Approximation Theory of Group Invariant Neural Networks

Nadav Dym

Technion

Is this your first time in Banff?

In July 2003 (age 16) I attended:

Mathematical Biology: From molecultes to ecosystems: the legacy of Lee Segel

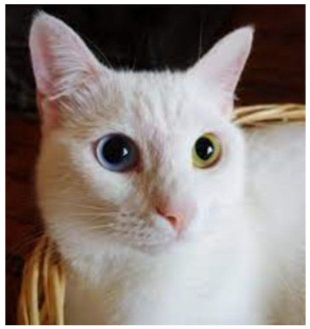
`While he liked talking about his work, he had the rare quality of actually being interested in hearing about other people's work (Daniel Segel, free translation)'

Approximation Theory of Group Invariant Neural Networks

Approximation Theory of Group Invariant Neural Networks

Supervised Machine Learning: Learn f from examples

x_i



$f(x_i)$ $`cat'$ $`dog'$ $`cat'$ $'dog'$

$$\min_{h \in \mathrm{H}} \sum_{i=1}^{N} |f(x_i) - h(x_i)|^2$$

?

Neural Networks

Approximation Theory for Group Invariant Neural Networks

Activation function: $\sigma: \mathbb{R} \to \mathbb{R}$

Induces $\sigma: \mathbb{R}^d \to \mathbb{R}^d$ $\quad \sigma(x_1, \dots, x_d) = (\sigma(x_1), \dots, \sigma(x_d))$

Affine functions $h^{(i)}(x) = A^{(i)}x + b^{(i)}$ where $h^{(i)}: \mathbb{R}^{w_{i-1}} \to \mathbb{R}^{w_i}$

<u>Definition:</u> We say that $\mathcal{N}: \mathbb{R}^d \to \mathbb{R}^m$ is a **fully connected neural network** if

$$\mathcal{N}(x) = h_{L+1} \circ \sigma \circ h_L \circ \sigma \circ \cdots \circ \sigma \circ h_0(x)$$

Depth of $\mathcal{N}$:= L

Width of $\mathcal{N}$:= Maximal dimension $\max_{1 \le i \le L} w_i$

Approximation Theory of Group Invariant Neural Networks

Universality Theorem [Cybenko 1989, Pinkus 1999,many others in between]

If the activation function: $\sigma: \mathbb{R} \to \mathbb{R}$ is continuous and not polynomial

then for every compact $K \subseteq \mathbb{R}^d$, continuous $f: K \to \mathbb{R}$ and $\epsilon > 0$,

There exists a **fully connected neural network** $\mathcal{N}: \mathbb{R}^d \to \mathbb{R}$ **of depth L=1** (and arbitrarily large width)

$$\mathcal{N}(x) = h_1 \circ \sigma \circ h_0(x)$$

Such that

$$|f(x) - \mathcal{N}(x)| < \epsilon, \qquad \forall x \in K$$

Universality- provides justification for choosing neural networks as a function space for any continuous learning task.

Approximation Theory of Group Invariant Neural Networks

Beyond universality- rates of approximation (More recent research)

Given $f: K \to \mathbb{R}$ which is Lispschitz/smooth/fractal and ϵ what width $W(\epsilon)$ and depth $L(\epsilon)$ are necessary to achieve an ϵ approximation?

Approximation Theory of Group Invariant Neural Networks

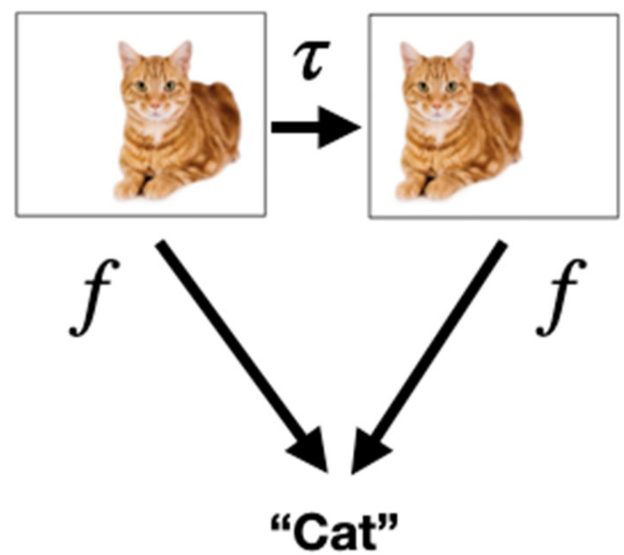
$$\min_{h \in H} \sum_i |f(x_i) - h(x_i)|^2$$

<u>Invariant networks:</u>

Construct $H = H_{inv}$ so that all $h \in H_{inv}$ are invariant to the symmetries of f

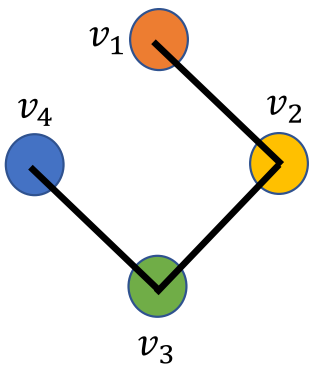
(e.g., Convolutional Neural Networks for translation invariance)

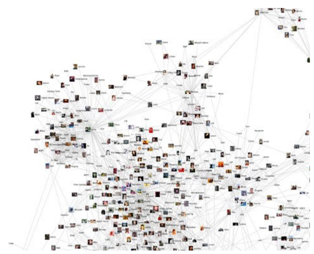
Many other examples..



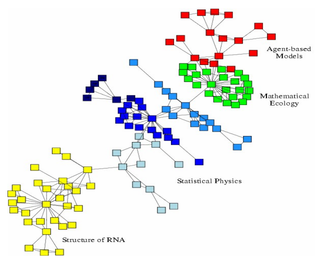
Popular model class: Convolutional Neural Networks

Invariant networks example 2: Learning on Graphs

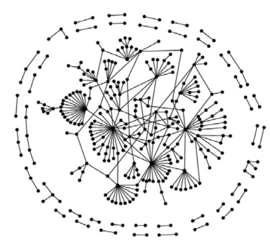




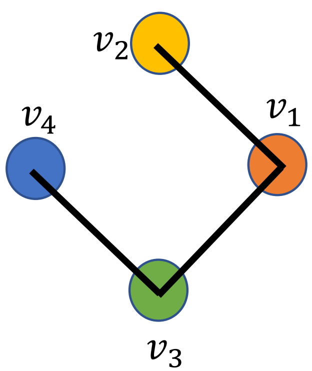
Social networks

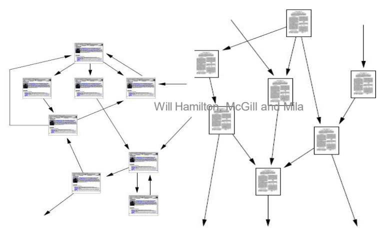


Economic networks

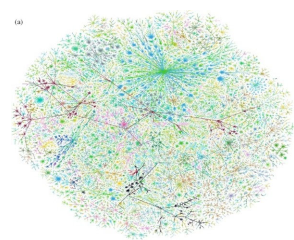


Biomedical networks

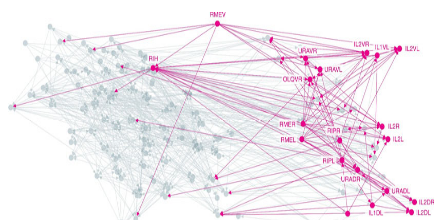




Information networks:
Web & citations



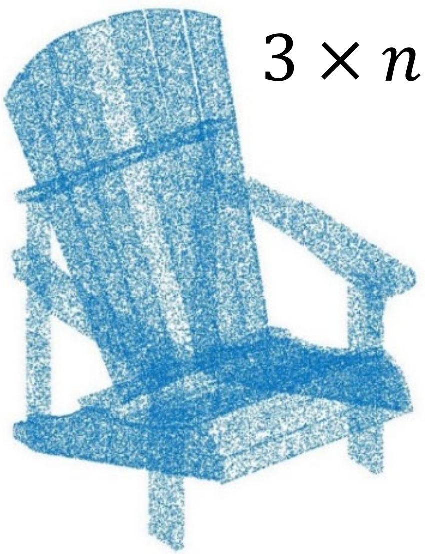
Internet



Networks of neurons

Graph Neural Networks : Graph valued functions typically invariant to node relabeling

Main example for today: point sets



$3 \times n$ points

$$X = \{x_1, x_2, \dots, x_n\}$$

$$X = (x_1, x_2, \dots, x_n) \sim \sigma_* X = (x_2, x_1, \dots x_n)$$

$$\sigma \in S_n = permutations$$

Orthogonal invariance

$$X = (x_1, x_2, \dots, x_n) \sim R_* X = (Rx_1, Rx_2, \dots Rx_n)$$

$$R \in O(d) = \{R \in \mathbb{R}^{d\times d} \mid RR^T = I_d\}$$

Special Orthogonal=Rotation invariance

$$X = (x_1, x_2, \dots, x_n) \sim R_* X = (Rx_1, Rx_2, \dots Rx_n)$$

$$R \in SO(d) = \{R \in \mathbb{R}^{d\times d} \mid RR^T = I_d, \det(R) = 1\}$$

Rotation+Permutation invariance

Point set symmetries:

