Enhancing Gradient-based Attacks With Symbolic Intervals

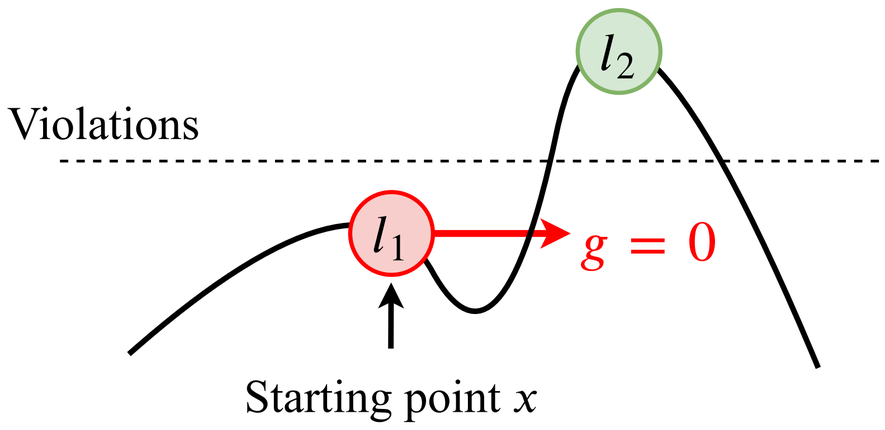
Shiqi Wang, Yizheng Chen, Ahmed Abdou, Suman Jana

GRADIENT-BASED ATTACKS

- Adversarial examples are defined in small Lp-norm bounded range
- First-order adversary: Gradient ascent the loss

LIMITATIONS

- NN is highly non-convex and non-linear => Easily get stuck at local optima
- Attack performance can be improved with multiple starting points.



MNIST

$||x - x_{adv}||_\infty \leq 0.3$

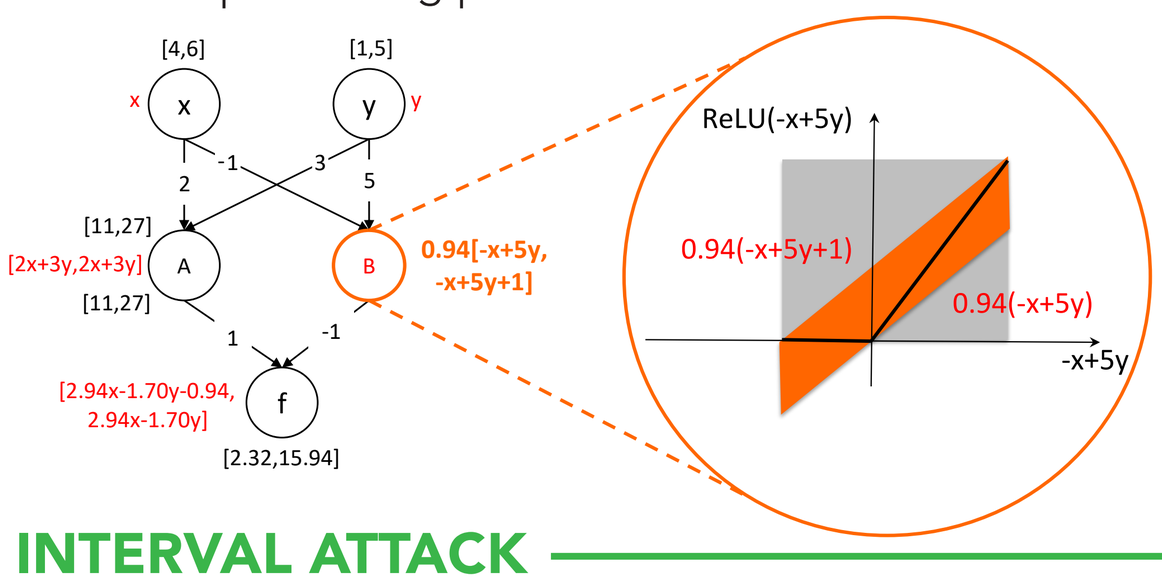
Clean accuracy 98%

Upper bound 92%

SAFE UNDER 100,000 RANDOM STARTING POINTS?

True bound?

