Control and Deep Learning: Some connections

Borjan Geshkovski & Enrique Zuazua

March 11, 2021

This note is an extended abstract for a talk given by the second author during the workshop "Challenges in Optimization with Complex PDE-Systems", at Oberwolfach, in February 2021.

It is superfluous to state the impact that deep learning has had on modern technology, as it powers many tools of modern society, ranging from web search to content filtering on social networks ([1]). 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 $f(\cdot) : \mathcal{X} \to \mathcal{Y}$ 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 linear parametric operators and simple nonlinearities. A canonical example of such models is the *perceptron*

$$f_{\text{approx}}(x) = \sum_{j=1}^{d} w_{1,j}\sigma(w_{2,j}x + b_j) \tag{1}$$

where $w_1 \in \mathbb{R}^d$, $w_2 \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$ are unknown parameters, with $\sigma : \mathbb{R} \to \mathbb{R}$ being a globally Lipschitz continuous function, defined element-wise, the so-called *activation function*.

A by-now classical result, Cybenko's *universal approximation theorem* ([2]) ensures that the set of functions which can be represented by formula (1) is a dense subset of $C^0([-1,1]^d)$. This theory has since flourished, and universal approximation results have been shown for more compound models than (1) (see [3]).

In practice however, one looks to use models wherein the compositions are iterated over multiple layers, namely *deep neural networks*. A staple of such models are the so-called *residual neural networks* (ResNets, [4]) which may often be cast as schemes of the mould

$$\begin{cases} \mathbf{x}_i^{k+1} = \mathbf{x}_i^k + w_1^k\sigma(w_2^k\mathbf{x}_i^k + b^k) & \text{for } k \in \{0, \dots, N_{\text{layers}} - 1\} \\ \mathbf{x}_i^0 = \vec{x}_i \end{cases} \tag{2}$$

for all $i \in [N]$, where $[N] := \{1, \dots, N\}$, $w_1^k, w_2^k \in \mathbb{R}^{d\times d}$ and $N_{\text{layers}} \geq 1$ designates the number of layers referred to as the *depth*. Due to the inherent dynamical nature of ResNets, several recent works have considered an associated continuous-time formulation, a trend started with the work [5]. This is motivated by the simple observation that for $T > 0$, (2) is the forward Euler approximation of the

neural ordinary differential equation (neural ODE)

$$\begin{cases} \dot{\mathbf{x}}_i(t) = w_1(t)\sigma(w_2(t)\mathbf{x}_i(t) + b(t)) & \text{for } t \in (0,T) \\ \mathbf{x}_i(0) = \vec{x}_i \in \mathbb{R}^d. \end{cases} \tag{3}$$

It should be noted that the origins of continuous-time supervised learning go back to the 1980s – in [6] back-propagation algorithms are connected to the adjoint method arising in optimal control (see also [7, 8]).

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.

This is the viewpoint adopted by our group. And here we present some of our main findings.

We first analyze neural ODEs from a control theoretical perspective to obtain a fundamental understanding of the working mechanisms behind the processes of classification (more precisely, how the neural ODE flow manages separation of the different classes of data according to their labels).

These objectives are tackled and achieved from the perspective of the simultaneous control of systems of neural ODEs. Namely, in [9] we prove that both separation and universal approximation (to arbitrary L^∞_{loc} or L^2_{loc} functions) are valid properties for the controlled neural ODE flow by means of genuinely nonlinear and constructive proofs, allowing us to also estimate the complexity of the developed control strategies. Indeed, the nonlinear nature of the activation function 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. We also present the counterparts in the context of the control of neural transport equations, establishing a link between optimal transport and deep neural networks.

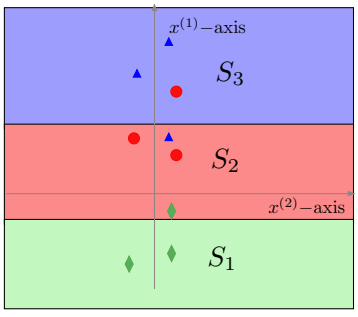

FIGURE 1. The constructive algorithm for designing controls which allow the flow to separate the labeled data into labeled strips.

