Adversarial Robustness is at Odds with Lazy Training

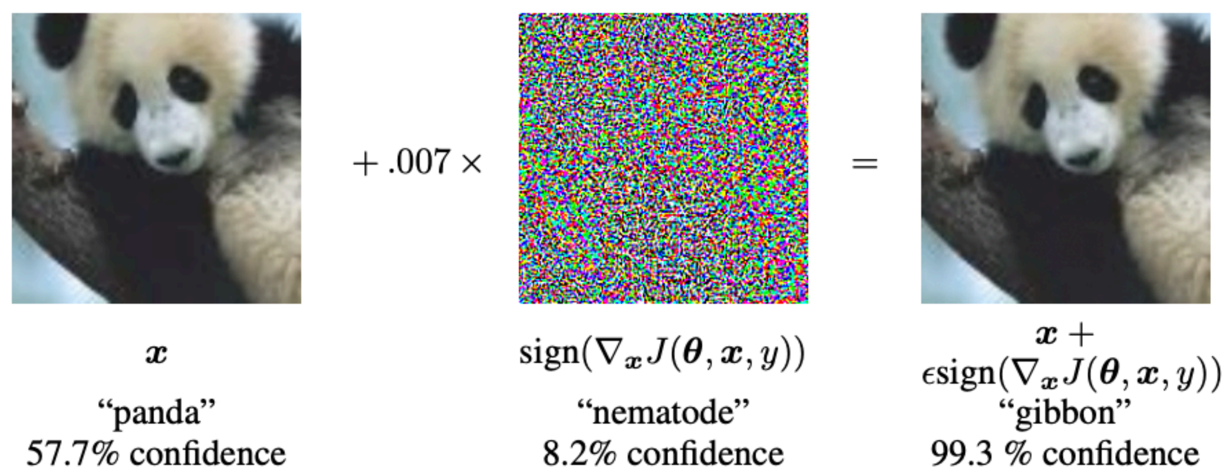
Yunjuan Wang, Enayat Ullah, Poorya Mianjy, Raman Arora
Department of Computer Science, Johns Hopkins University

JOHNS HOPKINS UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

1. Background

- ML systems are fragile and susceptible to imperceptible attacks [GSS15].



2. Motivation

[SSRD19]: Propose an algorithm to generate bounded ℓ_0-norm adversarial perturbation with guarantees for arbitrary deep networks.

[DS21]: Multi-step gradient ascent can find adversarial examples for random ReLU networks with small widths.

[BCGT21]: A single gradient step finds adversarial examples for sufficiently wide but not extremely wide randomly initialized ReLU networks.

[BBC21]: Extend the above to randomly initialized deep networks.

Question: why do trained neural networks remain vulnerable to adversarial attacks?

3. Problem Setup

- $\mathcal{X} \subseteq \mathbb{R}^d, \mathcal{Y} = \{\pm 1\}$.
- A two-layer width m ReLU net (a, W), $f(\mathrm{x}; \mathrm{a}, \mathrm{W}) := \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \sigma(\mathrm{w}_s^\top \mathrm{x})$.
- Attack model: ℓ_2 attack with perturbation budget R.

4. Lazy Regime

The dominant model for (non-robust) deep learning [JGH18, JT19, ADHL19].

Initialization: 1) $a_s \sim \mathrm{unif}(\{-1, +1\})$, fixed; 2) $\mathrm{w}_{s,0} \sim \mathcal{N}(0, \mathrm{I}_d), \forall s \in [m]$.

Why lazy regime?

1. **Provable generalization**: there exists $\overline{\mathrm{W}}$: $\|\bar{\mathrm{w}}_s - \mathrm{w}_{s,0}\|_2 = \mathcal{O}(1/\sqrt{m})$, $\forall s \in [m]$, such that the non-robust generalization error is small.
2. **Computational Tractability**: Such $\overline{\mathrm{W}}$ can be found by efficient first-order methods such as Stochastic Gradient Descent (SGD).

