Instructorship position in Applied and/or Computational Mathematics

Talks via zoom – Schedule

Thursday, 4th February 2021

17h15 - 17h45 **Dr Dmitrii OSTROVSKII**

University of Southern California, Los Angeles, USA

Zoom link:

Title: Near-optimal model discrimination with non-disclosure

Abstract: In the standard setup of parametric M-estimation with convex loss, we consider two population risk minimizers associated with two possible distributions of a random observation, and then ask the following question:

Given the value of one of the two minimizers and access to i.i.d. sampling from both distributions, what sample sizes allow to identify the ``right'' distribution, i.e., the one corresponding to the given value?

Making the first steps towards answering this question in full generality, we first consider the case of a well-specified linear model with squared loss. Here we provide nearly matching upper and lower bounds on the sample complexity, showing it to be min{1/Δ^2,\sqrt{r}/Δ} up to a constant factor, where Δ is a measure of separation between the two distributions and r is the maximum (resp., minimum) of the ranks of the design covariance matrices in the case of the upper (resp., lower) bound. This bound is dimension-independent and rank-independent for large enough separation. We then extend this result in two directions: (i) for the general parametric setup in asymptotic regime; (ii) for generalized linear models in the small-sample regime n < r and under weak moment assumptions. In both cases, we derive sample complexity bounds of a similar form, even under misspecification. Our testing procedures only access the ``known'' value through a certain functional of empirical risk. In addition, the number of observations that allows to reach statistical confidence for testing does not allow to "resolve" the two models -- that is, recover both minimizers up to O(Δ) prediction accuracy. These two properties open the prospect of applying our testing framework in practical tasks, where one would like to *identify* a prediction model, which can be proprietary, while guaranteeing that the model cannot be *inferred* by the identifying agent.