Internship: Estimation of the topology of decision boundaries: a neural ODE perspective.

Keywords: Topological Data Analysis (TDA), Invertible neural networks, neural Ordinary Differential Equation (ODE), Statistics.

City and Country: Champs-sur-Marne, France.

Team: SIGNAL, Laboratoire d'Informatique Gaspard Monge, Université Gustave Eiffel.

Internship Advisor: Théo Lacombe, `theo.lacombe@univ-eiffel.fr`.

Head of department:

(Team:) François-Xavier Vialard, `francois-xavier.vialard@univ-eiffel.fr`

(Lab:) Stéphane Vialette, `stephane.vialette@univ-eiffel.fr`.

General presentation of the topic: Let $\mathcal{X} \subset \mathbb{R}^d$ denote a space of observations with labels in $\{-1, 1\}$. Let $F : \mathcal{X} \to [-1, 1]$ be a binary classifier, that is, F assigns to each observation $x \in \mathcal{X}$ a value in $[-1, 1]$ which is in turn cast as a label $\mathrm{sgn}(F(x))$. Typically, F may be the map encoded by a neural network.

The map F splits $\mathcal{X}$ into two halves; while the set $\mathcal{B}(F) := \{x,\ F(x) = 0\}$ denotes the *decision boundary* of F. Intuitively, the geometric and topological properties of $\mathcal{B}(F)$ contain information regarding robustness properties of F: a complex (e.g. many connected components, high curvature, etc.) boundary may indicate that the classifier is prone to overfitting or may be sensitive to adversarial attacks, see Figure 1 for an illustration.

Studying (the properties of) $\mathcal{B}(F)$ and, ideally, regularizing it, is a natural goal. However, in practice, with most sophisticated modern models (e.g. deep neural networks), this boundary is mostly unknown, preventing a faithful computation of standard geometrical, topological, and other statistical descriptors. As such, works dealing with properties of such boundaries remain fairly limited to simple models (for which $\mathcal{B}(F)$ may be accessible in close form) or rely on naive sampling approaches that are unlikely to be reliable in practical settings.

Objectives of the internship: This project will focus on estimating topological properties of $\mathcal{B}(F)$ when F is a map encoded by a neural network. As the general case is likely to be

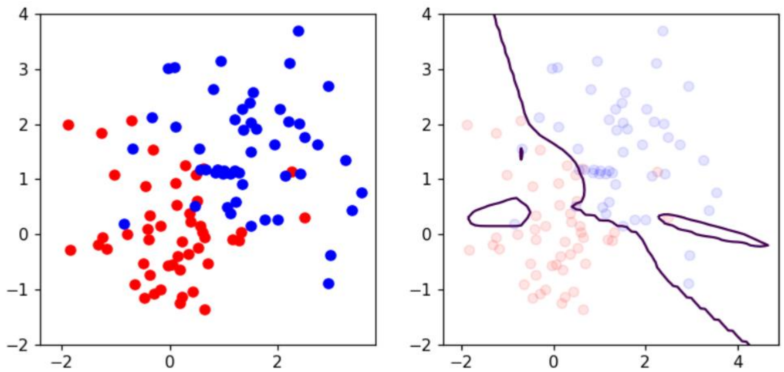

Figure 1: (Left) A simple 2D training dataset with two classes. (Right) The decision boundary reached by a network trained on this set. Thought the network reaches almost perfect training accuracy, the presence of spurious loops in the decision boundary that only catch few points of a given class indicates that the network is likely overfitting the training data.

infeasible, we propose to focus on the case of *invertible networks*, and particularly to *neural ODEs*. This particular type of neural networks can be understood as continuous extensions of *Residual Networks* (ResNet), which are among the most popular network architectures used to solve state-of-the-art learning problems.

