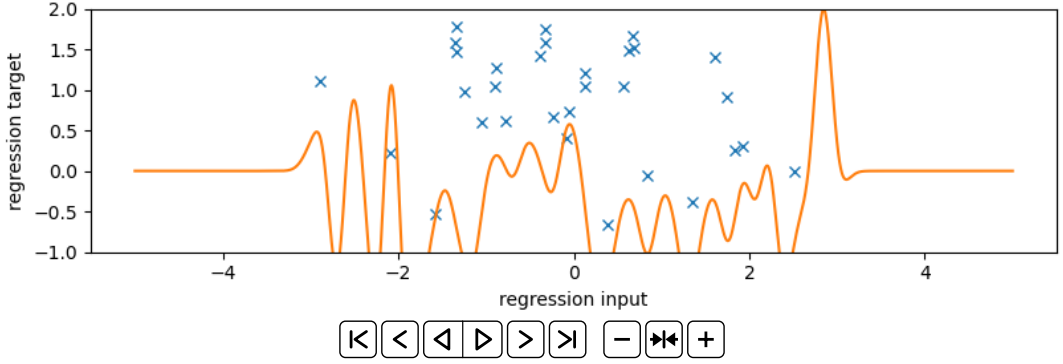
Designing a Smarter Neuron

Mark van der Wilk

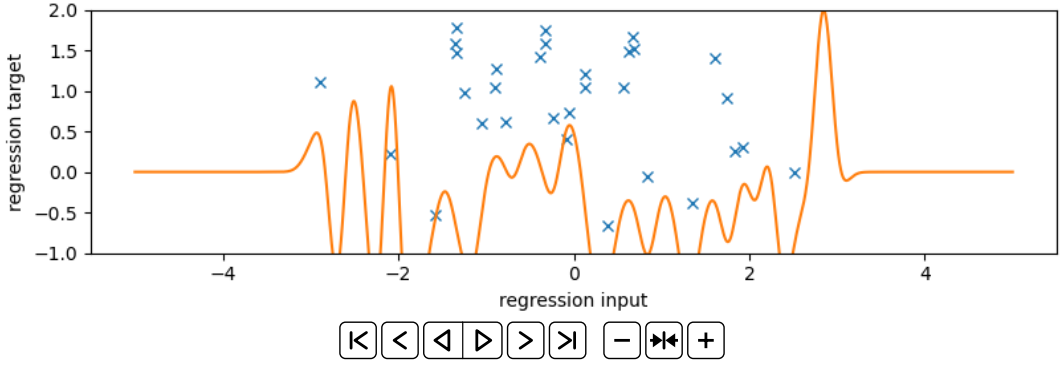
Machine Learning: What we have



Training procedure:

- Define a space of functions $\mathcal{F}$
- Find predictor $f^* = \text{argmin}_{f \in \mathcal{F}} \sum_n \text{loss}(x_n, y_n)$

Machine Learning: What we have

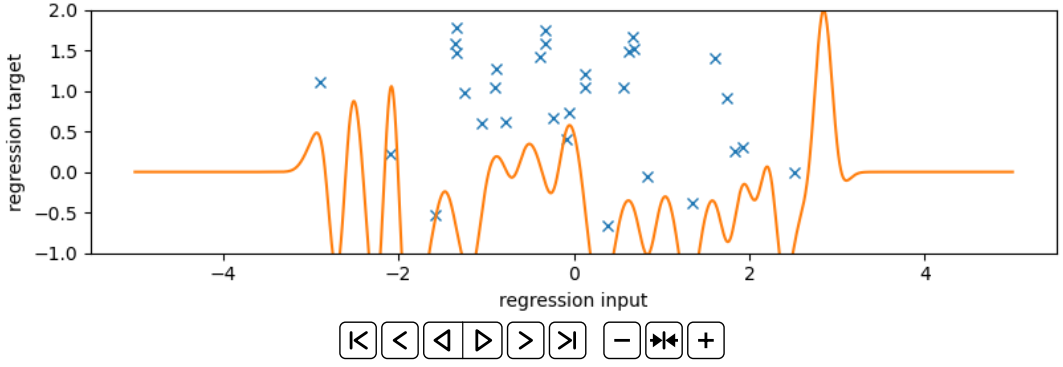


Training procedure:

- Define a space of functions $\mathcal{F}$
- Find predictor $f^* = \text{argmin}_{f \in \mathcal{F}} \sum_n \text{loss}(x_n, y_n)$

Theory states that somehow restricting $\mathcal{F}$ helps prediction.

