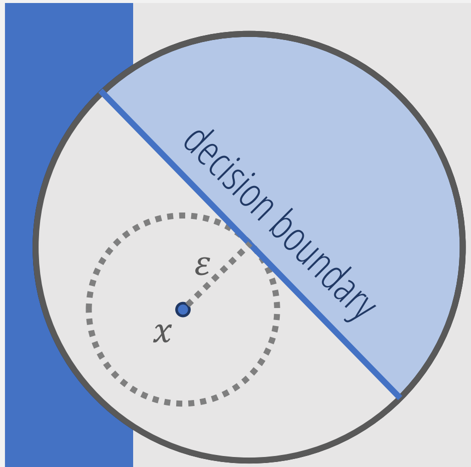
Globally Robust Neural Networks

Klas Leino, Zifan Wang, Matt Fredrikson | Carnegie Mellon University

Robustness Guarantees

Defense against ***adversarial examples***



A model $\boldsymbol{F}$ satisfies ***local robustness*** with robustness radius $\boldsymbol{\varepsilon}$ on a point $\boldsymbol{x}$ if

$$\forall x'.\, \|x - x'\|_p \leq \varepsilon \implies F(x) = F(x')$$

Our Contributions

We introduce a notion of *global* robustness

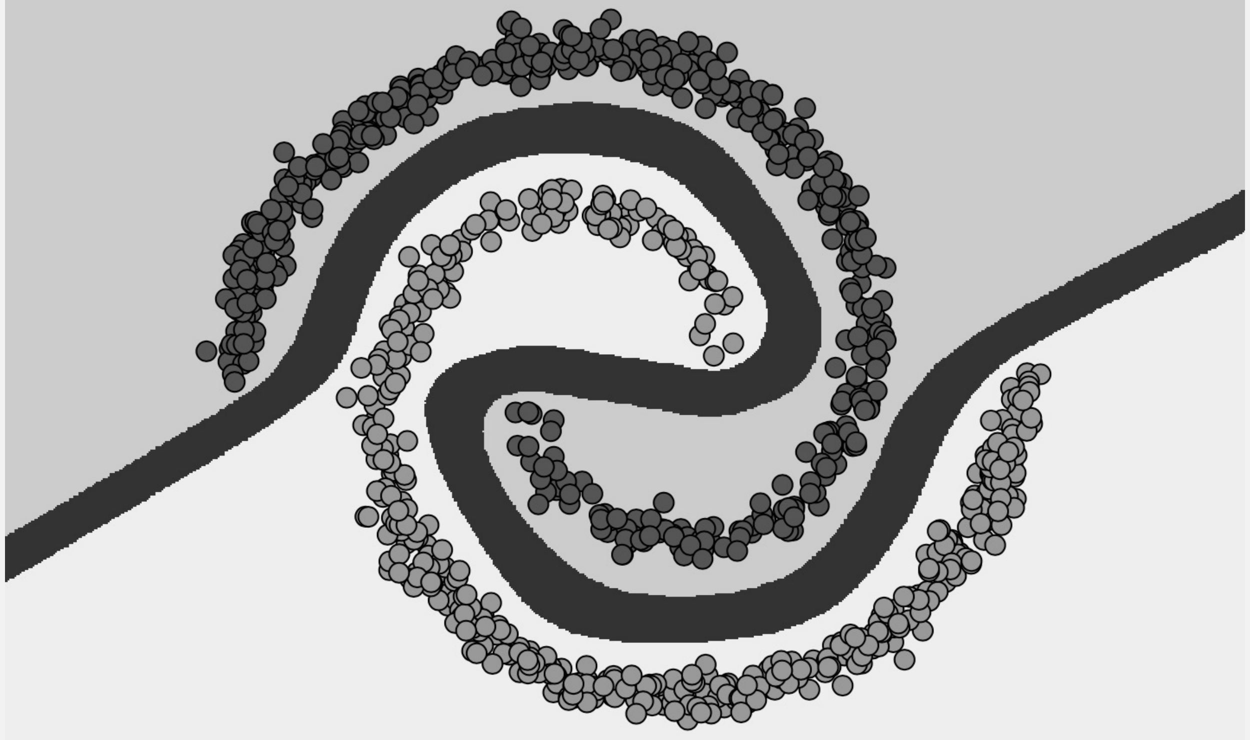
We devise a way to construct a type of network that is globally robust *by construction*