In practical applications however, the time-dependent parameters/controls are found by minimizing some cost functional rather than explicitly, via a process commonly referred to as *training*. Due to the ODE reformulation of ResNets, the training process is nothing else than an optimal control problem which consists in finding optimal parameters steering all of the network outputs $P\mathbf{x}_i(T)$ as close as possible to the corresponding labels $\vec{y}_i$, where $P : \mathbb{R}^d \to \mathbb{R}^m$ is a given affine and surjective map (e.g., a random matrix) which serves to match dimensions.

In [10, 11], we propose the training problem consisting in minimizing

$$\frac{1}{N}\sum_{i=1}^{N} \text{loss}\left(P\mathbf{x}_i(T), \vec{y}_i\right) + \int_0^T \|\mathbf{x}_i(t) - \overline{\mathbf{x}}_i\|^2 dt + \|u\|^2_{H^1(0,T;\mathbb{R}^{d_u})}, \tag{4}$$

where $\text{loss}(\cdot,\cdot)$ is a given continuous and nonnegative function which, in classification tasks (for simplicity, $\vec{y}_i \in \{0,1\}$), is usually $\text{loss}(x,y) := \|\frac{1}{1+e^{-x}} - y\|^2$ or $\text{loss}(x,y) = \log(1+\exp(-yx))$, and $\overline{\mathbf{x}}_i \in P^{-1}(\{\vec{y}_i\})$.

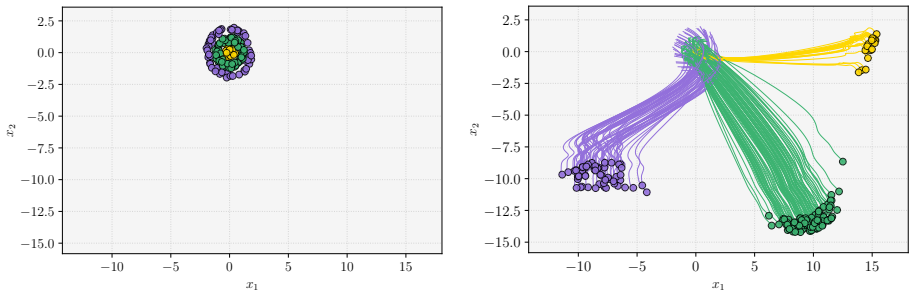

FIGURE 2. A depiction of the evolution of the trained neural ODE trajectories on a classification task with three labels – one sees the separation process over time.

As each time-step of a discretization to (3) may be seen to represent a different layer of the ResNet (2), the time horizon $T > 0$ in (3) may serve as an indicator of the number of layers N_{layers} in the discrete-time context (2). 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 [10, 12] (see [13] for the L^1–case), under controllability assumptions on the neural ODE (which are addressed in [9]), 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 $\mathbf{x}_T$ satisfy

$$\frac{1}{N}\sum_{i=1}^{N} \text{loss}\left(P\mathbf{x}_{T,i}(t), \vec{y}_i\right) + \|\mathbf{x}_{T,i}(t) - \overline{\mathbf{x}}_i\| \leq C\, e^{-\mu t} \tag{5}$$

and, moreover,

$$\|u_T(t)\| \leq Ce^{-\mu t} \tag{6}$$

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 ([14]).

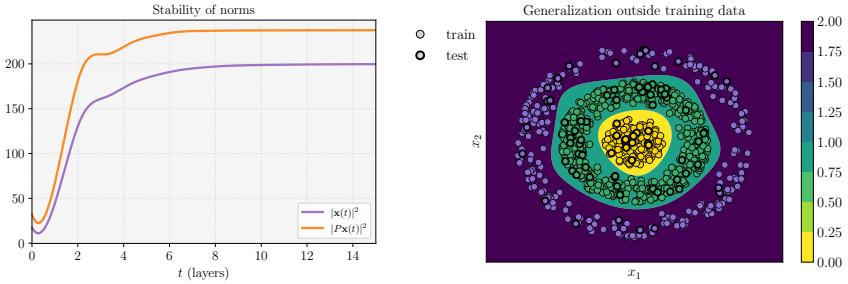

FIGURE 3. The trained neural ODE trajectories stabilize as per the turnpike property and generalize the shape of the dataset.

