UNIVERSITY OF TORONTO

VECTOR INSTITUTE

Sorting Out Lipschitz Function Approximation

Cem Anil*

James Lucas*

Roger Grosse

*Equal contribution

Pacific Ballroom

Poster #15

(6:30 – 9:00 PM)

Goal

Train neural networks subject to a **strict Lipschitz constraint** **while** **maintaining expressive power.**

Goal

Train neural networks subject to a **strict Lipschitz constraint** **while** **maintaining expressive power.**

$$\underbrace{||f(x_2) - f(x_1)||_p}_{\text{Norm of Output Change}} \leq \underbrace{K}_{\text{Lipschitz Constant}} \underbrace{||x_2 - x_1||_p}_{\text{Norm of Input Change}}$$

Goal

Train neural networks subject to a **strict Lipschitz constraint** **while maintaining expressive power.**

$$\underbrace{||f(x_2) - f(x_1)||_p}_{\text{Norm of Output Change}} \leq \underbrace{K}_{\text{Lipschitz Constant}} \underbrace{||x_2 - x_1||_p}_{\text{Norm of Input Change}}$$

$$\underbrace{||\nabla f(x)||_2}_{\text{Gradient Norm}} \leq \underbrace{K}_{\text{Lipschitz Constant}}$$

Goal

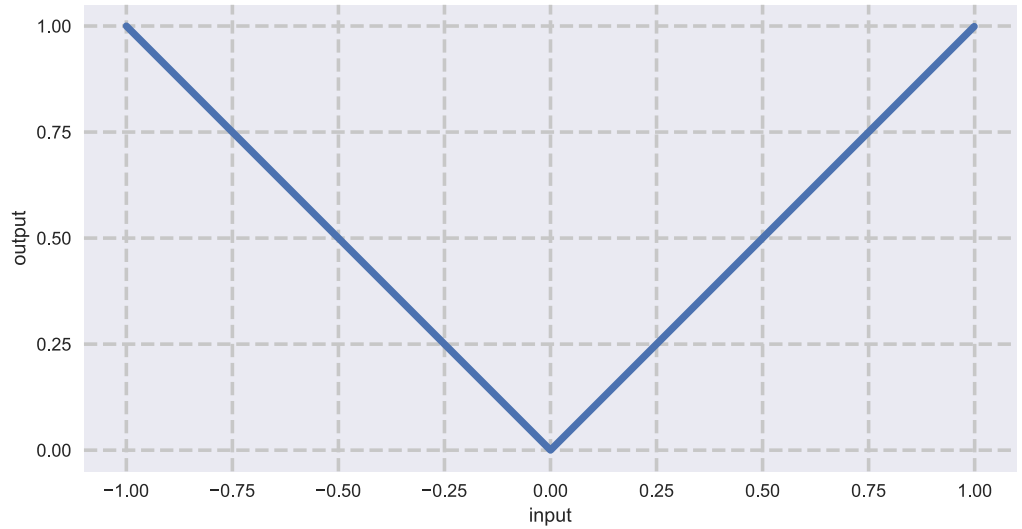
Train neural networks subject to a **strict Lipschitz constraint** **while maintaining expressive power.**

$$||f(x_2) - f(x_1)||_p \leq K\,||x_2 - x_1||_p$$

Norm of Output Change | Lipschitz Constant | Norm of Input Change

$$||\nabla f(x)||_2 \leq K$$

Gradient Norm | Lipschitz Constant



Why Care?

- Provable Adversarial Robustness (Cisse et. al., 2018)
- Wasserstein Distance Estimation(Arjovsky et. al., 2017)
- Training Generative Models (Arjovsky et. al., 2017) (Behrmann et. al., 2019)
- Computing Generalization Bounds (Bartlett et. al., 1998,2017)
- Stabilizing Neural Net Training(Xiao et. al., 2018) (Odena et. al., 2018)
- ...

Lipschitz via. Architectural Constraints

Design an architecture that is:

Constrained Enough

Never violates a prescribed K-Lipschitz constraint

Expressive Enough

Approximate any K-Lipschitz Function (universality).

Universal Lipschitz Function Approximation

Lipschitz via. Architectural Constraints

Design an architecture that is:

Constrained Enough
Never violates a prescribed K-Lipschitz constraint

Expressive Enough
Approximate any K-Lipschitz Function (universality).

Universal Lipschitz Function Approximation

Main Contributions

Propose an expressive Lipschitz constrained architecture that

- Overcomes a previously unidentified limitation in prior art.
- Can recover Universal Lipschitz function approximation.

Lipschitz via. Architectural Constraints

Design an architecture that is:

Constrained Enough
Never violates a prescribed K-Lipschitz constraint

Expressive Enough
Approximate any K-Lipschitz Function (universality).

Universal Lipschitz Function Approximation

Main Contributions

Propose an expressive Lipschitz constrained architecture that

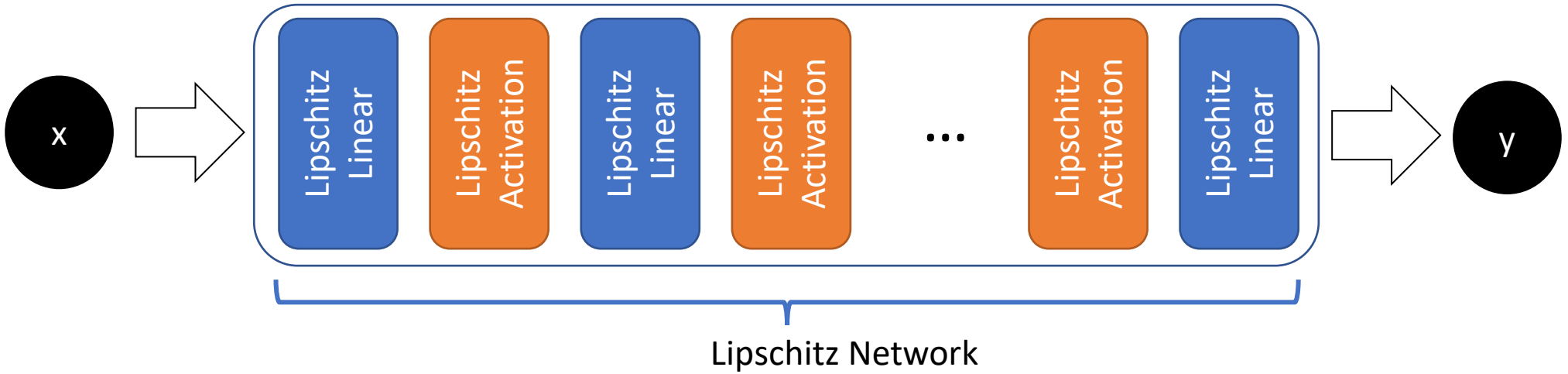
- Overcomes a previously unidentified limitation in prior art.
- Can recover Universal Lipschitz function approximation.

Apply this architecture to

- Train classifiers provably robust to adversarial perturbations.
- Obtain tight estimates of Wasserstein distance.

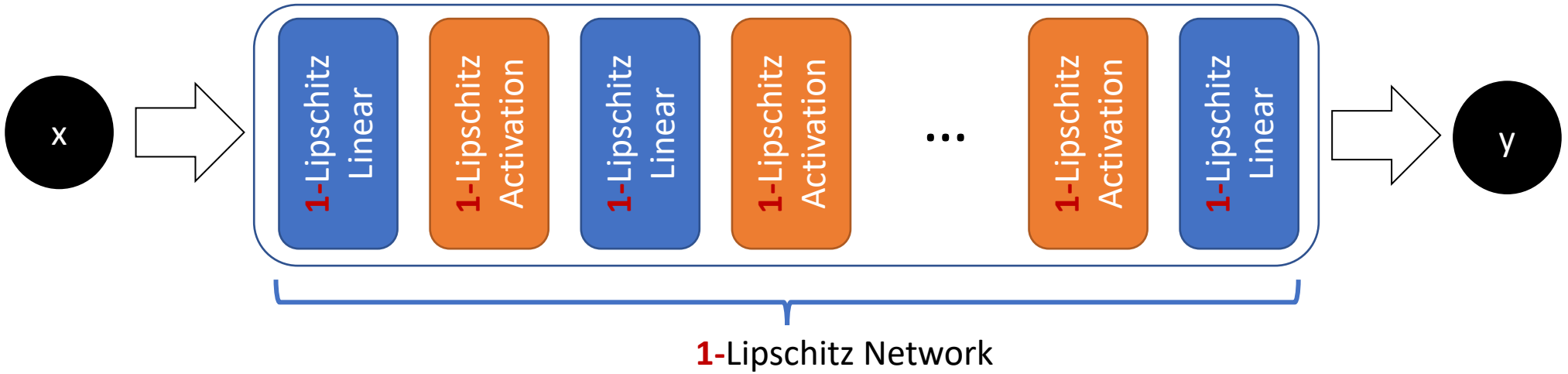
Lipschitz via. Architectural Constraints

- Compose Lipschitz linear layers and Lipschitz activations.



