FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

FACULTY OF SCIENCES

CHAIR IN APPLIED ANALYSIS FAU

MACHINE LEARNING VIA CONTROL OF NEURAL ODES

Borjan Geshkovski Universidad Autónoma de Madrid
Chair of Computational Mathematics, Fundación Deusto

Enrique Zuazua Applied Analysis - AvH Professorship
FAU Erlangen-Nürnberg

Introduction

A key paradigm of deep learning is that of *supervised learning*, which addresses the problem of predicting from labeled data, consisting in approximating an unknown function $f : \mathcal{X} \to \mathcal{Y}$ from N known but possibly noisy data samples $\{\vec{x}_i, \vec{y}_i\}_{i=1}^N$ with $\vec{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ and $\vec{y}_i \in \mathcal{Y}$. We shall mostly concentrate on *classification tasks*, wherein $\mathcal{Y} = \{1, \dots, m\}$.

The workhorse behind the recent successes of deep learning are models called *neural networks* for approximating f_{approx} of the unknown function f; these are parametrized computational architectures which propagate each individual sample $\vec{x}_i$ of the input data across a sequence of affine parametric operators composed with simple nonlinearities.

Residual Neural Networks (ResNets)

In practice, one looks to use models wherein the compositions of nonlinearities and affine parametric operators are iterated over multiple layers, namely *deep neural networks*. A staple of such models are the so-called *residual neural networks* (ResNets) which may often be cast as schemes of the mould

$$\begin{cases} \mathbf{x}_i^{k+1} = \mathbf{x}_i^k + w_1^k \sigma(w_2^k \mathbf{x}_i^k + b^k) & \text{for } k \in \{0, \dots, N_{\text{layers}} - 1\} \\ \mathbf{x}_i^0 = \vec{x}_i \end{cases} \tag{1}$$

for all $i \in \{1, \dots, N\}$, $w_1^k, w_2^k \in \mathbb{R}^{d \times d}$ and $N_{\text{layers}} \geqslant 1$ designates the number of layers referred to as the *depth*.

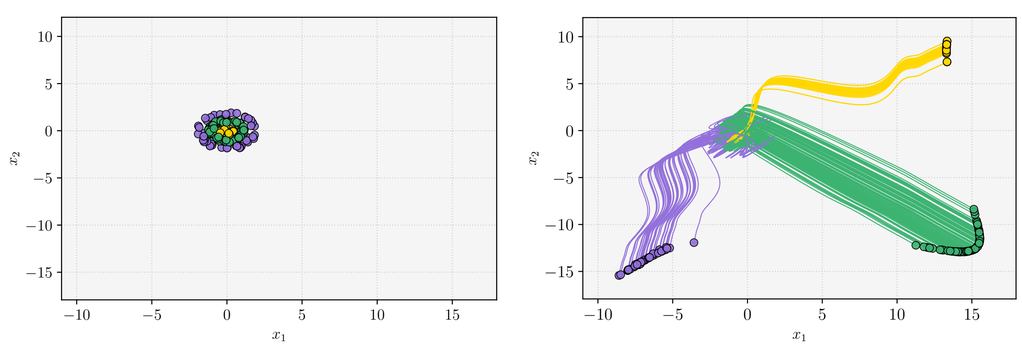
Supervised learning via control of Neural ODEs

Due to the inherent dynamical nature of ResNets, several recent works have considered an associated continuous-time formulation. This is motivated by the simple observation that for $T > 0$, (1) is the forward Euler approximation of the neural ordinary differential equation (neural ODE)

$$\begin{cases} \dot{\mathbf{x}}_i(t) = w_1(t)\sigma(w_2(t)\mathbf{x}_i(t) + b(t)) & \text{for } t \in (0, T) \\ \mathbf{x}_i(0) = \vec{x}_i \in \mathbb{R}^d. \end{cases} \tag{2}$$

One readily sees that the parameters w_2, w_1, b in the neural ODE play the role of *controls*, and thus, the supervised learning problem may be seen as a compound and high-dimensional simultaneous control problem.

The nonlinear nature of the activation function σ allows deforming half of the phase space while the other half remains invariant, a property that classical models in mechanics do not fulfill. This very property allows to build elementary controls inducing specific dynamics and transformations whose concatenation, along with properly chosen hyperplanes, allows achieving our goals in finitely many steps ([4]). This allows the neural ODE flow to efficiently separate the dataset into respective classes, as seen below.



The turnpike property

In [1], we propose the training problem consisting in minimizing

$$\frac{1}{N}\sum_{i=1}^N \text{loss}\left(P\mathbf{x}_i(T), \vec{y}_i\right) + \frac{1}{N}\int_0^T \|\mathbf{x}_i(t) - \overline{\mathbf{x}}_i\|^2 dt + \|u\|^2_{L^2(0,T;\mathbb{R}^{d_u})}, \tag{3}$$

where $\text{loss}(\cdot, \cdot)$ is continuous function which, in classification tasks ($\vec{y}_i \in \{-1, 1\}$), is usually $\text{loss}(x, y) := \|\tanh(x) - y\|^2$ or $\text{loss}(x, y) = \log(1 + \exp(-yx))$, and $\overline{\mathbf{x}}_i \in P^{-1}(\{\vec{y}_i\})$.

As each time-step of a discretization to (2) may be seen to represent a different layer of the ResNet (1), the time horizon $T > 0$ in (2) may serve as an indicator of the number of layers N_{layers} in the discrete-time context (1). A good understanding of the dynamics of the learning problem over longer time horizons would lead to potential rules for choosing the number of layers, and enlighten the possible generalization properties when the number of layers is large.

