Unification of Symmetries Inside Neural Networks: Transformer, Feedforward and Neural ODE

Koji Hashimoto (Kyoto U)

ArXiv:2402.02362 (w/ Yuji Hirono, Akiyoshi Sannai)

Unification of Symmetries Inside Neural Networks

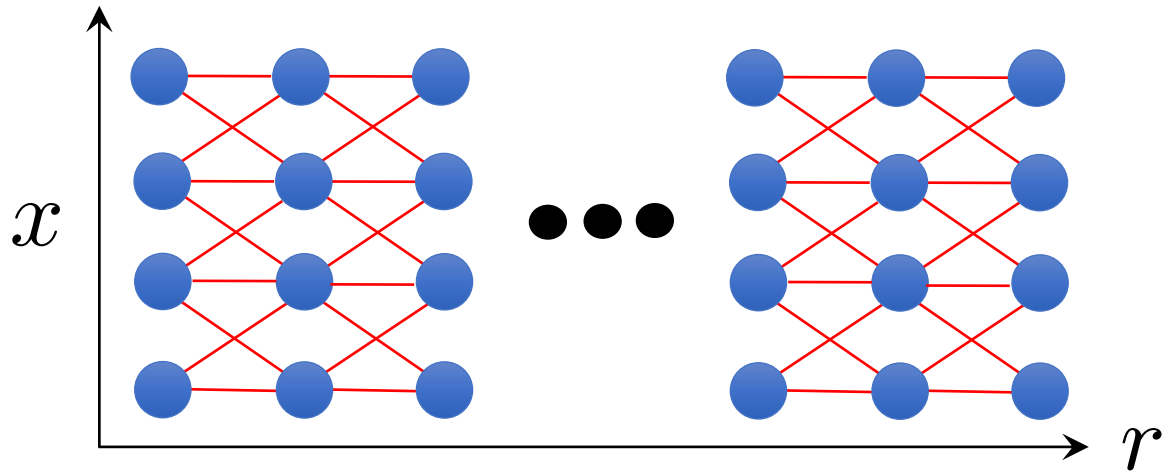
[Hirono, Sannai, KH 2402.02362]

1. Motivation: gauge sym in NN — 2 pages
2. Candidates for diffeo in NN — 4 pages
3. Diffeo in neural ODEs — 4 pages
4. Physics of NN symmetries? — 3 pages

1. Motivation: gauge sym in NN

There should be diffeo, a la AdS/DL

Deep feedforward NN = AdS spacetime



Bulk reconstruction by entanglement [Lam You] [You Yang Qi] …

Bulk reconstruction by QFT correlators

[Tanaka Tomiya Sughishita KH] [Akutagawa Sumimoto KH] [Hu You KH] [Tan Chen] [Song Oh Ahn Kim] [Yan Wu Ge Tian] …

