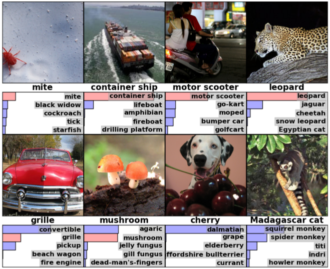
Lower Bounds on Neural Network Depth via Lattice Polytopes

Christian Haase **Christoph Hertrich** Georg Loho

Bielefeld, 6 September 2022

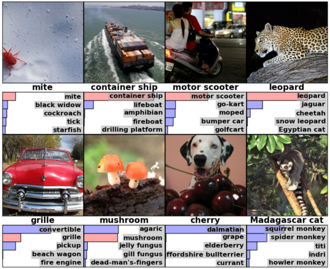
Neural Networks in Action



Krizhevsky et al. "Imagenet classification with deep convolutional neural networks" (NeurIPS 2012)

Neural Networks in Action

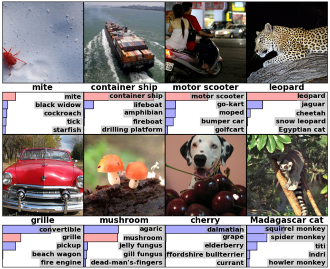
Gatys et al. "Image style transfer using convolutional neural networks" (CVPR 2016)



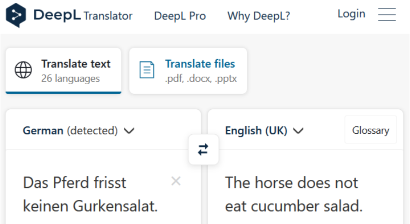
Krizhevsky et al. "Imagenet classification with deep convolutional neural networks" (NeurIPS 2012)

Neural Networks in Action

Gatys et al. "Image style transfer using convolutional neural networks" (CVPR 2016)

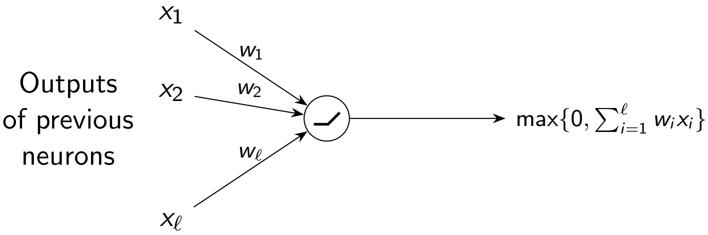


Krizhevsky et al. "Imagenet classification with deep convolutional neural networks" (NeurIPS 2012)

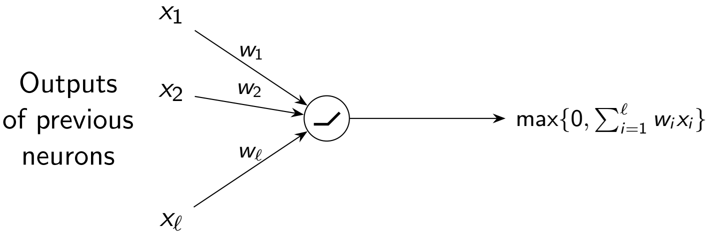


Screenshot `deepl.com` (Feb 18, 2022)

