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.