Approximation theory for 1-Lipschitz ResNets

Davide Murari
dm2011@cam.ac.uk

Department of Applied Mathematics and Theoretical Physics
University of Cambridge.

In collaboration with Takashi Furuya and Carola-Bibiane Schönlieb

Why 1-Lipschitz neural networks? $\|F(y) - F(x)\|_2 \leq \|y - x\|_2$

Adversarial robustness

1-Lipschitz $\implies$ lower sensitivity to input perturbation

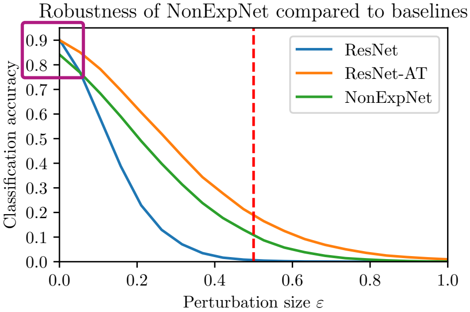
Wasserstein Generative Adversarial Networks (Kantorovich-Rubinstein duality)

$$W_1(\mu, \nu) = \sup_{\substack{f:\mathcal{X}\to\mathbb{R} \\ f\ 1-\text{Lipschitz}}} \mathbb{E}_{X\sim\mu}[f(X)] - \mathbb{E}_{Y\sim\nu}[f(Y)].$$

Convergent fixed point iterations

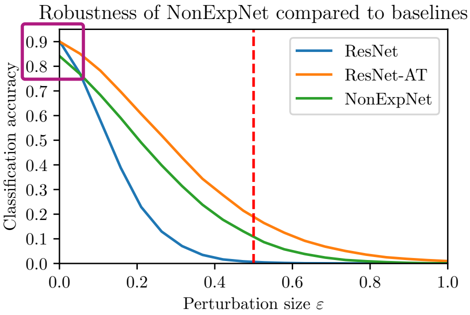
$T_\alpha(x) = (1-\alpha)x + \alpha F(x)$, $\alpha \in (0,1)$, F 1-Lipschitz $\implies x_{k+1} = T_\alpha(x_k) \to \mathrm{Fix}(F)$.

Motivation



[a] Ferdia Sherry et al. "Designing stable neural networks using convex analysis and ODEs". In: *Physica D: Nonlinear Phenomena* 463 (2024), p. 134159.

Motivation

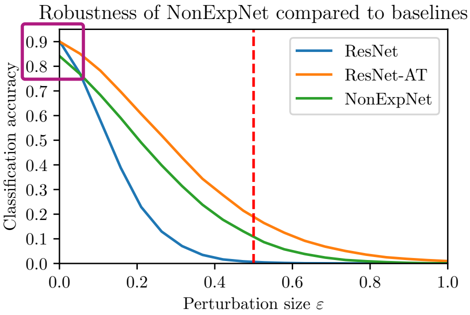


There exists a 1-Lipschitz network as accurate as the ResNet:

$$\mathcal{N} := \mathrm{ResNet}/\mathrm{Lip}(\mathrm{ResNet}).$$

[a] Sherry et al., "Designing stable neural networks using convex analysis and ODEs".

Motivation



There exists a 1-Lipschitz network as accurate as the ResNet:

$$\mathcal{N} := \mathrm{ResNet}/\mathrm{Lip}(\mathrm{ResNet}).$$

This experiment motivates the following two research questions:

1. Is our nonexpansive network[a] not expressive enough?
2. How can we improve the trainability of 1-Lipschitz networks?

[a] Sherry et al., "Designing stable neural networks using convex analysis and ODEs".

The 1-Lipschitz residual architecture

Non-expansive gradient flows

Gradient flows on $\mathbb{R}^d$

Consider the convex scalar function[a] $V_\theta(x) = \|\mathrm{ReLU}(Wx + b)\|_2^2/2$. Define

$$\mathcal{F}_\theta(x) = -\nabla V_\theta(x) = -W^\top \mathrm{ReLU}(Wx + b).$$

[a] $W \in \mathbb{R}^{h\times d}$, $b \in \mathbb{R}^h$, $h \in \mathbb{N}$, $\theta = (W, b)$.

