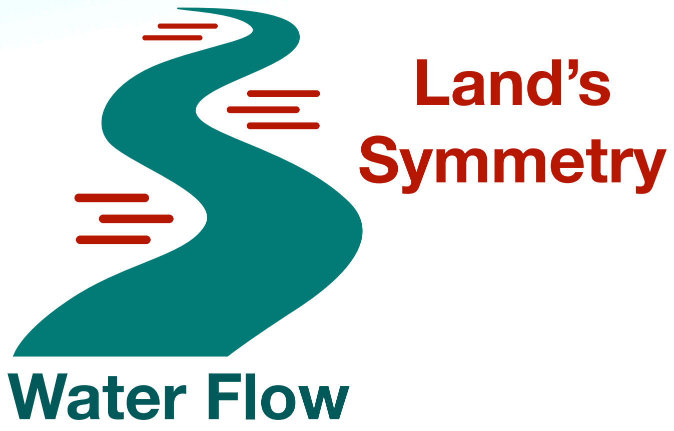
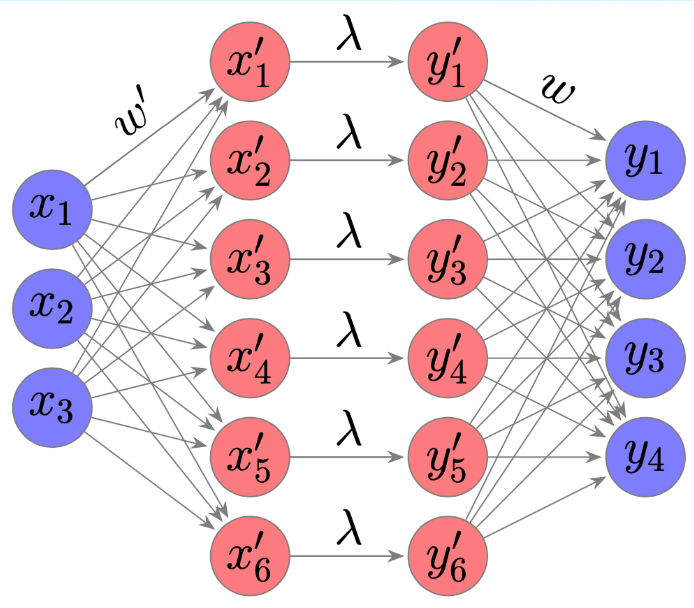
Question

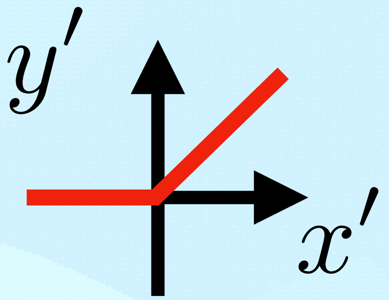
Can we improve neural networks by generalizing its symmetry constraints?



Activation Functions

ReLU: Rectified Linear Units





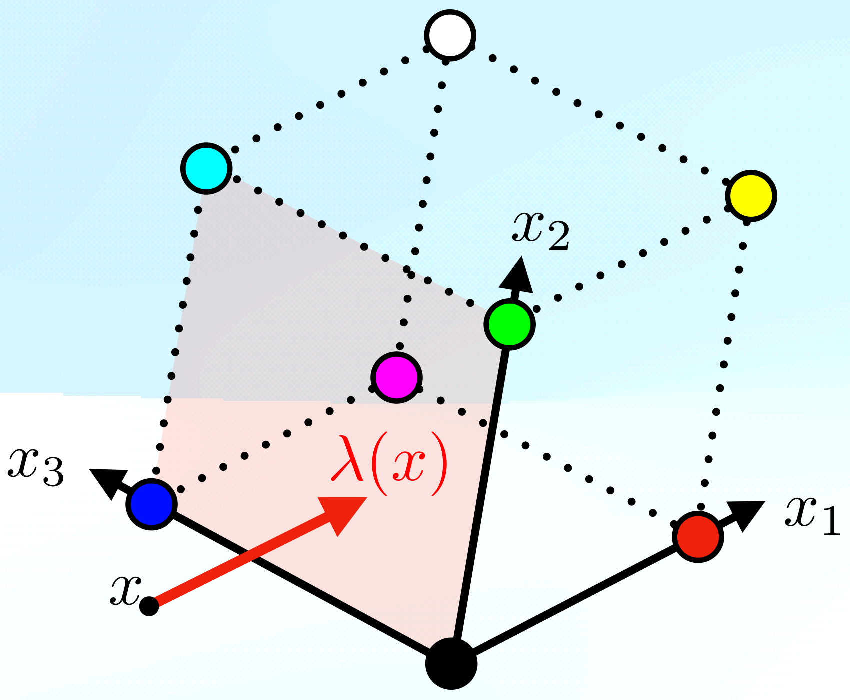
$$\begin{tikzcd}
x' \arrow[r, "\text{Activation Function}"'] & y' \arrow[d, "\text{Linear}"] \\
x \arrow[u, "\text{Linear}"] \arrow[r, dotted] & y
\end{tikzcd}$$

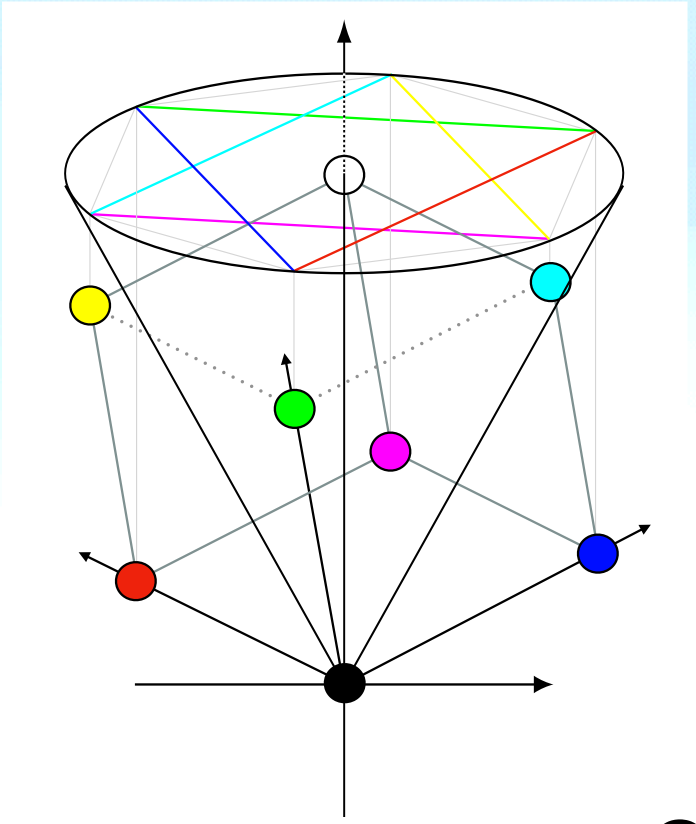
Characterization

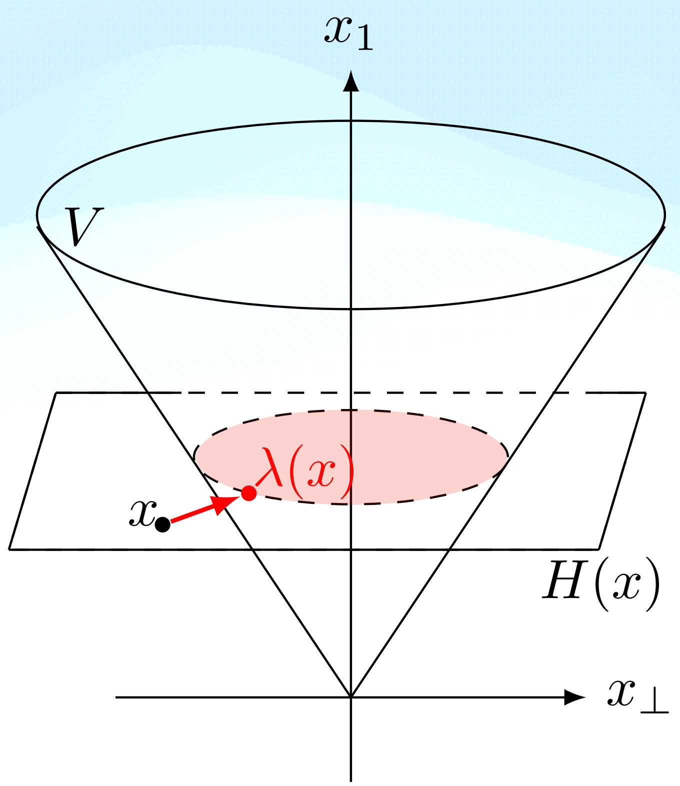
- Axis Homogeneity $\Lambda(x)_i = \Lambda(x_i)$ i.e. $\Lambda\pi_i = \pi_i\Lambda$ $\Lambda(x_1, x_2, \ldots) = (\Lambda(x_1), \Lambda(x_2), \ldots)$
- +Idempotence $\Lambda(\Lambda(x)) = \Lambda(x)$ $\Lambda\big|_O(x) = \mathrm{id}(x), O$ is Borel
- +Positive Homogeneity $\forall t > 0, \Lambda(tx) = t\Lambda(x)$ $\Lambda(x) := \lambda(x) = \max\{x, 0\}$

Solution

Allow orthogonal equivariance with a more symmetric invariant set







ReLU: loses the rotary symmetry

CoLU: keeps the rotary symmetry

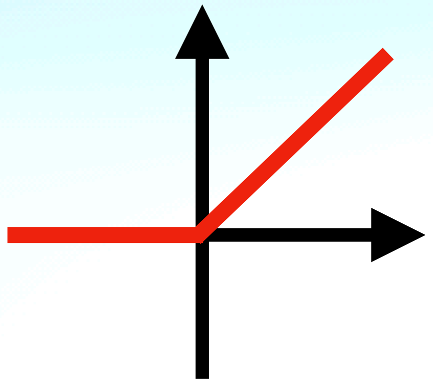
Previous works: spatial domain (Geometric Deep Learning)
Our work: feature space!

