Machine learning and symmetries

Soledad Villar (Johns Hopkins University)

Based on joint work with:

Z. Chen, L. Chen, J. Bruna, D.W. Hogg, K. Storey-Fisher, W. Yao, B. Blum-Smith

Mathematics, physics and machine learning seminar
@ Instituto Superior Técnico, Lisboa

Symmetries in deep learning

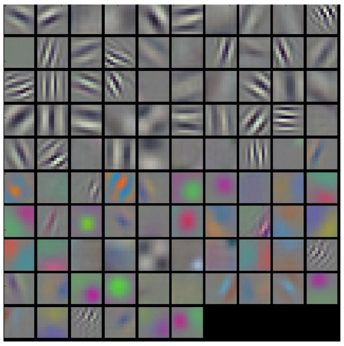
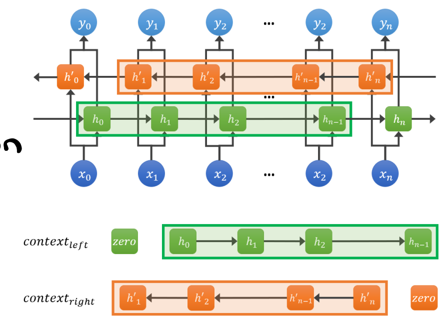


Image credit: Stanford CNN course

CNNs exploit translation and rotation symmetries in natural images by applying the same convolutional filters at different locations of the image

RNNs exploit time translation symmetry by applying the same recurrent unit at different locations

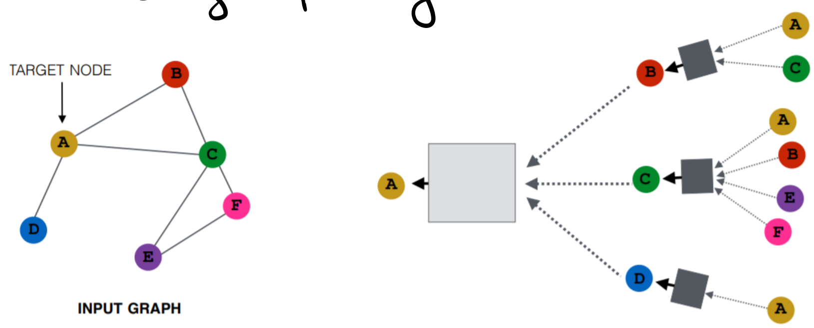


Credit: Zhao et al '19

Symmetries and graph neural networks

GNNs learn functions on graphs that are invariant to node relabeling. (Permutation invariance)
(equivariant) (equivariance)

Example: message passing neural networks (MPNN)



Credit : Leskovec

[Invariance / Equivariance implemented via weight sharing]

