Proceedings of the 9th. International Conference on Control and Optimization with Industrial Applications, August 27-29, 2024, Istanbul, Türkiye

CONTROL AND MACHINE LEARNING

E. ZUAZUA[1,2,3]

Abstract. In this talk, we discuss recent results that explore the relationship between control theory and machine learning, specifically applied to supervised learning. First, we study the classification and approximation properties of residual neural networks. Interpreting these problems as simultaneous or ensemble control ones, we build genuinely nonlinear and constructive algorithms, estimating the complexity of controls. Then, we analyze the multilayer perceptron architecture, characterizing the necessary depth and minimum width required to achieve simultaneous controllability. In the domain of large language models, residual networks are combined in an alternating manner with self-attention layers, whose role is to capture the "context". We view these layers as a dynamical system acting on a collection of points and characterize their asymptotic dynamics and convergence towards special points called *leaders*. We use our theoretical results to design an interpretable model to solve the task of sentiment analysis of movie reviews. Finally, we investigate federated learning, which enables multiple clients to collaboratively train models without sharing private data, thereby addressing data collection and privacy challenges. Within this framework, we address issues related to training efficiency, incentive mechanisms, and privacy concerns.

1. Introduction

The impact of machine learning in science is leading to rich and innovative lines of research in applied mathematics. There is a great need for theoretical foundations that guarantee the performance, reliability, and interpretability of machine learning methods. Specifically, mathematical models are required for understanding and optimizing rapidly emerging computational architectures like residual neural networks or transformers. This challenge can be addressed through the lens of control theory, a combination that offers great potential given the development in this area. In this lecture, we will discuss recent results from our group that explore the application of control tools to some of the main architectures and methods in machine learning, namely neural networks, self-attention, and federated learning.

2. Control-based supervised learning via neural networks

Supervised learning is one of the main paradigms of machine learning (ML), aiming to define a map that approximates an unknown function $f : \mathcal{X} \to \mathcal{Y}$ using a training dataset $\{(x_i, y_i)\}_{i=1}^N$. Neural networks form a widely used class of functions to approximate f, and among these, residual networks have proven to be particularly effective. In the continuous-time limit, these discrete systems transform into the so-called neural ODE:

$$\begin{cases} \dot{x} = W(t)\sigma(A(t)x + b(t)), & t \in (0,T), \\ x(0) = x_i, \end{cases} \tag{2.1}$$

for all $i \in [N] := \{1, \dots, N\}$. Here, x are the unknowns, $(W, A, b) \in L^\infty\left((0,T); \mathbb{R}^{d\times p} \times \mathbb{R}^{p\times d} \times \mathbb{R}^p\right)$ are piecewise constant controls taking L values; $L, p \geq 1$ are the depth and the width of the model, respectively. Moreover, $\sigma : \mathbb{R}^p \to \mathbb{R}^p$ is a Lipschitz-continuous non-linearity defined component-wise, a common example being the *rectified linear unit* (ReLU): $x \mapsto \max\{x, 0\}$.

One of the main advantages of neural ODEs, is the possibility to reinterpret several machine learning paradigms using tools from differential equations. For example, data classification can be formulated as a simultaneous control problem for (2.1), where our goal is to build controls $\{W, A, b\}$ that drive all initial data $\{x_i\}_{i=1}^N$ to their corresponding targets $\{y_i\}_{i=1}^N$ (prescribed according to the labels) through the flow map of the system (2.1).

In [7], we prove simultaneous controllability of (2.1) with an inductive algorithm that constructs explicit, piecewise constant controls (W, A, b) which sequentially guide each point x_i to its target y_i. Moreover, using similar techniques, we obtain a result of universal approximation for neural ODEs in norm $\|\cdot\|_{L^2}$. In [2], we reduce the complexity of the controls by proposing new algorithms for cluster-based classification. Our strategy aims to probabilistically reduce the number of parameters needed by leveraging the spatial structure of the data distribution.

[1]Chair for Dynamics, Control, Machine Learning, and Numerics, Alexander von Humboldt-Professorship, Department of Mathematics, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany.
[2]Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain.
[3]Chair of Computational Mathematics, Fundación Deusto. Av. de las Universidades, 24, 48007 Bilbao, Basque Country, Spain.

On the other hand, in [3], we focus on the role played by the architecture in controllability by studying the trade-off between depth and width. In the wide limit, (2.1) becomes autonomous, and we obtain an approximate control result with explicit error decay rates. This study on time-independent controls is closely related to the turnpike principle paradigm, which ensures that optimal control strategies remain almost steady over long time periods. In [4], we have analyzed the implications of this principle for designing more stable architectures for deep ResNets.

