EFFICIENT NEURAL NETWORK ROBUSTNESS CERTIFICATION WITH GENERAL ACTIVATION FUNCTIONS

MARCH 11, 2019

PREPARED BY: ALI HARAKEH

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING

HOW GOOD IS YOUR NEURAL NETWORK ?

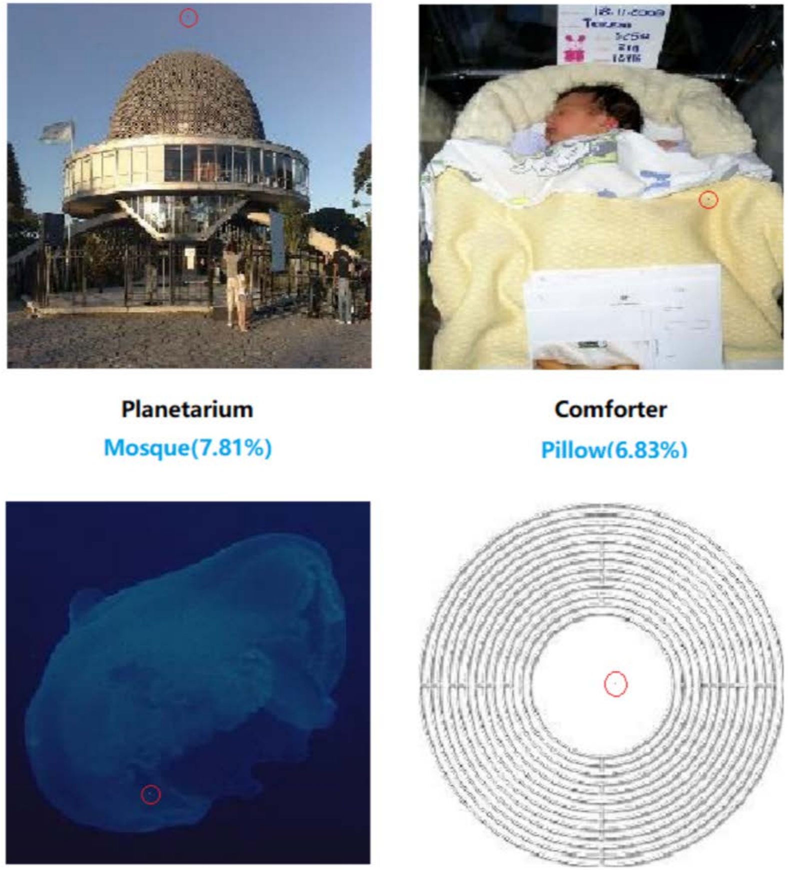
Pei, Kexin, et al. "Deepxplore: Automated whitebox testing of deep learning systems." *Proceedings of the 26th Symposium on Operating Systems Principles*. ACM, 2017.

HOW GOOD IS YOUR NEURAL NETWORK ?

- Neural networks are not robust to input perturbations.

- **Pushing the limit**: One Pixel Attack !
  - Su et. al. "One pixel attack for fooling deep neural networks." IEEE Transactions on Evolutionary Computation (2019).

- **Conclusion:** There is a need for an **automated** and **scalable** analysis to certify realistic neural networks against such input perturbations.