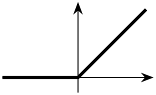
A Single ReLU Neuron



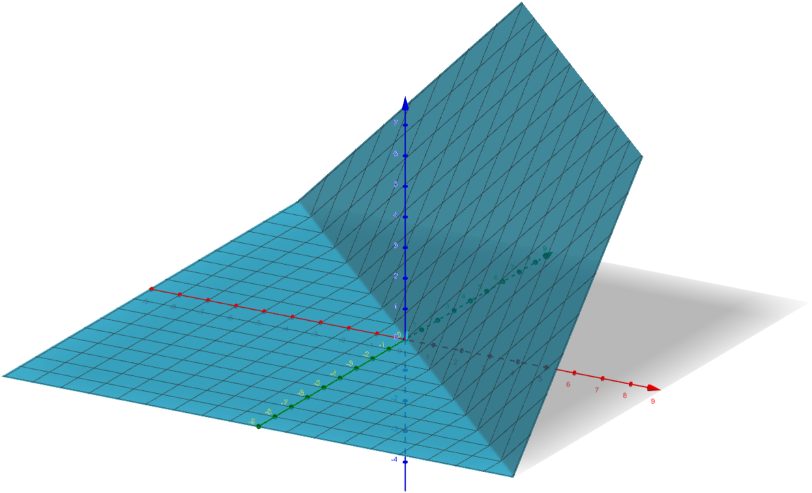
A Single ReLU Neuron



Rectified linear unit (ReLU): $\text{relu}(x) = \max\{0, x\}$

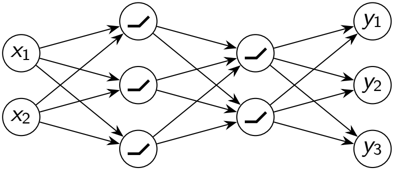


A Single ReLU Neuron



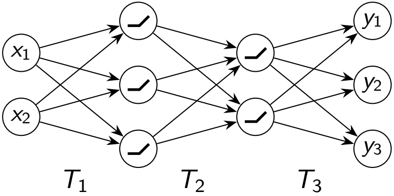
ReLU Feedforward Neural Networks

- Acyclic (layered) digraph of ReLU neurons



ReLU Feedforward Neural Networks

- Acyclic (layered) digraph of ReLU neurons



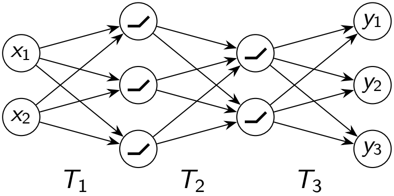
- Computes function

$$T_k \circ \mathrm{relu} \circ T_{k-1} \circ \cdots \circ T_2 \circ \mathrm{relu} \circ T_1$$

with linear transformations T_i.

ReLU Feedforward Neural Networks

- Acyclic (layered) digraph of ReLU neurons



- Computes function

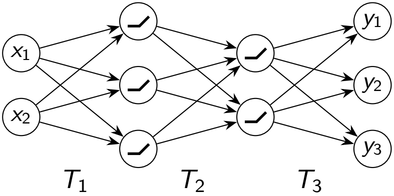
$$T_k \circ \mathrm{relu} \circ T_{k-1} \circ \cdots \circ T_2 \circ \mathrm{relu} \circ T_1$$

  with linear transformations T_i.

- Example: depth 3 (2 hidden layers).

ReLU Feedforward Neural Networks

- Acyclic (layered) digraph of ReLU neurons



- Computes function

$$T_k \circ \mathrm{relu} \circ T_{k-1} \circ \cdots \circ T_2 \circ \mathrm{relu} \circ T_1$$

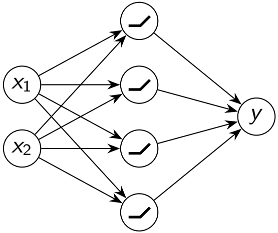
with linear transformations T_i.

- Example: depth 3 (2 hidden layers).
- Usage: Learn weights of T_i from given input-output pairs.

What is the class of functions computable by
ReLU Neural Networks
with a certain depth?

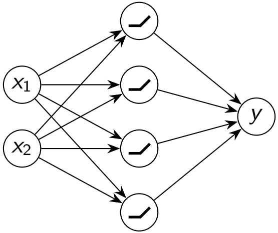
Universal approximation theorems:

One hidden layer enough to **approximate** any continuous function.



Universal approximation theorems:

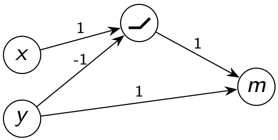
One hidden layer enough to **approximate** any continuous function.



What about **exact** representability?

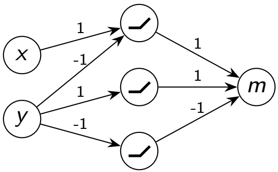
Example: Computing the Maximum of Two Numbers