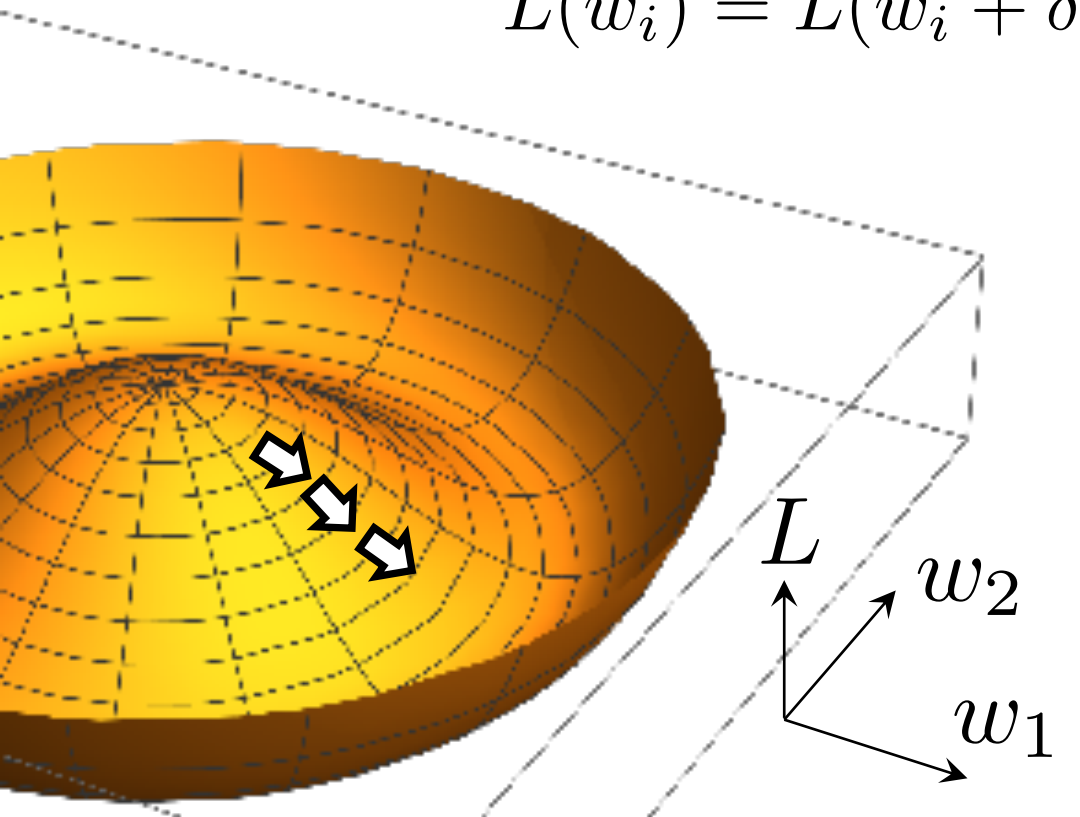
Deep Boltzmann machine = AdS/CFT [KH]

1. Motivation: gauge sym in NN

Symmetries = redundancy → trainability

NN symmetry : Invariance of loss function $L(w_i)$ under a transformation on weights w_i

$$L(w_i) = L(w_i + \delta_i(w))$$



Learning dynamics may depend on symmetries.

[Amari] [Amari Ozeki Karakida Yoshida Okada]
[Badrinarayanan Mishra Cipolla]
[Neyshabur Salakhutdinov Srebro]
[Ziyin 2023] …

Unification of Symmetries Inside Neural Networks

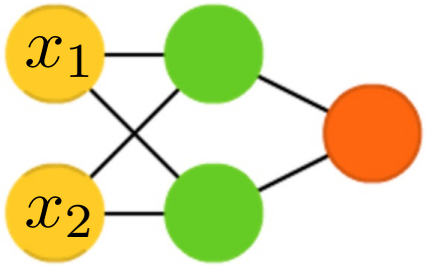
[Hirono, Sannai, KH 2402.02362]

1. Motivation: gauge sym in NN — 2 pages
2. Candidates for diffeo in NN — 4 pages
3. Diffeo in neural ODEs — 4 pages
4. Physics of NN symmetries? — 3 pages

2. Candidates for diffeo in NN

Review of Feedforward NN

Perceptron model



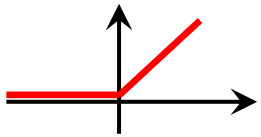
$W_i^{(2)}\varphi(W_{ij}^{(1)}x_j)$

"Unit" (circles) : Vector components

"Weight" W (lines) : Linear transformation to be optimized

"Activation function" φ (hidden line-end) : Nonlinear component-wise transf.

$$\mathrm{ReLU}(x) = x\,\theta(x)$$

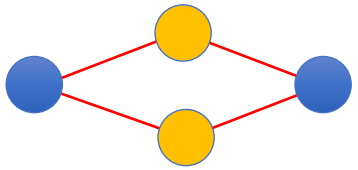


- Training protocol :
  1) Prepare many sets $\{(x_j, f)\}$: input and output
  2) Train the network (adjust W) by lowering

"Loss function" $E \equiv \sum_{\text{data}} \left| f - W_i^{(2)}\varphi\left(W_{ij}^{(1)}x_j\right)\right|$

2. Candidates for diffeo in NN

1. Permutation symmetry



$$f(x) = w_1\varphi(\tilde{w}_1 x + \tilde{b}_1) + w_2\varphi(\tilde{w}_2 + \tilde{b}_2) + b$$