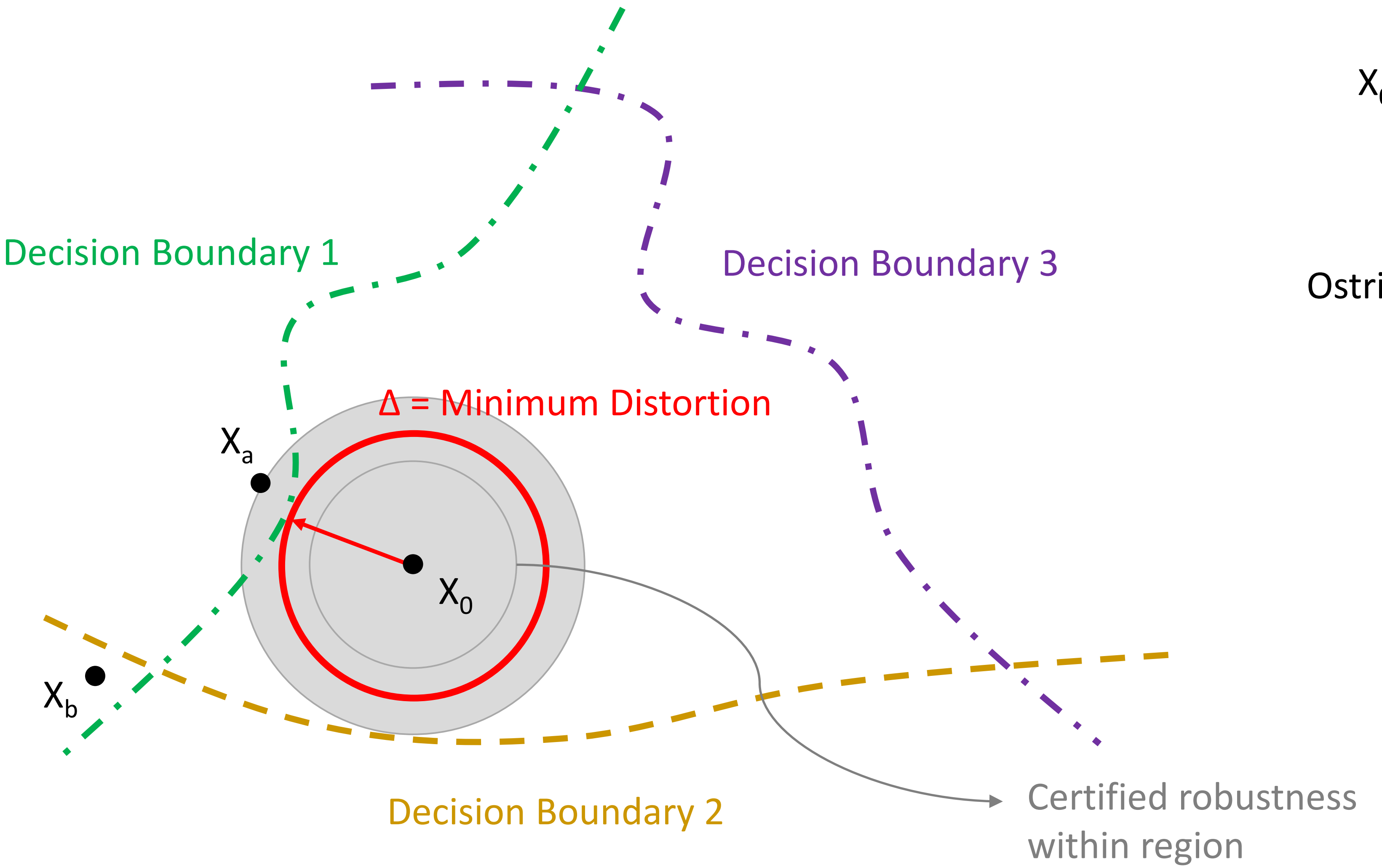
Planetarium
Mosque(7.81%)

Comforter
Pillow(6.83%)

Jellyfish
Bathing tub(21.18%)

Whorl
Blower (37.00%)

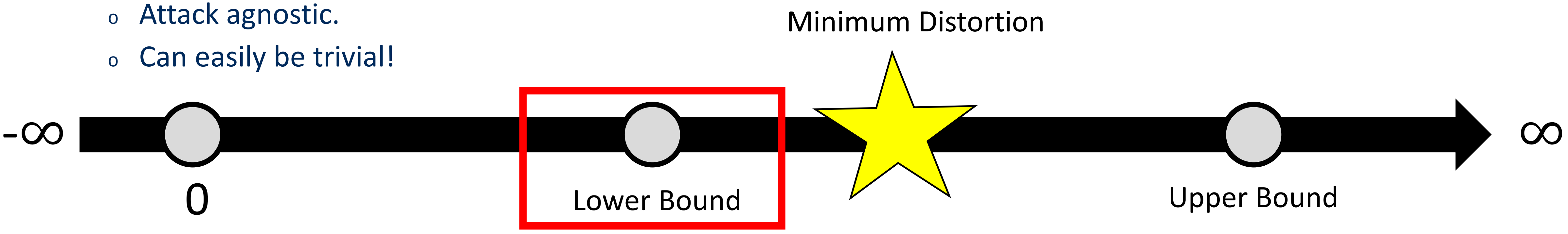
HOW TO CERTIFY NEURAL NETWORKS ?



