Can Implicit Bias Imply Adversarial Robustness?

YES, by choosing the right activation

Hancheng Min, René Vidal

Center for Innovation in Data Engineering and Science (IDEAS), University of Pennsylvania

ICML International Conference On Machine Learning

INTRODUCTION

- **Implicit bias of GD often benefits generalization**
- **Does it also improve adversarial robustness?**
- **If standard networks fail to be robust, can we alter the bias to favor more robust networks?**

pReLU network, $p \geq 1$; $\theta := \{w_j, v_j\}_j^h$

$$f_p(x;\theta) = \sum_{j=1}^h v_j \frac{\sigma^p(\langle x, w_j\rangle)}{\|w_j\|^{p-1}},\quad \sigma\text{: ReLU}$$

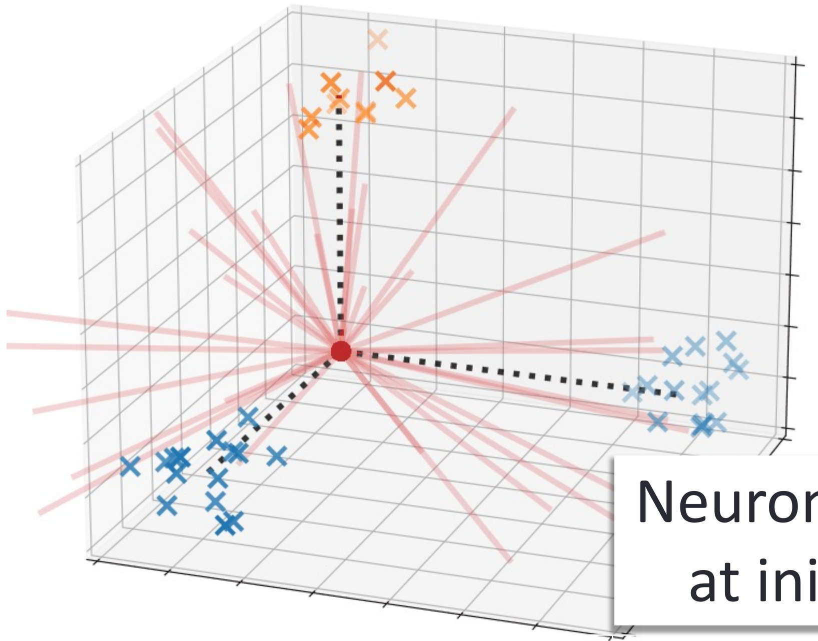
Loss: $\mathcal{L} = \sum_{i=1}^n \ell(y_i f_p(x_i;\theta))$

ℓ: exp. or logistic loss

Gradient flow (GF) with small initialization:

$$\dot{\theta} = -\nabla_\theta \mathcal{L}, \|\theta(0)\| \ll 1$$

Problem: Training shallow networks for binary classification problems with orthogonal data clusters



Neurons visualized at initialization

x : Positive data $\{x_i : y_i = +1\}$ **o** : Neurons $\{w_j\}$

x : Negative data $\{x_i : y_i = -1\}$ - : Neuron directions $\left\{\frac{w_j}{\|w_j\|}\right\}$

⋯ : Cluster centers

Data: samples from balanced mix. of Gaussians

$\mathcal{N}(\mu_1, \alpha^2 I), \cdots, \mathcal{N}(\mu_{K_1}, \alpha^2 I)$ K_1 pos. clusters

$\mathcal{N}(\mu_{K_1+1}, \alpha^2 I), \cdots \mathcal{N}(\mu_K, \alpha^2 I)$ K_2 neg. clusters

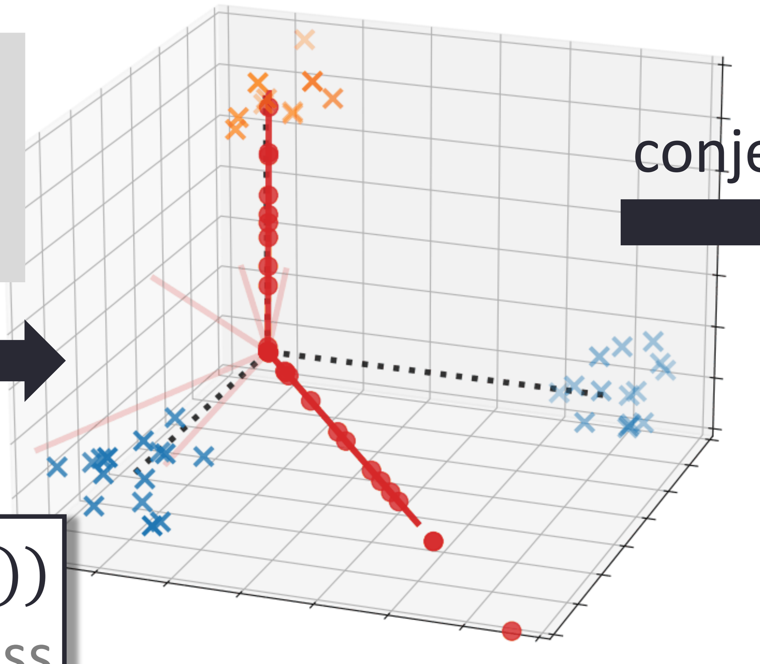
Cluster centers: $\mu_1, \cdots, \mu_K$ are orthonormal

