Understanding Certified Training with Interval Bound Propagation

Yuhao Mao, Mark Niklas Müller, Marc Fischer, Martin Vechev

Department of Computer Science

ETH zürich SRILAB

We theoretically analyze Interval Bound Propagation (IBP) in Certified Training:

- introduce a novel metric quantifying propagation tightness (PT)
- show that IBP training increases PT
- find that PT regularizes weight signs
- empirically confirm our theoretical analysis

Network Certification with Interval Bound Propagation (IBP)

Robustness: $f(x')_{i^*} - f(x')_i \geq 0, \forall i, x'$ s.t. $\|x' - x\|_\infty \leq \epsilon$.

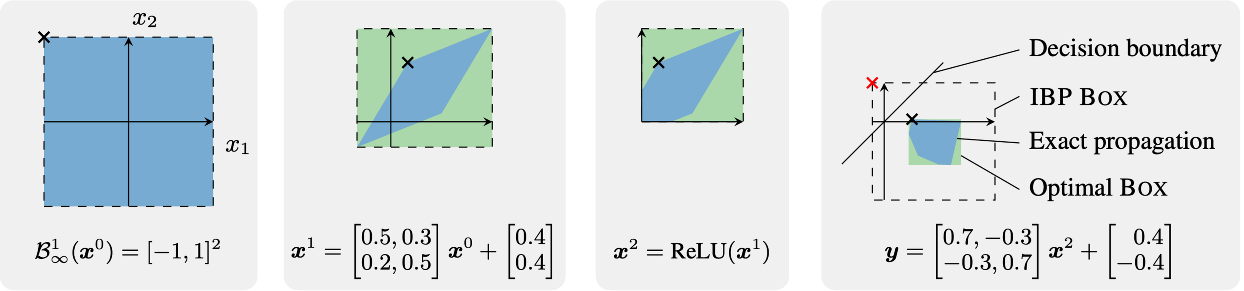
Interval Bound Propagation (IBP): compute output bounds layer-wisely, e.g., $[a, b] + [c, d] = [a + c, b + d]$.

Layer-wise Approximation $\text{Box}^\dagger(f, B^\epsilon(x)) = [\underline{z}^\dagger, \bar{z}^\dagger]$: apply optimal approximation layer-wisely, i.e., IBP approximation.

Optimal Approximation $\text{Box}^*(f, B^\epsilon(x))$: smallest hyper-box $[\underline{z}^*, \bar{z}^*]$ such that $f(x') \in [\underline{z}^*, \bar{z}^*], \forall x' \in B^\epsilon(x)$.

Propagation Invariance: a network is propagation invariant if $\text{Box}^\dagger(f, B^\epsilon(x)) = \text{Box}^*(f, B^\epsilon(x))$, i.e., IBP is exact.

Propagation Tightness: $\tau = (\underline{z}^* - \bar{z}^*)/(\bar{z}^\dagger - \underline{z}^\dagger)$, i.e., the ratio of optimal and layer-wise box sizes.



Explicit IBP for Deep Linear Network (DLN)

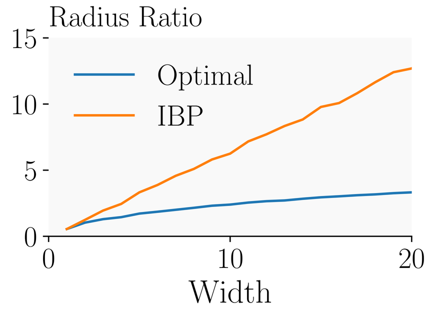
- For DLN $f = \prod_{k=1}^{L} W^{(k)}$, the size of approximations are: $\bar{z}^* - \underline{z}^* = 2\left|\Pi_{k=1}^{L} W^{(k)}\right|\epsilon$ and $\bar{z}^\dagger - \underline{z}^\dagger = 2\left(\Pi_{k=1}^{L}\left|W^{(k)}\right|\right)\epsilon$.
- DLN with all non-negative weights is propagation invariant.

Propagation Invariance

- A two-layer DLN $f = W^{(2)}W^{(1)}$ is propagation invariant if and only if $W^{(2)}_{i,k} \cdot W^{(1)}_{k,j} \geq 0$ for all k or $W^{(2)}_{i,k} \cdot W^{(1)}_{k,j} \leq 0$ for all k.
- A two-layer DLN $f = W^{(2)}W^{(1)}$ is not propagation invariant if $\left(W^{(2)}W^{(1)}\right)_{i,j}\left(W^{(2)}W^{(1)}\right)_{i,j'}\left(W^{(2)}W^{(1)}\right)_{i',j}\left(W^{(2)}W^{(1)}\right)_{i',j'} < 0$ for some i, j.
- $W^{(2)}W^{(1)} = \begin{pmatrix} 1 & 2 \\ -3 & 4 \end{pmatrix}$ → not propagation invariant.