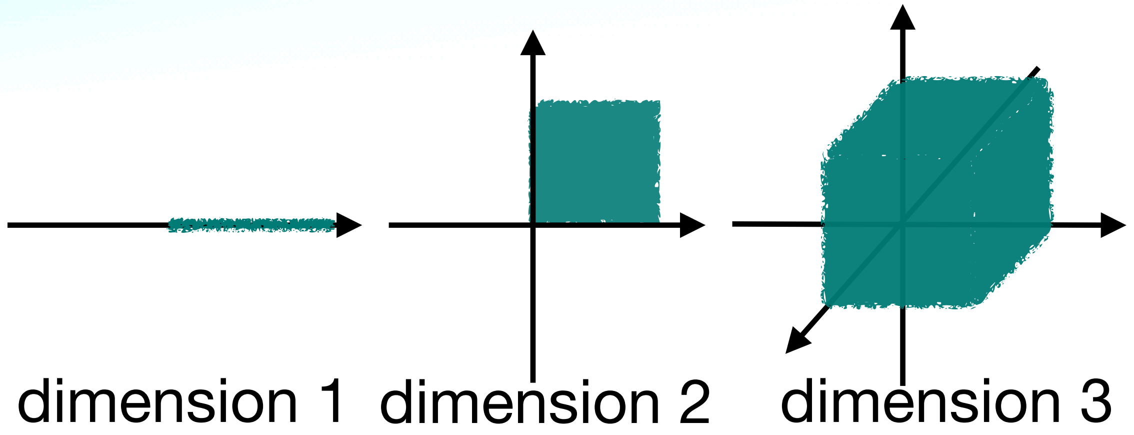
Conic Activation Functions

A symmetry constraint on generative models for improved **generalization property** and better **learning and performance**.

- **Semidefinite Program** $\Omega = \mathbb{R}_+^C$ $$\min_{y \geq 0} \frac{1}{2}\|x-y\|_2$$

Solution $\lambda(x) = \pi_\Omega(x) = x_+ = \max\{x, 0\}$





- **Conic Program?** $\Omega = \{x : x_1^2 \geq x_2^2 + x_3^2 + \ldots\}$ $$\min_{y \in \Omega} \frac{1}{2}\|x-y\|_2$$

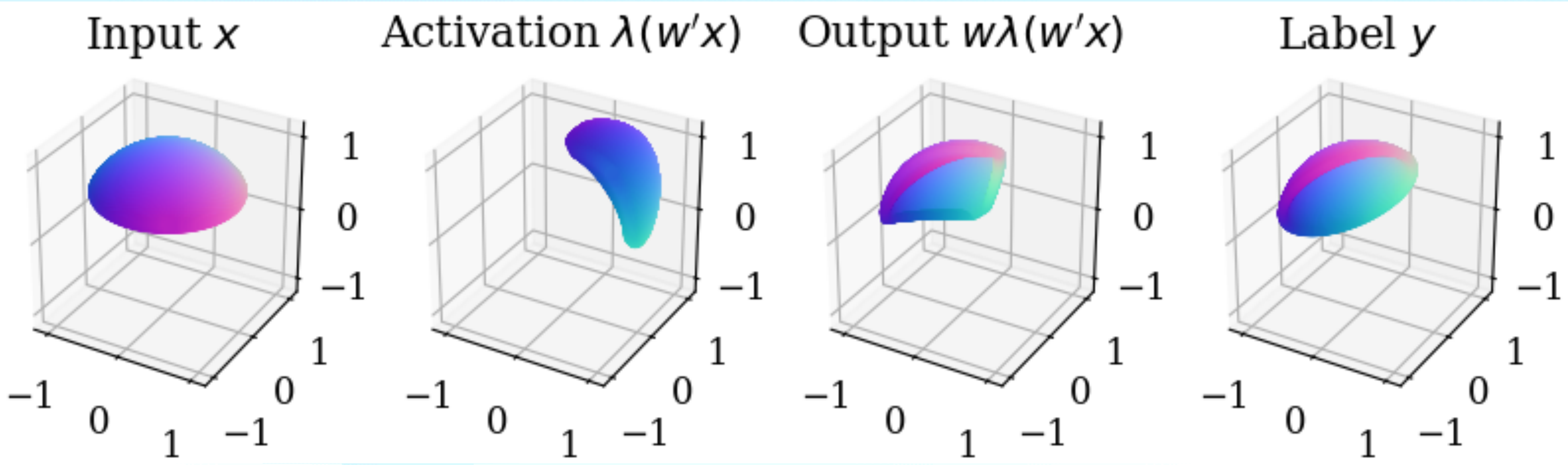
CoLU Symmetry is Compatible with Transformer

Nonlinearities	**Function**	**Group**	**Symmetry**	**Limiting**
Attention	$x \in \mathbb{R}^{C \times N} \mapsto Z^{-1} \exp(\frac{\langle x, x \rangle_C}{\sqrt{C}})x$	Orth	Entropic	ColorClusters
ReLU	$x \in \mathbb{R}^{C} \mapsto \max\{x, 0\}$	Perm	Simplex Δ^{C-1}	Orthant $\mathbb{R}^{C}_{+}$
CoLU	$x \in \mathbb{R}^{C} \mapsto \pi_{\widetilde{V} \cap H(x)}(x)$	**Orth**	Disk D^{C-1}	Cone $\widetilde{V}$

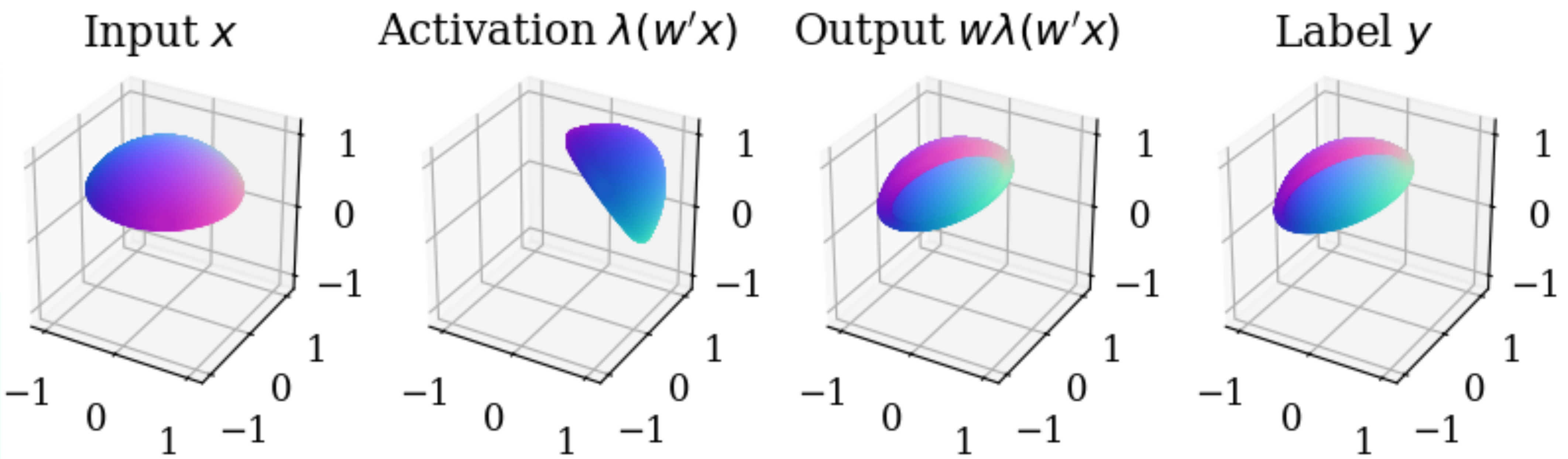
Conic Activation Functions

A symmetry constraint on generative models for improved **generalization property** and better **learning and performance**.

