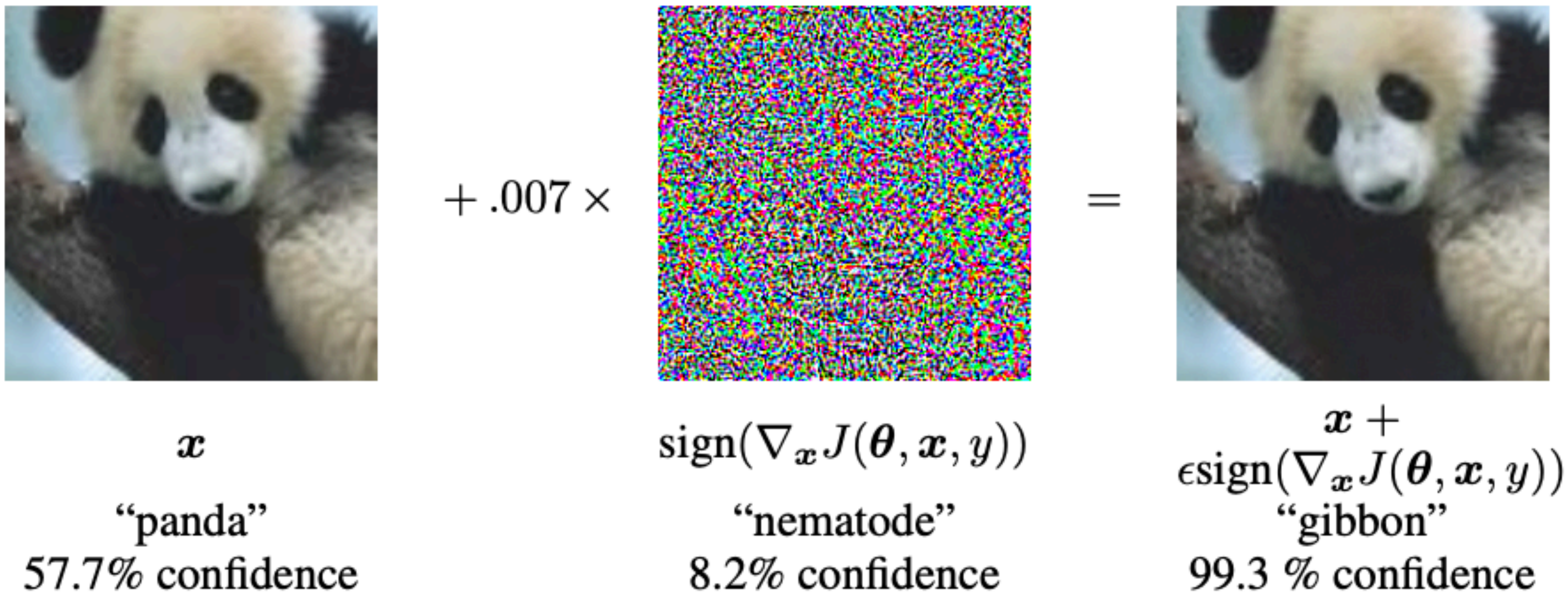
Adversarial Robustness is at Odds with Lazy Training

