Adversarial attacks on non-differentiable statistical models

Background: Adversarial attacks were first discovered in the context of deep neural networks (DNNs), where the networks' gradients were used to produce small bounded-norm perturbations of the input that significantly altered their output. Such attacks target the increase of the model's loss or the decrease of its accuracy and were shown to undermine the impressive performance of DNNs in multiple fields. Relevant papers: “Explaining and harnessing adversarial examples”.
Project Description: In this project, we aim to produce adversarial attacks on non-differentiable models such as statistical algorithms.
Prerequisites: Deep learning course
Supervisor: Yaniv Nemcovsky (yanemcovsky@gmail.com)