Invariance / Equivariance

Exact symmetries

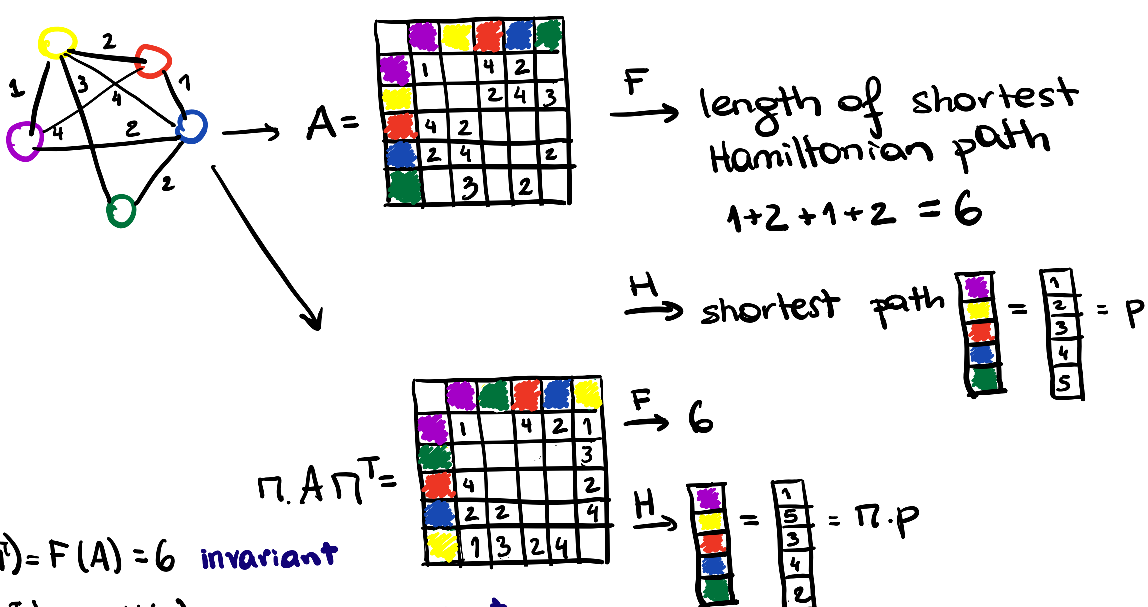
G a group acting on dataset X

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

If G also acts in Y

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

Example

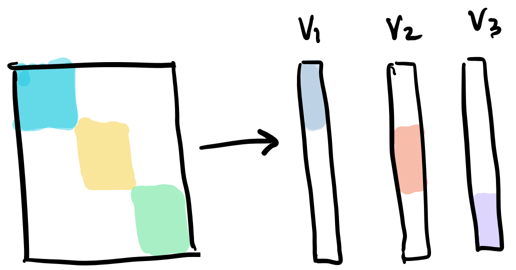


$F(\Pi A \Pi^T) = F(A) = 6$ invariant

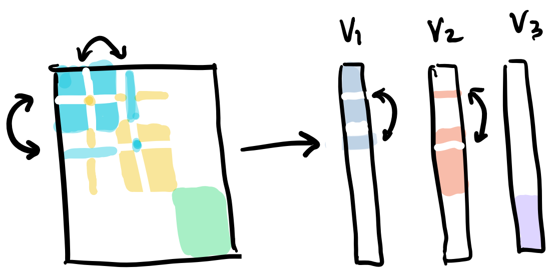
$H(\Pi A \Pi^T) = \Pi \cdot H(A) = \Pi \cdot p$ equivariant

Example (cont'd)

Spectral methods are permutation equivariant



This is why spectral clustering works



$f : \mathbb{R}^{n \times n} \to \mathbb{R}^n$

$f(\Pi A \Pi^T) = \Pi f(A)$

Π permutation matrix ($\Pi \in S_n$)

Spectral GNNs :

Given a graph G with adjacency A $(n \times n)$

Let $\mathcal{M} = \{I, D, A, A^2, A^3 \ldots\}$

Learn a "regularized spectral method" on $\Delta = \sum_{M \in \mathcal{M}} \alpha_M M$

unroll this to a GNN via power iteration $(v^{t+1} = \Delta v^t)$

$$v^{t+1} = \rho\left(\sum_{M \in \mathcal{M}} M v^t \alpha_M^t\right) \qquad \alpha_M^t \in \mathbb{R}^{d_t \times d_{t+1}}$$

$t = 1 \ldots T$ ← #LAYER

- Community detection : Chen, Li, Bruna '17
- Quadratic assignment : Nowak, V., Bandeira, Bruna '17
- Max-cut : Yao, Bandeira, V. '19

Symmetries in physics

Particular case of Physics-informed ML
Karniadakis et al. 21'

- **Motivation**
  - Laws of physics obey symmetries
  - Symmetric forms provide strong constraints on possible laws of physics
  - Symmetric forms and Einstein/Ricci summation notation enabled discovery of general relativity and various particle interactions

- What symmetries do we care about? $\{RR^T = I\} = O(d)$
  - Symmetries of classical physics $O(d), SO(d), E(d), O(1,d), IO(1,d), S_n$
  - Symmetries of quantum mechanics $U(1), U(2), SU(2), U(3), SU(3), C, P, T$

