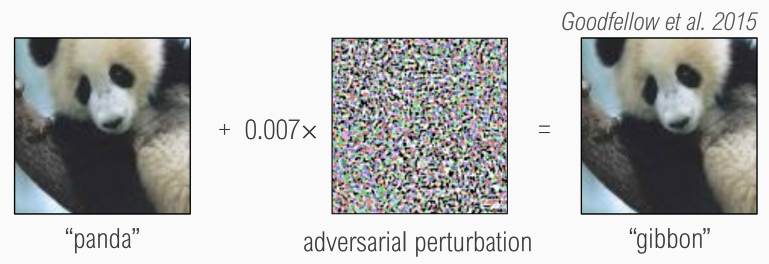
Globally-Robust Neural Networks

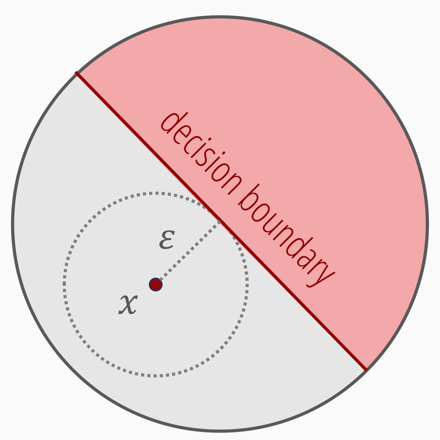
Klas Leino, Zifan Wang, Matt Fredrikson

Carnegie Mellon University

Adversarial Examples & Local Robustness

Deep networks have extensively been shown to be vulnerable to *adversarial examples*, wherein inconspicuous perturbations are chosen to cause arbitrary misclassifications.





local robustness

a model is *ε-locally-robust* at a point, x, if it classifies all points in the ε-ball centered at x consistently; i.e., there are no decision boundaries within ε from x

certified defenses

Certification of local robustness at a given point allows us to provably preclude small-norm adversarial examples at that point

Our Contributions

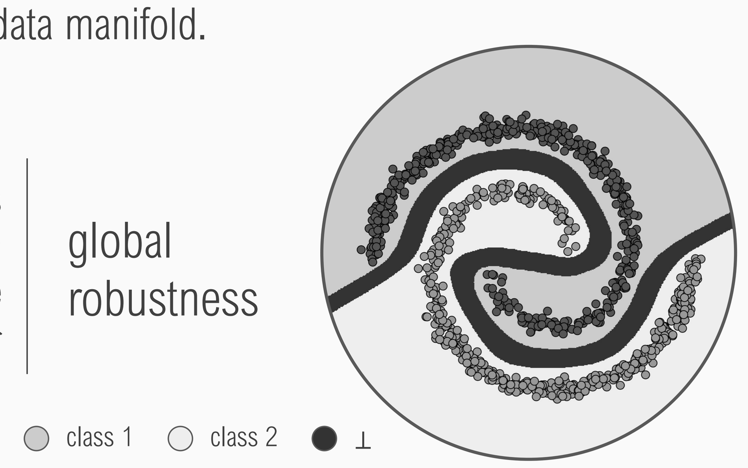
- We introduce a notion of *global* robustness
- We devise a way to construct a type of network that is globally robust *by construction*
- Our globally-robust networks are efficient to train and can certify points in a *single forward pass*

Global Robustness

A model cannot be locally robust at all points, as points near a decision boundary cannot be robust. Thus, a global notion of robustness must allow the model to reject some inputs. Ideally, the rejected inputs would lie off the data manifold.

