Gradient Flow Provably Learns Robust Classifiers for Orthonormal GMMs

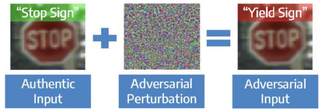
Hancheng Min, René Vidal

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

ICML International Conference On Machine Learning

INTRODUCTION

- NNs are often vulnerable to adversarial attacks
- [Pal et al., 2023]: If data is "concentrated and separated", robust classifiers exist without sacrificing clean accuracy



How can we find such robust classifiers by training NNs?

Problem: Training shallow networks for binary classification problems with orthogonal GMMs

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

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

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:

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

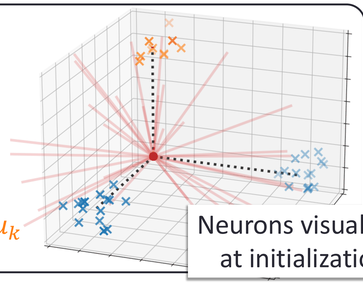
Data: samples from balanced mix. of Gaussians

$\mathcal{N}(\mu_1, \alpha^2 I), \cdots, \mathcal{N}(\mu_{K_1}, \alpha^2 I)$ 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: $\mu_+ := \sum_{k=1}^{K_1} \mu_k$, $\mu_- := \sum_{k=K_1+1}^{K} \mu_k$



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

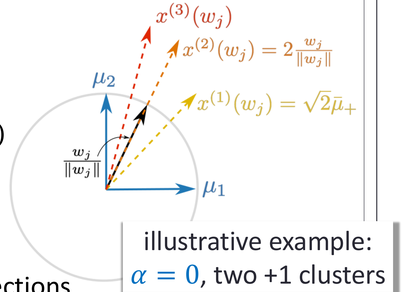
CRUCIAL ROLE OF IMPLICIT BIAS

- Small initialization ⟹ Two-phase neuron dynamics (direction $\frac{w_j}{\|w_j\|}$ first, then norm $\|w_j\|$)
- First alignment phase: Neuron move towards $x^{(p)}(w_j)$

$$\frac{d}{dt}\frac{w_j}{\|w_j\|} \approx \mathcal{P}_{w_j}^{\perp}\left(\underbrace{\sum_{i:\langle x_i, w_j\rangle > 0} x_i p \cos^{p-1}(x_i, w_j)\, y_i}_{x^{(p)}(w_j)}\right)$$

$\mathcal{P}_{w_j}^{\perp} := \left(I - \frac{w_j w_j^\top}{\|w_j\|^2}\right)$

- Depending the value of p, neurons learn different directions

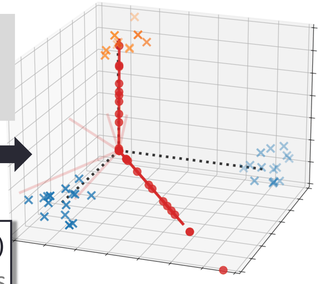


illustrative example: $\alpha = 0$, two +1 clusters

GRADIENT FLOW PROVABLY LEARNS ROBUST CLASSIFIERS

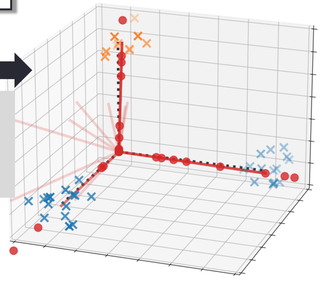
Observations

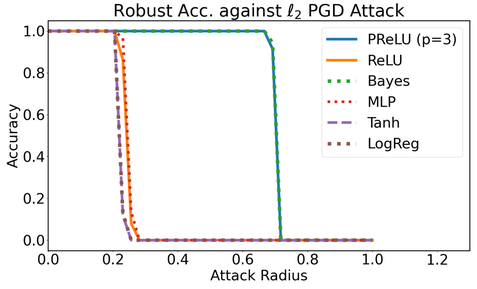
$p = 1$: ReLU Net
Neurons learn class average



Neurons visualized at the end of training

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





- Vulnerability of ReLU net persists even after: adding layers (MLP), change activations (Tanh, LogReg)
- Carefully chosen activation is the solution

Theorems

Provable vulnerability of ReLU (Prior works)

[Frei et al., 2023]: Any ReLU network trained by GF/GD is non-robust against $\mathcal{O}(1/\sqrt{K})$-radius ℓ_2 attacks

[Li et al., 2025]: ReLU network trained by GD with small initialization:

$f_1(x;\theta_T) \propto F(x) = \sigma(\langle x, \mu_+\rangle) - \sigma(\langle x, \mu_-\rangle)$

[Min and Vidal, 2024]: $F(x)$ is non-robust against $\mathcal{O}(1/\sqrt{K})$-attacks

Provable robustness of of pReLU (This work)

pReLU network ($p > 2$) trained by GF with small initialization:

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

$F^{(p)}(x)$ (p large) ≈ Bayes classifier ⟹ **Robust against $\mathcal{O}(1)$-attack:**

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

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

Optimal robust classifier: clusters are separated by $\sqrt{2}$ distance

$\sqrt{2}/2$ is the maximum achievable robustness w.o. clean acc drop

References

Pal et al., Adversarial examples might be avoidable: The role of data concentration in adversarial robustness. NeurIPS, 2023.

Frei et al., The double-edged sword of implicit bias: Generalization vs. robustness in ReLU networks. NeurIPS, 2023.

Li et al., Feature averaging: An implicit bias of gradient descent leading to non-robustness in neural networks. ICLR, 2025

Min and Vidal, Can implicit bias imply adversarial robustness? ICML, 2024