Adversarial robustness in classification via the lens of optimal transport

Jakwang Kim

PIMS, Mathematics of Uuiversity of British Columbia

In this talk, I introduce the recent advance of the adversarial training problems of classification via optimal transport perspective. Since neural networks revolutionized the machine learning community, there have been tons of research to understand these objects. One critical issue is their instability against well-designed perturbation, which potentially causes serious problems in the application of deep learning. For this reason, people introduce so-called the adversarial training model for achieving more stable (robust) classifiers. Despite its practical importance, there has been no rigorous framework to describe and understand this model even until recently. In series papers, with Nicolas Garcia Trillos, Matt Jacobs and Matt Werenski, we do the following: (1) to connect the multiclass adversarial training problem to optimal transport and generalized barycenter problem, which first illustrates the geometry of this problem, (2) to prove the existence of robust classifiers, and unify variants of adversarial training models, and (3) to propose an efficient numerical scheme based on the combination of our theory and entropic optimal transport from computational optimal transport.