Euler step (1-Lipschitz)

If $\tau \in [0, 2/\|W\|_2^2]$, the explicit Euler map $\varphi_\theta^\tau(x) = x + \tau\mathcal{F}_\theta(x)$ is 1-Lipschitz, i.e.,

$$\|\varphi_\theta^\tau(y) - \varphi_\theta^\tau(x)\|_2 \leq \|y - x\|_2, \; x, y \in \mathbb{R}^d.$$

Neural networks based on gradient flows

$$\mathcal{N}_\theta = \pi \circ \varphi_{\theta_L} \circ \ldots \circ \varphi_{\theta_1} \circ Q : \mathbb{R}^d \to \mathbb{R}^c$$

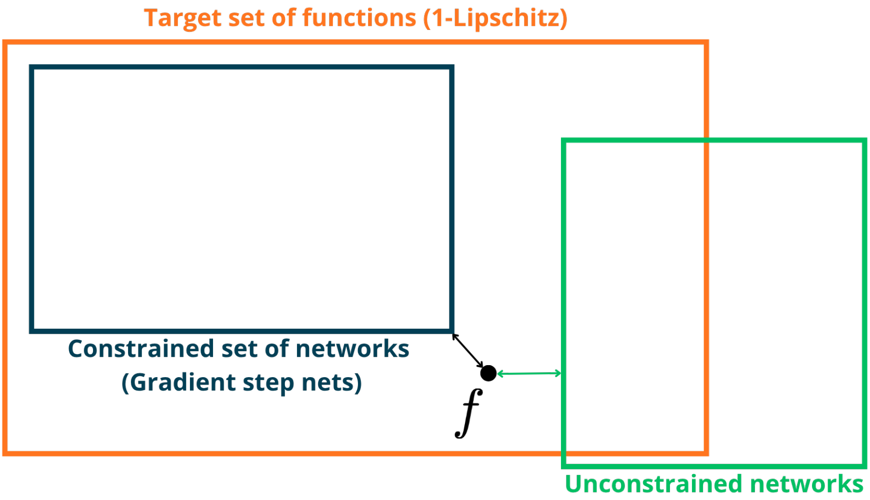
$$\varphi_{\theta_\ell} \in \mathcal{E}_h := \Big\{ \varphi : \mathbb{R}^h \to \mathbb{R}^h \,\Big|\, \varphi(x) = x - \tau W^\top \mathrm{ReLU}(Wx + b),\ W \in \mathbb{R}^{h' \times h}, b \in \mathbb{R}^{h'},$$
$$h' \in \mathbb{N}, \tau \in [0, 2/\|W\|_2^2] \Big\},$$

$Q : \mathbb{R}^d \to \mathbb{R}^h,\ \pi : \mathbb{R}^h \to \mathbb{R}^c$ affine maps.

Universal approximation theorem
Unbounded width and depth

Davide Murari, Takashi Furuya, and Carola-Bibiane Schönlieb. "Approximation theory for 1-Lipschitz ResNets". In: *NeurIPS* (2025)

The purpose of our theoretical analysis



Statement of the theorem

$$\mathcal{G}_d(\mathcal{X}) = \Big\{\pi \circ \varphi_{\theta_L} \circ \ldots \circ \varphi_{\theta_1} \circ Q : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R} \;\Big|\; \varphi_{\theta_\ell} \in \mathcal{E}_h,\ \ell = 1, \ldots, L,\ h, L \in \mathbb{N},$$
$$Q : \mathbb{R}^d \to \mathbb{R}^h,\ \pi : \mathbb{R}^h \to \mathbb{R},\ Q \text{ and } \pi \text{ affine}\Big\}.$$

$$\mathcal{C}_1(\mathcal{X}, \mathbb{R}) \iff 1 - \text{Lipschitz functions from } \mathcal{X} \text{ to } \mathbb{R}.$$

Statement of the theorem

$$\mathcal{G}_d(\mathcal{X}) = \Big\{\pi \circ \varphi_{\theta_L} \circ \dots \circ \varphi_{\theta_1} \circ Q : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R} \,\Big|\, \varphi_{\theta_\ell} \in \mathcal{E}_h,\ \ell = 1, \dots, L,\ h, L \in \mathbb{N},$$
$$Q : \mathbb{R}^d \to \mathbb{R}^h,\ \pi : \mathbb{R}^h \to \mathbb{R},\ Q \text{ and } \pi \text{ affine}\Big\}.$$

$$\mathcal{C}_1(\mathcal{X}, \mathbb{R}) \iff 1-\text{Lipschitz functions from } \mathcal{X} \text{ to } \mathbb{R}.$$