Class average: $\mu_+ := \sum_{k=1}^{K_1} \mu_k$, $\mu_- := \sum_{k=K_1+1}^{K} \mu_k$

HARNESSING IMPLICIT BIAS FOR ADVERSARIAL ROBUSTNESS

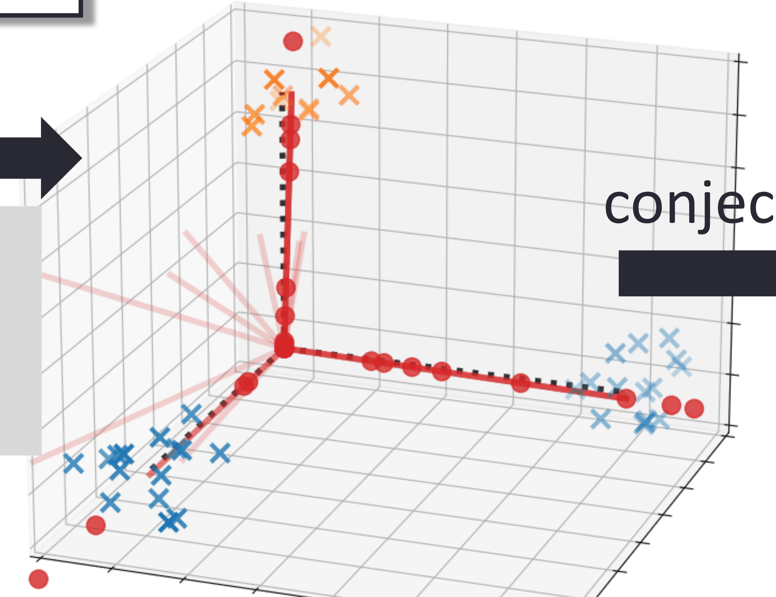
Observations

$p = 1$: ReLU Net
Neurons learn class average



Neurons visualized at the end of training

$p > 2$: pReLU Net
Neurons learn cluster centers



Conjectures

After training for sufficient time T

conjecture

$\boldsymbol{f_1(\cdot\,;\theta(T)) \approx F}$ up to a scaling factor

$F(x) = \sigma(\langle x, \mu_+\rangle) - \sigma(\langle x, \mu_-\rangle)$

- $\boldsymbol{F}$ is a ReLU network with two neurons μ_+ and μ_-
- GF on $\boldsymbol{f_1}$ converges to *ReLU classifier* $\boldsymbol{F}$

Implicit bias in shallow ReLU networks can harm adversarial robustness

Implicit bias of pReLU networks for p>3 leads to robust networks

conjecture

$\boldsymbol{f_p(\cdot\,;\theta(T)) \approx F^{(p)}}$ up to a scaling factor

$F^{(p)}(x) = \sum_{k=1}^{K_1} \sigma^p(\langle x, \mu_k\rangle) - \sum_{k=K_1+1}^{K} \sigma^p(\langle x, \mu_k\rangle)$

- $\boldsymbol{F^{(p)}}$ is a pReLU network with neurons $\mu_1, \cdots, \mu_K$
- GF on $\boldsymbol{f_p}$ ($p > 2$) converges to *pReLU classifier* $\boldsymbol{F^{(p)}}$