$O(d) = RR^T = I$
$O(d,1) = R\Lambda R^T = \Lambda$

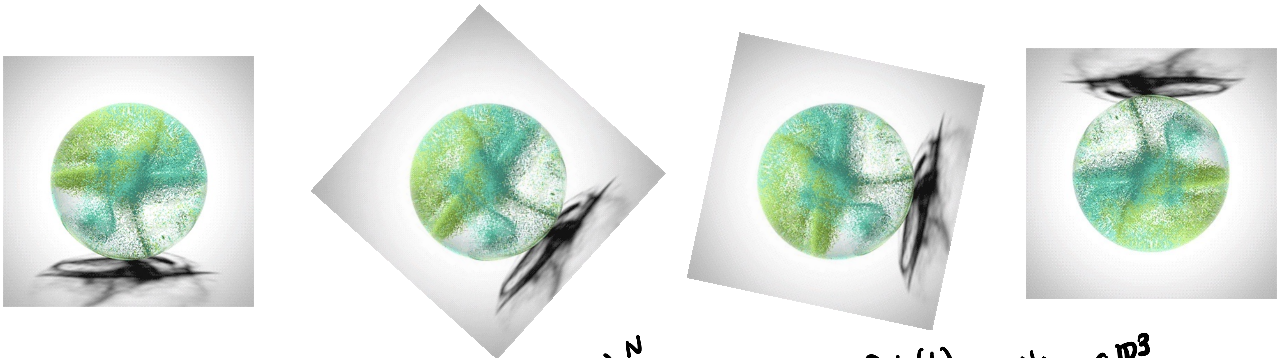
$\Lambda = \begin{pmatrix} -1 & & 0 \\ & 1 & \\ 0 & & 1 \end{pmatrix}$

- **Noether's theorem** [Emmy Noether 1915]

To every differentiable symmetry generated by local actions there corresponds a conservation law

Example: Rotation equivariance

- Dynamical systems, particle simulations (N-body problems)



Initial conditions $\left(q_i(0), p_i(0)\right)_{i=1}^{N}$

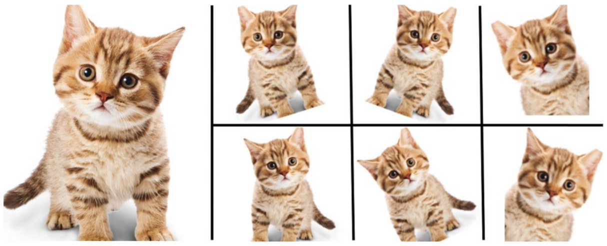
$q_i(t)$ position $\in \mathbb{R}^3$

$p_i(t)$ momentum $\in \mathbb{R}^3$

Predictions $\left(q_i(t_k), p_i(t_k)\right)_{i=1}^{N}$ $k=1,\dots T$

How are symmetries implemented?

- Data augmentation


Enlarge your Dataset
Credit: Bharath Raj

- Loss function penalties
- Architectural design
  - Approximate symmetries (CNN)
  - Exact symmetries
    - Weight sharing (message passing) Rose YU '21, '20. Weiler '21
    - Parameterization of symmetry preserving functions Kondor '18 Maron '18 Cohen '18
    - Symmetries as constraints Finzi et al '21
    - Irreducible representations Kondor, Thomas '18, Fuchs '20 Smidt
    - Steerable CNNs Cohen '17, Welling...

This talk : symmetries via architectural design

Pros :
- Better inductive bias
- Smaller generalization error
- Better sample complexity

Bietti et al. '21
Elesedy & Zaidy '21

Cons :
- Enforcing symmetries can be hard
- Trade off expressivity vs complexity

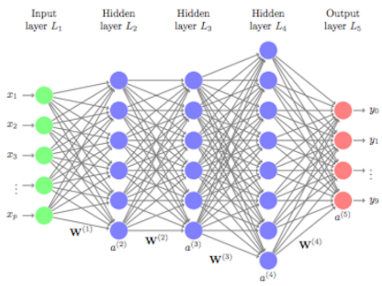
Typical approach: Equivariant architectures

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

Feed Forward NEURAL NETWORK

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

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



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

equivariant linear function

Maron '19

$\underbrace{v \otimes \dots \otimes v}_{v^{\otimes k_i}}$

$[\,]_x \rightarrow \rho(W_2 \rho(W_1 x)) = y$

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

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

θ: compatible activation function

Q: How to parametrize linear equivariant functions?

Finzi et al '21: solve a system of linear equations.