Universal approximation theorem

$\varepsilon > 0$, $\mathcal{X} \subset \mathbb{R}^d$ compact, $g \in \mathcal{C}_1(\mathcal{X}, \mathbb{R})$. Then, there exists $f \in \mathcal{G}_d(\mathcal{X}) \cap \mathcal{C}_1(\mathcal{X}, \mathbb{R})$ with

$$\max_{x \in \mathcal{X}} |f(x) - g(x)| < \varepsilon.$$

Two proof techniques

We prove this theorem in two different ways:

1. Show that each scalar piecewise affine 1-Lipschitz function belongs to $\mathcal{G}_d(\mathcal{X})$.

Two proof techniques

We prove this theorem in two different ways:

1. Show that each scalar piecewise affine 1-Lipschitz function belongs to $\mathcal{G}_d(\mathcal{X})$.

2. We prove that the **Restricted Stone-Weierstrass Theorem** holds.

Restricted Stone-Weierstrass Theorem

Let $\mathcal{X} \subset \mathbb{R}^d$ be compact and have at least two points. Let $\mathcal{A} \subset \mathcal{C}_1(\mathcal{X}, \mathbb{R})$ be a lattice[a] separating the points[b] of $\mathcal{X}$. Then $\mathcal{A}$ satisfies the universal approximation property for $\mathcal{C}_1(\mathcal{X}, \mathbb{R})$.

[a] Closed under max and min.

[b] For any pair of distinct elements $x, y \in \mathcal{X}$ and real numbers $a, b \in \mathbb{R}$ with $|a-b| \leq \|y-x\|_2$, there is an $f \in \mathcal{A}$ such that $f(x) = a$ and $f(y) = b$.

Representation of piecewise affine functions

Representation theorem

$$\mathcal{G}_d(\mathbb{R}^d) \cap \mathcal{C}_1(\mathbb{R}^d, \mathbb{R}) = 1 - \text{Lipschitz piecewise affine functions from } \mathbb{R}^d \text{ to } \mathbb{R}.$$

Representation of piecewise affine functions

Representation theorem

$$\mathcal{G}_d(\mathbb{R}^d) \cap \mathcal{C}_1(\mathbb{R}^d, \mathbb{R}) = 1 - \text{Lipschitz piecewise affine functions from } \mathbb{R}^d \text{ to } \mathbb{R}.$$

max-min representation

$f : \mathbb{R}^d \to \mathbb{R}$ piecewise affine 1-Lipschitz. Then[a]

$$f(x) = \max\{f_1(x), \dots, f_k(x)\},$$
$$f_i(x) = \min\{a_{i,1}^\top x + b_{i,1}, \dots, a_{i,l_i}^\top x + b_{i,l_i}\}.$$

[a] $k, l_i \in \mathbb{N}$, $a_{i,j} \in \mathbb{R}^d$, $b_{i,j} \in \mathbb{R}$, $\|a_{i,j}\|_2 \leq 1$

Three important identities with ReLU

Let $x, x_1, x_2 \in \mathbb{R}$.

- **Representation of the identity:**

$$x = \mathrm{ReLU}(x) - \mathrm{ReLU}(-x).$$

Three important identities with ReLU

Let $x, x_1, x_2 \in \mathbb{R}$.

- **Representation of the identity:**

$$x = \mathrm{ReLU}(x) - \mathrm{ReLU}(-x).$$

- **Representation of the maximum:**

$$\max\{x_1, x_2\} = x_1 + \mathrm{ReLU}(x_2 - x_1) = x_2 + \mathrm{ReLU}(x_1 - x_2).$$

- **Representation of the minimum:**

$$\min\{x_1, x_2\} = x_1 - \mathrm{ReLU}(x_1 - x_2) = x_2 - \mathrm{ReLU}(x_2 - x_1).$$

A warm-up example

What are the linear maps we can represent with our layers?

$M \in \mathbb{R}^{h\times h}$ satisfying $M^\top = M$ and $\lambda_i(M) \in [0,1]$.

$$M = R^\top (I_h - \Lambda^2) R = I_h - (\Lambda R)^\top (\Lambda R),$$

where $R \in \mathcal{O}(h)$, and $\Lambda^2 = \mathrm{diag}(1-\lambda_1, \dots, 1-\lambda_h)$.

A warm-up example

What are the linear maps we can represent with our layers?

