Aspects théoriques des réseaux de neurones

- Borjan Geshkovski (INRIA)

  Title: Dynamic metastability in the self-attention model

  Abstract: The pure self-attention model is a simplification of the celebrated Transformer architecture, which neglects multi-layer perceptron layers and includes only a single inverse temperature parameter. Despite its apparent simplicity, the model exhibits a remarkably similar qualitative behavior across layers to that observed empirically in a pre-trained Transformer. Viewing layers as a time variable, the self-attention model can be interpreted as an interacting particle system on the unit sphere. We show that when the temperature is sufficiently high, all particles collapse into a single cluster exponentially fast. On the other hand, when the temperature falls below a certain threshold, we show that although the particles eventually collapse into a single cluster, the required time is at least exponentially long. This is a manifestation of dynamic metastability: particles remain trapped in a "slow manifold" consisting of several clusters for exponentially long periods of time. Our proofs make use of the fact that the self-attention model can be written as the gradient flow of a specific interaction energy functional previously found in combinatorics.

- Pierre Marion (EPFL)

  Title: Three stories on deep linear networks

  Abstract: We discuss three related facets of optimization dynamics of deep linear networks with a quadratic loss, in connection to the largest eigenvalue of the loss Hessian, also known as the sharpness. The first result regards the maximal learning rate to ensure stable learning. We show that it is upper-bounded and explain this by a lower-bound on the sharpness of minimizers, which grows linearly with depth. Second, we study the properties of the minimizer found by gradient flow, which is the limit of gradient descent with vanishing learning rate, starting from a small-scale initialisation. We show that the learned weight matrices are approximately rank-one and that their singular vectors align. This implies an implicit regularization towards flat minima: the sharpness of the minimizer is no more than a constant times the lower bound. Finally, we study the case of a residual initialization. Convergence of the gradient flow for a Gaussian initialization of the residual network is proven in this case, as well as a bound on the sharpness of the minimizer.

- Paul Viallard (INRIA)

  Title: Uniform convergence bounds via PAC-Bayes and Wasserstein distances

  Abstract: In machine learning, practitioners may encounter overfitting when the model performs well on the training set but poorly on the learning task (represented by the test set). One way to assess overfitting is through generalization bounds, which provide upper bounds on the model's performance for the learning task. In this talk, I will first recap two types of bounds introduced in the literature: PAC-Bayesian and uniform convergence bounds. While these two types of bounds exhibit distinct natures, I will introduce a new approach to obtain generalization bounds that combine the strengths of both. More precisely, I will discuss how to leverage Wasserstein distances to convert PAC-Bayesian bounds into uniform convergence bounds.