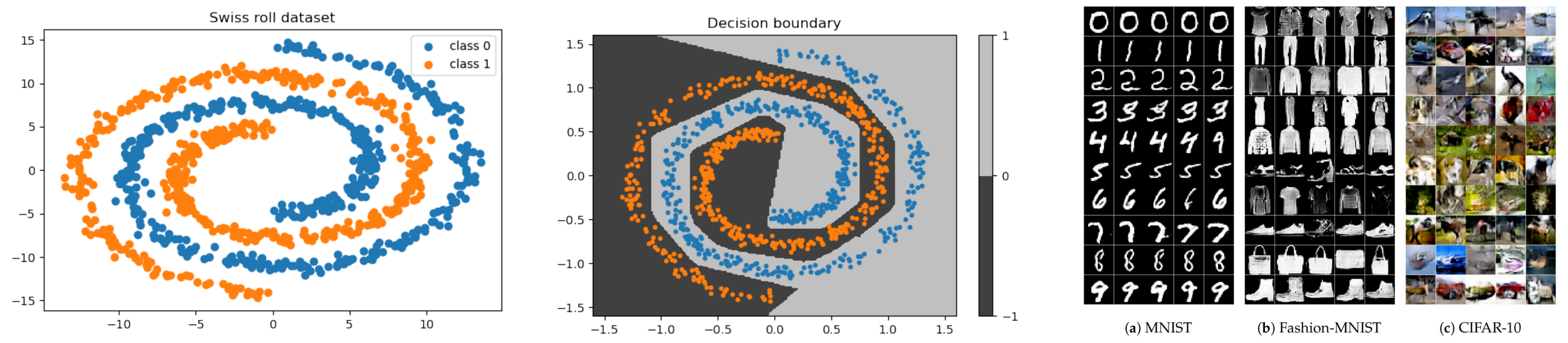
Defining Neural Network Architecture through Polytope Structures of Datasets

Sangmin Lee, Abbas Mammadov, Jong Chul Ye
KAIST

ICML International Conference On Machine Learning

Introduction



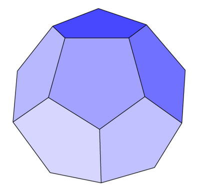
Q1. What is the minimal / available network architecture to solve this classification problem ?

Q2. How does the geometric complexity of datasets affect the network size ?

A. It can be answered through the polytope structure of the dataset.

Background

Definition. ***Convex polytope*** $\mathcal{C}$ with m-faces.



$$\mathcal{C} \;:= \cap_{k=1}^{m} \{ x \in \mathbb{R}^d \mid w_k^\top x + b_k \le 0 \}$$

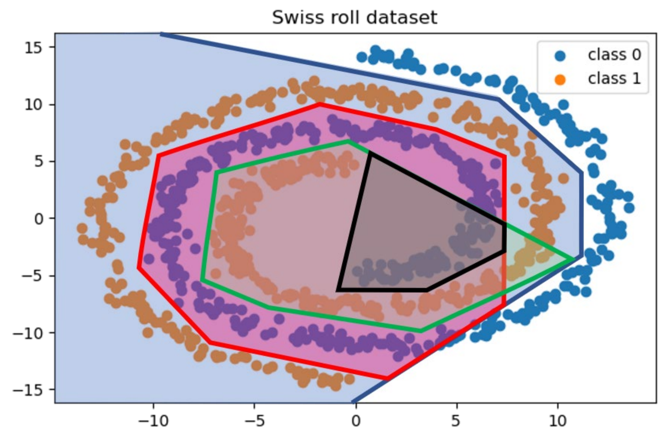
Definition. ***Polytope-basis cover*** $\mathcal{C}$ of a dataset $\mathcal{D} = \mathcal{D}_+ \cup \mathcal{D}_-$.

