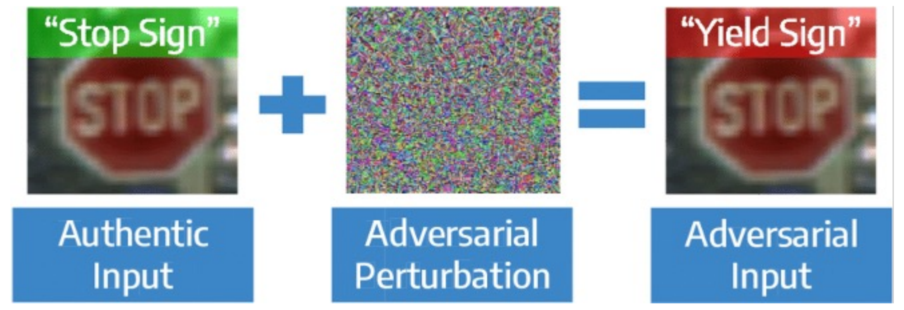
Gradient Flow Provably Learns Robust Classifiers for Orthonormal GMMs

Hancheng Min René Vidal

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

INTRODUCTION

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



How can we find such robust classifiers by training NNs?

PROBLEM

Problem: train shallow networks for binary classification of data from orthogonal GMMs

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

Normalized class centers:

$$\mu_+ := \frac{1}{\sqrt{K_1}}\sum_{k=1}^{K_1}\mu_k\,,\ \mu_- := \frac{1}{\sqrt{K_2}}\sum_{k=K_1+1}^{K}\mu_k$$

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 log. loss

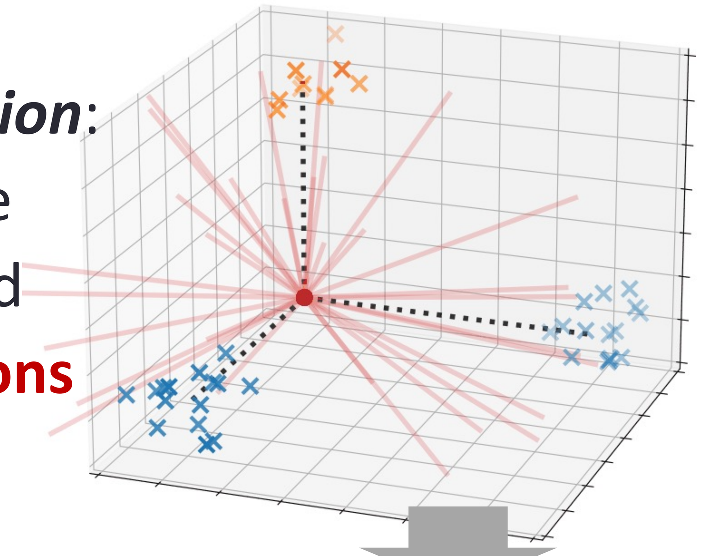
Gradient flow (GF) with small initialization:

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

GRADIENT FLOW LEARNS CLASS CENTERS (P=1) OR CLUSTER CENTERS (P>2)

Neurons visualized at initialization

Small Initialization: all neurons have **small norms** and **random directions**



x : Positive data $\{x_i: y_i = +1\}$

x : Negative data $\{x_i: y_i = -1\}$

⋯ : Cluster centers

o : Neurons $\{w_j\}$

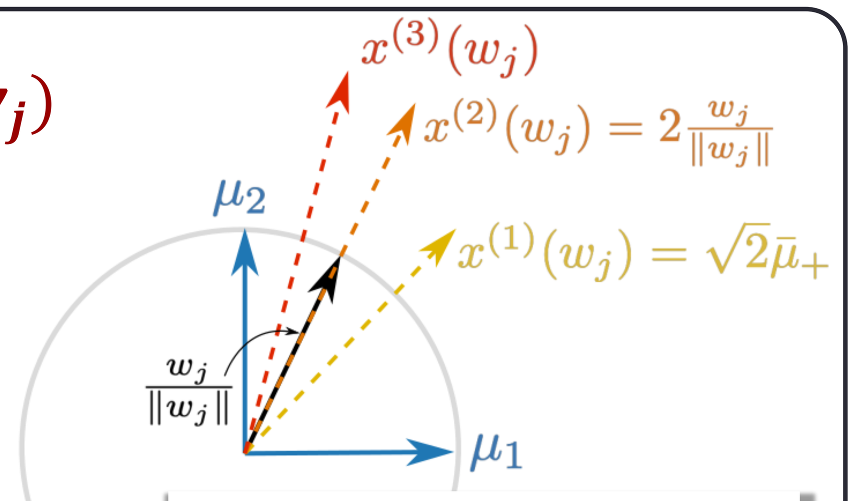
– : Neuron directions $\left\{\frac{w_j}{\|w_j\|}\right\}$