Permutation S_n

Orthogonal $O(d)$

Rotation $SO(d)$

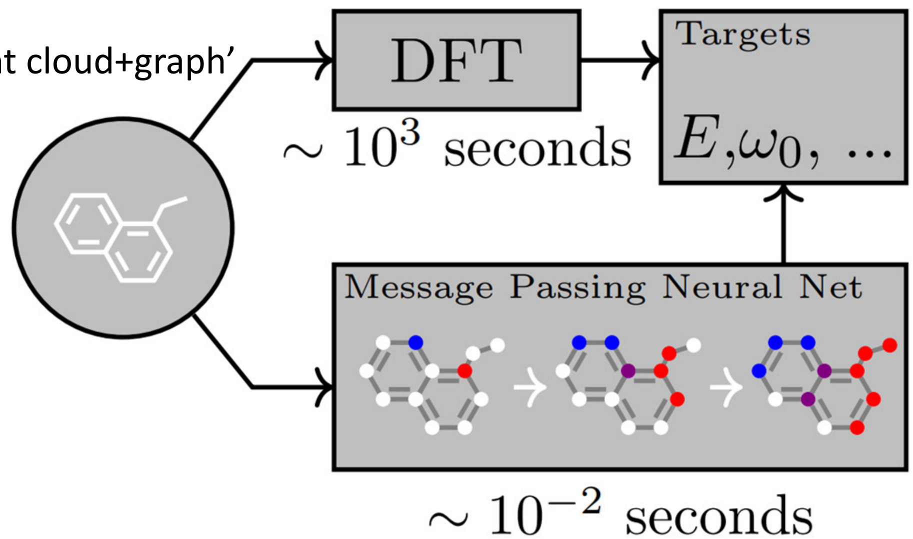
Orthogonal+Permutation

Rotation+Permutation

$$X = (x_1, x_2, \dots, x_n) \sim (R, \sigma)_*(X) = (Rx_2, Rx_1, \dots Rx_n)$$

Scientific applications (Chemistry, Physics)

Molecule=`point cloud+graph’



[Neural Message Passing for Quantum Chemistry Gilmer et al. 2017]

Symmetry preserving architectures for point sets

Point set networks (permutation invariant)
[PointNet: *Deep Learning on Point Sets for 3D Classification and Segmentation,* Qi et al. 2016]
[Deep sets, Zaheer et al. 2017]
[Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks, Lee et al. 2019]
…

PointNet/DeepSets On $\{x_1, \ldots, x_n\}$ consider **permutation invariant** functions of the form

$$\{x_1, \ldots, x_n\} \mapsto \mathcal{N}^{(2)}\left(\sum_{i=1}^{n} \mathcal{N}^{(1)}(x_i)\right)$$

Or $\{x_1, \ldots, x_n\} \mapsto \mathcal{N}^{(2)}\left(\max_i \{\mathcal{N}^{(1)}(x_i) \mid i = 1, \ldots n\}\right)$

Useful principle: Invariance cannot be \`ruined' by composition (by $\mathcal{N}^{(2)}$ in this example**)**

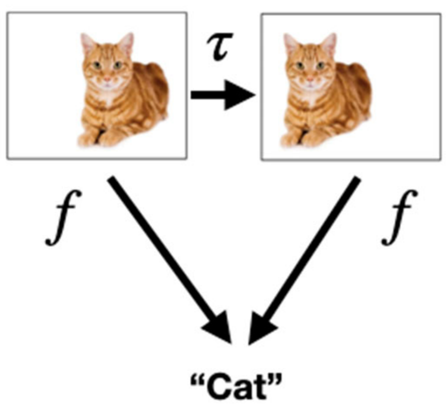
Symmetry preserving architectures for point sets 2

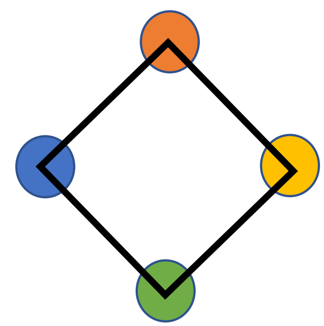
Point set networks (rotation invariant)
Not so much…

Point set networks (rotations+permutation invariant)
[Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds, Thomas et al. 2018]
[E(n) Equivariant Graph Neural Networks, Satorras et al. 2021]
[Directional Message Passing for Molecular Graphs, Gasteiger et al. 2020]
...

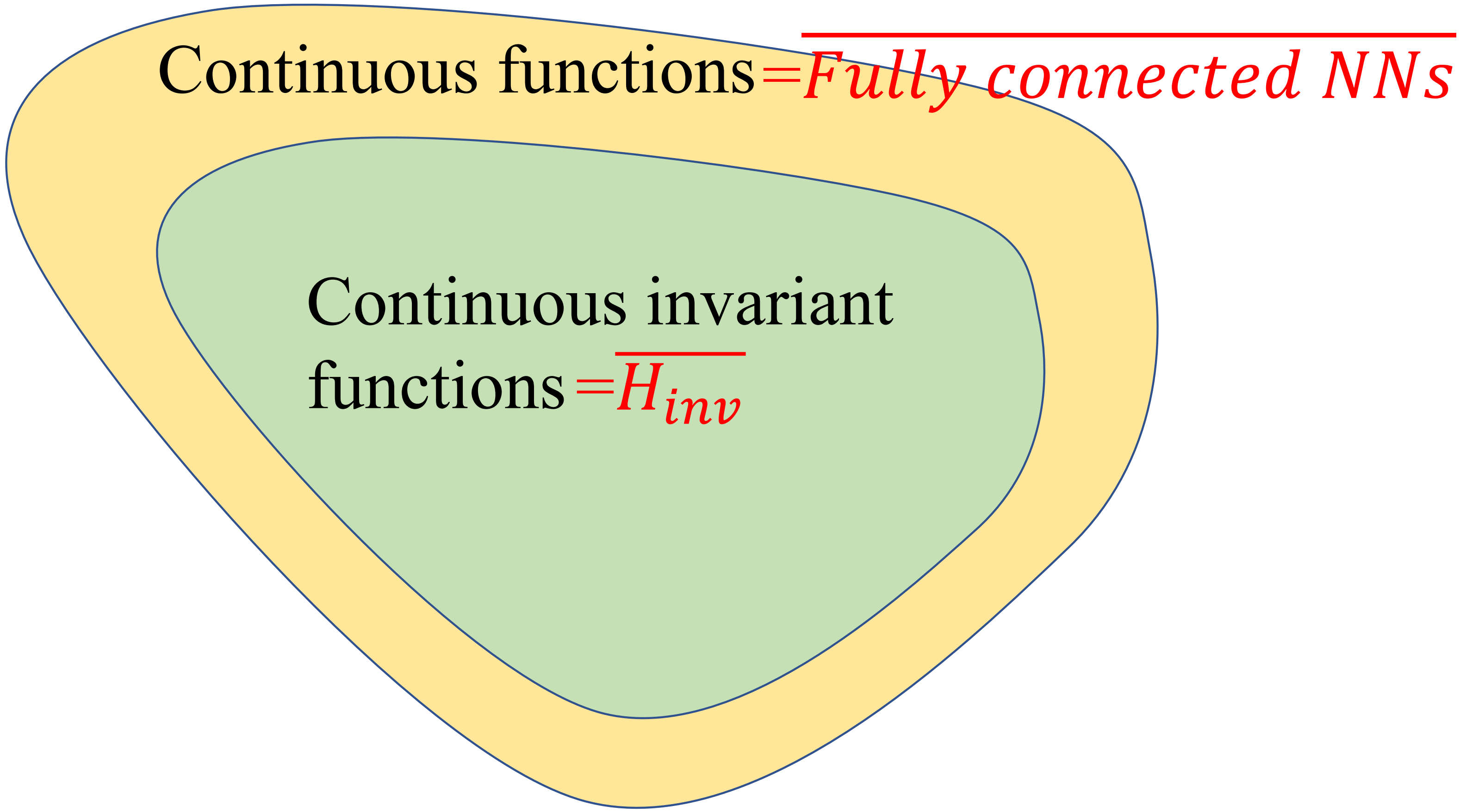
Approximation Theory of Group Invariant Neural Networks

Approximation Theory of Group Invariant Neural Networks





Universality of invariant machine learning



Example: Universality for permutation invariant point set functions

Question: Can any continuous permutation invariant $f\colon \mathbb{R}^{d\times n} \to \mathbb{R}$

$$f(x_1, \ldots, x_n) = f\left(x_{\tau(1)}, \ldots, x_{\tau(n)}\right) \text{ for every permutation } \tau$$

Be approximated by functions of the form

$$\{x_1, \ldots, x_n\} \mapsto \mathcal{N}^{(2)}\left(\sum_{i=1}^{n} \mathcal{N}^{(1)}(x_i)\right)$$

Throughout we will assume…

(G, V) are **nice**, meaning

- V is a real finite dimensional vector space

e.g., $V = \mathbb{R}^{d \times n}$

- G is a compact matrix group defined by polynomial equations

e.g., $O(d) = \{R \in \mathbb{R}^{d \times d} \,|\, RR^T = I_d\}$

- The map $(g, v) \mapsto gv$ is polynomial

e.g., $(R, X) \mapsto RX$

Standard approach: Invariant Universality via generators of the invariant ring

[Universal Approximations of Invariant Maps by Neural Networks, Yarotsky 2022]

Theorem [Hilbert, 1890]

Let (V,G) be **nice**, then there exist a finite number of invariant polynomials $F_1, \dots, F_N: V \to \mathbb{R}$ such that all invariant polynomials are of the form

$$q(v) = p\big(F_1(v), \dots, F_N(v)\big), \text{ for some } p: \mathbb{R}^N \to \mathbb{R}$$

Remark

$F_1, \dots, F_N$ are called the **generators** of the ring

$$R(V,G) = \{F: V \to \mathbb{R} \text{ are } G \text{ invariant polynomials}\}$$

Universality of invariant machine learning via generators of the invariant ring

[Universal Approximations of Invariant Maps by Neural Networks, Yarotsky 2022]

Corollary

Let (V, G) be **nice**, and $F_1, \dots F_N$ be generators of the invariant ring. Then any continuous invariant function $f: V \to \mathbb{R}$ can be approximated on compact subsets of V to arbitrary accuracy by

$$\mathcal{N}\big(F_1(v), \dots, F_N(v)\big), \text{ for some neural network } \mathcal{N}: \mathbb{R}^N \to \mathbb{R}$$

Universality of invariant machine learning via generators of the invariant ring

[Universal Approximations of Invariant Maps by Neural Networks, Yarotsky 2022]

<u>Issues</u>

- Can we explicitly compute the generators $F_1, \dots, F_N$?