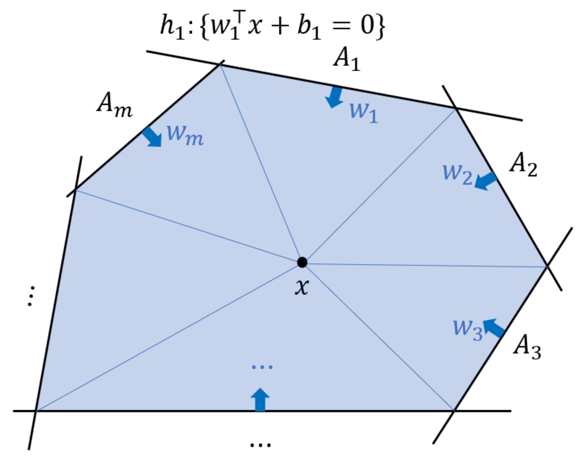
A collection of polytopes

$$\mathcal{C} := \{ P_1, \cdots, P_{n_P}, Q_1, \cdots, Q_{n_Q} \}$$

is a polytope-basis cover of $\mathcal{D}$ if

$$\sum_{k=1}^{n_P} \mathbb{I}_{\{x \in P_k\}} > \sum_{k=1}^{n_Q} \mathbb{I}_{\{x \in Q_k\}} \quad \text{if and only if} \quad x \in \mathcal{D}_+.$$





Proposition 3.1 & F.6. ***Lower and Upper bounds for network widths.***

Let C be a convex polytope with m faces. Then,

$$d \overset{\sigma}{\to} m \to 1$$

is the minimum width for universal approximation. Conversely, if $d \overset{\sigma}{\to} d_1 \overset{\sigma}{\to} \cdots \overset{\sigma}{\to} d_k \to 1$ is a feasible architecture on $\mathcal{X}$, then

$$d_1 \cdot \prod_{j=2}^{k}(2d_j+1) \ge \begin{cases} \left\lceil \frac{m}{2} \right\rceil + d - 2 & \text{if} \quad m \ge 2d+1, \\ 2d-1 & \text{if} \quad m = 2d-1, 2d, \\ d+1 & \text{if} \quad m < 2d-1. \end{cases}$$

Theory

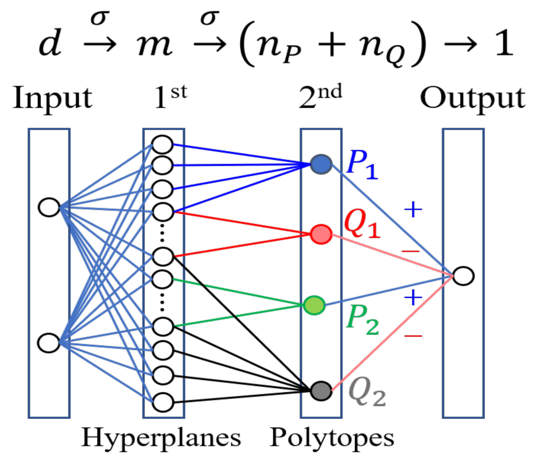
Theoretical results : Network architectures ∝ geometric complexity of training datasets