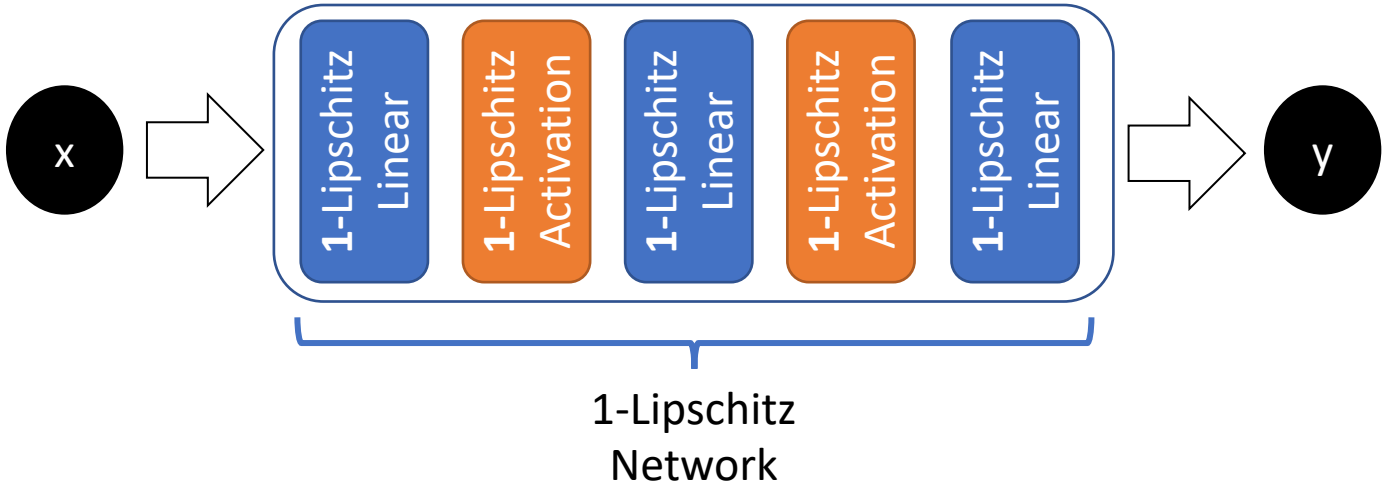
Lipschitz via. Architectural Constraints

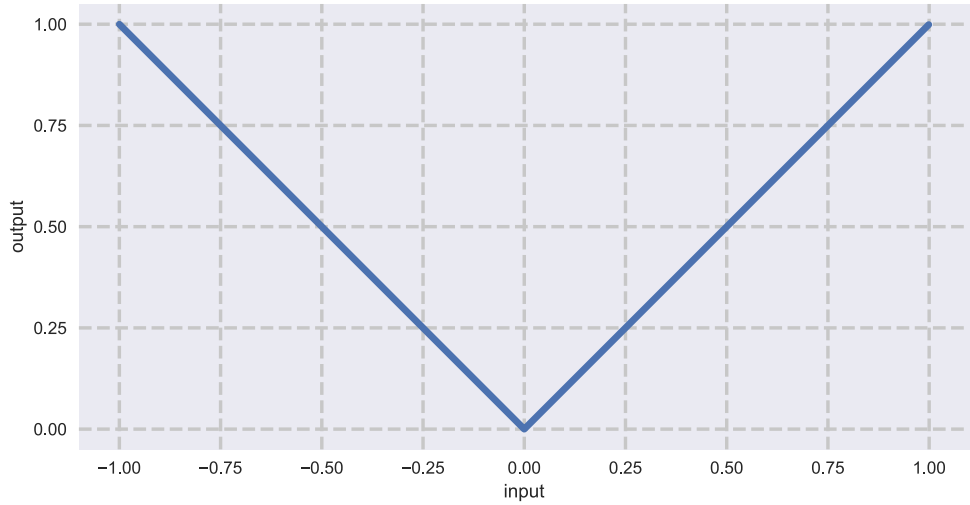
- Compose Lipschitz linear layers and Lipschitz activations.



Lipschitz via. Architectural Constraints

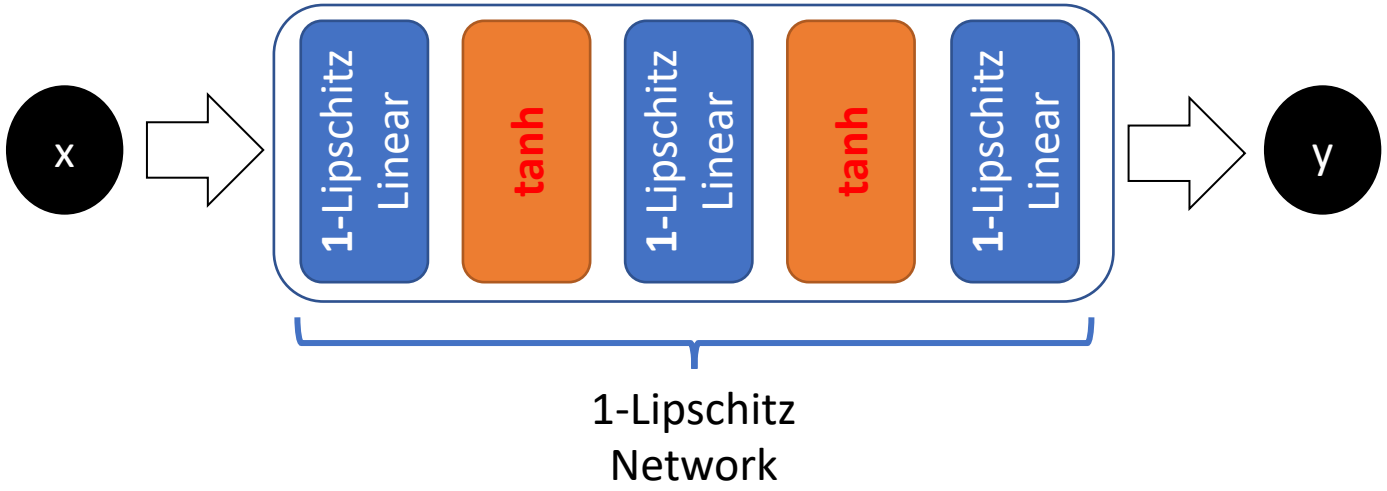
First thing to try: approximate absolute value function.

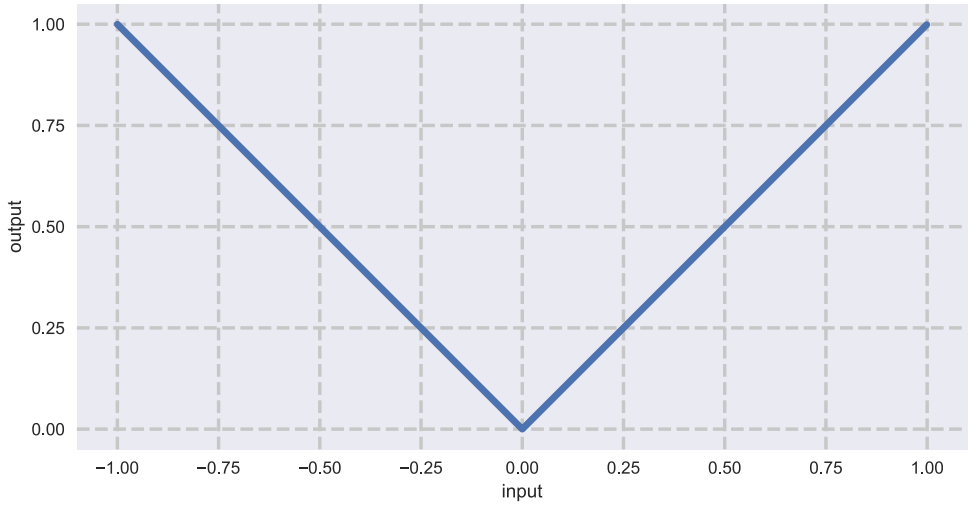




Lipschitz via. Architectural Constraints

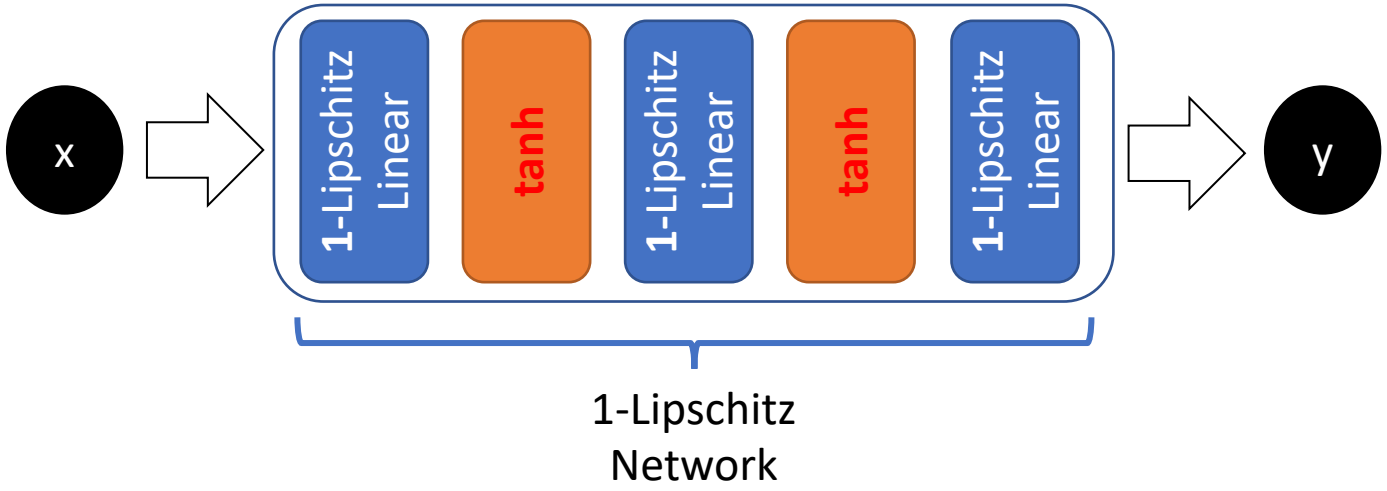
First thing to try: approximate absolute value function.

