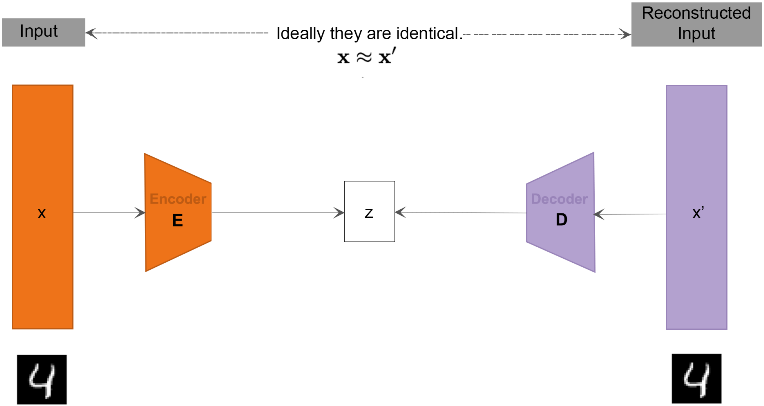
Deep Autoencoders: From Understanding to Generalization Guarantees

Romain Cosentino [1] Randall Balestriero [1] Richard Baraniuk [1] Behnaam Aazhang [1]

[1]Rice University


Input
Ideally they are identical.
$\mathbf{x} \approx \mathbf{x}'$
Reconstructed Input
x
Encoder
E
z
Decoder
D
x'

Autoencoders: Piecewise Affine Formalism

$$\boldsymbol{D} \circ \boldsymbol{E}(\boldsymbol{x}) = \sum_{\omega \in \Omega} 1_{\{\boldsymbol{x} \in \omega\}} (A_\omega^D A_\omega^E \boldsymbol{x} + A_\omega^D B_\omega^E + B_\omega^D),$$

- $\boldsymbol{x} \in \omega \subset \mathbb{R}^d$
- Ω is a partition of the space
- $A_\omega^D \in \mathbb{R}^{d \times h}$, $A_\omega^E \in \mathbb{R}^{h \times d}$, $B_\omega^E \in \mathbb{R}^h$ and $B_\omega^D \in \mathbb{R}^d$ with d being the dimension of the input data and h the bottleneck dimension.

$$A_\omega^E = W^L Q_\omega^{L-1} W^{L-1} \dots Q_\omega^1 W^1 \quad \text{and} \quad B_\omega^E = \boldsymbol{b}^L + \sum_{i=1}^{L-1} W^L Q_\omega^{L-1} W^{L-1} \dots Q_\omega^i \boldsymbol{b}^i.$$

- $W^\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}, \boldsymbol{b}^\ell \in \mathbb{R}^{d_\ell}$ the affine parameters of each layer,
- Q^ℓ the diagonal matrices encoding the region induced states of the nonlinearities, $(0, 1)$ for ReLU, $(-1, 1)$ for absolute value

2 Layers ReLU Network : Piecewise Affine Partitions

$$f(\boldsymbol{x}) = W^2 \mathrm{ReLU}(W^1 \boldsymbol{x}), \quad \text{where } W^1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, W^2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

- $\boldsymbol{x}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$,

$$f(\boldsymbol{x}_1) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \mathrm{ReLU}\left(\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}}_{Q_{\omega_1}} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

$$\rightarrow \boldsymbol{x}_1 \in \omega_1$$

- $\boldsymbol{x}_2 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$,

$$f(\boldsymbol{x}_2) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \mathrm{ReLU}\left(\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ -2 \end{pmatrix} \right) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{Q_{\omega_2}} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

$$\rightarrow \boldsymbol{x}_2 \in \omega_2$$

- $\boldsymbol{x}_3 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$

$$\rightarrow \boldsymbol{x}_3 \in \omega_1$$

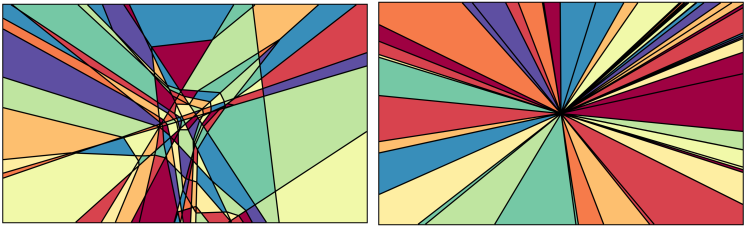


FIGURE 1: **2-dimensional visualizations of the input space partitioning** - To reconstruct its input, an AE achieves an affine map for each region - (*Left*) with bias (*Right*) zeros bias.

