Wasserstein adversarial robustness for NN

Jan Obloj
University of Oxford, United Kingdom
jan.obloj@maths.ox.ac.uk

We develop applications of tools from the optimal transport theory to adversarial robustness of NN. The latter refers to the phenomenon where a successfully trained NN architecture can be fooled by humanly imperceptible changes to the inputs. First discussed in the seminal work of Goodfellow et al (2015), the task of understanding sensitivity to such attacks and of developing training methods which are robust to such adversarial data attacks, is an important topic in the ML literature. We refer to RobustBench for a list of papers and a zoo of benchmarks. We re-interpret the problem as a distributionally robust optimization, interpret data as probability measures and employ Wasserstein balls to characterise potential adversarial perturbations. We then rely on the results in Bartl, Drapeau, Obloj and Wiesel (2021) to obtain explicit first-order approximations to the robust value and robust optimizers. This allows us to quantify, for small perturbations, the adversarial robustness and derive candidates for robust training methods. We show in particular how these allow to recover some classical approaches, such as the FGSM. We test our theoretical predictions on the model zoo available through RobustBench and report the observed empirical fit.

Joint work with Xingjian Bai (University of Oxford, UK), Yifan Jiang (University of Oxford, UK) and Guangyi He (University of Oxford, UK).