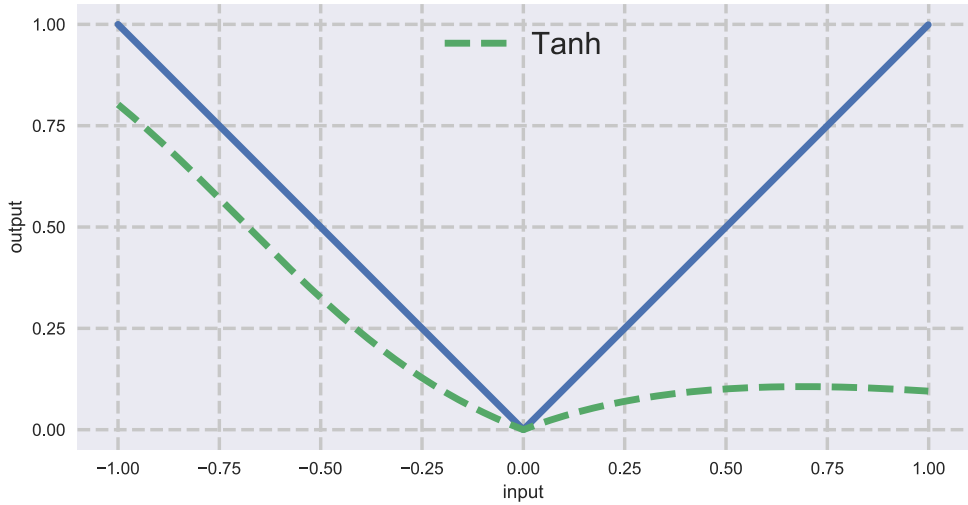




Lipschitz via. Architectural Constraints

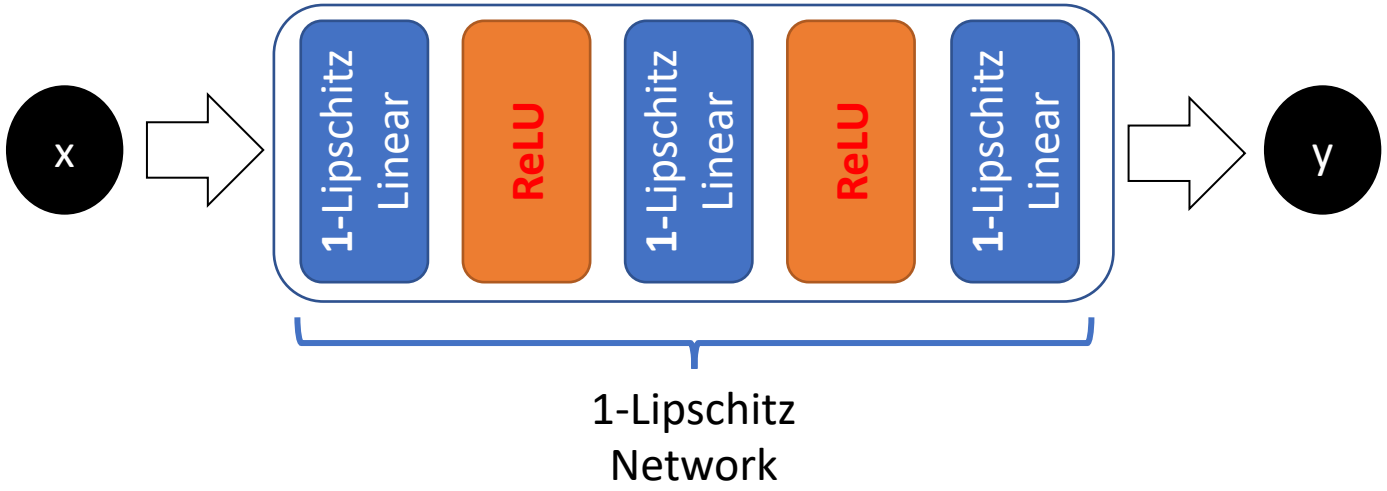
First thing to try: approximate absolute value function.

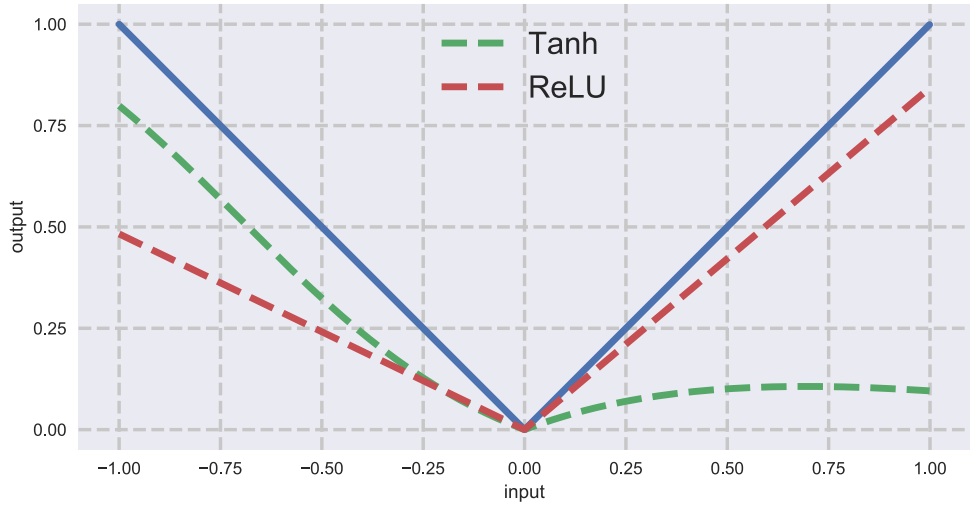




Lipschitz via. Architectural Constraints

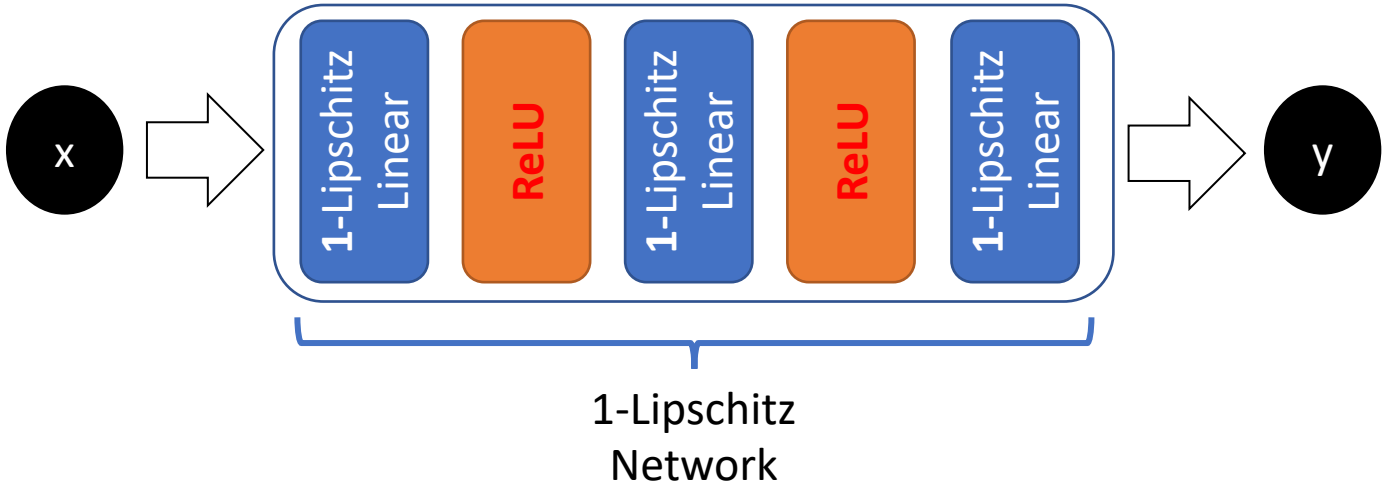
First thing to try: approximate absolute value function.

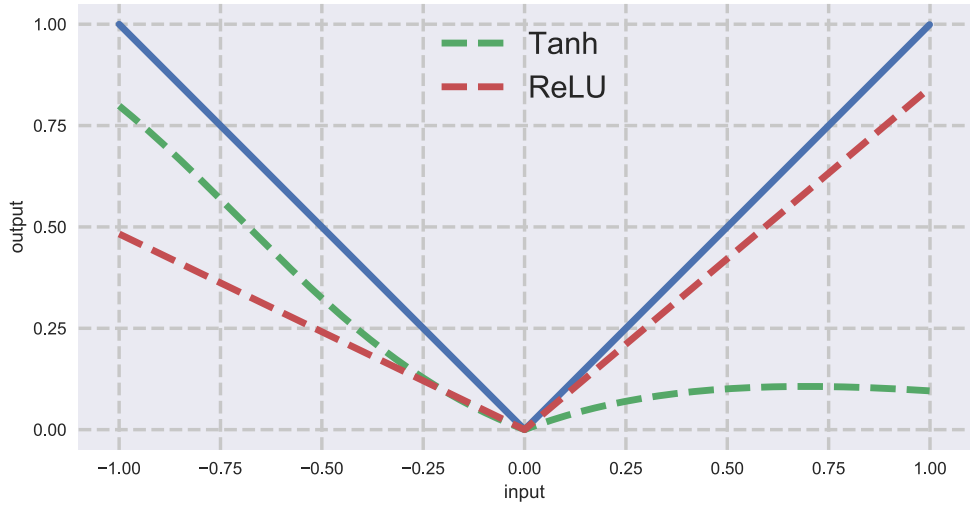




Lipschitz via. Architectural Constraints

What went wrong?