X_0	X_a	X_b
Ostrich	Vacuum	Shoe

HOW TO CERTIFY NEURAL NETWORKS ?

- **Upper bounds** on minimum distortion:
  - Attack dependent.
  - Is pretty non-informative in case of weak attacks that fail often.
- Formal Verification, **exact** minimum distortion:
  - NP-hard.
- **Lower bounds** on minimum distortion:
  - Attack agnostic.
  - Can easily be trivial!

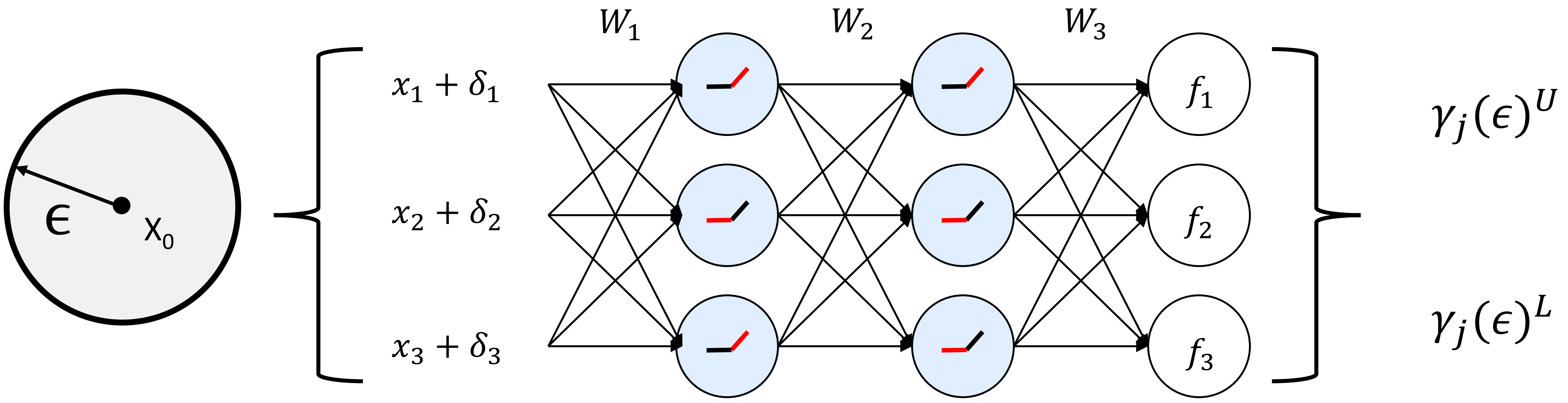


FAVORABLE PROPERTIES OF CERTIFICATION METHODS

Table 1: Comparison of methods for providing adversarial robustness certification in NNs.

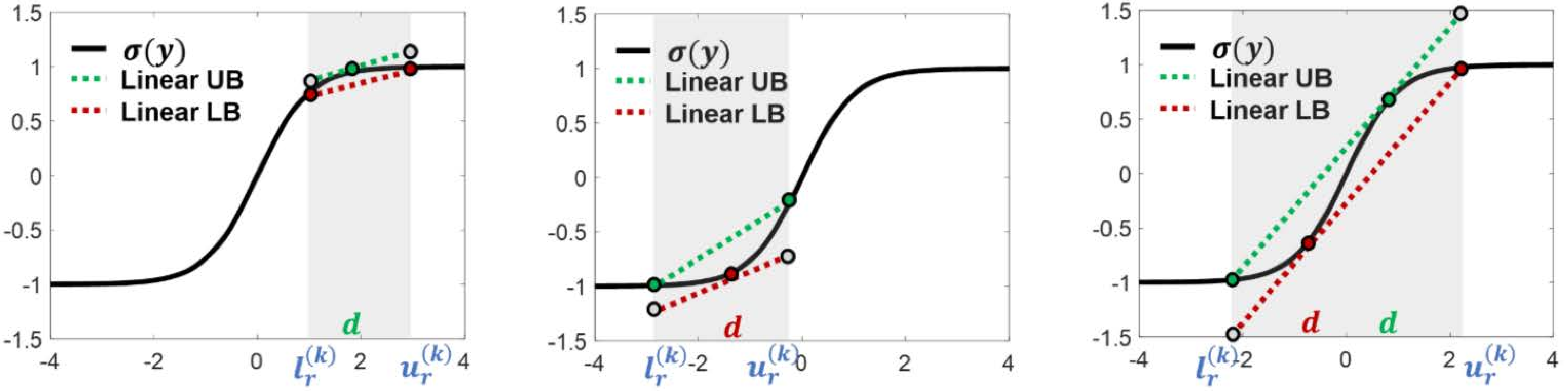
Method	Non-trivial bound	Multi-layer	Scalability	Beyond ReLU
Szegedy et. al. [3]	×	✓	✓	✓
Reluplex [15], Planet [25]	✓	✓	×	×
Hein & Andriushchenko [26]	✓	×	✓	differentiable*
Raghunathan et al. [19]	✓	×	×	×
Kolter and Wong [18]	✓	✓	✓	×
Fast-lin / Fast-lip [20]	✓	✓	✓	×
CROWN (ours)	✓	✓	✓	✓ (general)

