Stability and Generalization of Adversarial Training for Shallow Neural Networks with Smooth Activation

Kaibo Zhang, Yunjuan Wang, Raman Arora
Johns Hopkins University

NeurIPS 2024

Motivation

- Neural networks: highly vulnerable to adversarial attacks
- How to learn robust models? standard loss → robust loss
- Practical approach: Adversarial Training [1]
- Generalization guarantees of robust learning: uniform convergence [2] & algorithmic stability [3]
- Previous works focus on analyzing the stability for convex loss [4] or general non-convex loss [3]

Question *Can we give better stability guarantees of adversarial training for neural networks——a special instance of non-convex loss?*

[1] Madry et al. (2018). “Towards deep learning models resistant to adversarial attack.” In: International Conference on Learning Representations.

[2] Xiao et al (2023). “PAC-Bayesian adversarially robust generalization bounds for deep neural networks.”. In: The Second Workshop on New Frontiers in Adversarial Machine Learning.

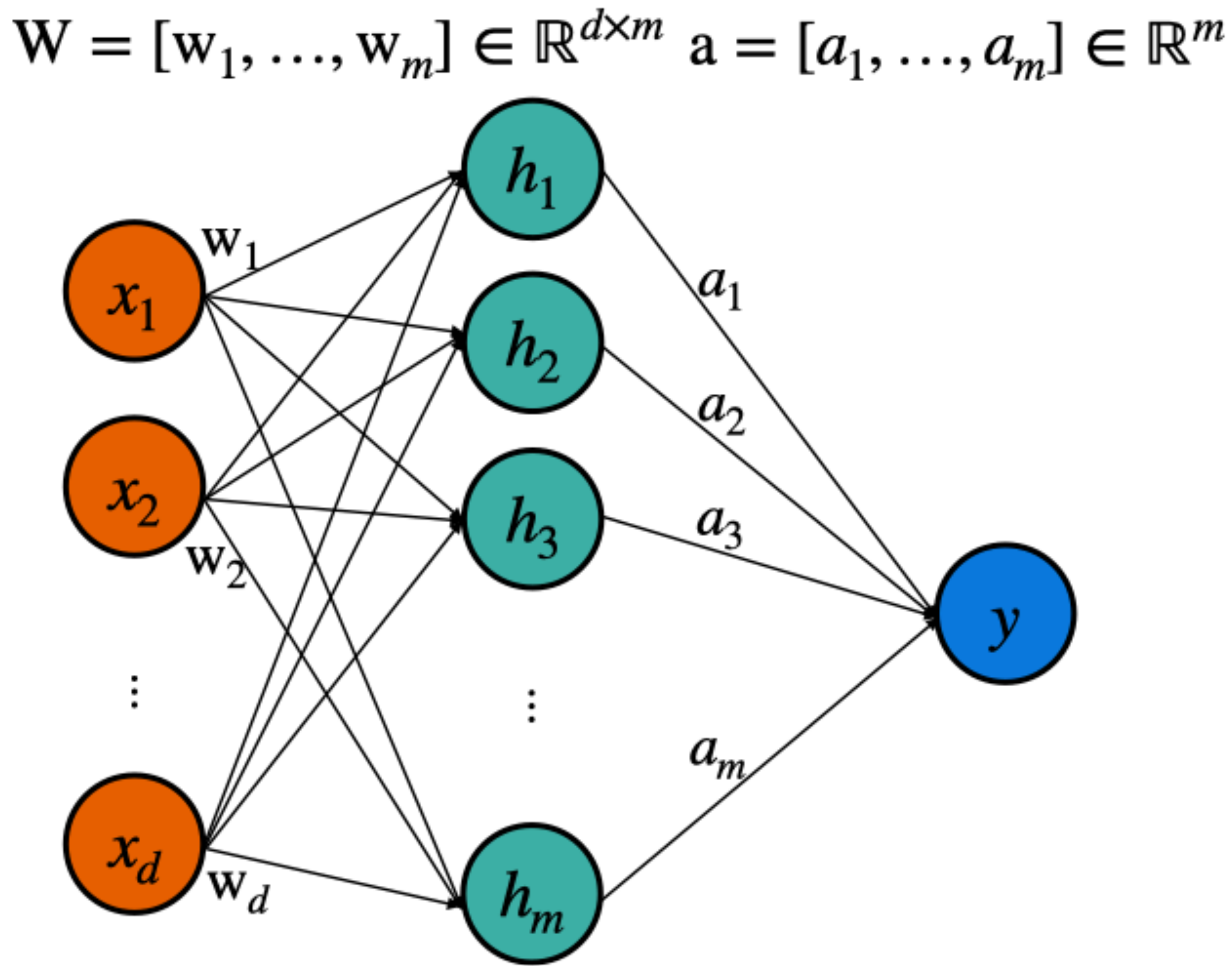
[3] Xiao et al. (2022). “Stability analysis and generalization bounds of adversarial training.” In: Advances in Neural Information Processing Systems.

[4] Xing et al. (2021). “On the algorithmic stability of adversarial training.” In: Advances in Neural Information Processing Systems.