Theorem 3.4. ***Explicit construction of a 3-layer network.***

Let $\mathcal{C}$ be a polytope-basis cover of $\mathcal{X}$. Then,

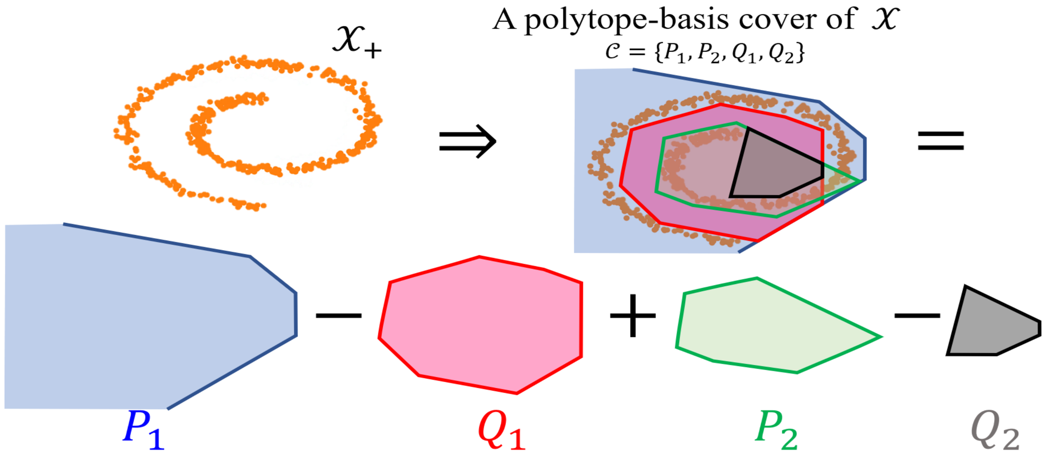
$$d \overset{\sigma}{\to} m \overset{\sigma}{\to} (n_P + n_Q) \to 1$$

is a feasible architecture on $\mathcal{X}$.



Remark. ***The significance of neurons.***

- Neurons in the 1st layer : ***hyperplane***
- Neurons in the 2nd layer : ***polytopes***
- Neurons in the 3rd layer : ***polytope-basis covers***



Theorem 3.5 & 3.6. ***Network architectures ∝ network widths.***

If $\mathcal{X}$ is a simplicial J-complex consists of k faces, then $d \overset{\sigma}{\to} d_1 \overset{\sigma}{\to} k \to 1$ is a feasible architecture with

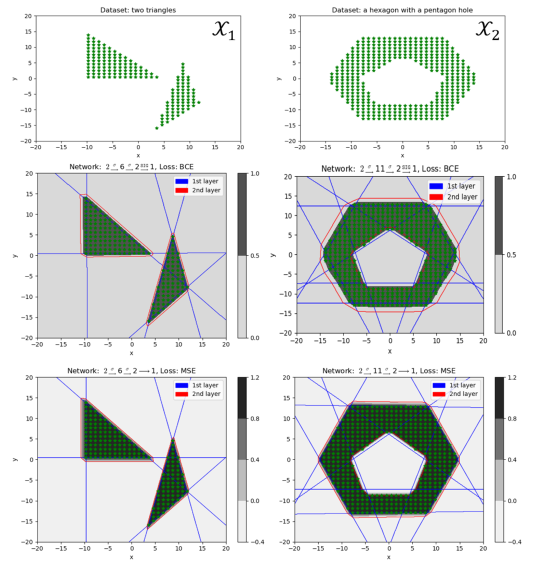
$$d_1 \;\le\; \min\left\{ k(d+1) - (d-1)\left\lfloor \sum_{j=0}^{\left\lfloor\frac{d-1}{2}\right\rfloor} \frac{k_j}{2} \right\rfloor, (d+1)\left\lfloor \sum_{j\le\frac{d}{2}} \left(k_j \frac{j+2}{d-j} + \frac{j+2}{j+1}\right) + \sum_{j>\frac{d}{2}} k_j \right\rfloor \right\} \;\approx\; O\left(k\frac{J+2}{d-J}+2\right).$$

If $\mathcal{X}$ can be separated by disjoint prismatic polytopes, then

$$d \overset{\sigma}{\to} \left(m + 2(\beta_0 - 1) + \sum_{k=1}^{d} (m - 2(d-k-1))\beta_k \right) \overset{\sigma}{\to} \left(\sum_{k=0}^{d} \beta_k \right) \to 1$$

is a feasible architecture.

