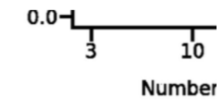
Exact and approximate symmetries in machine learning models

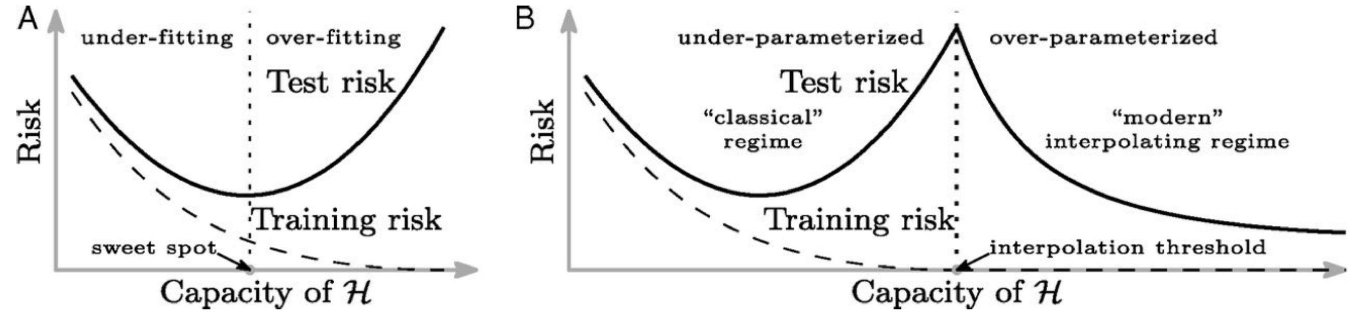
Soledad Villar
Johns Hopkins University

FANeSy School March 4 2024, Chile

Motivation: Deep learning inductive bias



Belkin et al '19
"double descent" →



- Very overparameterized models

  \# parameters >> # data points

- Many functions in the hypothesis class fit the data

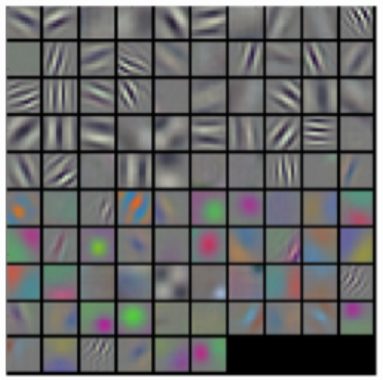
  How to choose the right one?

- Define the hypothesis class with the correct inductive bias

  Meaning that local optimization converges to a good solution

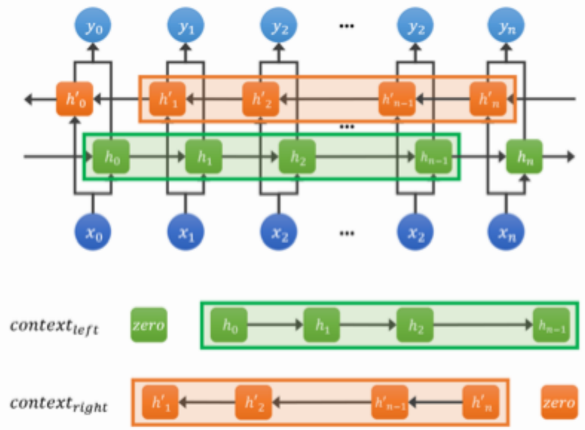
  (ie with good generalization)

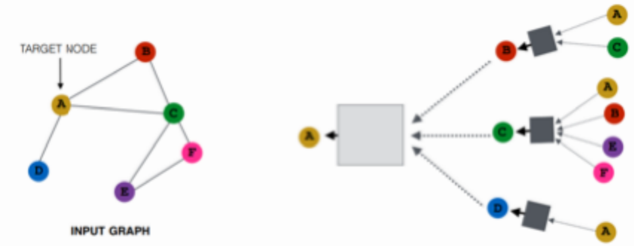
Deep learning architectures with good inductive bias



CNN
spatial translation symmetry

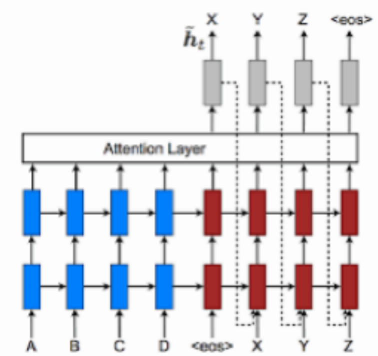
RNN
time translation symmetry





GNN
graph permutation symmetry

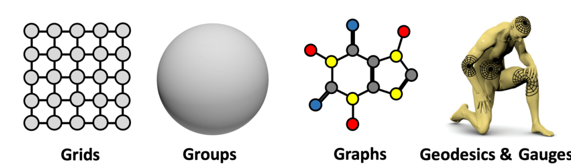
Transformers
flexible symmetries



(permutation and more!)

They exploit symmetries or approximate symmetries
[or more generally the physical structure of problems]

Related with geometric deep learning (Bronstein, Bruna, Cohen, Veličković)



Motivation

Symmetries in the physical sciences

2 types of symmetries:

ACTIVE

- Symmetries that come from observed regularities of physics

Emmy Noether

conservation of energy — time translation symmetry

conservation of momentum — translation symmetry

conservation of angular momentum — rotation symmetry

PASSIVE

- Symmetries that come from choice of mathematical representation of physical objects

coordinate freedom

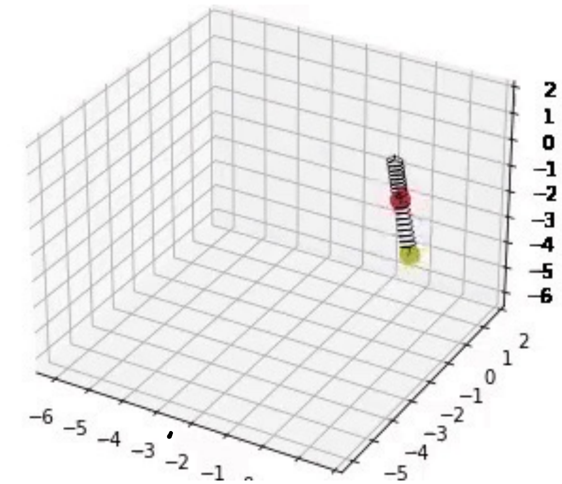
units equivariances

gauge invariances / equivariances

CLAIM: ML / data science methods should be consistent with these

Definition

Invariance / Equivariance



Exact symmetries

G a group acting on dataset X

$F: X \to Y$ invariant if $F(g \cdot x) = F(x) \quad \forall g \in G \quad x \in X$

→ "Rose"

If G also acts in Y

$H: X \to Y$ equivariant $H(g \cdot x) = g \cdot H(x) \quad \forall g \in G, \ x \in X$

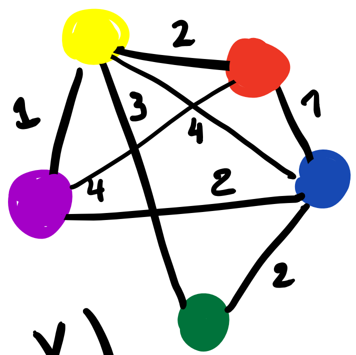
$$\begin{tikzcd} X \arrow[r, "H"] \arrow[d, "\varphi(g)"'] & Y \arrow[d, "\psi(g)"] \\ X \arrow[r, "H"] & Y \end{tikzcd}$$

All functions in this space are (approx) invariant or equivariant

EQUIVARIANT ML : $\mathcal{H} = \{ f_\theta : X \to Y \}$

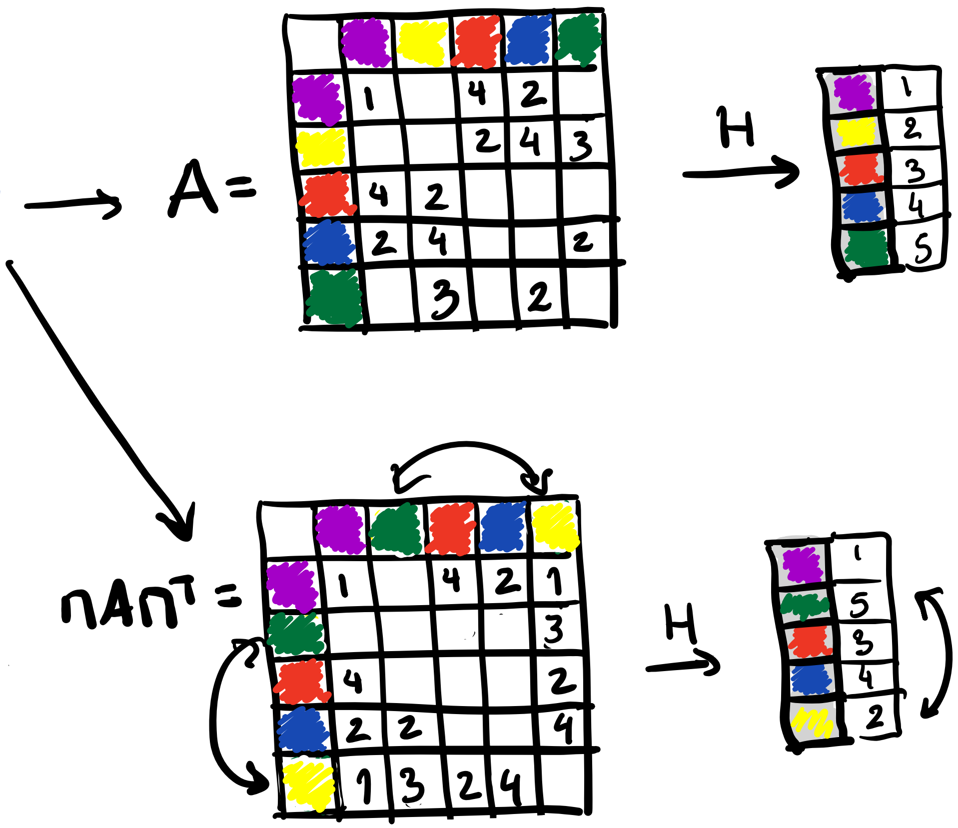
Example passive symmetries in graph learning

G
graph



$G = (A, X)$, $A \in \mathbb{R}^{n\times n}$, $X \in \mathbb{R}^{n\times d}$

$\Pi G = (\Pi A \Pi^T, \Pi X)$



Example:

shortest path starting at ●

- length (invariant)
- path (equivariant)

Graph classification / regression: $G \to \mathbb{R}$

$F(\Pi A \Pi^T, \Pi X) = F(A, X)$ **invariant**

Node embedding: $G \mapsto \mathbb{R}^{n\times d}$

$H(\Pi A \Pi^T, \Pi X) = \Pi \cdot H(A, X)$ **equivariant**

$\forall\ A \in \mathbb{R}^{n\times n},\ X \in \mathbb{R}^{n\times d}$
$\Pi \in$ permutation

We'll discuss Graph Neural Networks (GNNs) later...

How are symmetries implemented?

- Data augmentation
- Loss function penalties

Approximate

Credit: Bharath Raj

- Representation theory (Kondor, Cohen, Welling, ...)
- Group convolutions (Cohen, Trivedi ...)
- weight sharing (Duvenaud, Scarselli, ...)
- Invariant theory (Villar, Blum-Smith)
- ....

