Advancing Constrained Monotonic Neural Networks:

Achieving Universal Approximation Beyond Bounded Activations

Davide Sartor, Alberto Sinigaglia, Gian Antonio Susto

University of Padova

Monotonic Neural Networks

- Applications from regularization, to algorithmic fairness, quantile regression, density estimation and generative models

Monotonic Neural Networks

- Applications from regularization, to algorithmic fairness, quantile regression, density estimation and generative models
- Enforce positivity of jacobians via architectural constraints

$$\frac{\partial f_\theta(x)_j}{\partial x_i} \geq 0$$

Constrained Monotonic Neural Networks

- A Multi-Layer-Perceptron (MLP) is a composition of function, alternating affine transformations $l(x) = Wx + b$ and non-linearities $\sigma(x)$:

$$f(x) = l^1 \circ \sigma^1 \circ \cdots \circ l^N \circ \sigma^N(x)$$

- **Composition of monotone functions is a monotone functions**, thus monotonicity can be achieved by constraining each step to be monotonic:

Constrained Monotonic Neural Networks

- A Multi-Layer-Perceptron (MLP) is a composition of function, alternating affine transformations $l(x) = Wx + b$ and non-linearities $\sigma(x)$:

$$f(x) = l^1 \circ \sigma^1 \circ \cdots \circ l^N \circ \sigma^N(x)$$

- **Composition of monotone functions is a monotone functions**, thus monotonicity can be achieved by constraining each step to be monotonic:

$$\frac{\partial l(x)}{\partial x} \geq 0 \Rightarrow W \geq 0$$

$$\frac{\partial \sigma(x)}{\partial x} \geq 0 \Rightarrow \sigma'(x) \geq 0 \;\; \forall x$$

Universal Approximation Theorem

1. Enforcing weight-constraints guarantees monotonicity, however, the universal approximation theorem does not apply anymore

Universal Approximation Theorem

ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND CONTROL RESEARCH GROUP

1. Enforcing weight-constraints guarantees monotonicity, however, the universal approximation theorem does not apply anymore
2. Mikulincer & Reichman (2022) show that 4 layers are enough if the Heaviside-step function is used as activation, but does not hold for Rectified activations (ReLU, CELU, etc.)

Mikulincer, D. and Reichman, D.Size and depth of monotone neural networks: interpolation and approximation. Neurips 2022

Universal Approximation Theorem

1. Enforcing weight-constraints guarantees monotonicity, however, the universal approximation theorem does not apply anymore
2. Mikulincer & Reichman (2022) show that 4 layers are enough if the Heaviside-step function is used as activation, but does not hold for Rectified activations (ReLU, CELU, etc.)
3. We generalize the result to non bounded activation by proving that:

   MLPs with 4 layers, non-negative constrained weights and with monotonic activations that saturate on alternating sides are universal approximators for monotonic functions

Mikulincer, D. and Reichman, D.Size and depth of monotone neural networks: interpolation and approximation. Neurips 2022

Need for Activation Alternation

- The class of convex and non-decreasing functions is closed under composition
- Thus, **non-negative weight-constrained MLPs** with **monotonic and convex activations (*i.e.* ReLU)** are provably **not universal approximators**

$$f(x) = \ldots |W| ReLU(|W|x + b) + b \ldots$$

❌ Not a universal approximator

Non-Positive Constrained Monotonic MLPs

- The class of convex and non-decreasing/increasing functions is not closed under composition
- By a slight sign-rearrangement of the main theorem, it can be shown that **negatively weight-constrained MLPs are universal approximators**
- Thus, surprisingly, changing the weight sign results in a more expressive model