$M \in \mathbb{R}^{h\times h}$ satisfying $M^\top = M$ and $\lambda_i(M) \in [0,1]$.

$$M = R^\top(I_h - \Lambda^2)R = I_h - (\Lambda R)^\top(\Lambda R),$$

where $R \in \mathcal{O}(h)$, and $\Lambda^2 = \mathrm{diag}(1-\lambda_1,\dots,1-\lambda_h)$.

$$\begin{aligned} Mx &= x - (\Lambda R)^\top(\Lambda R)x \\ &= x - \begin{bmatrix}(\Lambda R)^\top & -(\Lambda R)^\top\end{bmatrix} \mathrm{ReLU}\left(\begin{bmatrix}\Lambda R \\ -\Lambda R\end{bmatrix} x\right). \end{aligned}$$

Remark: $\|\Lambda R\|_2 \leq 1$.

Proof of the representation theorem: An illustrating example

Target function: Fix $a_1, \dots, a_4 \in \mathbb{R}^d$ with $\|a_i\|_2 \leq 1$. Consider the 1-Lipschitz function

$$f(x) = \min\{\max\{a_1^\top x, a_2^\top x\}, \max\{a_3^\top x, a_4^\top x\}\}.$$

Proof of the representation theorem: An illustrating example

Target function: Fix $a_1, \dots, a_4 \in \mathbb{R}^d$ with $\|a_i\|_2 \leq 1$. Consider the 1-Lipschitz function

$$f(x) = \min\{\max\{a_1^\top x, a_2^\top x\}, \max\{a_3^\top x, a_4^\top x\}\}.$$

STEP 1: Extract all the affine pieces

$$Q(x) = \begin{bmatrix} a_1^\top x \\ a_2^\top x \\ a_3^\top x \\ a_4^\top x \end{bmatrix}$$

Proof: An illustrating example

STEP 2: Extract the max and min using the gradient steps

$$\varphi_{\theta_1}(v) := \begin{bmatrix} \max\{v_1, v_2\} \\ \min\{v_1, v_2\} \\ v_3 \\ v_4 \end{bmatrix} = v - \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \mathrm{ReLU}\left(\begin{bmatrix} -1 & 1 & 0 & 0 \end{bmatrix} v\right), \quad v \in \mathbb{R}^4.$$

$$\varphi_{\theta_1}(Qx) = \begin{bmatrix} \max\{a_1^\top x, a_2^\top x\} \\ \min\{a_1^\top x, a_2^\top x\} \\ a_3^\top x \\ a_4^\top x \end{bmatrix}$$

Proof: An illustrating example

STEP 3: Pick φ_{θ_2} to assemble $\max\{a_3^\top x, a_4^\top x\}$ and analogously φ_{θ_3} so that:

$$\varphi_{\theta_3} \circ \varphi_{\theta_2} \circ \varphi_{\theta_1}(Qx) = \begin{bmatrix} \min\{\max\{a_1^\top x, a_2^\top x\}, \max\{a_3^\top x, a_4^\top x\}\} \\ \max\{\max\{a_1^\top x, a_2^\top x\}, \max\{a_3^\top x, a_4^\top x\}\} \\ \max\{a_3^\top x, a_4^\top x\} \\ \min\{a_3^\top x, a_4^\top x\} \end{bmatrix}$$

Proof: An illustrating example

STEP 3: Pick φ_{θ_2} to assemble $\max\{a_3^\top x, a_4^\top x\}$ and analogously φ_{θ_3} so that:

$$\varphi_{\theta_3} \circ \varphi_{\theta_2} \circ \varphi_{\theta_1}(Qx) = \begin{bmatrix} \min\{\max\{a_1^\top x, a_2^\top x\}, \max\{a_3^\top x, a_4^\top x\}\} \\ \max\{\max\{a_1^\top x, a_2^\top x\}, \max\{a_3^\top x, a_4^\top x\}\} \\ \max\{a_3^\top x, a_4^\top x\} \\ \min\{a_3^\top x, a_4^\top x\} \end{bmatrix}$$

STEP 4: Extract the first component to conclude

$$e_1^\top \varphi_{\theta_2} \circ \varphi_{\theta_1}(Qx) = f(x).$$

Value and limitation of this result

Values:

- We have learned how to think of our networks: they are nothing more than a generic piecewise affine 1-Lipschitz function.

Value and limitation of this result

Values:

- We have learned how to think of our networks: they are nothing more than a generic piecewise affine 1-Lipschitz function.
- We understood that the seemingly too constrained architecture, still has enough flexibility.