  (often yes. In invariant theory this will be called `the first fundamental theorem for $(V, G)'$)

- How does N depend on $\dim(V)$?

  (**often this is very bad**… we will see examples)

- Do we want to use polynomials for approximation?

  (let's ignore this for now)

Point set `Orthogonal Universality via generators'

Group: $O(d) = \{R \in \mathbb{R}^{d\times d} \big| RR^T = I_d\}$

Action: $R_*(x_1, \dots, x_n) = (Rx_1, \dots, Rx_n)$

$\sim \boldsymbol{n^2}$ **Generators:**

$$\langle x_i, x_j \rangle \quad 1 \leq i < j \leq n$$

Universality: All continuous $O(d)$ invariant functions f can be approximated by functions of the form

$$\mathcal{N}(\langle x_1, x_1 \rangle, \langle x_1, x_2 \rangle, \dots, \langle x_n, x_n \rangle)$$

Where $\mathcal{N}$ is a (fully connected) neural network

Point set `Special Orthogonal Universality via generators'

Group: $SO(d) = \{R \in \mathbb{R}^{d\times d} \big| RR^T = I_d \;\; and \det(R) = 1\}$

Action: $R_*(x_1, \ldots, x_n) = (Rx_1, \ldots, Rx_n)$

$\sim \binom{n}{d}$ **Generators:**

$$\langle x_i, x_j \rangle, \quad 1 \le i \le j \le n \;\; \text{and} \det\left(x_{i_1}, \ldots, x_{i_d}\right) \;\; i_1 < i_2 < \cdots < i_d$$

Universality: All continuous $SO(d)$ invariant functions f can be approximated by functions of the form

$$\mathcal{N}\left(\langle x_1, x_1\rangle, \langle x_1, x_2\rangle, \ldots, \langle x_n, x_n\rangle, \det(x_1, \ldots, x_d), \ldots \det(x_{n-d+1}, \ldots, x_n)\right)$$

Where $\mathcal{N}$ is a (fully connected) neural network

Point set `Permutation Universality via generators'

Group: $\mathrm{S_n} = \{permutations\ \ \tau: \{1, \dots, n\} \to \{1, \dots, n\}\}$

Action: $\tau_*(x_1, \dots, x_n) = \left(x_{\tau^{-1}(1)}, \dots, x_{\tau^{-1}(n)}\right)$

$\boldsymbol{m(n,d)} = \binom{\boldsymbol{n+d}}{\boldsymbol{d}}$ **Generators:**

$(x_1, \dots, x_n) \mapsto \sum_{i=1}^{n} p_j(x_i)$ where $p_1, \dots, p_m$ form a basis for the space of polynomials of degree $\leq n$ in d variables

Universality: All continuous S_n invariant functions can be approximated by

$$\mathcal{N}\left(\sum_{i=1}^{n} p_1(x_i), \sum_{i=1}^{n} p_2(x_i), \dots \sum_{i=1}^{n} p_m(x_n)\right)$$

$$\text{Or } \mathcal{N}^{(2)}\left(\sum_{i=1}^{n} \mathcal{N}^{(1)}(x_i)\right)$$

Number of generators for point set actions

Group action on $\mathbb{R}^{d\times n}$	Num of generators
$O(d)$	$\sim n^2$
$SO(d)$	$\sim \binom{n}{d}$
S_n	$\binom{n+d}{d}$

Universality of invariant machine learning via ~~generating~~ *separating* invariants

Advocates: [Complete set of translation invariant measurements with Lipschitz bounds, Cahill et al. 2020]
[Group invariant max-filtering, Cahill et al. 2022]
[Low Dimensional Invariant Embeddings for Universal Geometric Learning, **Dym** and Gortler 2022]

Definition (Separating invariants)

Let G be a group acting on V. We say that $H_1, \dots, H_m : V \to \mathbb{R}$ are (V, G) separating invariants if

- **Invariant: if** $u =_G v$ **then** $H_i(u) = H_i(v), \forall i = 1, \dots, m$
- **Separating: if** $H_i(v) = H_i(u), \forall i = 1, \dots, m$ **then** $v =_G u$

Invariance means that $V/_G \ni [v] \mapsto (H_1(v), \dots, H_m(v))$ is well defined

Separating means that it is injective on $V/_G$

Example: $G = O(2)$ acts on $V = \mathbb{R}^{2\times 2}$ via $R_*(x_1, x_2) = (Rx_1, Rx_2)$

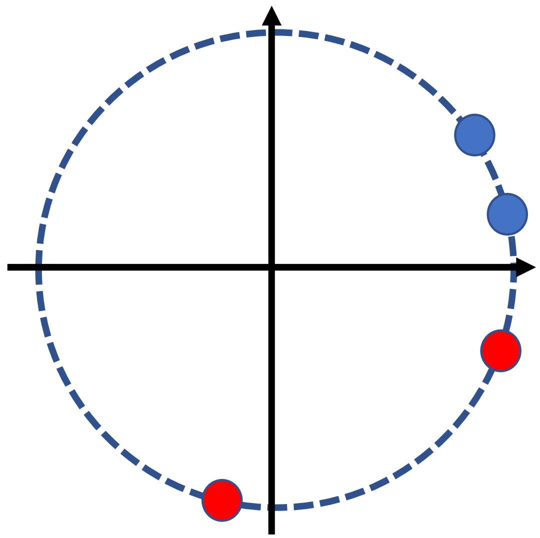
What invariants can we suggest? Are they separating?

How about:

$H_1(x_1, x_2) = \|x_1\|$ and $H_2(x_1, x_2) = \|x_2\|$?

We get separation by adding

$H_3(x_1, x_2) = \|x_1 - x_2\|$



Separation vs generation: sufficiency for universality

We saw

and $H_1, \dots, H_m$ be continuous separating invariants

Let (V, G) be **nice**, and ~~$F_1, \dots F_N$ be generators of the invariant ring~~. Then any continuous invariant function $f: V \to \mathbb{R}$ can be approximated on compact subsets of V to arbitrary accuracy by

$$\mathcal{N}\left(F_1(v), \dots, F_N(v)\right), \text{ for some neural network } \mathcal{N}: \mathbb{R}^N \to \mathbb{R}$$

$$\mathcal{N}(H_1(v), \dots H_m(v))$$

Remark: This in fact implies the generator-based theorem, since **generators are always separators**

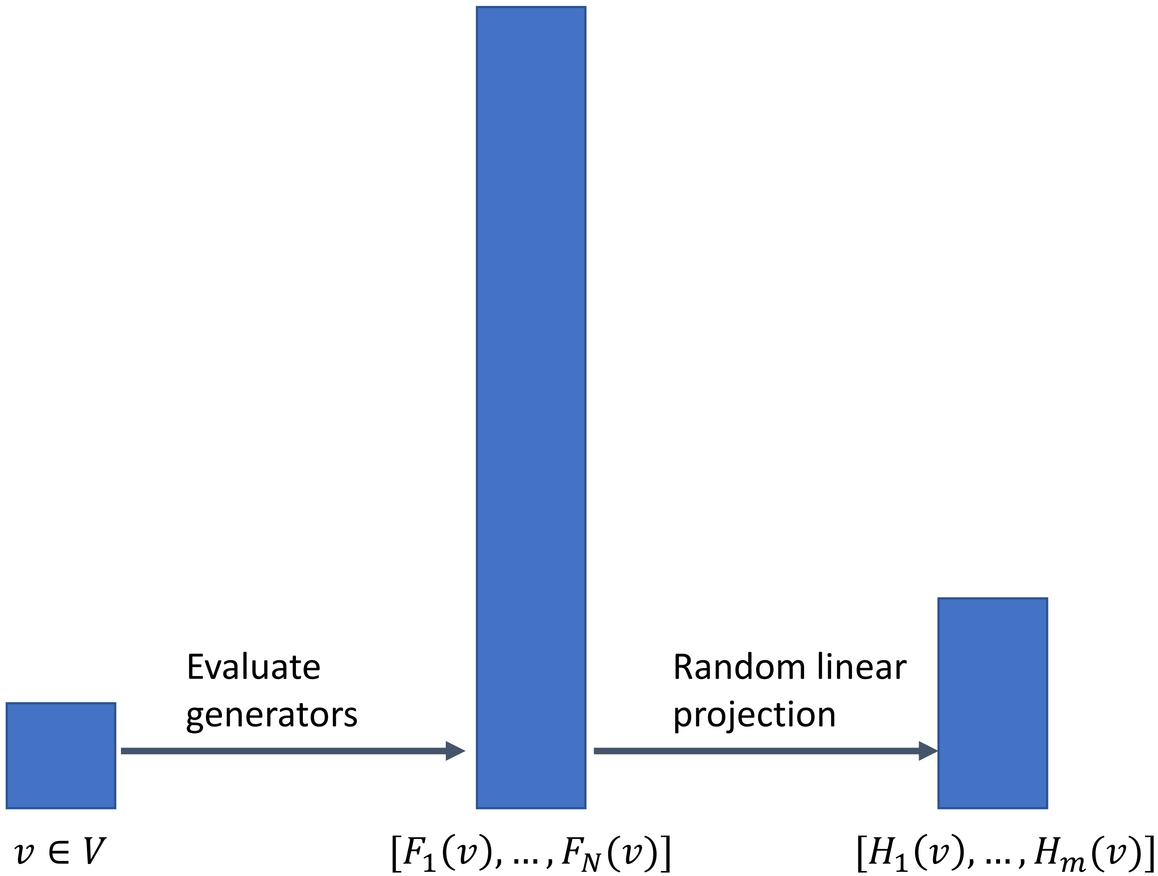
Separation vs generation: cardinality

Theorem [E. S. Dufresne 2008]