$$\max\{x, y\} = \max\{x - y, 0\} + y$$

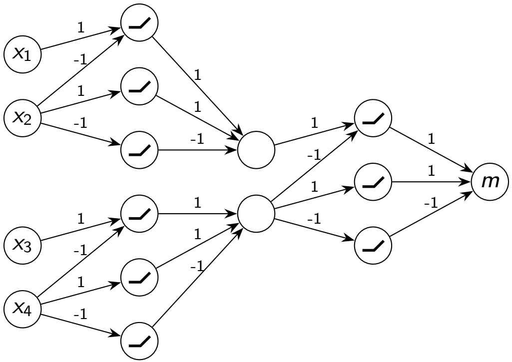


Example: Computing the Maximum of Two Numbers

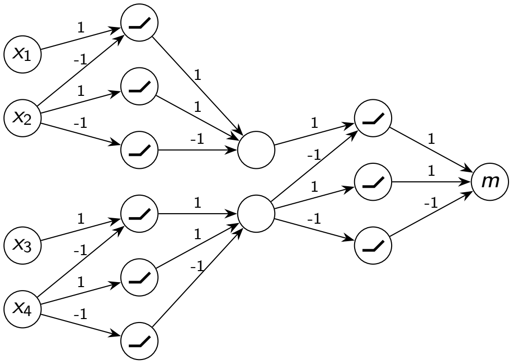
$$\max\{x, y\} = \max\{x - y, 0\} + y$$



Example: Computing the Maximum of Four Numbers



Example: Computing the Maximum of Four Numbers



- Inductively: Max of n numbers with $\lceil \log_2(n) \rceil$ hidden layers.

Representing Arbitrary Piecewise Linear Functions

Observation

Every function represented by a ReLU NN is continuous and piecewise linear (CPWL).

Representing Arbitrary Piecewise Linear Functions

Observation

Every function represented by a ReLU NN is continuous and piecewise linear (CPWL).

Theorem (Wang, Sun [WS05])

Any CPWL function $f : \mathbb{R}^n \to \mathbb{R}$ can be written as

$$f(x) = \sum_{i=1}^{p} \lambda_i \max\{a_{i,1}^T x, \ldots, a_{i,n+1}^T x\}.$$

Representing Arbitrary Piecewise Linear Functions

Observation

Every function represented by a ReLU NN is continuous and piecewise linear (CPWL).

Theorem (Wang, Sun [WS05])

Any CPWL function $f : \mathbb{R}^n \to \mathbb{R}$ can be written as

$$f(x) = \sum_{i=1}^{p} \lambda_i \max\{a_{i,1}^T x, \ldots, a_{i,n+1}^T x\}.$$

Theorem (Arora, Basu, Mianjy, Mukherjee [ABMM18])

Any CPWL function $f : \mathbb{R}^n \to \mathbb{R}$ can be represented by a ReLU NN with $\lceil \log_2(n+1) \rceil$ hidden layers.

Natural Question

Theorem (Arora, Basu, Mianjy, Mukherjee [ABMM18])
Any CPWL function $f : \mathbb{R}^n \to \mathbb{R}$ can be represented by a ReLU NN with $\lceil \log_2(n+1) \rceil$ hidden layers.

- Is logarithmic depth best possible?

Conjecture

Yes, there are functions which need $\lceil \log_2(n+1) \rceil$ *hidden layers!*

Conjecture

Yes, there are functions which need $\lceil \log_2(n+1) \rceil$ *hidden layers!*

Using [WS05], we show that this is equivalent to:

Conjecture

$\max\{0, x_1, \ldots, x_{2^k}\}$ *cannot be represented with* k *hidden layers.*

What lower bounds are known?

What lower bounds are known?

- Mukherjee, Basu (2017):
  $\max\{0, x_1, x_2\}$ not representable with 1 hidden layer.

What lower bounds are known?

- Mukherjee, Basu (2017):
  $\max\{0, x_1, x_2\}$ not representable with 1 hidden layer.

That's all!

What lower bounds are known?

- Mukherjee, Basu (2017):
  $\max\{0, x_1, x_2\}$ not representable with 1 hidden layer.

That's all!

- No function known that provably needs more than 2 hidden layers $\rightsquigarrow$ gap between 2 and $\lceil \log_2(n+1) \rceil$.

What lower bounds are known?

- Mukherjee, Basu (2017):
  $\max\{0, x_1, x_2\}$ not representable with 1 hidden layer.

That's all!

- No function known that provably needs more than 2 hidden layers $\rightsquigarrow$ gap between 2 and $\lceil \log_2(n+1) \rceil$.
- Smallest candidate: $\max\{0, x_1, x_2, x_3, x_4\}$.

Partial Results

- Hertrich, Basu, Di Summa, Skutella (NeurIPS 2021):

  2 hidden layers not enough for $\max\{0, x_1, x_2, x_3, x_4\}$
  under an additional assumption on the network.