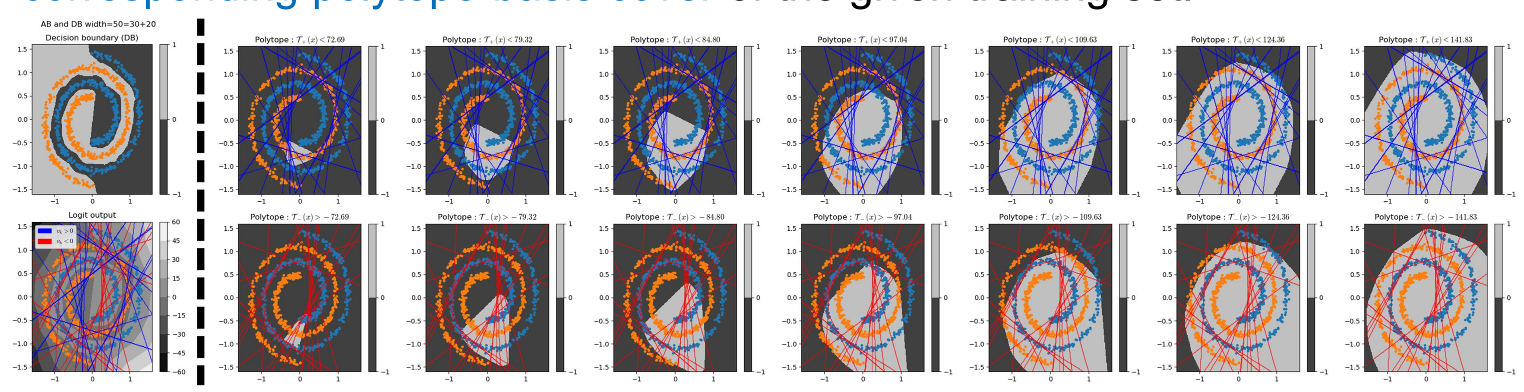
Convergence



Deriving a polytope-basis cover from a trained network

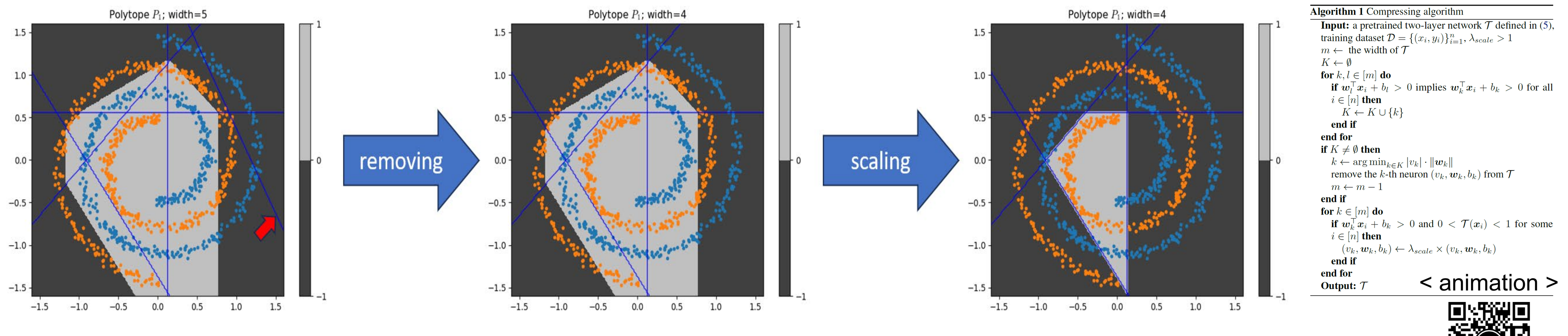
Theorem 3.7. ***Embedded polytope structures in a trained network.***

If a three-layer ReLU network $\mathcal{N}$ satisfies some conditions, then it induces a corresponding polytope-basis cover of the given training set.