- Each region (described by a specific color) has a particular: $A^D_\omega \in \mathbb{R}^{d \times h}$, $A^E_\omega \in \mathbb{R}^{h \times d}$, $B^E_\omega \in \mathbb{R}^h$ and $B^D_\omega \in \mathbb{R}^d$.
- The "code" of each region, $\omega \in \Omega$, is given by the Q^ℓ_ω.

Number of regions VS Number of data

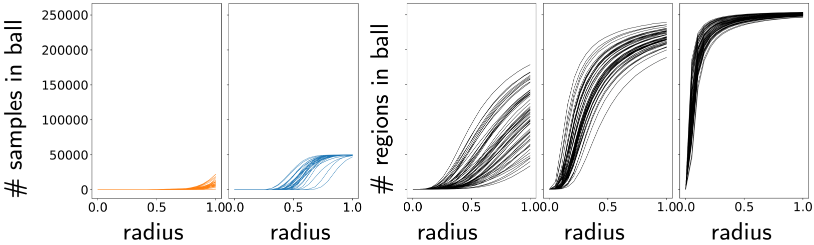
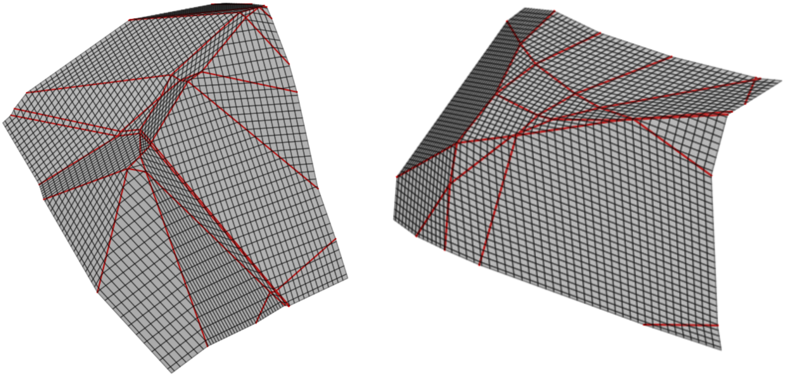


Figure 2: *Left:* data - Cifar10 and MNIST. *Right:* regions - small MLP, large MLP, ConvNet.

- Regardless of the dataset and neural network architecture: **the number of regions for any given ball is much larger than the number of data**

Objective

Understand how one can control these regions to equip *autoencoders* with **generalisation guarantees**



Decoder Continuous Piecewise Affine Surface

Per region Jacobian of the decoder

$$\forall \omega \in \Omega^D, J_\omega[\boldsymbol{D}] = A^D_\omega,$$

where the columns of A^D_ω form the basis of the tangent space induced by $\boldsymbol{D}$.

Assumption: Data Lie on The Orbit of a Group

Orbit of digit "7" w.r.t Rotation Group

Which group should we consider?

Lie Group: Is a group that is a differentiable manifold.

$$\text{Rotation Group: } SO(2) = \left\{ \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} | \theta \in \mathbb{R}/2\pi\mathbb{Z} \right\}.$$

Exponential Map: Any matrix Lie group can be defined via an exponential map.

$$\text{Rotation Group: } SO(2) = \left\{ \exp(\theta G) | G = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \theta \in \mathbb{R}/2\pi\mathbb{Z} \right\}.$$

Orbit w.r.t Lie Group: The data x are modeled by

$$\boldsymbol{x}(\theta) = \exp(\theta G)\boldsymbol{x}(0),$$

the orbit of $\boldsymbol{x}(0)$ with respect to the group induced by $\exp(\theta G)$.

Equivariance Lie Group Regularizations

$$\mathcal{L}_{\text{LieReg}} = \underbrace{\sum_{i=1}^{n} \|\mathbf{D}(\mathbf{E}(x_i)) - x_i\|}_{\textbf{Reconstruction Error}} + \min_{G} \underbrace{\sum_{\omega \in \Omega^D} \min_{\theta} \left\| \exp(\theta G) A^D_{\omega_\mathbf{0}} - A^D_{\omega} \right\|}_{\textbf{Lie Group Assumption}}$$

Learning with an exponential map

$$\exp(\theta G)$$

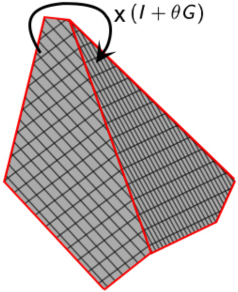
1. Non-Convex
2. Tedious Computation of the Gradient

