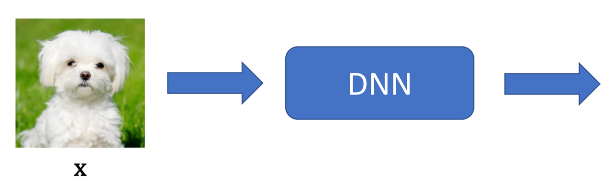
On the Convergence of Certified Robust Training with Interval Bound Propagation

Yihan Wang*, Zhouxing Shi*, Quanquan Gu, Cho-Jui Hsieh *Equal contribution

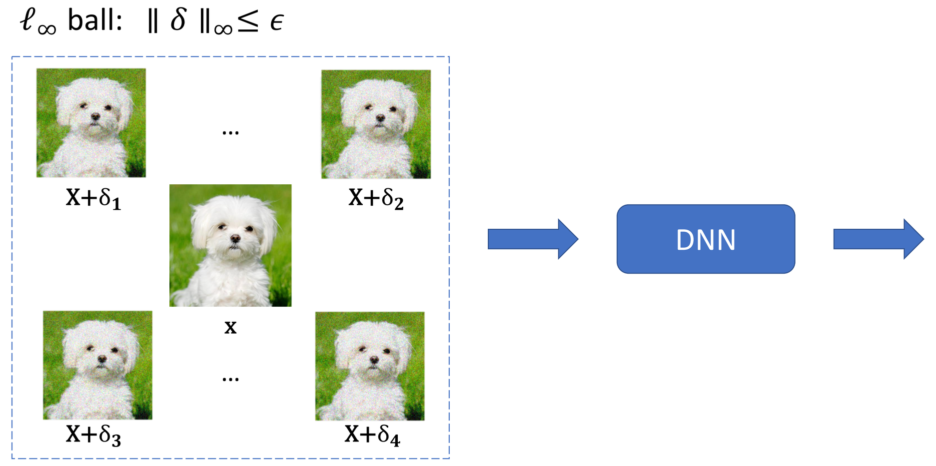
University of California, Los Angeles

Certified Robustness and Certified Robust Training



Logits

-5.889	airplane
0.1203	automobile
-4.2943	bird
1.3597	cat
0.1594	deer
6.032	**dog**
-6.2416	frog
-2.878	horse
-1.4488	ship
-9.332	truck



Worst-case logits

-0.889	airplane ↑
0.7203	automobile ↑
-0.2943	bird ↑
4.3597	cat ↑
1.1594	deer ↑
2.032	**dog ↓**
0.2416	frog ↑
-0.878	horse ↑
1.4488	ship ↑
-1.332	truck ↑

UNSAFE

Certified Robustness

- Certified Robustness studies whether a model is provably safe given a perturbation set.
- Safe if the score of the ground-truth label is provably larger than all other classes in the worst-case logits under perturbation.

Certified Robust Training

- Minimize the upper bound of worst-case loss to improve the certified robustness of models:

$$\min_{\theta} \overline{L}(f_\theta, \mathbf{x}, y, \epsilon), \quad \text{where } \overline{L}(f_\theta, \mathbf{x}, y, \epsilon) \geq \max_{\|\theta\|_\infty \leq \epsilon} L(f_\theta, \mathbf{x} + \delta, y).$$

Interval Bound Propagation Training (Mirman et al., 2018; Gowal et al., 2018)

- A simple but efficient method for computing the output bounds of neural networks.
- It computes the interval lower and upper bounds for each neuron and propagates bounds across layers.

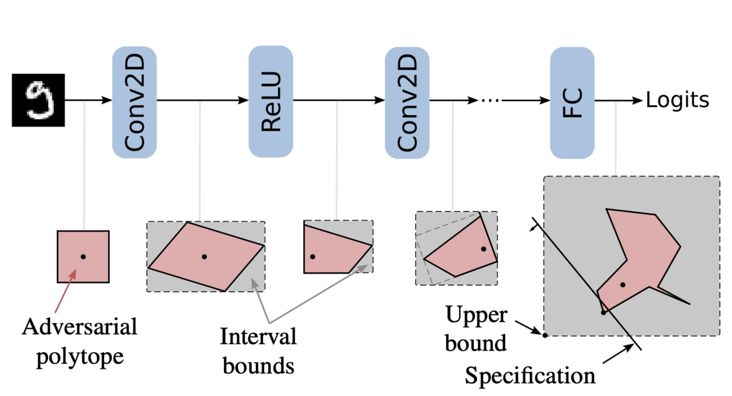


Figure 1. Illustration of IBP, from Gowal et al., 2018.