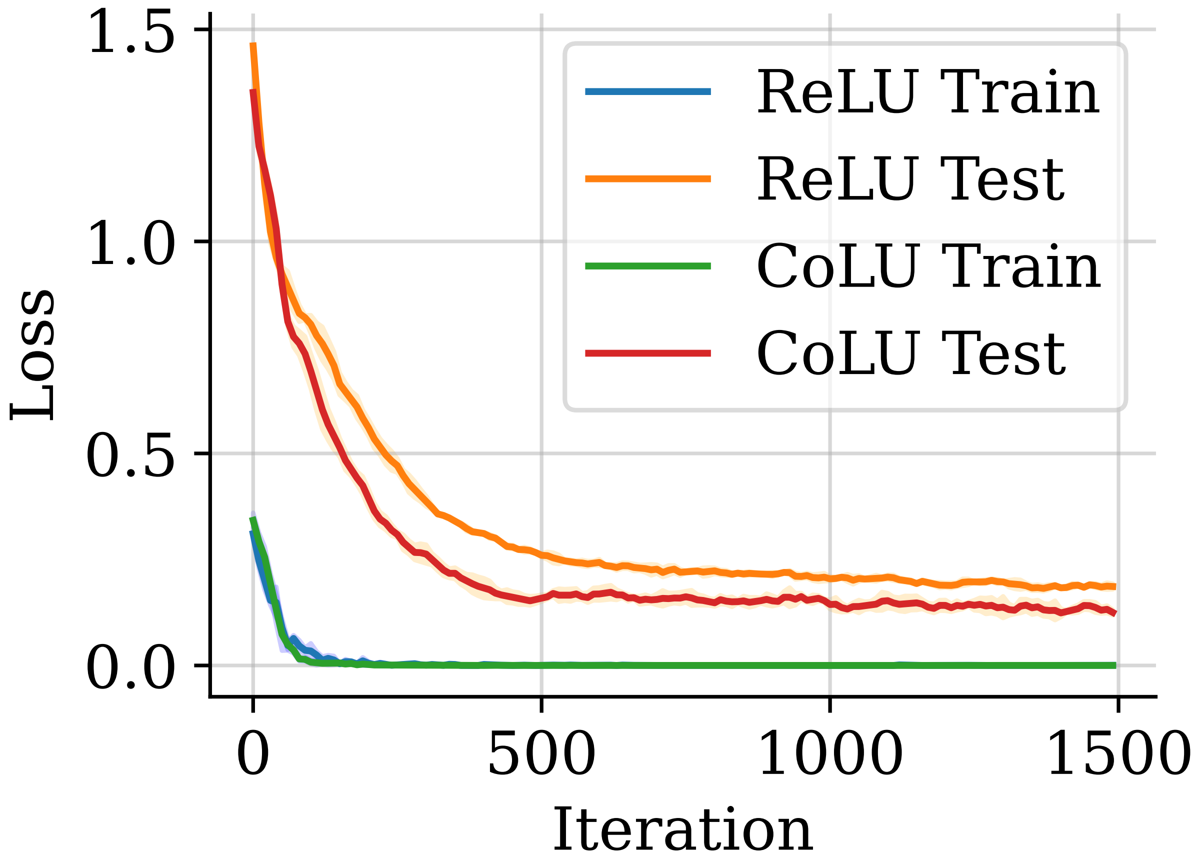
ReLU: permutation symmetry



CoLU: orthogonal/rotary symmetry



- Minimal Example
- Improved Generalization

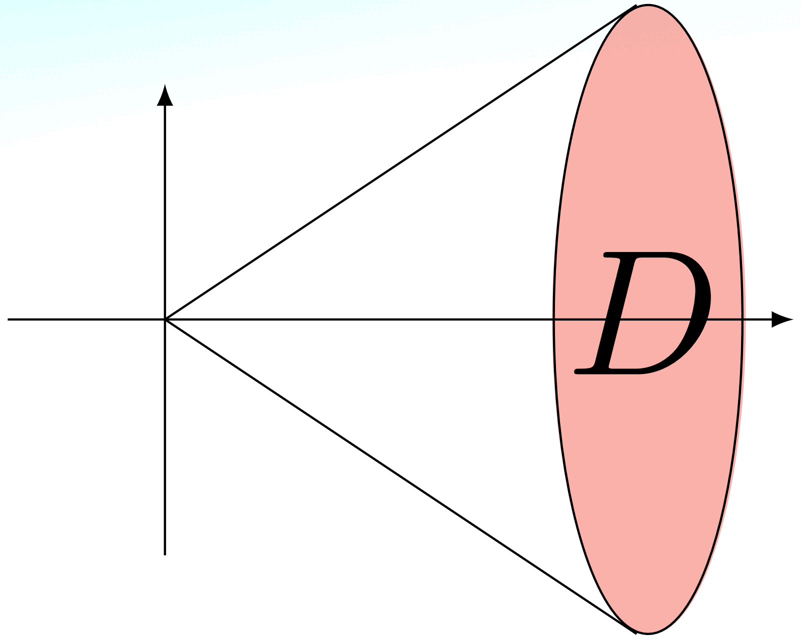


Conic Activation Functions

A symmetry constraint on generative models for improved **generalization property** and better **learning and performance**.

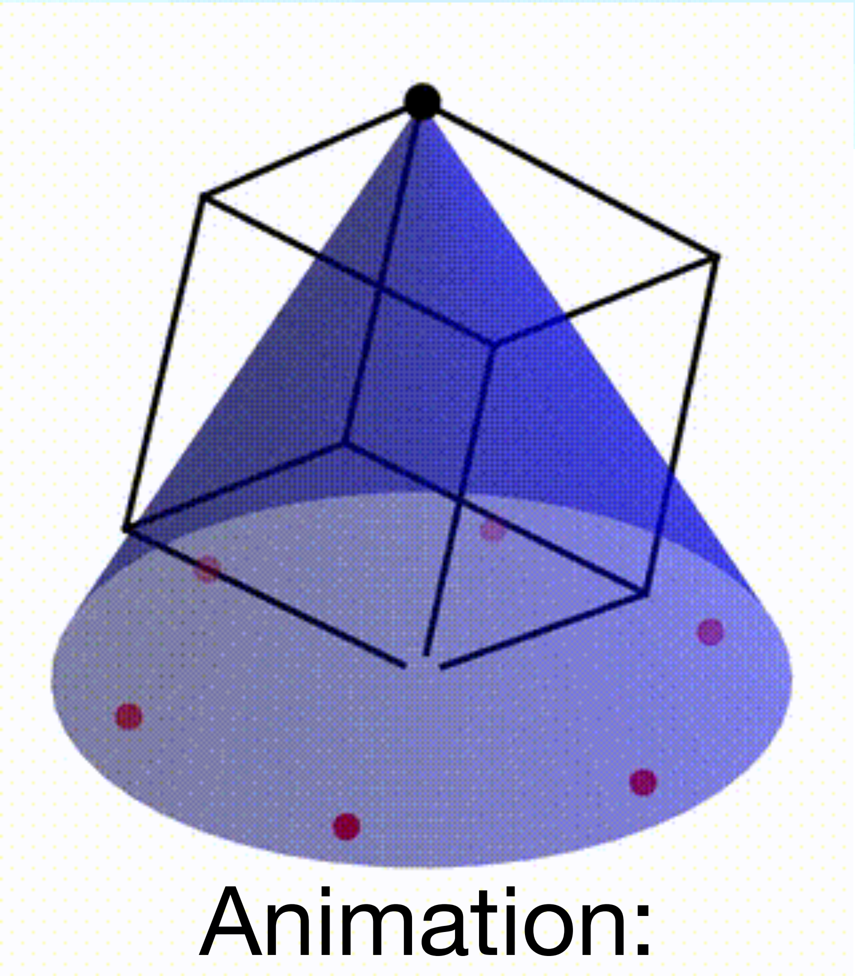
Closed Form

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



Projective Form

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



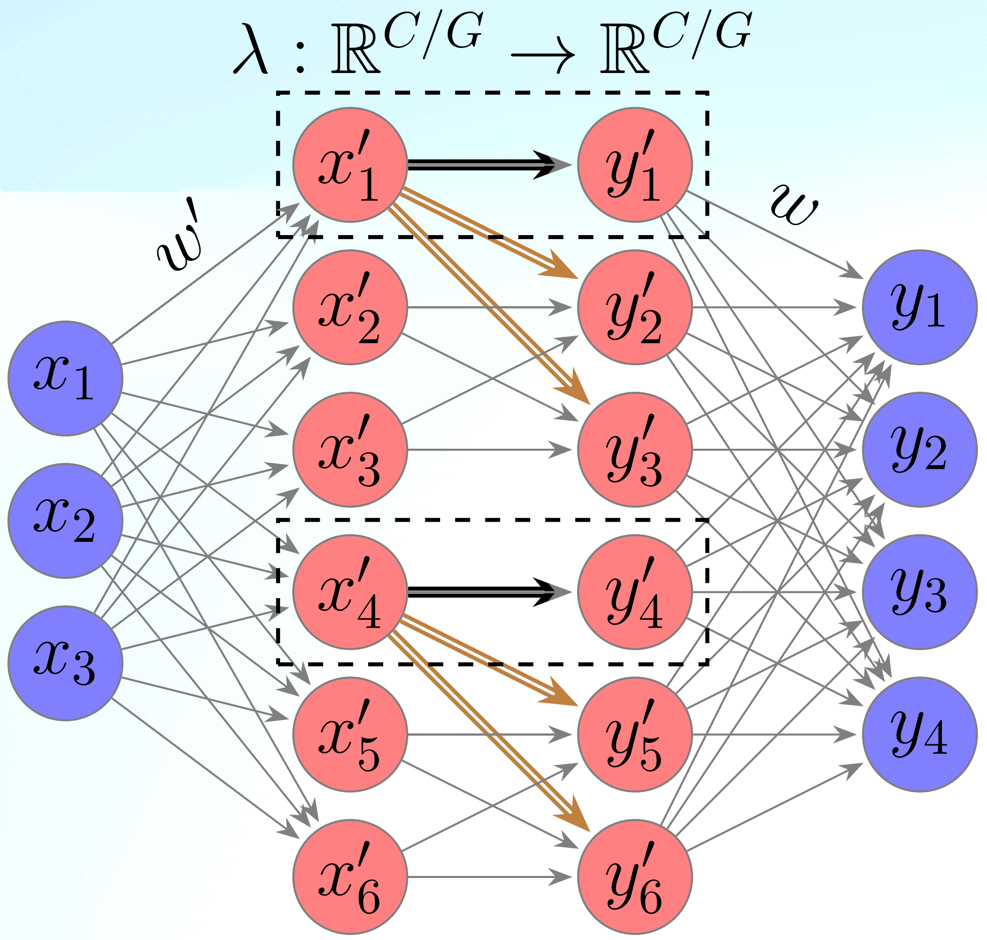
Animation: Conic Symmetry

Conic Activation Functions

A symmetry constraint on generative models for improved **generalization property** and better **learning and performance**.