Partial Results

- Hertrich, Basu, Di Summa, Skutella (NeurIPS 2021):

  2 hidden layers not enough for $\max\{0, x_1, x_2, x_3, x_4\}$
  under an additional assumption on the network.

- Haase, Hertrich, Loho (this talk!):

  Conjecture is true for networks with only integer weights.

How do Polytopes Come into Play?

How do Polytopes Come into Play?

$$\begin{aligned}\text{CPWL function} &= \text{difference of two convex CPWL functions} \\ &= \text{difference of two tropical polynomials} \\ &= \text{tropical rational function}\end{aligned}$$

How do Polytopes Come into Play?

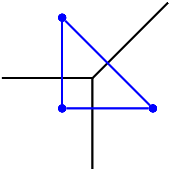
$$\begin{aligned}\text{CPWL function} &= \text{difference of two convex CPWL functions}\\ &= \text{difference of two tropical polynomials}\\ &= \text{tropical rational function}\end{aligned}$$

$\rightsquigarrow$ study **Newton polytopes**!

Newton Polytope of a Convex CPWL Function

- $f(x) = \max\{a_1^T x, \ldots, a_k^T x\} \quad \rightsquigarrow \quad P(f) = \mathrm{conv}\{a_1, \ldots, a_k\}$
- dual to underlying polyhedral complex of the CPWL function

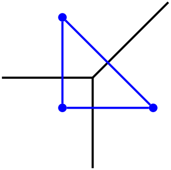
Example for $\max\{0, x_1, x_2\}$:



Newton Polytope of a Convex CPWL Function

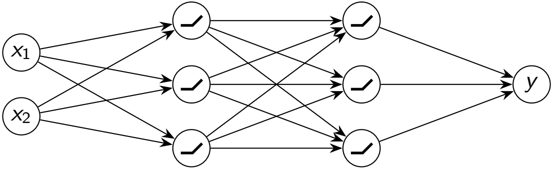
- $f(x) = \max\{a_1^T x, \ldots, a_k^T x\} \quad \rightsquigarrow \quad P(f) = \operatorname{conv}\{a_1, \ldots, a_k\}$
- dual to underlying polyhedral complex of the CPWL function

Example for $\max\{0, x_1, x_2\}$:

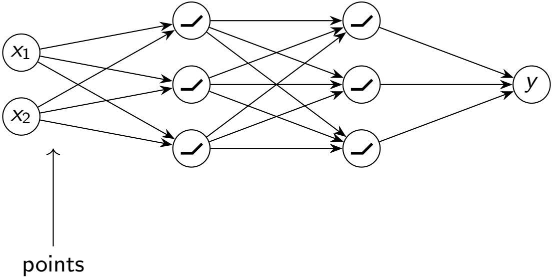


Convex CPWL functions	$\cong$	**Newton Polytopes**
(positive) scalar multiplication		scaling
addition		Minkowski sum
taking maximum		taking convex hull of union