- Irreducible representation approach:

$\rho : G \to GL(V)$ group representation

$\rho_k : G \to GL(\mathbb{R}^{d^{\otimes k}})$ $\quad \rho_k \longleftrightarrow \rho_{k'}$

$\rho(g)(v) = g * v$ extend to tensor product $\rho_k = \bigotimes_{r=1}^{k} G \to GL((\mathbb{R}^d)^{\otimes k})$

Linear equivariant map $\underline{\underline{Li}} \longleftrightarrow$ map between representations

$\underbrace{Li \circ \rho_{ki}(g) = \rho_{ki+1}(g) \circ Li \quad \forall\, g \in G}_{\text{Easy to parameterize using IRREDUCIBLE REPRESENTATIONS}}$

$$\rho_{ki} = \bigoplus_{\ell=1}^{T_{ki}} J_\ell$$

$$\bigotimes_{s=1}^{K} \rho_s = \bigoplus_{\ell=1}^{T} T_\ell$$

This identification is given by the Clebsh-Gordan coefficients
- Known for SO(3) but not in general

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

Our work : characterization of expressivity for GNNs

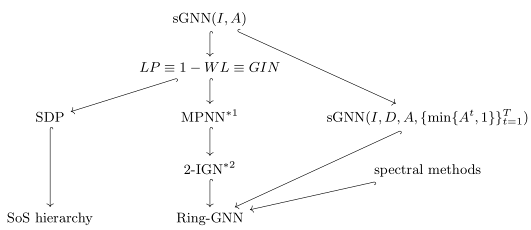
Z. Chen, V., L. Chen, J. Bruna NeurIPS 2019

Background : MPNNs cannot distinguish

(Xu et al '19)
(Morris et al '19)

$f(A) = f(B)$

Our main result: characterization of expressive power of GNNs based on ability to distinguish non-isomorphic graphs



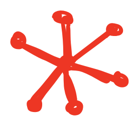
$f : \underset{\text{Adj}}{\mathbb{R}^{n\times n}} \times \boxed{\underset{\text{node features}}{\mathbb{R}^{n\times d}} \to \mathbb{R}}$

nets

$\left\{ \min_{X} \|AX - XB\|_1 \;\; \text{st} \;\; \begin{matrix} X\mathbf{1} = 1 \\ X^T\mathbf{1} = 1 \end{matrix} \right\}$

Extension (NeurIPS 2020): Can GNNs count substructures?

Answer : most architectures can only count star shape substructures



Scalars are universal: Equivariant machine learning structured like classical physics

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

NeurIPS 2021

Goal: Parameterize functions arising from physics, that obey classical physics symmetries: ROTATION, PARITY, BOOST, in any dimension

$f : (\mathbb{R}^d)^n \to \mathbb{R}$ invariant: $f(Q(v_1, \dots v_n)) = f(v_1, \dots v_n)$

f SCALAR INVARIANT FUNCTION

$h : (\mathbb{R}^d)^n \to \mathbb{R}^d$ equivariant: $h(Q(v_1, \dots v_n)) = Q \cdot h(v_1, \dots v_n)$

h VECTOR EQUIVARIANT FUNCTION

Groups and their actions on $(\mathbb{R}^d)^n$

Orthogonal	$\mathrm{O}(d) = \{Q \in \mathbb{R}^{d\times d} : Q^\top Q = Q\,Q^\top = I_d\}$,
Rotation	$\mathrm{SO}(d) = \{Q \in \mathbb{R}^{d\times d} : Q^\top Q = Q\,Q^\top = I_d,\ \det(Q) = 1\}$
Translation	$\mathrm{T}(d) = \{w \in \mathbb{R}^d\}$
Euclidean	$\mathrm{E}(d) = \mathrm{T}(d) \rtimes \mathrm{O}(d)$
Lorentz	$\mathrm{O}(1,d) = \{Q \in \mathbb{R}^{(d+1)\times(d+1)} : Q^\top \Lambda\, Q = \Lambda,\ \Lambda = \mathrm{diag}([1,-1,\ldots,-1])\}$
Poincaré	$\mathrm{IO}(1,d) = \mathrm{T}(d+1) \rtimes \mathrm{O}(1,d)$
Permutation	$\mathrm{S}_n = \{\sigma : [n] \to [n]$ bijective function$\}$