If (V, G) are **nice**, then there always exist **polynomial** separating invariants $H_1, \dots, H_m: V \to \mathbb{R}$ of cardinality

$$m = 2\dim(V) + 1$$

Partial solution: low dimensional-separation via generation+`linear compression’



Intermediate conclusions

Group action on $\mathbb{R}^{d\times n}$	Num of generators
$O(d)$	n^2
$SO(d)$	$n^2 + \binom{n}{d}$
S_n	$\binom{n+d}{d}$

Let's start here

Can we do better? Yes

Efficient invariants: $SO(d)$

Example: R $\in SO(d)$ acts on $\mathbf{X} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_n}) \in \mathbb{R}^{d\times n}$ $(d < n)$

$$R_*(\boldsymbol{x_1}, \dots \boldsymbol{x_n}) = (R\boldsymbol{x_1}, \dots, R\boldsymbol{x_n})$$

Generators: $\sim \binom{n}{d}$

$$\left|\boldsymbol{x_i} - \boldsymbol{x_j}\right|^2 \; and \; \left|\boldsymbol{x_j}\right|^2 \text{ and } \det\left(\boldsymbol{x_{i_1}}, \dots \boldsymbol{x_{i_d}}\right)$$

Continuous family of separating invariants:

$$H(\boldsymbol{x_1}, \dots, \boldsymbol{x_n}; \boldsymbol{w}, \boldsymbol{W}) = |\boldsymbol{w_1 x_1} + \cdots + \boldsymbol{w_n x_n}|^2 + \det(\boldsymbol{XW})$$

Random separators: For almost all $\boldsymbol{w}^{(1)}, \boldsymbol{W}^{(1)} \dots, \boldsymbol{w}^{(m)}, \boldsymbol{W}^{(m)}$ $, m = 2nd + 1$

$H\left(\boldsymbol{x_1}, \dots, \boldsymbol{x_n}; \boldsymbol{w}^{(i)}, \boldsymbol{W}^{(i)}\right)$ are invariant and separating!!!

$|\boldsymbol{x}_i - \boldsymbol{x}_j|^2$ *and* $|\boldsymbol{x}_j|^2$ and $\det(\boldsymbol{x}_{i_1}, \ldots \boldsymbol{x}_{i_d})$

Group action on $\mathbb{R}^{d \times n}$	Num of generators	Num of separators	Complexity per separator?
$O(d)$	n^2	$2n \cdot d + 1$	
$SO(d)$	$n^2 + \binom{n}{d}$	$2n \cdot d + 1$	nd^2
S_n	$\binom{n+d}{d}$	$2n \cdot d + 1$	

$|\boldsymbol{w}_1 \boldsymbol{x}_1 + \cdots + \boldsymbol{w}_n \boldsymbol{x}_n|^2 + \det(\boldsymbol{X}\boldsymbol{W})$

Efficient Invariants: SO(d) and beyond

For the action of SO(d) on $\mathbb{R}^{\mathrm{d\times n}}$, the following is a *continuous family of separating invariants*

$$H(\mathbf{x_1}, \dots, \mathbf{x_n}; \mathbf{w}, \mathbf{W}) = |\mathbf{w_1 x_1} + \cdots + \mathbf{w_n x_n}|^2 + \det(\mathbf{XW})$$

Definition: Let (V, G) be **nice**. We sat that a function $H: V \times \mathbb{R}^{d_W} \to \mathbb{R}$ is a *continuous family of separating invariants* if it satisfies the following conditions:

- **Invariance: If** $v =_G v'$ then $H(v; w) = H(v'; w)$ for all $w \in \mathbb{R}^{d_W}$
- **Separation:** If $v \neq_G v'$ then there exists $w \in \mathbb{R}^{d_W}$ such that $H(v; w) \neq H(v'; w)$

Finite Witness Theorem

Finite Witness Theorem [Dym and Gortler 2022] (weakened version):

Let (V, G) be **nice**. Let $H: V \times \mathbb{R}^{d_W} \to \mathbb{R}$ be a family of separating **polynomial** invariants.

Set $m = 2\dim(V) + 1$. Then for Lebesgue almost every $w^{(1)}, \dots, w^{(m)} \in \mathbb{R}^{d_W}$, the functions $H_1, \dots H_m$ defined by

$$H_i(v) = H\left(v; w^{(i)}\right)$$

are separating invariants.

Remarks

- Cardinality is often not optimal
- Proof idea comes from [On signal reconstruction without phase, Balan, Casazza and Edidin 2006] relies on Real Algebraic Geometry

Finite Witness Theorem [Dym and Gortler 2022] (weakened version):

Let (V, G) be **nice**. Let $H: V \times \mathbb{R}^{d_W} \to \mathbb{R}$ be a family of separating **polynomial** invariants.

Set $m = 2\dim(V) + 1$. Then for Lebesgue almost every $w^{(1)}, \dots, w^{(m)} \in \mathbb{R}^{d_w}$, the functions $H_1, \dots H_m$ defined by

$$H_i(v) = H\left(v; w^{(i)}\right)$$

are separating invariants.

Proof idea:

- Consider the **`lifted bad set'**

$$\boldsymbol{B} = \left\{\left(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{w}^{(1)}, \dots \boldsymbol{w}^{(m)}\right) \in \boldsymbol{V} \times \boldsymbol{V} \times \mathbb{R}^{d_w \times m} \middle| \boldsymbol{v} \neq_G \boldsymbol{v}' \;\; \boldsymbol{but} \;\; \boldsymbol{H}\left(\boldsymbol{v}; \boldsymbol{w}^{(i)}\right) = \boldsymbol{H}\left(\boldsymbol{v}'; \boldsymbol{w}^{(i)}\right), \forall \boldsymbol{i} = \mathbf{1} \dots \boldsymbol{m}\right\}$$

- This set is a subset of a $\mathbf{2}\dim(\boldsymbol{V}) + \boldsymbol{md_w}$ dimensional vector space defined by $\boldsymbol{m}$ equations

“therefore” $\dim(\boldsymbol{B}) = \boldsymbol{md_w} + (\mathbf{2}\dim(\boldsymbol{V}) - \boldsymbol{m}) = \boldsymbol{md_w} - \mathbf{1}$

- The dimension of the `projected bad set' is no larger

$$\boldsymbol{B_{proj}} = \left\{\left(\boldsymbol{w}^{(1)}, \dots \boldsymbol{w}^{(m)}\right) \in \mathbb{R}^{d_w \times m} \middle| \exists (\boldsymbol{v}, \boldsymbol{v}') \;\boldsymbol{s.t.}\; \boldsymbol{v} \neq_G \boldsymbol{v}' \;\; \boldsymbol{but} \;\; \boldsymbol{H}\left(\boldsymbol{v}; \boldsymbol{w}^{(i)}\right) = \boldsymbol{H}\left(\boldsymbol{v}'; \boldsymbol{w}^{(i)}\right), \forall \boldsymbol{i} = \mathbf{1} \dots \boldsymbol{m}\right\}$$

- $\boldsymbol{dim}\left(\boldsymbol{B_{proj}}\right) = \boldsymbol{md_w} - \mathbf{1} < \boldsymbol{dim}(\mathbb{R}^{d_w \times m})$
- Most $\left(\boldsymbol{w}^{(1)}, \dots \boldsymbol{w}^{(m)}\right)$ are not in $\boldsymbol{B_{proj}}$, and so are separating

Finite Witness Theorem [Dym and Gortler 2022] (weakened version):

Let (V, G) be **nice**. Let $H: V \times \mathbb{R}^{d_W} \to \mathbb{R}$ be a continuous family of separating polynomial invariants.

Set $m = 2\dim(V) + 1$. Then for Lebesgue almost every $w^{(1)}, \dots, w^{(m)} \in \mathbb{R}^{d_w}$, the functions $H_1, \dots H_m$ defined by

$$H_i(v) = H(v; w^{(i)})$$

are separating invariants.