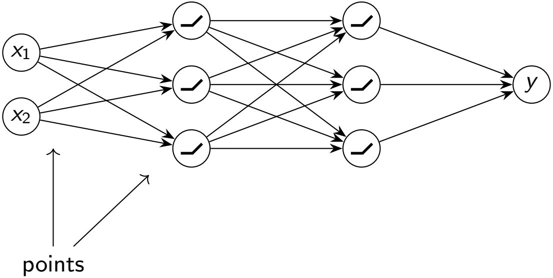
Newton Polytopes and Neural Networks



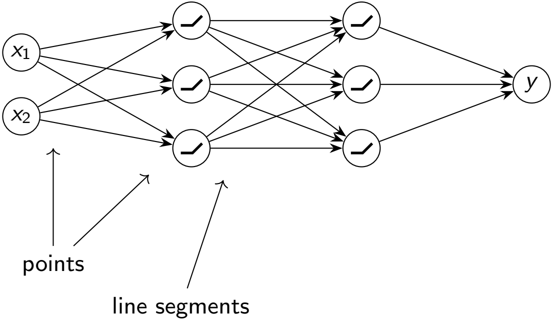
Newton Polytopes and Neural Networks



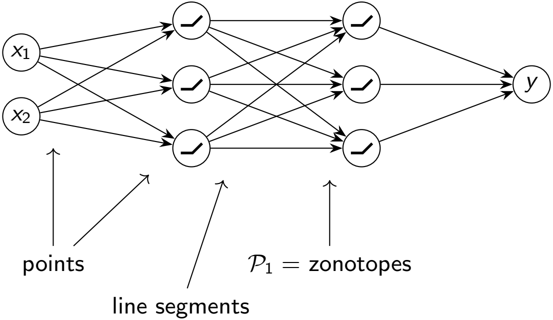
Newton Polytopes and Neural Networks



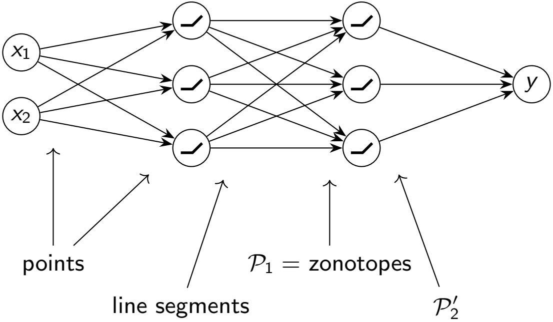
Newton Polytopes and Neural Networks



Newton Polytopes and Neural Networks

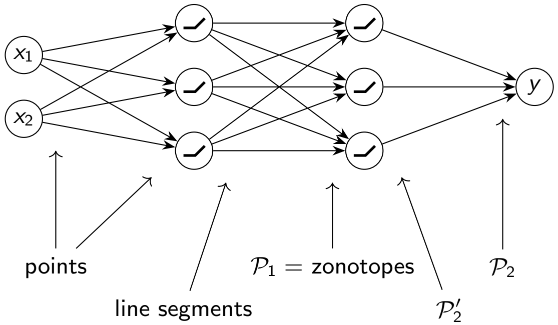


Newton Polytopes and Neural Networks



$$\mathcal{P}'_2 = \{P \text{ polytope} \mid P \text{ convex hull of union of two zonotopes}\}$$

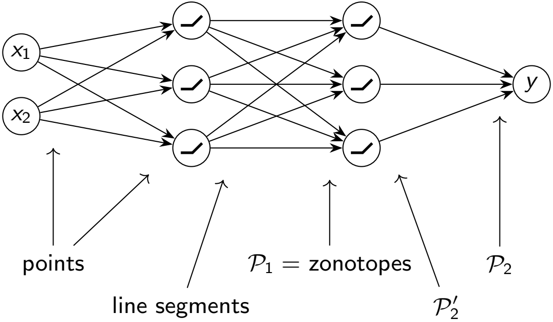
Newton Polytopes and Neural Networks



$$\mathcal{P}_2' = \{P \text{ polytope} \mid P \text{ convex hull of union of two zonotopes}\}$$

$$\mathcal{P}_2 = \{P \text{ polytope} \mid P \text{ finite Minkowski sum of polytopes in } \mathcal{P}_2'\}$$

Newton Polytopes and Neural Networks



$\mathcal{P}_2' = \{P \text{ polytope} \mid P \text{ convex hull of union of two zonotopes}\}$

$\mathcal{P}_2 = \{P \text{ polytope} \mid P \text{ finite Minkowski sum of polytopes in } \mathcal{P}_2'\}$

Newton polytope of $\max\{0, x_1, x_2, x_3, x_4\}$: 4-dim. simplex Δ^4.

Are there polytopes $Q, R \in \mathcal{P}_2$ with $Q + \Delta^4 = R$?

Polytopal Reformulation of the Conjecture

$$\mathcal{P}_0 := \{\text{points}\}$$

$$\mathcal{P}_1 := \{\text{zonotopes}\}$$

$$\mathcal{P}_k := \left\{ \sum_{i=1}^{m} \operatorname{conv}(P_i, Q_i) \;\middle|\; P_i, Q_i \in \mathcal{P}_{k-1}, m \in \mathbb{N} \right\}$$

$$\Delta^n := \operatorname{conv}\{0, e_1, e_2, \ldots, e_n\}$$

Polytopal Reformulation of the Conjecture

$$\mathcal{P}_0 := \{\text{points}\}$$
$$\mathcal{P}_1 := \{\text{zonotopes}\}$$
$$\mathcal{P}_k := \left\{ \sum_{i=1}^{m} \operatorname{conv}(P_i, Q_i) \;\middle|\; P_i, Q_i \in \mathcal{P}_{k-1}, m \in \mathbb{N} \right\}$$
$$\Delta^n := \operatorname{conv}\{0, e_1, e_2, \ldots, e_n\}$$

Conjecture

There is no pair of polytopes $P, Q \in \mathcal{P}_k$ such that $P + \Delta^{2^k} = Q$.