Local approximation

$$\exp(\theta G) \approx_{\theta \sim 0} I + \theta G$$

1. Only for small transformations
2. Need to know the neighbors of each sample

$$\mathcal{L}_{\mathrm{LieReg}} \approx \sum_{i=1}^{n} \|\mathbf{D}(\mathbf{E}(x_i)) - x_i\| + \min_{G} \sum_{\omega \in \Omega^D} \sum_{\omega' \in \mathcal{N}(\omega)} \min_{\theta} \left\| A_{\omega}^{D} - (I + \theta G) A_{\omega'}^{D} \right\|,$$



1. **If** the Decoder approximates the tangent space of the data at a position.
2. **If** the Lie group regularization is 0.

Then

The approximation of the data manifold is upper-bounded by the sum of the radius of each region.

THEOREM

If on a region $\omega' \in \Omega^D$ the matrix $A^D_{\omega'}$ forms a basis of the manifold tangent space on this region, and the Lie group regularization is 0 then for all region $\omega \in \Omega^D$ the basis vectors of A^D_{ω} are the basis vector of the tangent of the data manifold with

$$d\left(\cup_{\omega \in \Omega^D} \mathcal{T}_{AE}(\omega), \mathcal{X}\right) \leq \sum_{\omega \in \Omega^D} Rad(\omega),$$

where $\mathcal{T}_{AE}(\omega)$ the tangent space of the AE for the region ω, $\mathcal{X}$ denotes the data manifold, d defines the 2-norm distance, and $Rad(\omega_i)$ the radius of the region ω_i.

Results

Table 1: Comparison of the testing reconstruction errors ($\times 10^{-2} \pm$ std $\times 10^{-2}$)

Dataset \ Model	**AE**	**Den. AE**	**H.O.C. AE**	**Lie Group**
CIFAR10	5.6 ± 0.05	5.0 ± 0.05	-	$\mathbf{4.9} \pm 0.07$
MNIST	12.01 ± 0.003	12.01 ± 0.004	12.01 ± 0.004	$\mathbf{6.3} \pm 0.1$
CBF	62.38 ± 0.74	52.66 ± 0.76	51.09 ± 0.54	$\mathbf{43.99} \pm 1.2$
Yoga	33.76 ± 0.81	33.29 ± 0.72	32.08 ± 0.42	$\mathbf{20.28} \pm 1.1$
Trace	13.95 ± 0.45	11.28 ± 0.57	12.57 ± 0.21	$\mathbf{10.91} \pm 0.45$
Wine	63.06 ± 0.02	59.34 ± 0.02	49.94 ± 0.02	$\mathbf{19.01} \pm 0.02$
ShapesAll	67.98 ± 3.0	58.67 ± 1.4	61.42 ± 5.5	$\mathbf{52.97} \pm 1.9$
FiftyWords	64.91 ± 1.7	60.91 ± 1.0	60.92 ± 0.7	$\mathbf{57.89} \pm 1.0$
WordSynonyms	70.95 ± 1.5	66.02 ± 0.8	66.52 ± 0.5	$\mathbf{62.22} \pm 1.1$
InsectSounds	51.86 ± 0.6	40.24 ± 0.8	41.93 ± 0.6	$\mathbf{38.11} \pm 0.9$
ECG5000	21.92 ± 0.75	20.31 ± 0.39	20.31 ± 0.36	$\mathbf{18.06} \pm 0.9$
Earthquakes	56.23 ± 4.1	54.62 ± 4.1	51.79 ± 1.0	$\mathbf{50.20} \pm 0.5$
Haptics	37.25 ± 0.2	36.02 ± 1.8	27.21 ± 0.5	$\mathbf{16.94} \pm 3.4$
FaceFour	49.82 ± 1.0	48.51 ± 0.8	48.52 ± 0.7	$\mathbf{46.00} \pm 0.6$
Synthetic	95.61 ± 1.3	89.37 ± 1.0	88.47 ± 0.9	$\mathbf{55.87} \pm 0.8$

Conclusion & Directions

- We propose a way to learn an equivariant AE.
- The underlying group is learned via the Lie group generator G.
- Under the Lie group assumption on the data, we obtain generalization guarantees.
- Propose a way to develop contraints on the approximated manifold that are assumption driven.

- Learning Lie group is non trivial.
- Generalizing to "pancakes" and multiple orbits.
- Provide efficient and principled ways to sample neighboring regions.