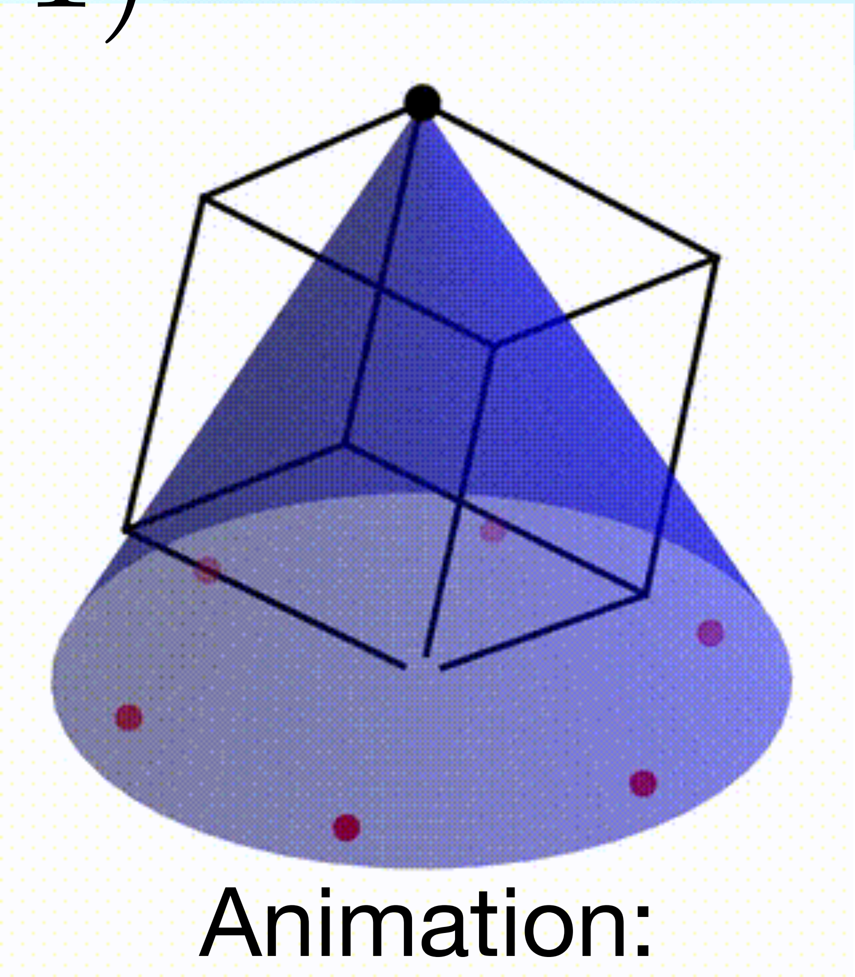
Multi-Head Structure $\pi_i^G \lambda = \lambda \pi_i^G, i = 1, 2, \ldots, G$ where $C = GS$