Machine Learning: What we have

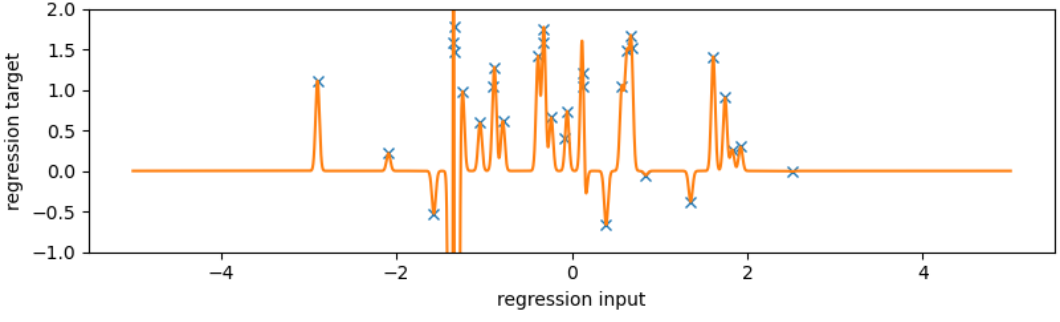


Training procedure:

- Define a space of functions $\mathcal{F}$
- Find predictor $f^* = \text{argmin}_{f \in \mathcal{F}} \sum_n \text{loss}(x_n, y_n)$

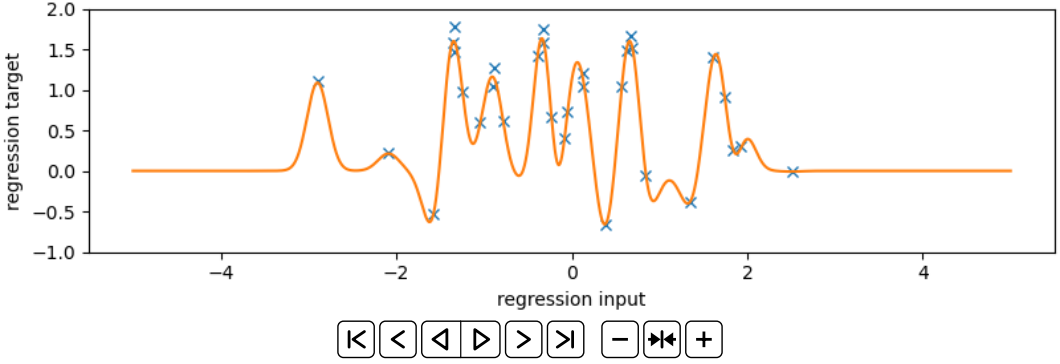
Theory states that somehow restricting $\mathcal{F}$ helps prediction. How to find $\mathcal{F}$?

Machine Learning: What can go wrong



- Predictions are now *closer* to the training data.
- But predictions are worse.

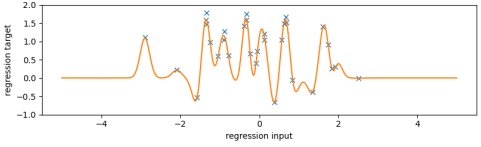
Machine Learning: What we want

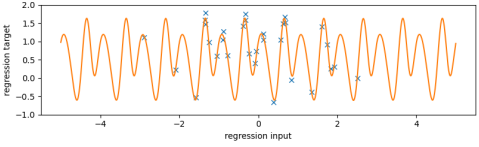


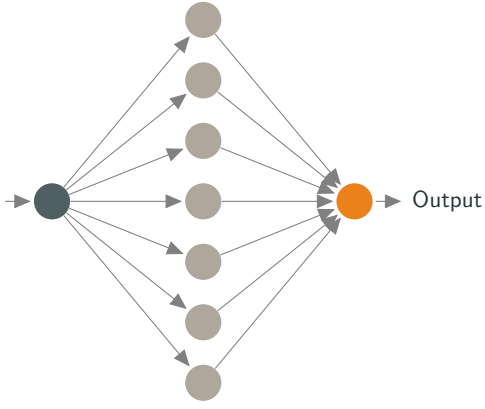
Training procedure:

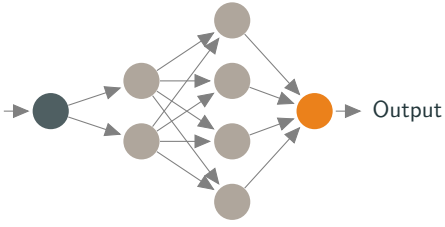
- Find predictor $f^* = \text{argmin}_{f \in \mathcal{F}} \sum_n \text{loss}(x_n, y_n)$
- But, select space of functions $\mathcal{F}$ *simultaneously*