Problem Settings

- (Du et al, 2019a;b) proved that on randomly initialized and overparameterized two-layer neural networks for standard training, SGD is guaranteed to converge to zero training error with high probability.
- But IBP training has a different training scheme compared to standard training and is often hard to achieve low errors in practice.
- We aim to theoretically analyze the convergence of IBP training under SGD.

Data

- Training set $\{(\mathbf{x}_i, y_i) : i \in [n]\}$.
- $\forall i \in [n]$, $\mathbf{x}_i \in [\epsilon, 1]^d$, $\|\mathbf{x}_i\|_2 \geq \xi > 0$.
- For perturbation radius ϵ,

$$\forall i, j \in [n], i \neq j, \forall \mathbf{x}'_i \in B_\infty(\mathbf{x}_i, \epsilon), \forall \mathbf{x}'_j \in B_\infty(\mathbf{x}_j, \epsilon), \quad \mathbf{x}'_i \not\parallel \mathbf{x}'_j,$$

where $B_\infty(\mathbf{x}_i, \epsilon)$ stands for the ℓ_∞-ball with radius ϵ centered at $\mathbf{x}_i$.

Model and Loss Function

- A two-layer neural network for a binary classification task:

$$f(\mathbf{W}, \mathbf{a}, \mathbf{x}_i) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(\mathbf{w}_r^\top \mathbf{x}_i),$$

with standard logistic loss:

$$L = \sum_{i=1}^{n} l(y_i f(\mathbf{W}, \mathbf{a}, \mathbf{x}_i)) = \sum_{i=1}^{n} \log\left(1 + \exp(-u_i(\mathbf{W}, \mathbf{a}, \mathbf{x}_i))\right).$$

$$\text{where } u_i(\mathbf{W}, \mathbf{a}, \mathbf{x}_i) = y_i f(\mathbf{W}, \mathbf{a}, \mathbf{x}_i).$$

- Upper bound of worst-case loss (IBP loss) $\overline{L}$ under perturbation radius ϵ:

$$\overline{L} \geq \sum_{i=1}^{n} \max_{\Delta_i} \left\{ \log\left(1 + \exp(-y_i f(\mathbf{W}, \mathbf{a}, \mathbf{x}_i + \Delta_i))\right) \;\middle|\; \|\Delta_i\|_\infty \leq \epsilon \right\}.$$

$$\overline{L} = \sum_{i=1}^{n} \log(1 + \exp(-\underline{u}_i)), \underline{u}_i = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \Big\{ \mathbb{1}(y_i a_r = 1)\sigma\left(\mathbf{w}_r^\top \mathbf{x}_i - \epsilon\|\mathbf{w}_r\|_1\right) + \mathbb{1}(y_i a_r = -1)\sigma\left(\mathbf{w}_r^\top \mathbf{x}_i + \epsilon\|\mathbf{w}_r\|_1\right) \Big\}.$$

Gradient Flow

- We consider gradient flow – gradient descent with infinitesimal step size, where

$$\forall r \in [m], \quad \frac{d\mathbf{w}_r(t)}{dt} = -\frac{\partial \overline{L}(t)}{\partial \mathbf{w}_r(t)}.$$

Main Results

Main Theorem

Suppose the assumptions hold for the training data, and the ℓ_∞ perturbation radius satisfies $\epsilon \leq O\left(\min\left(\frac{\delta^2\lambda_0^2}{d^{2.5}n^3}, \frac{\sqrt{2d}R}{\log(\sqrt{\frac{2\pi d}{R}}\xi)}\right)\right)$, where $R = \frac{c\delta\lambda_0}{d^{1.5}n^2}$, $c = \frac{\sqrt{2\pi}\xi}{384}$. For a two-layer ReLU network, suppose its width for the first hidden layer satisfies $m \geq \Omega\left(\left(\frac{d^{1.5}n^4\delta\lambda_0}{\delta^2\lambda_0^2 - \epsilon d^{2.5}n^4}\right)^2\right)$, and the network is randomly initialized as $a_r \sim unif[\{1, -1\}], \mathbf{w}_r \sim \mathbf{N}(0, \mathbf{I})$, with the second layer fixed during training. Then for any confidence $\delta(0 < \delta < 1)$, with probability at least $1 - \delta$, IBP training with gradient flow can converge to zero training error.