Symmetry Group: $\mathrm{Perm}(G) \times \mathrm{Orth}^G(S-1)$



$C = 6, S = 3, G = 2$

6 Neurons
3D Cone
2 Cones

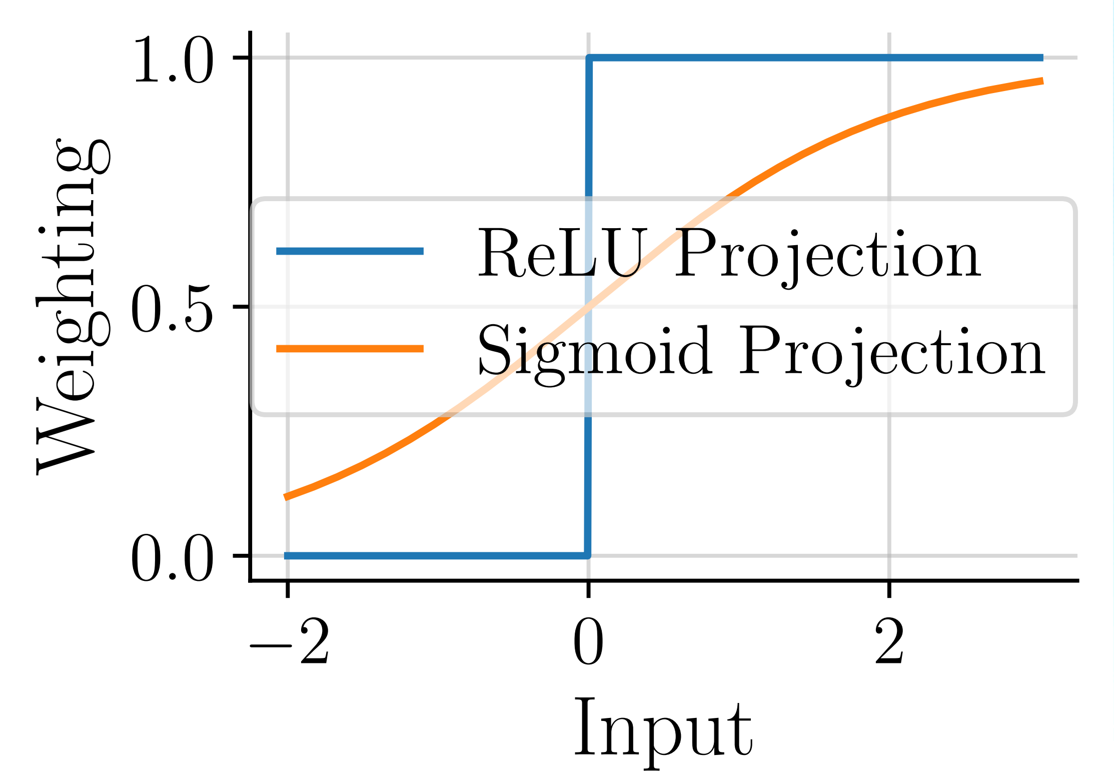


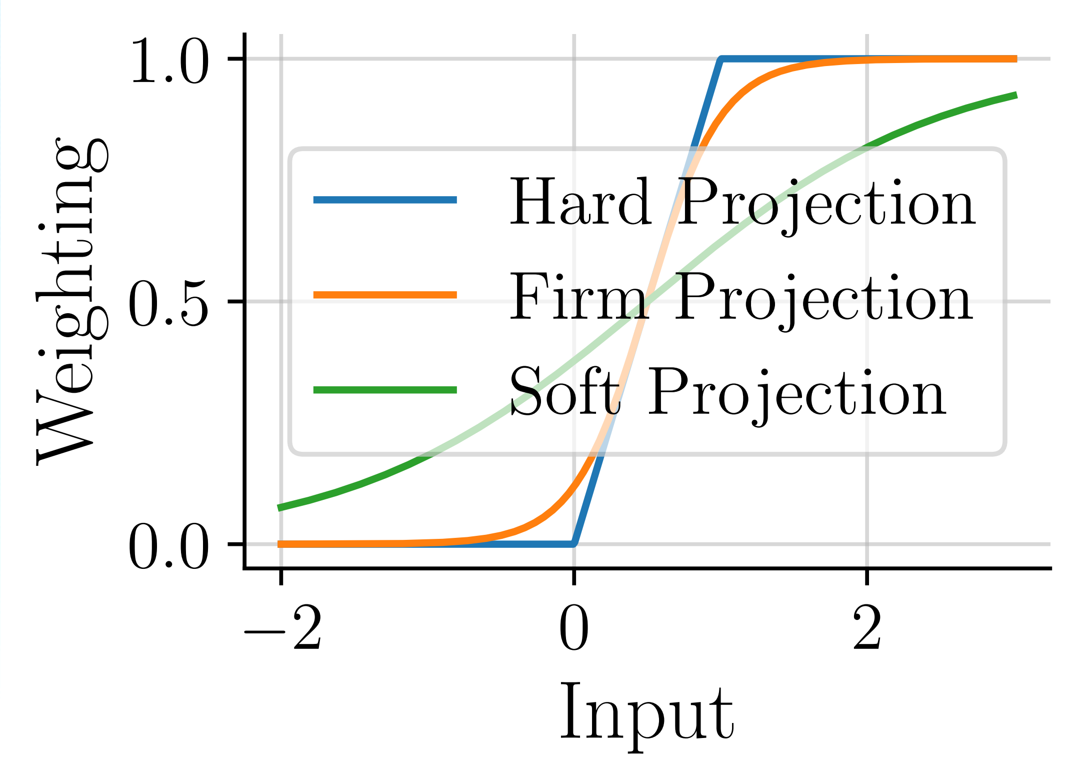
Animation:
Conic Symmetry

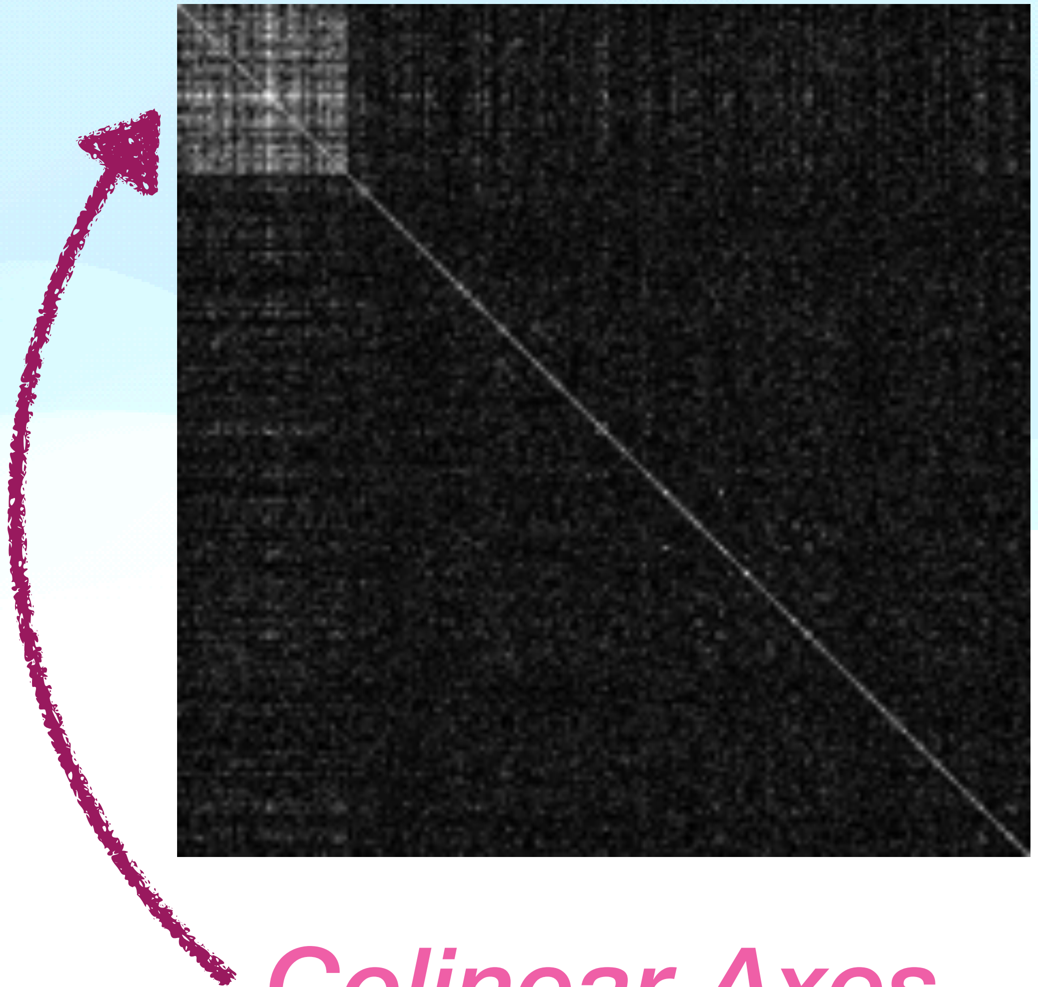
Conic Activation Functions

A symmetry constraint on generative models for improved **generalization property** and better **learning and performance**.

Soft Projection

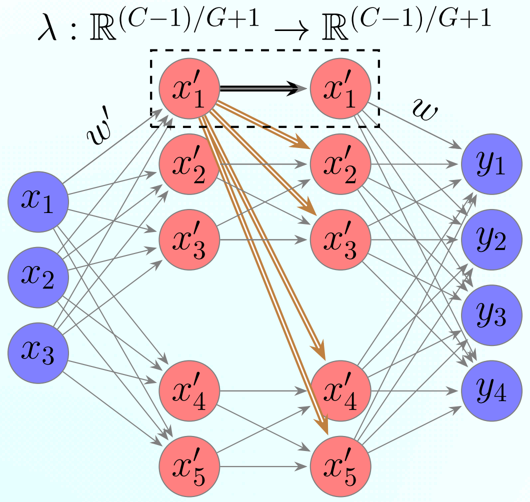






Colinear Axes

Axis Sharing



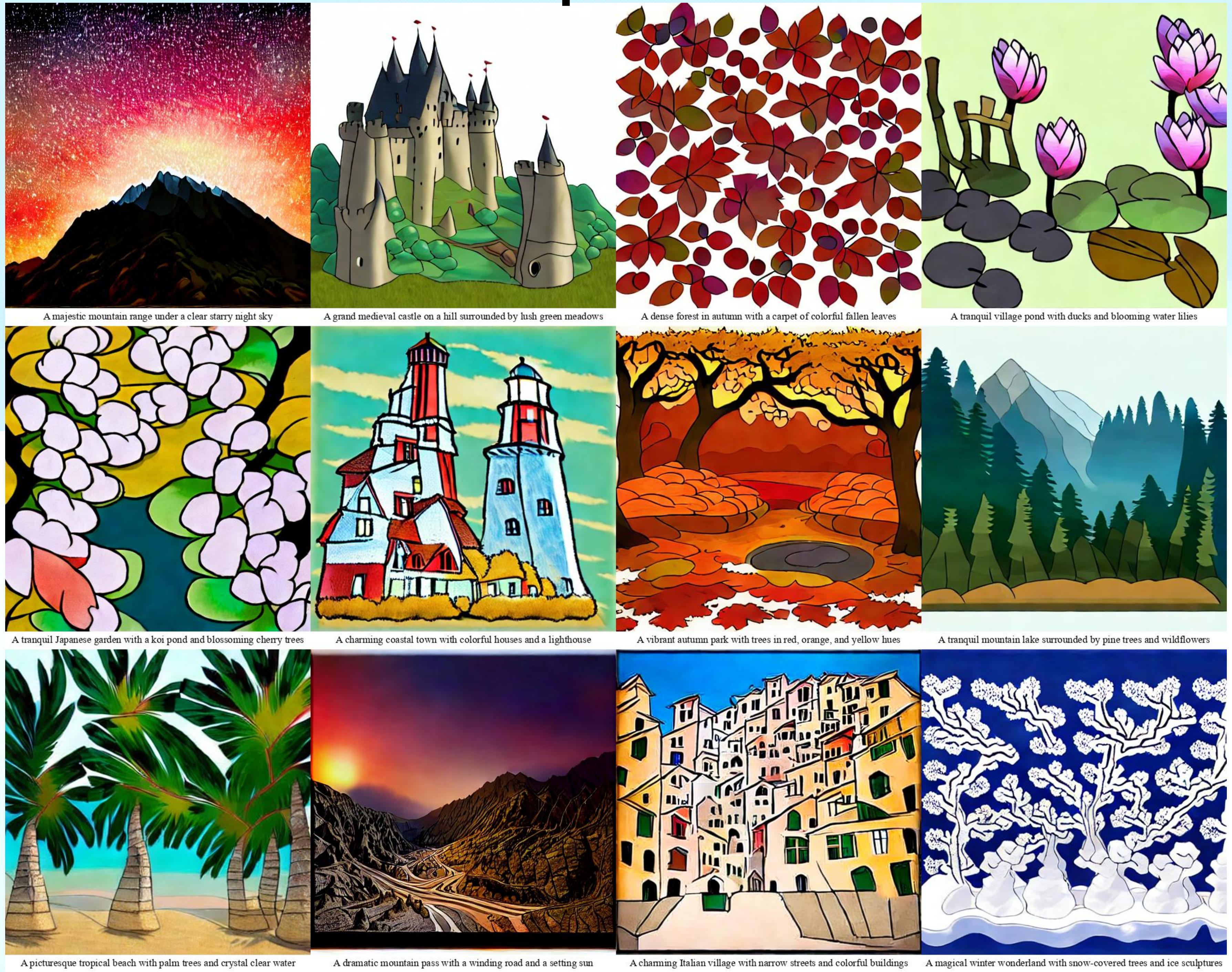
5 Neurons
3D Cone
2 Cones

Glue cone axes w/ $\pi_i^G = \pi_1 \times \pi_{\text{another}(S-1)\text{axes}}$

Conic Activation Functions

A symmetry constraint on generative models for improved **generalization property** and better **learning and performance**.

Generated Samples of CoLU-LDM



A majestic mountain range under a clear starry night sky

A grand medieval castle on a hill surrounded by lush green meadows

A dense forest in autumn with a carpet of colorful fallen leaves

A tranquil village pond with ducks and blooming water lilies

A tranquil Japanese garden with a koi pond and blossoming cherry trees

A charming coastal town with colorful houses and a lighthouse

A vibrant autumn park with trees in red, orange, and yellow hues

A tranquil mountain lake surrounded by pine trees and wildflowers

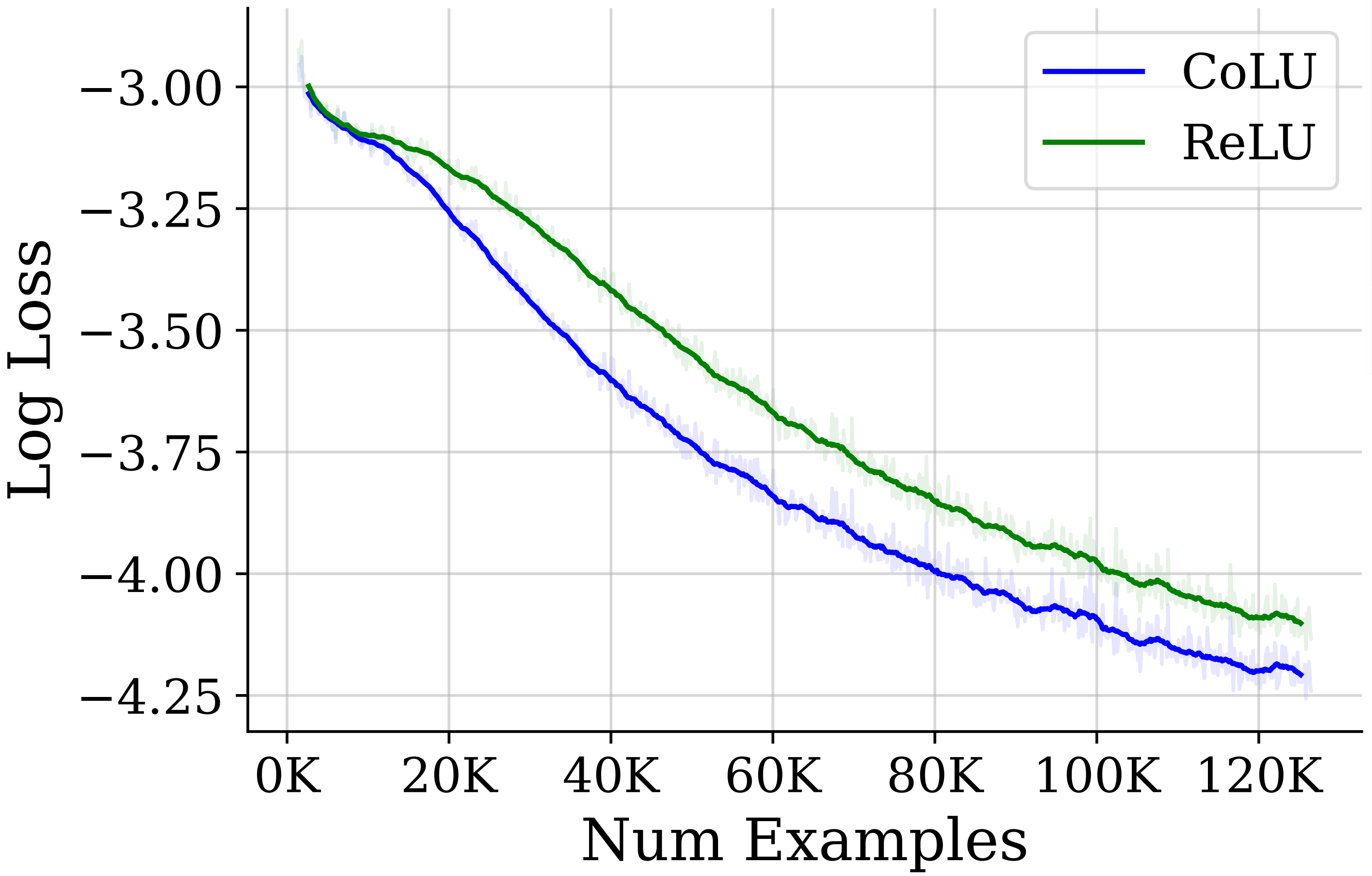
A picturesque tropical beach with palm trees and crystal clear water