Real algebraic geometry

Full Proof

Proof ~~idea~~ (inspired by phase retrieval paper):

- Consider the **`lifted bad set'**

$$\boldsymbol{B = \{(v, v', w^{(1)}, \dots w^{(m)}) \in V \times V \times \mathbb{R}^{d_w \times m} | v \neq_G v' \ \ but \ \ H(v; w^{(i)}) = H(v'; w^{(i)}), \forall i = 1 \dots m\}}$$

- This set is a subset of a $\mathbf{2\dim(V)} + \boldsymbol{md_w}$ dimensional vector space defined by $\boldsymbol{m}$ equations

"therefore" $\mathbf{dim}(\boldsymbol{B}) = \boldsymbol{md_w} + (\mathbf{2\,dim}(\boldsymbol{V}) - \boldsymbol{m}) = \boldsymbol{md_w} - \mathbf{1}$

- The dimension of the `projected bad set' is no larger

$$\boldsymbol{B_{proj} = \{(w^{(1)}, \dots w^{(m)}) \in \mathbb{R}^{d_w \times m} | \exists (v, v') \ s.t. \, v \neq_G v' \ \ but \ \ H(v; w^{(i)}) = H(v'; w^{(i)}), \forall i = 1 \dots m\}}$$

- $\boldsymbol{dim(B_{proj}) = md_w - 1 < dim(\mathbb{R}^{d_w \times m})}$
- Most $(\boldsymbol{w^{(1)}, \dots w^{(m)}})$ are not in $\boldsymbol{B_{proj}}$, and so are separating

Finite Witness Theorem-Applications

Group action on $\mathbb{R}^{d\times n}$	Num of generators	Num of separators	Complexity per separator?
$O(d)$	n^2	$2n \cdot d + 1$	$n \cdot d$
$SO(d)$	$\mathrm{n}^2 + \binom{n}{d}$	$2n \cdot d + 1$	$n \cdot d^2$
S_n	$\binom{n+d}{d}$	$2n \cdot d + 1$	$n \cdot log(n)$

Recent work- Analytic Finite Witness Theorem

Analtyic Finite Witness Theorem [Amir, Gortler, Avni, Ravina, Dym 2023] (weakened version): **Analytic**

Let (V, G) be **nice**. Let $H: V \times \mathbb{R}^{d_w} \to \mathbb{R}$ be a continuous family of separating ~~**polynomial**~~ invariants.

Set $m = 2\dim(V) + 1$. Then for Lebesgue almost every $w^{(1)}, \ldots, w^{(m)} \in \mathbb{R}^{d_w}$, the functions $H_1, \ldots H_m$ defined by

$$H_i(v) = H\left(v; w^{(i)}\right)$$

are separating invariants.

\`Proof'

Real Algebraic Geometry $\mapsto$ Real analytic geometry, o-minimal systems and related concepts

Application: Permutation invariant networks (with analytic activations)

Theorem [Amir, Gortler, Avni, Ravina, Dym 2023]

Let d, n be natural numbers and set $m = 2nd + 1$.

If $\sigma: \mathbb{R} \mapsto \mathbb{R}$ is **analytic** and not polynomial, then for Lebesgue almost every $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$

the permutation invariant function

$$\mathbb{R}^{d \times n} \ni (x_1, \ldots, x_n) \mapsto \sum_{i=1}^{n} \sigma(Ax_i + b)$$

is separating

Finite Witness Theorem

Analtyic Finite Witness Theorem-stronger (but not strongest) version

Let (V, G) be **nice**. Let $H: V \times \mathbb{R}^{d_w} \to \mathbb{R}$ be a continuous family of separating **analytic** invariants. Set $m = 2\dim(V) + 1$. Then for Lebesgue almost every $w^{(1)}, \dots, w^{(m)} \in \mathbb{R}^{d_w}$, the functions $H_1, \dots H_m$ defined by

$$H_i(v) = H\left(v; w^{(i)}\right)$$

are separating invariants.

Can be a low dimensional subset of some higher dimensional vector space, providing it is `reasonable'
e.g., a countable union of sets defined by polynomial and analytic equalities and inequalities
Or image of these sets under an analytic functions

Adding to the table…

Group action on $\mathbb{R}^{d\times n}$	Num of generators	Num of separators	Complexity per separator?
$O(d)$	n^2	$2n\cdot d+1$	$n\cdot d$
$SO(d)$	$\mathrm{n}^2+\binom{n}{d}$	$2n\cdot d+1$	$n\cdot d^2$
S_n	$\binom{n+d}{d}$	$2n\cdot d+1$	$n(d+\log(n))$
$O(d)\times S_n$	?	$2n\cdot d+1$	$\boldsymbol{n^d}$
$SO(d)\times S_n$	?	$2n\cdot d+1$	$\boldsymbol{n^d}$

Parting questions

- Separating invariants are injective mappings $f: V/G \to R^m$. Do they preserve distances?
- Separating invariants for surfaces? (one example: conformal welding)

Funding: This research was supported by the Israeli Science Foundation grant no. 272/23
N.D. is a Horev Fellow

Collaborators

Tal Amir

Snir Hordan

Ilai Avni

Ravina Ravina

Technion

Steven J. Gortler
Harvard

[Neural Injective Functions for Multisets, Measures and Graphs
via a Finite Witness Theorem.
Amir, Gortler, Avni, Ravina and Dym 2023]

Thank you!

[Low Dimensional Invariant Embeddings for Universal
Geometric Learning
Dym and Gortler 2022]

[Complete Neural Networks for Euclidean Graphs
Hordan, Amir, Gortler, and Dym 2023]

$O(d) \times S_n$ separation

<u>Generically separating</u>
Geometric Message Passing

<u>Fully separating</u>
$(d-1)$ order geometric message passing

Geometric message passing

e.g. EGNN [E(n) equivariant graph neural networks, Sattoras et al. 2021]

Input: $x_1, \ldots, x_n \in \mathbb{R}^d$

set $h_1^{(0)}, \ldots, h_n^{(0)} = 0$

$h_i^{(t)} = f_{agg}\left(h_i^{(t-1)}, \left\{h_j^{(t-1)}, |x_i - x_j|, j = 1, \ldots n\right\}\right)$ (repeat T times)

$h_{global}(x_1, \ldots, x_n) = f_{readout}(\{h_1^{(T)}, \ldots, h_n^{(T)}\})$

$h_{global}(x_1, \ldots, x_n)$ is $O(d) \times S_n$ invariant

Geometric message passing-separation

e.g. EGNN [E(n) equivariant graph neural networks, Sattoras et al. 2021]

Input: $x_1, \dots, x_n \in \mathbb{R}^d$

set $h_1^{(0)}, \dots, h_n^{(0)} = 0$

$h_i^{(t)} = f_{agg}\left(h_i^{(t-1)}, \left\{h_j^{(t-1)}, |x_i - x_j|, j = 1, \dots n\right\}\right)$ (repeat T times)

$h_{global} = f_{readout}(\{h_1^{(T)}, \dots, h_n^{(T)}\})$

Permutation invariant and separating

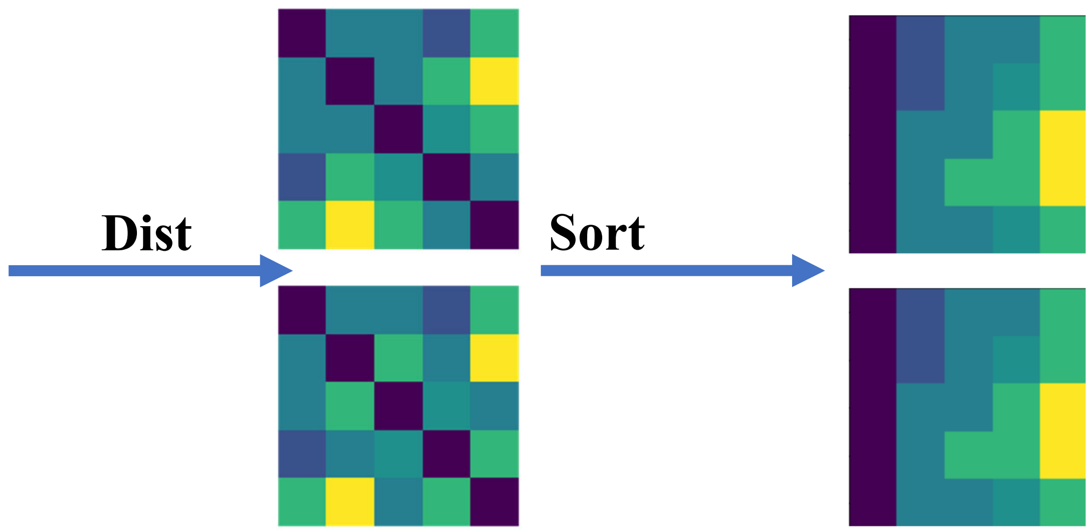
`Hard' to separate

[Incompleteness of Atomic Structure Representations. Physical Review Letters Pozdynakov et al. 2020]

0	1	2	-2	-1
1	1	0	0	-1
1	0	2	-2	0

$\neq_G$

0	1	2	-2	-1
1	1	0	0	-1
-1	0	2	-2	0

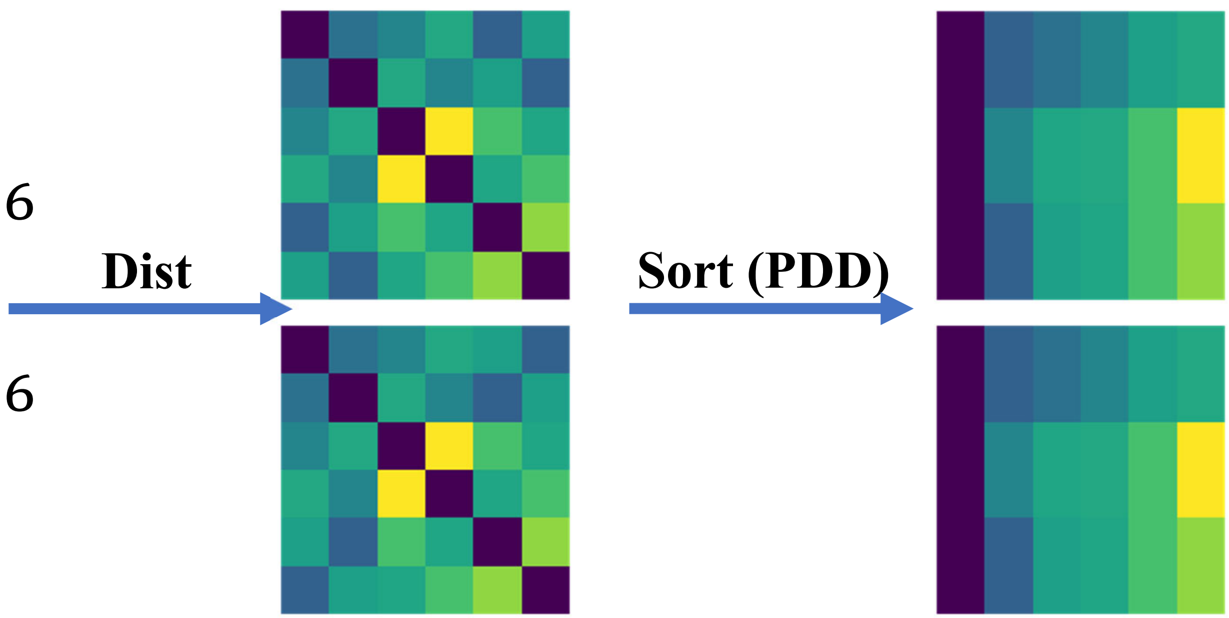


- **Cannot** be separated by MPNN with $T = 1$
- **Can** be separated by MPNN with $T \geq 2$