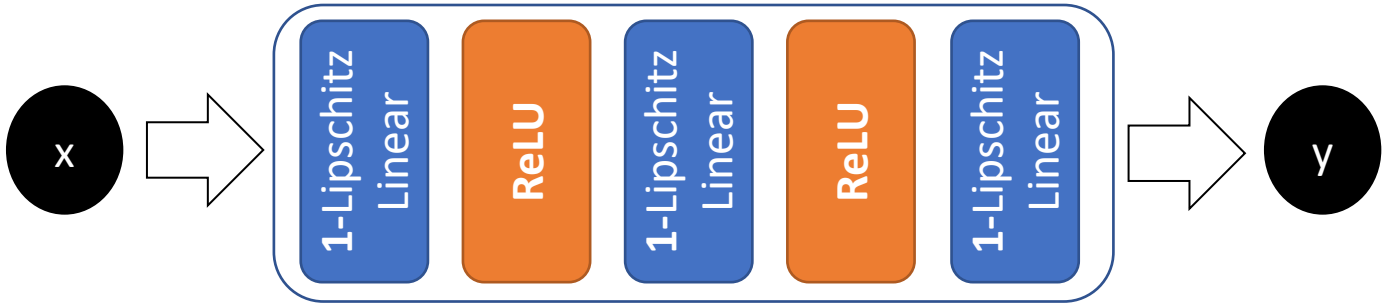


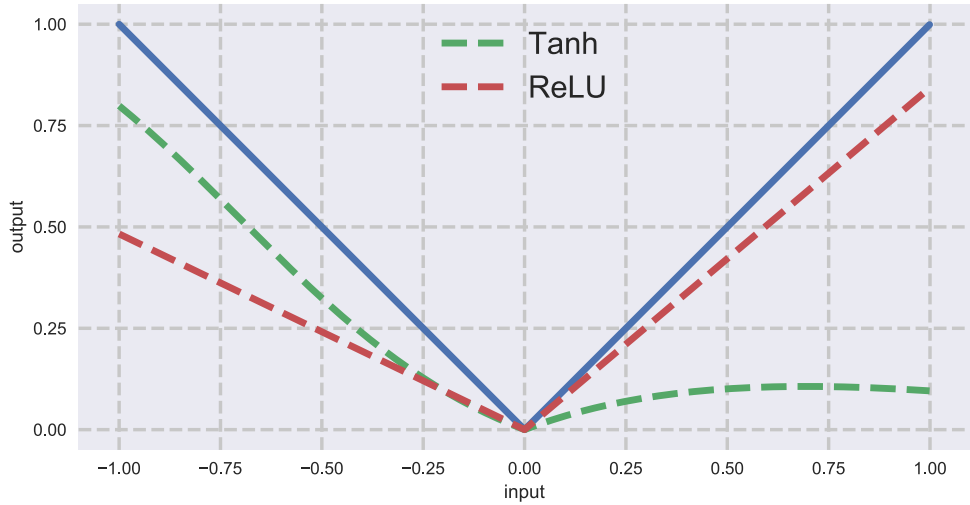


???

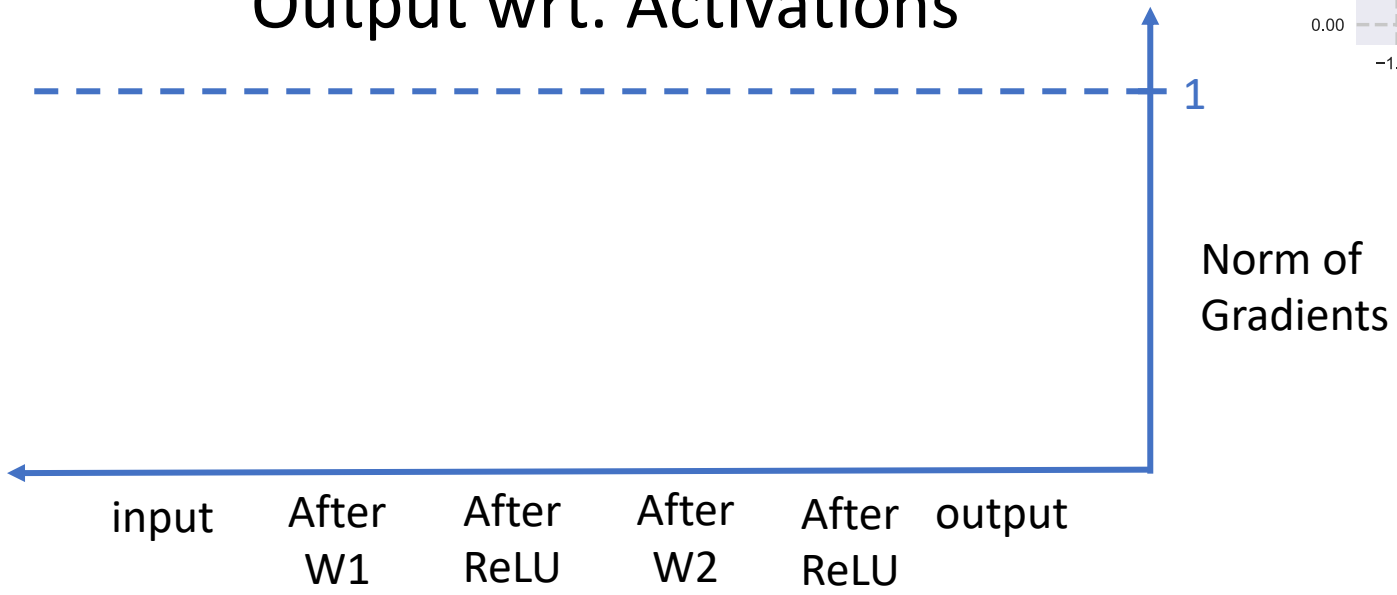
Lipschitz via. Architectural Constraints

- Diagnosing the issue: Inspect gradient norms!



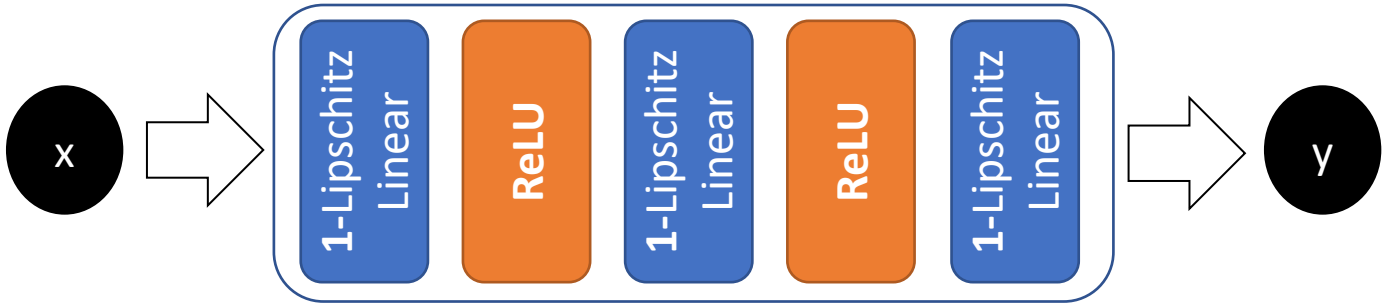


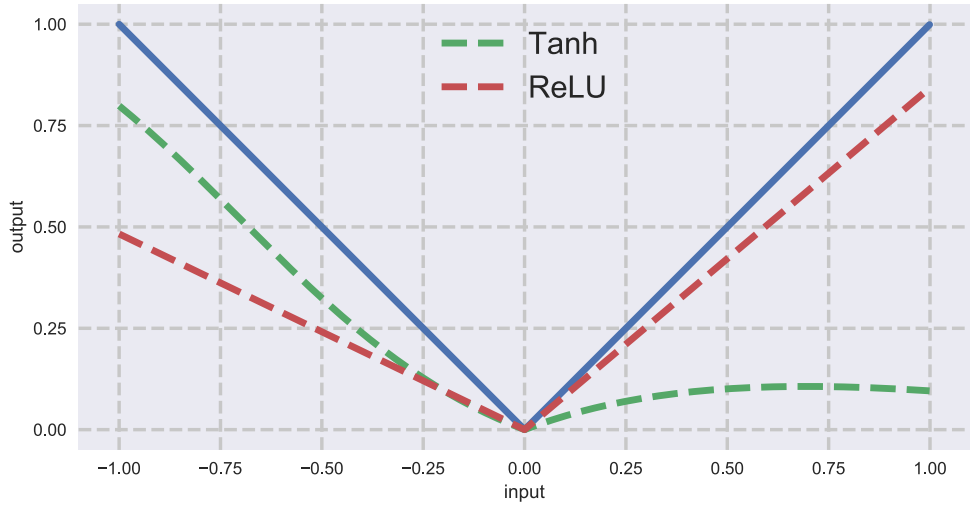
Gradient Norms of Output wrt. Activations



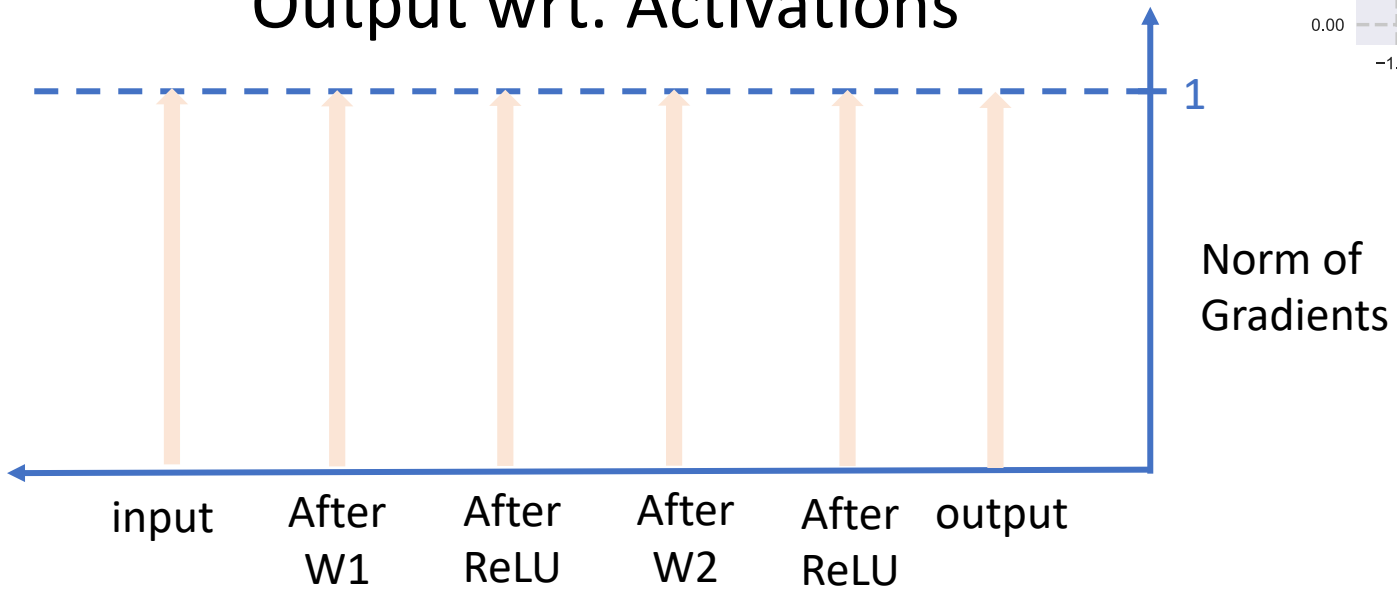
Lipschitz via. Architectural Constraints

- Diagnosing the issue: Inspect gradient norms!