A dramatic mountain pass with a winding road and a setting sun

A charming Italian village with narrow streets and colorful buildings

A magical winter wonderland with snow-covered trees and ice sculptures

Faster Learning



UNet with Attention (835M parameters)

Diffusion Model and Process Matching

Diffuse

Known $q \longrightarrow$

$$\underset{\text{Signal}}{} \quad \underset{\text{Noise}}{} \quad q(x(t)|x(0)) = \mathcal{N}(\overset{\text{Signal}}{\alpha(t)x(0)}; \overset{\text{Noise}}{\sigma^2(t)\mathbf{I}})$$

$$x(t) = \overset{\text{Signal}}{\alpha(t)x(0)} + \overset{\text{Noise}}{\underbrace{\sigma(t)\varepsilon}_{\int_0^t s_\tau \, \mathrm{d}B_\tau}} \quad \varepsilon \sim \mathcal{N}(0,1)$$

$x(0)$

$x(T)$

$$x(s) \sim q(x(s)|\underbrace{G(t, x(t))}_{\text{Approximates } x(0)}), s < t$$

Unknown p

$\longleftarrow$ *Denoise*

Negative Log Likelihood

$$-\log p(x(0)) \leq -\mathbb{E}_{x(t)\sim q(x(t)|x(0))}[\log p(x(t))] + \mathcal{D}_{\mathrm{KL}}\left(q(x(t)|x(0))|p(x(0)|x(t))\right)$$

Relative Entropy $\mathcal{D}_{\mathrm{KL}}(q|p) := -\int_{x\in M} \log(p(x)/q(x))\, \mathrm{d}q(x)$

Conclusion: Match p towards q

Dauphine | PSL UNIVERSITÉ PARIS CEREMADE UMR CNRS 7534 PR[AI]RIE PaRis Artificial Intelligence Research InstitutE cnrs

Diffusion Model and Process Matching

Diffuse

Known $q \longrightarrow$ $$q(x(t)|x(0)) = \mathcal{N}(\overset{\text{Signal}}{\alpha(t)x(0)}; \overset{\text{Noise}}{\sigma^2(t)\mathbf{I}})$$

$$x(t) = \overset{\text{Signal}}{\alpha(t)x(0)} + \overset{\text{Noise}}{\underbrace{\sigma(t)\varepsilon}_{\int_0^t s_\tau \,\mathrm{d}B_\tau}} \quad \varepsilon \sim \mathcal{N}(0,1)$$

$x(0)$

$x(T)$

$$x(s) \sim q(x(s)|\underbrace{G(t,x(t))}_{\text{Approximates } x(0)}), s < t$$

Unknown p

$\longleftarrow$ *Denoise*

Loss $$L(\theta) = \mathbb{E}_{\underset{\text{Time}}{t\sim\mathcal{U}(0,1)},\underset{\text{Dataset}}{x(0)\sim\pi},\underset{\text{Noisy Image}}{x_t\sim q(x(t)|x(0))}} |\underbrace{G(t,x(t);\theta)}_{\frac{x(t)-\sigma(t)\varepsilon_G}{\alpha(t)}} - \underbrace{x(0)}_{\frac{x(t)-\sigma(t)\varepsilon}{\alpha(t)}}|$$

Reparameterization

Residualization

Better Accuracy and Loss

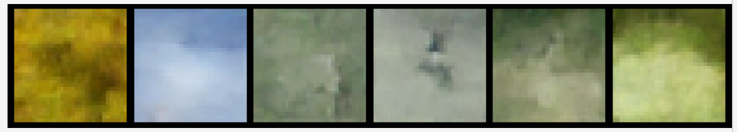
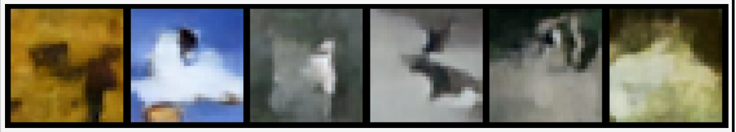
2-Layer MLP (MNIST, C=512)	ReLU	CoLU
Train Loss	0.0000 ± 0.0000	0.0000 ± 0.0000
Test Accuracy	97.17 ± 00.02	**97.23** ± 00.06
2-layer VAE (Shared&Soft)	ReLU	CoLU
Train Loss	84.29 ± 0.34	**83.88 ± 2.68**
Test Loss	98.14 ± 0.07	**97.64 ± 1.39**

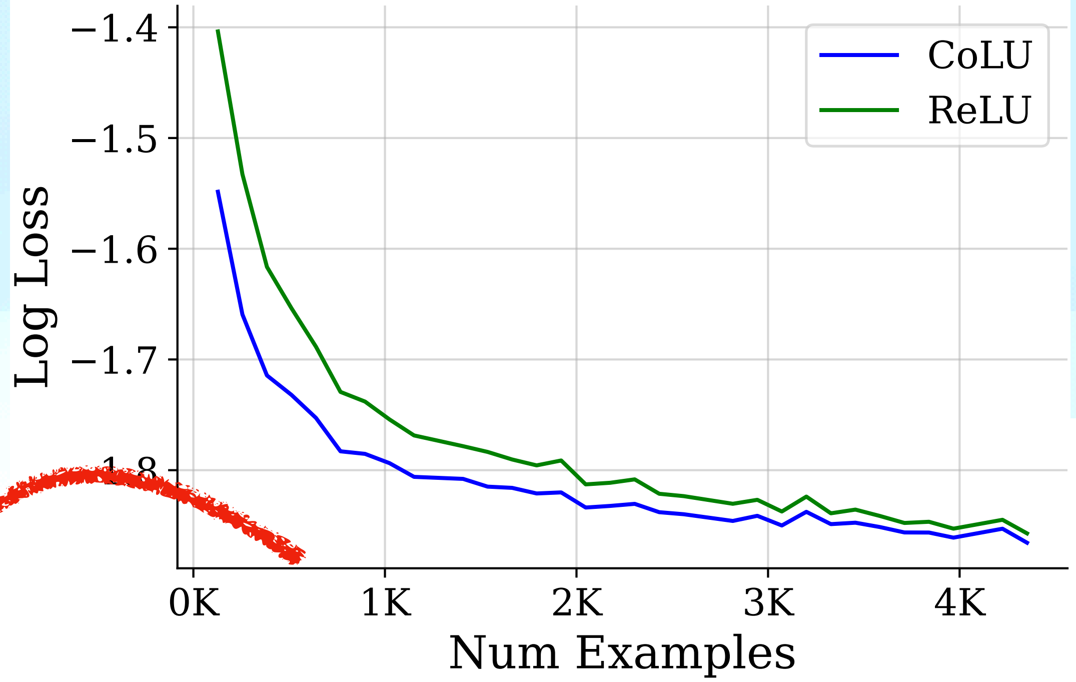
Conic Activation Functions

A symmetry constraint on generative models for improved **generalization property** and better **learning and performance**.