Yunjuan Wang, Enayat Ullah, Poorya Mianjy, Raman Arora
Johns Hopkins University

NeurIPS 2022

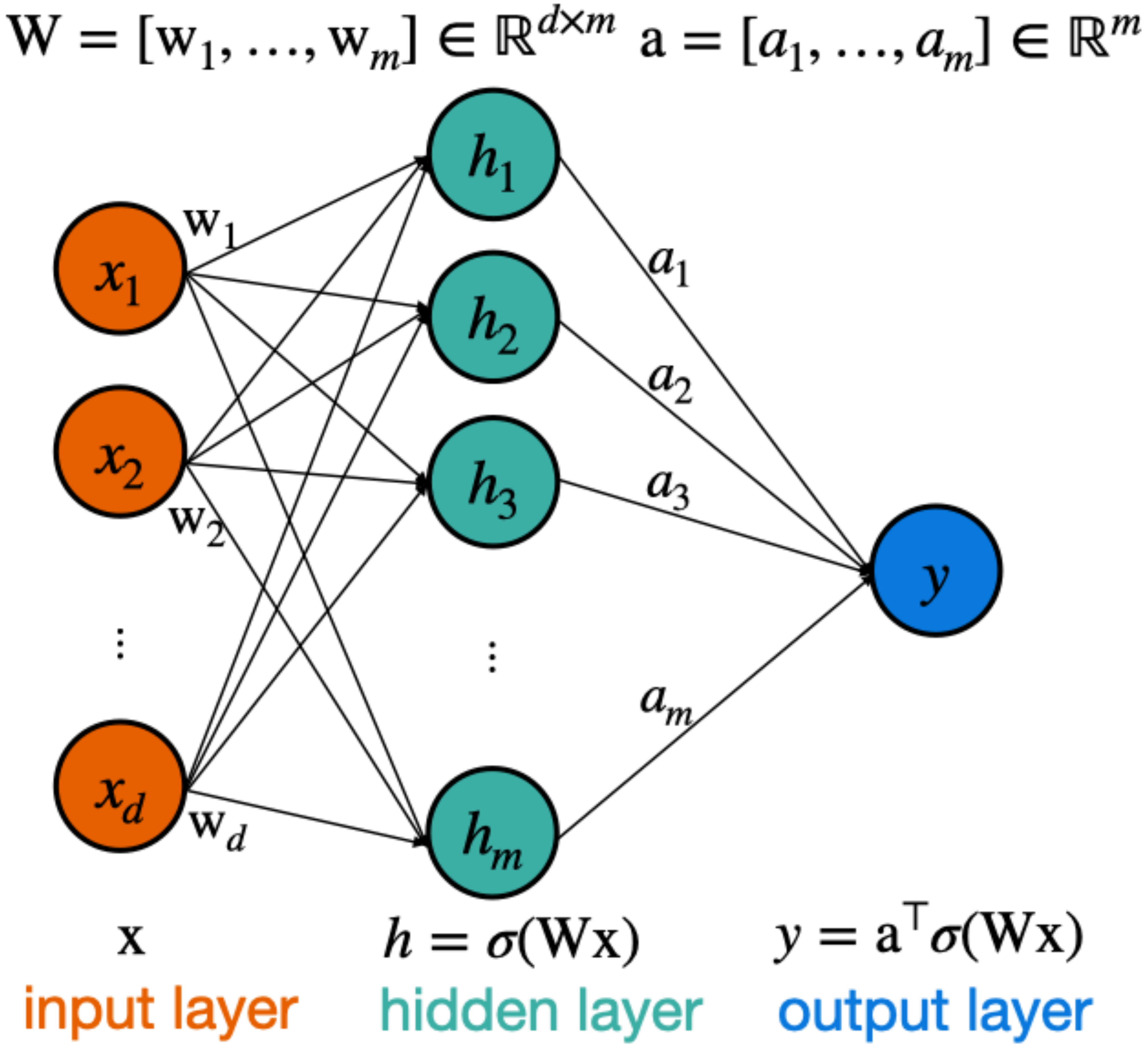
Motivation

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



Problem Setup

- $\mathcal{X} \subseteq \mathbb{R}^d, \mathcal{Y} = \{\pm 1\}$.
- A two-layer ReLU net parameterized by (a, W) $f(\mathrm{x}; \mathrm{a}, \mathrm{W}) := \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \sigma(\mathrm{w}_s^\top \mathrm{x}), \sigma(z)$ is ReLU.
- Attack model: ℓ_2 norm-bounded attack with perturbation budget R. $\mathrm{x}' \in \mathcal{B}_2(\mathrm{x}, R)$.



[SSRD19]: Propose an algorithm to generate bounded L0-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.

Lazy Training Regime

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

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

Key insights:

1. **Provable generalization**: there exists $\overline{\mathrm{W}} : \|\bar{\mathrm{w}}_s - \mathrm{w}_{s,0}\|_2 = \mathcal{O}\left(\frac{1}{\sqrt{m}}\right), \forall s \in [m]$ such that the 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 $\mathrm{W} \in \mathcal{B}_{2,\infty}\left(\mathrm{W}_0, \frac{C_0}{\sqrt{m}}\right) = \left\{\mathrm{W} : \|\mathrm{w}_s - \mathrm{w}_{s,0}\|_2 \leq \frac{C_0}{\sqrt{m}}, \forall s \in [m]\right\}$.

Question: Are networks in the lazy training regime susceptible to adversarial attacks?

Main Result

For any model in the lazy regime, a single step of gradient ascent on f suffices to find an adversarial example to flip the prediction sign.