Our globally-robust networks are efficient to train and can certify points in a *single forward pass*

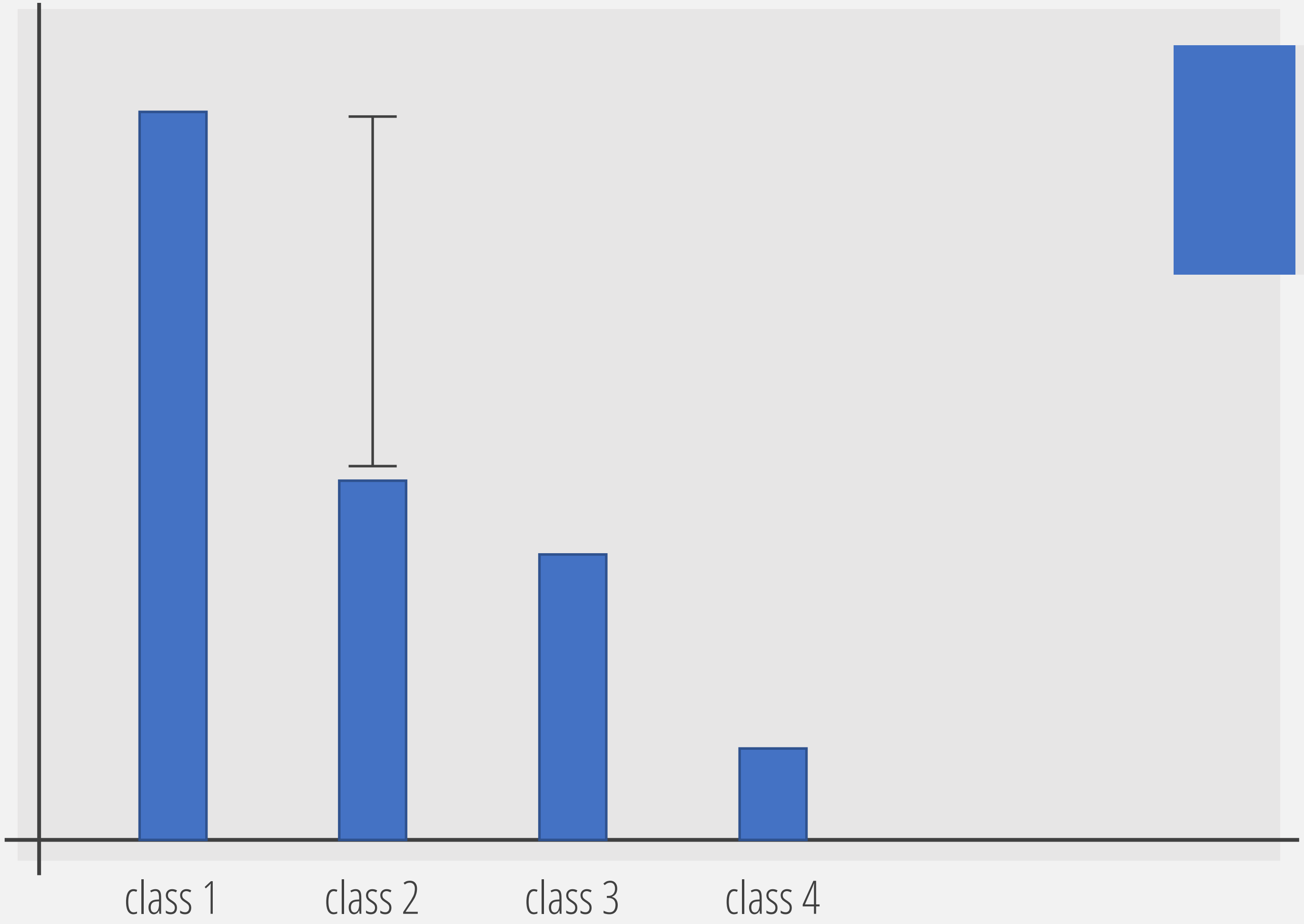
Global Robustness

A model F satisfies ***global robustness*** with robustness radius ε if $\forall x$