In [1,2] (see [3] for the L^1–regularization case), under controllability assumptions on the neural ODE (which are addressed in [4]), but without any smallness assumptions on the data, targets, or smoothness assumptions on the dynamics (we only assume $\sigma \in \text{Lip}(\mathbb{R})$), we conclude that the optimal controls $u_T = [w_{1,T}, w_{2,T}, b_T]$ and associated optimal trajectories $\mathbf{x}_T(\cdot)$ satisfy

$$\frac{1}{N}\sum_{i=1}^N \text{loss}\left(P\mathbf{x}_{T,i}(t), \vec{y}_i\right) + \frac{1}{N}\|\mathbf{x}_{T,i}(t) - \overline{\mathbf{x}}_i\| \leqslant C\, e^{-\mu t} \tag{4}$$

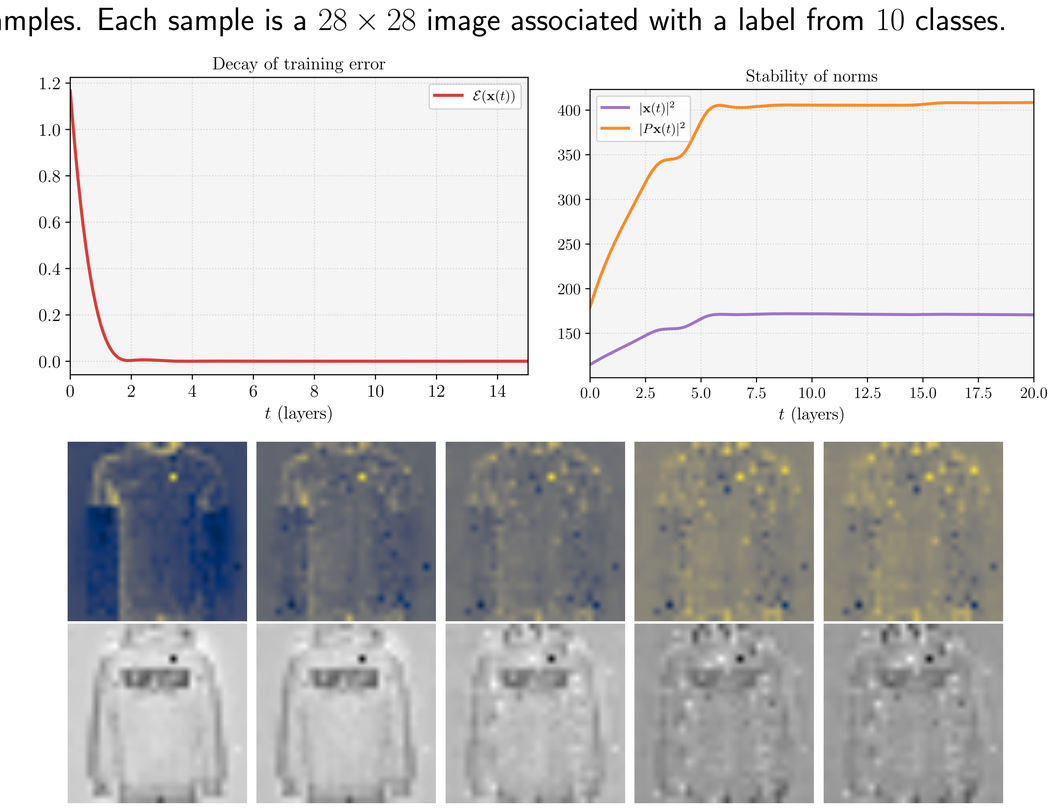
and, moreover,

$$\|u_T(t)\| \leqslant Ce^{-\mu t} \tag{5}$$

for some constant $C, \mu > 0$ independent of T and for all $t \in [0, T]$. This is a manifestation of the so-called *turnpike property*, well-known in optimal control and economics.

Experiments

Fashion-MNIST is a dataset of article images, consisting of a training set of 60000 samples. Each sample is a 28×28 image associated with a label from 10 classes.



Top: The decay of the training error and stabilization of optimal state trajectories as stipulated by turnpike. *Bottom:* The evolution of two individual samples $\mathbf{x}_i(t) \in \mathbb{R}^{784}$ at times $t \in \{0, 2, 8, 15, 20\}$. We see that each trajectory stabilizes to some stationary configuration.

Outlook

In the above presented works, we have studied a variety of supervised learning tasks from the continuous-time control theoretical perspective, allowing us to obtain fundamental understanding of the working mechanisms and properties of deep learning. We have, however, focused solely on supervised learning tasks, namely, wherein the dataset is labeled. A major challenge which ought to be formulated and addressed in a more control theoretical framework is the topic of *unsupervised learning*, wherein one only disposes of unlabeled data $\{\vec{x}_i\}_{i=1}^N$.

Selected publications

[1] Esteve-Yagüe, C., Geshkovski, B., Pighin, D., Zuazua, E. (2021). **Large-time asymptotics in deep learning.** arXiv preprint arXiv:2008.02491.

[2] Esteve-Yagüe, C., Geshkovski, B., Pighin, D., Zuazua, E. (2020). **Turnpike in Lipschitz-nonlinear optimal control.** arXiv preprint arXiv:2011.11091.

[3] Esteve-Yagüe, C., Geshkovski, B., (2021). **Sparse approximation in learning via neural ODEs.** arXiv preprint arXiv:2102.13566.

[4] Ruiz-Balet, D., Zuazua, E. **Neural ODE control for classification, approximation and transport.** In preparation.

caa-avh.nat.fau.eu
04/2021

Unterstützt von / Supported by
Alexander von Humboldt Stiftung/Foundation