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.

\[
\| f(x_2) - f(x_1) \|_p \leq K \| x_2 - x_1 \|_p
\]

- Norm of Output Change
- Lipschitz Constant
- Norm of Input Change
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
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
Main Contributions

Propose an expressive Lipschitz constrained architecture that

• Overcomes a previously unidentified limitation in prior art.
• Can recover 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 Network
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.

1-Lipschitz Linear tanh 1-Lipschitz Linear tanh 1-Lipschitz Linear

1-Lipschitz Network

Graph showing the output for input values ranging from -1.00 to 1.00.
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?

1-Lipschitz Linear ReLU 1-Lipschitz Linear ReLU 1-Lipschitz Linear

1-Lipschitz Network

???
Lipschitz via. Architectural Constraints

- Diagnosing the issue: Inspect gradient norms!

Gradient Norms of Output wrt. Activations

Graph showing the norm of gradients with different layers and activation functions.
Lipschitz via. Architectural Constraints

- Diagnosing the issue: **Inspect gradient norms!**

Gradient Norms of Output wrt. Activations

![Diagram showing gradient norms](image-url)
Lipschitz via. Architectural Constraints

- Diagnosing the issue: Inspect gradient norms!

Gradient Norms of Output wrt. Activations

Input $x$ to output $y$ through layers:
- 1-Lipschitz Linear
- ReLU
- 1-Lipschitz Linear
- ReLU
- 1-Lipschitz Linear

Graph showing gradient norms at different layers, indicating Lipschitz behavior through architectural constraints.
Lipschitz via. Architectural Constraints

• Diagnosing the issue: **Inspect gradient norms!**

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 $\Rightarrow$ Expressive Power
Gradient Norm Preservation $\Rightarrow$ Expressive Power
Gradient Norm Preservation $\Rightarrow$ Expressive Power
Gradient Norm Preservation $\Rightarrow$ 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)

<table>
<thead>
<tr>
<th>Linear</th>
<th>MNIST</th>
<th>CIFAR10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReLU</td>
<td>Spectral</td>
<td>0.95 ± 0.01</td>
</tr>
<tr>
<td>Maxout</td>
<td>Spectral</td>
<td>1.20 ± 0.03</td>
</tr>
<tr>
<td>MaxMin</td>
<td>Spectral</td>
<td>1.36 ± 0.07</td>
</tr>
<tr>
<td>GroupSort(4)</td>
<td>Spectral</td>
<td>1.64 ± 0.02</td>
</tr>
<tr>
<td>GroupSort(9)</td>
<td>Spectral</td>
<td>1.70 ± 0.02</td>
</tr>
<tr>
<td>ReLU</td>
<td>Björck</td>
<td>1.40 ± 0.01</td>
</tr>
<tr>
<td>Maxout</td>
<td>Björck</td>
<td>1.95 ± 0.01</td>
</tr>
<tr>
<td>MaxMin</td>
<td>Björck</td>
<td>2.16 ± 0.01</td>
</tr>
<tr>
<td>GroupSort(4)</td>
<td>Björck</td>
<td>2.31 ± 0.01</td>
</tr>
<tr>
<td>GroupSort(9)</td>
<td>Björck</td>
<td>2.31 ± 0.01</td>
</tr>
</tbody>
</table>
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.
Sorting Out
Lipschitz Function Approximation

Cem Anil*
James Lucas*
Roger Grosse

*Equal contribution

Pacific Ballroom
Poster #15
(6:30 – 9:00 PM)