‘Harder’

[Incompleteness of graph neural networks
for points clouds in three dimensions, Pozdnyakov and Ceriotti 2022]

$$X_+ \in \mathbb{R}^{3\times 6}$$
$$\neq_G$$
$$X_- \in \mathbb{R}^{3\times 6}$$



Cannot be separated by MPNN for any T

Geometric K-order message passing

[Sign and Basis Invariant Networks for Spectral Graph Representation Learning, Lim et al. 2022]
[Is distance matrix enough for geometric deep learning, Li et al. 2023]

Assume K= 3 for notation simplicity

$$h^{(0)}(i,j,k)(X) = \begin{pmatrix} \langle x_i, x_i\rangle & \langle x_i, x_j\rangle & \langle x_i, x_k\rangle \\ \langle x_j, x_i\rangle & \langle x_j, x_j\rangle & \langle x_j, x_k\rangle \\ \langle x_k, x_i\rangle & \langle x_k, x_j\rangle & \langle x_k, x_k\rangle \end{pmatrix}$$

$$h^{(t)}(i,j,k)(X) = f_{agg}\left(h^{(t-1)}(i,j,k), \left\{ \begin{pmatrix} h^{(t-1)}(s,j,k) \\ h^{(t-1)}(i,s,k) \\ h^{(t-1)}(i,j,s) \end{pmatrix}, \quad s = 1, \dots, n \right\} \right)$$

$$h_{global} = f_{readout}\left\{ h^{(T)}(i,j,k) \middle| (i,j,k) \in [n]^3 \right\}$$

Permutation invariant
+separating

Theorem [Hordan, Amir, Gortler, Dym, 2023]

For every $X, Y \in \mathbb{R}^{d \times n}$ we have that the d-order message passing with $T = 1$ is separating:

It gives the same output $h_{global}(X) = h_{global}(Y)$ if and only if X, Y are related by a permutation and orthogonal transformation.

A modified $d - 1$ message passing algorithm is also separating

Theorem [Rose et al. 2023]

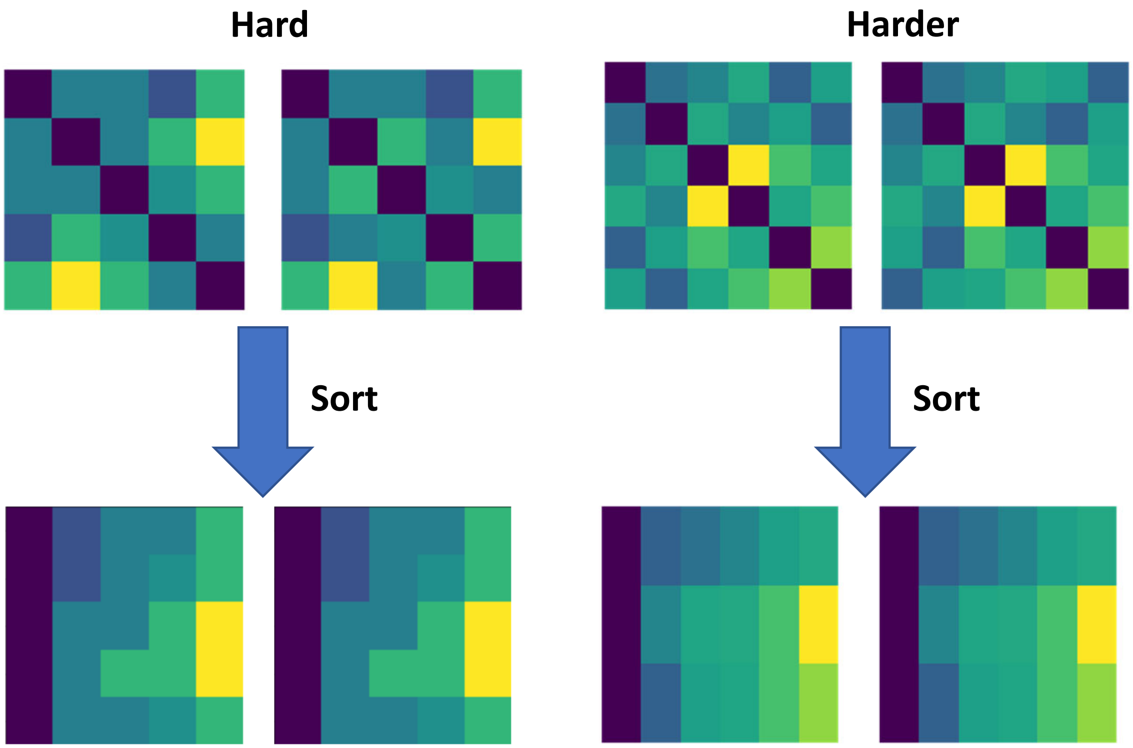
The original $d - 1$ message passing algorithm is also separating

Complexity

- Full $O(d) \times S_n$ separation with $(d-1) - WL$ requires computing $2nd + 1$ invariants with computational complexity of n^d each, using our permutation invariant separating functions

- This also uses the dependence of the theorem on intrinsic dimension. Considering extrinsic dimension only would lead to exponential blowup

Separation experiment



	Hard	Harder
MPNN	Yes	No
$(d-1)$ MPNN	Yes	Yes

Separation of existing invariant architectures:

$O(d) \times S_n$ Invariant architectures

$(d-1)MPNN$ $MPNN$

Point Clouds	GramNet	GeoEGNN	EGNN	LinearEGNN	MACE	TFN	DimeNet	GVPGNN
Hard1[2]	1.0	0.998	0.5	1.0	1.0	0.5	1.0	1.0
Hard2 [2]	1.0	0.97	0.5	1.0	1.0	0.5	1.0	1.0
Hard3 [2]	1.0	0.85	0.5	1.0	1.0	0.55	1.0	1.0
Harder [1]	1.0	0.899	0.5	0.5	1.0	0.5	1.0	1.0
Cholesky dim=6	1.0	Irrelevant	0.5	0.5	1.0	Irrelevant	Irrelevant	Irrelevant
Cholesky dim=8	1.0	Irrelevant	0.5	0.5	1.0	Irrelevant	Irrelevant	Irrelevant
Cholesky dim=12	N/A	Irrelevant	0.5	0.5	0.5	Irrelevant	Irrelevant	Irrelevant

Dataset composed of two point clouds which are hard to separate+rotations+permutations+noise

We didn't discuss…

- **Generic separation:** Separation up to a set of measure zero. Need only $\dim(V) + 1$ invariants
- **Stability:** Invariant and separating $H: V \to \mathbb{R}^m$ can be identified with $H: {}^V/_G \to \mathbb{R}^m$ injective.

  Is H bi-Lipschitz with respect to

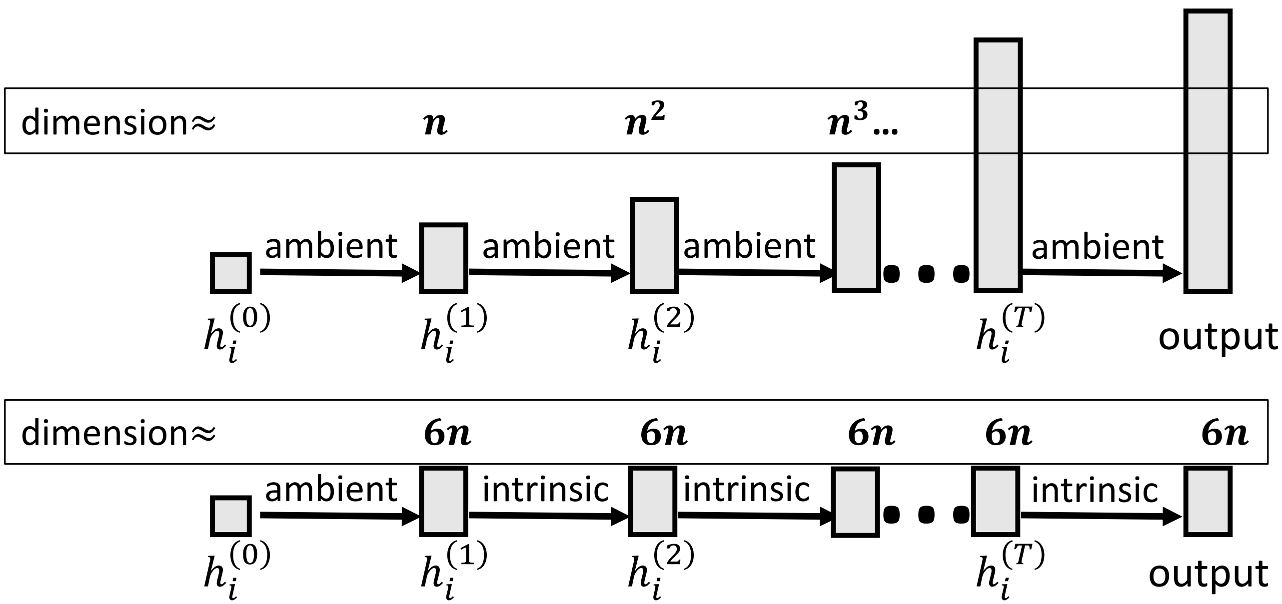
$$d([v],[v']) = \min_{g \in G} ||gv - v'||$$

[Permutation invariant representations with applications to graph deep learning, Balan Haghani Singh 2022]
[Group-invariant max flitering, Cahill Iverson Dixon and Packer]

TODO

$$h_i^{(t)} = f_{agg}\left(h_i^{(t-1)}, \left\{h_j^{(t-1)}, |x_i - x_j|, j = 1, \ldots n\right\}\right) \text{ (repeat } T \text{ times)}$$

Permutation invariant and separating



Separation of existing architectures: (when) does it happen?