STEP 1: EXPLICIT OUTPUT BOUNDS



LINEAR L/U BOUNDS FOR GENERAL ACTIVATION FUNCTIONS

- **Keyword**: Adaptive!

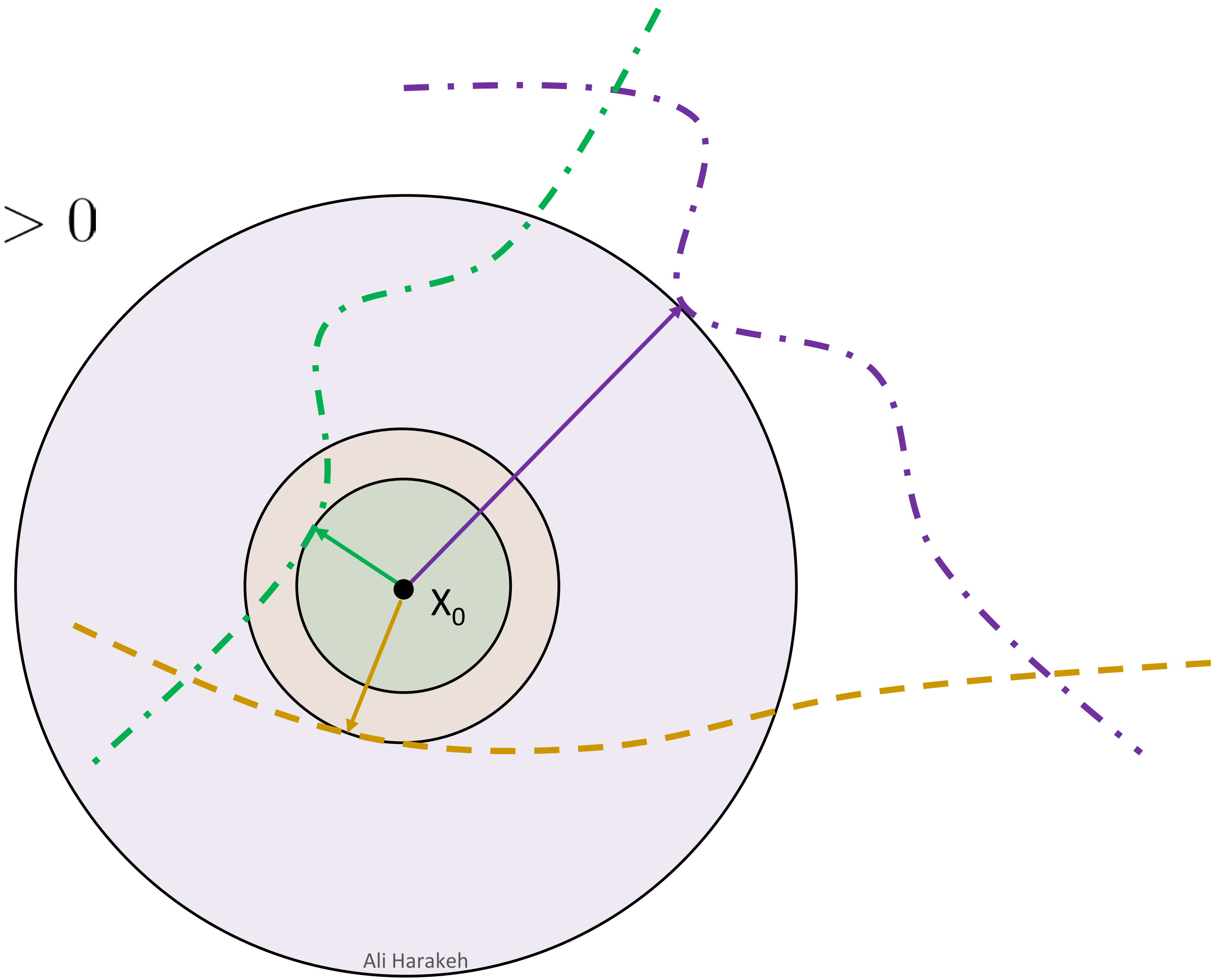


STEP 2: CERTIFIED LOWER BOUND FOR MINIMUM DISTORTION

$$\hat{\epsilon}_t = \max_{\epsilon} \ \epsilon$$

$$s.t. \ \gamma_c^L(\epsilon) - \gamma_t^U(\epsilon) > 0$$

$$\hat{\epsilon} = \min_{t \neq c} \hat{\epsilon}_t$$



RESULTS: TIGHTER LOWER BOUNDS

Table 4: Comparison of certified lower bounds on large ReLU networks. Bounds are the average over 100 images (skipped misclassified images) with random attack targets. Percentage improvements are calculated against Fast-Lin as Fast-Lip is worse than Fast-Lin.

Network	Certified Bounds				Improvement (%)	Average Computation Time (sec)		
	ℓ_p norm	Fast-Lin	Fast-Lip	CROWN-Ada	CROWN-Ada vs Fast-Lin	Fast-Lin	Fast-Lip	CROWN-Ada
MNIST $4 \times [1024]$	ℓ_1	1.57649	0.72800	**1.88217**	+19%	1.80	2.04	3.54
	ℓ_2	0.18891	0.06487	**0.22811**	+21%	1.78	1.96	3.79
	ℓ_∞	0.00823	0.00264	**0.00997**	+21%	1.53	2.17	3.57
CIFAR-10 $7 \times [1024]$	ℓ_1	0.86468	0.09239	**1.09067**	+26%	13.21	19.76	22.43
	ℓ_2	0.05937	0.00407	**0.07496**	+26%	12.57	18.71	21.82
	ℓ_∞	0.00134	0.00008	**0.00169**	+26%	8.98	20.34	16.66