GPT2 MLP (FineWeb10M)	ReLU	CoLU
Forward FLOPs	39.064M	39.101M
Test Loss	3.4569 ± 0.1182	**3.3804** ± 0.1159

ResNet-56 (CIFAR10)	ReLU	CoLU
Forward FLOPs	0.252M	0.257M
Test Accuracy	**92.7282** ± 0.357	93.5851 ± 0.442

Diffusion Model (CIFAR10)	ReLU	CoLU (Faster)
Train Loss	0.1653	**0.1458**
Early Samples		



Diffusion Model Training

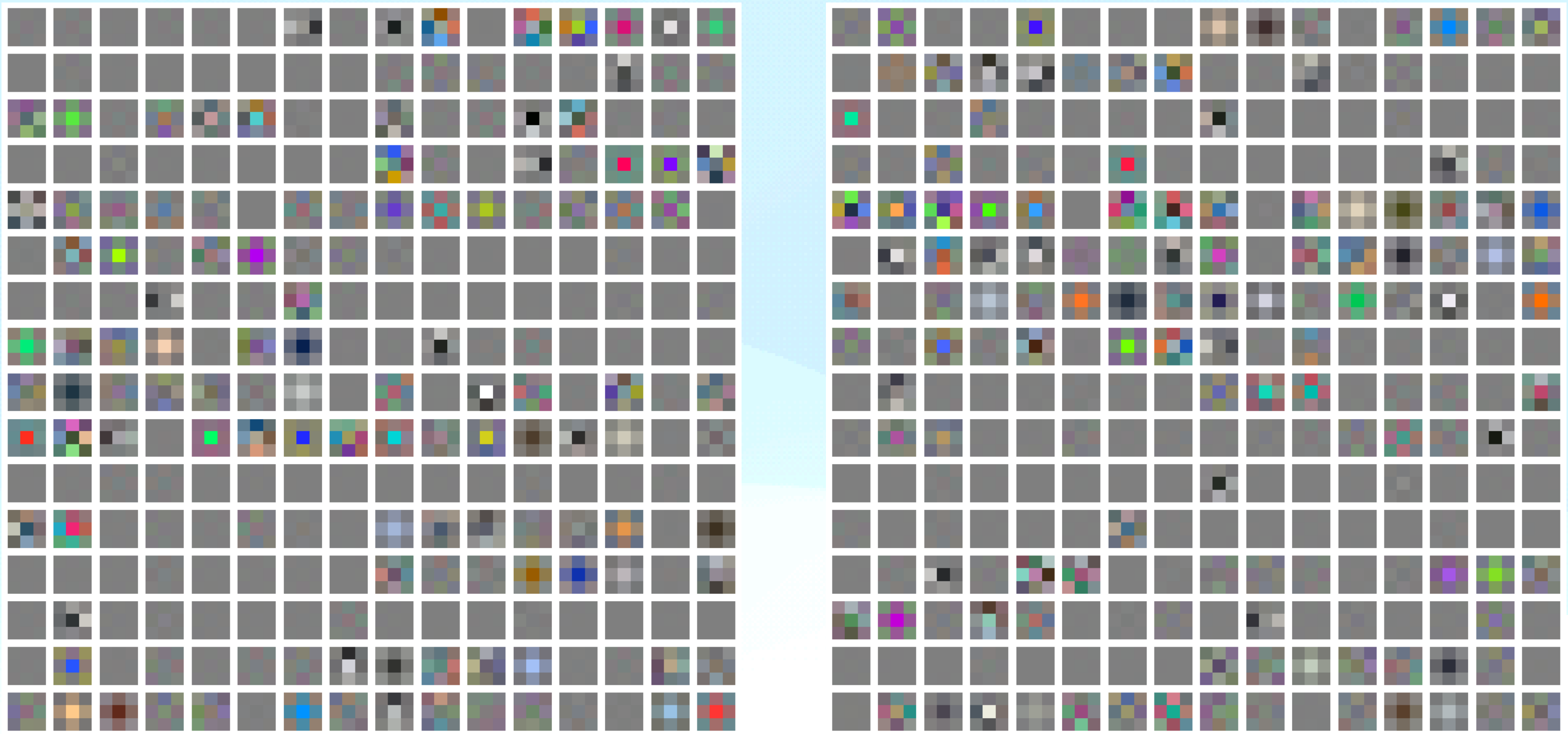
O(C) **Complexity**

Negligible **Overhead** vs ReLU

Conic Activation Functions

A symmetry constraint on generative models for improved **generalization property** and better **learning and performance**.