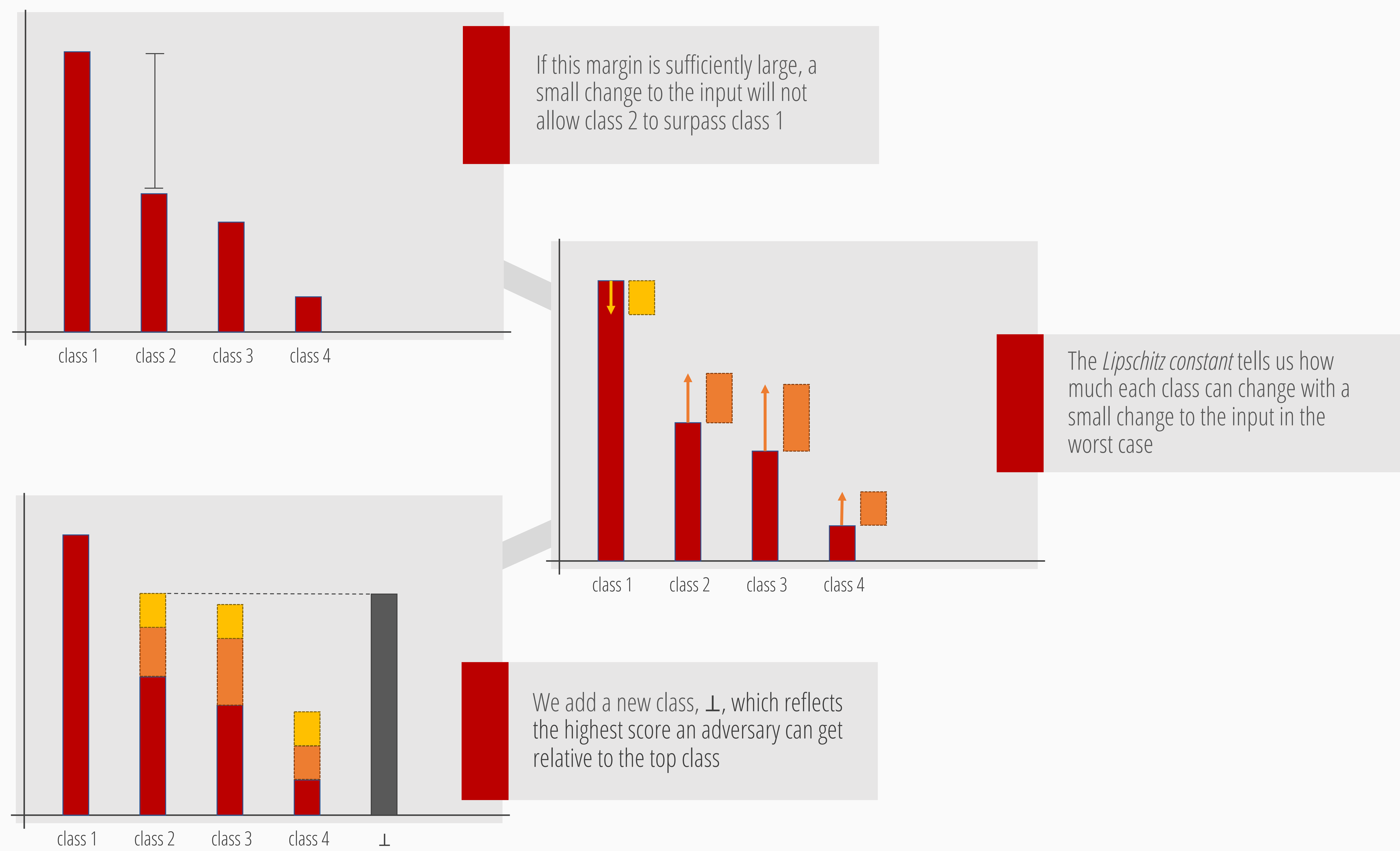
a model is *ε-globally-robust* if on all points it is either ε-locally-robust or abstains from prediction by outputting ⊥; i.e., all non-⊥ classes are separated by a margin of at least 2ε

global robustness



GloRo Nets: Globally Robust by Construction

We present **Glo**bally **Ro**bust Networks (GloRo Nets), which instrument the output of a standard neural network in such a way that the resulting network is guaranteed to be globally robust by construction.



safety

If the GloRo Net outputs a non-⊥ class on a given point, then the underlying network is guaranteed to be locally robust at that point

robust training

As ⊥ is never the correct label, by training the GloRo Net to be accurate, it avoids picking ⊥, which means it avoids being non-robust

Calculating the Lipschitz Bound

The global Lipschitz constant can be efficiently bounded by taking a layer-wise product of the spectral norm of each layer.

- Any layer with bounded Lipschitz constant can be used
- Layer-wise product may be loose; this bound may be able to be improved, effectively increasing the certified radius

