Proposition de stage 2022/2023

Low dimensional generalization properties of neural networks

Theme: Machine learning and statistical physics
Responsable: Cyril Furtlehner
Laboratoire: TAU team INRIA Saclay, LISN
Adresse: Université Paris-Saclay, 91405 Orsay Cedex
Mail: Cyril.Furtlehner@inria.fr

1 The double descent paradox in Machine learning

The practical success of deep neural networks in machine learning came also in apparent contradiction with the standard theory of statistical learning regarding overfitting mechanisms. The problem is stated as follows: suppose we want to learn a function

$$y = f_\theta(\mathbf{x})$$

based on a training data set $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_k, y_k), k = 1, \ldots N\}$, which relates some input data $\mathbf{x} \in \mathbb{R}^D$ (a time series, an image...) to some real observation $y \in \mathbb{R}$ for a regression problem or to a label $y \in \{0, 1\}$ for a classification problem. $\theta \in \mathbb{R}^M$ is a vector of parameters which maybe the weights of a linear regression, the parameters of a neural network or any ML model. The function is typically learned by performing a gradient descent on a loss function $\ell(\theta|\mathcal{D}_{\text{train}})$ and the central question then is to know whether the model will generalize well on unseen data, i.e. what is the behaviour of $\ell(\theta|\mathcal{D}_{\text{test}})$, where $\mathcal{D}_{\text{test}}$ is a test set of data. The traditional scenario expected from statistical learning theory is that for fixed N when varying the number of parameters M the train error will monotonically decrease to zero when reaching some critical value ρ_0 of the interpolation ratio $\rho = M/N$ ($= 1$ typically for a linear regression) while the generalization error will reach a minimum for some value $\rho^\star < \rho_0$ and increase again drastically for $\rho > \rho^\star$ because of overfitting. The surprise came when realizing that increasing further ρ may possibly yield a better solution at some point far in the over parameterized regime when some implicit regularization is present. Actually, in close analogy with the analysis of the spectrum of disordered quantum systems, this question can be investigated with help of a field theory formalism where the bare two-point function corresponds to the neural tangent kernel introduced recently [2]

$$G_0(\mathbf{x},\mathbf{x}') = \nabla_\theta^T f_{\theta_0}(\mathbf{x}) \nabla_\theta^T f_{\theta_0}(\mathbf{x}')$$

corresponding to an L_2 regularization of the solutions, while the central object of interest is obtained by including a random potential attached to the training data. Within this framework the generalization error as a function of ρ can be analyzed by means of the spectral properties of the disordered kernel. More precisely, under some general assumptions the generalization properties of the model can be obtained in the linerar regime of ridge regression, thanks to random matrix theory applied to the disordered kernel [2]. This random matrix regime is actually valid as long as the dimension d of $\mathbf{x}$ is large $(d \gg 1)$

2 Objectives of the internship

We would like to to investigate the generalization properties for low dimensional systems i.e. for $d = 1$ or 2. In particular for $d = 1$ we may expect to be able to obtain exact formulaes for the spectral density of the disordered kernel out of which train and test errors can be computed. Hence given a specific regularization we propose to:

- extract the spectral properties of the disordered kernel
- compute the generalization error in various regimes, namely under and over-parameterized ones.

Based on these "exact" model settings where the average over disorder can be done precisely, we would like to analyze the typical behavior of learning curves obtained with real data sets by looking in particular at the spectra of the associated disordered kernels.

Further reading

[1] Jacot, A. and Gabriel, F. and Hongler, C. “Neural Tangent Kernel: Convergence and Generalization in Neural Networks” NeurIPS (2018)

[2] C. Furtlehner, “Free dynamics of feature learning processes”, arXiv:2210.10702 (2022)