Exact

Restrict the class of functions to functions that satisfy the symmetries

Bonus if universal

This talk

- Symmetries by averaging
- How to use linear equiv layers to implement equivariant nerol networks (eg: representation theory)
  Kondor, Maron, Finzi, Welling, etc
- Equivariant functions as gradients of invariant functions Blum-Smith, V.
  - Examples
    - O(d), SO(d), O(1,d) equivariance V., Hogg, Storey-Fisher, Yao, Blum-Smith
    - Units equivariance V., Yao, Hogg, Blum-Smith, Dumitrescu
- Translation and rotation equivariance on images, vector and tensor fields Cohen, Welling Gregory et al
- Approximately equivariant contrastive learning Gupta, Robinson, Lim, V., Jegelka
- Invariant functions on point clouds with applications to cosmology
  Blum-Smith, Huang, Cuturi, V   Storey-Fisher, Hogg, Genel, V
- Approximate equivariance in graph neural networks Huang, Levie, V.
- Generalization gains of learning with symmetries Elesedy & Zaidi // Petrache & Trivedi Bietti, Venturi, Bruna // Tahmasebi & Jegelka

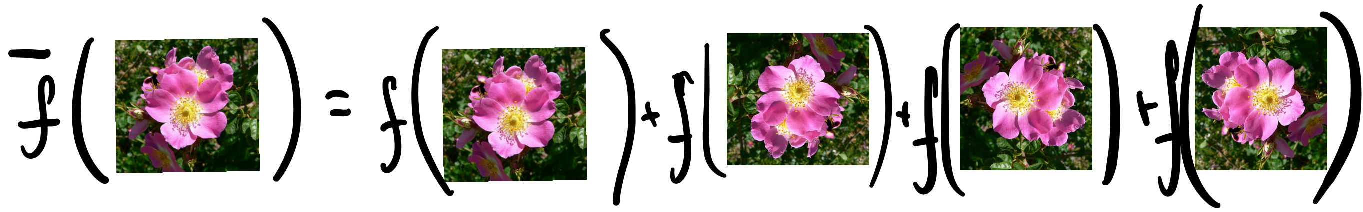
Implementation of symmetries via averaging

invariant

$$\bar{f}(x) = \frac{1}{|G|}\sum_{g\in G} f(g\cdot x)$$

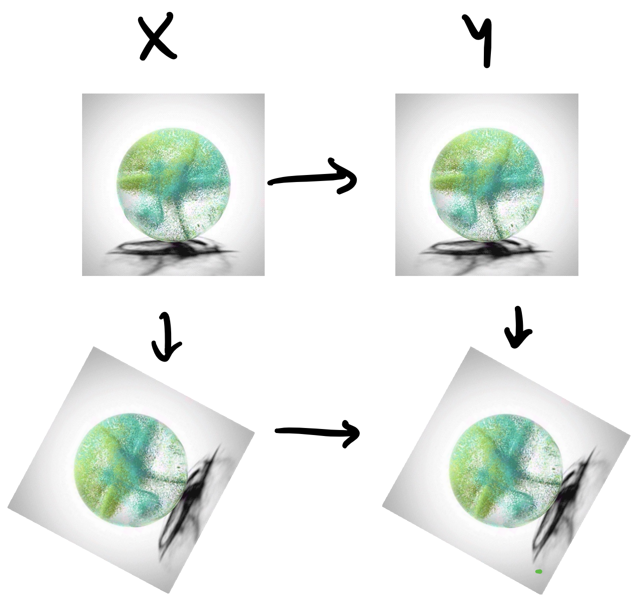
G-invariant function

any function



equivariant

$$\begin{tikzcd}
X \arrow[r, "\bar{f}"] \arrow[d, "\psi(g)"'] & Y \arrow[d, "\psi(g)"] \\
X \arrow[r, "\bar{f}"] & Y
\end{tikzcd}$$



$$\bar{f}(x) = \frac{1}{|G|}\sum_{g\in G} \psi(g)^{-1} f(\psi(g)\,x)$$

G-equivariant function

any function

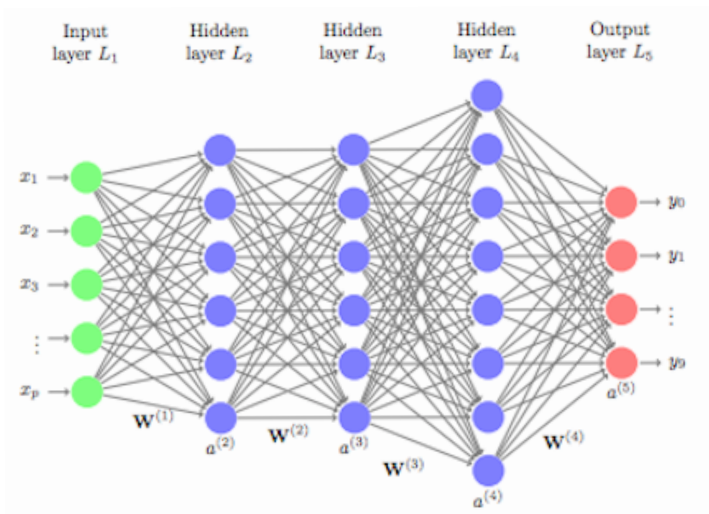
If G infinite but compact

$\Rightarrow \sum_{g\in G} \mapsto \int_G dg$ (Haar measure)

Representation theory approach (based on feed forward neural networks)

- Kondor 2018, Maron et al 2019 ....

Feed Forward NEURAL NETWORK



pointwise non-linear activations

input

$$F(v) = \theta \circ L_n \circ \dots L_2 \circ \theta \circ L_1 (v)$$

linear layers

Idea: Replace linear layers by linear equivariant/invariant layers

$$L_i(Qv) = Q . L_i(v)$$

Issue #1: Not many linear equivariant/invariant functions.
What linear rotation invariant functions can you think of?

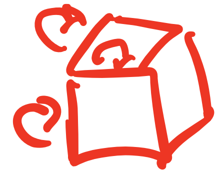
Issue #2: What are compatible activation functions?

They depend on the group: permutations (all work)
rotations (??)

Solution to Issue # 1 : Extend the action to tensors :

$$L_i : (\mathbb{R}^d)^{\otimes k_i} \rightarrow (\mathbb{R}^d)^{\otimes k_{i+1}}$$

equivariant linear function



Maron '19

$$L_i(Qv) = Q L_i(v)$$

$$Q(v \otimes \ldots \otimes v) = Qv \otimes \ldots \otimes Qv$$

Q: How to parametrize linear equivariant functions?

A: Representation theory

+ Schur's lemma

+ computing Clebsh-Gordan coefficients to decompose $(\mathbb{R}^d)^{\otimes K}$ in irreps

- Fuchs et al 2020
- Thomas, Smidt et al 2018
- E3NN (Smidt et al 2022)

Group representations

$G \circlearrowleft X$ $\quad \star : G \longrightarrow Sym(X)$ such that

Action

- identity $\quad e \star x = x \quad \forall x \in X$
- compatibility $\quad g \star (h \star x) = (g \cdot h) \star x$

group multiplication

Group representation

A group action $\rho : G \longrightarrow GL(V)$ is called a representation

It allows to represent the group elements as matrices and the group multiplication as matrix multiplication

Irreducible representations

$\rho : G \to W \to W$

$\rho : G \longrightarrow GL(V)$ representation has $\rho_W : G \longrightarrow GL(W)$ as a subrepresentation if $W \subset V$ linear subspace and $\rho|_V = \rho_W$. ρ is irreducible if it has no no-trivial subrepresentations

Schur's lemma

V, W vector spaces ρ_V, ρ_W irreducible representations of G on V and W.

1. If V and W are not isomorphic then there are no non-trivial linear equivariant maps $f : V \longrightarrow W$ (aka G-maps)

$$\rho_W(g) \circ f = f \circ \rho_V(g) \quad \forall g \in G$$

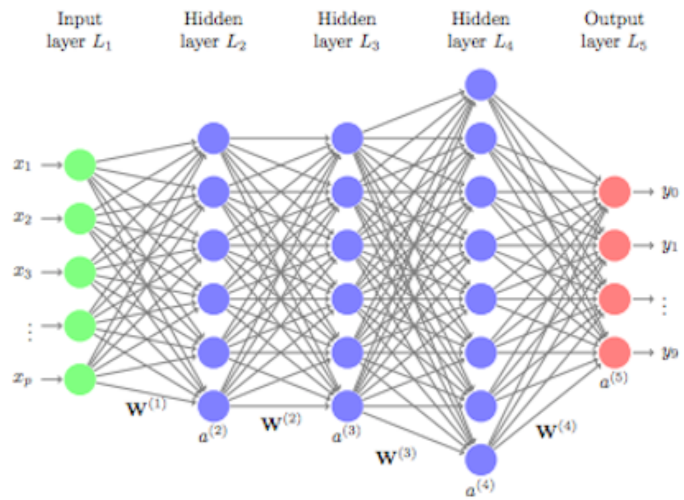
$$\iff \quad g \cdot f(v) = f(g \cdot v) \quad \forall g \in G \quad \forall v \in V$$

$$\begin{tikzcd} V \arrow[r, "f"] \arrow[d, "\rho_V(g)"'] & W \arrow[d, "\rho_W(g)"] \\ V \arrow[r, "f"] & W \end{tikzcd}$$

$$f(\rho_V(g)v) = \rho_W(g) f v$$

2. If $V = W$ finite dimensional (over $\mathbb{C}$) and $\rho_V = \rho_W$ then the only non trivial linear equivariant maps are scalar multiples of the identity.

Equivariant linear layers via irreps



pointwise non-linear activations   input

1. $F(v) = \theta \circ L_n \circ \dots L_2 \circ \theta \circ L_1 (v)$

linear layers

2. $L_i : (\mathbb{R}^d)^{\otimes k_i} \to (\mathbb{R}^d)^{\otimes k_{i+1}}$ linear equivariant

3. Decompose $(\mathbb{R}^d)^{\otimes k_i}$ and $(\mathbb{R}^d)^{\otimes k_{i+1}}$ in irreps

$$\bigoplus_{\ell=1}^{T_{k_i}} T_\ell \xrightarrow{\text{Schur's lemma}} \bigoplus_{\ell'=1}^{T_{k_{i+1}}} T_{\ell'}$$

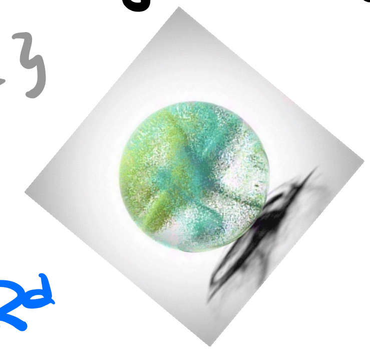
Given by Clebsch-Gordan coefficients

Dym and Maron 2021 - This approach universally approximates all SO(3) equiv functions If arbitrary high order tensors are involved

Invariant theory approach

Example **O(d)** Orthogonal group

$\{R \in \mathbb{R}^{d\times d} : RR^T = R^T R = I\}$



$f \to \mathbb{R}$

O(d)- invariant functions $f : (\mathbb{R}^d)^n \to \mathbb{R}$

$$f(Rv_1, \ldots Rv_n) = f(v_1, \ldots v_n) \quad \forall\ R \in O(d) \quad v_1 \ldots v_n \in \mathbb{R}^d$$

(Weyl 1946) first fundamental theorem of invariant theory

$f : (\mathbb{R}^d)^n \to \mathbb{R}$ is $O(d)$-invariant if and only if

$$f(v_1, \ldots v_n) = \tilde{f}\Big(\underbrace{(v_i^T v_j)_{i,j=1}^n}_{\text{inner products}}\Big)$$

- Similar characterizations for
  - Lorentz
  - Rotations
  - Symplectic group
  - Unitary group

$$\begin{tikzcd}
(\mathbb{R}^d)^n \arrow[r, "f"] & \mathbb{R} \\
(v_1, \ldots v_n) \arrow[d, "\Phi"] & \\
\mathbb{R}^{n\times n} \arrow[ruu, dashed, "\tilde{f}"'] &
\end{tikzcd}$$

$M \quad M_{ij} = v_i^T v_j$

(Villar et al '21)
Neurips '21

Characterization of SO(d)-invariant functions

$f : (\mathbb{R}^d)^n \to \mathbb{R}$ is SO(d)-invariant if and only if

$$f(v_1, \ldots, v_n) = \tilde{f}\left((v_i^T v_j)_{i,j=1}^n ,\ \det(v_{i_1} \ldots v_{i_d})_{i_1 \ldots i_d \in \binom{[n]}{d}} \right)$$

$$V = \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix}$$

Characterization of Lorentz-invariant functions

$f : (\mathbb{R}^{d+1})^n \to \mathbb{R}$ is Lorentz-invariant if and only if

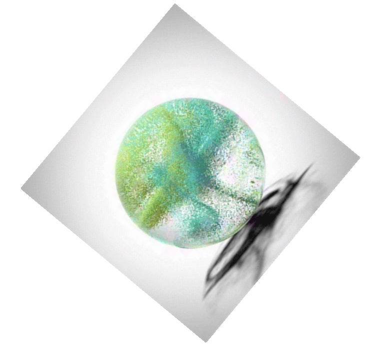
$$f(v_1, \ldots, v_n) = \tilde{f}\left(\langle v_i, v_j \rangle_M {}_{i,j=1}^{n} \right)$$

where $\langle (t, x), (t', x') \rangle_M = tt' - x^T x'$

Minkowski "inner product"

$$O(1,d) = \{ R : R \Lambda R^T = \Lambda \ : \ \Lambda = \begin{pmatrix} -1 & 0 \\ 0 & 1 \ddots 1 \end{pmatrix} \}$$

