equations to speed up the training of ResNet. These "continuum" ODEs corresponding to DNNs are often called Neural ODE [16], or ODE-Net [46, 62]. Specifically, the following parameterized dynamical system is often called ODE-Net:

$$\begin{aligned} &x_0 = g(x,\theta),\\ &\dot{x}_t = v(x_t,\theta_t), \quad t \in (0,T),\\ &y = h(x_T,\theta_T), \end{aligned} \tag{2.2}$$

where $x \in \mathbb{R}^d, y \in \mathcal{Y}, g\colon \mathbb{R}^d \times \mathbb{R}^{d\times d} \to \mathbb{R}^d, h\colon \mathbb{R}^d \times \mathbb{R}^{d_{\mathcal{Y}}\times d} \to \mathcal{Y}$ and v are defined as in (2.1). Note that for simplicity, it is assumed that $x_t \in \mathbb{R}^d$ for any $t \in [0,T]$, and accordingly, the neural network $v(\bullet, \theta_t)$ becomes a vector field on $\mathbb{R}^d$. Also, the finite-dimensional parameters $\theta_0, \theta_1, \dots,$ and θ_{L-1} in (2.1) are replaced with a (measurable) function on $[0,T]$. While $\theta\colon [0,T] \to \mathbb{R}^m$ is sometimes supposed to be continuous for theoretical reasons, the function θ on $[0,T]$ can be discontinuous during the learning process as seen in [39, Fig. 2] and [6, Fig. 1]. Thus, we impose the Lebesgue integrability condition on θ in our setting. The terminal time $T > 0$ is an arbitrary given constant.

Although there is not so much mathematical research on ODE-Net, the basic properties of general DNNs have also been studied for ODE-Net. For example, ODE-Net has universal approximation properties proved by [56] and that the objective functional J has no local minima shown in [19, 20, 38]. It is also known that specific additional assumptions (e.g. continuity of $\theta\colon [0,T] \to \mathbb{R}^m$) are necessary to regard ResNet as the discretization of ODE-Net (see, e.g. [31, 49, 57]) and to guarantee the convergence of learning algorithms [31].

2.2 Formulations of the learning of ODE-Net and existence results

Practically, people want ODE-Net to output a desired y for an input x. For this purpose, ODE-Net needs to learn, i.e. we optimize the parameter θ in ODE-Net (2.2). Thus, it is necessary to establish a theory of the learning of ODE-Net. E *et al.* [22, 23] were the first to attempt a general formulation of the learning of ODE-Net (2.2). They modeled the learning as a mean-field optimal control problem as follows.

Problem 2.1 (Learning Problem Constrained by ODE-Net, [23, Eq. 3])**.** Let $\mathcal{Y} = \mathbb{R}^l$, let Θ be a subset of $\mathbb{R}^m$ and let $v\colon \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d, \ell\colon \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ and $L\colon \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}_+$ be continuous. For a given input data $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d \times \mathcal{Y})$, the learning problem constrained by ODE-Net is posed as the following constrained minimization problem:

$$\min_{\theta \in L^\infty(0,T;\Theta)} \mathbb{E}\left[\ell(x_T, y) + \int_0^T L(x_t, \theta_t)\mathrm{d}t\right] \tag{2.3}$$

subject to

$$\begin{cases} \dot{x}_t = v(x_t, \theta_t), \quad t \in (0,T), \\ (x_0, y) \sim \mu_0. \end{cases} \tag{2.4}$$

The meanings of symbols appearing in Problem 2.1 are as follows. The given probability measure μ_0 is called training data, a probability distribution of input-output pairs of a random variable (x, y) in (2.2) used for the learning. The vector field $v(\bullet, \theta)$, $\theta \in \mathbb{R}^m$, on $\mathbb{R}^d$ represents the neural network explained in (2.2). After expanded by the linearity of the expected values, the first term of (2.3) represents the expected value of a loss function $\ell(x, y)$, which is the target we want to make as small as possible during the learning process. One often uses the squared loss $\ell(x, y) = |x - y|^2/2$ for regression problems or the cross-entropy for classification problems (see, e.g. Pytorch's document for the specific form). However, when using a neural network with many parameters, minimizing only the loss $\mathbb{E}[\ell(x_T, y)]$ can lead to the so-called overfitting, see basic statistics and machine learning

textbooks, e.g. [41, Section 1.4.7]. To avoid this overfitting, we also minimize the second expected value, which is called a regularization term. For example, some researchers use the L^2-regularization $L(x,\theta) = \lambda|\theta|^2/2$, L^1-regularization $L(x,\theta) = \lambda|\theta|$, and entropy regularization used in [26, 31]. In addition, the kinetic regularization $L(x,\theta) = \lambda|v(x,\theta)|^2/2$ that Finlay *et al.* [25] proposed with the help of the Benamou-Brenier formula can make the trajectories of ODEs' solutions well-behaved. Another way to deal with the overfitting is to restrict $\Theta \subset \mathbb{R}^m$ to compact sets. In optimal control theory, by virtue of the compactness of Θ, one can easily show the existence of optimal parameters (see, e.g. [13, Theorem 5.1.1]). It should be noted that these various regularizations require an assumption upon a function space to which the parameters θ belong. As is seen in the above Problem 2.1, E *et al.* [23] set the function space to L^∞-space.

Remark 2.1 (On Neglecting Input and Output Transformations in (2.2))**.** ODE-Net introduced in (2.2) contains input and output transformations g and h, leading to a learning problem corresponding to a minimization with respect to $\theta \in \mathbb{R}^{d\times d}, \theta_\bullet \in L^2(0,T;\mathbb{R}^m)$ and $\theta_L \in \mathbb{R}^{d_y\times d}$. However, current theoretical studies of ODE-Net often use formulations that ignore $g(x,\theta)$ and $h(x,\theta_L)$, and consider minimization only in θ_t as in Problem 2.1. In the author's view, the reason for this neglect is that the existence of minimizers for θ and θ_L is easy to check if one proposes a variational formulation that considers g and h. For example, if one imposes the L^2-regularization $|\theta|^2 + |\theta_L|^2$ for the parameters $\theta \in \mathbb{R}^{d\times d}$ and $\theta_L \in \mathbb{R}^{d_y\times d}$ associated with $g(\bullet,\theta)\colon x \mapsto x_0$ and $h(\bullet,\theta_L)\colon x_T \mapsto y$ respectively, the existence of minimizers $\theta^* \in \mathbb{R}^{d\times d}$ and $\theta_L^* \in \mathbb{R}^{d_y\times d}$ follows immediately by virtue of the direct method of the Calculus of Variations, a minimizing sequence of $((\theta^n,\theta_L^n))_n$ has a convergent subsequence thanks to the Bolzano-Weierstrass theorem. Hence, only the ODE $\dot{x}_t = v(x_t,\theta_t)$ in (2.2) is sometimes referred to as ODE-Net. On the other hand, g and h should not be ignored when we explore the learning process, that is, the dynamics of solving the problem with mathematical optimization methods such as GD. It is reported that singular values of a parameter defining g and h affect the convergence of GD [7, Theorem 2].

On the other hand, for Problem 2.1, Bonnet *et al.* [10, Section 1.4] brought a measure-theoretical formulation inspired by mean-field optimal control problems. A trick used in their formulation is that laws μ_t, $t \in (0,T)$, of random variables (x_t, y) subject to (2.4) satisfy the following continuity equation:

$$\begin{cases} \partial_t \mu_t + \operatorname{div}_x\left(\mu_t(x,y)v(x,\theta_t)\right) = 0, \quad (x,y) \in \mathbb{R}^d \times \mathcal{Y}, \quad t \in (0,T), \\ \mu_t|_{t=0} = \mu_0, \end{cases}$$

in the sense of distributions defined in Definition 3.2. They utilized this trick to translate Problem 2.1 into the following Problem 2.2 in the case of $L(x,\theta) = \lambda|\theta|^2$.

Problem 2.2 (Measure-Theoretical Learning Problem, [10, Eq. 1.8])**.** Let $\lambda > 0$ be constants and let $v\colon \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ and $\ell\colon \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ be continuous. For a given input data $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d \times \mathbb{R}^d)$, the learning problem constrained by ODE-Net is posed as the following constrained minimization problem:

$$\min_{\theta\in L^2(0,T;\mathbb{R}^m)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \ell(x,y)\mathrm{d}\mu_T(x,y) + \lambda \int_0^T |\theta_t|^2 \mathrm{d}t \tag{2.5}$$

subject to

$$\begin{cases} \partial_t \mu_t + \mathrm{div}_x \left(v(x,\theta_t)\mu_t\right) = 0, \quad (x,y) \in \mathbb{R}^d \times \mathbb{R}^d, \quad t \in (0,T), \\ \mu_t|_{t=0} = \mu_0. \end{cases} \tag{2.6}$$

In addition, μ belongs to $C_w([0,T];\mathcal{P}_c(\mathbb{R}^d \times \mathbb{R}^d))$ which is the space of narrowly continuous curves (see also Definition 3.1).

As for Problem 2.2, Bonnet *et al.* [12, Theorem 3.2] studied the unique existence of a minimizer θ^* under the assumption that $\lambda > 0$ is sufficiently large and the neural network $v(x,\theta)$ is bounded for θ. In practice, however, in order to minimize the loss, the regularization parameter λ is usually set to be a sufficiently small positive number rather than a large one.

The difficulty in obtaining existence theorems to Problem 2.2 is attributed to the variational formulation. From (2.5) and (2.6), we observe that the learning of ODE-Net has the following aspects:

(i) The objective functional J in (2.5) is minimized over an infinite-dimensional space $L^2(0,T;\mathbb{R}^m)$.

(ii) The minimization is constrained by the continuity equation (2.6) which is a differential equation on the infinite-dimensional space of probability measures $\mathcal{P}(\mathbb{R}^d \times \mathcal{Y})$.

When one tries to show the existence of a minimizer for a variational problem such as Problem 2.2 by using the direct method of the Calculus of Variations, it is difficult to obtain the strong compactness of minimizing sequences due to the infinite dimensionality in (i). In addition, even if minimizing sequences converge, it is not generally obvious whether limits satisfy the continuity equation (2.6) mentioned in (ii).

3 Preliminaries

This section presents fundamental mathematical tools.

3.1 Compactness lemma

For $T > 0$, we denote by $C([0,T];X)$ the set of continuous mappings from $[0,T]$ to a topological space X with the uniform convergence topology.

Lemma 3.1 (Ascoli-Arzelá's Theorem)**.** *Let (X,d) be a metric space. Then, a family $\mathcal{F} \subset C([0,T];X)$ is relatively compact in the uniform convergence topology if and only if*

- *for each $t \in [0,T]$, the set $\{x \in X \mid x = f(t) \text{ for some } f \in \mathcal{F}\}$ is relatively compact in X, and*
- *$\mathcal{F}$ is equi-continuous.*

Proof. A more general version of the above lemma in the case where X is a uniform space is proved in, e.g. [32, Chapter 7.17]. □

3.2 Probability measures and the Wasserstein space

Hereinafter, $\mathcal{P}(X)$ denotes the set of Borel probability measures on a separable metric space X. Here, we review some definitions and lemmas regarding properties and convergence of probability measures, as well as properties of the Wasserstein space.

Definition 3.1. *Let $p \geq 1$ and let (X, d) be a Polish space, i.e. a complete and separable metric space.*

(i) (narrow convergence) A sequence (μ^n) in $\mathcal{P}(X)$ is said to be narrowly convergent to $\mu \in \mathcal{P}(X)$ as $n \to \infty$ if

$$\lim_{n\to\infty} \int_X f \mathrm{d}\mu^n = \int_X f \mathrm{d}\mu \quad \text{for every function} \quad f \in C_b(X),$$

where $C_b(X)$ is the space of continuous and bounded real functions defined on X. A topology induced by the convergence is said to be the narrow topology.

(ii) (uniformly integrable p-moments) A subset K in $\mathcal{P}(X)$ has uniformly integrable p-moments if

$$\lim_{R\to\infty} \sup_{\mu\in K} \int_{X\setminus B_X(R,\overline{x})} d(x,\overline{x})^p \mathrm{d}\mu(x) = 0 \quad \text{for some} \quad \overline{x} \in X,$$

where $B_X(R, x)$ is the open ball of radius R and center x in X.