In initial GF training phase, **neuron w_j moves towards $x^{(p)}(w_j)$**

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

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

Depending the value of p, neurons learn different directions

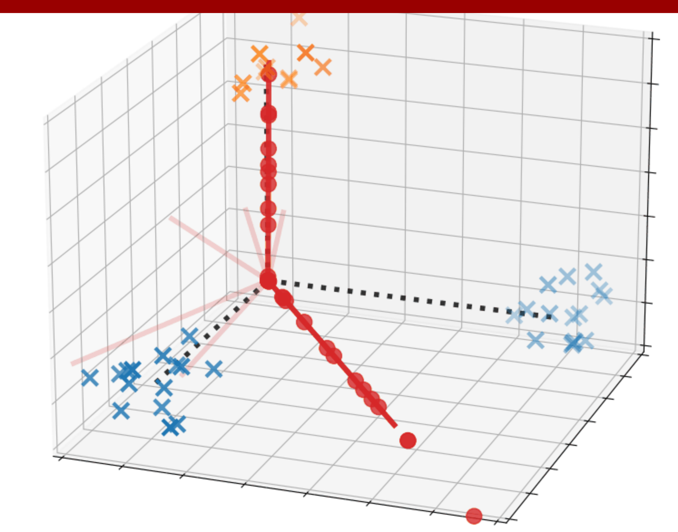


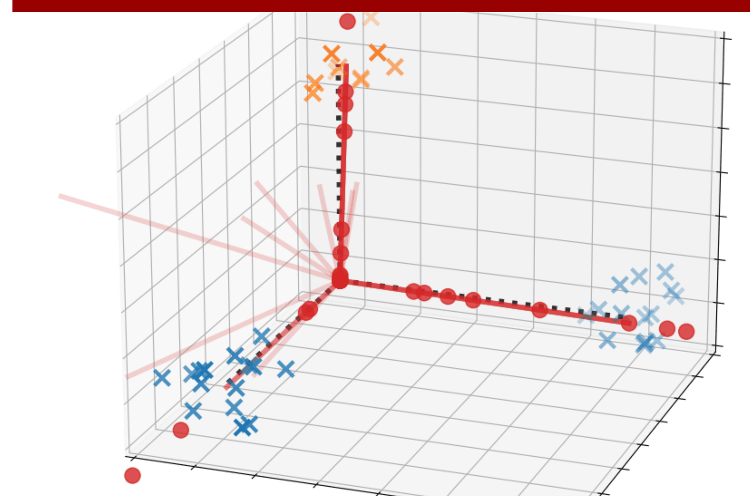
illustrative example: $\alpha = 0$, two +1 clusters

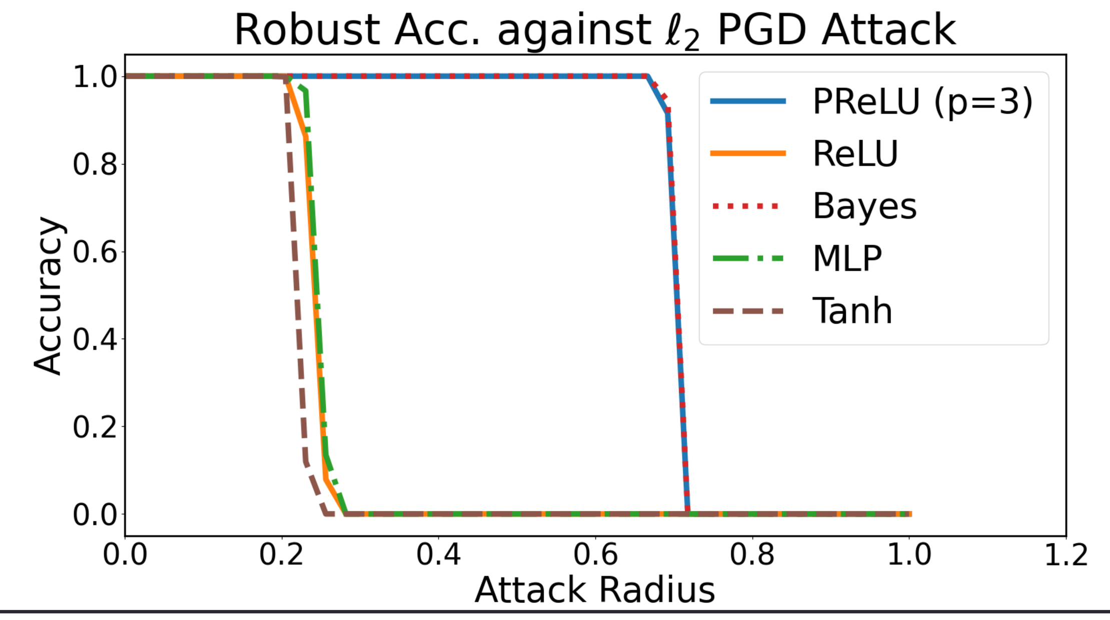
$p = 1$: ReLU net
Neurons learn class centers

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

Neurons visualized at the end of training







- ReLU network is more vulnerable to ℓ_2 adversarial attacks than pReLU network
- Vulnerability of ReLU network persists even: adding layers (MLP), or changing activations (Tanh)
- Carefully chosen activations (pReLU) needed

PROVABLE VULNERABILITY OF RELU (PRIOR WORKS)

[Frei et al., 2023]: Any limit point of GF/GD when training a ReLU network 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 PRELU (OUR WORK)

Convergence

pReLU network ($p > 2$) trained by GF with small init. and small α :

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

Robustness

$F^{(p)}(x)$ ($p > 2$) $\approx$ Bayes classifier $\Longrightarrow$ **Robust against $\mathcal{O}(1)$-attacks:**

$\forall \delta \in (0, \sqrt{2}]$, over new sample $(x, y) \in \mathbb{R}^D \times \{+1, -1\}$

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

Optimality

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

$\sqrt{2}/2$ is the maximum achievable ℓ_2-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