An extension of the developed theory reformulates the continuous model in terms of transport equations through the mean-field link between (2.1) and the hyperbolic transport PDE, leading to the neural transport equation:

$$\partial_t \rho + \mathrm{div}_x(W(t), \sigma(A(t), x + b(t))\rho) = 0. \tag{2.2}$$

We have studied the control problem for (2.2), aimed at transforming one given probability measure into another, up to an arbitrarily small Wasserstein-1 error [7, 3] or total variation error [8]. The first approach allows us to build a bridge with the theory of optimal transport, whereas the latter has applications in generative modeling via the technique known as normalizing flows.

In addition to the ResNet and its continuous counterpart, we have analyzed the so-called multilayer perceptron. This architecture is given by the following discrete dynamical system:

$$x_i^{k+1} = \sigma_k(A^k x_i^k + b^k), \quad k \in [L], \tag{2.3}$$

for all $i \in [N]$, where $x_i^1 = x_i$, $A^k \in \mathbb{R}^{d_k \times d_{k-1}}$, $b^k \in \mathbb{R}^{d_k}$, and $\{d_k\}_{k=1}^L$ is a sequence of positive integers for which d_k denote the width of (2.3) at the layer k. Here, $\sigma_k : \mathbb{R}^{d_k} \to \mathbb{R}^{d_k}$ denotes the (component-wise) ReLU function, and $\max_k\{d_k\}$ denotes the width of (2.3). In [5], for a dataset of N elements in $\mathbb{R}^d$ and M classes, we prove that (2.3) is simultaneously controllable with width 2 and at most $2N + 4M - 1$ layers. This is proven using an inductive algorithm that provides explicit values for the parameters. This result is sharp in the sense that (2.3) with width 1 cannot achieve simultaneous controllability. Additionally, in [5], for $p \in [1, \infty)$ and $\Omega \subset \mathbb{R}^d$, universal approximation (UA) for $L^p(\Omega; \mathbb{R}_+)$ functions is proven using (2.3) with width $d + 1$. Moreover, assuming $W^{1,p}$ regularity of the function, the necessary depth for UA is estimated in terms of the $W^{1,p}$ norm of the function to be approximated and the approximation error.

3. Self-attention as a clustering mechanism and its role in LLMs

For supervised learning tasks in large language models (LLMs), capturing "context" or how words relate to one another in a sentence, is a key feature. For this reason, the data samples used to train such models contain collections of words (i.e. sentences or paragraphs). More precisely, the training dataset is of the form $\{(Z_s, y_s)\}_{s=1}^N$, for matrices $Z_s \in \mathbb{R}^{n \times d}$, whose n rows encode words as points in Euclidean space $\mathbb{R}^d$. The *transformer* is a state-of-the-art neural networks in LLMs, which builds on ResNets by alternating with *self-attention* layers exploiting this data structure. Heuristically, these layers capture the "context" at the sample level by mixing its rows based on similarity between them.

For a fixed data sample $Z \in \mathbb{R}^{n \times d}$ with rows $z_1, \dots, z_n \in \mathbb{R}^d$, our model for (hardmax) self-attention is given by

$$z_i^{k+1} = z_i^k + \frac{\alpha}{1+\alpha} \frac{1}{|\mathcal{C}_i(Z^k)|} \sum_{j \in \mathcal{C}_i(Z^k)} \left(z_j^k - z_i^k\right), \quad k \geq 0, \tag{3.1a}$$

where $z_i^0 = z_i$, Z^k contains the rows $z_1^k, \dots, z_n^k$, $A \in \mathbb{R}^{d \times d}$ is a symmetric positive definite matrix, $\alpha > 0$, and

$$\mathcal{C}_i(Z^k) = \left\{ j \in [n] \ : \ \left\langle A z_i^k, z_j^k \right\rangle = \max_{\ell \in [n]} \left\langle A z_i^k, z_\ell^k \right\rangle \right\}. \tag{3.1b}$$

In the forthcoming preprint [1], we study the asymptotic dynamics of (3.1), that is, the behavior of self-attention as the number of layers $k \to \infty$. In particular, we prove that it exhibits clustering behaviour and that cluster points are determined by special points we call *leaders*. As an application, we use our clustering results to design a simple and interpretable transformer-based model to solve the supervised learning task in LLMs of *sentiment analysis*. We use a benchmark dataset with movie reviews, labeled as positive or negative. The proposed model contains only three components with distinct roles: the encoder, mapping words to points in $\mathbb{R}^d$, whose role is to select meaningful words as leaders; our transformer (3.1), whose role is to capture "context" by clustering the majority of words towards the few most meaningful ones; and the decoder, whose role is to project the final point values to a real prediction by dividing $\mathbb{R}^d$ in two half-spaces and identifying each half-space with each sentiment. After training the model, our interpretation is verified with examples (cf. Figure 1).