(iii) (finite p-th moment) A probability measure $\mu \in \mathcal{P}(X)$ is said to have the finite p-th moment if

$$\int_X d(x,\overline{x})^p \mathrm{d}\mu(x) < \infty \quad \text{for some} \quad \overline{x} \in X,$$

and the set of probability measures on X with the finite p-th moment is denoted by $\mathcal{P}_p(X)$.

(iv) (function of p-growth) A function $f\colon X \to \mathbb{R}$ is said to have p-growth if there exist $A, B \geq 0$ and $\overline{x} \in X$ such that

$$|f(x)| \leq A + B\big(d(x,\overline{x})\big)^p, \quad \forall\, x \in X.$$

(v) (Wasserstein distance) The (L^p-)Wasserstein distance between $\mu^1, \mu^2 \in \mathcal{P}_p(X)$ is defined by

$$W_p(\mu^1,\mu^2) := \inf\left\{ \left(\int_{X^2} d(x_1,x_2)^p \mathrm{d}\pi(x_1,x_2) \right)^{\frac{1}{p}} \,\middle|\, \pi \in \Gamma(\mu^1,\mu^2) \right\},$$

where $\Gamma(\mu^1, \mu^2)$ denotes the set of all Borel probability measures π on X^2 such that for any measurable subset $A \subset X$,

$$\pi[A \times X] = \mu^1[A], \quad \pi[X \times A] = \mu^2[A].$$

By using the Hölder inequality, one easily gets

Corollary 3.1. *Let $1 \leq p < q < \infty$, let X be a Polish space and let $\mu^1, \mu^2 \in \mathcal{P}_q(X)$. Then $W_p(\mu^1, \mu^2) \leq W_q(\mu^1, \mu^2)$.*

Lemma 3.2 (Kantorovich-Rubinstein Duality, [59, Theorem 1.14])**.** *Let (X, d) be a Polish space and let $\rho_0, \rho_1 \in \mathcal{P}_1(X)$. Then*

$$W_1(\rho_0, \rho_1) = \sup \left\{ \int_X \varphi \mathrm{d}(\rho_1 - \rho_0) \,\middle|\, \varphi \in L^1(|\rho_1 - \rho_0|),\ \operatorname*{Lip}_X(\varphi) := \sup_{x \neq y \in X} \frac{|\varphi(x) - \varphi(y)|}{d(x,y)} \leq 1 \right\}.$$

Proof. See [21, Section 11.8]. □

A sufficient condition for a family with the uniformly integrable p-moments is known, and the proof of the following lemma is given for the sake of the reader's convenience.

Lemma 3.3 ([4, Section 5.1.1])**.** *Let $p \geq 1$. If a subset $K \subset \mathcal{P}(X)$ satisfies*

$$\sup_{\mu \in K} \int_X d(x, \overline{x})^{p_1} \mathrm{d}\mu(x) < +\infty$$

for some $p_1 > p$ and $\overline{x} \in X$, then K has uniformly integrable p-moments.

The following lemma shows a fine criterion that reveals whether a sequence $(\mu^n) \subset \mathcal{P}(X)$ has the uniformly integrable p-moments.

Lemma 3.4 (Narrow Convergence for p-Growth Functions)**.** *A sequence (μ^n) in $\mathcal{P}(X)$ has uniformly integrable p-moments if and only if*

1. *the sequence is narrowly convergent to $\mu \in \mathcal{P}(X)$, and*
2. *for every continuous function $f \colon X \to \mathbb{R}$ of p-growth,*

$$\lim_{n \to \infty} \int_X f \mathrm{d}\mu^n = \int_X f \mathrm{d}\mu.$$

Proof. See [4, Lemma 5.1.7]. □

By Lemma 3.4 and [4, Proposition 7.1.5], convergence in W_p and narrow convergence for p-growth functions are equivalent.

3.3 Continuity equation

The following definition and lemma are based on a famous text [4, Chapter 4], to which we refer the reader who wants a general discussion of the continuity equations.

Definition 3.2 (Solutions in the Sense of Distributions)**.** *Let $T > 0$. A continuous curve $\mu \in C_w([0,T]; \mathcal{P}(\mathbb{R}^d \times \mathcal{Y}))$ is called a solution to the continuity equation*

$$\partial_t \mu_t + \operatorname{div}_x(v_t \mu_t) = 0 \quad \text{in } (0,T) \times \mathbb{R}^d \times \mathcal{Y}, \tag{3.1}$$

in the sense of distribution, if

$$\int_0^T \int_{\mathbb{R}^d \times \mathcal{Y}} \big(\partial_t \psi_t(x,y) + \nabla_x \psi_t(x,y) \cdot v_t(x)\big) \mathrm{d}\mu_t(x,y) \mathrm{d}t = 0 \tag{3.2}$$

for every $\psi \in C_c^\infty((0,T) \times \mathbb{R}^d \times \mathcal{Y})$. *Here a mapping* $v_t \colon \mathbb{R}^d \ni x \mapsto v_t(x) \in \mathbb{R}^d$, $t \in [0,T]$, *is a Borel vector field.*

In the following, we adopt Definition 3.2 as the solution of the continuity equation (3.1) with a vector field v.

Lemma 3.5 (Representation Formula for (3.1), [4, Proposition 8.1.8])**.** *Let* $T > 0$ *and let* $\mu \in C_w([0,T]; \mathcal{P}(\mathbb{R}^d \times \mathcal{Y}))$ *be a distributional solution of* (3.1) *with Borel vector fields* $v = (v_t)_t$. *Assume that* v *satisfies that*

$$\int_0^T \int_{\mathbb{R}^d \times \mathcal{Y}} |v_t| \mathrm{d}\mu_t \mathrm{d}t < \infty, \tag{3.3}$$

and

$$\int_0^T \Big(\sup_K |v_t| + \mathop{\mathrm{Lip}}_K (v_t) \Big) \mathrm{d}t < \infty \quad \textit{for every compact set} \quad K \subset \mathbb{R}^d. \tag{3.4}$$

Here $\mathrm{Lip}_K(v_t)$ *denotes a Lipschitz constant of the mapping* $v_t \colon \mathbb{R}^d \to \mathbb{R}^d$ *on* K, *i.e.*

$$\mathop{\mathrm{Lip}}_K (v_t) := \sup_{x \neq y \in K} \frac{|v_t(y) - v_t(x)|}{|y - x|}.$$

Then, for μ_0*-a.e.* $(x,y) \in \mathbb{R}^d \times \mathcal{Y}$, *there exists a unique solution* $X_\bullet(x) \in C([0,T]; \mathbb{R}^d)$ *such that*

$$\begin{aligned} &X_0(x) = x, \\ &\frac{\mathrm{d}}{\mathrm{d}t} X_t(x) = v_t\big(X_t(x)\big). \end{aligned}$$

Furthermore, the solution μ_t *is represented as*

$$\mu_t = (X_t \times \mathrm{Id}_{\mathcal{Y}})_\# \mu_0, \quad \forall\, t \in [0,T], \tag{3.5}$$

where $\mathrm{Id}_X \colon X \to X$ *is the identity mapping on* X.

Proof. The existence result can be shown by the use of the standard argument of the Picard iteration method. For the representation result, details are proved in, e.g. [4, Proposition 8.1.8]. □

The following lemma indicates the strong relation between the Wasserstein distance and the continuity equation.

Lemma 3.6 (Benamou-Brenier Formula, [8])**.** *Let* $\rho_0, \rho_1 \in \mathcal{P}_2(\mathbb{R}^d)$. *Then*

$$W_2(\rho_0,\rho_1)^2 = \inf \left\{ \int_0^1 \int_{\mathbb{R}^d} |v_t(x)|^2 \mathrm{d}\rho_t(x) \mathrm{d}t \;\middle|\; (\rho, v) \in V(\rho_0,\rho_1) \right\}, \tag{3.6}$$

where

$$V(\rho_0,\rho_1) := \left\{ (\rho, v) \in C\big([0,1]; \mathcal{P}_2(\mathbb{R}^d)\big) \times L^2(\mathrm{d}\rho_t \mathrm{d}t) \;\middle|\; \begin{array}{l} (3.1) \textit{ holds in the sense of} \\ \textit{Definition } 3.2, \textit{ and} \\ \rho_t|_{t=0} = \rho_0, \;\; \rho_t|_{t=1} = \rho_1. \end{array} \right\}.$$

Proof. See [3, Theorem 17.2]. □

4 Kinetic regularization and an existence theorem

In this section, we discuss the existence of a minimizer to the kinetic regularized learning problem introduced in Problem 1.1. The section begins with some background on kinetic regularization. Subsequently, we proceed to the proof of Theorem 1.1. Throughout the paper, we denote by C a generic non-negative constant which may vary from line to line.

4.1 Kinetic regularization

In general, to argue minimizers of a functional J as in (1.1) via the direct method of the calculus of variations, a minimizing sequence $((\mu^n, \theta^n))_n$ of the functional needs to be compact in an appropriate topological space. Moreover, the topology must be sufficiently strong to lead to a (lower semi) continuity of the functional. Driven by this necessity, some previous studies have tried to strengthen the topology of the space of the parameter $\theta\colon [0,T] \to \mathbb{R}^m$ in [10,29,57]. However, this strong topology leads to unusual assumptions, as reviewed in Section 2.2. Instead, we seek for compactness of the continuous curves $\mu^n\colon [0,T] \to \mathcal{P}(\mathbb{R}^d \times \mathcal{Y})$, rather than of the parameters θ^n ($n \in \mathbb{N}$). This idea is rarely seen in machine learning but often in the MFG theory (see, e.g. [43, Theorem 6.6.] and [11, Theorem 6]). To illustrate this idea, we need the following lemma derived from the Benamou-Brenier formula (Lemma 3.6).

Lemma 4.1 (Uniform Continuity Estimate)**.** *Let $\mu \in C_w([0,T]; \mathcal{P}(\mathbb{R}^d \times \mathcal{Y}))$ be a distributional solution to the continuity equation* (3.1) *with Borel vector fields $v_t\colon \mathbb{R}^d \ni x \mapsto v_t(x) \in \mathbb{R}^d$, $t \in [0,T]$. Then it holds that*

$$W_2(\mu_t, \mu_s)^2 \leq (s-t) \int_t^s \int_{\mathbb{R}^d \times \mathcal{Y}} |v(\tau, x)|^2 \mathrm{d}\mu_\tau(x,y) \mathrm{d}\tau$$

for $0 \leq t < s \leq T$.

In the rest of the paper, we often abbreviate $\int_{\mathbb{R}^d \times \mathcal{Y}} f(x) \mathrm{d}\mu(x,y)$ to $\int_{\mathbb{R}^d} f \mathrm{d}\mu$ for a function $f\colon \mathbb{R}^d \to \mathbb{R}$ independent of $y \in \mathcal{Y}$.

Proof. From Lemma 3.6, we have

$$\begin{aligned} W_2(\mu_t, \mu_s)^2 &\leq \inf_{\rho, w} \left\{ \int_0^1 \int_{\mathbb{R}^d \times \mathcal{Y}} |w_t(x)|^2 \mathrm{d}\rho_t(x,y) \mathrm{d}t \;\middle|\; \partial_t \rho + \mathrm{div}_x(w\rho) = 0, \rho_0 = \mu_t, \rho_1 = \mu_s. \right\} \\ &= \inf_{\rho, w} \left\{ \int_t^s \int_{\mathbb{R}^d} |w_\tau|^2 \mathrm{d}\rho_\tau \mathrm{d}\frac{\tau}{s-t} \;\middle|\; (s-t)\partial_\tau \rho + \mathrm{div}_x(w\rho) = 0, \rho_t = \mu_t, \rho_s = \mu_s. \right\} \end{aligned}$$

$$= (s-t)\inf\left\{\int_t^s \int_{\mathbb{R}^d} |w_\tau|^2 \mathrm{d}\rho_\tau \mathrm{d}\tau \;\middle|\; \partial_\tau \rho + \mathrm{div}_x(w\rho) = 0, \rho_t = \mu_t, \rho_s = \mu_s.\right\}$$

$$\leq (s-t)\int_t^s \int_{\mathbb{R}^d} |v_\tau|^2 \mathrm{d}\mu_\tau \mathrm{d}\tau$$

for $0 \leq t < s \leq T$. □

This lemma readily leads to the following:

Corollary 4.1. *Let $n \in \mathbb{N}$ and let $\mu^n \in C_w([0,T]; \mathcal{P}_c(\mathbb{R}^d \times \mathcal{Y}))$ be a distributional solution to the continuity equation* (3.1) *corresponding to Borel vector fields $(v_t^n)_{t\in[0,T]}$. If*

$$\sup_{n\in\mathbb{N}} \int_0^1 \int_{\mathbb{R}^d} \left|v_t^n\right|^2 \mathrm{d}\mu_t \mathrm{d}t < \infty,$$

then the family (μ^n) is equi-continuous.

