Approximation theory for 1-Lipschitz ResNets

Davide Murari

Department of Applied Mathematics and Theoretical Physics
University of Cambridge

In collaboration with Takashi Furuya and Carola-Bibiane Schönlieb
dm2011@cam.ac.uk

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

Adversarial robustness

Constraining the Lipschitz constant leads to a reduced sensitivity to input perturbations.

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

If $\|F(y) - F(x)\|_2 < \|y - x\|_2$ for every $x, y \in \mathbb{R}^d$, then $x_{k+1} = F(x_k)$ admits a unique and attractive fixed point. If $T_\alpha(x) = (1-\alpha)x + \alpha F(x)$, $\alpha \in (0,1)$ and F 1-Lipschitz, then whenever $x_{k+1} = T_\alpha(x_k)$ has a fixed point, the sequence converges.

Negative gradient flows

Let $V : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable convex function. We consider vector fields of the form

$$\mathcal{F}(x) = -\nabla V(x).$$

Given two solution curves, $\dot{x}(t) = \mathcal{F}(x(t))$ and $\dot{y}(t) = \mathcal{F}(y(t))$, we see that

$$\frac{d}{dt}\|x(t) - y(t)\|_2^2 = -\left(\nabla V(x(t)) - \nabla V(y(t))\right)^\top (x(t) - y(t)) \leq 0.$$

Thus, the flow map $\phi^t_{\mathcal{F}} : \mathbb{R}^d \to \mathbb{R}^d$ defined by $\phi^t_{\mathcal{F}}(x(0)) = x(t)$ is 1-Lipschitz.

Non-expansive gradient flows

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

Consider the scalar function[a] $V_\theta(x) = 1^\top \mathrm{ReLU}^2(Wx + b)/2$. Define

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

If $\dot{x} = \mathcal{F}_\theta(x)$ and $\dot{y} = \mathcal{F}_\theta(y)$, we have $\|y(t) - x(t)\|_2 \leq \|y(0) - x(0)\|_2$ for every $t \geq 0$.

[a] $W \in \mathbb{R}^{h\times d}$, $b \in \mathbb{R}^h$, $h \in \mathbb{N}$, $\theta = (W, b)$, and $1 \in \mathbb{R}^h$ a vector of ones.

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.,

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

ResNets based on gradient flows

We study the approximation properties of scalar ResNets of the form

$$\mathcal{N}_\theta = \pi \circ \varphi_{\theta_L} \circ \ldots \circ \varphi_{\theta_1} \circ Q : \mathbb{R}^d \to \mathbb{R},\ \varphi_{\theta_\ell} \in \mathcal{E}_h,$$

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

where $Q : \mathbb{R}^d \to \mathbb{R}^h$ and $\pi : \mathbb{R}^h \to \mathbb{R}$ are affine maps.

Q

$$\dot{x}(t) = -A(t)^\top \mathrm{ReLU}(A(t)x(t) + b(t))$$

π

0.87

First approximation theorem: Unbounded width and depth

Let $d \in \mathbb{N}$, $\mathcal{X} \subseteq \mathbb{R}^d$, and fix $c = 1$, i.e., consider scalar-valued networks. The networks we just derived define the following set

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

We denote with $\mathcal{C}_1(\mathcal{X}, \mathbb{R})$ the set of 1-Lipschitz functions from $\mathcal{X}$ to $\mathbb{R}$.

Universal approximation theorem

Let $\varepsilon > 0$, $\mathcal{X} \subset \mathbb{R}^d$ compact, and $g \in \mathcal{C}_1(\mathcal{X}, \mathbb{R})$ a 1-Lipschitz function. Then, there exists $f \in \mathcal{G}_d(\mathcal{X}) \cap \mathcal{C}_1(\mathcal{X}, \mathbb{R})$ such that

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

Two proof techniques

We prove this theorem in two different ways:

1. First, we prove that the **Restricted Stone-Weierstrass Theorem** (see below) holds,
2. Second, we prove that each **piecewise affine 1-Lipschitz** function belongs to our set of networks, and conclude thanks to their density in the set of 1-Lipschitz functions.