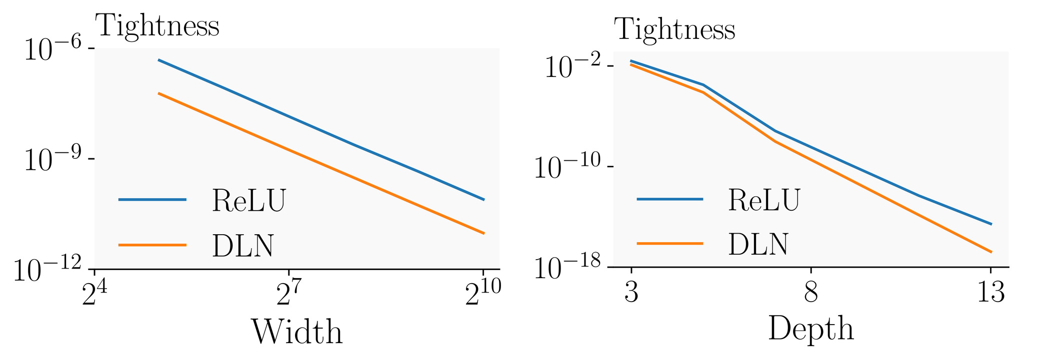
Box Reconstruction Error

For linearly separable data, PCA (optimal) weights lead to linear growth of layer-wise box size and sqrt growth of optimal box size.



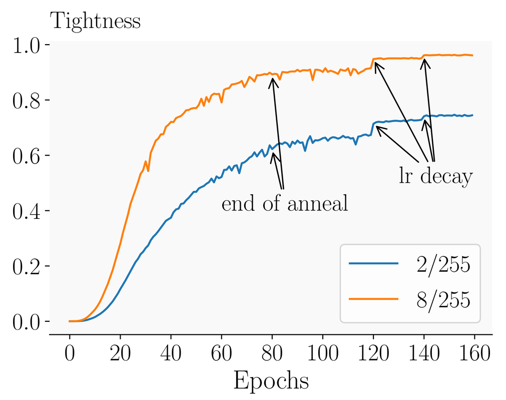
Tightness at Initialization

- For two-layer DLN with weights sampled from i.i.d. Gaussian distribution and hidden dimension d, tightness decreases in squared root order of d: $\tau = \Theta(d^{-1/2})$.
- For L-layer DLN randomly initialized with i.i.d. Gaussian and minimum hidden dimension d, tightness decreases in exponential order of L: $\tau = O(d^{-\lfloor L/4 \rfloor})$.



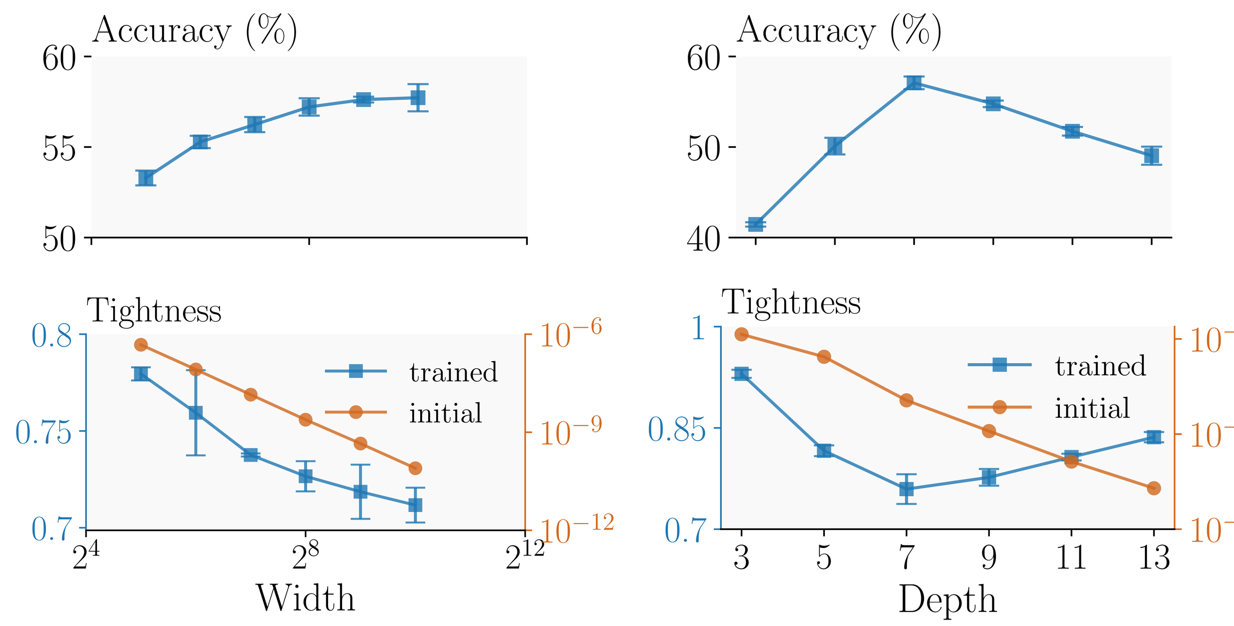
IBP Increases Tightness

If $\text{Box}^\dagger(f, B^\epsilon(x))$ deviates too much from $\text{Box}^*(f, B^\epsilon(x))$, then the gradient difference between IBP and standard loss is aligned with an increase in tightness, i.e., IBP-trained models have larger tightness.