To use Corollary 4.1 explicitly, we add a term

$$\frac{\lambda}{2}|v(x,\theta)|^2, \quad \lambda > 0, \tag{4.1}$$

to the objective functional J (1.1) in Problem 1.1. This regularization term $|v(x,\theta)|^2$ is reported to be effective in generative models in [25]. Here, we use kinetic regularization for simplicity, but in fact, one can prove the existence of a minimizer without a kinetic regularization term. See Remark 4.2 for details.

4.2 Existence theorem

Our strategy is to use the direct method of the calculus of variations, containing the following three steps:

(i) Take a minimizing sequence $((\mu^n, \theta^n))_n$ and extract a convergent subsequence in suitable topologies.

(ii) Check that J is lower semicontinuous with respect to those topologies.

(iii) Verify that the limits of convergent subsequences satisfy the constraint (1.2).

As for a minimizing sequence, we get a weakly convergent subsequence of (θ^n) in $L^2(0,T;\mathbb{R}^m)$ and the strongly convergent subsequence of (μ^n) in $C([0,T]; (\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y}), W_2))$ by virtue of Corollary 4.1 and the Ascoli-Arzelá theorem. From these convergences and Assumption 1.2, we observe that the functional J is lower semicontinuous in (μ, θ). Also, we can verify that the limits solve the continuity equation again. This is why we impose the kinetic regularization term onto the functional J.

For the proof of Theorem 1.1, we need a lemma on the boundedness of the support of μ_t uniformly in $t \in [0,T]$.

Lemma 4.2. *Let $\theta \in L^2(0,T;\mathbb{R}^m)$ and let $\mu \in C([0,T];(\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y}), W_2))$ be a distributional solution of* (1.2) *corresponding to vector fields $(v(\bullet,\theta_t))_{t\in[0,T]}$. If Assumption* 1.2 *holds, there exists a radius $R^* = R^*(\mu_0, f, T, \|\theta\|_{L^2(0,T;\mathbb{R}^m)}) > 0$ such that*

$$\operatorname{supp}\mu_t \subset B_{\mathbb{R}^d\times\mathcal{Y}}(R^*,0) := B_{\mathbb{R}^d\times\mathcal{Y}}(R^*), \quad \forall\, t\in[0,T].$$

Proof. To use Lemma 3.5, we check the assumptions (3.3) and (3.4). By Assumption 1.2 and the Lipschitz continuity of f, we have

$$\begin{aligned}
&\int_0^T \int_{\mathbb{R}^d\times\mathcal{Y}} |v(x,\theta_t)| \mathrm{d}\mu_t \mathrm{d}t \\
&\le \int_0^T \int_{\mathbb{R}^d\times\mathcal{Y}} |\theta_t|\,|f(x)| \mathrm{d}\mu_t \mathrm{d}t \\
&\le \|\theta\|_{L^2(0,T;\mathbb{R}^m)} \sqrt{\int_0^T \left(\int_{\mathbb{R}^d\times\mathcal{Y}} |f(x)| \mathrm{d}\mu_t(x,y)\right)^2 \mathrm{d}t} \\
&\le \sqrt{T}\|\theta\|_{L^2(0,T;\mathbb{R}^m)} \sqrt{\sup_{t\in[0,T]} \int_{\mathbb{R}^d\times\mathcal{Y}} |f(x)|^2 \mathrm{d}\mu_t(x,y)} \\
&\le C\|\theta\|_{L^2(0,T;\mathbb{R}^m)} \left(1 + \sup_{t\in[0,T]} \sqrt{\int_{\mathbb{R}^d\times\mathcal{Y}} |x|^2 \mathrm{d}\mu_t(x,y)}\right) < \infty.
\end{aligned}$$

Similarly, it holds that

$$\int_0^T \Big(\sup_K |v(\bullet,\theta_t)| + \operatorname*{Lip}_K |v(\bullet,\theta_t)|\Big)\mathrm{d}t \le C\left(T + \|\theta\|^2_{L^2(0,T;\mathbb{R}^m)}\right) < \infty$$

for every compact set $K \subset \mathbb{R}^d \times \mathcal{Y}$. We thus find from Lemma 3.5 that μ_t can be represented as $\mu_t = (X_t, \mathrm{Id}_{\mathcal{Y}})_{\#}\mu_0$ where $X_t\colon \mathbb{R}^d \to \mathbb{R}^d$ is the flow maps of the corresponding ODE satisfying

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}X_t(x)} t = v\big(X_t(x),\theta_t\big), \quad t\in(0,T), \\ X_0(x) = x \end{cases}$$

for almost all $(x,y) \in \operatorname{supp}\mu_0$. By Grönwall's inequality and Assumption 1.2, we have

$$\begin{aligned}
|X_t(x)| &\le \left(|x| + C\int_0^T |\theta_s|\mathrm{d}s\right) \exp\left(C\int_0^T |\theta_s|\mathrm{d}s\right) \\
&\le C\Big(\operatorname{diam}(\operatorname{supp}\mu_0) + T + \sqrt{T}\|\theta\|_{L^2(0,T;\mathbb{R}^m)}\Big) \exp\left(C\sqrt{T}\|\theta\|_{L^2(0,T;\mathbb{R}^m)}\right) \le R^*
\end{aligned}$$

for some $R^* > 0$ independent of t since $\theta \in L^2(0,T;\mathbb{R}^m)$, whence follows $\operatorname{supp}\mu_t \subset B(R^*)$ for all $t \in [0,T]$. □

Remark 4.1. Assumption 1.2 can be generalised to Assumption 6.1 when one only proves Lemma 4.2. See also Lemma A.1.

With Lemma 4.2, one can now proceed with the proof of Theorem 1.1.

Proof of Theorem 1.1. Set

$$S = \left\{ (\mu, \theta) \in C\big([0,T]; \big(\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y}), W_2\big)\big) \times L^2(0,T;\mathbb{R}^m) \;\middle|\; (\mu,\theta) \text{ satisfies (1.2)} \right\}.$$

Since $(\mu_0, 0) \in S$ is a trivial and regular solution to the continuity equation, we see that $S \neq \emptyset$. It is also obvious that $0 \leq \inf J < +\infty$ because the integrand ℓ is non-negative. Then, there exists a minimizing sequence $((\mu^n, \theta^n))_{n=1}^{\infty} \subset S$ such that $J(\mu^n, \theta^n) \to \inf_S J$ as $n \to \infty$. For the sequence, there exists a constant $C > 0$ independent of n such that

$$\frac{\lambda}{2} \int_0^T \int_{\mathbb{R}^d} \left|\theta_t^n f(x)\right|^2 \mathrm{d}\mu_t^n(x)\mathrm{d}t \leq C, \tag{4.2}$$

$$\frac{\epsilon}{2} \int_0^T \left|\theta_t^n\right|^2 \mathrm{d}t \leq C. \tag{4.3}$$

From Lemma 4.1 and (4.2), we have for $0 \leq t < s \leq T$,

$$W_2\big(\mu_t^n, \mu_s^n\big)^2 \leq (s-t) \int_t^s \int_{\mathbb{R}^d} \left|\theta_\tau^n f(x)\right|^2 \mathrm{d}\mu_\tau^n(x)\mathrm{d}\tau \leq \frac{2C}{\lambda}(s-t).$$

Hence, it follows that $(\mu^n) \subset C([0,T]; (\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y}), W_2))$ is equi-continuous. Also, by Lemma 4.2 and (4.3), there exists a constant $R^* > 0$ independent of n and t such that

$$\mu_t^n \in \left\{ \mu \in \mathcal{P}_c(\mathbb{R}^d \times \mathcal{Y}) \;\middle|\; \operatorname{supp}\mu \subset B_{\mathbb{R}^d \times \mathcal{Y}}(R^*) \right\} \quad \forall\, n \in \mathbb{N}, \quad \forall\, t \in [0,T].$$

In addition, the set $\{\mu \in \mathcal{P}_c(\mathbb{R}^d \times \mathcal{Y}) | \operatorname{supp}\mu \subset B_{\mathbb{R}^d \times \mathcal{Y}}(R^*)\}$ is compact with respect to L^2-Wasserstein topology because of [4, Proposition 7.1.5]. Hence, Lemma 3.1 and (4.3) imply that there exist a subsequence of (n), still denoted by n, $\mu^* \in C([0,T]; (\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y}), W_2))$ and $\theta^* \in L^2(0,T;\mathbb{R}^m)$ such that

$$\mu^n \;\to\; \mu^* \quad \text{strongly in } C\big([0,T]; \big(\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y}), W_2\big)\big), \tag{4.4}$$

$$\theta^n \;\to\; \theta^* \quad \text{weakly in } L^2(0,T;\mathbb{R}^m). \tag{4.5}$$

By the following Claim 4.1, we can deduce that (μ^*, θ^*) solves (1.2) in the sense of distribution.

Claim 4.1. For the limits μ^* and θ^*, it holds that

$$\int_0^T \int_{\mathbb{R}^d \times \mathcal{Y}} \big(\partial_t \zeta_t + \nabla_x \zeta_t \cdot v\big(\bullet, \theta_t^*\big)\big) \mathrm{d}\mu_t^* \mathrm{d}t = 0 \tag{4.6}$$

for all $\zeta \in C_c^\infty((0,T) \times \mathbb{R}^d \times \mathcal{Y})$. Moreover, $\operatorname{supp}\mu^* \subset B_{\mathbb{R}^d \times \mathcal{Y}}(R)$ for some $R > 0$.

Proof. We already know that

$$\int_0^T \int_{\mathbb{R}^d \times \mathcal{Y}} \partial_t \zeta_t \mathrm{d}\mu_t^n \mathrm{d}t + \int_0^T \int_{\mathbb{R}^d \times \mathcal{Y}} \nabla_x \zeta_t \cdot v\big(x, \theta_t^n\big) \mathrm{d}\mu_t^n \mathrm{d}t = 0 \tag{4.7}$$

for all $\zeta \in C_c^\infty((0,T)\times\mathbb{R}^d\times\mathcal{Y})$ and $n\in\mathbb{N}$. It follows that

$$\begin{aligned}
0 &= \int_0^T \int_{\mathbb{R}^d\times\mathcal{Y}} \partial_t\zeta_t \,\mathrm{d}\big(\mu_t^n - \mu_t^*\big)\mathrm{d}t \\
&\quad + \int_0^T \int_{\mathbb{R}^d\times\mathcal{Y}} \nabla_x\zeta_t(x)\cdot v\big(x,\theta_t^*\big)\mathrm{d}\big(\mu_t^n-\mu_t^*\big)(x)\mathrm{d}t \\
&\quad + \int_0^T \int_{\mathbb{R}^d\times\mathcal{Y}} \nabla_x\zeta_t(x)\cdot \big(v\big(x,\theta_t^n\big)-v\big(x,\theta_t^*\big)\big)\,\mathrm{d}\mu_t^*(x)\mathrm{d}t \\
&\quad + \int_0^T \int_{\mathbb{R}^d\times\mathcal{Y}} \nabla_x\zeta_t(x)\cdot \big(v\big(x,\theta_t^n\big)-v\big(x,\theta_t^*\big)\big)\,\mathrm{d}\big(\mu_t^n(x)-\mu_t^*(x)\big)\mathrm{d}t \\
&\quad + \int_0^T \int_{\mathbb{R}^d\times\mathcal{Y}} \partial_t\zeta_t \,\mathrm{d}\mu_t^*\mathrm{d}t + \int_0^T \int_{\mathbb{R}^d\times\mathcal{Y}} \nabla_x\zeta_t(x)\cdot v\big(x,\theta_t^*\big)\mathrm{d}\mu_t^*(x)\mathrm{d}t \\
&=: I_1+I_2+I_3+I_4+I_5. \qquad (4.8)
\end{aligned}$$