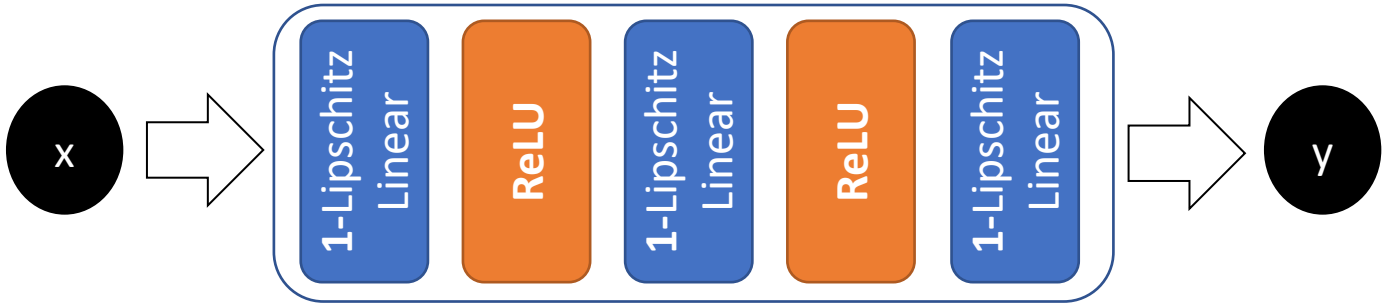


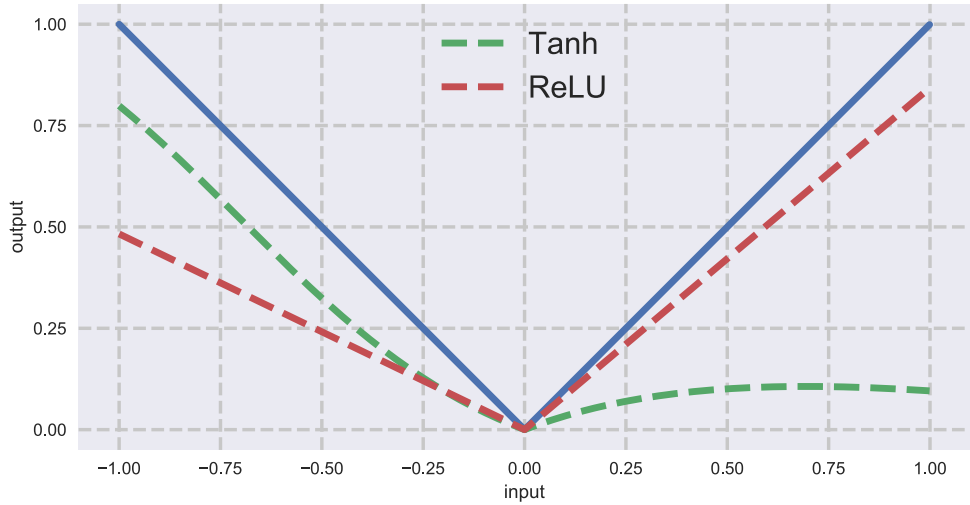
Gradient Norms of Output wrt. Activations



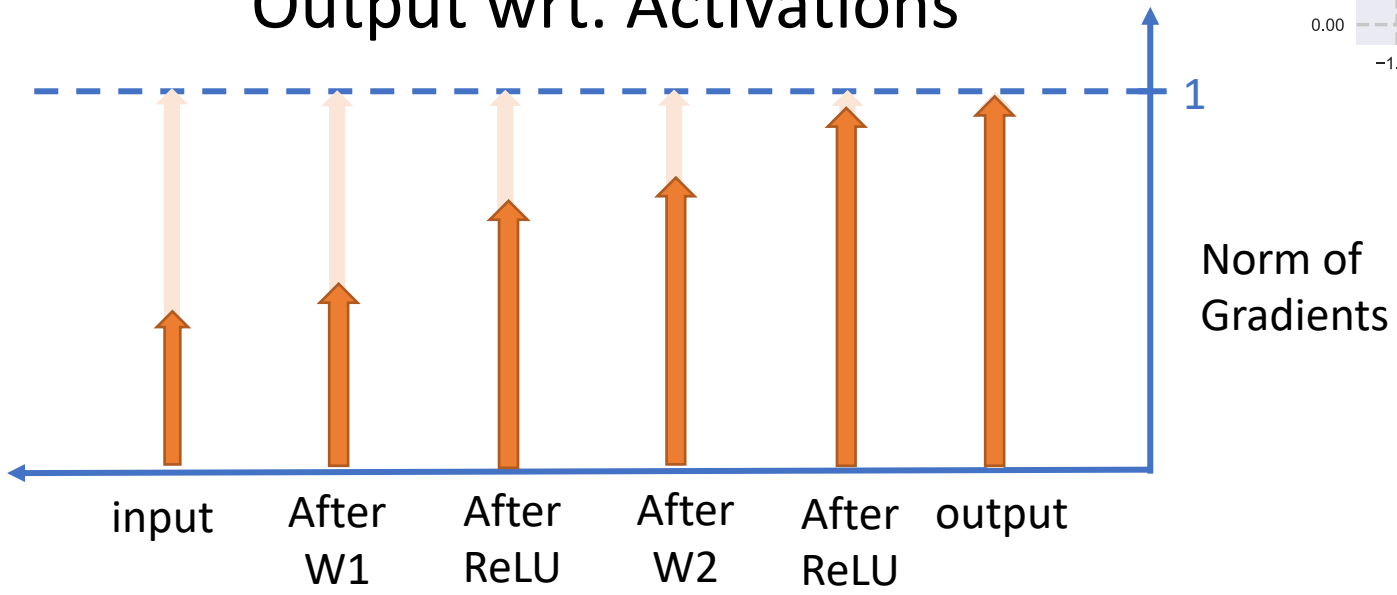
Lipschitz via. Architectural Constraints

- Diagnosing the issue: Inspect gradient norms!



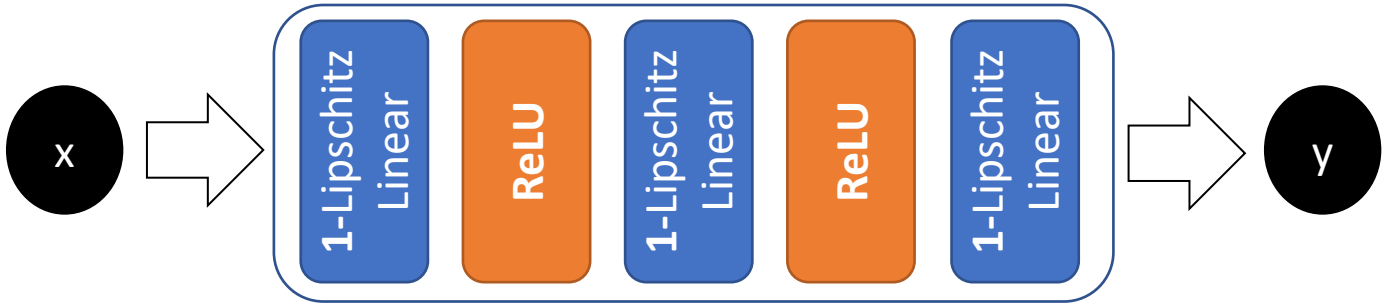


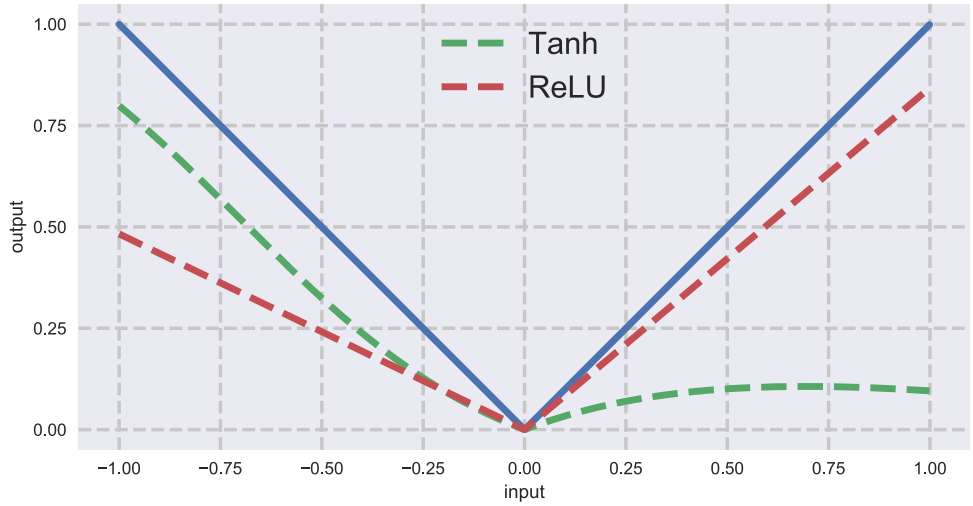
Gradient Norms of Output wrt. Activations