- F is $(\epsilon/_2)$-locally robust at x or
- $F(x) = \bot$

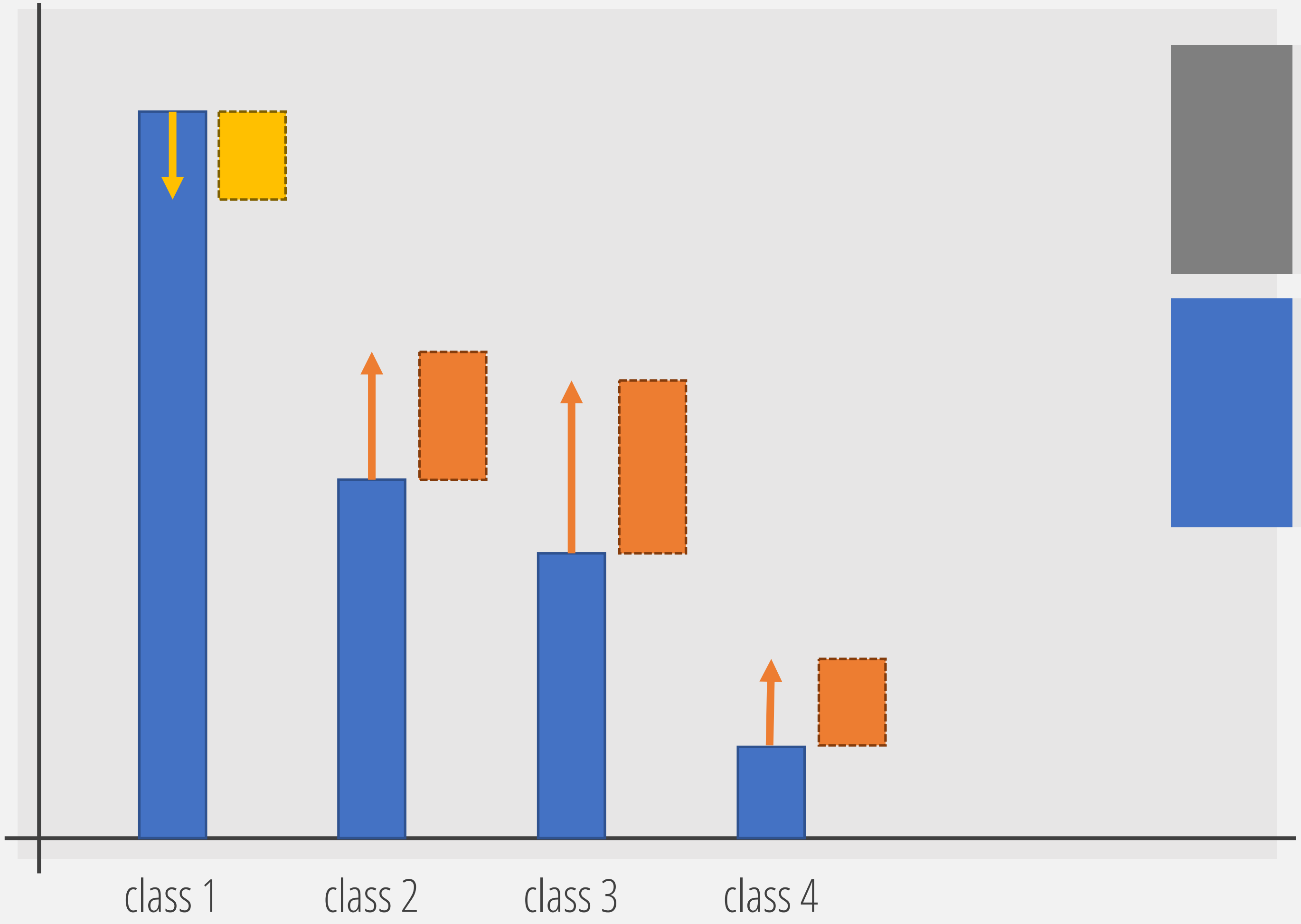


Globally Robust Neural Networks (GloRo Nets)