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

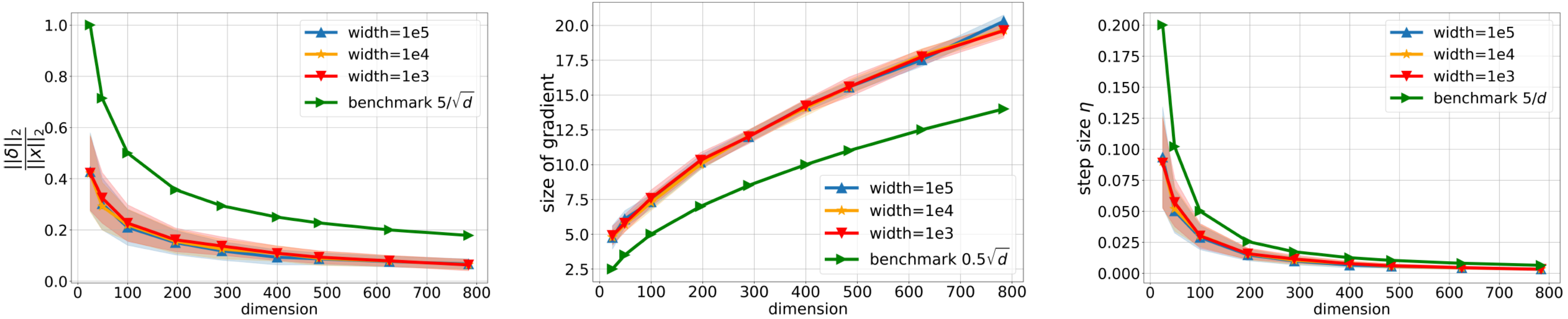
$$\mathrm{sign}(f(\mathrm{x};\mathrm{a},\mathrm{W})) \neq \mathrm{sign}(f(\mathrm{x}+\delta;\mathrm{a},\mathrm{W}))$$

where $\delta = \eta \nabla_{\mathrm{x}} f(\mathrm{x};\mathrm{a},\mathrm{W})$ with $|\eta| = \mathcal{O}(1/d)$,
$\max\left\{d^{2.4}, \mathcal{O}\left(\log\left(1/\gamma\right)\right)\right\} \leq m \leq \mathcal{O}\left(\exp(d^{0.24})\right)$.

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

Experiment

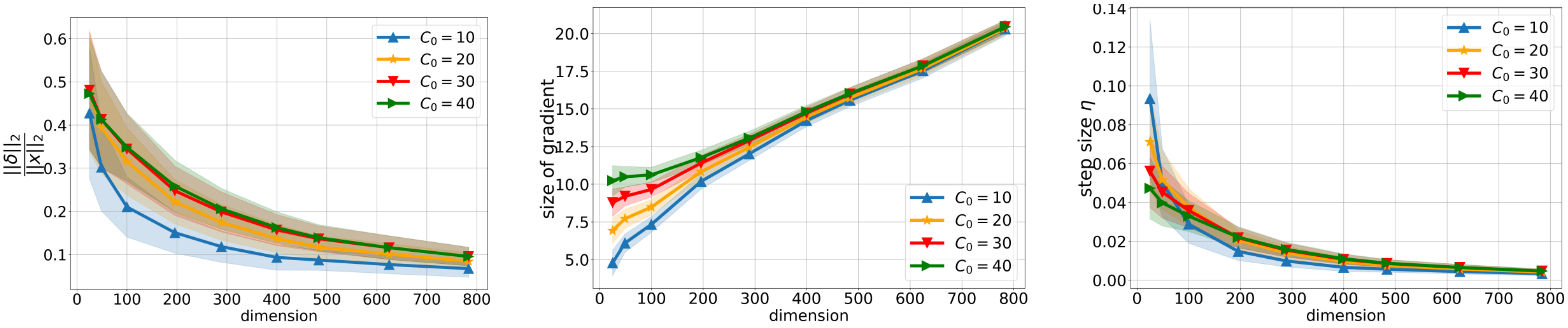
- Binary MNIST. Networks trained using **SGD** in the lazy regime



Main Takeaway: Our theoretical bound can be **tight** as experiments show: $\|\delta\| = \mathcal{O}(1/\sqrt{d})$ (left), $\|\nabla f_x(x; W)\| = \Omega(\sqrt{d})$ (middle), $|\eta| = \mathcal{O}(1/d)$ (right) for **different network widths**.