Definition: The lazy regime is the set of all networks parameterized by (a, W), such that

$$\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, C_0/\sqrt{m}\right) = \left\{\mathrm{W} : \|\mathrm{w}_s - \mathrm{w}_{s,0}\|_2 \leq C_0/\sqrt{m}, \forall s \in [m]\right\}.$$

Question: Are lazy regime networks susceptible to adversarial attacks?

5. Main Result

A single gradient step suffices to attack lazy regime networks.

Theorem: With probability at least $1-\gamma$, for all $\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, C_0/\sqrt{m}\right)$

$$\mathrm{sign}(f(\mathrm{x}; \mathrm{a}, \mathrm{W})) \neq \mathrm{sign}(f(\mathrm{x}+\delta; \mathrm{a}, \mathrm{W}))$$

where $\delta = \eta \nabla_{\mathrm{x}} f(\mathrm{x}; \mathrm{a}, \mathrm{W})$ with $|\eta| = \mathcal{O}(1/d)$,

$$\max\left\{d^{2.4}, \mathcal{O}\log\left(1/\gamma\right)\right\} \leq m \leq \mathcal{O}\left(\exp(d^{0.24})\right)$$

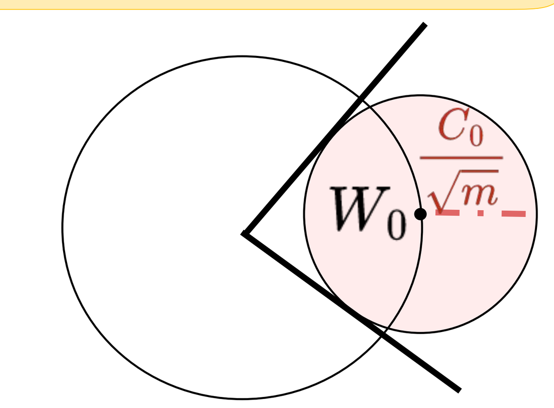
Remark: Imperceptible perturbation $\|\delta\| = \mathcal{O}(1/\sqrt{d})$.

Corollary: W.h.p., for all $\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, C_0/\sqrt{m}\right)$ the robust error is greater than 0.9.

6. Beyond Lazy Regime

- Scaling the network weights with a positive factor for binary classification does not change the prediction.
- Our results can be extended to the following cone

$$C = \left\{\mathrm{W} : \exists r > 0 : \|r\mathrm{W} - \mathrm{W}_0\|_{2,\infty} \leq C_0/\sqrt{m}\right\}$$



7. Experiments

- Binary MNIST. Networks trained using **SGD** in the lazy regime.
- Recall notation: $\mathbf{m}$: network width; $\|\delta\|$: L2 perturbation budget; η: step size;
  $\mathbf{V}$: maximal weight deviation $\max_{s\in[m]} \|\bar{\mathrm{w}}_s - \mathrm{w}_{s,0}\|_2 = C_0/\sqrt{m}$