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

To prove the universal approximation theorem, we first show that our networks can represent all piecewise affine 1-Lipschitz functions.

Representation theorem

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

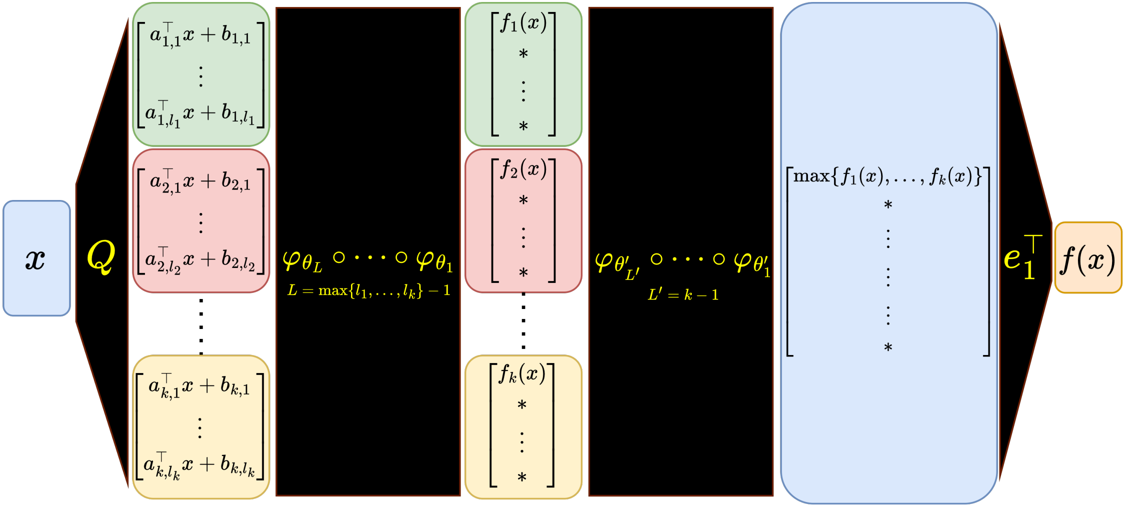
This theorem follows from the max-min representation of piecewise affine functions:

max-min representation of 1-Lipschitz piecewise affine scalar functions

Let $f:\mathbb{R}^d\to\mathbb{R}$ be a 1-Lipschitz piecewise affine scalar function. Then, there exists a choice of scalars $b_{i,j}\in\mathbb{R}$ and vectors $a_{i,j}\in\mathbb{R}^d$, $\|a_{i,j}\|_2\leq 1$, such that

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

Visualisation of the derivation in the proof



Key layers used in our proof

We then extract the maxima and minima as needed via maps of the form

$$\varphi(x) = \begin{bmatrix} \max\{x_1, x_2\} \\ \min\{x_1, x_2\} \\ x_3 \\ \vdots \\ x_h \end{bmatrix} = x - 2 \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \mathrm{ReLU}\left(\begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} & 0 & \cdots & 0 \end{bmatrix} x \right),$$

which can be written as $\varphi(x) = x - \tau W^\top \mathrm{ReLU}(Wx)$, with $\tau = 2$, and

$$\mathbb{R}^{1\times h} \ni W = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} & 0 & \cdots & 0 \end{bmatrix}.$$

Notice that $\tau = 2/\|W\|_2^2$ since $\|W\|_2 = 1$.

Second approximation theorem: Bounded width and unbounded depth

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

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

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

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

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

Theorem statement

Characterisation of the set of networks

Let $d, h \in \mathbb{N}$ with $h \geq 3$. All the functions in $\widetilde{\mathcal{G}}_{d,h}(\mathbb{R}^d, \mathbb{R})$ 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 $\widetilde{\mathcal{G}}_{d,h}(\mathcal{X}, \mathbb{R})$ satisfies the universal approximation property for $\mathcal{C}_1(\mathcal{X}, \mathbb{R})$ if $h \geq d + 3$.

THANK YOU FOR THE ATTENTION