Proposition de stage 2021/2022

Analysis of learning curves via disordered kernel functions

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 [1]. The problem is stated as follows: suppose we want to learn a function

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

based on a training data set $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_k, y_k), k = 1, \dots N\}$, 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 α_0 of the interpolation ratio $\alpha = M/N$ ($= 1$ typically for a linear regression) while the generalization error will reach a minimum for some value $\alpha^\star < \alpha_0$ and increase again drastically for $\alpha > \alpha^\star$ because of overfitting. The surprise came when realizing that increasing further α 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 (as e.g. in [2]) where the bare two-point function corresponds to the neural tangent kernel introduced recently [3]

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

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 α can be analyzed by means of the spectral properties of the disordered kernel.

2 Objectives of the internship

We would like to investigate a certain number of models in this framework in combinations with various regularization by following these steps:

- extracting the spectral properties of the disordered kernel
- computing the generalization error in various regimes

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] M. Belkin, D. Hsu, S. Ma and S. Mandal, "Reconciling modern machine-learning practice and the classical biasâĂŞvariance trade-off", PNAS (2019)

[2] O. Cohen, O. Malka, Z. Ringel, "Learning curves for overparametrized deep neural networks: A field theory perspective", Phys.Rev. Res. (2021)

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