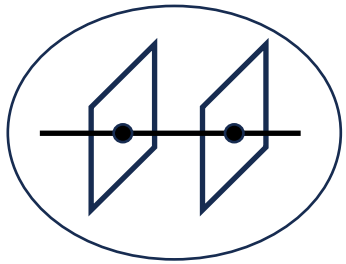
Multi-Layer
Perceptron

NN symmetry : swapping of neurons

$$w_1 \leftrightarrow w_2, \quad \tilde{w}_1 \leftrightarrow \tilde{w}_2, \quad \tilde{b}_1 \leftrightarrow \tilde{b}_2$$

Note : this Z_2 sym enhances at singularity

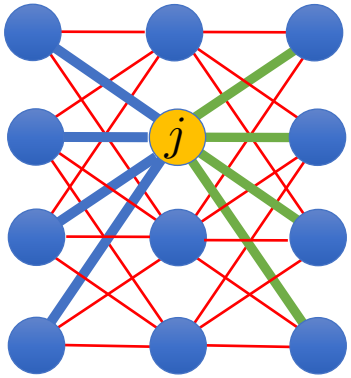
[Amari Ozeki Karakida Yoshida Okada 2016]



2. Candidates for diffeo in NN

2. Rescaling symmetry

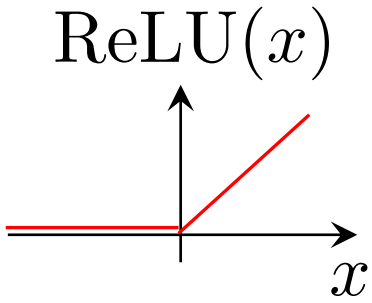
x_k



$$f_i = w_{ij}\varphi(\tilde{w}_{jk}x_k + \tilde{b}_j) + b_i$$

ReLU activation

$\varphi(x) = \mathrm{ReLU}(x)$



NN symmetry : For any fixed j, rescale

$$w_{ij} \mapsto \alpha w_{ij}, \quad \tilde{w}_{jk} \mapsto \alpha^{-1}\tilde{w}_{jk}, \quad \tilde{b}_j \mapsto \alpha^{-1}\tilde{b}_j$$

due to ReLU scaling property

$$\mathrm{ReLU}(\alpha^{-1}x) = \alpha^{-1}\mathrm{ReLU}(x)$$

2. Candidates for diffeo in NN

3. Self-attention in transformers

$$h_i = \sum_{j=1}^{n} \mathrm{ReLU}\left((xw^{(q)})_i (xw^{(k)})_j^{\mathrm{T}} \right) (xw^{(v)})_j$$

$x \in \mathbf{R}^{n\times d}$: set of data $x_i \in \mathbf{R}^d (i = 1, 2, \cdots, n)$

$w^{(q),(k),(v)} \in \mathbf{R}^{d\times d}$: query, key and value weights

NN sym 1 : rescaling $w^{(a)} \mapsto \alpha^{(a)} w^{(a)}, \quad \alpha^{(q)}\alpha^{(k)}\alpha^{(v)} = 1$

NN sym 2 : Internal sym

$$w^{(q)} \mapsto w^{(q)} A, \quad w^{(k)} \mapsto w^{(k)} (A^{-1})^T, \quad A \in SL(d, \mathbf{R})$$