Example O(d) Orthogonal group



$h \xrightarrow{} \mathbb{R}^d$

equivariant functions $h:(\mathbb{R}^d)^n \to \mathbb{R}^d$

$$h(Rv_1, \dots Rv_n) = R \cdot h(v_1, \dots v_n) \quad \forall\, R \in O(d) \quad v_1 \dots v_n \in \mathbb{R}^d$$

Proposition

$h:(\mathbb{R}^d)^n \to \mathbb{R}^d$ is $O(d)$-equivariant if and only if

$$h(v_1, \dots v_n) = \sum_{s=1}^{n} \underbrace{f_s\Big(\underbrace{(v_i^T v_j)_{i,j=1}^{n}}_{\text{inner products}} \Big)}_{\text{invariant scalar functions}} v_s$$

Proof (by Schwarz, Malgrange)

$h:(\mathbb{R}^d)^n \to \mathbb{R}^d$ equivariant

$f:(\mathbb{R}^d)^n \times \mathbb{R}^d \to \mathbb{R}$ invariant

$$f(v_1 \dots v_n, v^*) = \langle h(v_1 \dots v_n), v^* \rangle \qquad \frac{\partial}{\partial v^*}$$

$$f(v_1 \dots v_n, v^*) = \xi\left(\langle v_i, v_j \rangle_{i,j=1 \dots n}, \langle v_s, v^* \rangle_{s=1 \dots n} \right)$$

$$= \sum f_s(\langle v_i, v_j \rangle) \langle v_s, v^* \rangle \qquad \frac{\partial}{\partial v^*}$$

$$= \sum f_s(\langle v_i, v_j \rangle) v_s$$

General theory INVARIANCE → EQUIVARIANCE

(Malgrange) Schwarz ('75, '80)

Ben Blum-Smith

Knowledge of invariant maps $\underline{V \times W^* \to \mathbb{R}}$

$\Rightarrow$ Knowledge of equivariant maps $V \to W$

If $h: V \to W$ equivariant $\Rightarrow$ $f: V \times W^* \to \mathbb{R}$ invariant

$$(v, \ell) \longmapsto \ell(f(v))$$

$$\underset{}{\mathrm{Hom}(V, W)} \cong \underset{\text{(linear maps)}}{\mathrm{Hom}(V \otimes W^*, \mathbb{R})} \cong \underset{\text{(bilinear maps)}}{\mathrm{Hom}(V \times W^*, \mathbb{R})}$$

Formally

$\left.\begin{array}{l} G \text{ acts on } V \text{ by } \phi \\ G \text{ acts on } W \text{ by } \psi \end{array}\right\} \Rightarrow G$ acts on maps(V, W)

$$gf := \psi(g) \circ f \circ \phi^{-1}(g)$$

$$\begin{tikzcd} V \arrow[r, "f"] \arrow[d, "\phi(g)"'] & W \arrow[d, "\psi(g)"] \\ V \arrow[r, "gf"] & W \end{tikzcd}$$

f is a fixed point of the action on maps(V, W) $\Longleftrightarrow$ f is equivariant

$$f = gf \Leftrightarrow f \text{ equivariant}$$

$$\begin{tikzcd} V \arrow[r, "f"] \arrow[d, "\phi(g)"'] & W \arrow[d, "\psi(g)"] \\ V \arrow[r, "f"] & W \end{tikzcd}$$

Algorithm invariants to equivariance

Input: bihomogeneous generators $f_1 \dots f_m$ for $\mathbb{R}[V \times W^*]^G$ G-invariants

① $f_1 \dots f_r$ deg 0 in W^*
$f_{r+1} \dots f_s$ deg 1 in W^*
discard the rest

② Choose basis $e_1, \dots, e_n$ for W
$e_1^T \dots e_n^T$ for W^* dual basis

$$\ell = \sum_{i=1}^{n} \ell_i e_i^T$$

$$\ell(e_i) = \ell_i$$

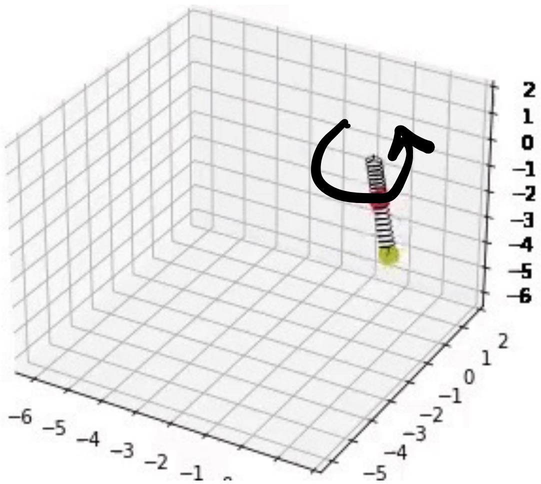
③ Define $F_j : V \to W$ $j = r+1 \dots s$ derivative of $f_j(v, \ell)$ wrt ℓ

$$F_j(v) := \sum_{i=1}^{n} \left(\frac{\partial}{\partial \ell_i} f_j(v, \ell) \right) e_i$$

Output: Equivariant functions h can be written as

$$h = \sum_{j=r+1}^{s} P_j(f_1 \dots f_r) F_j$$

Toy example : double pendulum with springs



Credit : EMLP (Finzi et al '21)

data: $(q_1(t), p_1(t))$, m_1, L_1, K_1

$(q_2(t), p_2(t))$ m_2, L_2, K_2

Weichi Yao

problem: predict the dynamics

$$KE = \frac{|p_1|^2}{2m_1} + \frac{|p_2|^2}{2m_2} \qquad PE = \frac{1}{2}K_1(|q_1| - L_1)^2 - m_1 p_1 \cdot g + \frac{1}{2}K_2(|q_1 - q_2| - L_2)^2 - m_2 p_2 \cdot g$$

$H = KE + PE$ conserved quantity $\longleftrightarrow$ time translation symmetry

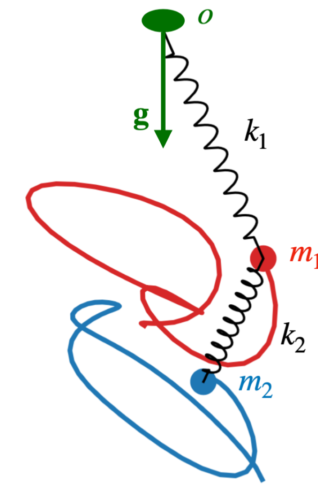
(Hamiltonian)

$F : (\mathbb{R}^3)^5 \times \mathbb{R} \longrightarrow (\mathbb{R}^3)^4$

O(3)-equivariant (passive symmetry)

$$(q_1(0), p_1(0), q_2(0), p_2(0), g, \Delta t) \longmapsto (q_1(\Delta t), p_1(\Delta t), q_2(\Delta t), p_2(\Delta t))$$

Computational approaches:



Goal: Predict $z(t) = (q_1(t), q_2(t), p_1(t), p_2(t))$

- Neural ODEs

$$(z(0), g) \longrightarrow O(3)\text{-equivariant function } F_\theta \longrightarrow \frac{dz(t)}{dt} = F_\theta(z, g) \xrightarrow[\text{ODE solver}]{} \hat{z}(t)$$

$$\text{input} \longrightarrow \text{scalars} \xrightarrow{f_1} \vdots \,(\text{MLP}) \longrightarrow \oplus \longrightarrow \text{output}$$

q_1, q_2, p_1, p_2, g $\qquad$ $q_1^T q_1$, $q_1^T q_2$, $q_1^T p_1$

$$F_\theta(q_1, q_2, p_1, p_2, g) = \sum_v f_v\big((v_i, v_j)_{v_i, v_j \in V}\big) \cdot v$$

$$v_i \in V = \{q_1, q_2, p_1, p_2, g\}$$

- Hamiltonian neural network (HNNs)

$$(z(0), g) \rightarrow \text{O(3) - invariant function } H_\theta \rightarrow \frac{dp_1}{dt} = -\frac{dH_\theta}{dq_1}, \quad \frac{dq_1}{dt} = \frac{dH_\theta}{dp_1} \xrightarrow[\text{ODE solver}]{} \hat{z}(t)$$

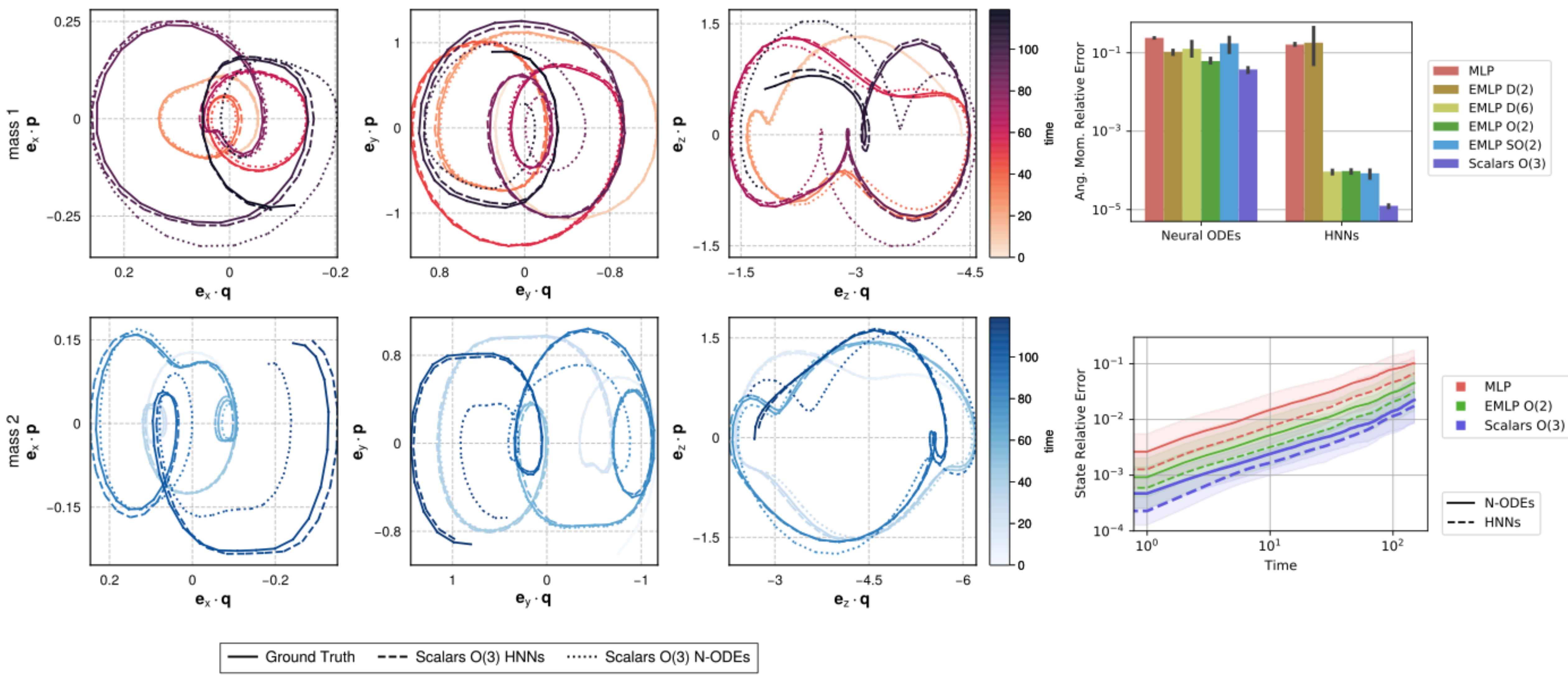
(symplectic integrator)

$$H_\theta(q_1, q_2, p_1, p_2, q_0, g) = h\,(\text{inner products})$$

Learned scalar invariant function

Results

	Scalars O(3)	EMLP				MLP
		O(2)	SO(2)	D_2	D_6	
N-ODEs:	**.009 ± .001**	.020 ± .002	.051 ± .036	.023 ± .002	.036 ± .025	.048 ± .000
HNNs:	**.005 ± .002**	.012 ± .002	.016 ± .003	.111 ± .167	.013 ± .002	.028 ± .001



Units - equivariance

Non-compact groups (Villar et al '22)

Double pendulum:

$$PE = \frac{1}{2} k_1 (|q_1| - L_1)^2 - m_1 p_1 \cdot g + \frac{1}{2} k_2 (|q_1 - q_2| - L_2)^2 - m_2 p_2 \cdot g$$

$$KE = \frac{|p_1|^2}{2 m_1} + \frac{|p_2|^2}{2 m_2}$$

Energy has units: $kg\ m^2\ s^{-2}$

Predictions should be equivariant with respect to rescalings

↳ Passive symmetry from dimensional analysis

Approach:

inputs: $\mathbb{R}^d$ $\xrightarrow{\Phi}$ dimensionless features $\mathbb{R}^s$ $\xrightarrow{\text{NNET}}$ dimensionless labels $\xrightarrow{\text{Fixed scaling}}$ dimensional outputs

each feature has units

Φ ↓ Based on classical dimensional analysis (Buckingham PI theorem)

