Conic Activation Functions

Changqing Fu and Laurent D. Cohen, CEREMADE, Paris Dauphine University PSL, PRAIRE Institute

Replace Pointwise Activations with Improved Performance and Training

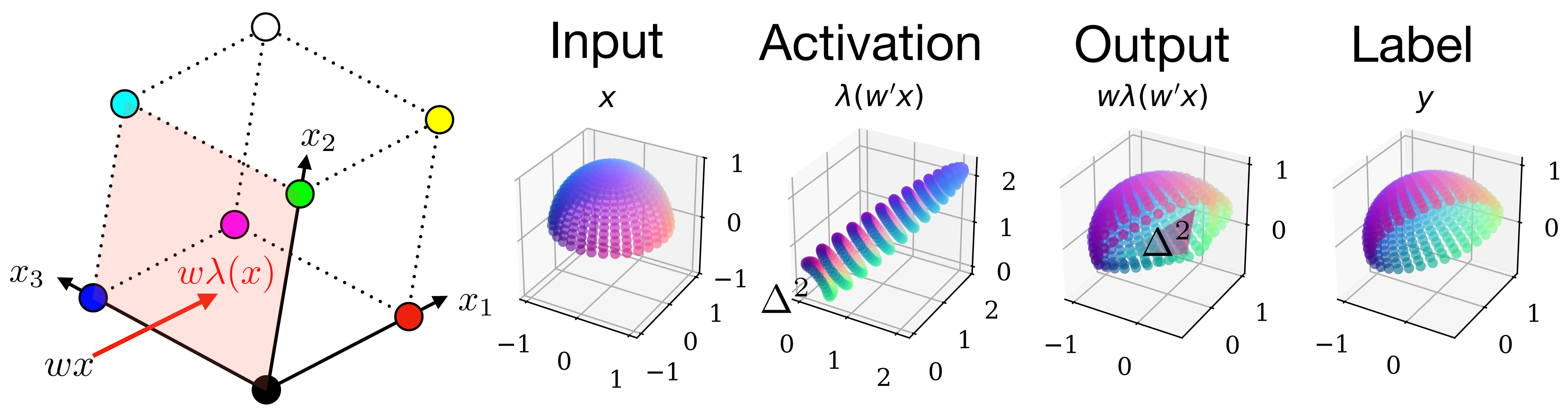
Observations

1. Component-Wise activations is only permutation equivariant
2. ReLU is a conic projection onto the positive orthant
   $\lambda=\max\{\cdot,0\}$=idempotent+component-wise+positive-scalable

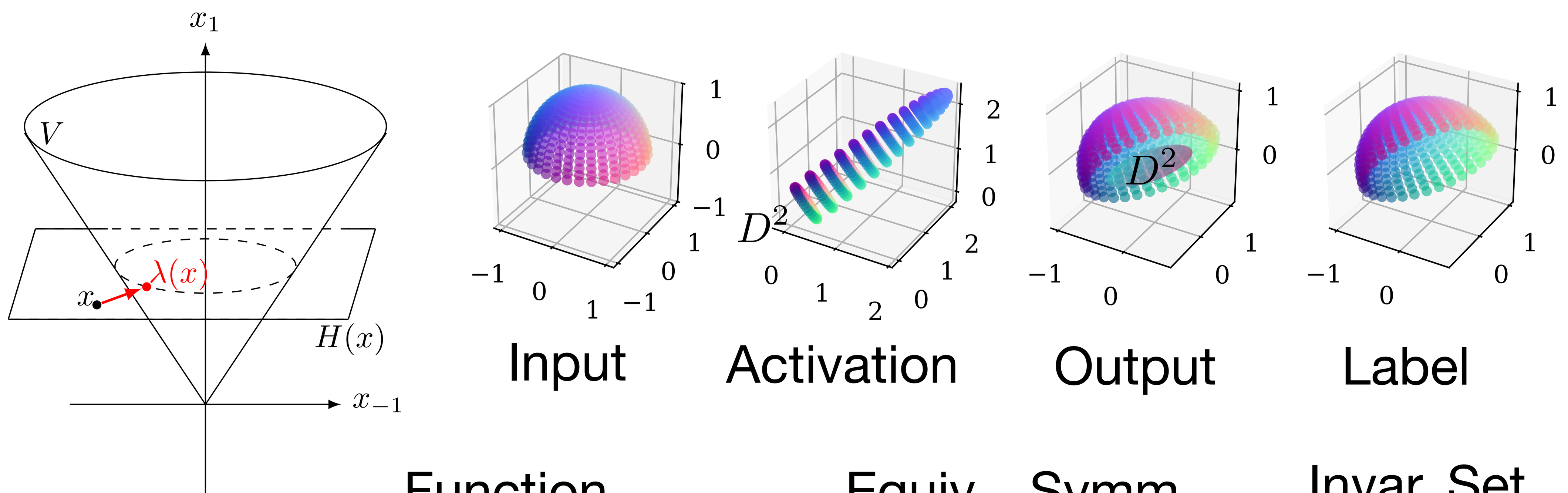
Solution

Allow orthogonal equivariance in neural networks by replacing the positive orthant with the Lorentz cone