If this margin is sufficiently large, a small change to the input will not allow class 2 to surpass class 1

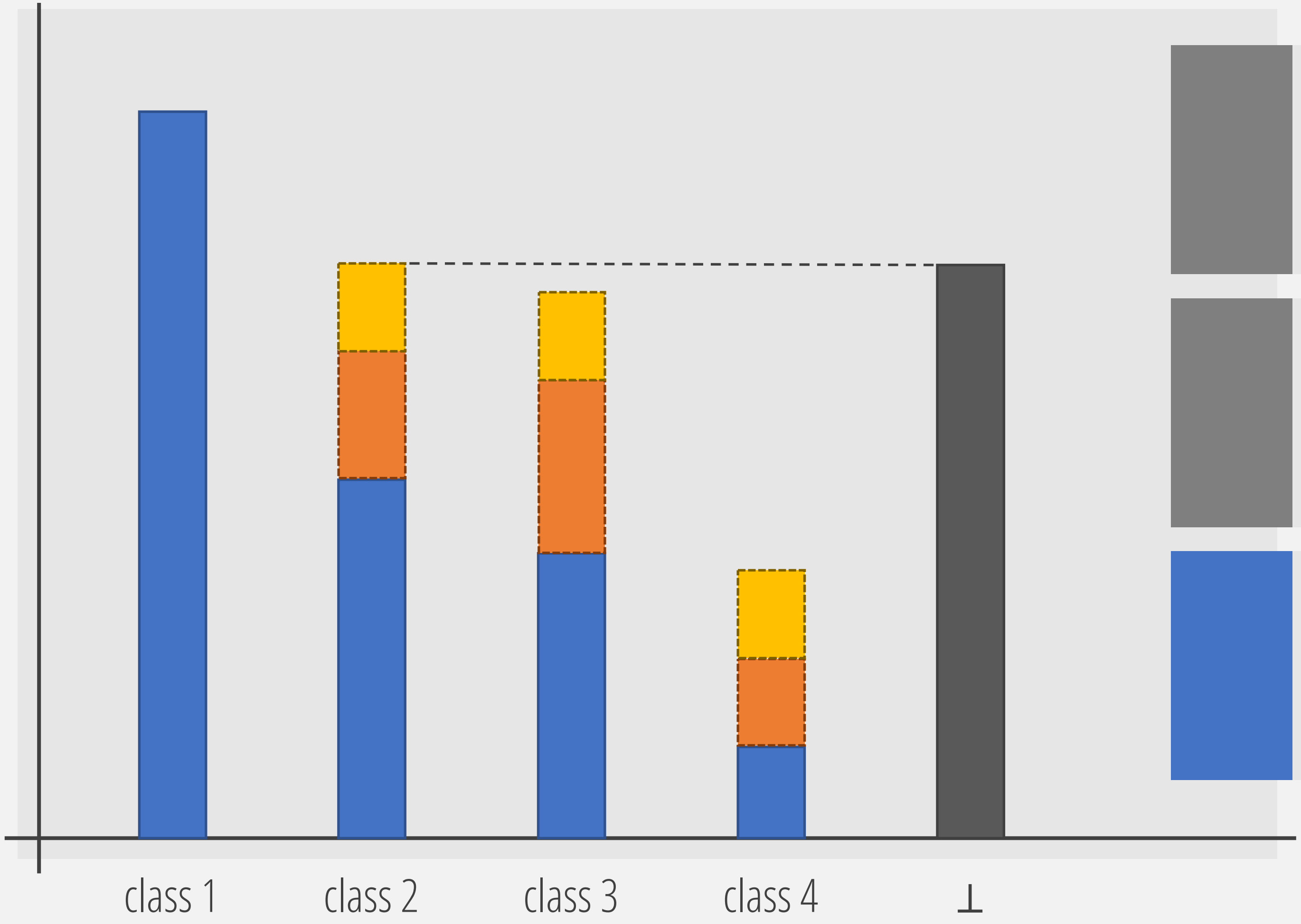
Globally Robust Neural Networks (GloRo Nets)