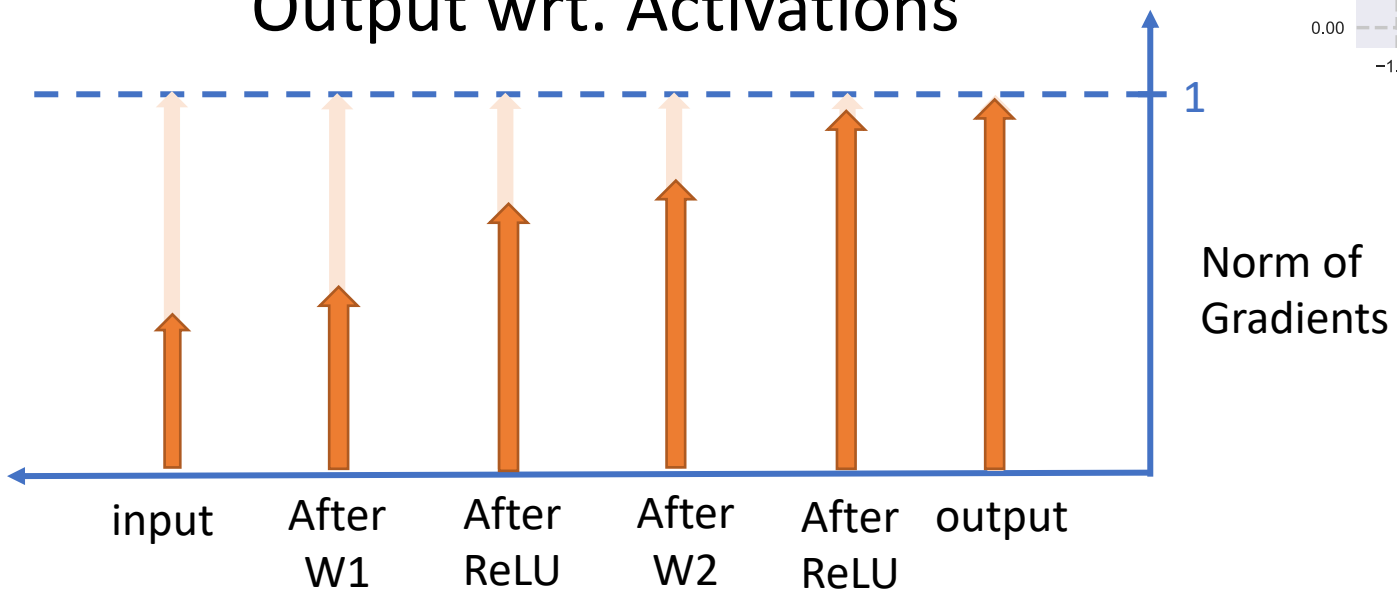
Lipschitz via. Architectural Constraints

- Diagnosing the issue: Inspect gradient norms!





Gradient Norms of Output wrt. Activations

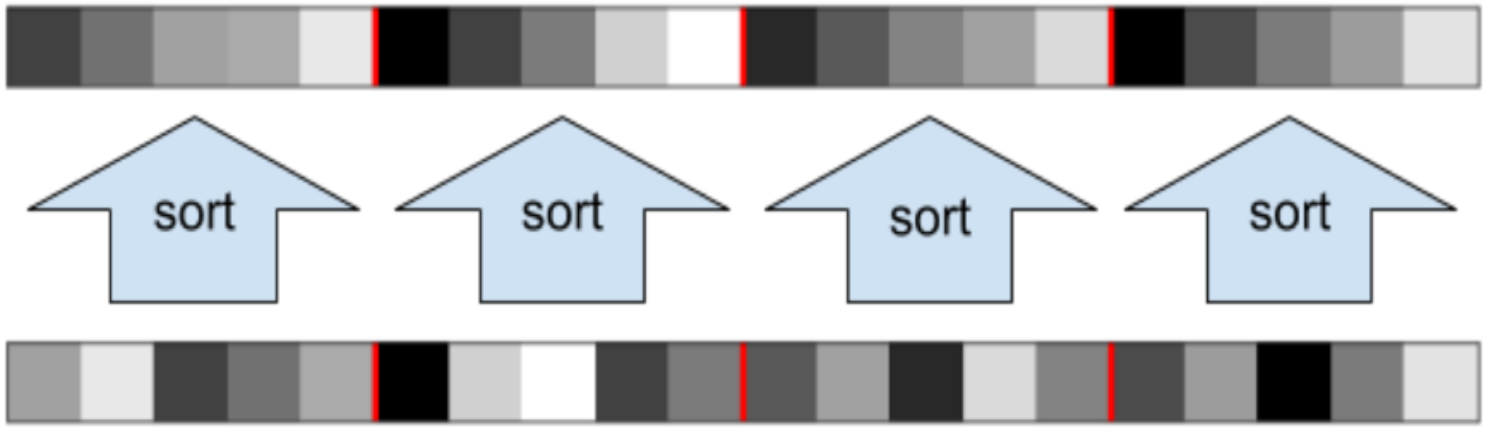


Problem:
Architecture is losing gradient norm!

Solution: Gradient Norm Preservation

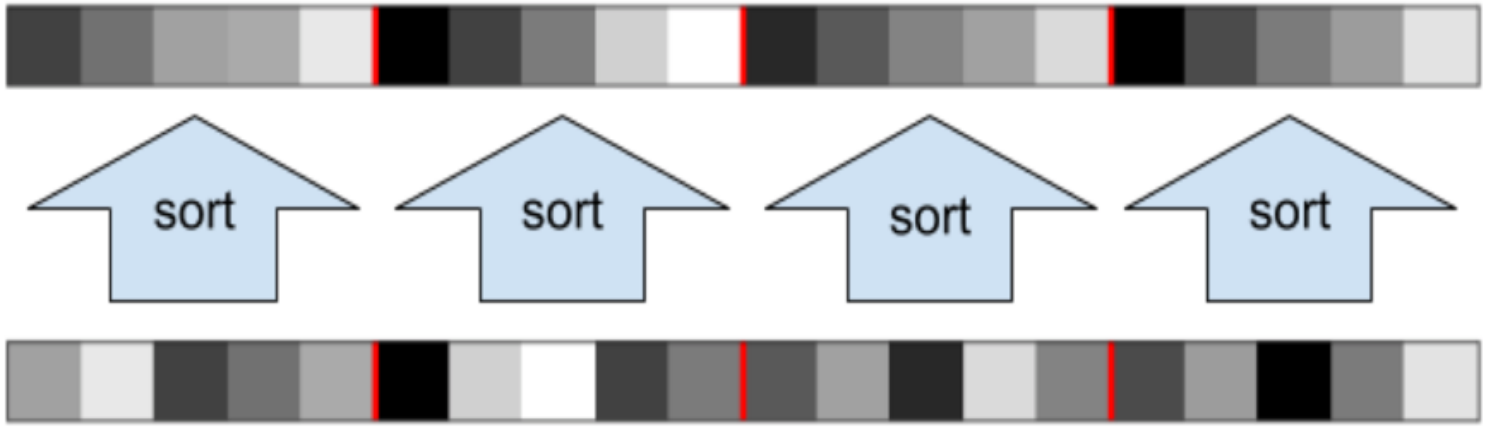
Solution: Gradient Norm Preservation

- Activation: GroupSort



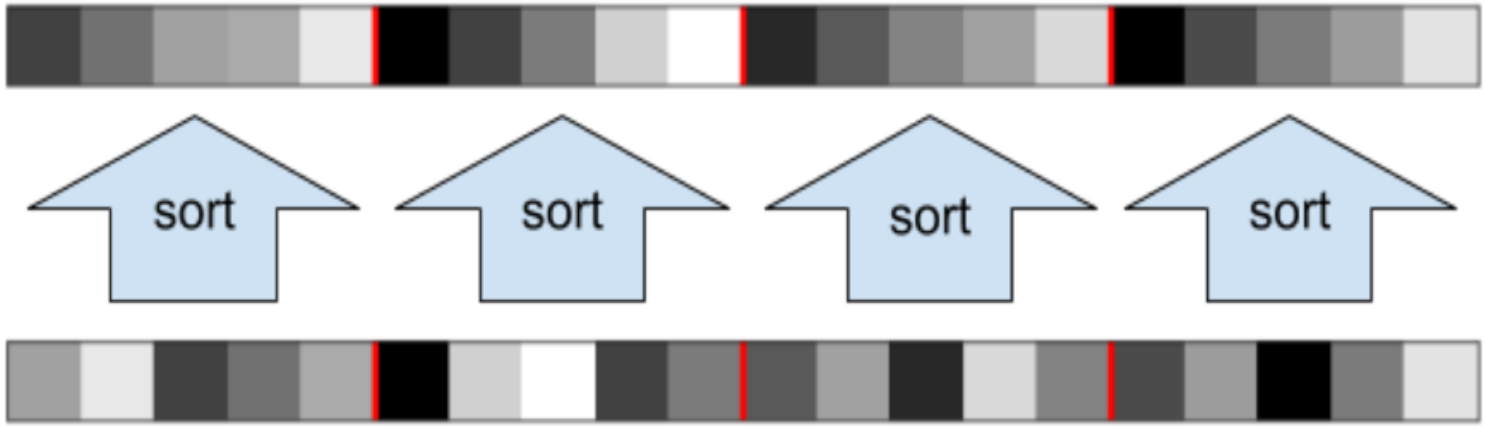
Solution: Gradient Norm Preservation

- Activation: GroupSort
  - Nonlinear, continuous and differentiable almost everywhere.
  - **Gradient Norm Preserving**