Unification of Symmetries Inside Neural Networks

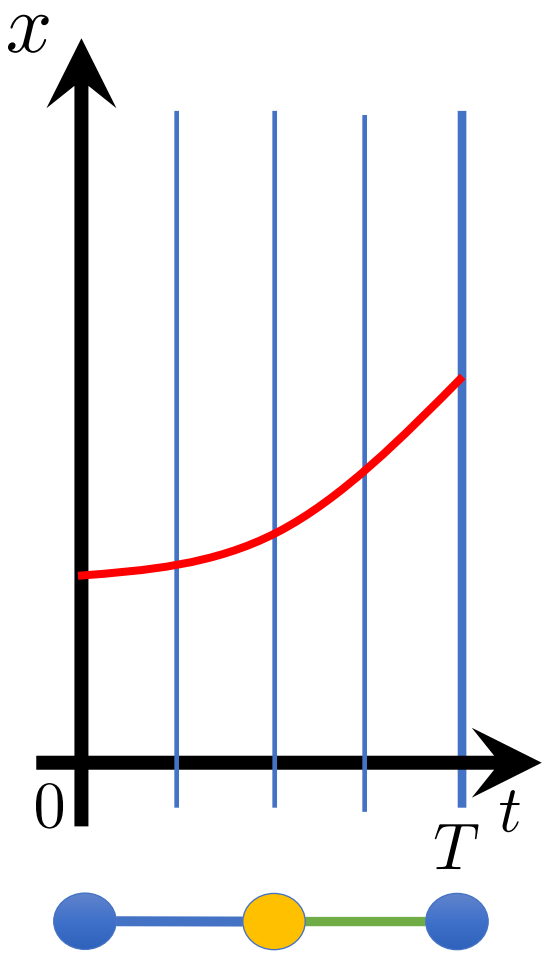
[Hirono, Sannai, KH 2402.02362]

1. Motivation: gauge sym in NN — 2 pages
2. Candidates for diffeo in NN — 4 pages
3. Diffeo in neural ODEs — 4 pages
4. Physics of NN symmetries? — 3 pages

3. Diffeos in neural ODEs

Neural ODEs = continuous ver. of NN

Chen, Rubanova, Bettencourt, Duvenaud
ArXiv:1806.07366 [cs.LG]



$$\dot{x}(t) = F(t, x(t))$$

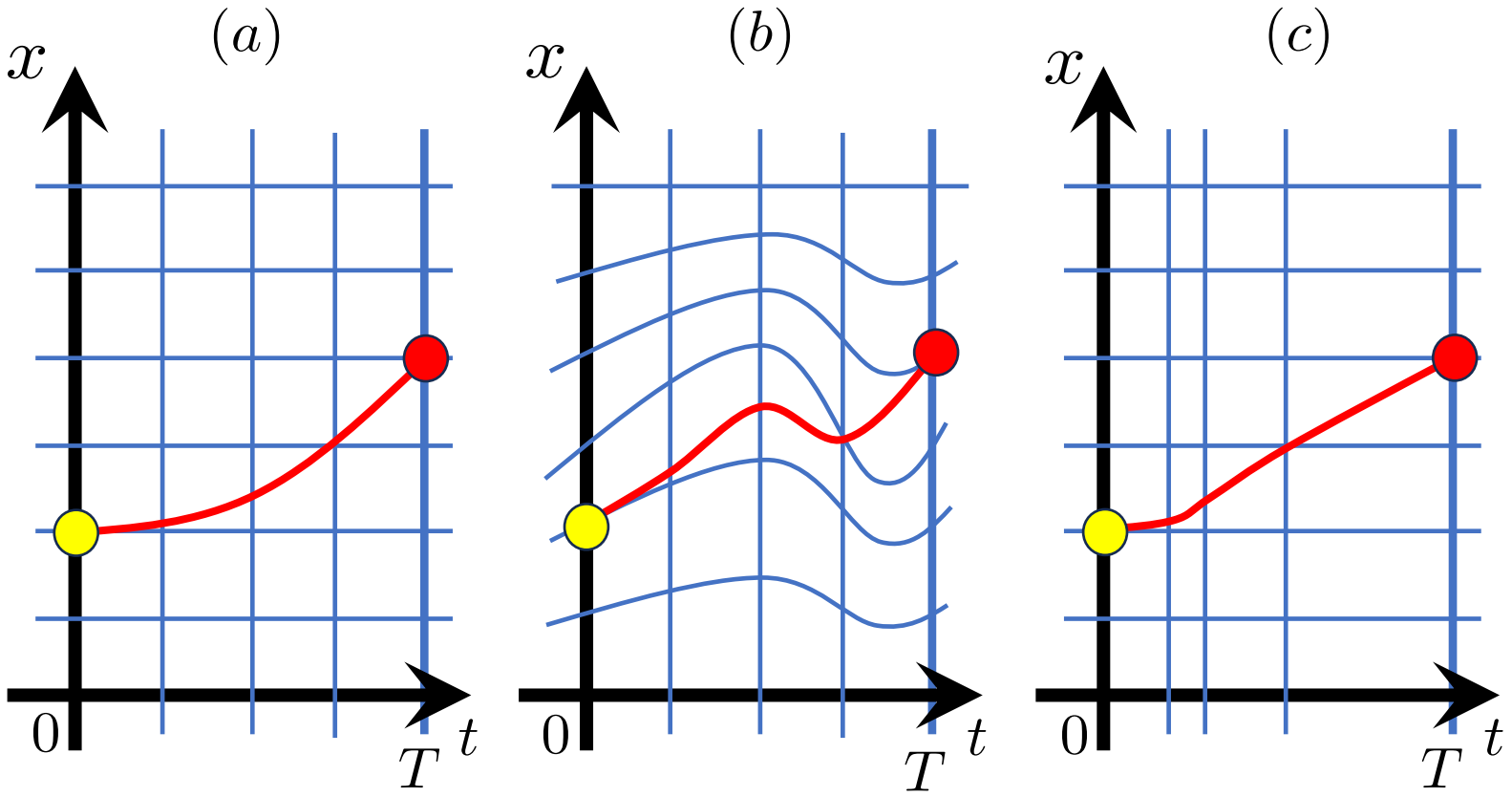
Training : train $F(t, x)$ such that the relation $(x(0), x(T))$ reproduces given set of data $\{(x_i, x_f)\}$.