Search over Network Architecture

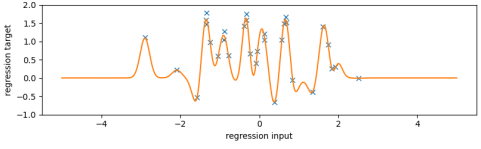


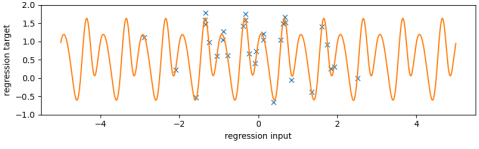


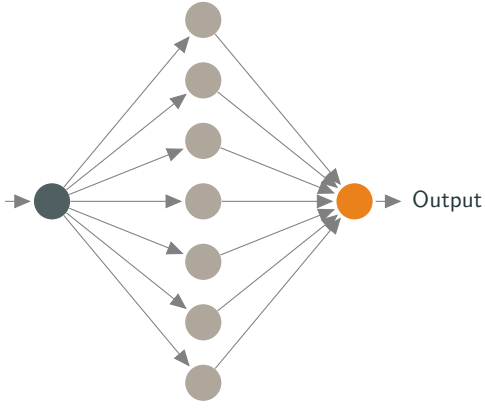


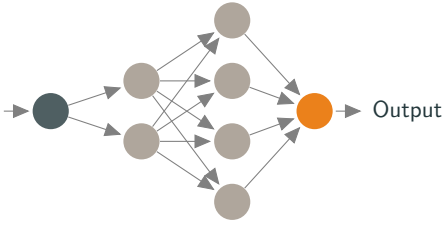


Search over Network Architecture









- More adaptive
- Better predictions
- Less human intervention

Model Selection: The Underappreciated Capability of Bayes

$$p(w, \text{arch} \mid \mathcal{D}) = \underbrace{\frac{p(\mathcal{D} \mid w, \text{arch})p(w \mid \text{arch})}{p(\mathcal{D} \mid \text{arch})}}_{p(w \mid \text{arch}, \mathcal{D})} \underbrace{\frac{p(\mathcal{D} \mid \text{arch})p(\text{arch})}{p(\mathcal{D})}}_{p(\text{arch} \mid \mathcal{D})}$$

Model Selection: The Underappreciated Capability of Bayes

$$p(w, \text{arch} \mid \mathcal{D}) = \underbrace{\frac{p(\mathcal{D} \mid w, \text{arch}) p(w \mid \text{arch})}{p(\mathcal{D} \mid \text{arch})}}_{p(w \mid \text{arch}, \mathcal{D})} \underbrace{\frac{p(\mathcal{D} \mid \text{arch}) p(\text{arch})}{p(\mathcal{D})}}_{p(\text{arch} \mid \mathcal{D})}$$

- Can approximate with key quantity, the *marginal likelihood*:

$$p(\mathcal{D}|\text{arch}) = \int p(\mathcal{D}|w, \text{arch}) p(w|\text{arch}) \mathrm{d}w \tag{1}$$

Model Selection: The Underappreciated Capability of Bayes

$$p(w, \text{arch} \mid \mathcal{D}) = \underbrace{\frac{p(\mathcal{D} \mid w, \text{arch}) p(w \mid \text{arch})}{p(\mathcal{D} \mid \text{arch})}}_{p(w \mid \text{arch}, \mathcal{D})} \underbrace{\frac{p(\mathcal{D} \mid \text{arch}) p(\text{arch})}{p(\mathcal{D})}}_{p(\text{arch} \mid \mathcal{D})}$$

- Can approximate with key quantity, the *marginal likelihood*:

$$p(\mathcal{D}|\text{arch}) = \int p(\mathcal{D}|w, \text{arch}) p(w|\text{arch}) \mathrm{d}w \tag{1}$$

- Information theory: How many *bits* are needed to encode architecture *and* data.