Intuition: Learn a Rotation



ReLU: loses the rotary symmetry near the equator
CoLU: keeps the rotary symmetry everywhere



	Function	Equiv.	Symm.	Invar. Set
Self-Attn	$x\in\mathbb{R}^{CHW}\mapsto\sigma(x\otimes x/\sqrt{C})x$	Orth	Entropic	UniColor
ReLU	$x\in\mathbb{R}^{C}\mapsto\max\{x,0\}$	Perm	SimplexΔ^{C-1}	Orthant $\mathbb{R}^C_+$
CoLU	$x\in\mathbb{R}^{C}\mapsto\pi_{\widetilde{V}\cap H(x)}(x)$	Orth	DiskD^{C-1}	Cone$\widetilde{V}$

Computation

The complexity of CoLU is $O(C)$ and the introduced overhead is negligible in practice.

Conic Activation Functions

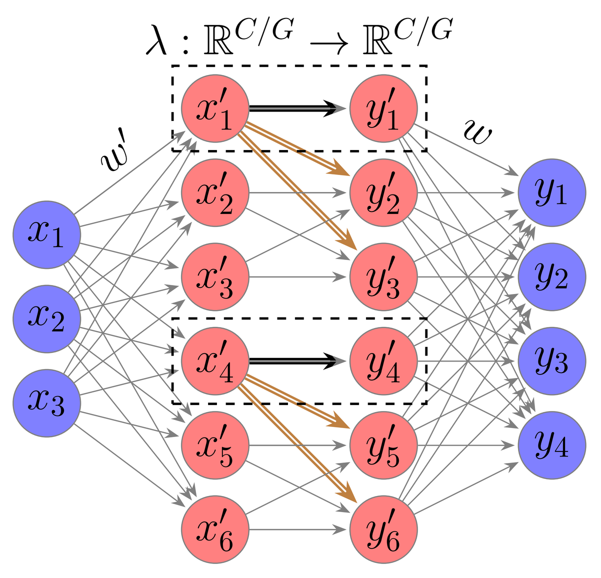
$$\lambda(x)_i=\begin{cases}x_1, & x_{-1}=(x_2,\dots,x_C) \quad i=1\\ \min\{\max\{x_1/(|x_{-1}|+\varepsilon),0\},1\}x_i, & i=2,\dots,C\end{cases}$$

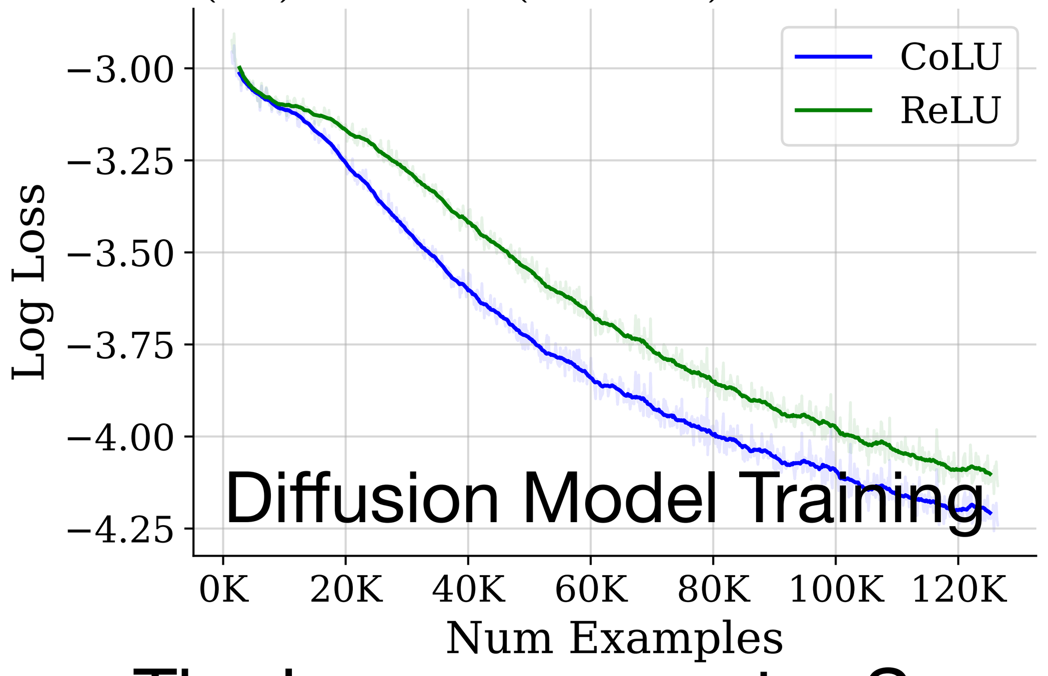
$$\lim_{\varepsilon\to 0}\lambda(x)=\pi_{\widetilde{V}\cap H(x)}(x)=\pi_{\max\{x_1,0\}D+\min\{x_1,0\}\mathbf{e}_1}(x)$$