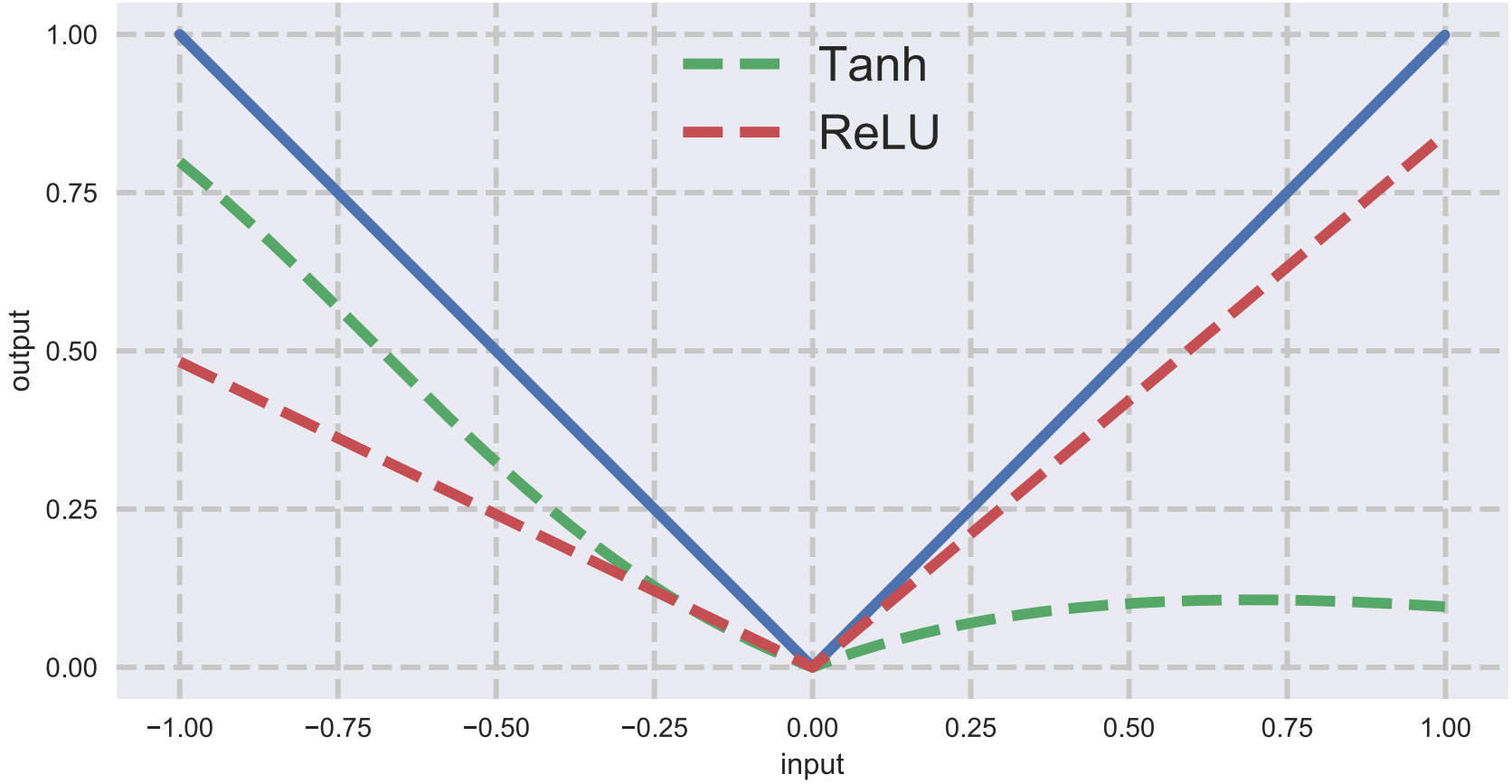
Solution: Gradient Norm Preservation

- Activation: GroupSort
  - Nonlinear, continuous and differentiable almost everywhere.
  - **Gradient Norm Preserving**

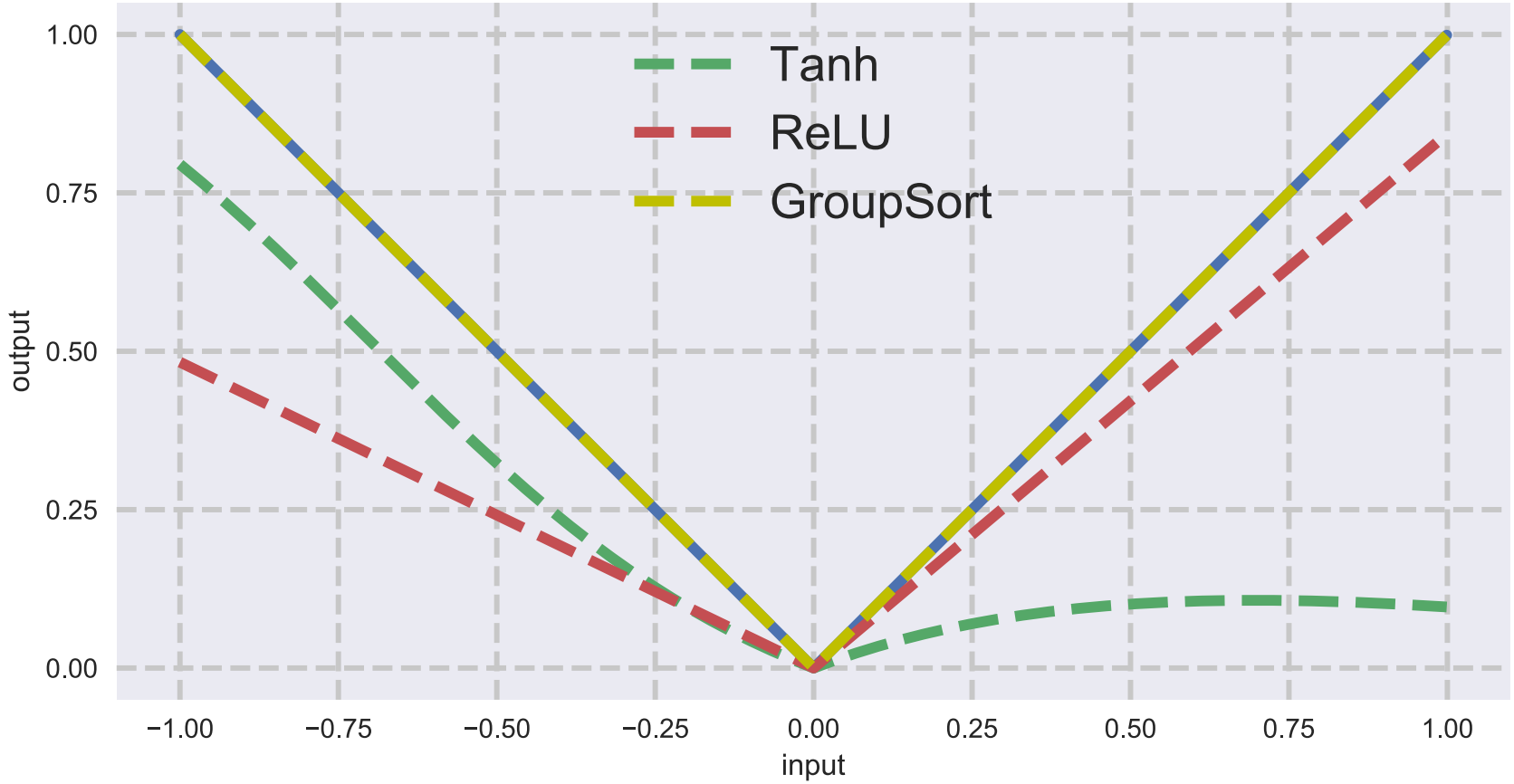


- Linear Transformation: Described in the paper.

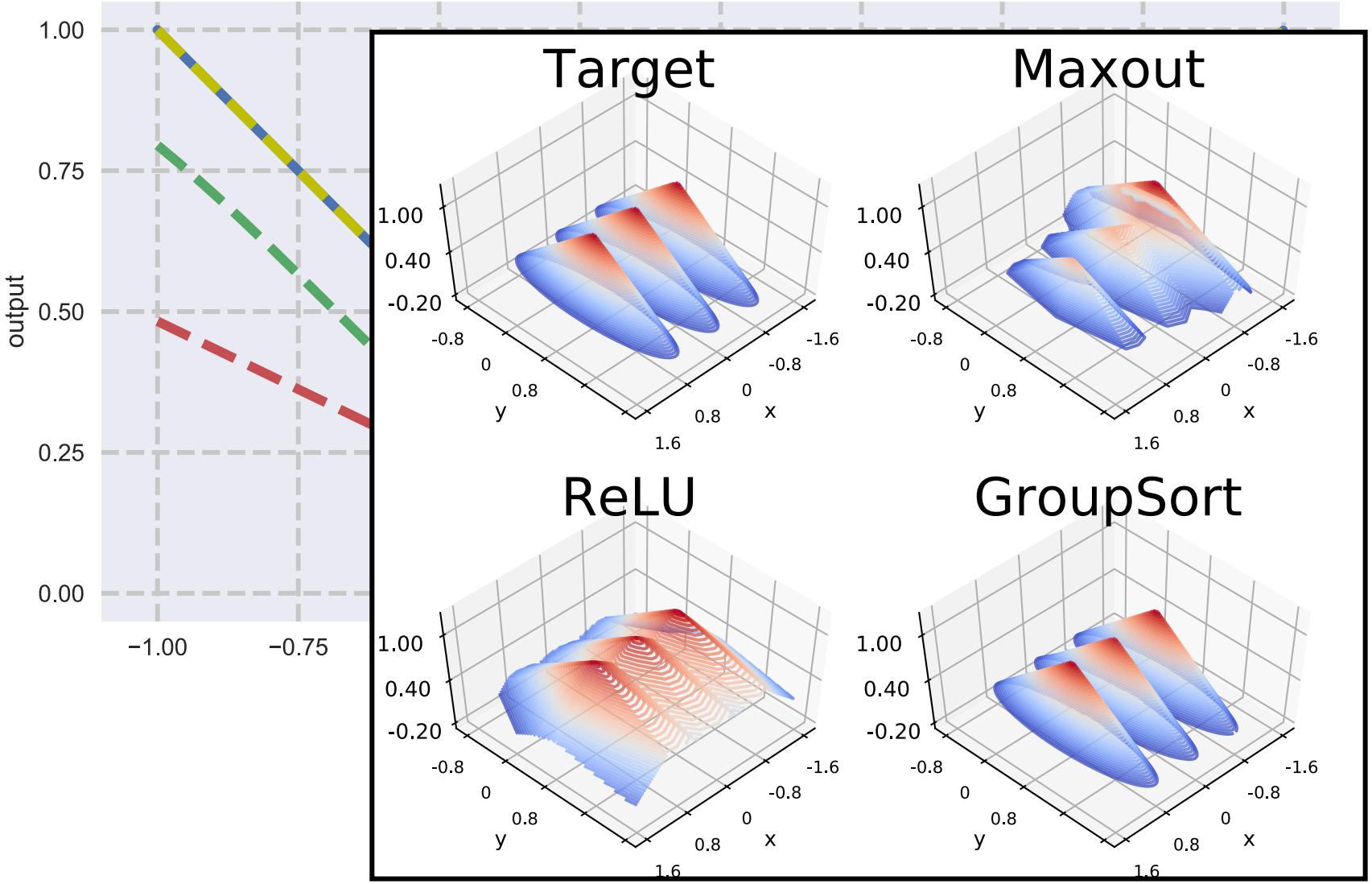
Gradient Norm Preservation => Expressive Power



