1151-62-110 **Ethan X Fang*** (xxf13@psu.edu), 516 E Irvin Ave, State College, PA 16801, and **Yan Li**, **Huan Xu** and **Tuo Zhao**. *Inductive Bias of Gradient Descent based Adversarial Training on Separable Data.*

Adversarial training is a principled approach for training robust neural networks. Despite of tremendous successes in practice, its theoretical properties still remain largely unexplored. In this paper, we provide new theoretical insights of gradient descent based adversarial training by studying its computational properties, specifically on its inductive bias. We take the binary classification task on linearly separable data as an illustrative example, where the loss asymptotically attains its infimum as the parameter diverges to infinity along certain directions. Specifically, we show that when the adversarial perturbation during training has bounded ℓ_2-norm, the classifier learned by gradient descent based adversarial training converges in direction to the maximum ℓ_2-norm margin classifier at the rate of $\tilde{\mathcal{O}}(1/\sqrt{T})$, significantly faster than the rate $\mathcal{O}\{1/\log T\}$ of training with clean data. In addition, when the adversarial perturbation during training has bounded ℓ_q-norm, the resulting classifier converges in direction to a maximum mixed-norm margin classifier, which has a natural interpretation of robustness, as being the maximum ℓ_2-norm margin classifier under worst-case ℓ_q-norm perturbation to the data. (Received August 12, 2019)