Main Theorems

Frei et al., 23: similar results hold for any ReLU net learned via GF/GD

New sample $(x, y) \in \mathbb{R}^D \times \{+1, -1\}$

ReLU classifier F

- ***Generalize on clean data***

$$\mathbb{P}(F(x)y > 0) \geq 1 - 2\exp\left(-\frac{CD}{4\alpha^2 K}\right)$$

- ***Non-robust against $\mathcal{O}(1/\sqrt{K})$-attack***

There exist $d_0 \in \mathbb{S}^{D-1}$, such that $\forall \rho > 0$

$$\mathbb{P}\left(F\left(x + \frac{1+\rho}{\sqrt{K}} d_0\right) y > 0\right) \leq 2\exp\left(-\frac{CD\rho^2}{\alpha^2 K}\right)$$

pReLU classifier $F^{(p)}$

- ***Generalize on clean data***

$$\mathbb{P}(F^{(p)}(x)y > 0) \geq 1 - 2(K+1)\exp\left(-\frac{CD}{\alpha^2 K}\right)$$

- ***Robust against $\mathcal{O}(1)$-attack***

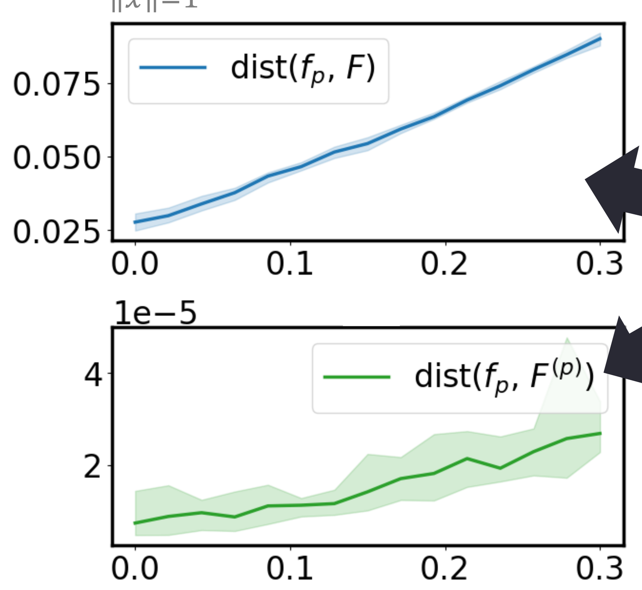
Let $p > 2$, then $\forall \delta \in (0, \sqrt{2})$,

$$\mathbb{P}\left(\min_{\|d\| \leq 1}\left[F^{(p)}\left(x + \frac{\sqrt{2}-\delta}{2} d\right) y\right] > 0\right) \geq 1 - 2(K+1)\exp\left(-\frac{CD\delta^2}{2\alpha^2 K^2}\right)$$

NUMERICAL EXPERIMENTS

Our conjectures are empirically verified

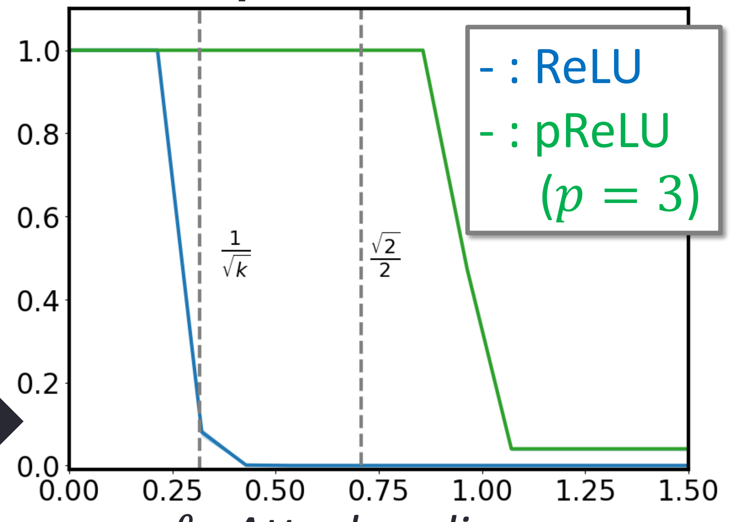
$\mathrm{dist}(f_p, F) = \inf_{c>0} \sup_{\|x\|=1} |cf_p(x) - F(x)|$



intra-cluster variance α

- $D = 1000, K_1 = 6, K_2 = 4$
- $p = 1$ (ReLU): $f_p \approx F$
- $p = 3$ (pReLU): $f_p \approx F^{(p)}$
- Robustness of f_p is well predicted by our theory

Robust Accuracy of f_p on new (x, y)



ℓ_2 Attack radius

pReLU improves the robustness of trained network on MNIST without adversarial training

Experiment details:

- Digits are centered by the mean digit of the training set
- CE loss; Kaiming initialization; Adam with batch size 1000
- Attack model: AutoAttack (Croce&Hein, 20)