Units-typed space

(x, u)

$x \in \mathbb{R}$ feature, $u \in \mathbb{Z}^k$ units

eg (kg, m, s) exponents

Energy $kg\, m^2 s^{-2}$: $[1, 2, -2]$

$(\lambda_1\ \lambda_2\ \lambda_3) \cdot x$

$\lambda_1^1 \cdot \lambda_2^2 \cdot \lambda_3^{-2}\, x$

$\alpha \cdot (x, u) = (\alpha x, u)$, $\alpha \in \mathbb{R}$

$$(x, u) + (x', u') = \begin{cases} (x + x', u) & \text{if } u = u' \\ \nexists & \text{otherwise} \end{cases}$$

$$(x, u)(x', u') = (xx', u + u')$$

$$(x, u)^\gamma = (x^\gamma, \gamma \cdot u), \quad \gamma \in \mathbb{Z}$$

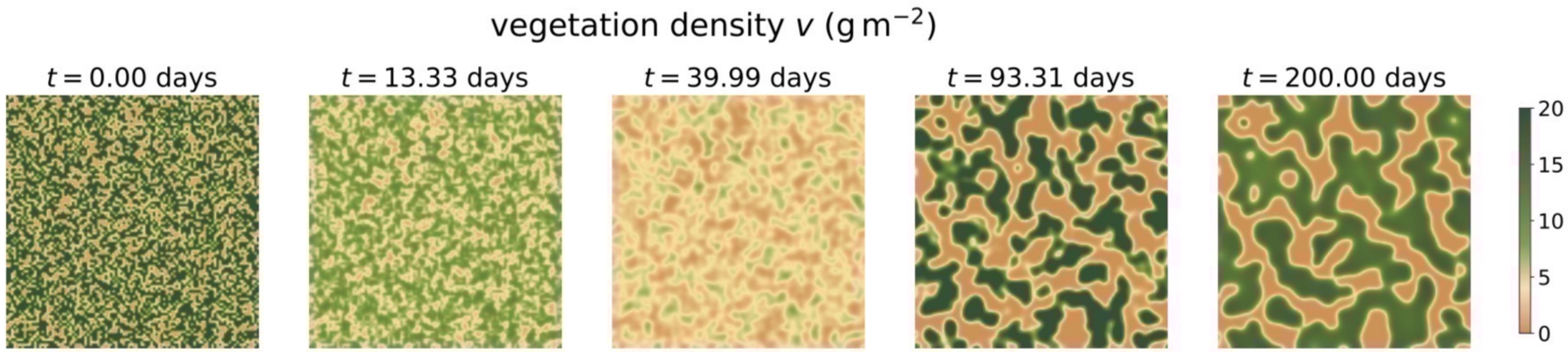
Dimensionless features

$\mathbf{x} = (x_i, u_i)_{i=1 \ldots d}$

$$\xi = \Phi(\mathbf{x}) = \prod_{i=1}^{d} x_i^{\alpha_i} \quad \text{where} \quad \sum_{i=1}^{d} \alpha_i u_i = 0, \quad u_i \in \mathbb{Z}^k$$

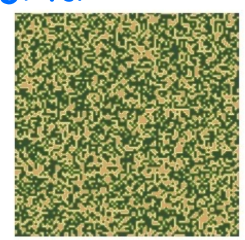
dimensionles features = # input variables − # independent units

Example: vegetation dynamics



Bianca Dumitrascu

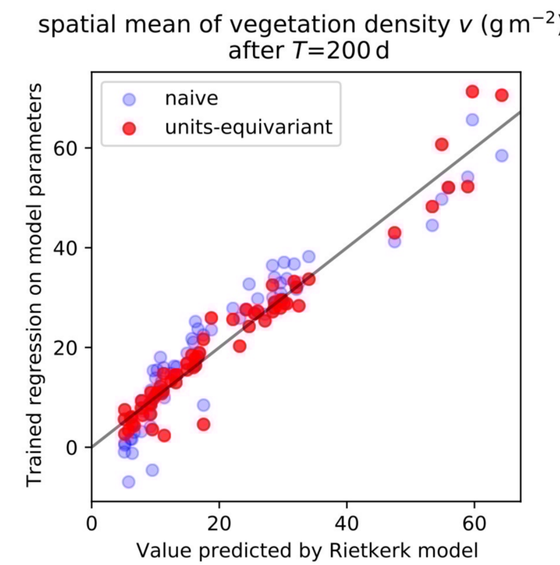
Rietkerk model



$$\frac{\mathrm{d}u}{\mathrm{d}t} = R - \alpha \frac{v + k_2 W_0}{v + k_2} u + D_u \nabla^2 u$$
$$\frac{\mathrm{d}w}{\mathrm{d}t} = \alpha \frac{v + k_2 W_0}{v + k_2} u - g_m \frac{v\, w}{k_1 + w} - \delta_w w + D_w \nabla^2 w$$
$$\frac{\mathrm{d}v}{\mathrm{d}t} = c\, g_m \frac{v\, w}{k_1 + w} - \delta_v v + D_v \nabla^2 v,$$

	description	default	units
R	rainfall	0.375	$\ell\, \mathrm{d}^{-1}\, \mathrm{m}^{-2}$
α	infiltration rate	0.2	d^{-1}
k_2	saturation const.	5	$\mathrm{g\, m}^{-2}$
W_0	water infiltration const.	0.1	—
D_u	surface water diffusion	100	$\mathrm{d}^{-1}\, \mathrm{m}^2$
g_m	water uptake	0.05	$\ell\, \mathrm{g}^{-1}\, \mathrm{d}^{-1}$
k_1	water uptake constant	5	$\ell\, \mathrm{m}^{-2}$
δ_w	soil water loss	0.2	d^{-1}
D_w	soil water diffusion	0.1	$\mathrm{d}^{-1}\, \mathrm{m}^2$
c	water to biomass	20	$\ell^{-1}\, \mathrm{g}$
δ_v	vegetation loss	0.25	d^{-1}
D_v	vegetation diffusion	0.1	$\mathrm{d}^{-1}\, \mathrm{m}^2$
T	total integration time	200	d
δt	integration time step	0.005	d
L	integration patch length	200	m
δl	spatial step size	2	m

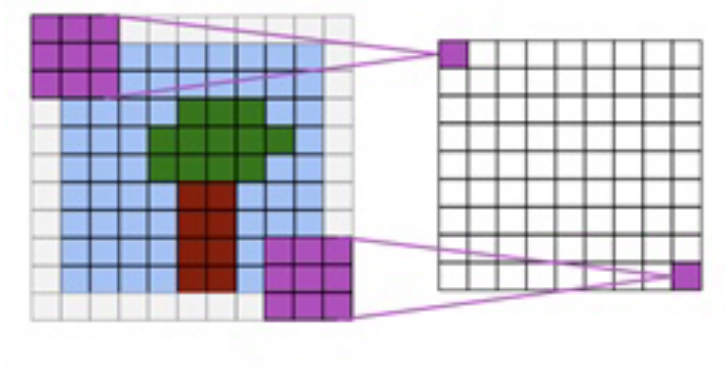
Dimensionless features
$c\, \alpha^{-1}\, g_m$
$R^{-1} \alpha\, k_1$
$R^{-1} c^{-1} \alpha\, k_2$
$\alpha^{-1} \delta_w$
$\alpha^{-1} \delta_v$
W_0
$\alpha^{-1} D_v\, L^{-2}$
$\alpha^{-1} D_u\, L^{-2}$
$\alpha\, T$
$\alpha\, \delta t$
$\alpha^{-1} D_w\, L^{-2}$
$L^{-1} \delta l$

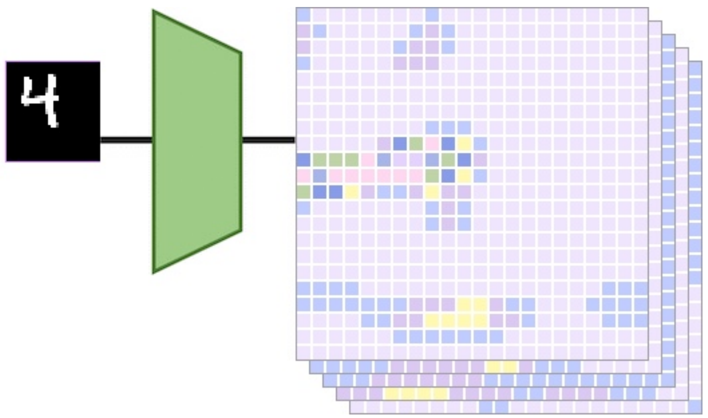


Problem: learn the differential equations from data using equivariant symbolic regression & equivariant convolutions