Discretizing it gives a residual NN :

$$x(t + \delta t) = x(t) + F(t, x(t))\delta t$$

3. Diffeos in neural ODEs

Diffeos in ADM-like decomposition in neural ODE



Original frame

Spatial diffeo

$x(t) \mapsto x(t) + \epsilon(t, x(t))$

Time reparam.

$t \mapsto t + f(t)$

3. Diffeos in neural ODEs

Linear neural ODE can be solved

Linear neural ODE : $\dot{x}(t) = w(t)x(t) + b(t)$

Explicit solution of the ODE

$$x(T) = e^{\int_0^T w(t')dt'} x(0) + \left(\int_0^T e^{-\int_0^{t'} w(t'')dt''} b(t')dt' \right) e^{\int_0^T w(t'')dt''}$$

"Wilson loop"

Spatial diffeo $x(t) \mapsto x(t) - g(t)x(t)$ is weight transf.

$$w(t) \mapsto w(t) + \dot{g}(t), \quad b(t) \mapsto e^{g(t)-g(t=0)}\, b(t)$$

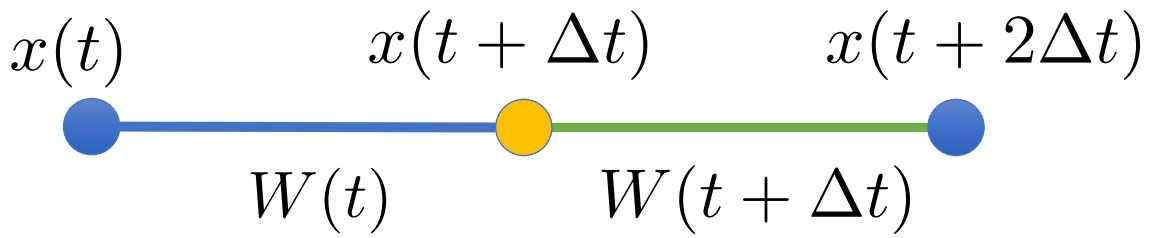
"Gauge field" "Higgs field"

Time reparam. $t \mapsto t + f(t)$ is weight transf.

$$w(t) \mapsto w(t) - \frac{d}{dt}(w(t)f(t)), \quad b(t) \mapsto b(t) - \frac{d}{dt}(b(t)f(t))$$

3. Diffeos in neural ODEs

Rescaling = a spatial diffeo



Spatial diffeo : Integrated weight transforms as Wilson line

$$W(t) \mapsto e^{-g(t)} W(t) e^{g(t+\Delta t)}$$

$$W(t+\Delta t) \mapsto e^{-g(t+\Delta t)} W(t) e^{g(t+2\Delta t)}$$

which reproduces the rescaling symmetry

$$w_{ij} \mapsto \alpha w_{ij}, \quad \tilde{w}_{jk} \mapsto \alpha^{-1} \tilde{w}_{jk}$$