$$f(x) = \ldots |W| ReLU(|W|x + b) + b \ldots$$

❌ Not a universal approximator

$$f(x) = \ldots -|W| ReLU(-|W|x + b) + b \ldots$$

✅ A universal approximator

Activation Switch

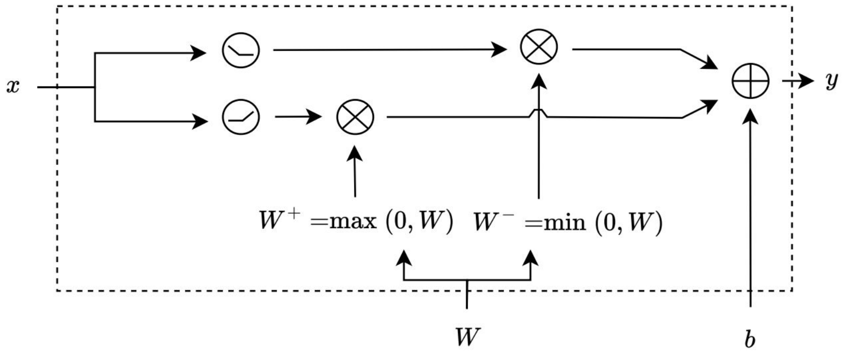
- Though non-positive-constrained MLPs are universal approximators, **their initialization is fundamental** for an effective optimization

Activation Switch

- Though non-positive-constrained MLPs are universal approximators, **their initialization is fundamental** for an effective optimization
- A novel parametrization is presented, **switching activation based on the parameters sign**:

Algorithm 1 Forward pass of a Monotonic MLP with post-activation switch

Input: data $x \in \mathbb{R}^d$, weight matrix $W \in \mathbb{R}^{d \times d'}$, bias vectors $b \in \mathbb{R}^{d'}$, monotonic activation function σ
Output: prediction $\hat{y} \in \mathbb{R}^{d'}$
$W^+ := \max(W, 0)$
$W^- := \min(W, 0)$
$z^+ := W^+ \sigma(x)$
$z^- := W^- \sigma(-x)$
$\hat{y} := z^+ + z^- + b$



Evaluation

- This novel formulation achieves **state of the art performances**.

Method	COMPAS (Test Accuracy)	Blog Feedback (Test RMSE)	Loan Defaulter (Test Accuracy)	AutoMPG (Test MSE)	Heart Disease (Test Accuracy)
XGBoost	$68.5\% \pm 0.1\%$	0.176 ± 0.005	$63.7\% \pm 0.1\%$	-	-
Certified	$68.8\% \pm 0.2\%$	0.159 ± 0.001	$65.2\% \pm 0.1\%$	-	-
Non-Neg-DNN	$69.3\% \pm 0.1\%$	0.154 ± 0.001	$65.2\% \pm 0.1\%$	10.31 ± 1.86	$89\% \pm 1\%$
DLN	$67.9\% \pm 0.3\%$	0.161 ± 0.001	$65.1\% \pm 0.2\%$	13.34 ± 2.42	$86\% \pm 2\%$
Min-Max Net	$67.8\% \pm 0.1\%$	0.163 ± 0.001	$64.9\% \pm 0.1\%$	10.14 ± 1.54	$75\% \pm 4\%$
Constrained MNN	$69.2\% \pm 0.2\%$	0.154 ± 0.001	$65.3\% \pm 0.1\%$	8.37 ± 0.08	$89\% \pm 0\%$
Scalable MNN	$69.3\% \pm 0.9\%$	0.150 ± 0.001	$65.0\% \pm 0.1\%$	7.44 ± 1.20	$88\% \pm 4\%$
Expressive MNN	$69.3\% \pm 0.1\%$	0.160 ± 0.001	$\mathbf{65.4\% \pm 0.1\%}$	7.58 ± 1.20	$90\% \pm 2\%$
Ours	$\mathbf{69.5\% \pm 0.1\%}$	$\mathbf{0.149 \pm 0.001}$	$\mathbf{65.4\% \pm 0.1\%}$	$\mathbf{7.34 \pm 0.46}$	$\mathbf{94\% \pm 1\%}$

AMCO
ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND CONTROL RESEARCH GROUP

ICML
International Conference On Machine Learning

Project Page:
https://amco-unipd.github.io/monotonic/

Contacts:
<davide.sartor.4@phd.unipd.it>
<alberto.sinigaglia@phd.unipd.it>

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DEGLI STUDI DI PADOVA

HUMAN INSPIRED TECHNOLOGY Research Centre