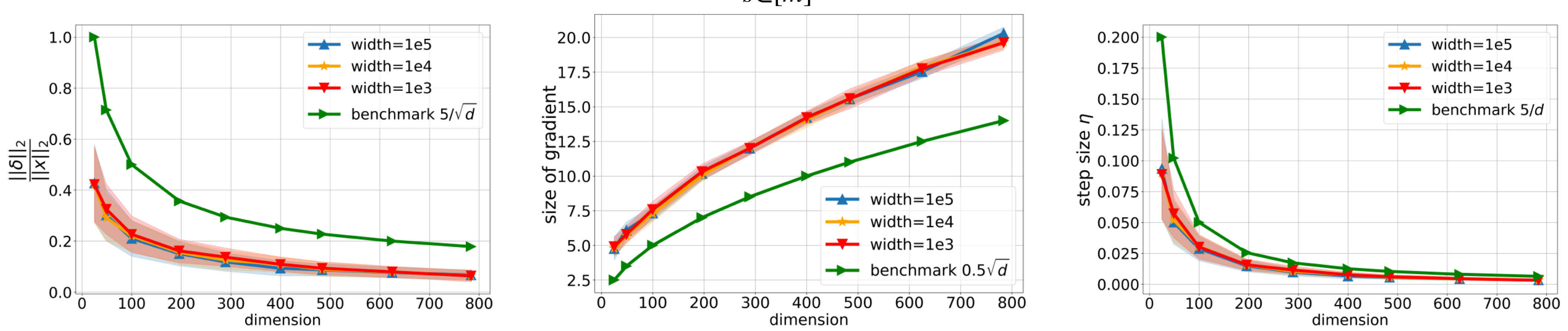


Figure: Minimal normalized $\|\delta\|$ needed to flip the predicted label **(left)**; the norm of the gradient $\|\nabla_{\mathrm{x}} f(\mathrm{x}; \mathrm{a}, \mathrm{W})\|$ for W in the lazy regime **(middle)**; corresponding step size **(right)**, as a function of the input dimension for a fixed value of $C_0 = 10$ and different network widths m.

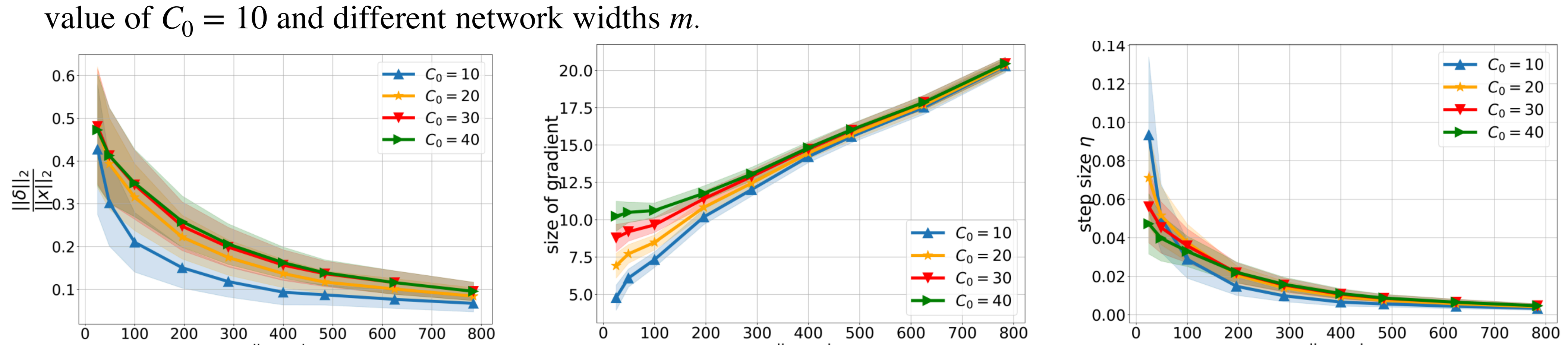


Figure: Minimal normalized $\|\delta\|$ needed to flip the predicted label **(left)**; the norm of the gradient $\|\nabla_{\mathrm{x}} f(\mathrm{x}; \mathrm{a}, \mathrm{W})\|$ for W in the lazy regime **(middle)**; corresponding step size **(right)**, as a function of the input dimension for a fixed network width of $m = 10^5$ and different value of C_0.

Observations:

- Experimental results **closely match** the theoretical bound for **different network widths** and **weight deviations**:

1. $\|\delta\| = \mathcal{O}(1/\sqrt{d})$; 2. $|\eta| = \mathcal{O}(1/d)$

Main Takeaway: Networks within the lazy training regime are vulnerable to adversarial attacks.

[SSRD19]: Shamir, Adi, et al. "A simple explanation for the existence of adversarial examples with small hamming distance." *arXiv preprint arXiv:1901.10861* (2019).