Model Selection: The Underappreciated Capability of Bayes

$$p(w, \text{arch} \mid \mathcal{D}) = \underbrace{\frac{p(\mathcal{D} \mid w, \text{arch}) p(w \mid \text{arch})}{p(\mathcal{D} \mid \text{arch})}}_{p(w \mid \text{arch}, \mathcal{D})} \underbrace{\frac{p(\mathcal{D} \mid \text{arch}) p(\text{arch})}{p(\mathcal{D})}}_{p(\text{arch} \mid \mathcal{D})}$$

- Can approximate with key quantity, the *marginal likelihood*:

$$p(\mathcal{D}|\text{arch}) = \int p(\mathcal{D}|w, \text{arch}) p(w|\text{arch}) \mathrm{d}w \tag{1}$$

- Information theory: How many *bits* are needed to encode architecture *and* data.
- Statistical Learning Theory: Related to bounds on *generalisation error*.

Model Selection: The Underappreciated Capability of Bayes

$$p(w, \text{arch} \mid \mathcal{D}) = \underbrace{\frac{p(\mathcal{D} \mid w, \text{arch}) p(w \mid \text{arch})}{p(\mathcal{D} \mid \text{arch})}}_{p(w \mid \text{arch}, \mathcal{D})} \underbrace{\frac{p(\mathcal{D} \mid \text{arch}) p(\text{arch})}{p(\mathcal{D})}}_{p(\text{arch} \mid \mathcal{D})}$$

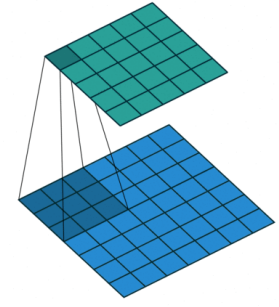
- Can approximate with key quantity, the *marginal likelihood*:

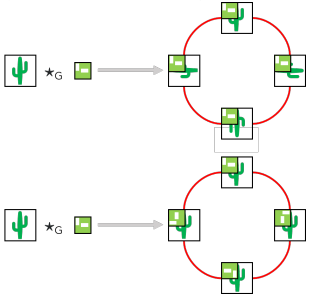
$$p(\mathcal{D}|\text{arch}) = \int p(\mathcal{D}|w, \text{arch}) p(w|\text{arch}) \mathrm{d}w \tag{1}$$

- Information theory: How many *bits* are needed to encode architecture *and* data.
- Statistical Learning Theory: Related to bounds on *generalisation error*.
- Key difficulty: Computing marginal likelihood.

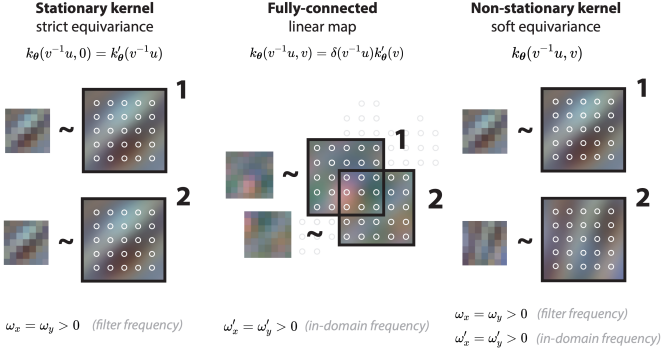
Geometric Deep Learning [v.d.Ouderaa, Romero & **vdW**, NeurIPS 2022]

Geometric Deep Learning provides a mathematical framework for specifying architectures.

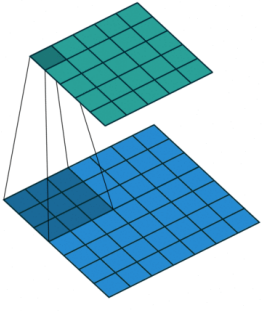


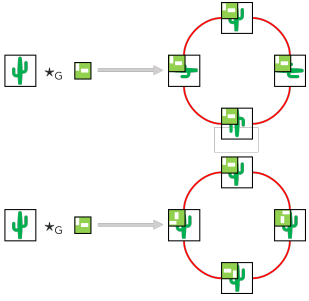


Geometric Deep Learning provides a mathematical framework for specifying architectures.

