The Geometry of Adversarial Training

Leon Bungert
Hausdorff Center for Mathematics, University of Bonn
leon.bungert@hcm.uni-bonn.de

In this talk I will show that "Adversarial Training" [1]—a methodology designed for the training of adversarially robust classifiers—is equivalent to a variational regularization problem involving a nonlocal perimeter term. Using this structure one can show that adversarial training admits a convex relaxation which is reminiscent of the Chan-Esedoglu model from image denoising [2]. Furthermore, this allows to prove existence of solutions and study finer properties and regularity. Finally, I hint at how to modify adversarial training to an Almgren-Taylor-Wang [3] like scheme for mean curvature flow.

Joint work with: Nicolás García Trillos, Ryan Murray.

References

[1] A. Madry, et al. Towards deep learning models resistant to adversarial attacks. *ICLR'18*.

[2] Tony F. Chan, S. Esedoglu. Aspects of total variation regularized L^1 function approximation. *SIAM Journal on Applied Mathematics*, 65(5): 1817–1837, 2005.

[3] F. Almgren, J. E. Taylor, L. Wang. Curvature-driven flows: a variational approach. *SIAM Journal on Control and Optimization*, 31(2), 387–438, 1993.