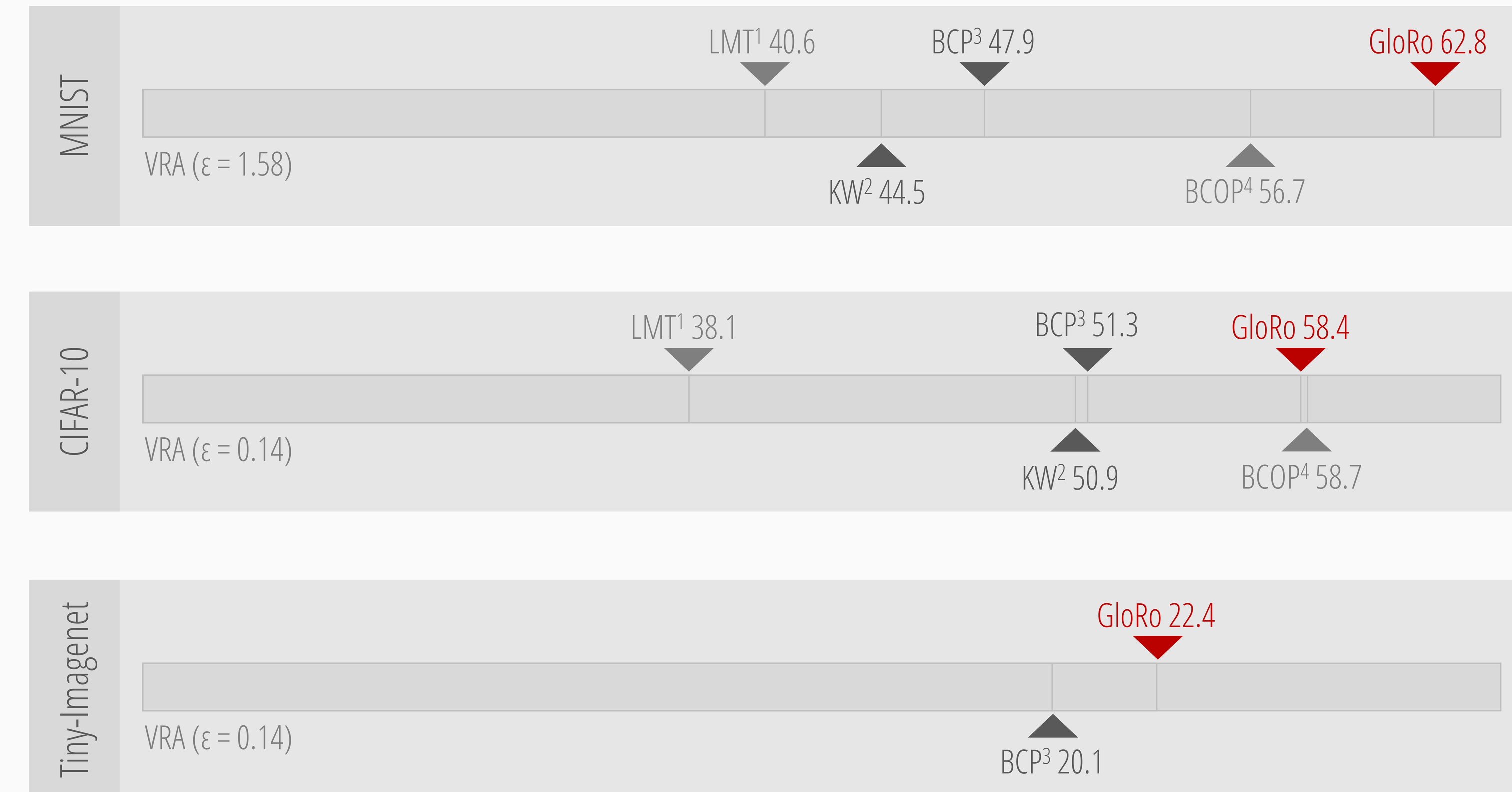
completeness of the global bound

We use the *global* Lipschitz constant to implement GloRo Nets. Although only *local* Lipschitzness is required for robustness[1], we show that in theory, the global bound is equally powerful for robustness certification.

[1]Yang et al., 2020

Summary of Results

GloRo Nets match or exceed the Verified Robust Accuracy (VRA) achieved by previous state-of-the art deterministic certification methods.



[1]Tsuzuku et al., 2018; [2]Wong & Kolter, 2018; [3]Lee et al., 2020; [4]Li et al., 2019

GloRo Net certification and training is significantly more time- and memory-efficient than other certification methods, and more scalable than any other deterministic method.

	method	time per training epoch (s)	memory per instance during training (MB)
CIFAR-10	GloRo	6.9	3.6
	KW[1]	516.8	100.9
	BCP[2]	47.5	12.7

	method	time to certify test set (s)	memory per instance (MB)
CIFAR-10	GloRo	0.4	1.8
	KW[1]	2,500.0	1,400.0
	BCP[2]	5.8	19.1
	RS[3]	36,800.0	19.8

[1]Wong & Kolter, 2018; [2]Lee et al., 2020; [3]Cohen et al., 2019

learn more

check out our talk and the full paper for more!

code available on GitHub

full paper
https://tinyurl.com/gloro-icml2021

implementation
https://github.com/klasleino/gloro