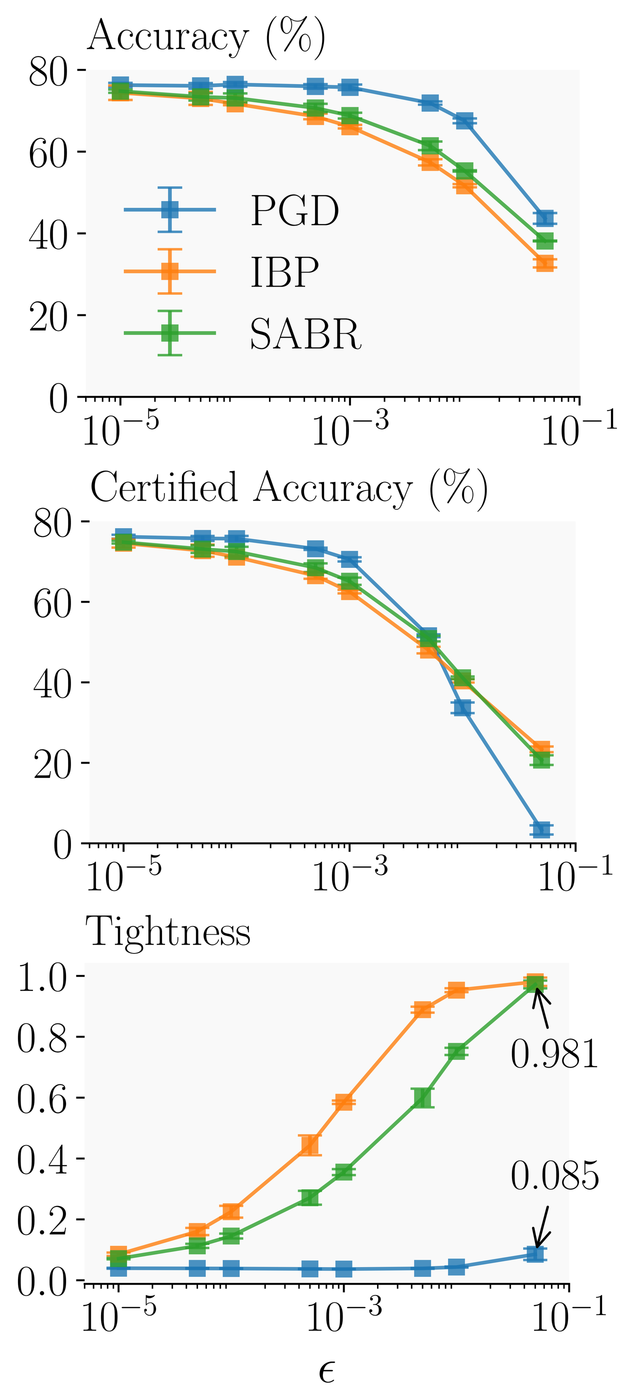


Results for ReLU networks

IBP Training w.r.t. Network Width and Depth



Effect of Input Radius on Accuracies and Tightness



Width-scale Rule Predicts Better Models

Dataset	ϵ	Method	Width	Accuracy	Certified
MNIST	0.1	IBP	1×	98.83	98.10
			4×	98.86	98.23
		SABR	1×	98.99	98.20
			4×	**98.99**	**98.32**
	0.3	IBP	1×	97.44	93.26
			4×	97.66	93.35
		SABR	1×	**98.82**	93.38
			4×	98.48	**93.85**
CIFAR-10	$\frac{2}{255}$	IBP	1×	67.93	55.85
			2×	68.06	56.18
		IBP-R	1×	78.43	60.87
			2×	**80.46**	62.03
		SABR	1×	79.24	62.84
			2×	79.89	**63.28**
	$\frac{8}{255}$	IBP	1×	47.35	34.17
			2×	47.83	33.98
		SABR	1×	50.78	34.12
			2×	**51.56**	**34.95**
TinyImageNet	$\frac{1}{255}$	IBP	0.5×	24.47	18.76
			1×	25.33	19.46
			2×	25.40	19.92
		SABR	0.5×	27.56	20.54
			1×	28.63	21.21
			2×	**28.97**	**21.36**

Accuracies and Tightness for Different Methods

Method	ϵ	Accuracy	Tightness	Certified
PGD	2/255	81.2	0.001	-
	8/255	69.3	0.007	-
COLT	2/255	78.4*	0.009	60.7*
	8/255	51.7*	0.057	26.7*
IBP-R	2/255	78.2*	0.033	62.0*
	8/255	51.4*	0.124	27.9*
SABR	2/255	75.6	0.182	57.7
	8/255	48.2	0.950	31.2
IBP	2/255	63.0	0.803	51.3
	8/255	42.2	0.977	31.0

* Literature result.