Table 1: **The groups considered in this work.**

Orthogonal; Lorentz	$Q \star (v_1,\ldots,v_n) = (Q\,v_1,\ldots,Q\,v_n)$
Translation	$w \star (v_1,\ldots,v_n) = (v_1 + w,\ldots,v_k + w, v_{k+1},\ldots,v_n)$ (where the first k vectors are position vectors)
Euclidean; Poincaré	$(w,Q) \star (v_1,\ldots,v_n) = (Q\,v_1 + w,\ldots,Q\,v_k + w, Q\,v_{k+1},\ldots,Q\,v_n)$
Permutation	$\sigma \star (v_1,\ldots,v_n) = (v_{\sigma(1)},\ldots,v_{\sigma(n)})$

Table 2: **The actions of the groups on vectors.** For the Euclidean group, the position vectors are positions of points; for the Poincaré group, the position vectors are positions of *events*.

Example: particle (m, r, v)

mass position velocity

$$\underset{\text{total mechanical energy}}{T} = \sum_{i=1}^{n} \frac{1}{2} m_i |v_i|^2 - \frac{1}{2} \sum_{i=1} \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{G m_i m_j}{|r_i - r_j|}$$

- O(d) invariant
- translation invariant

} ⇒ E(d) invariant

- permutation invariant

Our approach

Characterization of invariant functions:

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

Also true for Lorentz

- $f(v_1,\dots v_n) = \tilde{f}\left((v_i^T v_j)_{i,j=1}^{n} \right)$

Proof • $V = \begin{pmatrix} | & & | \\ v_1 & \dots & v_n \\ | & & | \end{pmatrix}$, $M = V^T V$. Consider the Cholesky decomposition of $M = L^T L$ then $L = V \cdot Q$ for some $Q \in O(d)$. In words you can recover $v_1 \dots v_n$ from the inner products up to orthogonal transformations.

- Physics point of view: All scalars can be written in Einstein summation notation

Lorentz : $\langle (t_1,x_1),(t_2,x_2) \rangle = t_1 t_2 - \langle x_1, x_2 \rangle$

Do we need all the inner products?

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

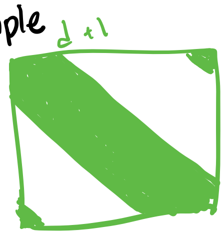
$f(Qv_1, \dots Qv_n) = f(v_1 \dots v_n)$

f is invariant if and only if $f(\underbrace{v_1 \dots v_n}_{n \text{ d-vectors}}) = \tilde{f}(\underbrace{\langle v_i, v_j \rangle_{i,j=1}^n}_{n \times n \text{ scalars}})$

Do we need all $n \times n$ scalars? NO:

- Rigity theory of Gram matrices
- Low rank matrix completion

Example

$\hat{M} =$ 

determines $M = V^T V$

$f(v_1, \dots v_n) = \tilde{f}(M) = \hat{f}(\hat{M})$

Equivariant vector functions:

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

$$h(v_1 \dots v_n) = \sum_{i=1}^{n} \underbrace{f_i(v_1 \dots v_n)}_{\text{O(d) invariant scalar function (Lorentz)}} \cdot v_i$$

O(d) invariant scalar function
(Lorentz)

d is any dimension!!

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

$h(v_1, v_2) = \underline{v_1 \times v_2}$

Proof (sketch)

- h O(d)-equivariant $\Rightarrow h(v_1 \dots v_n) \in \mathrm{span}(v_1 \dots v_n)$
- $h(v_1, \dots v_n) = \sum_{i=1}^{n} f_i(v_1 \dots v_n)\, v_i$ coefficients functions can taken to be invariant
- If h polynomial $\Rightarrow f_i$'s can be chosen to be polynomials

↓ EXAMPLE

APPROACH: input → scalars → f_1 … f_n (MLP) → ⊕ → output (UNIVERSAL)

Example: Electromagnetic force law.