Value and limitation of this result

Values:

- We have learned how to think of our networks: they are nothing more than a generic piecewise affine 1-Lipschitz function.
- We understood that the seemingly too constrained architecture, still has enough flexibility.

Limitations:

- No control over how **wide** and deep the network is.

Value and limitation of this result

Values:

- We have learned how to think of our networks: they are nothing more than a generic piecewise affine 1-Lipschitz function.
- We understood that the seemingly too constrained architecture, still has enough flexibility.

Limitations:

- No control over how **wide** and deep the network is.
- We do not have a concrete (implementable) way to constrain Q while preserving the model's flexibility.

Universal approximation theorem

Unbounded depth and <u>fixed width</u>

A revised set of networks

Fix $h \geq 3$. We now consider the set

$$\widetilde{\mathcal{G}}_{d,h}(\mathcal{X}) := \Big\{ v^\top \circ \varphi_{\theta_L} \circ A_{L-1} \circ \cdots \circ A_1 \circ \varphi_{\theta_1} \circ Q : \mathcal{X} \to \mathbb{R} \;\Big|\; m = (1,1,1,h-3), \\ Q \in \widetilde{\mathcal{R}}_{d,m}, v \in \mathbb{R}^h, \|v\|_1 \leq 1, A_1, ..., A_{L-1} \in \widetilde{\mathcal{L}}_m, \varphi_{\theta_\ell} \in \widetilde{\mathcal{E}}_{h-3},\, L \in \mathbb{N} \Big\}.$$

A revised set of networks

Fix $h \geq 3$. We now consider the set

$$\widetilde{\mathcal{G}}_{d,h}(\mathcal{X}) := \Big\{ v^\top \circ \varphi_{\theta_L} \circ A_{L-1} \circ \cdots \circ A_1 \circ \varphi_{\theta_1} \circ Q : \mathcal{X} \to \mathbb{R} \,\Big|\, m = (1,1,1,h-3), \\ Q \in \widetilde{\mathcal{R}}_{d,m}, v \in \mathbb{R}^h, \|v\|_1 \leq 1, A_1, \dots, A_{L-1} \in \widetilde{\mathcal{L}}_m, \varphi_{\theta_\ell} \in \widetilde{\mathcal{E}}_{h-3}, \, L \in \mathbb{N} \Big\}.$$

$$\mathcal{L}_m = \left\{ \begin{bmatrix} A_{11} & \dots & A_{1k} \\ \vdots & \ddots & \vdots \\ A_{k1} & \dots & A_{kk} \end{bmatrix} \in \mathbb{R}^{\alpha_m \times \alpha_m} \,\middle|\, A_{ij} \in \mathbb{R}^{m_i \times m_j}, \sum_{j=1}^k \|A_{ij}\|_2 \leq 1, \; i = 1, \dots, k \right\}, \; m \in \mathbb{N}^k,$$

$$\mathcal{R}_{d,m} = \left\{ \begin{bmatrix} B_1^\top & \cdots & B_k^\top \end{bmatrix}^\top \in \mathbb{R}^{\alpha_m \times d} \,\middle|\, B_i \in \mathbb{R}^{m_i \times d}, \; \|B_i\|_2 \leq 1, \; i = 1, \dots, k \right\}, \; \alpha_m := \|m\|_1,$$

A revised set of networks

Fix $h \geq 3$. We now consider the set

$$\widetilde{\mathcal{G}}_{d,h}(\mathcal{X}) := \Big\{ v^\top \circ \varphi_{\theta_L} \circ A_{L-1} \circ \cdots \circ A_1 \circ \varphi_{\theta_1} \circ Q : \mathcal{X} \to \mathbb{R} \,\Big|\, m = (1,1,1,h-3),$$
$$Q \in \widetilde{\mathcal{R}}_{d,m}, v \in \mathbb{R}^h, \|v\|_1 \leq 1, A_1, \dots, A_{L-1} \in \widetilde{\mathcal{L}}_m, \varphi_{\theta_\ell} \in \widetilde{\mathcal{E}}_{h-3}, L \in \mathbb{N} \Big\}.$$