It follow from (4.4) that

$$\begin{aligned}
|I_1| &\le \int_0^T \operatorname*{Lip}_{\mathbb{R}^d\times\mathcal{Y}}(\partial_t\zeta_t)\int_{\mathbb{R}^d\times\mathcal{Y}} \frac{\partial_t\zeta_t}{\mathrm{Lip}_{\mathbb{R}^d\times\mathcal{Y}}(\partial_t\zeta_t)}\,\mathrm{d}\big(\mu_t^n-\mu_t^*\big)\mathrm{d}t \\
&\le C\int_0^T W_1\big(\mu_t^n,\mu_t^*\big)\mathrm{d}t \le C\int_0^T W_2\big(\mu_t^n,\mu_t^*\big)\mathrm{d}t \\
&\le CT\sup_{t\in[0,T]} W_2\big(\mu_t^n,\mu_t^*\big) \;\to\; 0,
\end{aligned}$$

as $n\to\infty$ by the Kantorovich-Rubinstein duality (Lemma 3.2) and Corollary 3.1. By Assumption 1.2, the function $\partial_t\zeta_t(x)v(x,\theta_t^*)$ is Lipschitz continuous in x and y over $\mathbb{R}^d\times\mathcal{Y}$, and thus we see again from Lemma 3.2 and Corollary 3.1 that

$$\begin{aligned}
|I_2| &\le C\int_0^T W_1\big(\mu_t^n,\mu_t^*\big)\mathrm{d}t \le C\int_0^T W_2\big(\mu_t^n,\mu_t^*\big)\mathrm{d}t \\
&\le CT\sup_{t\in[0,T]} W_2\big(\mu_t^n,\mu_t^*\big) \;\to\; 0 \quad\text{as}\quad n\to\infty.
\end{aligned}$$

For I_3, we use Assumption 1.2 to apply (4.5). We set

$$\varphi_\bullet := \int_{\mathbb{R}^d\times\mathcal{Y}} \nabla_x\zeta_\bullet f^\top \mathrm{d}\mu_\bullet^* \in L^2(0,T;\mathbb{R}^m).$$

In fact, it is shown that

$$\|\varphi\|_{L^2(0,T;\mathbb{R}^m)}^2 \le \int_0^T\int_{\mathbb{R}^d\times\mathcal{Y}} \big|\nabla_x\zeta_t f^\top\big|^2\mathrm{d}\mu_t^n\mathrm{d}t \le C\bigg(1+\sup_{t\in[0,T]} W_2\big(\mu_t^n,\delta_0\big)\bigg) < \infty.$$

Then, we can deduce that

$$I_3 = \int_0^T \big\langle \varphi_t, \theta_t^n-\theta_t^*\big\rangle\mathrm{d}t \;\to\; 0 \quad\text{as}\quad n\to\infty,$$

where $\langle A,B\rangle := \mathrm{Tr}(A^\top B)$, $A,B\in\mathbb{R}^{d\times p}$ is the inner product on $\mathbb{R}^{d\times p}$.

As for I_4, it follows from (4.4) and (4.5) that

$$\begin{aligned}
|I_4| &= \left| \int_0^T \left\langle \theta_t^n - \theta_t^*, \int_{\mathbb{R}^d \times \mathcal{Y}} \nabla_x \zeta_\bullet f^\top \mathrm{d}\big(\mu_t^n - \mu_t^*\big) \right\rangle \mathrm{d}t \right| \\
&\leq C \int_0^T \big|\theta_t^n - \theta_t^*\big| W_1\big(\mu_t^n, \mu_t^*\big)\, \mathrm{d}t \\
&\leq C \|\theta^n - \theta^*\|_{L^2(0,T;\mathbb{R}^m)} \sup_{t\in[0,T]} W_2\big(\mu_t^n, \mu_t^*\big) \\
&\leq C \sup_{t\in[0,T]} W_2\big(\mu_t^n, \mu_t^*\big) \;\; \to \;\; 0.
\end{aligned}$$

Passing to the limit as $n \to \infty$ in (4.8), we get (4.6).

By the lower semicontinuity of the L^2-norm, we have

$$\|\theta_\bullet^*\|_{L^2(0,T;\mathbb{R}^m)} \leq \sup_{n\in\mathbb{N}} \|\theta_\bullet^n\|_{L^2(0,T;\mathbb{R}^m)} < \infty.$$

Then Lemma 4.2 implies that there exists $R > 0$ such that $\operatorname{supp} \mu_t^* \subset B_{\mathbb{R}^d\times\mathcal{Y}}(R)$ for all $t \in [0,T]$, whence follows the conclusion. □

We resume the proof of Theorem 1.1. From Claim 4.1, we have $(\mu^*, \theta^*) \in S$. We then show that $J(\mu^n, \theta^n) \to J(\mu^*, \theta^*)$ as $n \to \infty$. First, from (4.4), Assumption 1.1, Lemma 3.4 and [4, Proposition 7.1.5], it follows that

$$\int_{\mathbb{R}^d\times\mathcal{Y}} \ell \mathrm{d}\mu_T^n \;\; \to \;\; \int_{\mathbb{R}^d\times\mathcal{Y}} \ell \mathrm{d}\mu_T^* \tag{4.9}$$

as $n \to \infty$. We next estimate the regularization term. Again from (4.3) and (4.4), we infer that

$$\begin{aligned}
&\left| \int_0^T \int_{\mathbb{R}^d} \big|\theta_t^n f(x)\big|^2 \mathrm{d}\mu_t^n(x)\mathrm{d}t - \int_0^T \int_{\mathbb{R}^d} \big|\theta_t^n f(x)\big|^2 \mathrm{d}\mu_t^*(x)\mathrm{d}t \right| \\
&= \left| \int_0^T \left\langle \theta_t^n \int_{\mathbb{R}^d} ff^\top \mathrm{d}\big(\mu_t^n - \mu_t^*\big), \theta_t^n \right\rangle \mathrm{d}t \right| \\
&\leq \int_0^T \left| \int_{\mathbb{R}^d} ff^\top \mathrm{d}\big(\mu_t^n - \mu_t^*\big) \right| |\theta_t^n|^2 \mathrm{d}t \\
&\leq \frac{2C}{\epsilon} \max_{t\in[0,T]} \left| \int_{\mathbb{R}^d} ff^\top \mathrm{d}\big(\mu_t^n - \mu_t^*\big) \right| \;\; \to \;\; 0 \quad \text{as} \quad n \to \infty.
\end{aligned} \tag{4.10}$$

Now we set

$$|||*|||^2_{L^2(0,T;\mathbb{R}^m)} := \int_0^T \left\langle \theta_t \left(\lambda \int_{\mathbb{R}^d} ff^\top \mathrm{d}\mu_t^* + \epsilon \right), \theta_t \right\rangle \mathrm{d}t$$

for $\theta \in L^2(0,T;\mathbb{R}^m)$. Since $\lambda \int_{\mathbb{R}^d} ff^\top \mathrm{d}\mu_t^* + \epsilon$ is positive definite matrix, the function $|||\bullet|||_{L^2(0,T;\mathbb{R}^m)}$ defines an equivalent norm of $L^2(0,T;\mathbb{R}^m)$. Hence, it follows from [14, Proposition 3.5] that $|||\bullet|||_{L^2(0,T;\mathbb{R}^m)}$ is weakly lower semicontinuous. Thus, we conclude that

$$
\begin{aligned}
\inf_S J &= \liminf_{n\to\infty} J(\mu^n, \theta^n) \\
&= \int_{\mathbb{R}^d\times\mathcal{Y}} \ell \mathrm{d}\mu_T^* + \liminf_{n\to\infty} \int_0^T \int_{\mathbb{R}^d} \left(\frac{\lambda}{2} \left|\theta_t^n f(x)\right|^2 + \frac{\epsilon}{2} \left|\theta_t^n\right|^2 \right) \mathrm{d}\mu_t^n(x)\mathrm{d}t \\
&= \int_{\mathbb{R}^d\times\mathcal{Y}} \ell \mathrm{d}\mu_T^* + \liminf_{n\to\infty} \int_0^T \int_{\mathbb{R}^d} \left(\frac{\lambda}{2} \mathrm{Tr}\left(\theta_t^n f(x) f(x)^\top \theta_t^{n\top}\right) + \frac{\epsilon}{2} \left|\theta_t^n\right|^2 \right) \mathrm{d}\mu_t^n(x)\mathrm{d}t \\
&= \int_{\mathbb{R}^d\times\mathcal{Y}} \ell \mathrm{d}\mu_T^* + \liminf_{n\to\infty} \int_0^T \int_{\mathbb{R}^d} \left(\frac{\lambda}{2} \left\langle \theta_t^n f(x) f(x)^\top, \theta_t^n \right\rangle + \frac{\epsilon}{2} \left\langle \theta_t^n, \theta_t^n \right\rangle \right) \mathrm{d}\mu_t^n(x)\mathrm{d}t \\
&= \int_{\mathbb{R}^d\times\mathcal{Y}} \ell \mathrm{d}\mu_T^* + \frac{1}{2} \liminf_{n\to\infty} \int_0^T \left\langle \theta_t^n \left(\lambda \int_{\mathbb{R}^d} f f^\top \mathrm{d}\mu_t^n + \epsilon \right), \theta_t^n \right\rangle \mathrm{d}t \\
&\geq \int_{\mathbb{R}^d\times\mathcal{Y}} \ell \mathrm{d}\mu_T^* + \frac{1}{2} \liminf_{n\to\infty} |||\theta^n|||_{L^2(0,T;\mathbb{R}^m)}^2 \\
&\quad + \frac{\lambda}{2} \liminf_{n\to\infty} \left(\int_0^T \int_{\mathbb{R}^d} \left|\theta_t^n f(x)\right|^2 \mathrm{d}\mu_t^n(x)\mathrm{d}t - \int_0^T \int_{\mathbb{R}^d} \left|\theta_t^n f(x)\right|^2 \mathrm{d}\mu_t^*(x)\mathrm{d}t \right) \\
&\geq \int_{\mathbb{R}^d\times\mathcal{Y}} \ell \mathrm{d}\mu_T^* + \frac{1}{2} |||\theta^*|||_{L^2(0,T;\mathbb{R}^m)}^2 + 0 \\
&= J(\mu^*, \theta^*) \geq \inf_S J,
\end{aligned} \tag{4.11}
$$

i.e. $J(\mu^*, \theta^*) = \inf_S J$, and the proof of Theorem 1.1 is complete. □

Remark 4.2 (The Case of $\lambda = 0$)**.** If one only wants to show the existence of a minimizer, it is sufficient to use only $\epsilon|\theta|^2/2$ as the regularization term. In other words, we can prove the theorem when $\lambda = 0$. Indeed, by Lemma 4.2 it is apparent that

$$
\int_{\mathbb{R}^d} |x|^2 \mathrm{d}\mu_t^n \leq (R^*)^2
$$

holds for every $t \in [0, T]$ and $n \in \mathbb{N}$, where $R^* > 0$ is the same as the radius in Lemma 4.2. Thus, we obtain

$$
\begin{aligned}
W_2\left(\mu_t^n, \mu_s^n\right)^2 &\leq (s-t) \int_t^s \int_{\mathbb{R}^d} \left|\theta_\tau^n f(x)\right|^2 \mathrm{d}\mu_\tau^n(x)\mathrm{d}\tau \\
&\leq (s-t) \left(\|\theta\|_{L^2(s,t;\mathbb{R}^m)}^2 \left\| \int_{\mathbb{R}^d} |f(x)|^2 \mathrm{d}\mu_\bullet^n(x) \right\|_{L^\infty(s,t)} \right) \\
&\leq \frac{2C}{\epsilon} (s-t) \left\| \int_{\mathbb{R}^d} |f(x)|^2 \mathrm{d}\mu_\bullet^n(x) \right\|_{L^\infty(s,t)} \\
&\leq \frac{2C}{\epsilon} (s-t) \left\| (\mathrm{Lip}\, f)^2 \int_{\mathbb{R}^d} |x|^2 \mathrm{d}\mu_\bullet^n(x) + |f(0)|^2 \right\|_{L^\infty(s,t)} \\
&\leq \frac{2C}{\epsilon} \left((R^* \,\mathrm{Lip}\, f)^2 + |f(0)|^2 \right) (s-t)
\end{aligned}
$$

from Lemma 4.1 and Assumption 1.2, or (6.1), and (4.3). Here $\mathrm{Lip}\, f \geq 0$ is a Lipschitz constant of f. Consequently, we can guarantee the equi-continuity of the curve μ without

the kinetic regularization term. Then, we complete the proof using an argument similar to the one above. In this case, the proof of the lower semicontinuity (4.11) becomes rather simple. It is noteworthy, however, that even in this case, deriving the convergence of I_3 in the proof of Claim 4.1 from the weak convergence of θ is difficult without imposing Assumption 1.2. In this sense, it seems essential under L^2-regularization that the neural network is linear with respect to the parameters. If continuity of θ can be obtained, e.g. by H^1-regularization, then the convergence of I_3 can be easily shown. See the proof of Theorem A.1.