Particle (q, r, v)
charge position velocity

$$F = \underbrace{\sum_{i=1}^{n} k\, q\, q_i \frac{(r - r_i)}{|r - r_i|^3}}_{\text{electrostatic force}} + \underbrace{\sum_{i=1}^{n} k\, q\, q_i \frac{v \times (v_i \times (r - r_i))}{c^2\, |r - r_i|^3}}_{\text{magnetic force}}$$

⬇ using $a \times (b \times c) = (a^T c) b - (a^T b) c$

$$F = \sum_{i=1}^{n} k\, q\, q_i \frac{(r - r_i)}{|r - r_i|^3} + \sum_{i=1}^{n} k\, q\, q_i \frac{(v^\top (r - r_i))\, v_i - (v^\top v_i)\,(r - r_i)}{c^2\, |r - r_i|^3}$$

$$= \sum_{i=1}^{n} k\, q\, q_i \left(1 - \frac{v^\top v_i}{c^2}\right) \frac{(r - r_i)}{|r - r_i|^3} + \sum_{i=1}^{n} k\, q\, q_i \frac{(v^\top (r - r_i))\, v_i}{c^2\, |r - r_i|^3},$$

* Not a field formulation anymore

Translations and permutations

Euclidean group: includes translation symmetry
(also Poincaré)

$$h(v_1, \dots, v_n) = \tilde{h}(v_1 - v, \dots, v_n - v) \quad O(d)\text{-invariant}$$

where v is the center of mass $\frac{1}{n}\sum v_i$ or any weighted mean position

Permutation invariance: If h is $O(d)$-equivariant (or Lorentz) and

$$h(v_1 \dots v_n) = h(v_{\sigma(1)} \dots v_{\sigma(n)}) \quad \sigma \in S_n \text{ (permutation)}$$

$$h(v_1 \dots v_n) = \sum_{i=1}^{n} \underbrace{f(v_i, v_{[-i]})}_{O(d)\text{-invariant (or Lorentz)}} \cdot v_i$$

perm inv wrt n-1 last inputs

$\{v_1 \dots v_{i-1}, v_{i+1}, \dots v_n\}$

- Easy to implement with message passing graph neural networks

Parameterization for general groups?

Gripaios

- G reductive group over $\mathbb{C}$ (or $\mathbb{R}$)
- The algebra of invariant polynomials P is a graded Cohen-Macaulay algebra

ie: $\exists$ $P = \mathbb{C}[f_1, \dots f_n]$ where $f_1 \dots f_n$ - homogeneous
- algebraically independent
- elements of A

where A finitely generated free module over P

ie: $x = p_1(f_1 \dots f_n) g_1 + \dots + p_m(f_1 \dots f_n) g_m$

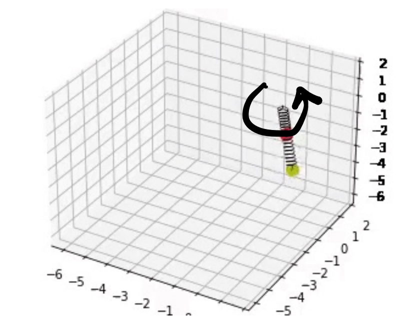
$f_1 \dots f_n$ "primary invariants"

$g_1 \dots g_m$ "secondary invariants" basis of A as a P-module

Example: $O(d)$ $d > n$ $f_1 \dots f_m$ = scalar products
$g_1 \dots g_m = 1$

Ongoing work with B. Blum-Smith parameterization for more general groups.

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

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

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

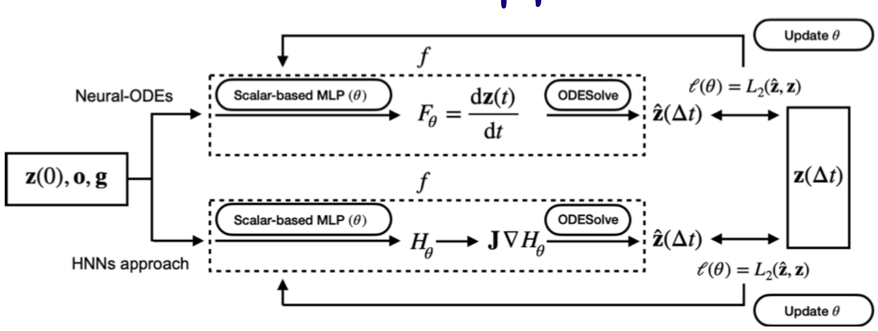
$H = KE + PE$ conserved quantity $\longleftrightarrow$ time translation symmetry