Equivariance on images

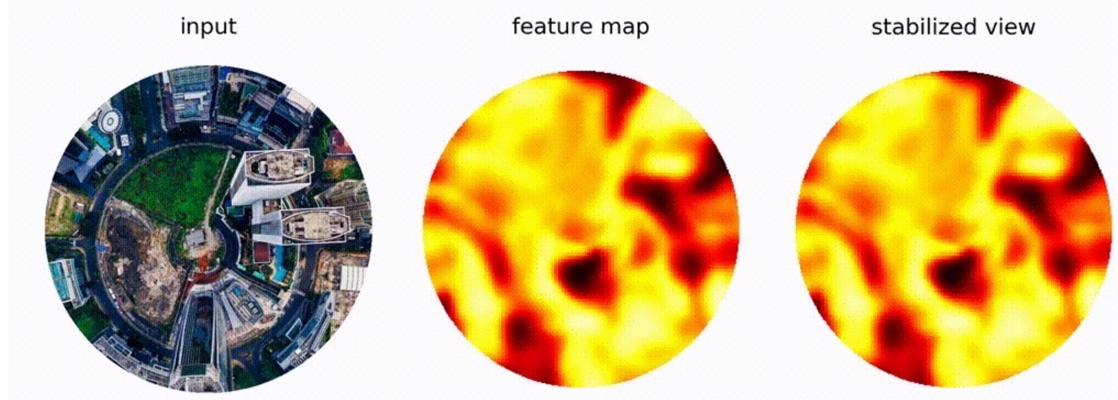
LeCun '89
Cohen & Welling '16



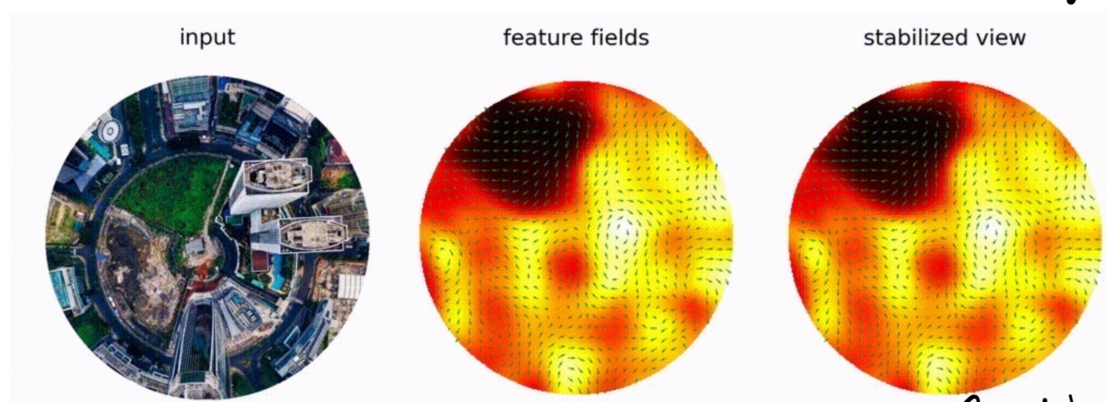


Credit: Christian Wolf

Convolutions are translation-equivariant

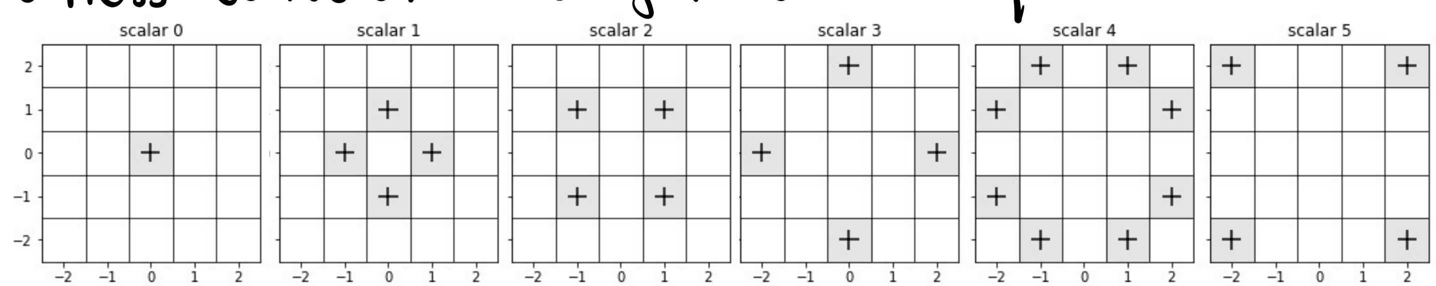


Convolutions are NOT rotation equivariant



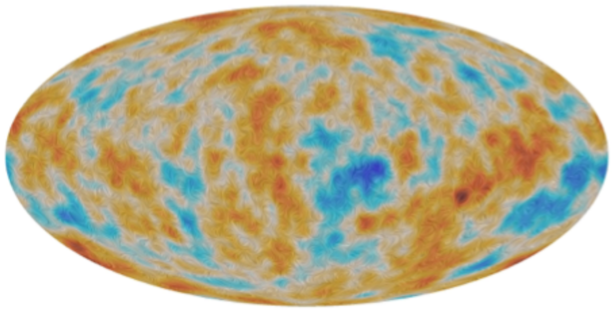
Credit: Weiler & Cesa

Unless convolutional filters are symmetric

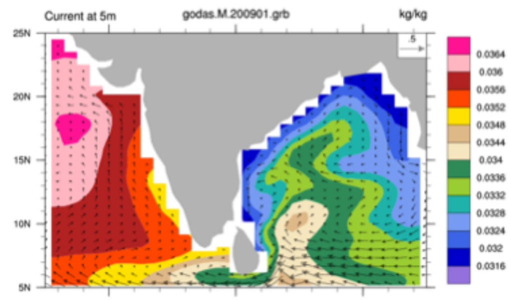


Natural extension to vector and tensor fields

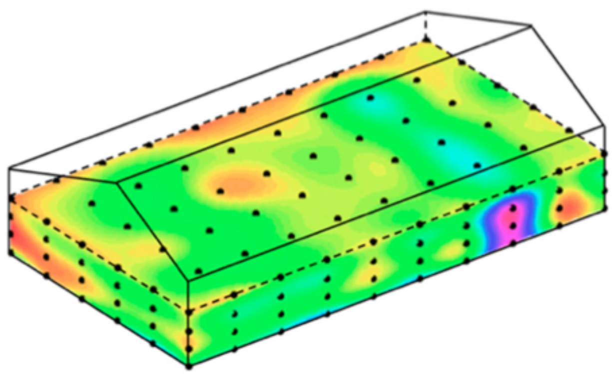
Gregory et al '23



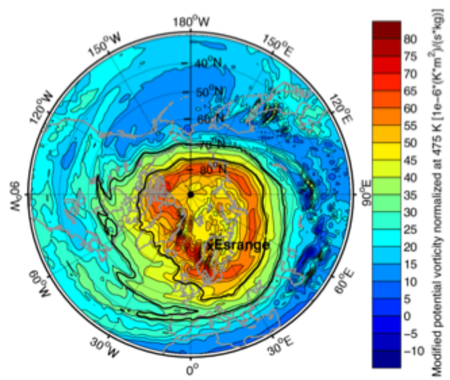
(a) temperature and polarization



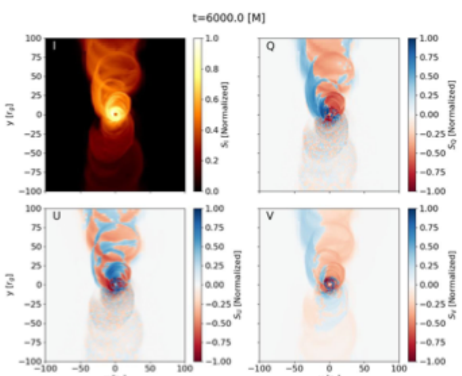
(b) salinity and current



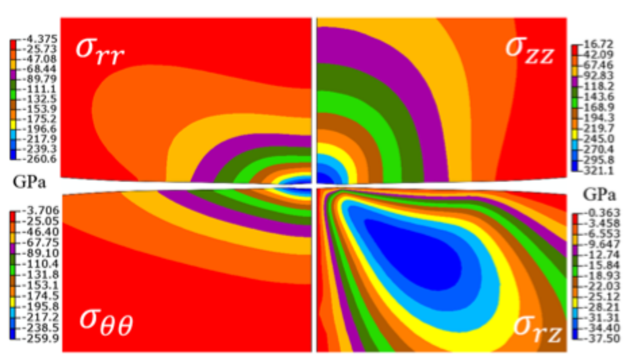
(c) temperature



(d) vorticity

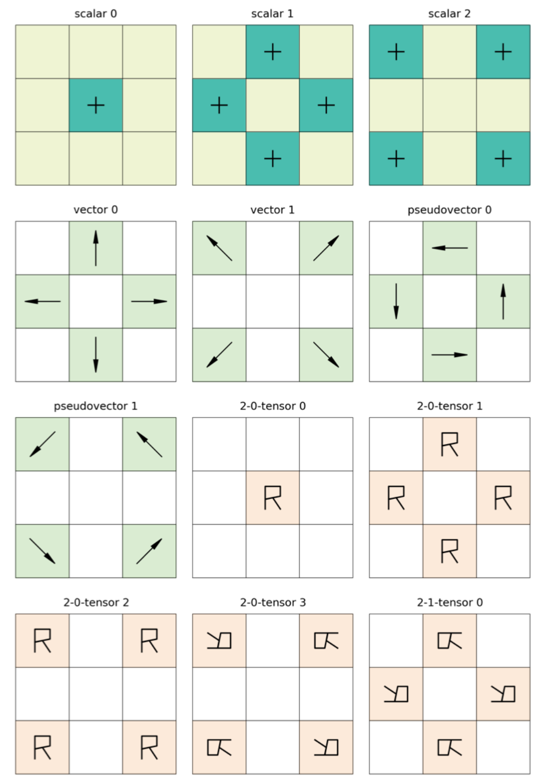


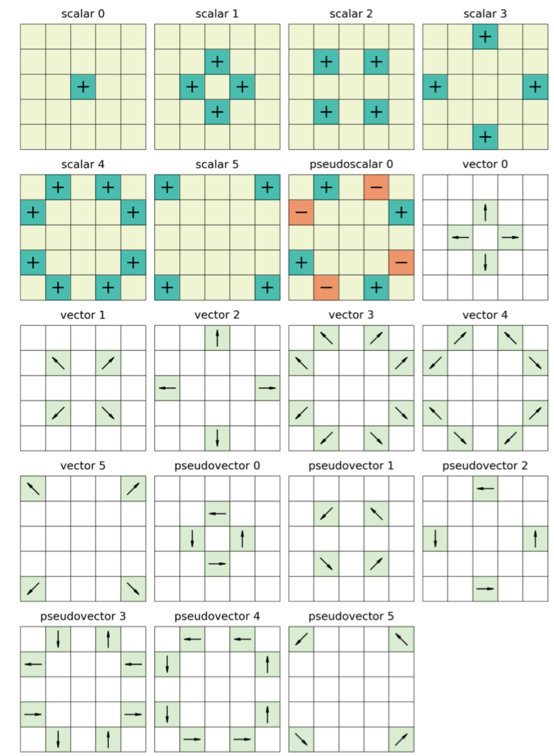
(e) intensity and polarization



(f) stress tensor

Combine equivariant formulation based on scalars with convolutions





- Linear equivariant layers are convolutions with symmetric vector, tensor and (pseudo) filters
- Discretization of coordinate-free operators (grad, curl)
- Combine with tensor products and Einstein summation contractions

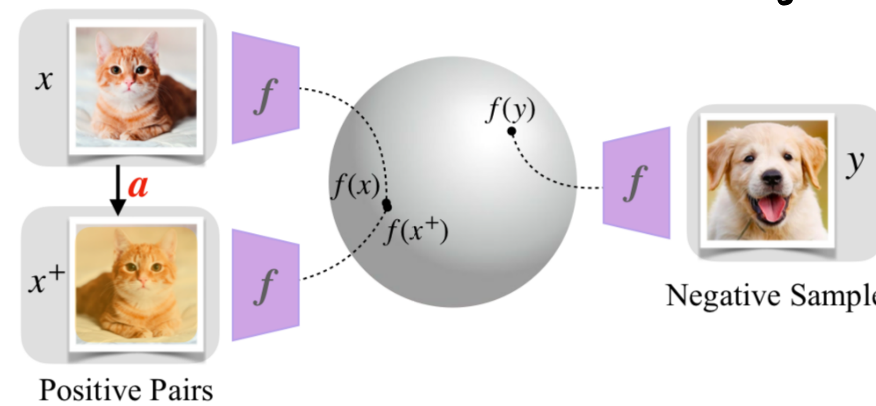
Application

Approximate symmetries in contrastive learning

S. Gupta J. Robinson D. Lim S. Jegelka

Contrastive learning (self-supervised)

$$f : \underset{\substack{\uparrow \\ \text{data}}}{X} \longrightarrow \underset{\substack{\uparrow \\ \text{embedding}}}{S^{d-1}}$$



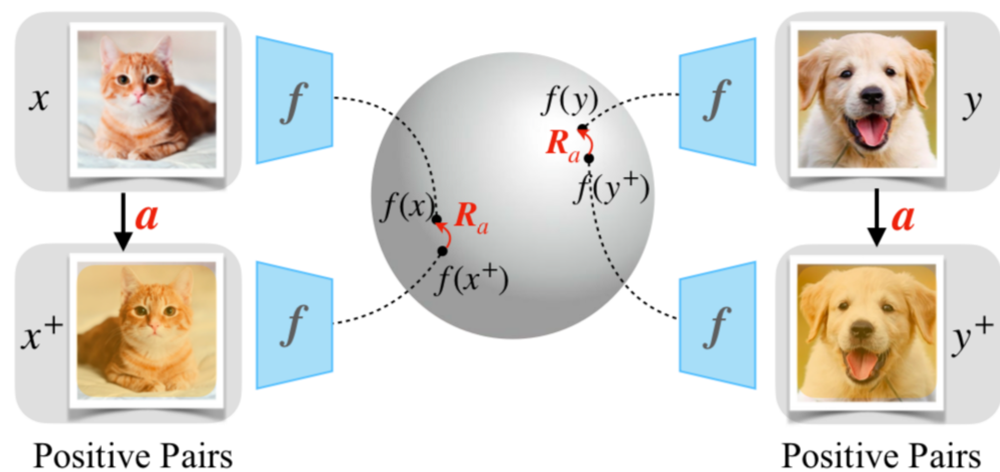
Invariant Contrastive Learning (SimCLR)

$$\mathcal{L}_{\text{infoNCE}} = \underset{x, x^+, \{x_i^-\}_{i=1}^N}{\mathbb{E}} - \log \frac{e^{\frac{f(x)^T f(x^+)}{\tau}}}{e^{\frac{f(x)^T f(x^+)}{\tau}} + \sum_{i=1}^{N} e^{\frac{f(x)^T f(x_i^-)}{\tau}}}$$

$\uparrow$ $a(x)$ augmented versions of points

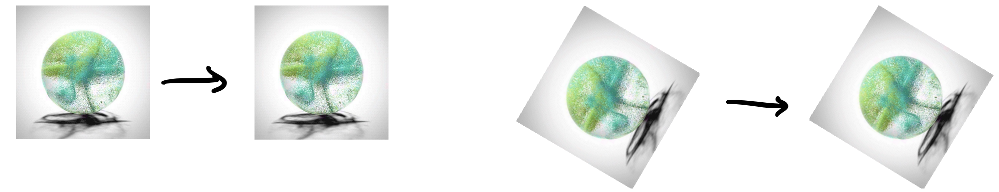
Equivariance to augmentation

- $f(a(x)) = T_a f(x)$
- Equivariance should be expressed in terms of pairs of data points



Orthogonally Equivariant Contrastive Learning (CARE)