Roughly speaking, a neural ODE will encode a map $F = C \circ F_0$, where $F_0 : \mathbb{R}^d \to \mathbb{R}^d$ is invertible and $C : \mathbb{R}^d \to \mathbb{R}$ is a simple binary classifier, for instance a linear one. Intuitively, F_0 will describe a "flow" that will push the data until C is able to separate them, while F_0^{-1} is the backward flow. The interesting part here is that $\mathcal{B}(F)$ becomes much simpler to understand: $F(x) = 0$ yields $x \in F_0^{-1}(C^{-1}(\{0\}))$. Furthermore, $C^{-1}(\{0\})$ describes an hyperplan $\mathcal{H} \subset \mathbb{R}^d$ accessible in close form, hence $\mathcal{B}(F)$ is the image of H by the backward flow F_0^{-1}. One can thus sample points in $\mathcal{H}$, and push them by F_0^{-1} to get a sample on $\mathcal{B}(F)$ which can in turn be used to estimate some topological properties of this boundary.

Of course, things will not be that easy; in particular, $\mathcal{H}$ is high-dimensional so it does not make sense to sample "uniformly on $\mathcal{H}$", yielding several challenges of different nature (computational, statistical, topological...). The steps of this internship may be, in what would be a natural chronological order[1]:

- Getting familiar with neural ODEs and their implementation.
- Getting familiar with previous literature in TDA involving estimation of the topology of classification boundaries.
- Empirical study of classification boundaries in low-dimensional settings; for instance, numerically observing the influence of the model parameters (number of parameters, penalization terms, etc.).

[1]Of course, this list is purely indicative and may change depending on the intermediate results obtained and the student appetence.

- Estimation of the topology (through the Čech persistence diagram) of the boundary in low dimensional settings; if the student is interested, this could be approached from both a numerical and theoretical perspectives.
- Trying to extend this approach to higher dimensional settings (e.g. starting with the MNIST dataset). This will require to sample points in $\mathcal{H}$ in a non-naive way; to do so we will first investigate an approach based on adversarial attacks. The goal would be to obtain an efficient and reliable method to estimate the topology of a neural ODE decision boundary in a fairly general setting.

Note also that this approach may be declined in various way: other type of geometrical or topological descriptors (e.g. curvature, different diagrams, etc.) and models, depending on the student wills and the progress of the project.

Expected abilities of the student: The student must be familiar with standard statistical and machine learning notions (classification, estimation, overfitting, etc.). A background in Topological Data Analysis is appreciated. A background in Deep Learning / neural ODE is *not* required (but of course would be appreciated as well), but the will to implement and experiment with such models is of importance.

References

[1] Chen, Ricky TQ and Rubanova, Yulia and Bettencourt, Jesse and Duvenaud, David: *Neural ordinary differential equations.* arXiv preprint arXiv:1806.07366, 2018.

[2] Chen, Chao and Ni, Xiuyan and Bai, Qinxun and Wang, Yusu: *A topological regularizer for classifiers via persistent homology.* The 22nd International Conference on Artificial Intelligence and Statistics, 2019.

[3] Vialard, François-Xavier and Kwitt, Roland and Wei, Susan and Niethammer, Marc: *A shooting formulation of deep learning.* Advances in Neural Information Processing Systems, 2020.

[4] Ramamurthy, Karthikeyan Natesan and Varshney, Kush and Mody, Krishnan: *Topological data analysis of decision boundaries with application to model selection.* International Conference on Machine Learning, 2019.

[5] Li, Weizhi and Dasarathy, Gautam and Ramamurthy, Karthikeyan Natesan and Berisha, Visar: *Finding the homology of decision boundaries with active learning.* arXiv preprint arXiv:2011.09645, 2020.

[6] Petri, Giovanni and Leitão, António: *On The Topological Expressive Power of Neural Networks.* NeurIPS 2020 Workshop on Topological Data Analysis and Beyond, 2020.

[7] Guss, William H and Salakhutdinov, Ruslan: *On characterizing the capacity of neural networks using algebraic topology.* arXiv preprint arXiv:1802.04443, 2018.