$O(d) \times S_n$ Invariant architectures

Theoretical separation	Yes (ours)	Yes (ours)	No	No	?	Sort of	?	?
Point Clouds	GramNet	GeoEGNN	EGNN	LinearEGNN	MACE	TFN	DimeNet	GVPGNN
Hard1[2]	1.0	0.998	0.5	1.0	1.0	0.5	1.0	1.0
Hard2 [2]	1.0	0.97	0.5	1.0	1.0	0.5	1.0	1.0
Hard3 [2]	1.0	0.85	0.5	1.0	1.0	0.55	1.0	1.0
Harder [1]	1.0	0.899	0.5	0.5	1.0	0.5	1.0	1.0
Cholesky dim=6	1.0	Irrelevant	0.5	0.5	1.0	Irrelevant	Irrelevant	Irrelevant
Cholesky dim=8	1.0	Irrelevant	0.5	0.5	1.0	Irrelevant	Irrelevant	Irrelevant
Cholesky dim=12	N/A	Irrelevant	0.5	0.5	0.5	Irrelevant	Irrelevant	Irrelevant

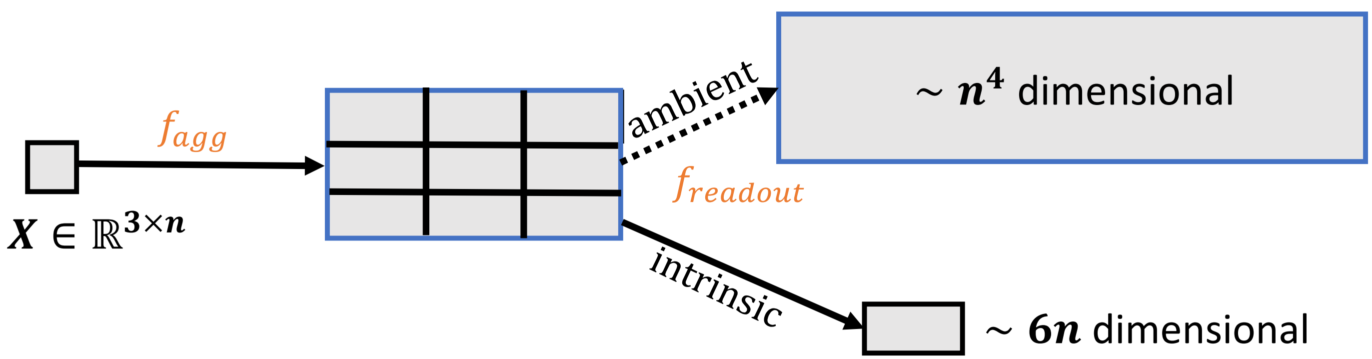
Dataset composed of two point clouds which are hard to separate+rotations+permutations+noise

Proof of theorem (intuition)

Full $O(d) \times S_n$ separation (1): Cardinality

$$h^{(1)}(i,j,k)(X) = f_{agg}\left(h^{(0)}(i,j,k), \left\{\begin{pmatrix} h^{(0)}(s,j,k) \\ h^{(0)}(i,s,k) \\ h^{(0)}(i,j,s) \end{pmatrix}, \quad s = 1, \dots, n\right\}\right)$$

$$h_{global} = f_{readout}\left\{h^{(1)}(i,j,k) \middle| (i,j,k) \in [n]^3\right\}$$



Phase retrieval

Better solution: imported from phase retrieval

Phase retrieval: we want to reconstruct a signal $\boldsymbol{z} \in \mathbb{C}^n$ from phaseless linear measurements

$$H_i(\boldsymbol{z}) = \left|\left\langle \boldsymbol{w}^{(i)}, \boldsymbol{z} \right\rangle\right|^2, i = 1, \ldots m$$

$\boldsymbol{S^1}$ **invariance:** For all θ we have that $\left|H_i\left(e^{i\theta}\boldsymbol{z}\right)\right| = |H_i(\boldsymbol{z})|$ so we can only hope for reconstruction up to a global phase factor, that is

$$H_i(\boldsymbol{z}) = H_i(\hat{\boldsymbol{z}}) \quad \Longleftrightarrow \quad \boldsymbol{z} = e^{i\theta}\hat{\boldsymbol{z}} \text{ for some } \theta$$

In other words, we would like $H_1, \ldots H_m$ to be separating

Better solution: imported from phase retrieval

Theorem [On signal reconstruction without phase, Balan, Casazza and Edidin 2006]

If $m = 4n - 2$ then for Lebesgue almost all $w^{(1)}, \ldots, w^{(m)} \in \mathbb{R}^n$ the functions $H_1, \ldots H_m$ defined by

$$H_i(\boldsymbol{z}) = \left|\left\langle \boldsymbol{w}^{(i)}, \boldsymbol{z} \right\rangle\right|^2, i = 1, \ldots m$$

are separating with respect to the action of S^1on $\mathbb{C}^n$

Remark: In our context we think of $V = \mathbb{C}^n$ is a real vector space of dimension $2n$. So

$$m = 4n - 2 < 2\dim(V) + 1 = 4n + 1$$

Remark: Note that all invariant are obtained by taking sample of $H(\boldsymbol{z}; \boldsymbol{w})$ which is polynomial in both its argument $\boldsymbol{z}$ and its parameters $\boldsymbol{w}$

Separation vs. generation for phase retrieval

Theorem [On signal reconstruction without phase, Balan, Casazza and Edidin 2006]

If $m = 4n - 2$ then for Lebesgue almost all $w^{(1)}, \ldots, w^{(m)} \in \mathbb{R}^n$ the functions $H_1, \ldots H_m$ defined by

$$H_i(\boldsymbol{z}) = \left|\left\langle \boldsymbol{w}^{(i)}, \boldsymbol{z} \right\rangle\right|^2, i = 1, \ldots m$$

are separating with respect to the action of S^1on $\mathbb{C}^n$

In contrast, there are $\sim n^2$ generators for the ring of invariant polynomials:

$$H_{s,t}(z_1, \ldots, z_n) = z_s \overline{z_t}$$

Invariant universality rephrased

Assume $\boldsymbol{G}$ acts on $\boldsymbol{V}$

Orbit: $[v] = \{w \in V | \exists g \in G,\ w = gv\ \}$

Quotient space: $V/_G = \{[v]|\ v \in V\}$

If $\boldsymbol{f}: \boldsymbol{V} \to \boldsymbol{Y}$ is invariant then it induces a well-defined $\hat{\boldsymbol{f}}: \boldsymbol{V}/_{\boldsymbol{G}} \to \boldsymbol{Y}$ via

$$\hat{f}([v]) = f(v)$$

Invariant universality via Invariant embeddings

If $\widehat{\boldsymbol{F}}: {}^{\boldsymbol{V}}/_{\boldsymbol{G}} \to \mathbb{R}^{\boldsymbol{m}}$ is invariant and *injective, then any* $\widehat{\boldsymbol{f}}: {}^{\boldsymbol{V}}/_{\boldsymbol{G}} \to \boldsymbol{Y}$ is of the form

$\widehat{\boldsymbol{f}}([\boldsymbol{v}]) = \boldsymbol{h} \circ \widehat{\boldsymbol{F}}([\boldsymbol{v}])$, for an appropriate $\boldsymbol{h}: \mathbb{R}^{\boldsymbol{m}} \to \boldsymbol{Y}$

On the image of $\widehat{\boldsymbol{F}}$ we have $\boldsymbol{h} = \widehat{\boldsymbol{f}} \circ \left(\widehat{\boldsymbol{F}}\right)^{-\mathbf{1}}$

Goal: Find injective $\widehat{\boldsymbol{F}}: {}^{\boldsymbol{V}}/_{\boldsymbol{G}} \to \mathbb{R}^{\boldsymbol{m}}$

Invariant embeddings and separating invariants

Goal: Find injective $\widehat{\boldsymbol{F}}: {}^{\boldsymbol{V}}/_{\boldsymbol{G}} \to \mathbb{R}^{\boldsymbol{m}}$