Unification of Symmetries Inside Neural Networks

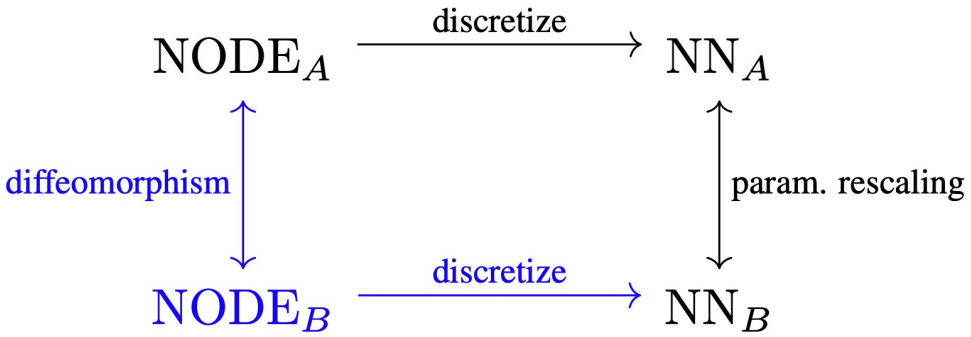
[Hirono, Sannai, KH 2402.02362]

1. Motivation: gauge sym in NN — 2 pages
2. Candidates for diffeo in NN — 4 pages
3. Diffeo in neural ODEs — 4 pages
4. Physics of NN symmetries? — 3 pages

4. Physics of NN symmetries?

NN gauge symmetries are broken in general

- Discretization breaks spacetime sym, as in lattice QFT.



- Other than (leaky) ReLU, no rescaling symmetry.

[Godfrey Brown Emerson Kvinge 2205.14258(cs.LG)]

- Special neural ODE allows a metric interpretation.

"Neural geodesic equation"
$$\begin{cases} \dot{x}^i = v^i \\ \dot{v}^i = -\Gamma^i_{jk}(x) v^j v^k \end{cases}$$

4. Physics of NN symmetries?

Training and symmetry breaking

Nevertheless, amusing to see similarities to gauge theory!