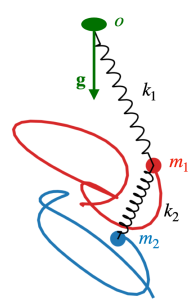
$$F: (\mathbb{R}^3)^5 \times \mathbb{R} \to (\mathbb{R}^3)^4$$

O(3)-equivariant

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

Computational approaches:



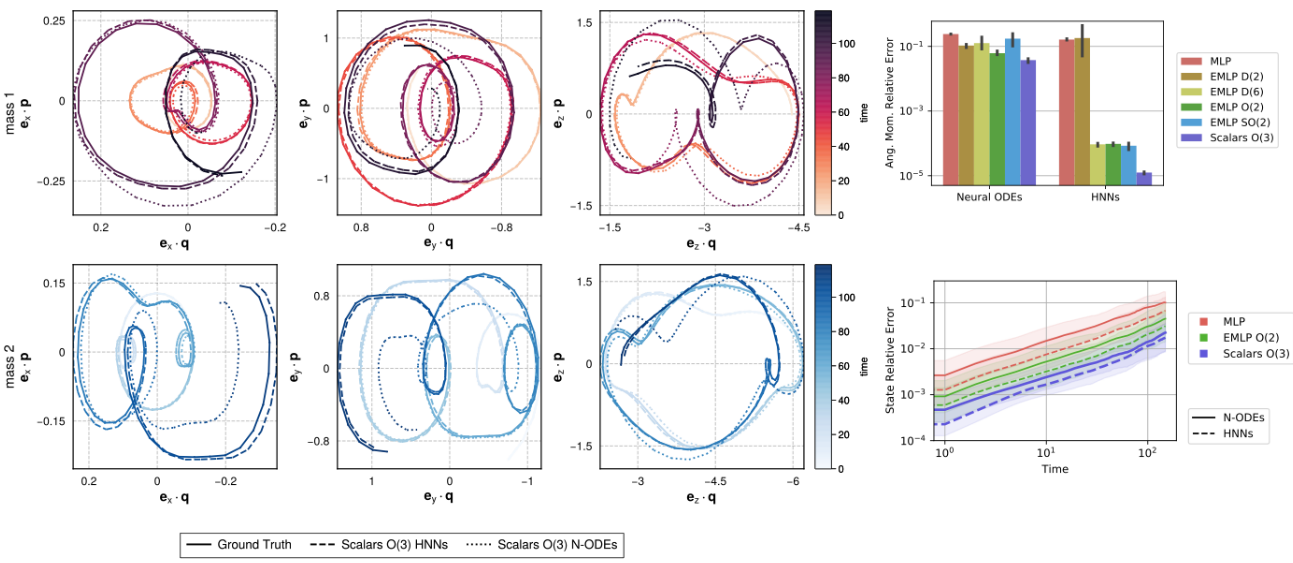


- Neural ODEs
- Hamiltonian neural network (HNNs)

$(q_1, p_1, q_2, p_2) \to$ Energy (H)

Results

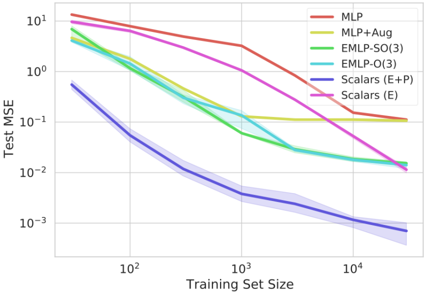
	Scalars O(3)	EMLP				MLP
		O(2)	SO(2)	D_2	D_6	
N-ODEs:	**.009 ± .001**	.020 ± .002	.051 ± .036	.023 ± .002	.036 ± .025	.048 ± .000
HNNs:	**.005 ± .002**	.012 ± .002	.016 ± .003	.111 ± .167	.013 ± .002	.028 ± .001