Experiments

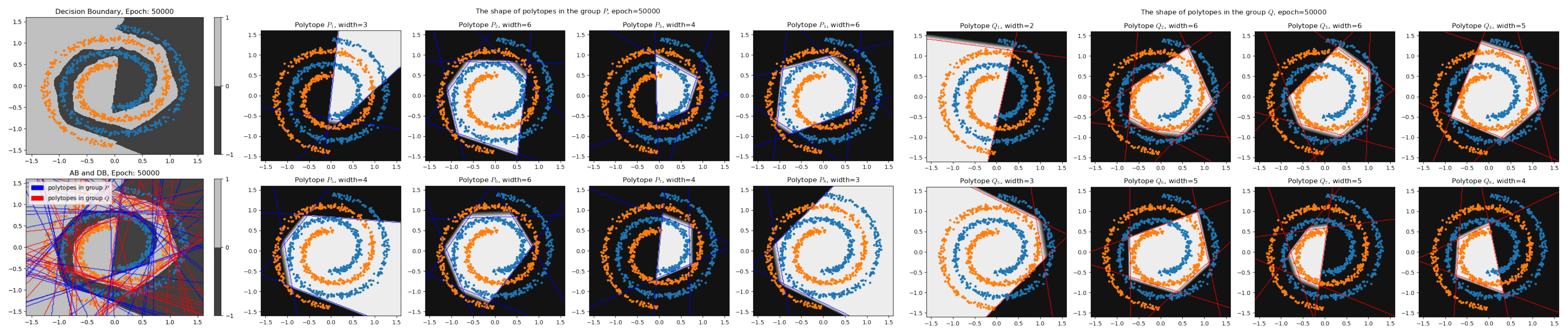
Compressing algorithm (Algorithm 1)



(a) (b) (c)

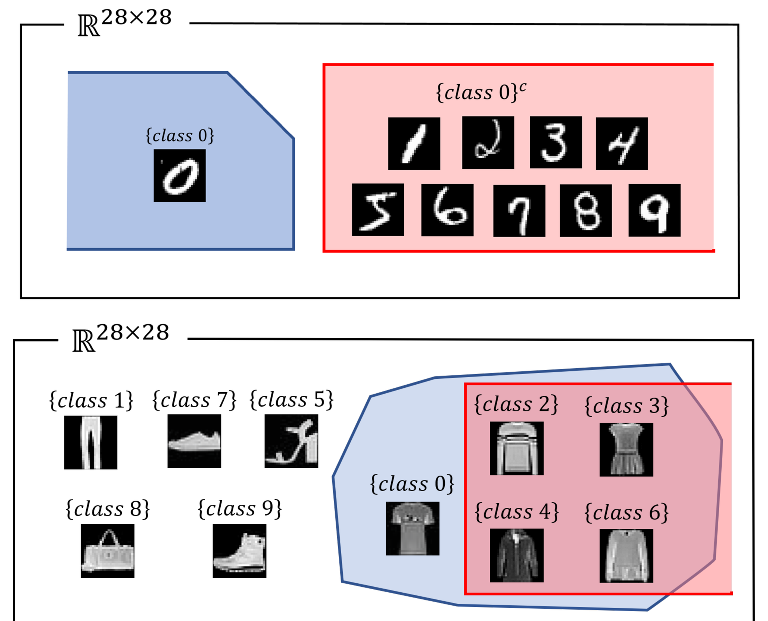
< animation >

Experimental results : Polytope-basis covers on toy datasets



Experimental results : Polytope-basis covers on real datasets (MNIST, F-MNIST, CIFAR10)

Datasets		Class 0	1	2	3	4	5	6	7	8	9
MNIST	{class}	4	4	7	8	5	7	4	8	8	7
	$\{class\}^c$	3	3	4	5	4	5	4	4	9	9
Fashion-MNIST	{class}	9+3	4	9+5	9+3	9+6	8	9+7	9+1	6	10
	$\{class\}^c$	16	3	22	11	20	4	28	6	4	5
CIFAR10	{class}	29+3	19	23+3	24+4	19	16+3	21	21	18	21
	$\{class\}^c$	29	7	27+3	26	26+4	17	13	10	20+4	8



❑ Uniqueness of the polytope covers