[DS21]: Daniely, Amit, and Hadas Schacham. "Most ReLU Networks Suffer from $\ell 2$ Adversarial Perturbations." *NeurIPS* (2020).

[BCGT21]: Bubeck, Sébastien, et al. "A single gradient step finds adversarial examples on random two-layers neural networks." *NeurIPS* (2021).

[BBC21]: Bartlett, Peter, Sébastien Bubeck, and Yeshwanth Cherapanamjeri. "Adversarial examples in multi-layer random relu networks." *NeurIPS* (2021).

[JT19]: Ji, Ziwei, and Matus Telgarsky. "Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks." *In ICLR* (2020).

- Binary MNIST. Networks trained using **adversarial training** in the lazy regime.

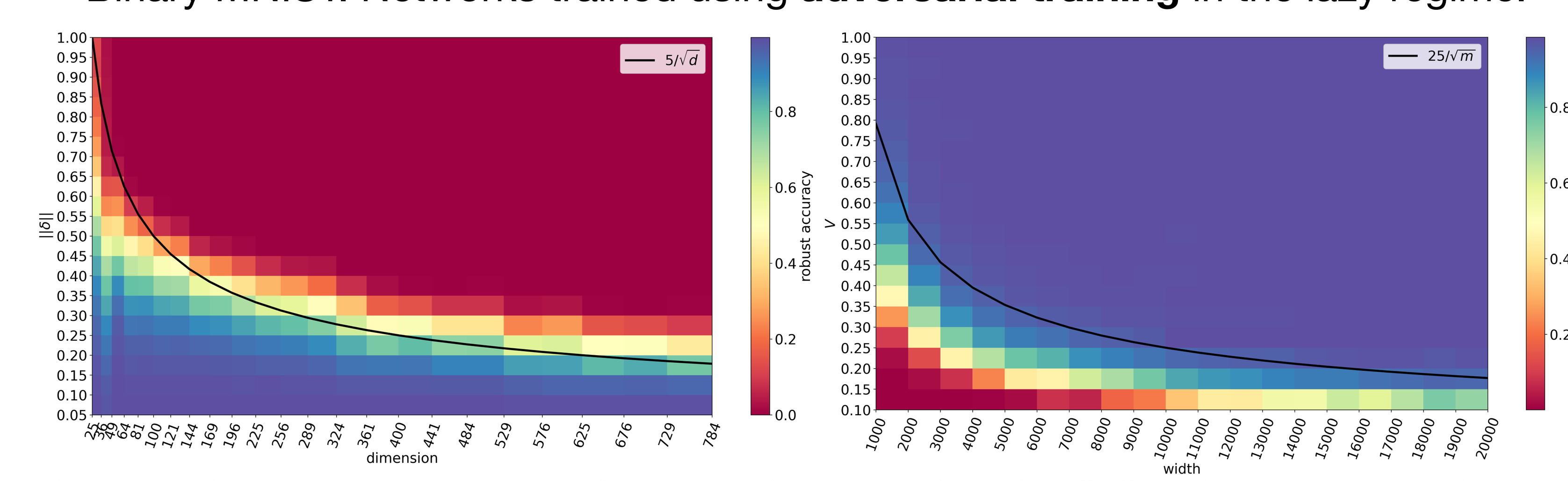


Figure: Robust test accuracy as a function of perturbation size $\|\delta\|$ and dimension d given a fixed network width and fixed value of the maximal deviation of weight vectors **(left)**; Robust test accuracy as a function of width m and maximal deviation V given a fixed dimension and fixed perturbation size **(right)**.

Observations:

1. a sharp drop in robust accuracy about the $\mathcal{O}(1/\sqrt{d})$ threshold for the perturbation budget $\|\delta\|$ as predicted by the theorem.
2. a phase transition in the robust test accuracy at the value of V around $\mathcal{O}(1/\sqrt{m})$ as required by the theorem.