- Weights = gauge field, Biases = Higgs field
- Weight decay = Gauge mass term

$$\sum_{i,j} w_{ij}^2 \sim \int dt\, w(t)^2$$

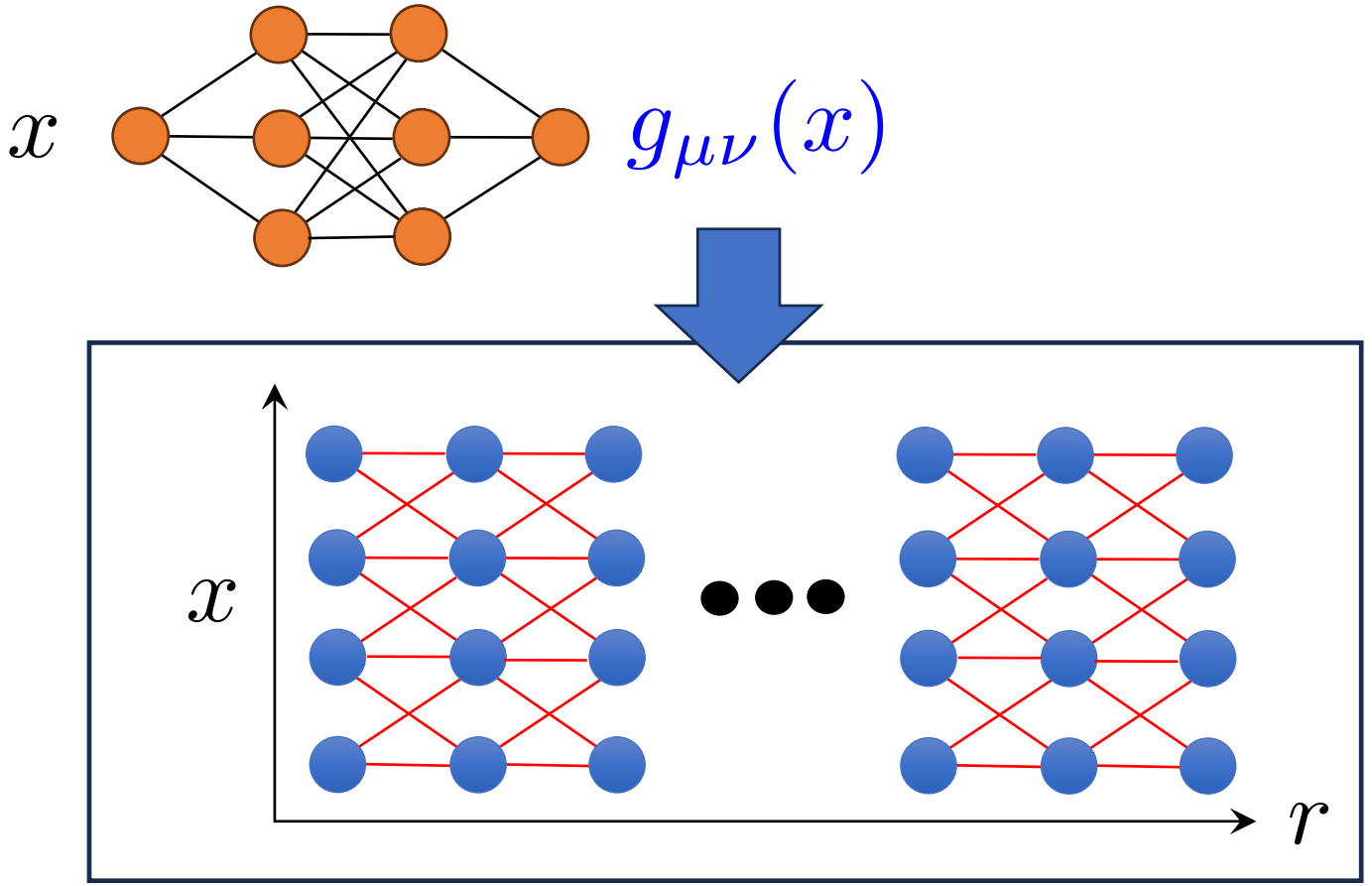
- Linear transport condition = Lorentz gauge fixing term

$$L_{\mathrm{R}} = \lambda \int_0^T dt \left[(\dot{w} + w^2)^2 + \left[(\partial_t + w) b \right]^2 \right]$$

4. Physics of NN symmetries?

Generative spacetimes

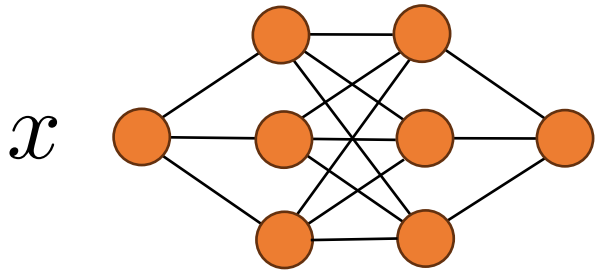
We may add other NN to generate the spacetime lattice