$$\mathcal{L}_m = \left\{ \begin{bmatrix} A_{11} & \dots & A_{1k} \\ \vdots & \ddots & \vdots \\ A_{k1} & \dots & A_{kk} \end{bmatrix} \in \mathbb{R}^{\alpha_m \times \alpha_m} \,\middle|\, A_{ij} \in \mathbb{R}^{m_i \times m_j}, \sum_{j=1}^{k} \|A_{ij}\|_2 \leq 1,\ i = 1, \dots, k \right\}, m \in \mathbb{N}^k,$$

$$\mathcal{R}_{d,m} = \left\{ \begin{bmatrix} B_1^\top & \cdots & B_k^\top \end{bmatrix}^\top \in \mathbb{R}^{\alpha_m \times d} \,\middle|\, B_i \in \mathbb{R}^{m_i \times d}, \|B_i\|_2 \leq 1,\ i = 1, \dots, k \right\}, \alpha_m := \|m\|_1,$$

$$\widetilde{\mathcal{E}}_h = \left\{ \varphi_\theta : \mathbb{R}^{h+3} \to \mathbb{R}^{h+3} \,\middle|\, \varphi_\theta(x) = \begin{bmatrix} \max\{x_1, x_2\} & \min\{x_1, x_2\} & x_3 & \widetilde{\varphi}_\theta(x_{4:})^\top \end{bmatrix}, \ \widetilde{\varphi}_\theta \in \mathcal{E}_h \right\}.$$

Theorem statement

Characterisation of the set of networks

Let $d, h \in \mathbb{N}$ with $h \geq 3$. All the functions in $\widetilde{\mathcal{G}}_{d,h}(\mathbb{R}^d)$ are 1-Lipschitz.

Theorem statement

Characterisation of the set of networks

Let $d, h \in \mathbb{N}$ with $h \geq 3$. All the functions in $\widetilde{\mathcal{G}}_{d,h}(\mathbb{R}^d)$ are 1-Lipschitz.

Representation Theorem

Any piecewise affine 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ can be represented by a network in $\widetilde{\mathcal{G}}_{d,h}(\mathbb{R}^d, \mathbb{R})$ with $h \geq d + 3$.

Theorem statement

Characterisation of the set of networks

Let $d, h \in \mathbb{N}$ with $h \geq 3$. All the functions in $\widetilde{\mathcal{G}}_{d,h}(\mathbb{R}^d)$ are 1-Lipschitz.

Representation Theorem

Any piecewise affine 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ can be represented by a network in $\widetilde{\mathcal{G}}_{d,h}(\mathbb{R}^d, \mathbb{R})$ with $h \geq d + 3$.

Universal Approximation Theorem

Let $d \in \mathbb{N}$, and $\mathcal{X} \subset \mathbb{R}^d$ be compact. The set $\widetilde{\mathcal{G}}_{d,h}(\mathcal{X})$ satisfies the universal approximation property for $\mathcal{C}_1(\mathcal{X}, \mathbb{R})$ if $h \geq d + 3$.

Idea of the proof for the representation theorem

Example: $f(x) = \min\{\max\{a_1^\top x, a_2^\top x\}, \max\{a_3^\top x, a_4^\top x\}\}$:

$$\mathbb{R}^d \ni x \underbrace{\mapsto}_{Q} \begin{bmatrix} a_1^\top x \\ a_2^\top x \\ 0 \\ x \end{bmatrix}$$

Idea of the proof for the representation theorem

Example: $f(x) = \min\{\max\{a_1^\top x, a_2^\top x\}, \max\{a_3^\top x, a_4^\top x\}\}$:

$$\mathbb{R}^d \ni x \underbrace{\mapsto}_{Q} \begin{bmatrix} a_1^\top x \\ a_2^\top x \\ 0 \\ x \end{bmatrix} \underbrace{\mapsto}_{\varphi_{\theta_1}} \begin{bmatrix} \max\{a_1^\top x, a_2^\top x\} \\ \min\{a_1^\top x, a_2^\top x\} \\ 0 \\ x \end{bmatrix}$$

Idea of the proof for the representation theorem

Example: $f(x) = \min\{\max\{a_1^\top x, a_2^\top x\}, \max\{a_3^\top x, a_4^\top x\}\}$:

$$\mathbb{R}^d \ni x \underbrace{\mapsto}_{Q} \begin{bmatrix} a_1^\top x \\ a_2^\top x \\ 0 \\ x \end{bmatrix} \underbrace{\mapsto}_{\varphi_{\theta_1}} \begin{bmatrix} \max\{a_1^\top x, a_2^\top x\} \\ \min\{a_1^\top x, a_2^\top x\} \\ 0 \\ x \end{bmatrix} \underbrace{\mapsto}_{A_1} \begin{bmatrix} a_3^\top x \\ a_4^\top x \\ \max\{a_1^\top x, a_2^\top x\} \\ x \end{bmatrix}$$