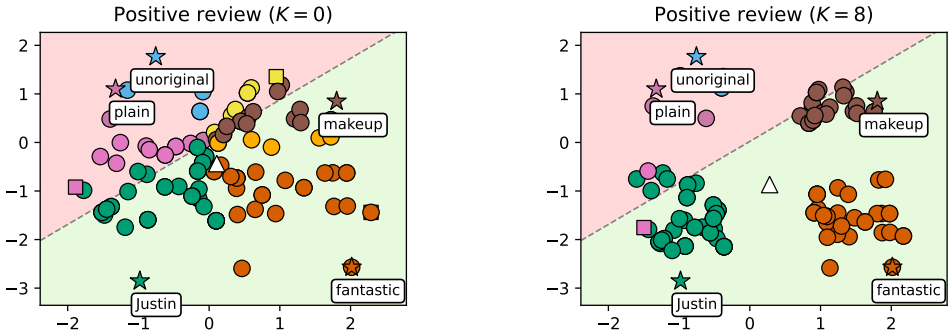

FIGURE 1. Evolution of the encoded words of a positive review as they are processed by $K = 8$ transformer layers. Points are colored according to the point they follow, leaders are stars and tagged with the word they encode, squares are non-leaders who are followed by other points, circles are the remaining points, and the triangle is the mean word. The dashed line is the hyperplane separating the negative class (red) from the positive class (green).

4. Federated learning: training, incentive, and privacy

With the growing amount of distributed data, *federated learning* (FL) has emerged as a promising paradigm to address challenges like data collection and privacy protection in centralized learning approaches. As in supervised learning, FL aims to learn a model to approximate $f : \mathcal{X} \to \mathcal{Y}$, but under the constraint that training data and labels are stored across distributed clients. Given m clients, the training of FL can be formulated as

$$\min_{\theta \in \mathcal{W}} \sum_{k=1}^{m} p_k \ell_k(\theta), \tag{4.1}$$

where $\theta \in \mathcal{W}$ are trainable parameters, $\ell_k : \mathcal{W} \to \mathbb{R}$ is client k's local loss function, commonly set as the empirical risk over its local dataset, and $p_k \geq 0$ with $\sum_{k=1}^{m} p_k = 1$ specifies the relative impact of client k.

To solve (4.1) efficiently, we propose in [10] an inexact and self-adaptive algorithm termed FedADMM-InSa. We design an inexactness criterion to guide each client to independently adjust its local training accuracy, leading to personalized training and better adaptation to heterogeneous data. Additionally, we present a self-adaptive scheme that dynamically adjusts each client's penalty parameter to enhance the robustness of our algorithm.

As in [10], existing research on FL primarily focuses on designing efficient learning algorithms, without considering that clients may be reluctant to engage without appropriate compensation (rewards from the server) for their training efforts. We address this issue in [6] by formulating incentive mechanisms in FL within a potential game framework. We investigate the uniqueness of the Nash equilibrium in these games and offer the server an easily calculable threshold for the reward, under which it can achieve effective incentives concerning clients' training efforts.

Moreover, the privacy benefits of FL (exchanging model parameters instead of data) can be compromised by data reconstruction attacks. In [9], we propose an approximate and weighted attack method to recover clients' private data under the widely used multiple-step local update scenarios. Experimental results validate the superiority of our attack method, emphasizing the need for effective defense mechanisms in FL to enhance privacy.

References

[1] A. Alcalde, G. Fantuzzi, and E. Zuazua. Clustering in pure-attention hardmax transformers and its role in sentiment analysis. In preparation (2024).
[2] A. Álvarez-López, R. Orive-Illera, and E. Zuazua. Optimized classification with neural odes via separability. *arXiv preprint arXiv:2312.13807*, 2023.
[3] A. Álvarez-López, A. H. Slimane, and E. Zuazua. Interplay between depth and width for interpolation in neural odes. *arXiv preprint arXiv:2401.09902*, 2024.
[4] B. Geshkovski and E. Zuazua. Turnpike in optimal control of pdes, resnets, and beyond. *Acta Numerica*, 31:135–263, 2022.
[5] M. Hernández and E. Zuazua. Deep neural networks: Multi-classification and universal approximation. In preparation (2024).
[6] K. Liu, Z. Wang, and E. Zuazua. Game theory in federated learning: A potential game perspective. In preparation (2024).
[7] D. Ruiz-Balet and E. Zuazua. Neural ode control for classification, approximation, and transport. *SIAM Review*, 65(3):735–773, 2023.
[8] D. Ruiz-Balet and E. Zuazua. Control of neural transport for normalising flows. *Journal de Mathématiques Pures et Appliquées*, 181:58–90, 2024.
[9] Y. Song, Z. Wang, and E. Zuazua. Approximate and weighted data reconstruction attack in federated learning. *arXiv preprint arXiv:2308.06822*, 2023.
[10] Y. Song, Z. Wang, and E. Zuazua. Fedadmm-insa: An inexact and self-adaptive admm for federated learning. *arXiv preprint arXiv:2402.13989*, 2024.