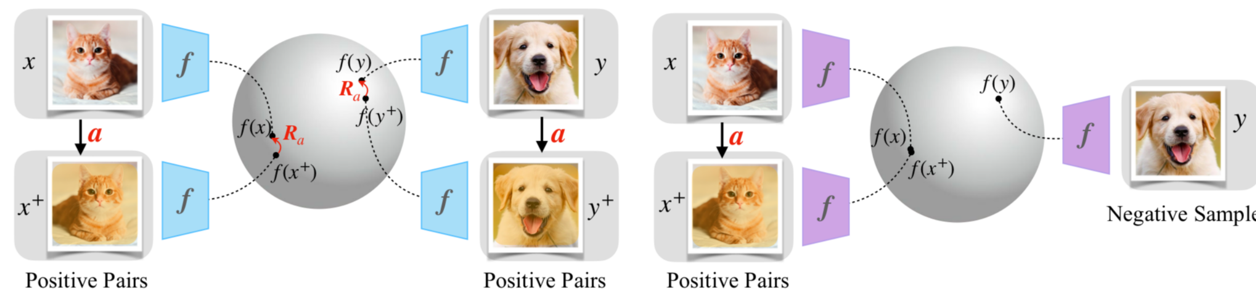
O(d)-equivariant contrastive learning



Recall $y_1, y_2, z_1, z_2 \in S^{d-1}$ satisfy $\overbrace{y_1^T y_2 = z_1^T z_2}^{\text{same inner product}}$

iff there exists $R \in O(d)$ such that $Ry_1 = z_1$, $Ry_2 = z_2$

Equivariant contrastive learning:



Orthogonally Equivariant Contrastive Learning (CARE) **Invariant Contrastive Learning (SimCLR)**

$f: X \to \mathbb{R}^d$

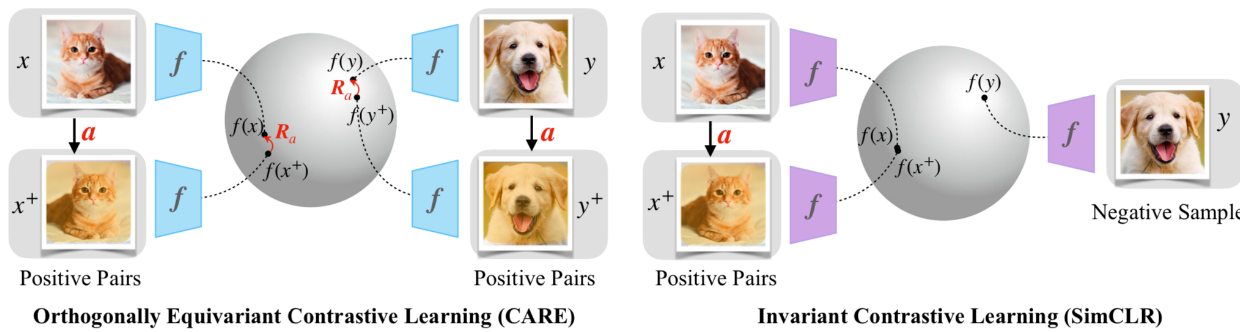
$$f(a(x'))^T f(a(x)) = f(x')^T f(x) \quad \forall x, x' \in X$$

$\iff$ there exists $R_a \in O(d)$ st $f(a(x)) = R_a f(x) \quad \forall x \in X$

Augmentations in X $\Rightarrow$ Orthogonal transformations in embedding space

Contrastive Augmentation-induced Rotation Equivariance (CARE)

Augmentations on inputs $\Longleftrightarrow \approx O(d)$ transformations in embeddings



$$\mathcal{L}_{CARE}(f) = \mathcal{L}_{inv}(f) + \mathcal{L}_{unif}(f) + \lambda \mathcal{L}_{equiv}(f)$$

$$\mathcal{L}_{inv}(f) = \mathbb{E}_{a,a' \sim A} \|f(a(x)) - f(a'(x))\|, \quad \mathcal{L}_{unif}(f) = \log \mathbb{E}_{x,x' \in \mathcal{X}} \exp(f(x)^T f(x'))$$

$$\mathcal{L}_{equi}(f) = \mathbb{E}_{a \sim A} \mathbb{E}_{x, x' \sim \mathcal{X}} \left[f(a(x'))^T f(a(x)) - f(x')^T f(x) \right]^2$$

PROP $\mathcal{L}_{equi}(f) = 0 \Longleftrightarrow$ for almost all $a \in A$ $\exists R_a \in O(d)$ st $f(a(x)) = R_a f(x)$

$\Rightarrow \rho : A \to O(d)$, $a \mapsto R_a$ is a group homomorphism $\Rightarrow A \leq O(d)$

Results :

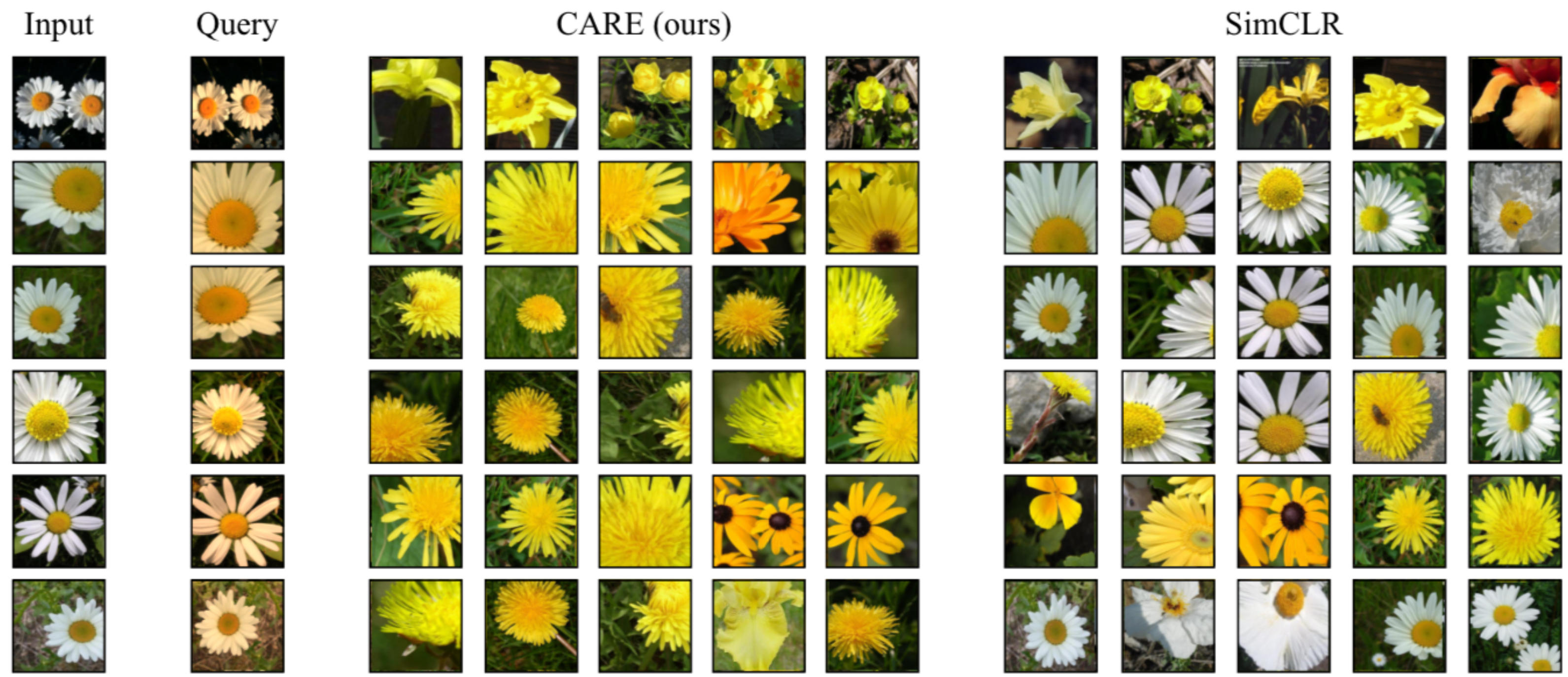


Figure 5: CARE exhibits sensitivity to features that invariance-based contrastive methods (e.g., SimCLR) do not. For each input we apply color jitter to produce the query image. We then retrieve the 5 nearest neighbors in the embedding space of CARE and SimCLR.

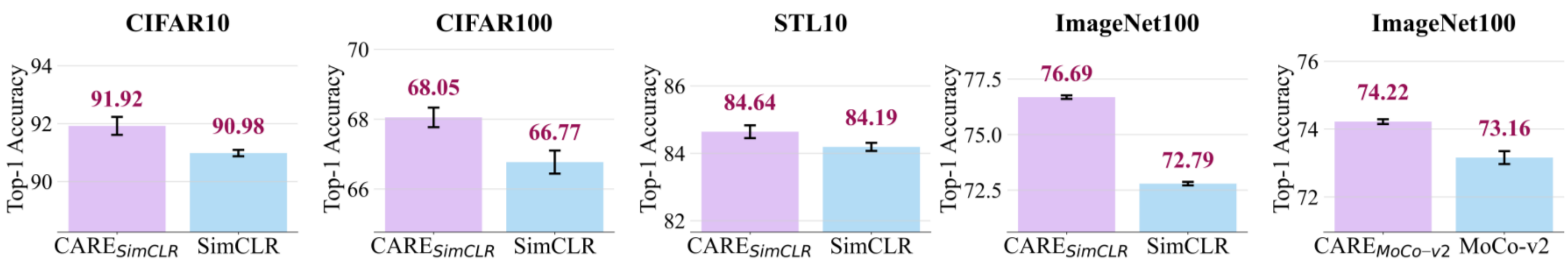
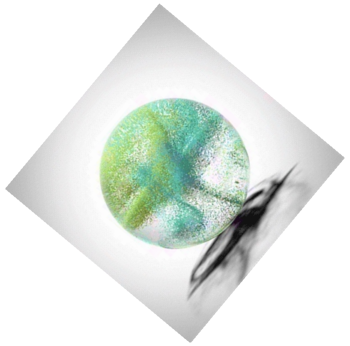


Figure 8: Top-1 linear readout accuracy (%) on CIFAR10, CIFAR100, STL10 and ImageNet100. All results are from 5 independent seed runs for the linear probe.

Extension

Invariant functions on point clouds

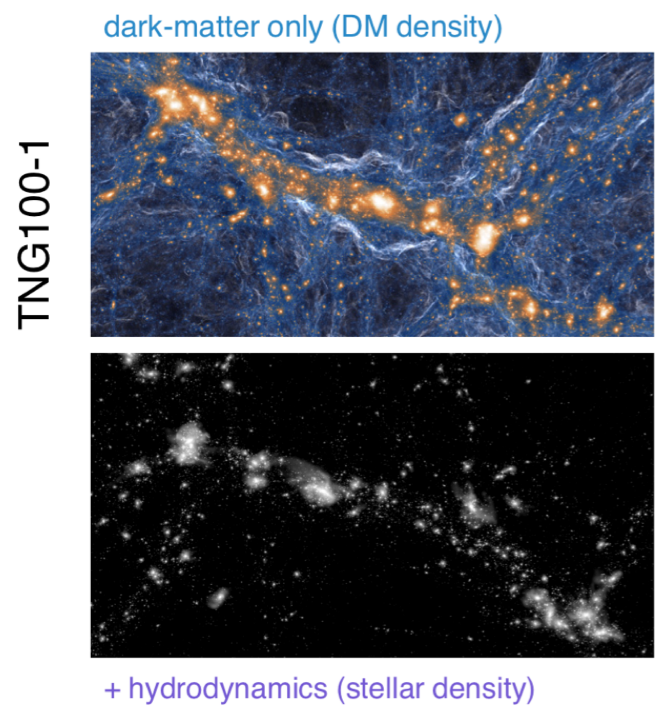
O(d) + permutation symmetry

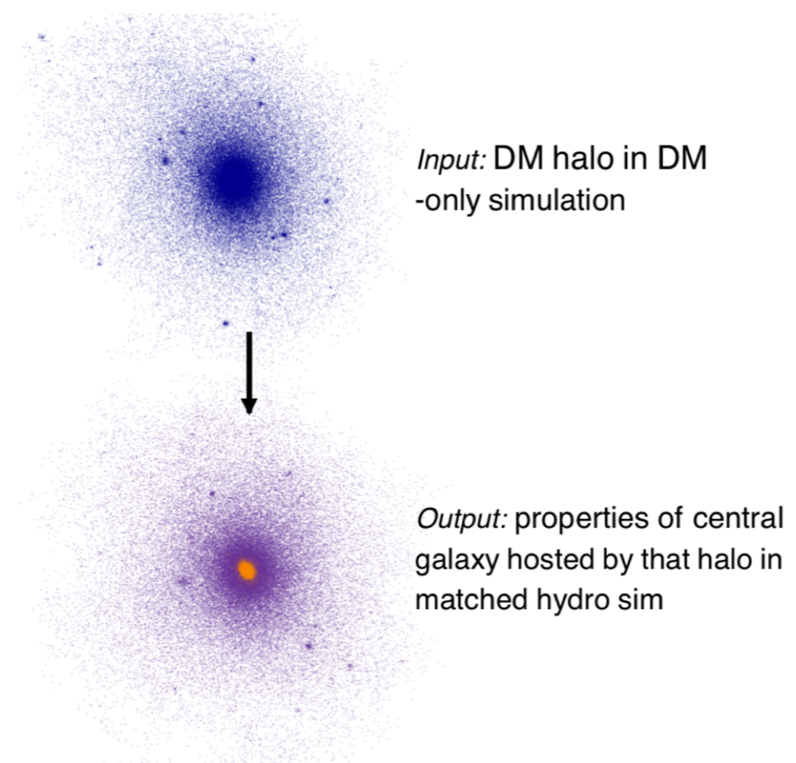


Kate Storey-Fisher

Motivation:

Emulation of cosmological simulations





Galaxy properties predictions from dark-matter only sims:

Storey-Fisher, Hogg, Genel, Hattori, V.