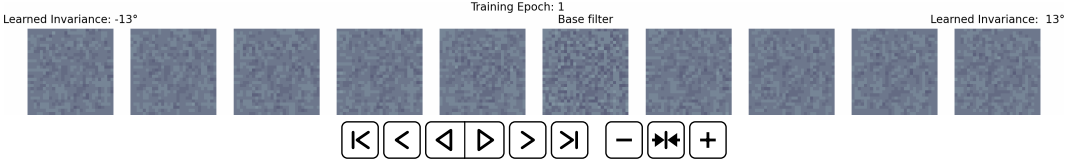


$$\int_G k_w(v^{-1}u)f(v)\mathrm{d}\mu(v) \to \int_G k_w(v^{-1}u, v)f(v)\mathrm{d}\mu(v)$$

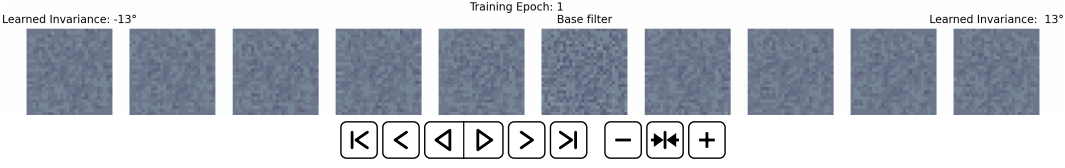




Proof-of-Concept on Shallow Neural Networks



Proof-of-Concept on Shallow Neural Networks



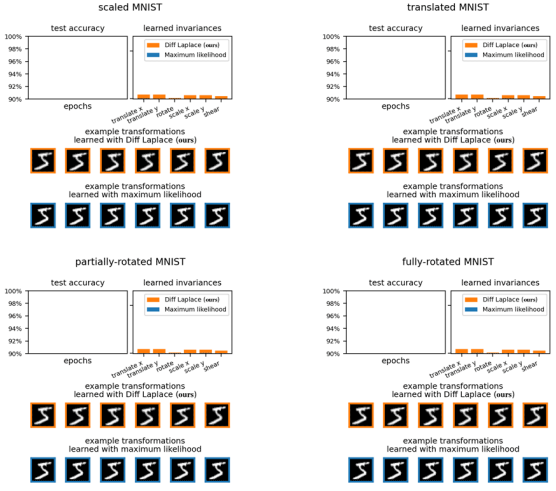
(a) Sampled filters of affine model trained on regular mnist.

(b) Sampled filters of affine model trained on rotated mnist.

(c) Sampled filters of affine model trained on scaled mnist.

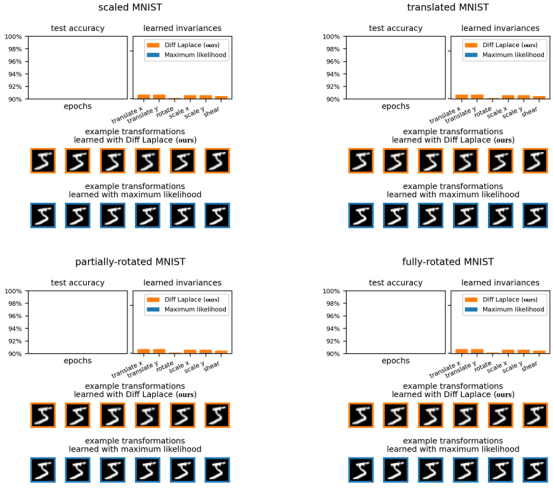
(d) Sampled filters of affine model trained on translated mnist.

Making things work for Deep NNs



Making things work for Deep NNs

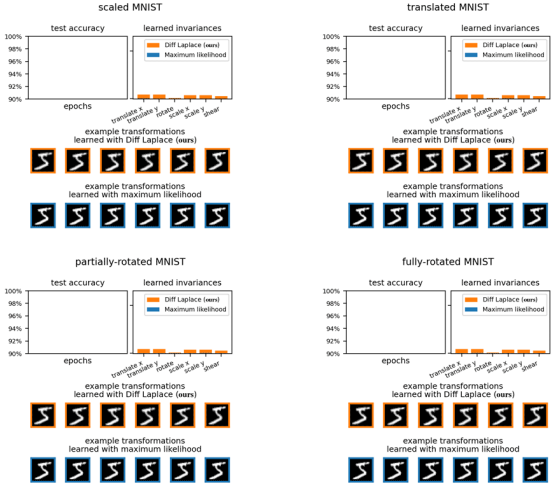
- NTK theory



Making things work for Deep NNs

- NTK theory

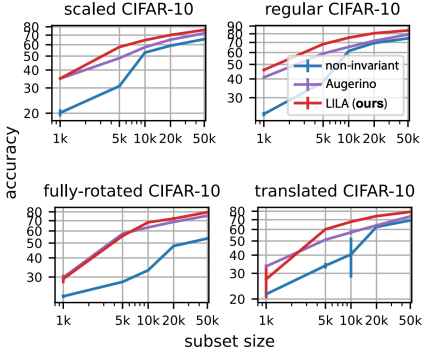
$$f_w(\mathbf{x}) \approx f_{w_0}(\mathbf{x}) + (w - w_0)^\intercal \nabla_w f_w(\mathbf{x})$$