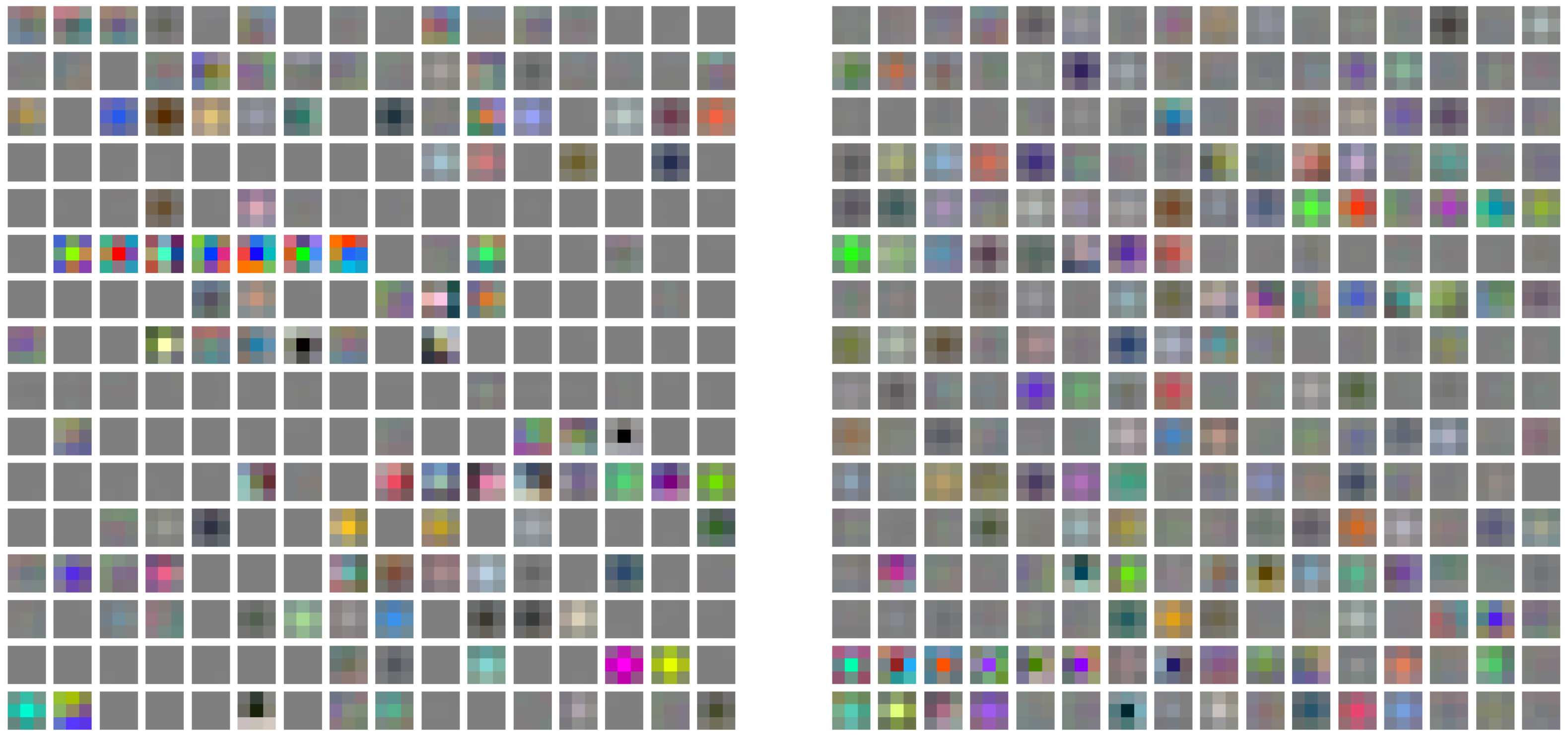
ReLU:
matches each other
with swapping



Palettes

Last Convolution Layer of Diffusion Model with Different Seeds

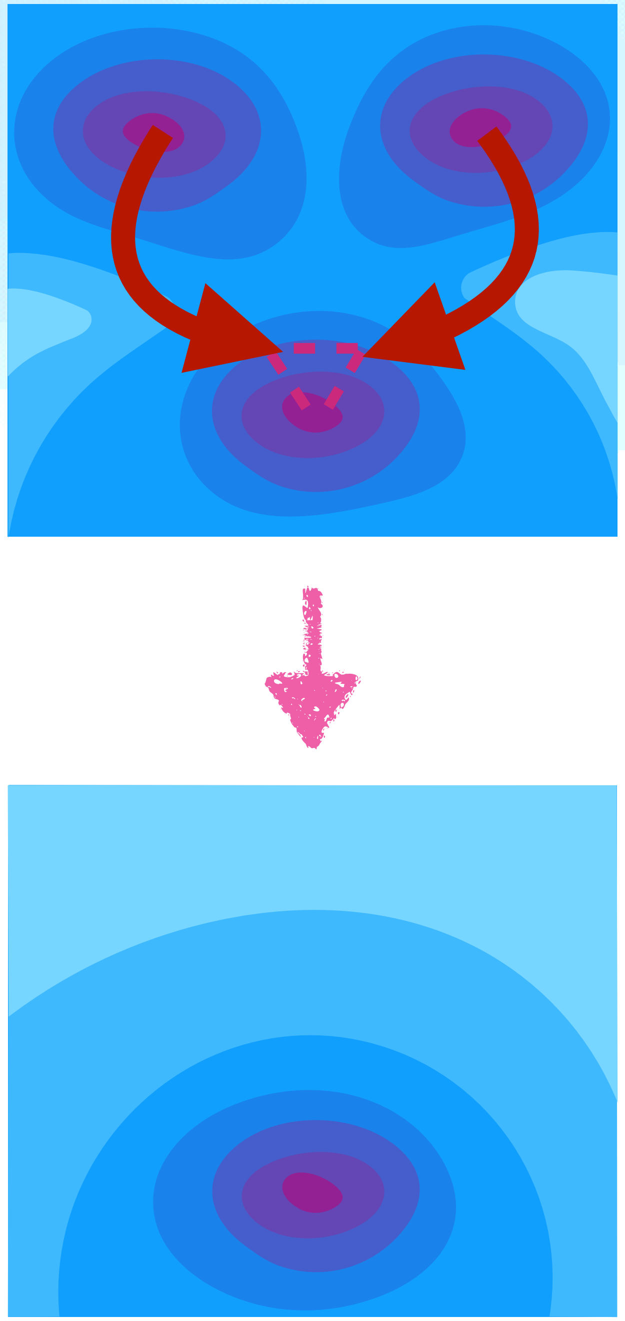
CoLU:
with swapped color
rotation

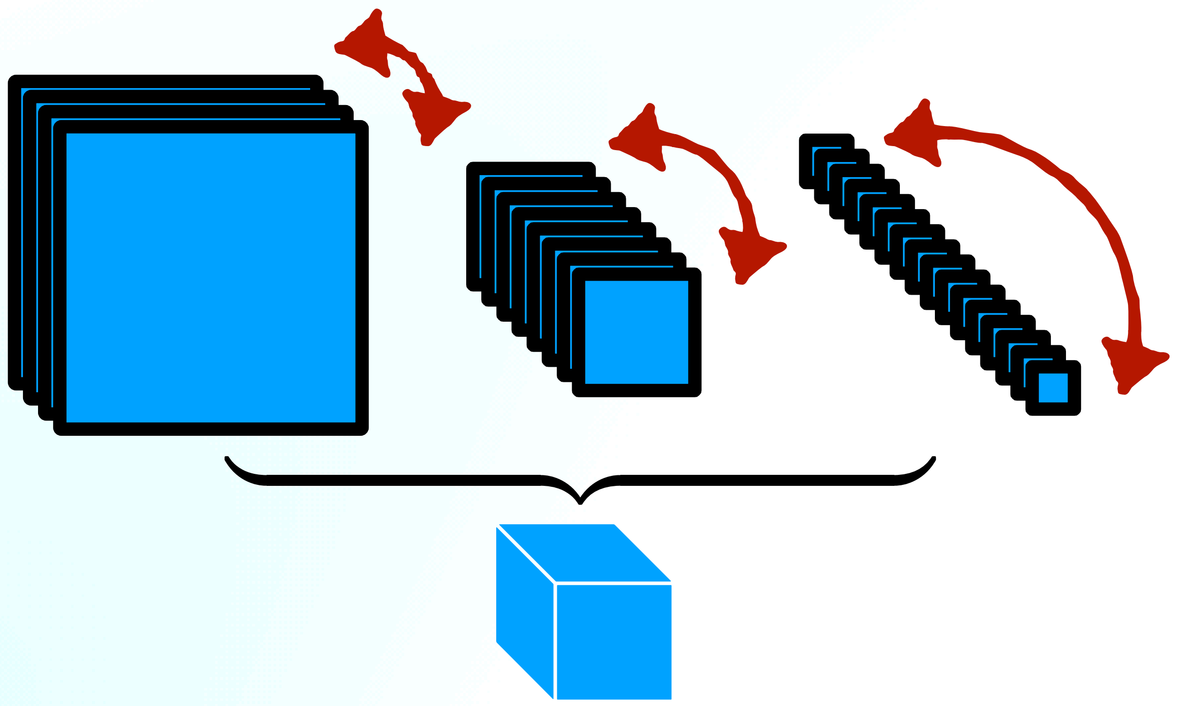


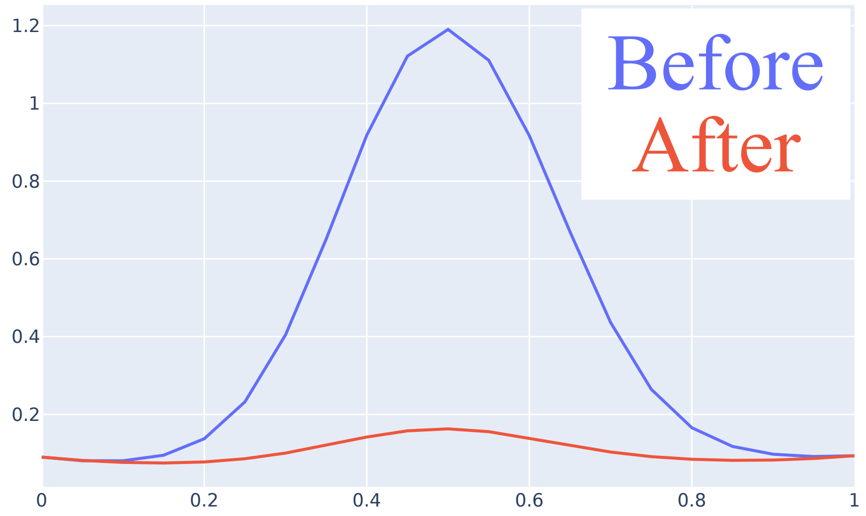
Implication: Generalize NN Symmetry

Previous works: Linear Mode Connectivity

- **Optimization**: Non-convex loss is convexified by the quotient.
- **Geometry**: Neural network symmetry induced by activation.
- **Probability**: Optimal mixture irrelevant of initializations.







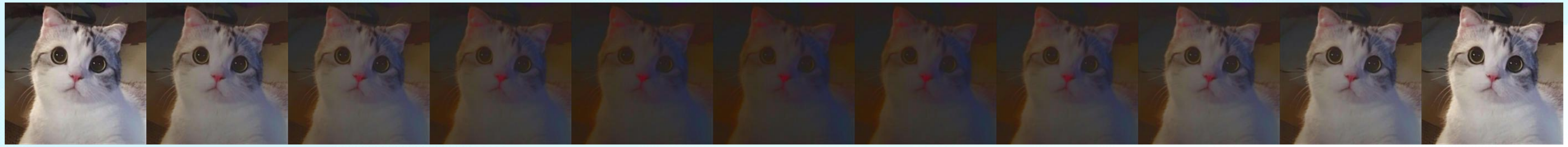
Loss Barrier

Conic Activation Functions

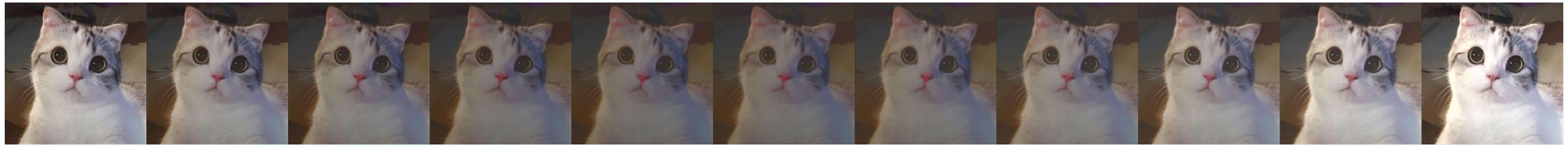
A symmetry constraint on generative models for improved **generalization property** and better **learning and performance**.

Linear Mode Connectivity: Generative Models

Before



After



Outputs of CNN with interpolated parameters

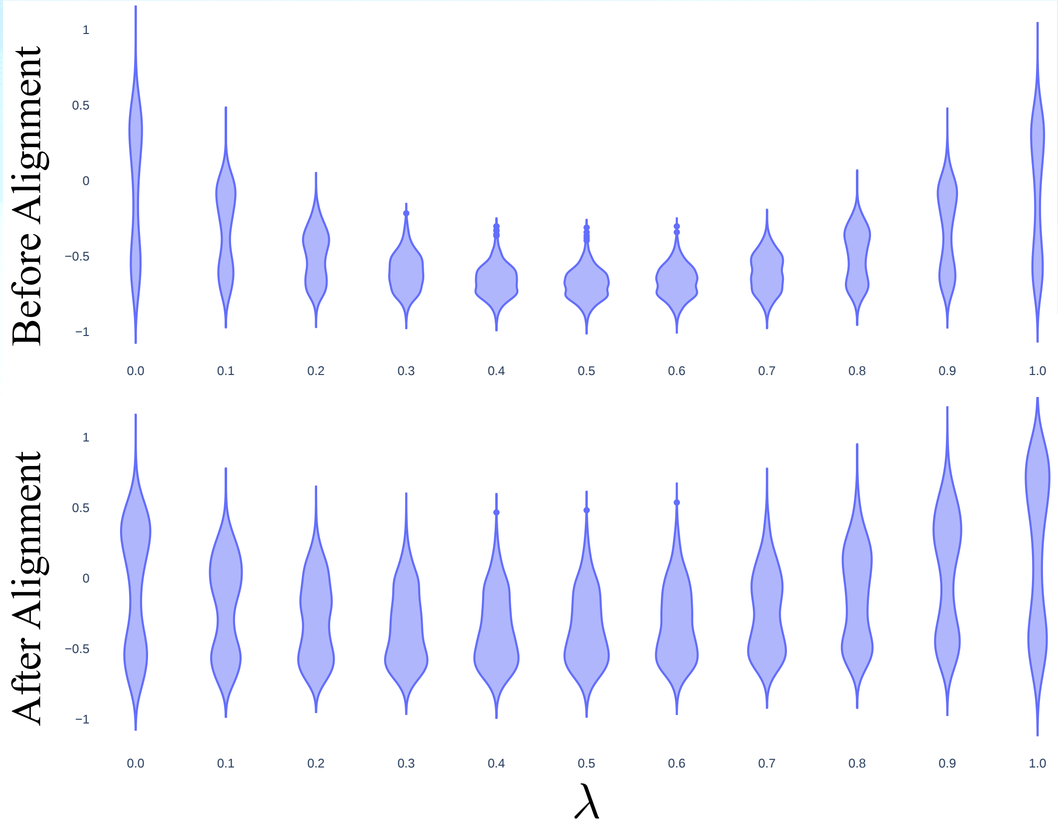


Image Histogram

Linear Mode Connectivity: Generative Models

Animation: interpolation between parameters in a finetuned diffusion model

Conclusion

CoLU is a symmetry constraint on generative models for improved **generalization property** and better **learning and performance**.