Implications

- For a given ϵ, as long as it satisfies an upper bound on ϵ which is dependent on the training dataset, with a sufficiently large width m, convergence of IBP training is guaranteed with high probability.
- When ϵ is larger than the upper bound, IBP training is not guaranteed to converge under our analysis even with arbitrarily large m, which is essentially different from analysis on standard training and implies a possible limitation of IBP training.

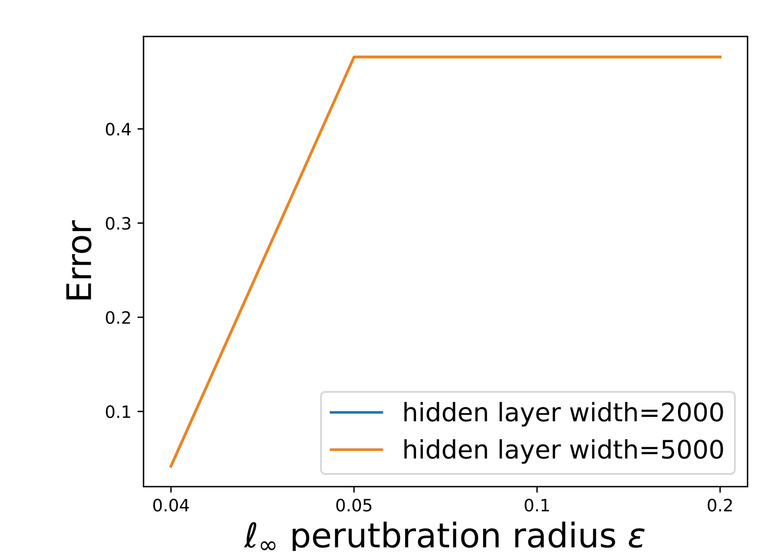
Proof Summary

To prove this theorem:

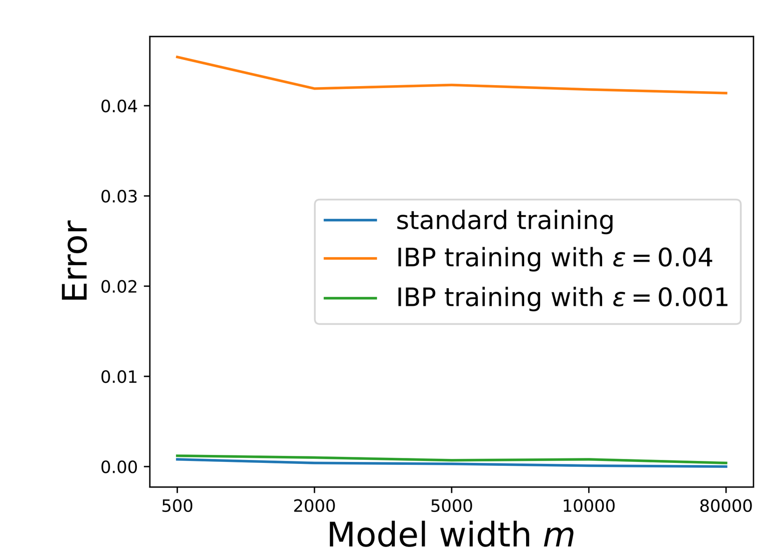
- We first analyze the stability of Gram matrix $\mathbf{H}_{ij}(t) = \sum_{r=1}^{m} \left\langle \frac{\partial \underline{u}_i(t)}{\partial \mathbf{w}_r(t)}, \frac{\partial \underline{u}_j(t)}{\partial \mathbf{w}_r(t)} \right\rangle$ during IBP training, and we show that $\lambda_{\min}(\mathbf{H}(t))$ remains positive with high probability.
- When $\mathbf{H}(t)$ remains positive definite, IBP loss descends in a linear convergence rate.
- We then reach constraints on ϵ and requirement on network width m to guarantee the convergence of IBP training.

Experiments

- MNIST 2 v.s. 5 binary classification.
- Compared to standard training, for the same width m, IBP has higher training errors.
- For relatively large ϵ ($\epsilon = 0.04$), even if we enlarge m up to 80,000 limited by the memory of a single GPU, IBP error remains high.



(a) Final training error of standard training and IBP.



(b) Final training error of IBP training.

Contact: { yihanwang, zshi, qgu, chohsieh }@cs.ucla.edu