If this margin is sufficiently large, a small change to the input will not allow class 2 to surpass class 1

The *Lipschitz constant* tells us how much each class can change with a small change to the input in the worst case

Globally Robust Neural Networks (GloRo Nets)



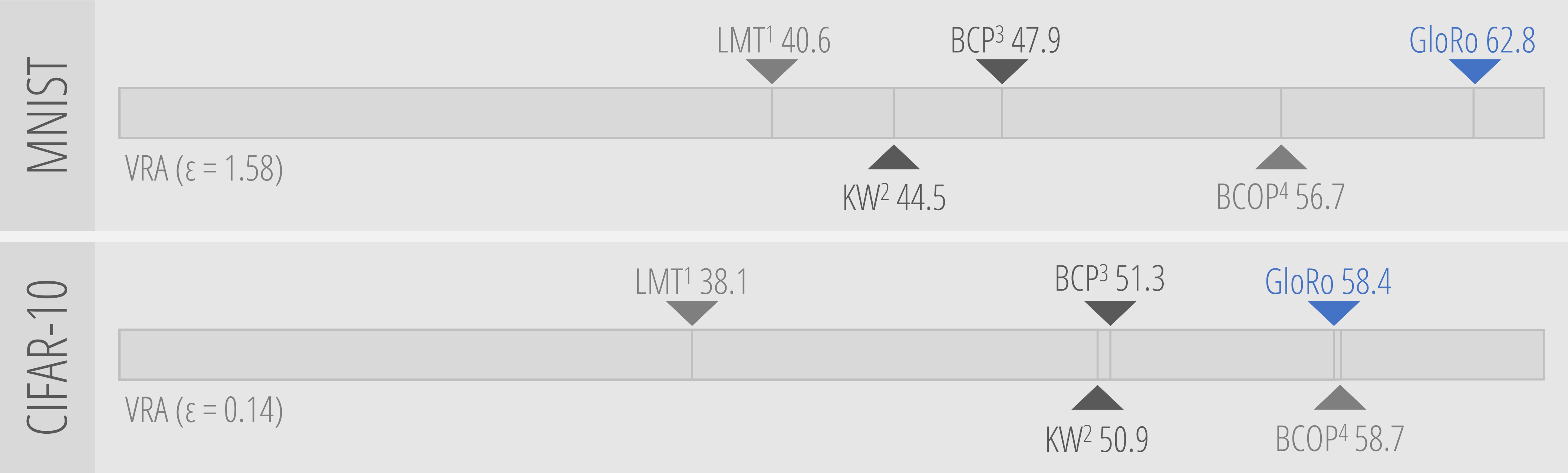
If this margin is sufficiently large, a small change to the input will not allow class 2 to surpass class 1

The *Lipschitz constant* tells us how much each class can change with a small change to the input in the worst case

We add a new class, ⊥, which reflects the highest score an adversary can get relative to the top class

Summary of Results

GloRo Nets match or exceed VRA of 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

Summary of Results

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

	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

Conclusion

Summary

We provide a **scalable** approach to deterministic robustness certification that achieves **state-of-the-art VRA** using only **a single forward pass** of the network for certification.

Check Out Our Paper!

- Paper on ArXiv
- Implementation on GitHub

https://github.com/klasleino/gloro

full paper

https://tinyurl.com/gloro-icml2021