Les mathématiques de l'apprentissage machine
« Atelier sur les défis modernes en théorie d'apprentissage »
23 au 26 avril 2018

Mathematics of Machine Learning
"Workshop on Modern Challenges of Learning Theory"
April 23-26, 2018

Figuring out the magic behind backprop and SGD

Yoshua Bengio [*]

yoshua.umontreal@gmail.com

Why is it that an algorithm as simple as back-propagation with stochastic gradient descent SGD so successful to train deep neural networks? One the one hand, the optimization problem in very high-dimensional problems may be easier than initially thought, with bad local minima probably less of an issue than was feared for many decades. On the other hand, back-propagation, even when crippled with lots of noise, still leads to very fast convergence, and much faster convergence than perturbation-based methods, but open frontier is how to deal with estimation of gradients through non-differentiable black boxes or discrete transformations. SGD variants end up having surprisingly good properties, both in terms of optimization and in terms of generalization. SGD focusses first on what the training examples have in common, on the factors that can explain many examples, and only much later on the exceptions, outliers and noisy examples which would otherwise lead to overfitting.

[*] DIRO, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal, QC H3C 3J7, CANADA