RACHEL MORRIS, Concordia University

Regularity guarantees for adversarially robust learning

While neural network image classification enjoys high success rates in most settings, recent work discovered that well-targeted adversarial attacks can transform a correctly classified image into one that is visually indistinguishable from the original but that completely fools the classification algorithm. This has sparked many new approaches to classification which include an adversary in the training process: such an adversary can improve robustness and generalization properties, at the cost of decreased accuracy and increased training time. By considering a "worst-case" adversary, the resulting mathematical model for adversarial training can be understood as an energy minimization problem with a regularizing nonlocal perimeter term. In this presentation, I will discuss my current work studying regularity guarantees for the decision boundary of an adversarially robust minimizer. In particular, for a continuous and bounded underlying density, the decision boundary is C^2 smooth. I will discuss using explicit geometric perturbations and second variation analysis to show singular points (i.e. corners, cusps) are suboptimal. For the smoother points, I will demonstrate how leveraging necessary conditions allows one to upgrade C^1 regularity to C^2 regularity.