Two-layer neural networks

- Two-layer network parameterized by (a, W): $f_W(x) = \sum_{s=1}^{m} a_s \phi(\langle w_s, x \rangle)$.
- m: number of hidden units/width
- $\phi(x)$: H-smooth and Lipschitz activation
- $|a_s| = \dfrac{1}{\sqrt{m}}$ kept fixed throughout training
- No restriction on the initialization of weight W_0
- Binary classification: input $\|x\|_2 \leq C_x$, label $y \in \{\pm 1\}$
- Loss function: logistic loss $\ell(z) = \ln(1 + e^{-z})$

—— smooth and Lipschitz



Adversarial Training

- General attack model: $B(x) \supseteq \{x\}$ Special instance: $B(x) = \{\tilde{x} \mid \|\tilde{x} - x\| \leq \alpha\}$
- Robust loss: $\ell_{\mathrm{rob}}\left(W; (x, y)\right) = \sup_{\tilde{x} \in B(x)} \ell\left(W; (\tilde{x}, y)\right)$
- Adversarial Training: $\min_{W} \hat{L}_{\mathrm{rob}}(W) = \frac{1}{n} \sum_{(x,y) \in S} \ell_{\mathrm{rob}}\left(W; (x, y)\right)$

  Step 1: for $(x, y) \in S$, generate $\tilde{x} \in B(x)$, s.t. $\ell\left(W; (\tilde{x}, y)\right) \geq \ell_{\mathrm{rob}}\left(W; (x, y)\right) - \beta$

  Step 2: do one step gradient descent for $\sum \ell\left(W; (\tilde{x}, y)\right)$, and go back to step 1
- Remark: β captures the precision of the attack algorithm
- Difficulty: the robust loss $\hat{L}_{\mathrm{rob}}(W)$ is non-convex and non-smooth

  ——hard to establish computational guarantees for a non-convex and non-smooth loss

Main Result: Guarantees of Adversarial Training

Theorem (informal): If width $m \geq O(\eta^2 T^2)$,

1. generalization guarantee:

$$\mathbb{E} L_{\text{rob}}(W_T) \leq \frac{1}{1 - O\left(\eta\sqrt{T} + \frac{\eta T}{n} + \sqrt{\beta \eta T}\right)} \mathbb{E} \hat{L}_{\text{rob}}(W_T).$$

2. Optimization guarantee:

$$\min_{0 \leq t \leq T} \hat{L}_{\text{rob}}(W_t) \leq \min_{W} \left(\hat{L}_{\text{rob}}(W) + \frac{2}{\eta T}\|W - W_0\|_F^2\right) + O(\eta)$$

Remark 1: A small β and $\eta T \ll \min\{\sqrt{m}, n\}$ suffices to guarantee a small generalization gap

Remark 2: $\eta T \ll \min\{\sqrt{m}, n\}$ can be viewed as early stopping

Remark 3: A very small learning rate η is required for $\eta\sqrt{T}$ to be small

Technical Insights

1. ***Uniform Argument Stability***

- Stability captures the difference in outputs, if the inputs differ in one example

——for neighboring S_1, S_2, $\delta_{\mathscr{A}}(S_1, S_2) = \|\mathscr{A}(S_1) - \mathscr{A}(S_2)\|_F$

- Better stability gives better generalization

$$\mathbb{E}L_{\mathrm{rob}}(\mathscr{A}(S)) \leq \frac{1}{1 - C_x \sup_{S_1 \simeq S_2} \delta_{\mathscr{A}}(S_1, S_2)} \mathbb{E}\hat{L}_{\mathrm{rob}}(\mathscr{A}(S))$$

- For neighboring S_1, S_2, if width $m \geq O(\eta^2 T^2)$,

$$\delta_{\mathscr{A}}(S_1, S_2) = O\left(\eta\sqrt{T} + \frac{\eta T}{n} + \sqrt{\beta\eta T}\right)$$

Technical Insights

2. *Weakly Convex Robust Loss*

- $f(x)$ is called $-l$-weakly convex, if $f(x) + \frac{l}{2}||x||_2^2$ is convex
- If $l \approx 0$, then $f(x)$ is approximately convex
- $\hat{L}_{\text{rob}}(W)$ is $-\frac{HC_x^2}{\sqrt{m}}$-weakly convex
- If width $m \geq O(\eta^2 T^2)$, $\hat{L}_{\text{rob}}(W)$ behaves similarly as a convex loss

——stability is established based on the weakly convex property

Improvement: Smoothing Using Moreau Envelope

$\eta\sqrt{T}$ in stability upper bound arises due to non-smoothness of robust loss.

Question: Can we remove this term?

- For $\mu < O(\sqrt{m})$, define Moreau Envelope $M^{\mu}(W) = \min_{U}\left(\hat{L}_{\mathrm{rob}}(U) + \frac{\|U - W\|_F^2}{2\mu}\right)$

- $M^{\mu}(W)$ is smooth, and it has the same global minimizer as $\hat{L}_{\mathrm{rob}}(W)$

- $\hat{L}_{\mathrm{rob}}(W) - O(\mu) \leq M^{\mu}(W) \leq \hat{L}_{\mathrm{rob}}(W)$
- Doing gradient descent on $M^{\mu}(W)$ guarantees:

for neighboring S_1, S_2, if $m \geq O(\eta^2 T^2)$, $\eta \leq \mu$, $\delta_{\mathscr{A}}(S_1, S_2) = O\left(\frac{\eta T}{n}\right)$