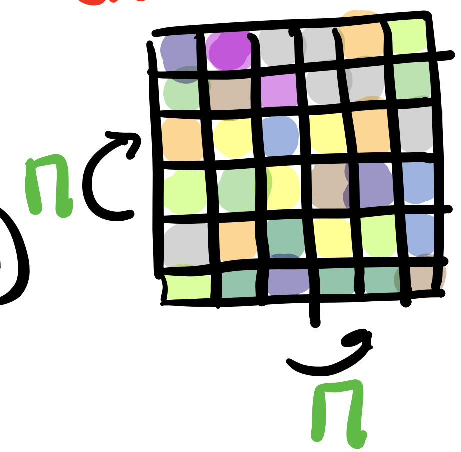
Invariant functions on point clouds

$$X = \begin{pmatrix} | & & | \\ p_1 & \cdots & p_n \\ | & & | \end{pmatrix} \in \mathbb{R}^{d \times n} \qquad O(d) \backslash X / S_n$$

$$f : X \to \mathbb{R} \quad O(d)\text{-invariant} \iff f(X) = \tilde{f}(\underbrace{X^T X}_{\text{Gram matrix}})$$

$$f(X\Pi) = f(X) \Rightarrow \tilde{f}(\Pi^T X^T X \Pi) \underset{\uparrow}{=} \tilde{f}(X^T X)$$

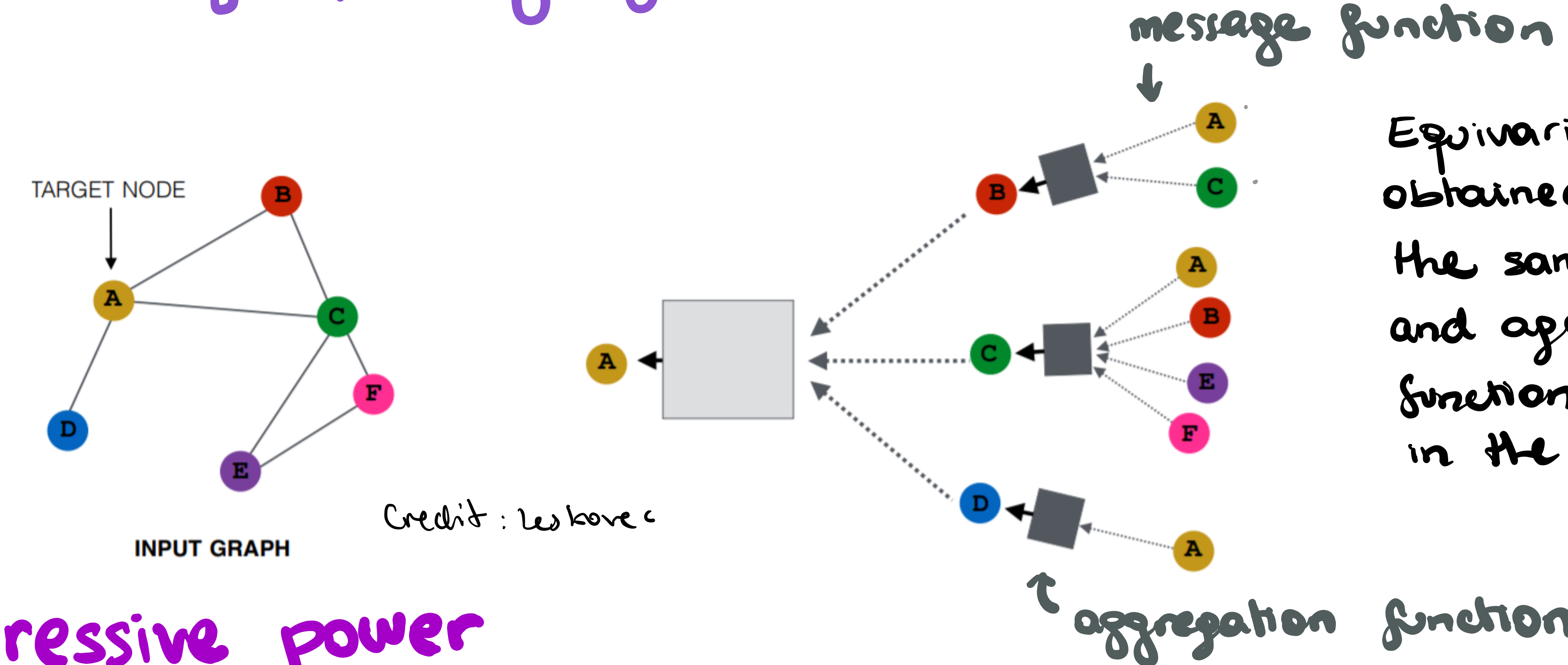
($\tilde{f}$ invariant by permutations acting by conjugation)



Symmetry of GNNs!

→ Current work: $n(d+1)$ generators for the field of invariant functions

Message passing graph neural networks

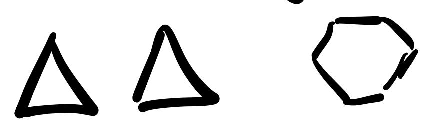


Credit : Leskovec

Equivariance is obtained by using the same message and aggregation function everywhere in the graph

Expressive power

- Not every invariant / equivariant function can be expressed this way example :

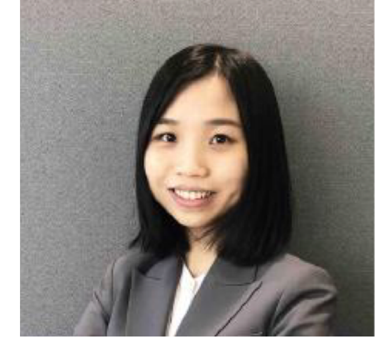


(Morris et al , Xu et al '19)

- Connections to graph isomorphism
- They cannot count substructures

Check out position paper by Morris et al '24 on theoretical directions !!

Our approach : Break the symmetries



Teresa Huang

Approximate equivariant graph networks

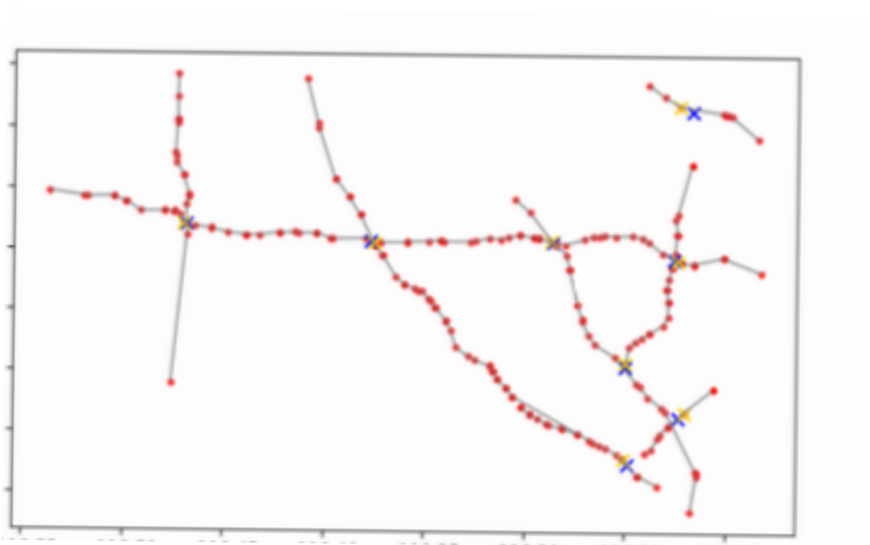
- When the graph is fixed decouple the action from the domain and the signal

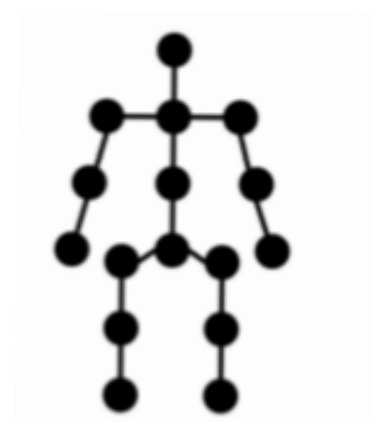
$$\pi G = (\pi A \pi^T, \pi X) \rightarrow \pi G = (A, \pi X)$$

$$\pi \in S_n \qquad\qquad \pi \in \mathcal{G} < S_n$$

$$H(\pi G) = \pi H(G) \qquad\qquad H(\pi G) \approx \pi H(G)$$

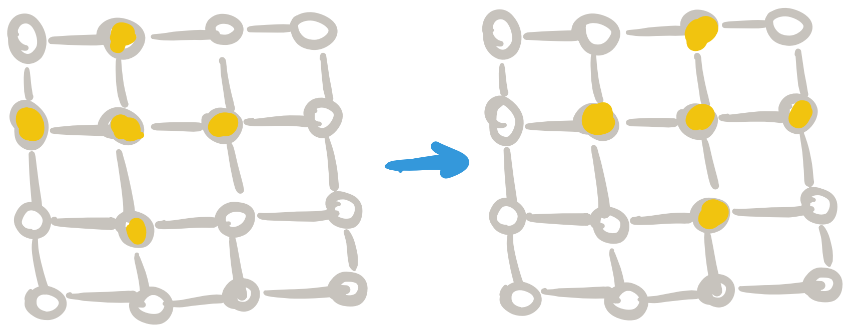
- Choose different subgroups $\mathcal{G}$





Intuition

CNN symmetry

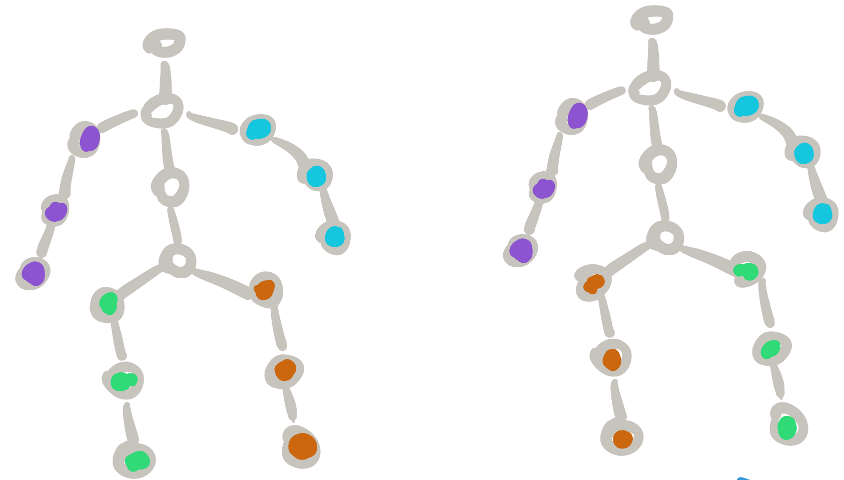


domain stays the same
signal (image) shifts

$X : \mathbb{Z}/n\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z} \to \mathbb{R}$

grid signal (image)

GNN symmetry



both domain (graph) and signal (color) move together (the object doesn't change)