4. Physics of NN symmetries?

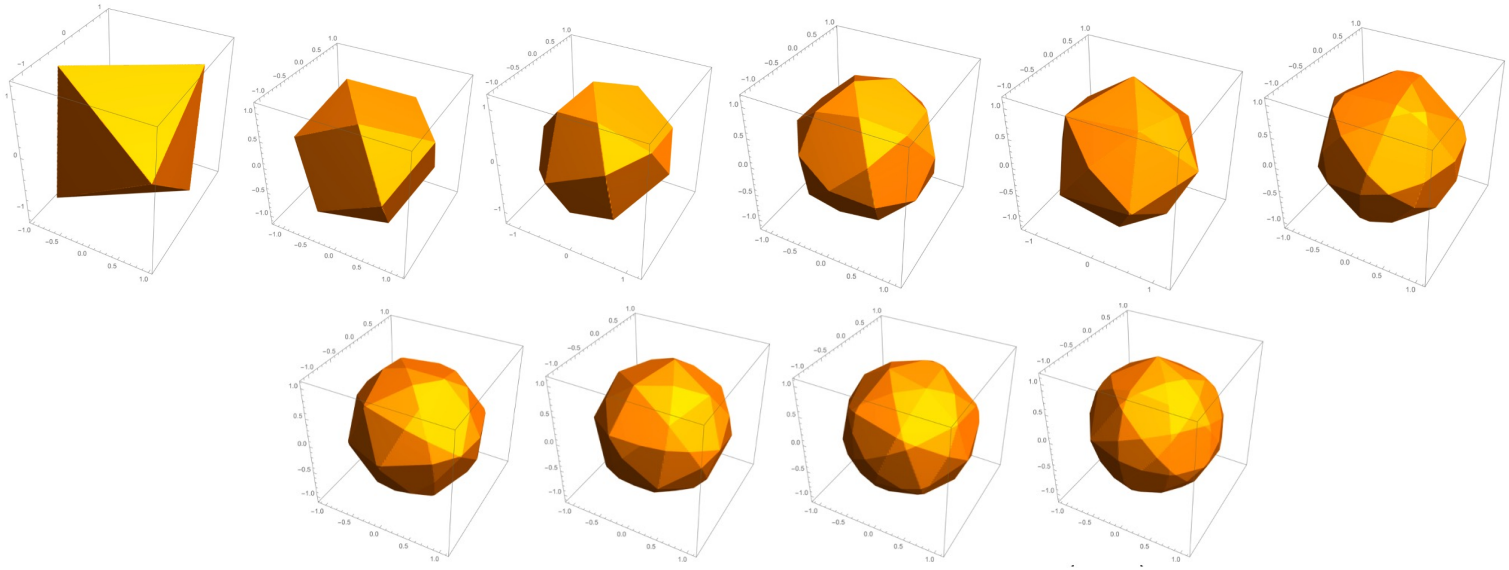
Generative spacetimes

We may add other NN to generate the spacetime lattice



$g_{\mu\nu}(x)$

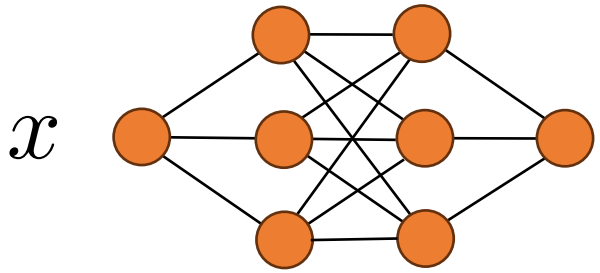
Naito, Naito, KH
ArXiv:2307.00721 [cs.LG]



4. Physics of NN symmetries?

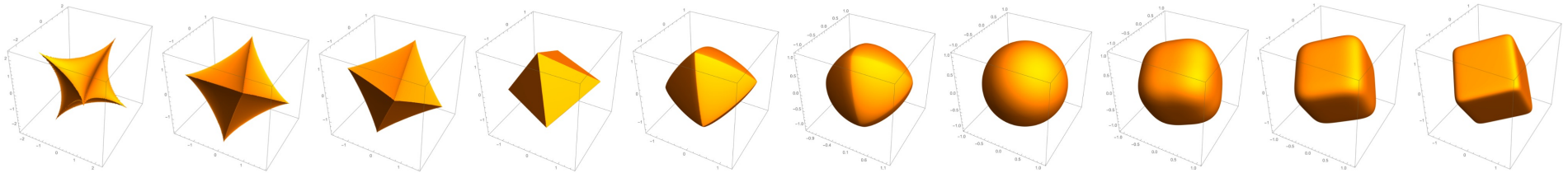
Generative spacetimes

We may add other NN to generate the spacetime lattice



$$g_{\mu\nu}(x)$$

Naito, Naito, KH
ArXiv:2307.00721 [cs.LG]



"Neural Polytopes" :
Continuous generalization of polytopes

Unification of Symmetries Inside Neural Networks

[Hirono, Sannai, KH 2402.02362]

1. Motivation: gauge sym in NN — 2 pages
2. Candidates for diffeo in NN — 4 pages
3. Diffeo in neural ODEs — 4 pages
4. Physics of NN symmetries? — 3 pages