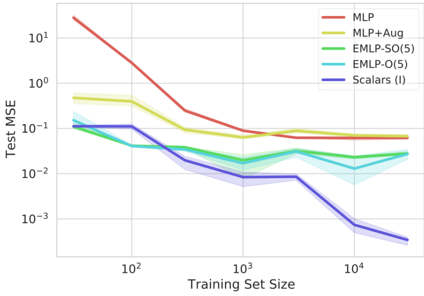
numerical performance - Sample complexity

comparison to Finzi et al. 21'

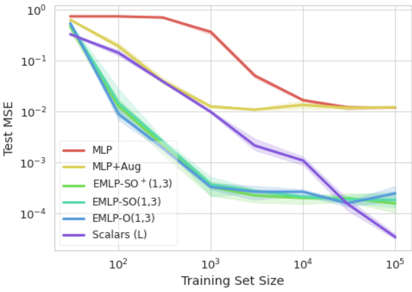
Simple O(3)-equivariant task + permutation invariant



Simple O(5)- invariant task



Simple Lorentz-equivariant



$$h(\{x_i\}_{i=1}^{5}) = \sum_{i=1}^{5} m_i (x_i^T x_i I - x_i x_i^T)$$

$$f(x_1, x_2) = \sin(\|x_2\|) - \frac{\|x_2\|^3}{2} + \frac{x_1^T x_2}{\|x_1\| \|x_2\|}$$

Summary

GOAL:

Enforcing exact symmetries in machine learning models

- Better sample complexity
- Smaller generalization error

GNNs (permutation equivariance)

- Characterization of expressive power of GNNs via graph isomorphism

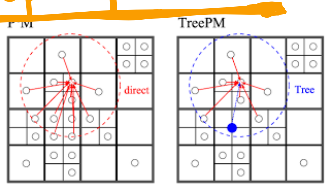
Symmetries in classical physics

- We provide a simple characterization of all equivariant functions wrt physically relevant group actions

universal approximation ↔ (based on Einstein summation notation & classical invariant theory)

Open problems

- Design a subset of permutation-invariant scalars that are universally expressive
- Explore connections with matrix completion
- Incorporate multi-scale information (FMM, k-d tree)



Source: Ishiyama et al 21'

- Dimensions and units symmetry
- Generalization bounds
- Extension to general groups

- Application domains (ongoing work)

- Computational biology (collaboration with Bianca Dumitrascu, Cambridge University, UK)
- Astrophysics (collaboration with David Hogg, Flatiron, NYU)
  - Benchmark dataset for ML for dynamics.

Thank you!

- Chen, Villar, Chen, Bruna
  NeurIPS 2019
- Chen, Chen, Villar, Bruna
  NeurIPS 2020
- Villar, Hogg, Storey-Fischer, Yao, Blum-Smith
  NeurIPS 2021
- Yao, Storey-Fischer, Hogg, Villar
  NeurIPS workshop
  ML for physics 2021

Quantum many body systems

Points $\{\pm 1\}$ magnets SK-model Ising model

+1 -1 -1 +1 -1

Hamiltonian $\sum_{\text{neighbors}}$ interactions

→ translation symmetry
(indistinguishability)

⟹ Changing coupling become NP-hard
(change objective function)

learning task → learning $\overbrace{\text{ground state}}^{\text{Not symmetry}}$ for a given objective function

→ interested in local properties on ground state which are invariant with respect to certain permutations

- Chemistry & material science

$$f(x) = \int_{\sigma \in G} \tilde{f}(\sigma x)\, d\sigma$$

$\boxed{\text{Convolution}}$

input $x(u)$ $u \in \mathbb{R}^3$ $f : B^{\mathbb{R}^3} \to \mathbb{R}$

1st layer $x * T_w f = x^{(1)}(w)$ $w \in G$
rotated copy of my function

$$\int_{u \in \mathbb{R}^3} T_w f^{(1)}(v)\, x(u+v)\, dv = x^{(1)}(u, w)$$

$x^{(1)} : \mathbb{R}^3 \times G \to \mathbb{R}$

2nd layer define $T_\rho x(u, w) = x(\rho u, \rho^{-1} w)$ $\rho \in G$

$f^{(2)} : \mathbb{R}^3 \times G \to \mathbb{R}$

//

$$x^{(1)} \circledast T_w f^{(2)} = x^{(2)}$$

$$\int_{v \in \mathbb{R}^3} \int_{\rho \in G}$$

universal?