Table 5: Comparison of certified lower bounds by CROWN-Ada on ReLU networks and CROWN-general on networks with tanh, sigmoid and arctan activations. CIFAR models with sigmoid activations achieve much worse accuracy than other networks and are thus excluded.

Network	Certified Bounds by CROWN-Ada and CROWN-general					Average Computation Time (sec)			
	ℓ_p norm	ReLU	tanh	sigmoid	arctan	ReLU	tanh	sigmoid	arctan
MNIST $3 \times [1024]$	ℓ_1	3.00231	2.48407	2.94239	2.33246	1.25	1.61	1.68	1.70
	ℓ_2	0.50841	0.27287	0.44471	0.30345	1.26	1.76	1.61	1.75
	ℓ_∞	0.02576	0.01182	0.02122	0.01363	1.37	1.78	1.76	1.77
CIFAR-10 $6 \times [2048]$	ℓ_1	0.91201	0.44059	-	0.46198	71.62	89.77	-	83.80
	ℓ_2	0.05245	0.02538	-	0.02515	71.51	84.22	-	83.12
	ℓ_∞	0.00114	0.00055	-	0.00055	49.28	59.72	-	58.04

CROWN: WEAK POINTS

1. Feed-Forward Neural Networks with fully connected layers only.
   - CNN-Cert: https://arxiv.org/abs/1811.12395

2. Input should be in the form of an epsilon bound norm ball.
   - Usually not an issue. Common assumption.

3. Single input certification. Results averaged over 100 points of input.
   - A2I and derivatives? Covering arguments?

DISCUSSION QUESTIONS

What do you think of the provided comparison method?

Do you think the authors overpromise scalability?

Can we argue safety of a DNN using CROWN?