Idea of the proof for the representation theorem

Example: $f(x) = \min\{\max\{a_1^\top x, a_2^\top x\}, \max\{a_3^\top x, a_4^\top x\}\}$:

$$\mathbb{R}^d \ni x \underbrace{\mapsto}_{Q} \begin{bmatrix} a_1^\top x \\ a_2^\top x \\ 0 \\ x \end{bmatrix} \underbrace{\mapsto}_{\varphi_{\theta_1}} \begin{bmatrix} \max\{a_1^\top x, a_2^\top x\} \\ \min\{a_1^\top x, a_2^\top x\} \\ 0 \\ x \end{bmatrix} \underbrace{\mapsto}_{A_1} \begin{bmatrix} a_3^\top x \\ a_4^\top x \\ \max\{a_1^\top x, a_2^\top x\} \\ x \end{bmatrix} \underbrace{\mapsto}_{\varphi_{\theta_2}} \begin{bmatrix} \max\{a_3^\top x, a_4^\top x\} \\ \min\{a_3^\top x, a_4^\top x\} \\ \max\{a_1^\top x, a_2^\top x\} \\ x \end{bmatrix}$$

Idea of the proof for the representation theorem

Example: $f(x) = \min\{\max\{a_1^\top x, a_2^\top x\}, \max\{a_3^\top x, a_4^\top x\}\}$:

$$\mathbb{R}^d \ni x \underbrace{\mapsto}_{Q} \begin{bmatrix} a_1^\top x \\ a_2^\top x \\ 0 \\ x \end{bmatrix} \underbrace{\mapsto}_{\varphi_{\theta_1}} \begin{bmatrix} \max\{a_1^\top x, a_2^\top x\} \\ \min\{a_1^\top x, a_2^\top x\} \\ 0 \\ x \end{bmatrix} \underbrace{\mapsto}_{A_1} \begin{bmatrix} a_3^\top x \\ a_4^\top x \\ \max\{a_1^\top x, a_2^\top x\} \\ x \end{bmatrix} \underbrace{\mapsto}_{\varphi_{\theta_2}} \begin{bmatrix} \max\{a_3^\top x, a_4^\top x\} \\ \min\{a_3^\top x, a_4^\top x\} \\ \max\{a_1^\top x, a_2^\top x\} \\ x \end{bmatrix}$$

$$\underbrace{\mapsto}_{A_2} \begin{bmatrix} \max\{a_1^\top x, a_2^\top x\} \\ \max\{a_3^\top x, a_4^\top x\} \\ 0 \\ x \end{bmatrix}$$

Idea of the proof for the representation theorem

Example: $f(x) = \min\{\max\{a_1^\top x, a_2^\top x\}, \max\{a_3^\top x, a_4^\top x\}\}$:

$$\mathbb{R}^d \ni x \underbrace{\mapsto}_{Q} \begin{bmatrix} a_1^\top x \\ a_2^\top x \\ 0 \\ x \end{bmatrix} \underbrace{\mapsto}_{\varphi_{\theta_1}} \begin{bmatrix} \max\{a_1^\top x, a_2^\top x\} \\ \min\{a_1^\top x, a_2^\top x\} \\ 0 \\ x \end{bmatrix} \underbrace{\mapsto}_{A_1} \begin{bmatrix} a_3^\top x \\ a_4^\top x \\ \max\{a_1^\top x, a_2^\top x\} \\ x \end{bmatrix} \underbrace{\mapsto}_{\varphi_{\theta_2}} \begin{bmatrix} \max\{a_3^\top x, a_4^\top x\} \\ \min\{a_3^\top x, a_4^\top x\} \\ \max\{a_1^\top x, a_2^\top x\} \\ x \end{bmatrix}$$

$$\underbrace{\mapsto}_{A_2} \begin{bmatrix} \max\{a_1^\top x, a_2^\top x\} \\ \max\{a_3^\top x, a_4^\top x\} \\ 0 \\ x \end{bmatrix} \underbrace{\mapsto}_{\varphi_{\theta_3}} \begin{bmatrix} \max\{a_1^\top x, a_2^\top x, a_3^\top x, a_4^\top x\} \\ \min\{\max\{a_1^\top x, a_2^\top x\}, \max\{a_3^\top x, a_4^\top x\}\} \\ 0 \\ x \end{bmatrix}$$