Remark 4.3 (Uniqueness of a Minimizer)**.** When a regularization parameter ϵ is sufficiently large, the uniqueness of the minimizer is proved by Bonnet *et al.* [10, Theorem 3.2]. For self-containdness, we will present details in Appendix B. On the contrary, we do not refer to the uniqueness in Theorem 1.1 since ϵ might be small in most practical cases.

5 Ideal learning problem

This section discusses the existence of a minimizer to the ideal learning problem as introduced in Problem 1.2. At the beginning of this section, we explain why we consider this problem before we prove the main theorem (Theorem 1.2).

5.1 Idealization of learning problems

Neural networks v in (1.2) have, in general, a complex structure, while Assumption 1.2 imposes the simplicity of linearity of v in θ to prove Theorem 1.1. However, it is difficult to show Claim 4.1 under the general assumption because the trick described in (4.11) is unavailable.

Thus, we assume that $v(\bullet, \theta)$ can be any square-integrable vector field. This assumption might be justified by universal approximation properties resulting from the complexity. Universal approximation means that the set of functions expressed by a neural network $\{v(\bullet, \theta) \colon \mathbb{R}^d \to \mathbb{R}^d | \theta \in \mathbb{R}^m\}$ is dense in appropriate function spaces (for example, Lebesgue spaces $L^p(\mathbb{R}^d)$). We refer the reader to [18, 30, 44] for details. In light of these results, one can infer that

$$\overline{\left\{\mathbb{R}^d \times [0,T] \ni (x,t) \longmapsto v(x,\theta_t) \in \mathbb{R}^d \;\middle|\; \theta \in L^2(0,T;\mathbb{R}^m)\right\}}^{\|\bullet\|_{L^2(\mathrm{d}\mu_t\mathrm{d}t)}} = L^2(\mathrm{d}\mu_t\mathrm{d}t)$$

holds where, by abuse of notation, we set for a fixed $\mu \in C([0,T];\mathcal{P}(\mathbb{R}^d))$,

$$L^2(\mathrm{d}\mu_t\mathrm{d}t) := \Bigg\{ (v_t)_{t\in[0,T]} \text{ is a family of Borel vector fields on } \mathbb{R}^d \;\Bigg|\; \|v\|^2_{L^2(\mathrm{d}\mu_t\mathrm{d}t)} := \int_0^T \int_{\mathbb{R}^d} |v_t|^2 \mathrm{d}\mu_t\mathrm{d}t < +\infty \Bigg\}. \tag{5.1}$$

This abuse is referred to in the notation in Villani's text [59, Eq. (8.6)]. Furthermore, if $\epsilon = 0$ in (1.1), θ only appears via v in Problem 1.1.

Therefore, we can regard Problem 1.1 as a problem about vector fields (neural network) v, rather than parameters θ. From the above, we consider Problem 1.2 as the further idealized learning problem. The problem is similar to the variational form of MFG introduced by Lasry and Lions [34,35]. We also refer the reader to more comprehensive lecture notes by Santambrosio [50, Section 2.2].

5.2 Proof of the existence via the Lagrangian framework

We then consider the existence of a minimizer in Problem 1.2. For this problem, we want to apply a similar argument to Theorem 1.1. However, unlike the previous problem, the space $L^2(\mathrm{d}\mu_t\mathrm{d}t)$ depends on μ, rendering it intractable. In such cases, it is helpful to rewrite the problem with "probability measure on curves" as the variable instead of "curve on probability measures" or $\mu \in C([0,T];\mathcal{P}(\mathbb{R}^d))$ as the variable. That is, we consider $Q \in \mathcal{P}(\mathbb{R}^d \times C([0,T];\mathbb{R}^d))$ as presented in Proposition 5.1. Here $AC^2([0,T];\mathbb{R}^d)$ denotes the set of an absolutely continuous curve $\gamma\colon [0,T] \to \mathbb{R}^d$ such that there exists $m \in L^2(0,T)$ satisfying

$$|\gamma(s)-\gamma(t)| \le \int_s^t m(\tau)\mathrm{d}\tau, \quad \forall\, s,t \in [0,T], \quad s<t.$$

Proposition 5.1 (Probabilistic Representation)**.** *Let $\mu \in C([0,T];(\mathcal{P}_2(\mathbb{R}^d),W_2))$ satisfy the continuity equation $\partial_t\mu_t + \mathrm{div}(v_t\mu_t) = 0$ in the distributional sense for a Borel vector field v_t such that*

$$\int_0^T \int_{\mathbb{R}^d} |v_t|^2 \mathrm{d}\mu_t\mathrm{d}t < +\infty. \tag{5.2}$$

Then, there exists $Q \in \mathcal{P}(\mathbb{R}^d \times C([0,T];\mathbb{R}^d))$ such that

(i) Q is concentrated on the set of pairs (x,γ) such that $\gamma \in AC^2([0,T],\mathbb{R}^d)$ is an absolutely continuous solution of $\dot\gamma(t) = v_t(\gamma(t))$ for a.a. $t \in (0,T)$ with $\gamma(0) = x$;

(ii) $\mu_t = \mu_t^Q$ for any $t \in [0,T]$, where μ_t^Q is defined as

$$\int_{\mathbb{R}^d} \varphi \mathrm{d}\mu_t^Q := \int_{\mathbb{R}^d \times C([0,T];\mathbb{R}^d)} \varphi\big(\gamma(t)\big)\mathrm{d}Q(x,\gamma) \tag{5.3}$$

for all $\varphi \in C_b(\mathbb{R}^d)$.

Conversely, if $Q \in \mathcal{P}(\mathbb{R}^d \times C([0,T];\mathbb{R}^s))$ satisfies (i) and

$$\int_{\mathbb{R}^d \times C([0,T];\mathbb{R}^d)} \int_0^T |\dot\gamma(t)|^2 \mathrm{d}t\mathrm{d}Q(x,\gamma) < +\infty, \tag{5.4}$$

then there exists $\mu^Q \in C([0,T];(\mathcal{P}_2(\mathbb{R}^d),W_2))$ induced via (5.3)*, which is a solution of the continuity equation with the following vector field:*

$$\tilde v_t(x) := \int_{\{\gamma \in C([0,T];\mathbb{R}^d)\,|\,\gamma(t)=x\}} \dot\gamma(t)\mathrm{d}Q_x(\gamma) \in L^2\big(\mathrm{d}\mu_t^Q\mathrm{d}t\big), \quad \mu_t^Q\text{-a.e. } x \in \mathbb{R}^d,$$

where Q_x is the disintegrated measures with respect to the evaluation map $\mathsf{e}_t\colon \mathbb{R}^d \times C([0,T];\mathbb{R}^d) \ni (x,\gamma) \longmapsto \gamma(t) \in \mathbb{R}^d$.

Proof. See proofs of [4, Theorem 8.2.1], [2, Theorem 5.3], and [36, Theorems 4 and 5]. □

Recently, in studies for MFG [9, 50], considering μ instead of Q is called the Lagrange perspective. This perspective is summed up in the slogan "Think Eulerian, prove Lagrangian" in [60, Chapter 15], which is widely applied in, e.g. [51]. We refer the reader to [4, Section 8.2] and [36] for a more general theory.

Referring to the formulation in MFG, we rewrite the ideal Problem 1.2 in terms of Q. Before starting a more ideal problem, we introduce an evaluation map

$$\mathsf{e}_t\colon (y,\gamma) \in \mathcal{Y} \times C([0,T];\mathbb{R}^d) \;\longmapsto\; \big(\gamma(t),y\big) \in \mathbb{R}^d \times \mathcal{Y}, \quad t \in [0,T].$$

With the above, one can rewrite Problem 1.2 as follows:

Problem 5.1 (Ideal Learning Problem in the Lagrangian Framework)**.** Let $\lambda > 0$ be a constant, let $\ell\colon \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ be continuous, and let $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d \times \mathcal{Y})$ be a given input data. Set

$$\widetilde{J}(Q) := \int_{\mathcal{Y}\times C([0,T];\mathbb{R}^d)} \left(\ell\big(\gamma(1),y\big) + \int_0^T \frac{\lambda}{2}|\dot\gamma(t)|^2 \mathrm{d}t \right) \mathrm{d}Q(y,\gamma) \tag{5.5}$$

for $Q \in \mathcal{P}(\mathcal{Y} \times C([0,T];\mathbb{R}^d))$. Then, the ideal learning problem in the Lagrangian framework is posed as the following constrained minimization problem:

$$\inf \left\{ \widetilde{J}(Q) \;\middle|\; Q \in \mathcal{P}\big(\mathcal{Y} \times C([0,T];\mathbb{R}^d)\big) \text{ such that } (\mathsf{e}_0)_\# Q = \mu_0 \right\}.$$

Comparing Problems 1.2 and 5.1, the functional $\widehat{J}$ in (1.4) and $\widetilde{J}$ in (5.5) have the correspondence such that

$$\begin{aligned}
&\int_{\mathbb{R}^d\times\mathcal{Y}} \ell(x,y)\mathrm{d}\mu_T(x,y) \iff \int_{\mathcal{Y}\times C([0,T];\mathbb{R}^d)} \ell\big(\gamma(1),y\big)\mathrm{d}Q(y,\gamma),\\
&\int_0^T \int_{\mathbb{R}^d\times\mathcal{Y}} \frac{\lambda}{2}|v(x,t)|^2 \mathrm{d}\mu_t(x,y)\mathrm{d}t \iff \int_{\mathcal{Y}\times C([0,T];\mathbb{R}^d)} \int_0^T \frac{\lambda}{2}|\dot\gamma(t)|^2 \mathrm{d}t\mathrm{d}Q(y,\gamma).
\end{aligned}$$

We see that Problem 5.1 has fewer constraints and fewer variables than Problem 1.2. This is because, according to Problem 1.2, μ and v satisfying the continuity equation (1.2) can be recovered as long as Q is obtained. This fact leads us to the existence of a minimizer for Problem 5.1.

Lemma 5.1 (Existence Result for Problem 5.1)**.** *Under Assumption* 1.1*, there exists a minimizer* $Q \in \mathcal{P}(\mathcal{Y} \times C([0,T];\mathbb{R}^d))$ *for Problem* 5.1.