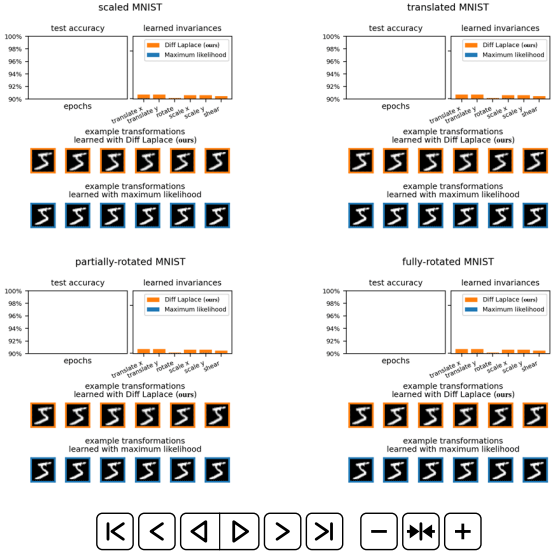


Making things work for Deep NNs

- NTK theory

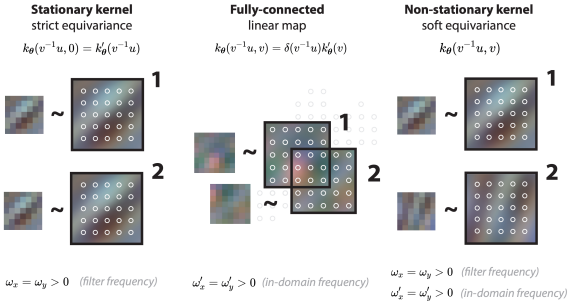
$$f_w(\mathbf{x}) \approx f_{w_0}(\mathbf{x}) + (w - w_0)^\top \nabla_w f_w(\mathbf{x})$$



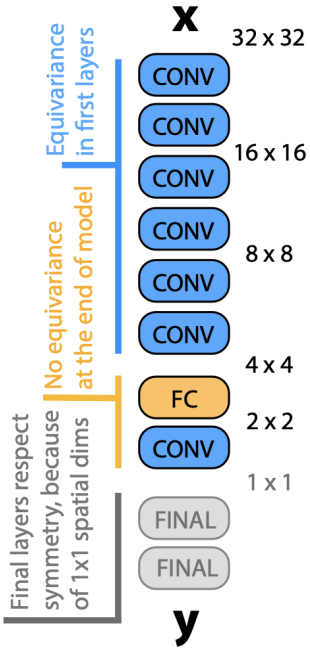


Putting it All Together: Learning Architecture

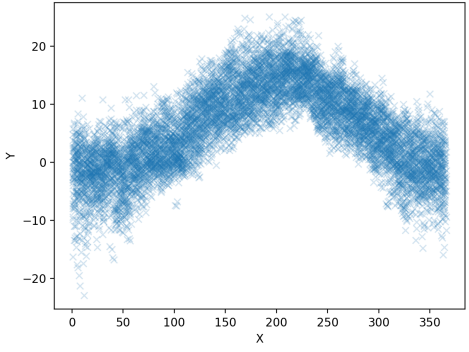
1. Use layers that can interpolate between different equivariance characteristics.
2. Optimise *marginal likelihood* estimate to find the right ones.

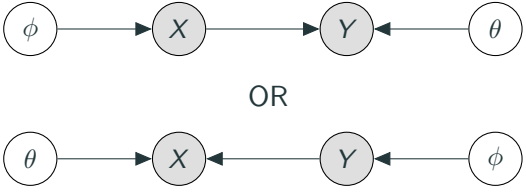


$p(\text{arch}|\mathcal{D}) \approx$ Width of minimum.