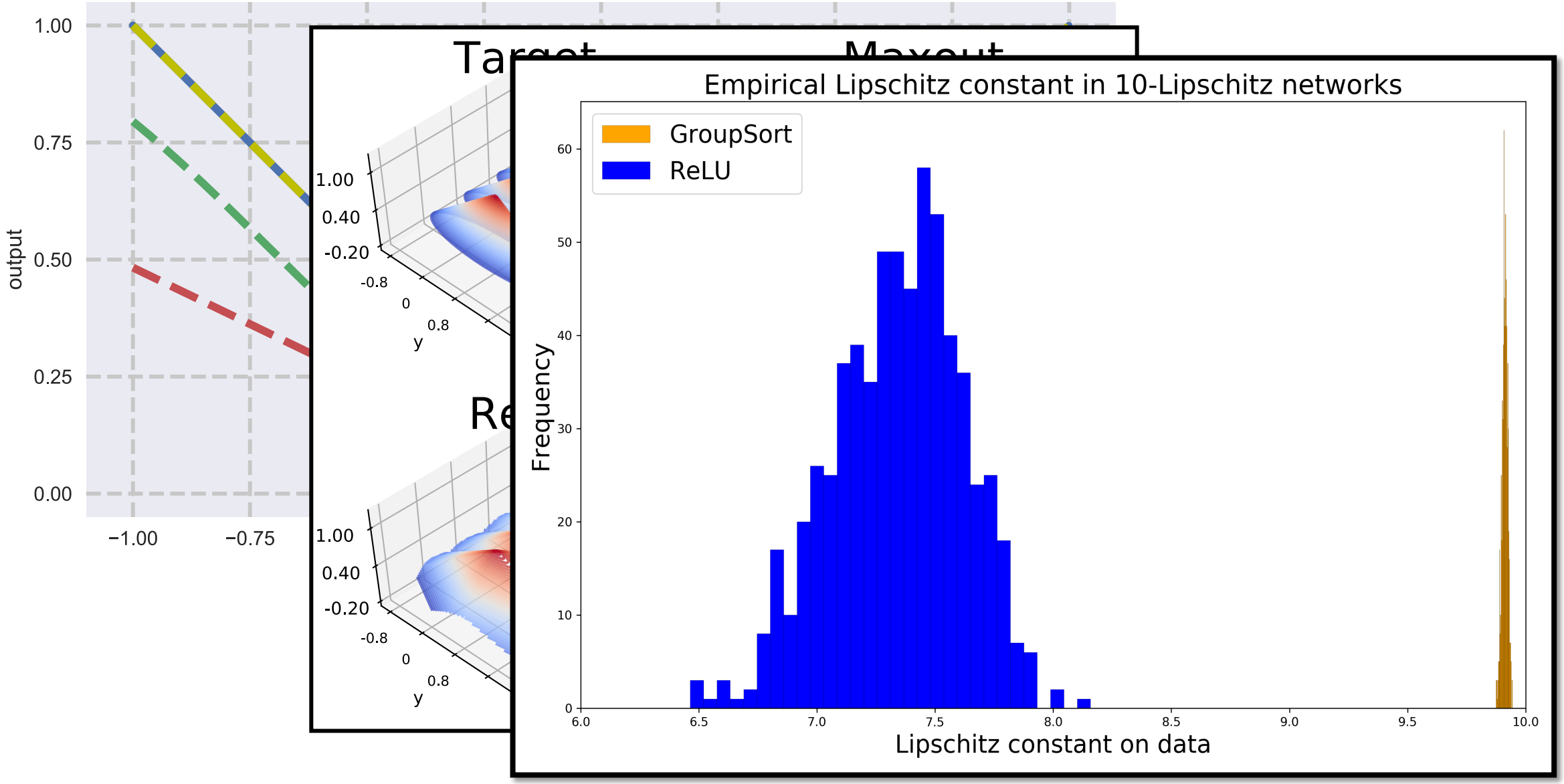
Gradient Norm Preservation => Expressive Power



Gradient Norm Preservation => Expressive Power



Gradient Norm Preservation => Expressive Power



Universal Lipschitz Function Approximation

- Norm constrained GroupSort architectures can recover Universal Lipschitz Function Approximation!

Subtleties and details in the paper/poster

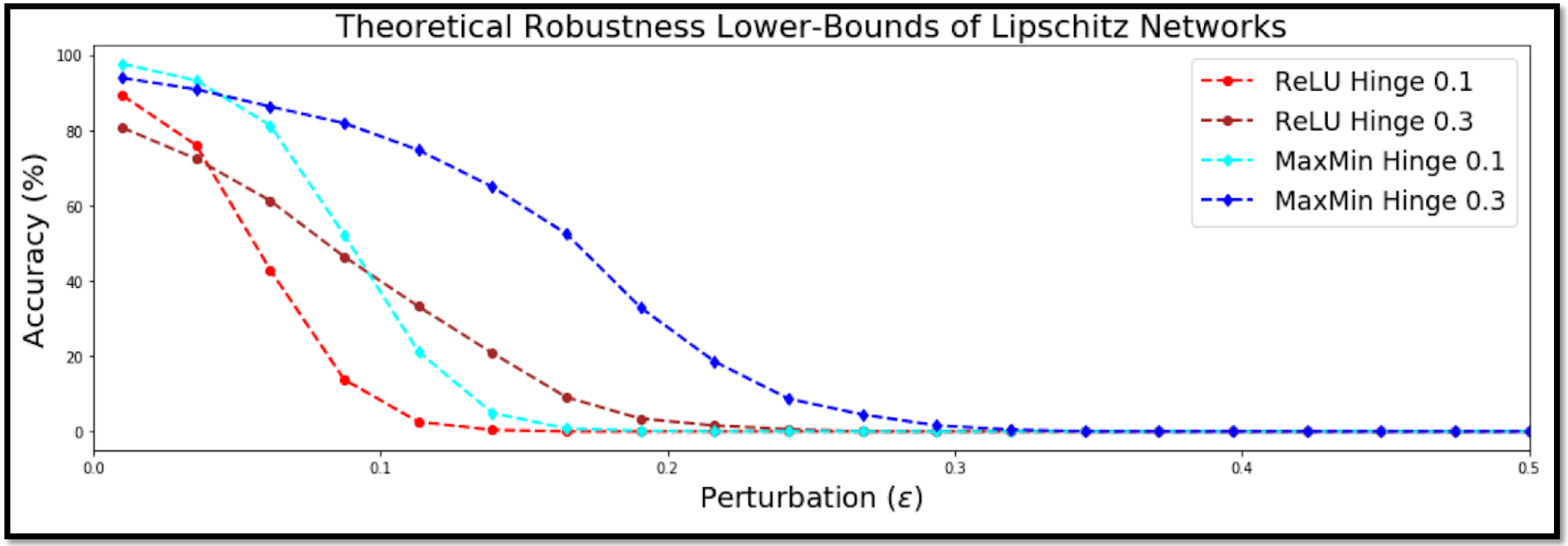
Wasserstein Distance Estimation

- Much tighter estimates of Wasserstein distance
- Training Wasserstein GANs (Arjovsky et. al. 2017)

	Linear	MNIST	CIFAR10
ReLU	Spectral	0.95 ± 0.01	1.12 ± 0.02
Maxout	Spectral	1.20 ± 0.03	1.40 ± 0.01
MaxMin	Spectral	1.36 ± 0.07	1.62 ± 0.04
GroupSort(4)	Spectral	1.64 ± 0.02	1.63 ± 0.03
GroupSort(9)	Spectral	1.70 ± 0.02	1.41 ± 0.04
ReLU	Björck	1.40 ± 0.01	1.39 ± 0.01
Maxout	Björck	1.95 ± 0.01	1.76 ± 0.02
MaxMin	Björck	2.16 ± 0.01	2.08 ± 0.02
GroupSort(4)	Björck	$\mathbf{2.31 \pm 0.01}$	2.17 ± 0.02
GroupSort(9)	Björck	2.31 ± 0.01	$\mathbf{2.23 \pm 0.02}$

Provable Adversarial Robustness

- L-inf constrained GroupSort networks + multi-class hinge loss gets us provable adversarial robustness with little hit to accuracy.



Main Contributions

Propose an Lipschitz GroupSort Networks that

- Buy us expressivity via. Gradient norm preservation.
- Can recover Universal Lipschitz function approximation.

Apply GroupSort Networks to

- Train classifiers provably robust to adversarial perturbations.
- Obtain tight estimates of Wasserstein distance.

UNIVERSITY OF TORONTO

VECTOR INSTITUTE

Sorting Out Lipschitz Function Approximation

Cem Anil*

James Lucas*

Roger Grosse

*Equal contribution

Pacific Ballroom

Poster #15

(6:30 – 9:00 PM)