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 that 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 $\{\vec{x}_i\}$.

Acknowledgments. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No.765579-ConFlex, the Alexander von Humboldt-Professorship program, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement NO. 694126-DyCon), the Transregio 154 Project "Mathematical Modelling, Simulation and Optimization Using the Example of Gas Networks" of the German DFG, grant MTM2017-92996-C2-1-R COSNET of MINECO (Spain) and by the Air Force Office of Scientific Research (AFOSR) under Award NO. FA9550-18-1-0242.

REFERENCES

[1] LeCun, Y., Bengio, Y., and Hinton, G. *Deep learning.* Nature 521, 7553 (2015), 436–444.
[2] Cybenko, G. *Approximation by superpositions of a sigmoidal function.* Math. Control Signals Systems (1989), 303–314.
[3] Pinkus, A. *Approximation theory of the mlp model in neural networks.* Acta Numer. 8, 1 (1999), 143–195.

[4] He, K., Zhang, X., Ren, S., and Sun, J. *Deep residual learning for image recognition.* In Proceedings of the IEEE conference on computer vision and pattern recognition (2016), pp. 770–778.
[5] E, W. A proposal on machine learning via dynamical systems. *Commun. Math. Stat. 5*, 1 (2017), 1–11.
[6] LeCun, Y., Touresky, D., Hinton, G., and Sejnowski, T. *A theoretical framework for back-propagation.* In Proceedings of the 1988 connectionist models summer school (1988), vol. 1, CMU, Pittsburgh, Pa: Morgan Kaufmann, pp. 21–28.
[7] Sontag, E., and Sussmann, H. *Complete controllability of continuous-time recurrent neural networks.* Systems Control Lett. 30, 4 (1997), 177–183.
[8] Sontag, E. D., and Qiao, Y. *Further results on controllability of recurrent neural networks.* Systems Control Lett. 36, 2 (1999), 121–129.
[9] Ruiz-Balet, D., and Zuazua, E. *Neural ODE control for classification, approximation and transport.* In preparation (2021).
[10] Esteve, C., Geshkovski, B., Pighin, D., and Zuazua, E. *Large-time asymptotics in deep learning.* arXiv preprint arXiv:2008.02491 (2020).
[11] Geshkovski, B. *Control in moving interfaces and deep learning.* PhD Thesis (2021).
[12] Esteve, C., Geshkovski, B., Pighin, D., and Zuazua, E. *Turnpike in Lipschitz-nonlinear optimal control.* arXiv preprint arXiv:2011.11091 (2020).
[13] Esteve Yagüe, C., and Geshkovski, B. *Sparse approximation in learning via neural ODEs.* arXiv preprint arXiv:2102.13566 (2021).
[14] Trélat, E., and Zuazua, E. *The turnpike property in finite-dimensional nonlinear optimal control.* J. Differ. Equ. 258, 1 (2015), 81–114.

Borjan Geshkovski
Departamento de Matemáticas
Universidad Autónoma de Madrid
28049 Madrid, Spain
and
Chair of Computational Mathematics
Fundación Deusto
Av. de las Universidades, 24
48007 Bilbao, Basque Country, Spain
Email address: borjan.geshkovski@uam.es

Enrique Zuazua
Chair in Applied Analysis
Alexander von Humboldt-Professorship
Department of Mathematics
Friedrich-Alexander-Universität Erlangen-Nürnberg
91058 Erlangen, Germany
and
Chair of Computational Mathematics
Fundación Deusto
Av. de las Universidades, 24
48007 Bilbao, Basque Country, Spain
and
Departamento de Matemáticas
Universidad Autónoma de Madrid
28049 Madrid, Spain
Email address: enrique.zuazua@fau.de