Results for the Integer Case

Haase, Hertrich, Loho (work in progress)

From now on: all weights integer!

Note: logarithmic upper bound only uses integer weights

Results for the Integer Case

Haase, Hertrich, Loho (work in progress)

From now on: all weights integer!

Note: logarithmic upper bound only uses integer weights

$\Rightarrow$ All Newton polytopes are lattice polytopes

Results for the Integer Case

Haase, Hertrich, Loho (work in progress)

From now on: all weights integer!

Note: logarithmic upper bound only uses integer weights

$\Rightarrow$ All Newton polytopes are lattice polytopes

Let $\mathcal{P}_k^{\mathbb{Z}} := \mathcal{P}_k \cap \{\text{lattice polytopes}\}$.

Results for the Integer Case

Haase, Hertrich, Loho (work in progress)

From now on: all weights integer!

Note: logarithmic upper bound only uses integer weights

$\Rightarrow$ All Newton polytopes are lattice polytopes

Let $\mathcal{P}_k^{\mathbb{Z}} \coloneqq \mathcal{P}_k \cap \{\text{lattice polytopes}\}$.

Theorem

There is no pair of polytopes $P, Q \in \mathcal{P}_k^{\mathbb{Z}}$ *such that* $P + \Delta^{2^k} = Q$.

Results for the Integer Case

Haase, Hertrich, Loho (work in progress)

From now on: all weights integer!

Note: logarithmic upper bound only uses integer weights

$\Rightarrow$ All Newton polytopes are lattice polytopes

Let $\mathcal{P}_k^{\mathbb{Z}} := \mathcal{P}_k \cap \{\text{lattice polytopes}\}$.

Theorem

There is no pair of polytopes $P, Q \in \mathcal{P}_k^{\mathbb{Z}}$ such that $P + \Delta^{2^k} = Q$.

Corollary

The minimum number of hidden layers to represent $\max\{0, x_1, \ldots, x_{2^k}\}$ with integer weights is precisely $k + 1$.

Proof Idea

- Consider normalized volume of lattice polytopes ($\in \mathbb{Z}$)
- By carefully subdividing Minkowski sums and convex hulls:

Lemma

A 2^k-dimensional polytope in $\mathcal{P}_k^{\mathbb{Z}}$ has even normalized volume.

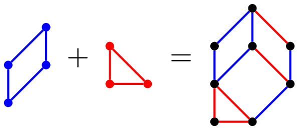
Proof Idea

- Consider normalized volume of lattice polytopes ($\in \mathbb{Z}$)
- By carefully subdividing Minkowski sums and convex hulls:

Lemma

A 2^k-dimensional polytope in $\mathcal{P}_k^{\mathbb{Z}}$ has even normalized volume.

- Theorem follows because Δ^{2^k} has normalized volume one.
- Example in 2D:



Outlook

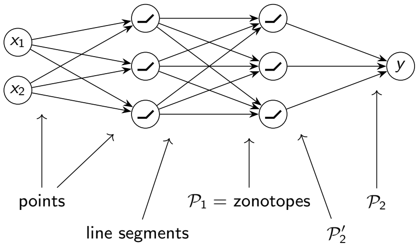
- Volume unsuitable for non-integer case
- Find different way to separate $\mathcal{P}_k$ from each other!

Outlook

- Volume unsuitable for non-integer case
- Find different way to separate $\mathcal{P}_k$ from each other!

Conjecture

There is no pair of polytopes $P, Q \in \mathcal{P}_k$ such that $P + \Delta^{2^k} = Q$.

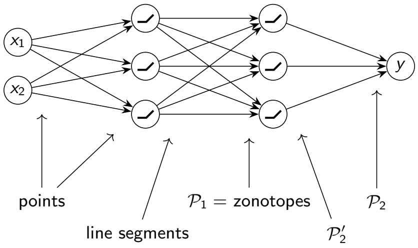


Outlook

- Volume unsuitable for non-integer case
- Find different way to separate $\mathcal{P}_k$ from each other!

Conjecture

There is no pair of polytopes $P, Q \in \mathcal{P}_k$ *such that* $P + \Delta^{2^k} = Q$.



Thank you!