Proof. Set

$$S = \left\{ Q \in \mathcal{P}\big(\mathcal{Y} \times C([0,T];\mathbb{R}^d)\big) \;\middle|\; (\mathsf{e}_0)_\# Q = \mu_0 \right\},$$

here the probability measures $\mathcal{P}(\mathcal{Y} \times C([0,T];\mathbb{R}^d))$ are endowed with the narrowly convergence topology. We can easily check that a measure $(\mathsf{e}_\mathcal{Y} \times \mathsf{c}_\bullet)_\#\mu_0$ belongs to S, where we set

$$\mathsf{e}_{\mathcal{Y}}\colon \mathbb{R}^d \times \mathcal{Y} \ni (x,y) \longmapsto y \in \mathcal{Y},$$
$$\mathsf{c}\colon \mathbb{R}^d \times \mathcal{Y} \ni (x,y) \longmapsto [0,T] \ni t \longmapsto x \in \mathbb{R}^d \in C([0,T];\mathbb{R}^d).$$

It is clear that $0 \leq \widetilde{J} < +\infty$ since the integrand ℓ is non-negative. Thus, we take a minimizing sequence $(Q^n)_n \subset S$ such that $\widetilde{J}(Q^n) \to \inf J \in \mathbb{R}$ as $n \to \infty$. From the second term of (5.5), there exists a constant C independent of n such that

$$\frac{\lambda}{2} \int_{\mathcal{Y}\times C([0,T];\mathbb{R}^d)} \int_0^T |\dot{\gamma}(t)|^2 \mathrm{d}t \mathrm{d}Q^n(y,\gamma) \leq C. \tag{5.6}$$

Next, we claim that $(Q^n)_n$ is tight. We choose the maps r^1 and r^2 defined on $\mathcal{Y} \times C([0,T];\mathbb{R}^d)$ as

$$r^1\colon (y,\gamma) \longmapsto y \in \mathcal{Y}, \quad r^2\colon (y,\gamma) \longmapsto \gamma \in C([0,T];\mathbb{R}^d).$$

It is clear that $(r^1_{\#}Q^n)_n$ is tight because of Assumption 1.1 and Prokhorov's theorem. In addition, the functional

$$A\colon C([0,T];\mathbb{R}^d) \ni \gamma \longmapsto \begin{cases} \int_0^T \frac{\lambda}{2}|\dot{\gamma}(t)|^2 \mathrm{d}t, & \text{if } \gamma \text{ is an absolutely continuous} \\ & \text{curve with } |\dot{\gamma}| \in L^2(0,T) \\ & \text{and } \gamma(0) \in \operatorname{supp}\mu_0, \\ +\infty, & \text{otherwise} \end{cases} \tag{5.7}$$

has a compact sublevel sets in $C([0,T];\mathbb{R}^d)$ because of the Ascoli-Arzelá theorem. Hence, we can see that $(r^2_{\#}Q^n)_n$ is also tight thanks to an integral condition for tightness [4, Remark 5.1.5] and (5.6). Then, we obtain the tightness of $(Q^n)_n$ by applying a tightness criterion [4, Lemma 5.2.2] for the maps r^1 and r^2.

Therefore, there exists a subsequence (n), still denoted by n, and $Q^* \in \mathcal{P}(\mathcal{Y} \times C([0,T];\mathbb{R}^d))$ such that

$$Q^n \rightharpoonup Q^* \quad \text{in } \mathcal{P}\big(\mathcal{Y} \times C([0,T];\mathbb{R}^d)\big)$$

by Prokhorov's theorem. It remains to verify that the limit Q^* satisfies $(\mathsf{e}_0)_{\#}Q = \mu_0$ and

$$\widetilde{J}(Q^*) = \inf \widetilde{J}\big(= \lim_{n\to\infty} \widetilde{J}(Q^n)\big).$$

The former is obtained by the continuity of the evaluation map e_t, $t \in [0,T]$. The latter is shown as follows. By the continuity of ℓ and e_T, we obtain that

$$\begin{aligned} \lim_{n\to\infty} \int_{\mathbb{R}^d\times\mathcal{Y}} \ell \mathrm{d}(\mathsf{e}_T)_{\#}Q^n &= \lim_{n\to\infty} \int_{\mathbb{R}^d\times\mathcal{Y}} \min \ell \circ \mathsf{e}_T, C' \mathrm{d}Q^n \\ &= \int_{\mathbb{R}^d\times\mathcal{Y}} \min \ell \circ \mathsf{e}_T, C' \mathrm{d}Q^* = \int_{\mathbb{R}^d\times\mathcal{Y}} \ell \mathrm{d}(\mathsf{e}_T)_{\#}Q^* \end{aligned}$$

for a sufficiently large constant $C' > 0$. In addition, the functional A in (5.7) is lower semicontinuous, and we can choose $A^k \in C_b(C([0,T];\mathbb{R}^d)), k = 1, 2, \ldots$, such that $A^k \nearrow A$ as $k \to \infty$ by [3, Theorem 10.2]. Then, we get that for each $k \in \mathbb{N}$,

$$\liminf_{n\to\infty} \int_{C([0,T];\mathbb{R}^d)} A \mathrm{d}Q^n \geq \liminf_{n\to\infty} \int_{C([0,T];\mathbb{R}^d)} A^k \mathrm{d}Q^n = \int_{C([0,T];\mathbb{R}^d)} A^k \mathrm{d}Q^*.$$

Hence, passing to the limit as $k \to \infty$ in the above inequality, we obtain

$$\liminf_{n\to\infty} \int_{C([0,T];\mathbb{R}^d)} A \mathrm{d}Q^n \geq \lim_{k\to\infty} \int_{C([0,T];\mathbb{R}^d)} A^k \mathrm{d}Q^* = \int_{C([0,T];\mathbb{R}^d)} A \mathrm{d}Q^*$$

by virtue of Fatou's lemma. □

From Lemma 5.1 and Proposition 5.1 we immediately obtain Theorem 1.2.

Proof of Theorem 1.2. Let Q^* denote the minimizer of $\widetilde{J}$. From Proposition 5.1, we can get $\mu^{Q^*} \in C([0,T];\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y}))$ satisfying (5.4) and $\tilde{v} \in L^2(\mathrm{d}\mu_t^{Q^*}\mathrm{d}t)$. By [36, Theorem 5], we have $\widetilde{J}(Q^*) = \widehat{J}(\mu^{Q^*}, \tilde{v})$. From this equality and Proposition 5.1, it follows that

$$\widehat{J}(\mu^{Q^*}, \tilde{v}) \leq \widehat{J}(\mu, v), \quad \forall\, \mu \in C\big([0,T];\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y})\big), \quad v \in L^2(\mathrm{d}\mu_t \mathrm{d}t).$$

The proof is complete. □

6 Conclusion

In this paper, we introduced the kinetic regularized learning problem (Problem 1.1) and proved the existence of its minimizer in Theorem 1.1. A key idea in the proof is to show that a sequence of curves $(\mu^n) \subset C([0,T];(\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y}), W_2))$, rather than a parameter $(\theta^n) \subset L^2(0,T;\mathbb{R}^m)$, converges strongly. Furthermore, we attempted to idealize Problem 1.1 as Problem 1.2, although the relationship between this idealization and the existing neural network is unclear. However, considering the minimizers of Problem 1.2 will provide essential clues for understanding deep learning in the future.

Our results can be further developed through a generalization of neural networks and regularization terms. The directions of each generalization are described below and will be subjects of future work.

6.1 For general neural network architectures

It remains to establish an existence result for neural networks more general than Assumption 1.2. A general l-layer neural network v is a continuous vector field satisfying the following assumptions.

Assumption 6.1 (General l-Layer Neural Network)**.** There exists $C > 0$, it holds that

$$|v(x,\theta)| \leq C|\theta|^l(1+|x|), \quad x \in \mathbb{R}^d, \tag{6.1}$$

$$|v(x_1,\theta)-v(x_2,\theta)| \le C|\theta|^l|x_1-x_2|, \quad x_1,x_2\in\mathbb{R}^d \tag{6.2}$$

for $\theta\in\mathbb{R}^m$.

Note that we also assume that $v\colon \mathbb{R}^d\times\mathbb{R}^m\to\mathbb{R}^d$ is continuous in Problem 1.1.

For example, v satisfying Assumption 1.2 is a 1-layer neural network. In practice, 2, 3-layer neural network is often used. We mentioned in Section 1.2 that the nonlinearity of such l-layer neural networks hinders the proof of existence theorems, especially Claim 4.1. To relax this nonlinearity, it may be effective to consider a mean-field neural network

$$\mathcal{V}(x,\vartheta)=\int_{\mathbb{R}^m} v(x,\theta)\mathrm{d}\vartheta(\theta), \tag{6.3}$$

where ϑ is a learnable probability measure on $\mathbb{R}^m$. This assumption has long been known as the Young measure [15,45] in optimal control theory, but it has recently been recognized again as a helpful approach to shallow neural networks [1,17,40] and ODE-Nets [20,31,38]. The author is in the process of conducting further theoretical research using this network $\mathcal{V}$.

Appendix A H^1-Regularization

We discuss the existence of a minimizer in the same problem setting as [57].

Problem A.1 (H^1-Regularized Learning Problem)**.** Let $\lambda>0$ be a constant, let $\mathcal{Y}$ be a subset of $\mathbb{R}^d$ and let $v\colon\mathbb{R}^d\times\mathbb{R}^m\to\mathbb{R}^d$ and $\ell\colon\mathbb{R}^d\times\mathcal{Y}\to\mathbb{R}_+$ be continuous. Let $\mu_0\in\mathcal{P}_c(\mathbb{R}^d\times\mathcal{Y})$ a given input data. Set

$$J_{H^1}(\mu,\theta):=\int_{\mathbb{R}^d\times\mathcal{Y}}\ell\mathrm{d}\mu_T+\frac{\lambda}{2}\|\theta\|^2_{H^1(0,T;\mathbb{R}^m)} \tag{A.1}$$

for $\mu\in C([0,T];(\mathcal{P}_2(\mathbb{R}^d\times\mathcal{Y}),W_2))$ and $\theta\in H^1(0,T;\mathbb{R}^m)$. The H^1-regularized learning problem constrained by ODE-Net is posed as the following constrained minimization problem:

$$\begin{aligned}&\inf\left\{J_{H^1}(\mu,\theta)\;\middle|\;\mu\in C\big([0,T];\big(\mathcal{P}_2(\mathbb{R}^d\times\mathcal{Y}),W_2\big)\big),\ \theta\in H^1(0,T;\mathbb{R}^m)\right\}\\&\text{subject to}\\&\begin{cases}\partial_t\mu_t+\mathrm{div}_x\left(v(\bullet,\theta_t)\mu_t\right)=0,\\ \mu_t|_{t=0}=\mu_0.\end{cases}\end{aligned} \tag{A.2}$$

We note that the constraint (A.2) is the same as (1.2).

For Problem A.1, we can obtain an existence result without Assumption 1.2.

Theorem A.1 (Existence Theorem for Problem A.1)**.** *Under Assumptions* 1.1 *and* 6.1*, there exists a minimizer* $(\mu,\theta)\in C([0,T];(\mathcal{P}_2(\mathbb{R}^d\times\mathcal{Y}),W_2))\times H^1(0,T;\mathbb{R}^m)$ *of* (A.1) *in Problem* A.1*.*

Before the proof of Theorem A.1, we prepare a lemma similar to Lemma 4.2.