Idea of the proof for the representation theorem

Example: $f(x) = \min\{\max\{a_1^\top x, a_2^\top x\}, \max\{a_3^\top x, a_4^\top x\}\}$:

$$\mathbb{R}^d \ni x \underbrace{\mapsto}_{Q} \begin{bmatrix} a_1^\top x \\ a_2^\top x \\ 0 \\ x \end{bmatrix} \underbrace{\mapsto}_{\varphi_{\theta_1}} \begin{bmatrix} \max\{a_1^\top x, a_2^\top x\} \\ \min\{a_1^\top x, a_2^\top x\} \\ 0 \\ x \end{bmatrix} \underbrace{\mapsto}_{A_1} \begin{bmatrix} a_3^\top x \\ a_4^\top x \\ \max\{a_1^\top x, a_2^\top x\} \\ x \end{bmatrix} \underbrace{\mapsto}_{\varphi_{\theta_2}} \begin{bmatrix} \max\{a_3^\top x, a_4^\top x\} \\ \min\{a_3^\top x, a_4^\top x\} \\ \max\{a_1^\top x, a_2^\top x\} \\ x \end{bmatrix}$$

$$\underbrace{\mapsto}_{A_2} \begin{bmatrix} \max\{a_1^\top x, a_2^\top x\} \\ \max\{a_3^\top x, a_4^\top x\} \\ 0 \\ x \end{bmatrix} \underbrace{\mapsto}_{\varphi_{\theta_3}} \begin{bmatrix} \max\{a_1^\top x, a_2^\top x, a_3^\top x, a_4^\top x\} \\ \min\{\max\{a_1^\top x, a_2^\top x\}, \max\{a_3^\top x, a_4^\top x\}\} \\ 0 \\ x \end{bmatrix} \underbrace{\mapsto}_{e_2^\top} f(x).$$

The vector valued extension

$$f : \mathbb{R}^d \to \mathbb{R}^c$$

1-Lipschitz in ℓ^2. Then each component $f_1, ..., f_c : \mathbb{R}^d \to \mathbb{R}$ is also 1-Lipschitz.

The vector valued extension

$$f : \mathbb{R}^d \to \mathbb{R}^c$$

1-Lipschitz in ℓ^2. Then each component $f_1, ..., f_c : \mathbb{R}^d \to \mathbb{R}$ is also 1-Lipschitz.

$\implies$ All the networks we saw before are dense also in the vector valued set.

The vector valued extension

$$f : \mathbb{R}^d \to \mathbb{R}^c$$

1-Lipschitz in ℓ^2. Then each component $f_1, ..., f_c : \mathbb{R}^d \to \mathbb{R}$ is also 1-Lipschitz.

$\implies$ All the networks we saw before are dense also in the vector valued set.

Question

Does this density persist for networks which are constrained[a] to be 1-Lipschitz?

[a] In a practically implementable way

The vector valued extension

$$f : \mathbb{R}^d \to \mathbb{R}^c$$

1-Lipschitz in ℓ^2. Then each component $f_1, ..., f_c : \mathbb{R}^d \to \mathbb{R}$ is also 1-Lipschitz.

$\implies$ All the networks we saw before are dense also in the vector valued set.

Question

Does this density persist for networks which are constrained[a] to be 1-Lipschitz?

[a] In a practically implementable way

Answer

I don't know, but I believe so...
(ongoing with Takashi Furuya, and Carola-Bibiane Schönlieb)

Extensions and related projects

Current extension:

- **Extension to scalar Transformers/In-context maps**
  Takashi Furuya, Davide Murari, and Carola-Bibiane Schönlieb. "Approximation Theory for Lipschitz Continuous Transformers." ICML 2026.

Ongoing related projects:

- **1-Lipschitz networks on Hadamard manifolds**
  (with Elena Celledoni, Marta Ghirardelli, Brynjulf Owren, and Carola-Bibiane Schönlieb)

- **Trainability of 1-Lipschitz networks (scaling-laws for constrained models)**
  (with Kevin Huang, Teresa Klatzer, Carola-Bibiane Schönlieb, and Zak Shumaylov)

THANK YOU FOR THE ATTENTION!

Link to the paper: davidemurari.com/onelip