Research internship in machine learning theory (2019)

Advisor: Sébastien Gerchinovitz
Institut de Mathématiques de Toulouse (Université Paul Sabatier)

I have one Master 2 research internship opportunity in machine learning theory, which can be about one of the following two problems. If the internship is successful, we may consider applying for a PhD grant on the same research topic. Very motivated Master 1 students can also consider applying. If you are interested, please write an email (together with a CV) to:

sebastien.gerchinovitz@math.univ-toulouse.fr

Problem 1: Theoretical challenges in deep learning Over the past years, deep neural networks have proved remarkably efficient in machine learning tasks such as speech recognition, image or video processing, natural language processing, or solving board games (Goodfellow et al., 2016). However, the reasons why deep networks and the associated learning algorithms work so well in these applications is far from being well understood from a theoretical viewpoint. Indeed theoretical results about the geometry or statistical complexity of the function spaces they generate, as well as the optimization guarantees of algorithms such as stochastic gradient descent for deep networks, are still preliminary. In this internship, we will adress a few research questions about the approximation properties of feedforward neural networks.

An important part of the internship will be to understand the following three papers. Eldan and Shamir (2016) proved the existence of a 'phase transition' between feedforward neural networks (NN for short) with 1 versus 2 hidden layers. Specifically they show that when the dimension d of the inputs is large, there exists a function that can be approximated with a 2-hidden-layer NN of polynomial width, while it cannot be approximated by any 1-hidden-layer NN unless its width is exponential in the dimension d. The second paper, by Daniely (2017), provides a short proof of a very similar depth separation result. The third paper, by Yarotsky (2018), is about approximation properties of feedforward neural networks with possibly much more than 2 hidden layers. More precisely, given the dimension d of the inputs and a total number W of weights, Yarotsky characterized the best possible approximation of continuous functions $f : [0,1]^d \to \mathbb{R}$ in terms of their modulus of continuity and the computational budget W. An interesting byproduct of this result is that, in Yarotsky's setting, only very deep networks provide optimal approximation guarantees among feedforward neural networks.

Goals of the internship. The first goal of this internship is to learn and summarize state-of-the-art results in approximation theory for deep learning, in connection with several works in statistical learning (using, e.g., the concept of VC-dimension). A lot of research questions remain open or can be generalized, so the intern will also work on a few open problems.

Problem 2: An open problem in sequential learning theory Another research area in which the intern could work is sequential learning theory. More precisely, we will study adversarial online learning problems, which are iterative and robust versions of classical statistical learning problems. They are more robust in the sense that their theoretical guarantees no longer rely on the usual i.i.d. assumption on the observed data, which may fail in practice. This setting has various applications, e.g., in electricity consumption forecasting, meteorological forecasting, ad auction optimization, etc. The techniques involve simple concepts from statistics, optimization, and information theory.

More precisely, we will consider the following sequential prediction problem, where the goal is to predict the value of an outcome $y_t \in \mathbb{R}$ given a context vector $x_t \in \mathbb{R}^d$. At every round $t \geqslant 1$, the learner first chooses a function $\widehat{f}_t : \mathbb{R}^d \to \mathbb{R}$, then observes a context vector $x_t \in \mathbb{R}^d$ and makes the prediction $\widehat{f}_t(x_t)$, and finally observes the true outcome $y_t \in \mathbb{R}$; the error is measured though the square loss $\big(y_t - \widehat{f}_t(x_t)\big)^2$. Let $\mathcal{F}$ be a set of functions from $\mathbb{R}^d$ to $\mathbb{R}$. The learner's average performance on $T \geqslant 1$ rounds is quantified through its regret

$$\mathrm{Reg}_T(\mathcal{F}) \triangleq \frac{1}{T}\sum_{t=1}^{T}\big(y_t - \widehat{f}_t(x_t)\big)^2 - \inf_{f\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^{T}\big(y_t - f(x_t)\big)^2 \,,$$

which is the difference between the learner's average error and the average error of the best function $f \in \mathcal{F}$ in hindsight, which is unknown to the learner. The main goal is to design an online algorithm that provably guarantees a small regret $\mathrm{Reg}_T(\mathcal{F})$ for all sequences $(x_t)_{1\leqslant t\leqslant T}$ of (possibily bounded) contexts and all sequences $(y_t)_{1\leqslant t\leqslant T}$ of bounded outcomes.

Goals of the internship. We will mainly consider the particular case of *online linear regression*, which corresponds to a set $\mathcal{F}$ of linear predictors $f_u(x) = u \cdot x$, with $u \in \mathbb{R}^d$. The intern will first read Chapter 2 of the reference book by Cesa-Bianchi and Lugosi (2006), and then the very recent paper by Gaillard et al. (2018) about optimal regret bounds for online linear regression with unbounded comparison vectors $u \in \mathbb{R}^d$. The goal is to try to solve an open problem stated in the latter paper. If time allows, of if a PhD project is suggested, there will be several opportunities to solve other open problems for more complex nonparametric settings, i.e., when $\mathcal{F}$ is a subset of an infinite-dimensional function space.

References

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Amit Daniely. Depth separation for neural networks. In *Proceedings of the 2017 Conference on Learning Theory*, pages 690–696, 2017. URL http://proceedings.mlr.press/v65/daniely17a.html.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *29th Annual Conference on Learning Theory*, pages 907–940, 2016. URL http://proceedings.mlr.press/v49/eldan16.html.

Pierre Gaillard, Sébastien Gerchinovitz, Malo Huard, and Gilles Stoltz. Uniform regret bounds over R^d for the sequential linear regression problem with the square loss. 2018. URL https://arxiv.org/abs/1805.11386. arXiv:1805.11386.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL http://www.deeplearningbook.org.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In *Proceedings of the 31st Conference On Learning Theory*, pages 639–649, 2018. URL https://arxiv.org/abs/1802.03620.