$G = ([n], A)$ $\quad X : [n] \to \mathbb{R}^d$

nodes

adj matrix

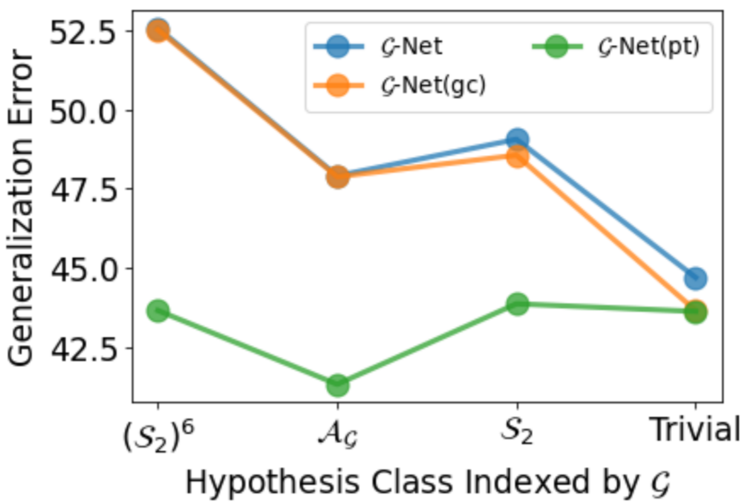
graph signal

What we do: Fix the graph and implement different symmetries on the signal

- How to implement the approximate symmetries?
  - → Cluster graph nodes and share functions on clusters
  - → Representation theory

Group $= (S_{c_1} \times \dots S_{c_k}) \rtimes A_G$

- Bias - variance tradeoff



Lemma $G \le S_n$ $\quad X \sim \mu$ $\quad S_n$-invariant

$$Y = f^*(X) + \xi \qquad f = \bar{f}_G + f_G^\perp$$

$$\underbrace{\Delta(f, \bar{f}_G)}_{\text{risk gap}} = \mathop{\mathbb{E}}_{X\sim\mu} \| Y - f(X) \|^2 - \mathop{\mathbb{E}}_{X\sim\mu} \| Y - \bar{f}_G(X) \|^2$$

$$= \underbrace{-2 \langle f^*, f_G^\perp \rangle_\mu}_{\text{mismatch}} + \underbrace{\| f_G^\perp \|_\mu^2}_{\text{constraint}}$$

Applications

- Explicit bias // variance tradeoff for linear regression with approx symmetry
- Approx guarantees for graph coarsening
- Examples showing imposing more symmetry may reduce the risk (bias ↑ variance ↓)

cf Elesedy & Zaidi ICML '21

Bias-variance for linear regression

$$\begin{tikzcd} \mathcal{X} \cong \mathbb{R}^N \arrow[r] \arrow[d, "\phi"'] & \mathcal{Y} \cong \mathbb{R}^N \arrow[d, "\psi"] \\ \mathcal{X} \arrow[r] & \mathcal{Y} \end{tikzcd}$$

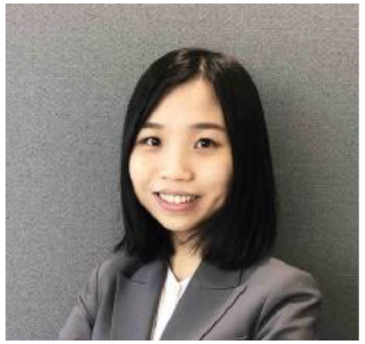
$$Y = X\theta + \eta$$

$$\mathbb{E}(\eta\eta^T) = \sigma^2$$

$$\hat{\theta} = \underset{\theta}{\arg\min} \; \|X\theta - Y\|^2$$

$\Psi_G(\hat{\theta})$ = projection of $\hat{\theta}$ onto equivariant maps

$$\Psi_G(\hat{\theta}) = \int_G \phi(g)\,\theta\,\psi(g)^{-1}\,dg$$



Teresa Huang

$$\mathbb{E}\left(\Delta\left(\hat{\theta}, \Psi_G(\hat{\theta})\right)\right) = \underbrace{-\sigma^2 \|\Psi_G^{\perp}(\theta)\|_F^2}_{\text{bias}^2} + \underbrace{\frac{\sigma^2 N^2 - (\chi_\psi | \chi_\phi)}{n - Nd - 1}}_{\text{variance}}$$

$$(\chi_\psi, \chi_\phi) = \int_G \chi_\psi(g)\,\chi_\phi(g)\,dg \leftarrow$$ dimension of space of linear equivariant functions

Numerical experiments

human pose estimation

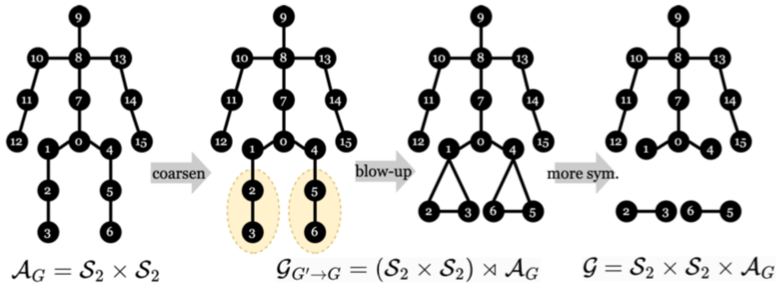
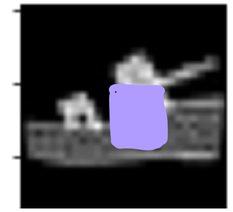
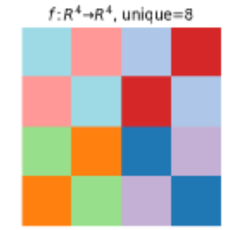


Figure 3: Human skeleton graph G, its coarsened graph G' (clustering leg joints), and blow-up of G'

$\mathcal{G}$-Net(gc+ew)	$\mathcal{S}_{16}$	$(\mathcal{S}_6)^2$	Relax-$\mathcal{S}_{16}$	$\mathcal{A}_G = (\mathcal{S}_2)^2$	Trivial
MPJPE ↓	42.55 ± 0.88	43.33 ± 0.99	$\mathbf{39.87 \pm 0.46}$	42.18 ± 0.49	41.60 ± 0.3
P-MPJPE ↓	34.48 ± 0.44	34.87 ± 0.48	$\mathbf{31.38 \pm 0.14}$	32.08 ± 0.20	31.69 ± 0.1

Image inpainting





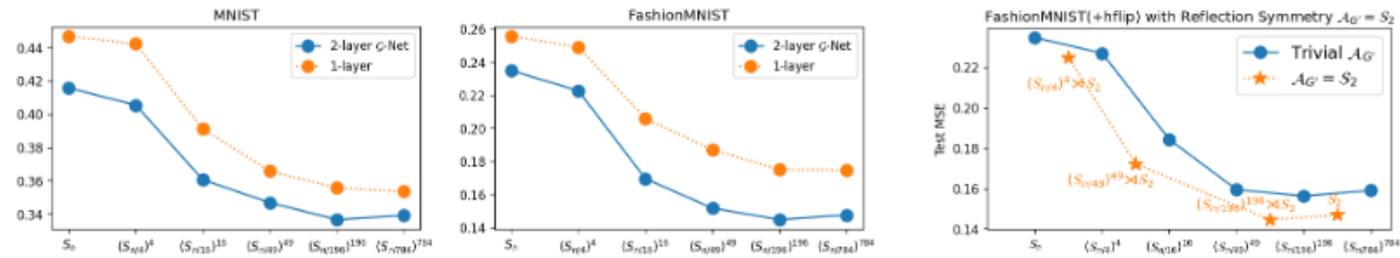
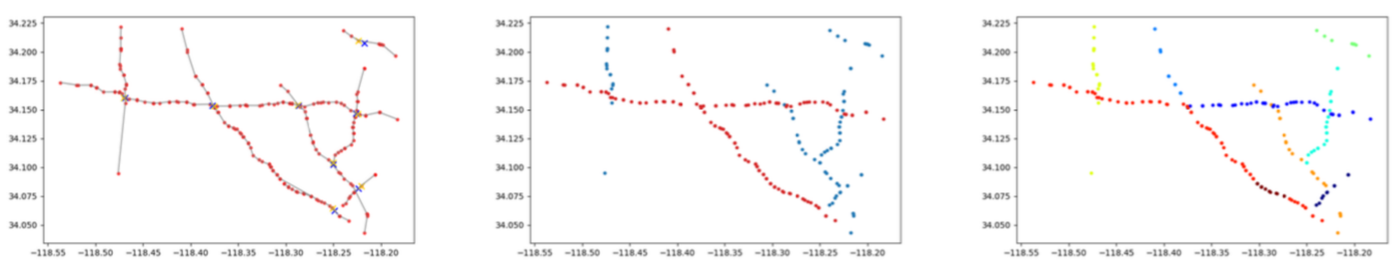


Figure 6: Bias-variance tradeoff via graph coarsening. Left:2-layer $\mathcal{G}$-Net(blue) and 1-layer linear $\mathcal{G}$-equivariant functions (orange), assuming the coarsened graph is asymmetric; Right: 2-layer $\mathcal{G}$-Net with both trivial and non-trivial coarsened graph symmetry.

MSE $(\times 1e^{-2})\downarrow$	$\mathcal{S}_{28^2} = \mathcal{S}_n$	$(\mathcal{S}_{14^2})^4 = (\mathcal{S}_{n/4})^4$	$(\mathcal{S}_{7^2})^{16} = (\mathcal{S}_{n/16})^{16}$	$(\mathcal{S}_{4^2})^{49} = (\mathcal{S}_{n/49})^{49}$	$(\mathcal{S}_{2^2})^{196} = (\mathcal{S}_{n/196})^{196}$	Trivial $= (\mathcal{S}_{n/784})^{784}$
MNIST	41.56 ± 0.16	40.53 ± 0.26	36.06 ± 0.24	34.68 ± 0.5	$\mathbf{33.67 \pm 0.07}$	33.92 ± 0.04
Fashion	23.48 ± 0.14	22.26 ± 0.02	16.94 ± 0.08	15.16 ± 0.1	$\mathbf{14.47 \pm 0.11}$	14.75 ± 0.11

Traffic flow prediction



(a) Our faithful traffic graph (b) Graph clustering (2 clusters) (c) Graph clustering (9 clusters)

$\mathcal{G}$-Net(gc)	$\mathcal{S}_N$	$\mathcal{S}_{c_1} \times \mathcal{S}_{c_2}$	$\mathcal{S}_{c_1} \times \ldots \times \mathcal{S}_{c_9}$
Graph G_s	3.173 ± 0.013	$\mathbf{3.150 \pm 0.008}$	3.204 ± 0.006
Graph G	3.106 ± 0.013	$\mathbf{3.092 \pm 0.008}$	3.174 ± 0.013

How much do we gain by imposing symmetries?

Elesedy Zaidi '21
see also: Petrache & Trivedi '23
Bietti et al '22
Tahmasebi & Jegelka '23
...

$G \curvearrowright \mathbb{R}^d$ compact group, $X \sim \mu$ supported in $\mathbb{R}^d$, μ G-invariant

Training data $(x_i, y_i = f^*(x_i) (+ \eta_i))$

f^*: invariant target; η_i: noise

$$\mathrm{Risk}(f) = \mathbb{E}_{x\sim\mu} \| f(x) - y \|^2$$

$$\Delta(f, \bar{f}) = \mathrm{Risk}(f) - \mathrm{Risk}(\bar{f}) = \| f^{\perp} \|_{\mu}^2$$

Δ: generalization gap

$\bar{f}$: proj of f onto space of invariant functions

key property $\bar{f}(x) = \int_{g\in G} f(g\cdot x)\, dg$

$= \arg\min_{h \text{ invariant}} \| f - h \|_{\mu}^2$

Not true for non-compact groups
what is the "right" notion of projection?

Note that equivariant ML doesn't perform any proj

Open problem: model to define "baseline" and quantify "gains"

Conclusions

- (Approximate) symmetries give a good inductive bias for ML
  → Physical sciences (cosmology)
  → Engineering (self-supervised learning)
- Tools: invariant theory, representation theory
- Rethinking the roles of symmetries as "model selection"
  - bias-variance tradeoff on graph learning by relaxing symmetries
- Future work
  - Coordinate free models on vector fields
    → Ocean dynamics
  - Interactions with differential geometry

Thank you!

Villar, Hogg, Storey-Fisher, Yao, Blum-Smith Neurips 2021

Yao, Hogg, Storey-Fisher, Villar Neurips ML 4 Physics 2021

Villar, Yao, Hogg, Blum-Smith, Dumitrascu JMLR, 2023

Blum-Smith, Villar Notices of the AMS

Villar, Hogg, Yao, Kevrekidis, Schölkopf TMLR 2024

Gupta, Robinson, Lim, Villar, Jegelka ICLR 2024

Gregory, Hogg, Blum-Smith, Arias, Wong, Villar arXiv 2023

Huang, Levie, Villar NeurIPS 2023

Elesedy, Zaidi ICML 2021

Questions?

Funding

NSF

SIMONS FOUNDATION

Office of Naval Research
ONR
Science & Technology

amazon