Causality [Dhir, Power & **vdW**, ICML 2024]

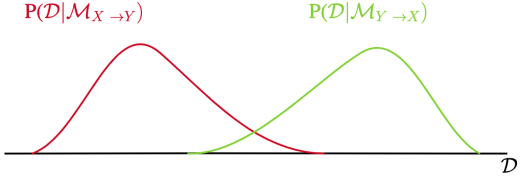


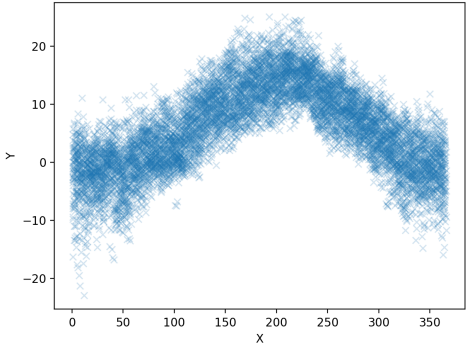


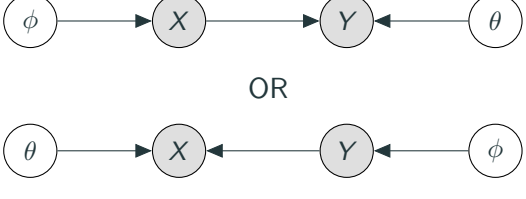
Causality [Dhir, Power & **vdW**, ICML 2024]

We show:

- Independent Causal Mechanism alone implies asymmetry in code lengths
- Direct Bayesian principles are successful at identifying causality

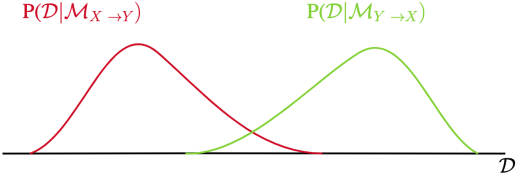


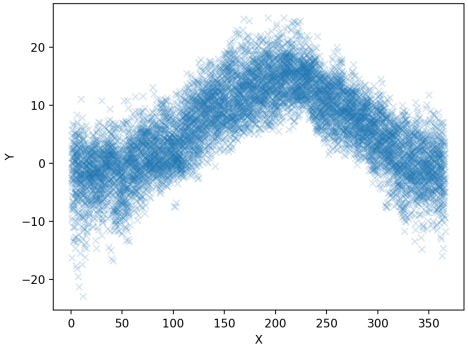




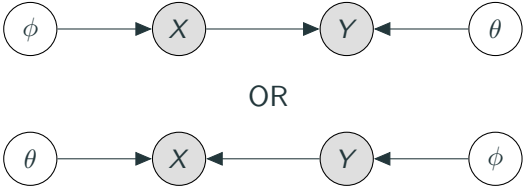
We show:

- Independent Causal Mechanism alone implies asymmetry in code lengths
- Direct Bayesian principles are successful at identifying causality





Methods	CE-Cha	CE-Multi	CE-Net	CE-Gauss	CE-Tueb
LiNGAM	57.8	62.3	3.3	72.2	31.1
ANM	43.7	25.5	87.8	90.7	63.9
PNL	78.6	51.7	75.6	84.7	73.8
IGCI	55.6	77.8	57.4	16.0	63.1
RECI	59.0	94.7	66.0	71.0	70.5
SLOPPY	60.1	95.7	79.3	71.4	65.3
CGNN	76.2	94.7	86.3	89.3	76.6
GPI	71.5	73.8	88.1	90.2	70.6
CDCI (best method reported)	72.2	96.0	94.3	**91.8**	-
GPLVM - closed form	**81.9**	**97.7**	**98.9**	89.3	-
GPLVM - stochastic	-	-	-	-	**78.3**



Compression & Computation

Compression & Computation

- Gaussian processes need $O(N^{\frac{2D}{2\nu - D}})$ neurons, for smoothness ν

  [Burt, Rasmussen, **vdW**, JMLR, 2020]

Compression & Computation

- Gaussian processes need $O(N^{\frac{2D}{2\nu - D}})$ neurons, for smoothness ν

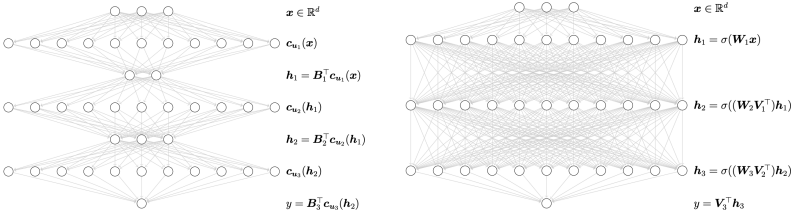
  [Burt, Rasmussen, **vdW**, JMLR, 2020]
- Hints at: Finding the right architecture $\implies$ smaller number of neurons.