Group-Wise Activation

$$\pi_i^G\lambda=\lambda\pi_i^G, i=1,2,\dots,G \text{ where } C=GS$$

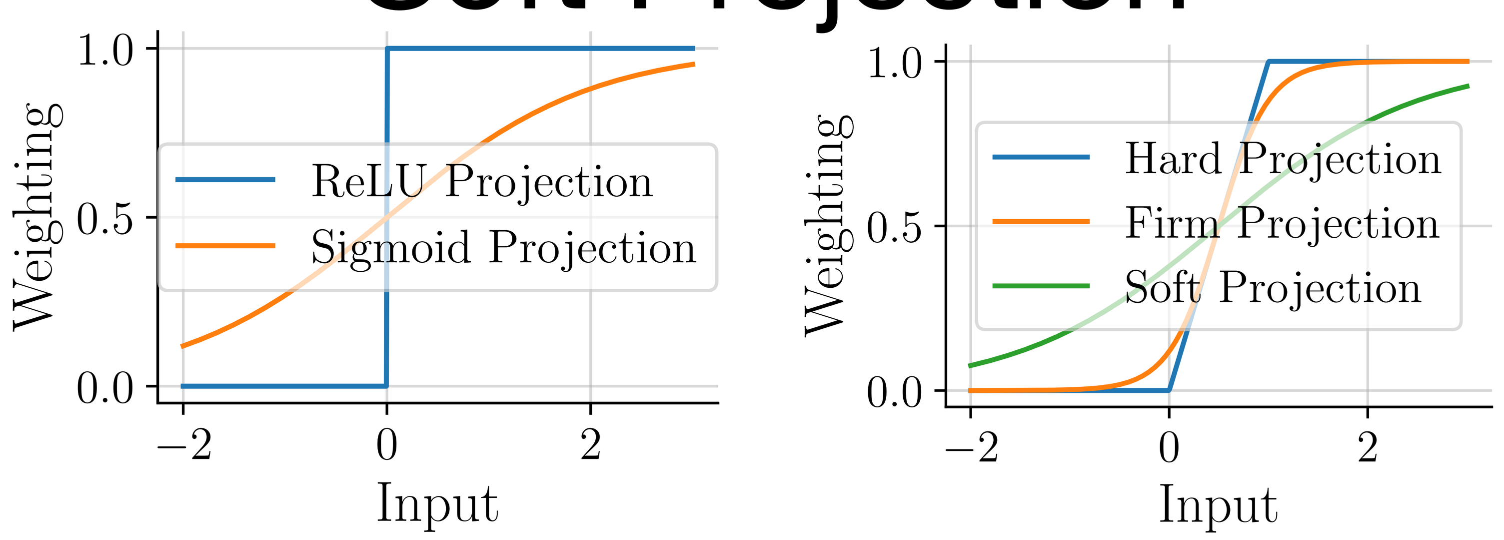
$$\mathcal{G}^*_\lambda=\mathcal{G}^*_{\mathcal{I}_\lambda}=\mathrm{Perm}(G)\times\mathcal{O}^G(S-1)$$





Labels are G Values. The hyper-parameter S (dim_cone) or G (num_cones) balances between pure-permutation and orthogonal symmetries

Soft Projection



The soft mask replaces $x\in\mathbb{R}\mapsto\min\{\max\{x,0\},1\}$

Axis Sharing

Gluing cone axes $\pi_i^G=\pi_1\times$ other $S-1$ dimensions
Saves params. Further improves 4D spacetime

Experiments (S=4)

2-Layer MLP (MNIST)

	ReLU	CoLU
Train Loss	(6.85 ± 0.0)e-5	(**0.0** ± 0.0)e-5
Test Accuracy	(93.99 ± 0.19)%	(**94.89** ± 0.25)%

ResNet56 (CIFAR10)

	ReLU	CoLU
Train Loss	0.0051 ± 0.0014	**0.0032** ± 0.0001
Test Accuracy	(90.65 ± 1.00)%	(**91.01** ± 0.39)%

2-Layer VAE (MNIST)

(Shared&Soft)	ReLU	CoLU
Train Loss	84.29 ± 0.34	**83.88 ± 2.68**
Test Loss	98.14 ± 0.07	**97.64 ± 1.39**

GPT2 MLP (Shakespeare Corpus)

	ReLU	CoLU
Train Loss	1.256	1.263
Test Loss	1.482	**1.481**

Diffusion UNet (CIFAR10)

	ReLU	CoLU
Train Loss	0.1653	**0.1458**
Early Samples		

Latent Diffusion (Fine-Tune)

Research supported by Google's TPU Research Cloud (TRC)
PRAIRIE (ANR-19-P3IA-0001) and HPC from GENCI-IDRIS (Jean-Zay)
Code: https://github.com/EvergreenTree/di-f-fu-sion