Lemma A.1. *Let $\theta \in H^1(0,T;\mathbb{R}^m)$ and let $\mu \in C([0,T];(\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y}), W_2))$ be a distributional solution of* (A.2) *corresponding to a vector fields $(v(\bullet,\theta_t))_t$. If Assumption* 6.1 *holds, there exists a radius $R^* = R^*(\mu_0, f, T, \|\theta\|_{H^1(0,T;\mathbb{R}^m)}) > 0$ such that*

$$\operatorname{supp} \mu_t \subset B_{\mathbb{R}^d \times \mathcal{Y}}(R^*), \quad \forall\, t \in [0,T].$$

Proof. The strategy of the proof is the same as that of the proof of Lemma 4.2. By Assumption 6.1 and Sobolev inequality, we have

$$\begin{aligned}
&\int_0^T \int_{\mathbb{R}^d \times \mathcal{Y}} |v(x,\theta_t)| \mathrm{d}\mu_t \mathrm{d}t \\
&\leq C \int_0^T \int_{\mathbb{R}^d \times \mathcal{Y}} |\theta_t|^l (1+|x|) \mathrm{d}\mu_t \mathrm{d}t \\
&\leq C \|\theta\|^l_{L^l(0,T;\mathbb{R}^m)} \left(1 + \sup_{t\in[0,T]} \int_{\mathbb{R}^d \times \mathcal{Y}} |x| \mathrm{d}\mu_t \right) \\
&\leq C \|\theta\|^l_{H^1(0,T;\mathbb{R}^m)} \left(1 + \sup_{t\in[0,T]} \sqrt{\int_{\mathbb{R}^d \times \mathcal{Y}} |x|^2 \mathrm{d}\mu_t} \right) \\
&= C \|\theta\|^l_{H^1(0,T;\mathbb{R}^m)} \left(1 + \sup_{t\in[0,T]} W_2(\mu_t, \delta) \right) < \infty.
\end{aligned}$$

Also, again using Assumption 6.1 and Sobolev inequality, we can estimate $|X_t(x)|$ in the proof of Lemma 4.2 by $\|\theta\|_{H^1(0,T;\mathbb{R}^m)}$. □

Proof of Theorem A.1. Set

$$S = \left\{ (\mu,\theta) \in C\big([0,T];\big(\mathcal{P}_2(\mathbb{R}^d \times \mathcal{Y}), W_2\big)\big) \times H^1(0,T;\mathbb{R}^m) \;\middle|\; (\mu,\theta) \text{ satisfies (A.2)} \right\}.$$

It is obvious that $S \neq \emptyset$ and $0 \leq J_{H^1} < \infty$ on S. Then, we can take a minimizing sequence $((\mu^n,\theta^n))_{n=1}^\infty \subset S$ such that $J_{H^1}(\mu^n,\theta^n) \to \inf_S J_{H^1}$ as $n \to \infty$. By the second term of (A.1), there exists a constant $C > 0$ such that

$$\frac{\lambda}{2} \|\theta^n\|^2_{H^1} \leq C \tag{A.3}$$

for all $n \in \mathbb{N}$. From Lemma 4.1, Assumption 6.1, (A.3) and the Sobolev inequality, we have for $0 \leq t < s \leq T$,

$$\begin{aligned}
W_2\big(\mu^n_t, \mu^n_s\big)^2 &\leq (s-t) \int_t^s \int_{\mathbb{R}^d} \big|v\big(x,\theta^n_\tau\big)\big|^2 \mathrm{d}\mu^n_\tau(x) \mathrm{d}\tau \\
&\leq C(s-t) \int_t^s \int_{\mathbb{R}^d} |\theta_\tau|^{2l} (1+|x|^2) \mathrm{d}\mu^n_\tau(x) \mathrm{d}\tau \\
&\leq C(s-t) \int_t^s \int_{\mathbb{R}^d} |\theta_\tau|^{2l} \big(1+R^{*2}\big) \mathrm{d}\mu^n_\tau(x) \mathrm{d}\tau \\
&\leq C \|\theta\|^{2l}_{L^{2l}(0,T;\mathbb{R}^m)} (s-t)
\end{aligned}$$

$$\begin{aligned}&\leq C\|\theta\|^{2l}_{H^1(0,T;\mathbb{R}^m)}(s-t)\\&\leq C(s-t),\end{aligned}$$

where $R^* > 0$ is the constant appeared in Lemma A.1. Hence, there exist a subsequence

$$\begin{aligned}&(n') := \big(n(k)\big)_{k=1}^{\infty} \subset \mathbb{Z}_{>0},\\&(\mu^*, \theta^*) \in C\big(0, T; \big(\mathcal{P}(\mathbb{R}^d \times \mathcal{Y}), W_2\big)\big) \times H^1(0, T; \mathbb{R}^m)\end{aligned}$$

such that

$$\theta^{n'} \to \theta^* \quad \text{weakly in} \quad H^1(0,T;\mathbb{R}^m), \tag{A.4}$$

$$\theta^{n'} \to \theta^* \quad \text{strongly in} \quad C(0,T;\mathbb{R}^m), \tag{A.5}$$

$$\mu^{n'} \to \mu^* \quad \text{strongly in} \quad C\big(0,T;\big(\mathcal{P}(\mathbb{R}^d \times \mathcal{Y}), W_2\big)\big). \tag{A.6}$$

Here, we used the Sobolev embedding theorem in (A.5). By the above, we can deduce the following claim.

Claim A.1. The limits μ^* and θ^* satisfy (4.6) for all $\zeta \in C_c^\infty((0,T) \times \mathbb{R}^d \times \mathcal{Y})$.

Proof. As in the proof of Claim 4.1, the proof is completed by taking the limits of I_1 to I_4 in (4.8). From the proof of Claim 4.1, $I_1, I_2 \to 0$ as $n \to \infty$. Also, $I_3 \to 0$ as $n \to \infty$ by the uniform convergence (A.5), and the continuity of $v(x,\theta)$ with respect to θ. For I_4, it follows from (6.2), (A.4) and (A.6) that

$$\begin{aligned}|I_4| &\leq \int_0^T \mathrm{Lip}\left(\nabla_x \zeta_t \cdot \left(v\big(\bullet, \theta_t^n\big) - v\big(\bullet, \theta_t^*\big)\right)\right) W_1\big(\mu_t^n, \mu_t^*\big)\mathrm{d}t\\&\leq C\Big(\|\theta^n\|^l_{L^l(0,T;\mathbb{R}^m)} + \|\theta^*\|^l_{L^l(0,T;\mathbb{R}^m)}\Big) \sup_{t\in[0,T]} W_1\big(\mu_t^n, \mu_t^*\big)\\&\leq C\Big(\|\theta^n\|^l_{H^1(0,T;\mathbb{R}^m)} + \|\theta^*\|^l_{H^1(0,T;\mathbb{R}^m)}\Big) \sup_{t\in[0,T]} W_2\big(\mu_t^n, \mu_t^*\big)\\&\leq C \sup_{t\in[0,T]} W_2\big(\mu_t^n, \mu_t^*\big) \to 0 \quad \text{as} \quad n \to \infty.\end{aligned}$$

Thus we obtain the conclusion. □

We resume the proof of Theorem A.1. From Claim A.1, we have $(\mu^*, \theta^*) \in S$. In addition, J_{H^1} is lower semicontinuous from (4.9) and the weak lower semi-continuity of the H^1-norm $\| \bullet \|_{H^1(0,T;\mathbb{R}^m)}$. The proof is complete. □

Appendix B Convexity assumptions

For comparison, using the proof technique by Bonnet *et al.* [10], we show that a unique minimizer to Problem 1.1 exists. This proof technique makes use of the idea that we can regard the functional J as a univariate functional $\widetilde{J}(\theta) := J(\mu^\theta, \theta)$, where $\mu^\theta \in C([0,T];$

$\mathcal{P}(\mathbb{R}^d \times \mathcal{Y}))$ is a solution of (1.2) for a given $\theta \in L^2(0,T;\mathbb{R}^m)$. The existence and uniqueness of the solution can be proved by showing the convexity of $\widetilde{J}$. For this purpose, we evaluate the Lipschitz constant of the Fréchet derivative $\nabla_\theta \widetilde{J}$ of $\widetilde{J}$.

First, recall from Lemma 3.5 that μ^θ is represented as $\mu_t^\theta = (\Phi^\theta(0,t;\bullet) \times \mathrm{Id}_{\mathcal{Y}})_{\#}\mu_0$ using a flow map $\Phi^\theta \colon [0,T]^2 \times \mathbb{R}^d \to \mathbb{R}^d$, $\theta \in L^2(0,T;\mathbb{R}^m)$ according to the ODE

$$\begin{cases} \partial_t \Phi^\theta(t_0,t;x) = v\big(\Phi^\theta(t_0,t;x),\theta_t\big), \\ \Phi^\theta(t_0,t_0;x) = x. \end{cases} \tag{B.1}$$

We note here that from (1.3) it can be verified that the Lipschitz continuity assumption (3.4) is satisfied, as in the proof of Lemma 4.2. The derivative of Φ with respect to θ can be described by a linearization of (B.1).

Lemma B.1 (Taylor Expansion of Φ^θ)**.** *Suppose that the neural network v satisfies Assumption* 1.2 *and $f \colon \mathbb{R}^d \to \mathbb{R}^p$ is differentiable. Then, for every $\theta, \vartheta \in L^2(0,T;\mathbb{R}^m)$, the Taylor expansion*

$$\Phi^{\theta+\epsilon\vartheta}(t_0,t;x) = \Phi^\theta(t_0,t;x) + \epsilon \int_0^t \Delta^\theta_{(s,t)}(x)\vartheta_s f\big(\Phi^\theta(t,s;x)\big)\mathrm{d}s + o(\epsilon)$$

holds in $C([0,t] \times \operatorname{supp}\mu_0;\mathbb{R}^d \times \mathcal{Y})$, where, for $(t_0,x) \in [0,T] \times \mathbb{R}^d$, the map $[0,T] \ni t \longmapsto \Delta^\theta_{(t_0,t)}(\bullet) \in C(\mathbb{R}^d;\mathbb{R}^{d\times d})$ is the unique solution of the linearized Cauchy problem

$$\begin{cases} \partial_t \Delta^\theta_{(t_0,t)}(x) = \theta_t J f\big(\Phi^\theta(t_0,t;x)\big)\Delta^\theta_{(t_0,t)}(x), \\ \Delta^\theta_{(t_0,t_0)}(x) = \mathrm{Id}_{\mathbb{R}^d}, \end{cases} \tag{B.2}$$

where $Jf \colon \mathbb{R}^d \to \mathbb{R}^p$ denotes the Jacobian matrix of f.

Proof. See [13, Theorem 3.2.6]. □

To evaluate the Lipscitz continuity of $\nabla_\theta \widetilde{J}$, we need to estimate the variation of $\Delta^\theta_{(0,t)}$ with respect to θ. This evaluation requires us to impose a further assumption on $v(x,\theta) = \theta f(x)$ in addition to Assumption 1.2. In the following, we will denote $R^*(\|\theta\|)$ the same radius as in Lemma 4.2.

Assumption B.1 (Strong Smoothness on v)**.** The function f is twice continuously differentiable, and for given $\theta \in L^2(0,T;\mathbb{R}^m)$ and $(x,y) \in B(R^*(\|\theta\|))$, f satisfies $\|f\|_{C^1(\mathbb{R}^p;\mathbb{R}^d)} < \infty$.

This assumption correspond to [10, Assumption 2]. Under this assumption, we can show the Lipschitz continuity by the same argument as in the proof of [10, Lemma 3.1].

Lemma B.2 (Fréchet-Differentiablity of the Loss Functional)**.** *The sum of loss and kinetic regularization*

$$J_\ell \colon \theta \longmapsto \int_{\mathbb{R}^d \times \mathcal{Y}} \ell \mathrm{d}\mu_T^\theta + \frac{\lambda}{2}\int_0^T \int_{\mathbb{R}^d} |v(x,\theta)|^2 \mathrm{d}\mu_t^\theta \mathrm{d}t$$

is Fréchet-differentiable. In addition, for $\theta^1, \theta^2 \in L^2(0,T;\mathbb{R}^m)$, there exists $C(\lambda, \|\theta^1\|, \|\theta^2\|) > 0$ such that

$$\big\|\nabla J_\ell(\theta^1) - \nabla J_\ell(\theta^2)\big\| \le C(\lambda, \|\theta^1\|, \|\theta^2\|)\|\theta^1 - \theta^2\|.$$

From Lemma B.2, the following corollary follows immediately.

Corollary B.1 (Semiconvexity for the Parameter θ)**.** *The functional*

$$\widetilde{J}\colon \theta \longmapsto J(\mu^\theta,\theta)=\int_{\mathbb{R}^d\times\mathcal{Y}}\ell\mathrm{d}\mu_T^\theta+\int_0^T\int_{\mathbb{R}^d}\left(\frac{\lambda}{2}|v(x,\theta_t)|^2+\frac{\epsilon}{2}|\theta_t|^2\right)\mathrm{d}\mu_t^\theta(x)\mathrm{d}t \tag{B.3}$$

satisfies

$$\begin{aligned}\widetilde{J}\big((1-\zeta)\theta^1+\zeta\theta^2\big)\leq{}&(1-\zeta)\widetilde{J}(\theta^1)+\zeta\widetilde{J}(\theta^2)\\&-\big(\epsilon-C(\lambda,\|\theta^1\|,\|\theta^2\|)\big)\frac{\zeta(1-\zeta)}{2}\|\theta^1-\theta^2\|^2\end{aligned}$$

for any $\theta^1,\theta^2\in L^2(0,T;\mathbb{R}^m)$ and $\zeta\in[0,1]$. Here $C(\lambda,\|\theta^1\|,\|\theta^2\|)$ is the same positive number as in Lemma B.2*.*

By this corollary, $\widetilde{J}$ is strongly convex on a L^2 ball if ϵ is sufficiently large compared to other parameters such as λ and T. This plays an essential role in the proof of Theorem B.1 below.