Experiment

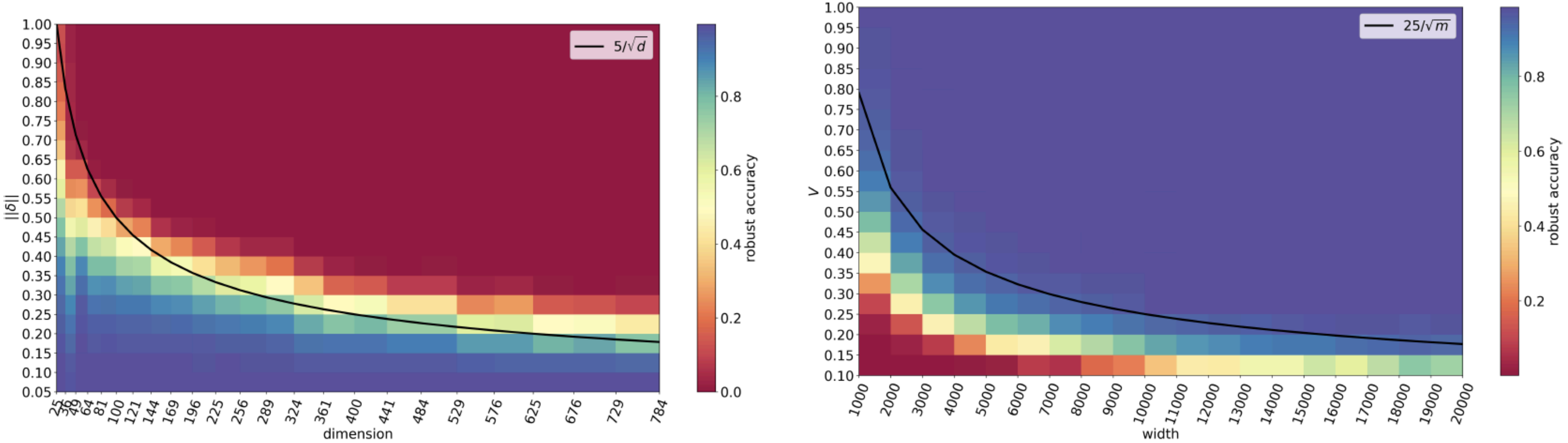
- Binary MNIST. Networks trained using **SGD** in the lazy regime



Main Takeaway: Our theoretical bound can be **tight** as experiments show: $||\delta|| = \mathcal{O}(1/\sqrt{d})$ (left), $||\nabla f_{\mathrm{x}}(\mathrm{x}; \mathrm{W})|| = \Omega(\sqrt{d})$ (middle), $|\eta| = \mathcal{O}(1/d)$ (right) for **different weight deviations.**

Experiment

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



Main Takeaway: a sharp drop in robust accuracy about the $\mathcal{O}(1/\sqrt{d})$ threshold for the perturbation budget $||\delta||$ as predicted by the main theorem (left); a phase transition in the robust test accuracy for maximal weight deviation V around $\mathcal{O}(1/\sqrt{m})$ as required by the main theorem (right).

Conclusion

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

Future directions:

1. Extend to multi-layer networks.
2. Consider stronger attacks, i.e. gradient ascent-based attack that is run to convergence.
3. Understand the relationship between the width, the input dimension, maximal weight deviation from the initialization, and robust accuracy.

Thanks!

Reference

[GSS15]: Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *In 3rd International Conference on Learning Representations* (2015).

[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." *Advances in Neural Information Processing Systems* (2020).

[BCGT21]: Bubeck, Sébastien, et al. "A single gradient step finds adversarial examples on random two-layers neural networks." *Advances in Neural Information Processing Systems* 34 (2021): 10081-10091.

[BBC21]: Bartlett, Peter, Sébastien Bubeck, and Yeshwanth Cherapanamjeri. "Adversarial examples in multi-layer random relu networks." *Advances in Neural Information Processing Systems* 34 (2021): 9241-9252.

[JGH18]: Jacot, Arthur, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks." *Advances in neural information processing systems* 31 (2018).

[JT19]: Ji, Ziwei, and Matus Telgarsky. "Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks." *In 8th International Conference on Learning Representations* (2020).

[ADHL19]: Arora, Sanjeev, et al. "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks." *International Conference on Machine Learning*. PMLR, 2019.