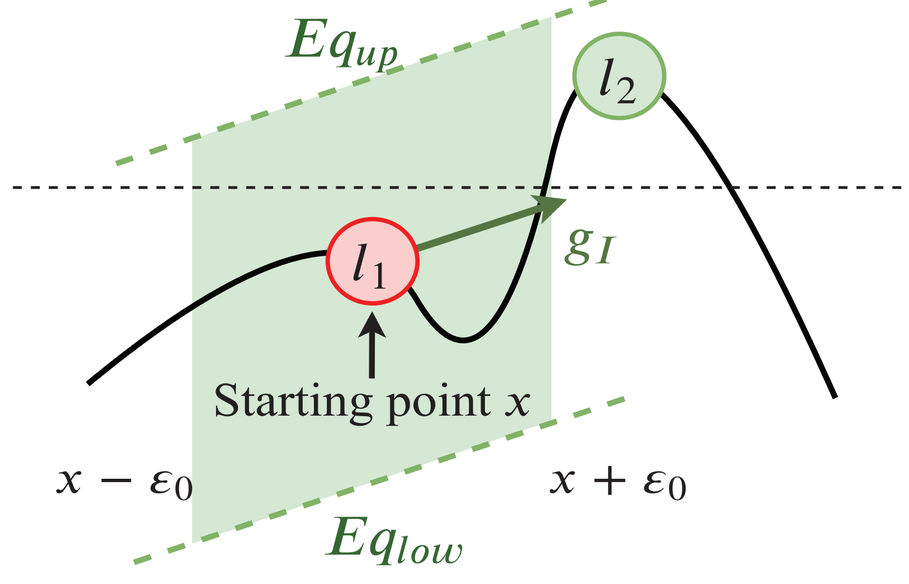
Lower bound 0%



VERIFICATION METHODS: SOUND BOUND PROPAGATION

- Relax nonlinearity with convex function
- Over-approximate the output range
- Provide a broader view within surrounding area

INTERVAL ATTACK



BROADER VIEW IN INTERVAL ATTACK

- It's likely to guide towards the worst-case behavior within surrounding area
- Higher chance to avoid local optima

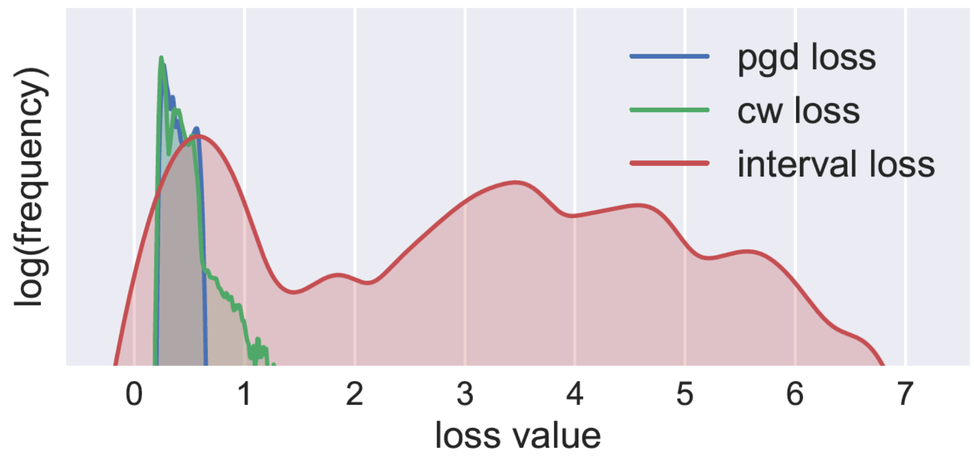
INTERVAL GRADIENT

- The slope of the two parallel symbolic intervals
- A generic framework that can adapt other sound propagation methods (e.g., worst-case or average gradients)

OVERESTIMATION ERROR

(1) Cannot converge to optima if only rely on interval gradient => Using interval gradient ascent to locate interesting area and then use regular gradient ascent to converge

(2) The error is proportional to the input range => Dynamically balance the range used for each step



Network	# Hidden units	# Parameters	ACC (%)	Attack success rate (%)			
				PGD	CW	Interval Attack	Interval Attack Gain
MNIST_FC1	1,024	668,672	98.1	39.2	42.2	**56.2**	+17 (43%)
MNIST_FC2	10,240	18,403,328	98.8	34.4	32.2	**44.4**	+10.0 (38%)
MNIST_Conv	38,656	3,274,634	98.4	7.2	7.3	**11.6***	+4.4 (61%)

* Interval attack achieves the best attack success rate in MadryLab MNIST Challenge (Madry et al., 2018b).

EXPERIMENTAL RESULTS

- 100,000 might still not enough to locate all adversaraial examples=>Stronger attacks are still needed before model is verified
- The strongest attack so far on MadryLab MNIST challenge
- On average 47% relatively more than PGD attack

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK
PennState
CLSec COLUMBIA UNIVERSITY SECURITY GROUP

CODE AVAILABLE AT:
https://github.com/tcwangshiqi-columbia/Interval-Attack