Theorem B.1 (Existence and Uniqueness of Minimizer of $\widetilde{J}$)**.** *Suppose that Assumptions* 1.2 *and* B.1 *If $\epsilon>0$ in* (B.3) *is sufficiently large, there exists $\theta\in L^2(0,T;\mathbb{R}^m)$ which minimize $\widetilde{J}$, and θ is a unique minimizer of $\widetilde{J}$.*

The proof is carried out using the direct method of calculus of variations as in Theorem 1.1.

Proof. It is clear that $0\leq\inf\widetilde{J}<+\infty$, then we can take a minimizing sequence $(\theta^n)_{n=1}^\infty\subset L^2(0,T;\mathbb{R}^m)$ such that $\widetilde{J}(\theta^n)\to\inf\widetilde{J}$ as $n\to\infty$. Thus, there exists $C>0$ independent of n such that (4.3) holds. Then there exists a subsequence $(n')\subset\mathbb{Z}_{>0}$ such that (4.5). In addition, by Corollary B.1, we see that there exists $\epsilon>0$ such that $\widetilde{J}$ is convex on $B(2C/\epsilon)$. Therefore, by Mazur's lemma, there exists another minimizing sequence $(\hat{\theta}^n)_{n=1}^\infty$ such that $\hat{\theta}_n\to\theta$ in $L^2(0,T;\mathbb{R}^m)$. Because $\widetilde{J}$ is lower semicontinuous, we conclude that

$$J(\theta)\leq\liminf_{n\to\infty}J(\hat{\theta}_n)=\inf_{L^2(0,T;\mathbb{R}^m)}\widetilde{J},$$

i.e. θ is a minimizer of $\widetilde{J}$. The uniqueness is immediately obtained from the strong convexity of $\widetilde{J}$. □

Acknowledgement

The author, N. Isobe, would like to thank his supervisor, Norikazu Saito, for his encouragement and advice during the preparation of the paper.

References

[1] S. Akiyama and T. Suzuki, On learnability via gradient method for two-layer ReLU neural networks in teacher-student setting, in: *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Vol. 139, 152–162, 2021.

[2] L. Ambrosio, Transport equation and Cauchy problem for BV vector fields, *Invent. Math.*, 158(2):227–260, 2004.

[3] L. Ambrosio, E. Brué, and D. Semola, *Lectures on Optimal Transport*, Springer, 2021.

[4] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*, in: *Lectures in Mathematics. ETH Zürich*, Birkhäuser Verlag, 2008.

[5] V. I. Arnol'd, On functions of three variables, *Dokl. Akad. Nauk SSSR*, 114:679–681, 1957.

[6] G. Baravdish, G. Eilertsen, R. Jaroudi, B. T. Johansson, L. Mal, and J. Unger, Learning via nonlinear conjugate gradients and depth-varying neural ODEs, *arXiv:2202.05766v1*, 2022.

[7] R. Barboni, G. Peyr, and F.-X. Vialard, Global convergence of ResNets: From finite to infinite width using linear parameterization, *Adv. Neural Inf. Process. Syst.*, 35:16385–16397, 2021.

[8] J.-D. Benamou and Y. Brenier, A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem, *Numer. Math.*, 84(3):375–393, 2000.

[9] J.-D. Benamou, G. Carlier, and F. Santambrogio, *Variational Mean Field Games*, in: *Active Particles*, Birkhäuser, Vol. 1, 141–171, 2017.

[10] B. Bonnet, C. Cipriani, M. Fornasier, and H. Huang, A measure theoretical approach to the mean-field maximum principle for training neurODEs, *Nonlinear Anal.*, 227:113161, 2022.

[11] B. Bonnet and H. Frankowska, Differential inclusions in Wasserstein spaces: The Cauchy-Lipschitz framework, *J. Differential Equations*, 271:594–637, 2021.

[12] B. Bonnet and H. Frankowska, On the properties of the value function associated to a mean-field optimal control problem of Bolza type, in: *60th IEEE Conference on Decision and Control*, IEEE Press, 4558–4563, 2021.

[13] A. Bressan and B. Piccoli, *Introduction to the Mathematical Theory of Control*, American Institute of Mathematical Sciences, 2007.

[14] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, 2011.

[15] C. Castaing, P. R. de Fitte, and M. Valadier, *Young Measures on Topological Spaces: With Applications in Control Theory and Probability Theory*, Springer, 2004.

[16] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, Neural ordinary differential equations, *NeurIPS*, 6572–6583, 2018.

[17] L. Chizat and F. Bach, On the global convergence of gradient descent for over-parameterized models using optimal transport, *Adv. Neural Inf. Process. Syst.*, Vol. 31, 2018.

[18] G. Cybenko, Approximation by superpositions of a sigmoidal function, *Math. Control Signals Systems*, 2(4):303–314, 1989.

[19] Z. Ding, S. Chen, Q. Li, and S. Wright, On the global convergence of gradient descent for multi-layer ResNets in the mean-field regime, *J. Mach. Learn. Res.*, 23(48):1–65, 2021.

[20] Z. Ding, S. Chen, Q. Li, and S. J. Wright, Overparameterization of deep ResNet: Zero loss and mean-field analysis, *J. Mach. Learn. Res.*, 23(48):1–65, 2022.

[21] R. M. Dudley, *Real Analysis and Probability*, Cambridge University Press, 2002.

[22] W. E, A proposal on machine learning via dynamical systems, *Commun. Math. Stat.*, 5(1):1–11, 2017.

[23] W. E, J. Han, and Q. Li, A mean-field optimal control formulation of deep learning, *Res. Math. Sci.*, 6(1):10, 2019.

[24] C. Esteve, B. Geshkovski, D. Pighin, and E. Zuazua, Large-time asymptotics in deep learning, *arXiv: 2008.02491*, 2021.

[25] C. Finlay, J.-H. Jacobsen, L. Nurbekyan, and A. Oberman, How to train your neural ODE: The world of Jacobian and kinetic regularization, in: *Proceedings of the 37th International Conference on Machine Learning*, PMLR, Vol. 119, 3154–3164, 2020.

[26] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, Reinforcement learning with deep energy-based policies, in: *Proceedings of the 34th International Conference on Machine Learning*, PMLR, Vol. 70, 1352–1361, 2017.

[27] E. Haber and L. Ruthotto, Stable architectures for deep neural networks, *Inverse Problems*, 34(1):014004, 2017.
[28] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 770–778, 2016.
[29] M. Herty, A. Thuenen, T. Trimborn, and G. Visconti, Continuous limits of residual neural networks in case of large input data, *Commun. Appl. Ind. Math.*, 13(1):96–120, 2022.
[30] K. Hornik, Approximation capabilities of multilayer feedforward networks, *Neural Netw.*, 4(2):251–257, 1991.
[31] J.-F. Jabir, D. Šiška, and L. Szpruch, Mean-field neural ODEs via relaxed optimal control, *arXiv: 1912.05475v3*, 2021.
[32] J. L. Kelley, *General Topology*, Springer, 1975.
[33] A. N. Kolmogorov, On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition, *Dokl. Akad. Nauk SSSR*, 114:953–956, 1957.
[34] J.-M. Lasry and P.-L. Lions, Jeux à champ moyen. II – Horizon fini et contrôle optimal, *Comptes Rendus Math.*, 343(10):679–684, 2006.
[35] J.-M. Lasry and P.-L. Lions, Mean field games, *Jpn. J. Math.*, 2(1):229–260, 2007.
[36] S. Lisini, Characterization of absolutely continuous curves in Wasserstein spaces, *Calc. Var. Partial Differential Equations*, 28(1):85–120, 2007.
[37] E. Lorin, Derivation and analysis of parallel-in-time neural ordinary differential equations, *Ann. Math. Artif.*, 88(10):1035–1059, 2020.
[38] Y. Lu, C. Ma, Y. Lu, J. Lu, and L. Ying, A mean field analysis of deep ResNet and beyond: Towards provably optimization via overparameterization from depth, in: *Proceedings of the 37th International Conference on Machine Learning*, PMLR, Vol. 119, 6426–6436, 2020.
[39] S. Massaroli, M. Poli, J. Park, A. Yamashita, and H. Asama, Dissecting neural ODEs, *Adv. Neural Inf. Process. Syst.*, 33:3952–3963, 2020.
[40] S. Mei, A. Montanari, and P.-M. Nguyen, A mean field view of the landscape of two-layer neural networks, *Proc. Natl. Acad. Sci. USA*, 115(33):E7665–E7671, 2018.
[41] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2013.
[42] V. Nair and G. E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: *Proceedings of the 27th International Conference on Machine Learning*, PMLR, 807–814, 2010.
[43] C. Orrieri, A. Porretta, and G. Savar, A variational approach to the mean field planning problem, *J. Funct. Anal.*, 277(6):1868–1957, 2019.
[44] A. Pinkus, Approximation theory of the MLP model in neural networks, *Acta Numer.*, 8:143–195, 1999.
[45] N. Pogodaev, Optimal control of continuity equations, *NoDEA Nonlinear Differential Equations Appl.*, 23(2):21, 2016.
[46] A. Queiruga, N. B. Erichson, L. Hodgkinson, and M. W. Mahoney, Stateful ODE-Nets using basis function expansions, *Adv. Neural Inf. Process. Syst.*, 34:21770–21781, 2021.
[47] L. Ruthotto, S. J. Osher, W. Li, L. Nurbekyan, and S. W. Fung, A machine learning framework for solving high-dimensional mean field game and mean field control problems, *Proc. Natl. Acad. Sci. USA*, 117(17):9183–9193, 2020.
[48] M. E. Sander, P. Ablin, M. Blondel, and G. Peyré, Momentum residual neural networks, in: *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Vol. 139, 9276–9287, 2021.
[49] M. E. Sander, P. Ablin, and G. Peyré, Do residual neural networks discretize neural ordinary differential equations?, *Adv. Neural Inf. Process. Syst.*, 35:36520–36532, 2022.
[50] F. Santambrogio, *Lecture Notes on Variational Mean Field Games*, in: *Mean Field Games. Lecture Notes in Mathematics*, Vol. 2281, Springer, 2020.
[51] C. Sarrazin, Lagrangian discretization of variational mean field games, *SIAM J. Control Optim.*, 60(3):1365–1392, 2022.
[52] A. Scagliotti, *Ensembles of Affine-Control Systems with Applications to Deep Learning*, PhD Thesis, Scuola Internazionale Superiore di Studi Avanzati, 2022.
[53] A. Scagliotti, Optimal control of ensembles of dynamical systems, *ESAIM: COCV*, 29:22, 2023.
[54] S. Sonoda and N. Murata, Double continuum limit of deep neural networks, in: *ICML Workshop Princi-*

pled Approaches to Deep Learning, Vol. 1740, 2017.

[55] D. A. Sprecher, On the structure of continuous functions of several variables, *Trans. Amer. Math. Soc.*, 115:340–355, 1965.

[56] T. Teshima, K. Tojo, M. Ikeda, I. Ishikawa, and K. Oono, Universal approximation property of neural ordinary differential equations, *arXiv:2012.02414*, 2020.

[57] M. Thorpe and Y. van Gennip, Deep limits of residual neural networks, *Res. Math. Sci.*, 10(1):6, 2023.

[58] F.-X. Vialard, R. Kwitt, S. Wei, and M. Niethammer, A shooting formulation of deep learning, *Adv. Neural Inf. Process. Syst.*, Vol. 33, 11828–11838, 2020.

[59] C. Villani, *Topics in Optimal Transportation (Graduate Studies in Mathematics)*, AMS, 2003.

[60] C. Villani, *Optimal Transport: Old and New*, in: *A Series of Comprehensive Studies in Mathematics*, Springer, 2009.

[61] G. Yang, D. Yu, C. Zhu, and S. Hayou, Tensor programs VI: Feature learning in infinite-depth neural networks, in: *The Twelfth International Conference on Learning Representations*, ICLR, 2024.

[62] Y. D. Zhong, B. Dey, and A. Chakraborty, Symplectic ODE-Net: Learning hamiltonian dynamics with control, in: *International Conference on Learning Representations*, ICLR, 2020.