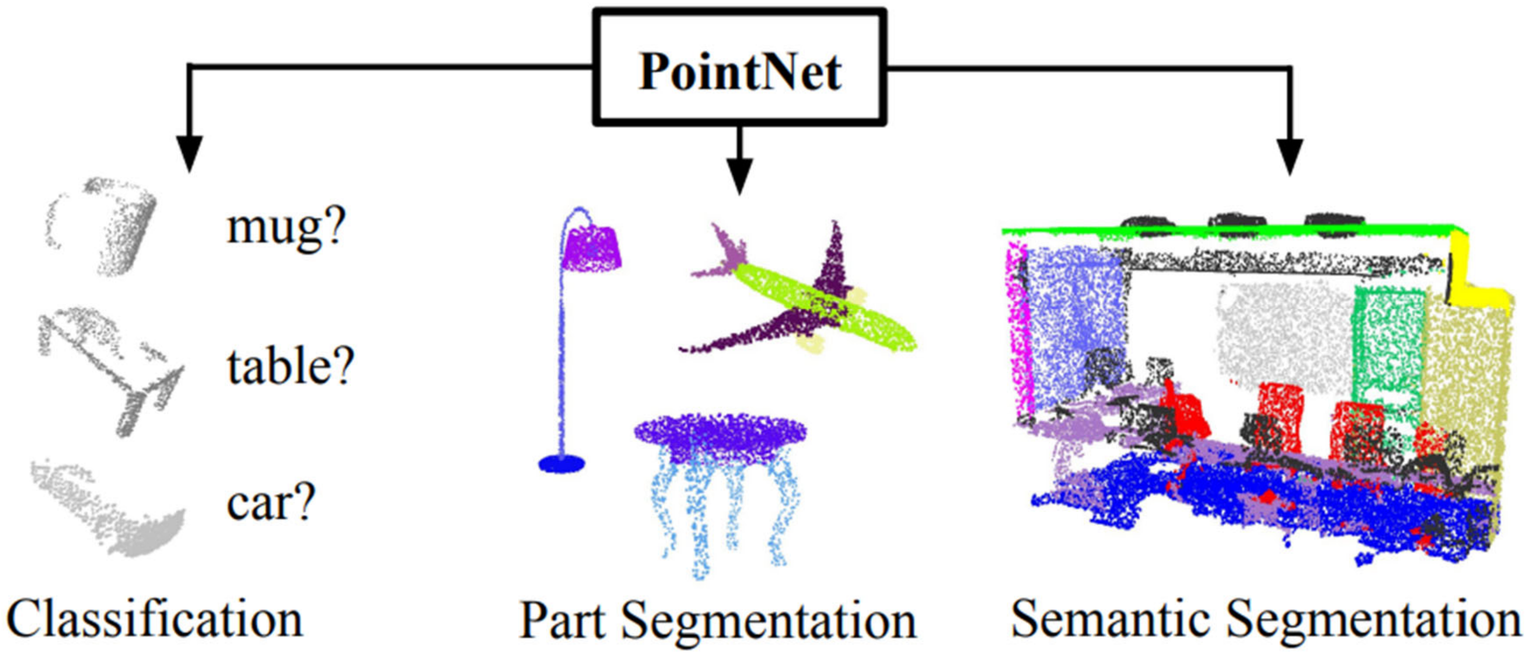
Goal: Find invariant and separating $\boldsymbol{F}: \boldsymbol{V} \to \mathbb{R}^{\boldsymbol{m}}$

- **Invariant:** if $[w] = [v]$ then $F(v) = F(w)$
- **Separating:** If $F(v) = F(w)$ then $[w] = [v]$

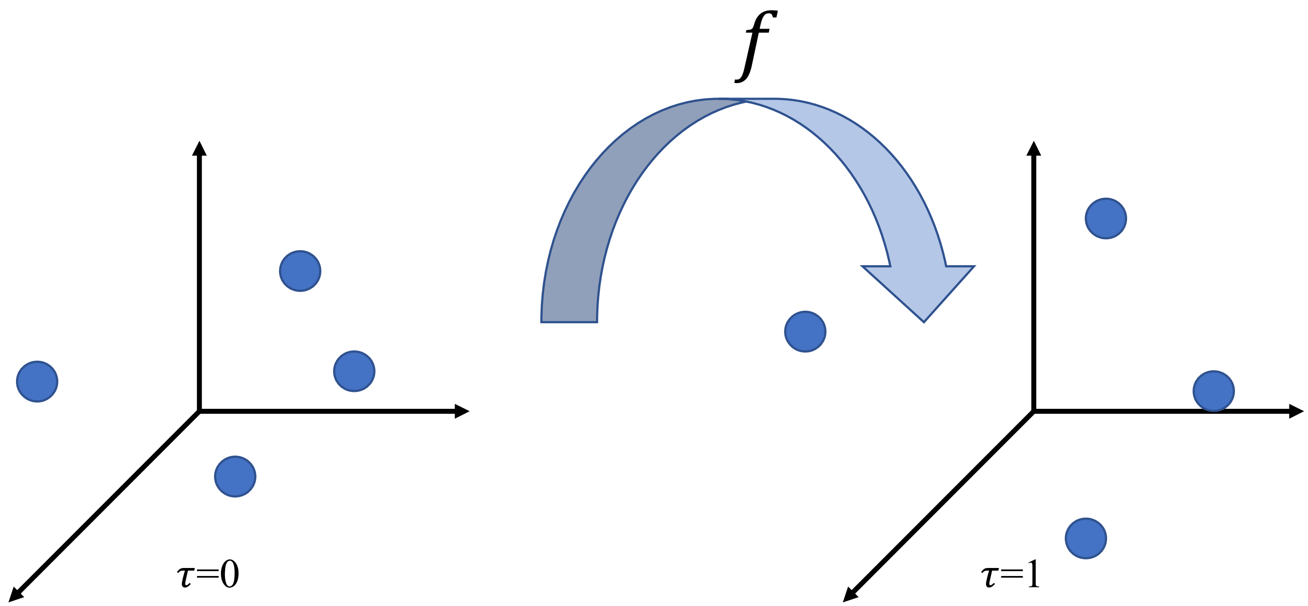
Conclusion: things we didn't discuss

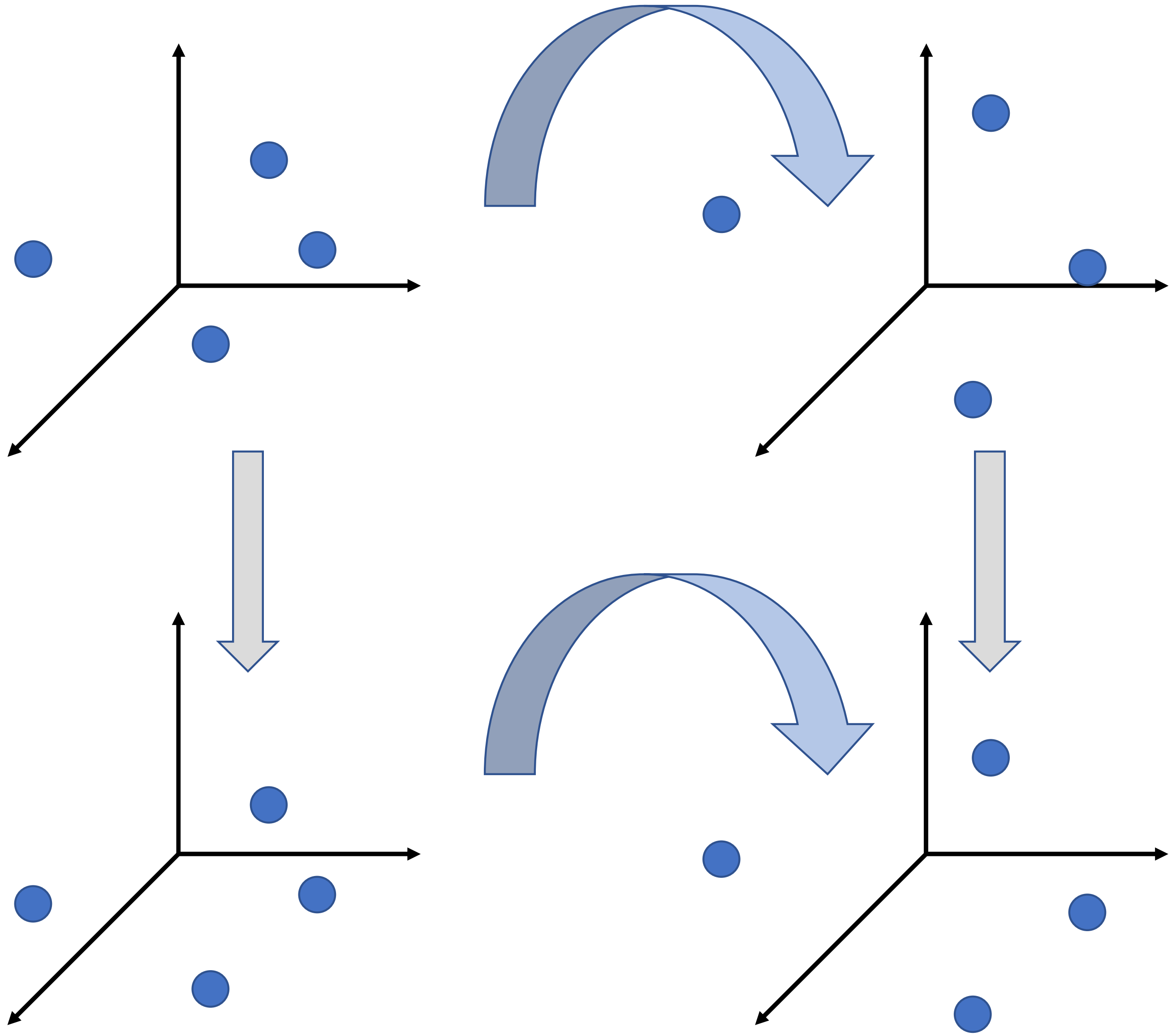
- Stability
- Equivariance
- Performance

Invariance vs. equivariance



Equivariance: For Physics simulation





N-body problem

Equivariant to
- Permutation
- Translation
- Orthogonal
- Lorenz!