Compression & Computation

- Gaussian processes need $O(N^{\frac{2D}{2\nu - D}})$ neurons, for smoothness ν

  [Burt, Rasmussen, **vdW**, JMLR, 2020]
- Hints at: Finding the right architecture $\implies$ smaller number of neurons.
- Gaussian processes are equivalent to Deep GPs

  [Dutordoir, Hensman, **vdW**, Ek, Ghahramani, Durrande, NeurIPS 2021]



(a) Deep Gaussian process posterior

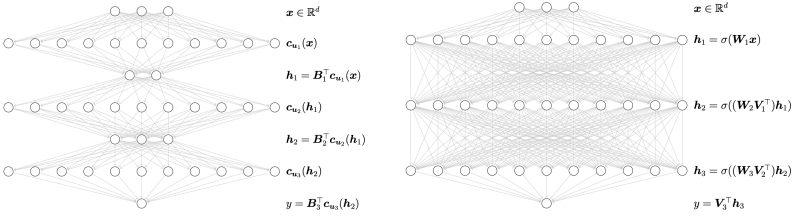
(b) Fully-connected deep neural net

Compression & Computation

- Gaussian processes need $O(N^{\frac{2D}{2\nu - D}})$ neurons, for smoothness ν

  [Burt, Rasmussen, **vdW**, JMLR, 2020]
- Hints at: Finding the right architecture $\implies$ smaller number of neurons.
- Gaussian processes are equivalent to Deep GPs

  [Dutordoir, Hensman, **vdW**, Ek, Ghahramani, Durrande, NeurIPS 2021]

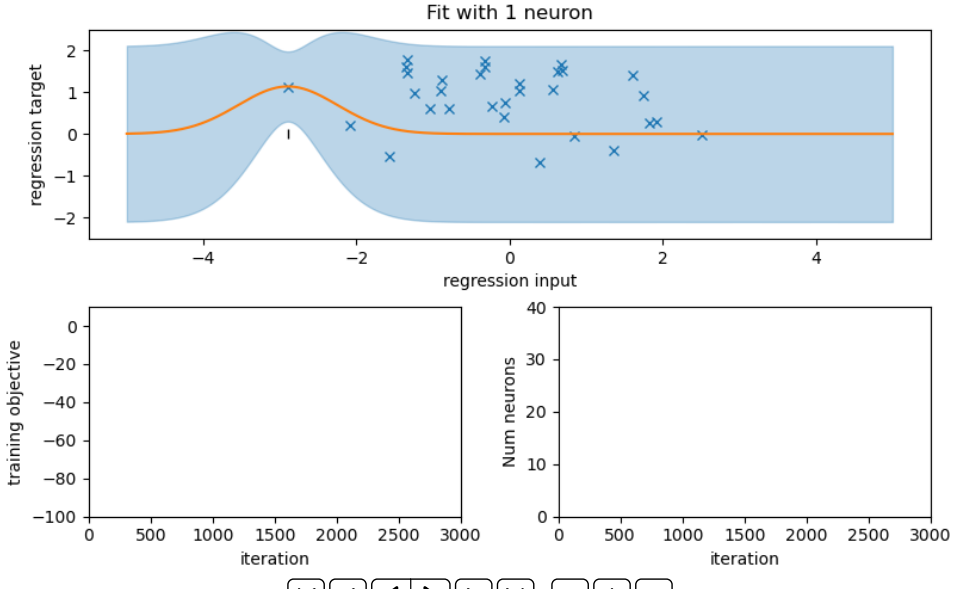


(a) Deep Gaussian process posterior (b) Fully-connected deep neural net

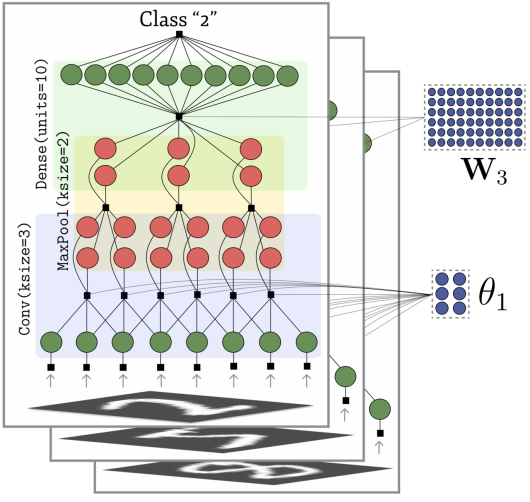
Key Question:

Can we reduce model size, by finding the right inductive bias?

Removing Unnecessary Neurons



Local Learning Rules



Research Plan

What? We want a smarter neuron that:

Research Plan

What? We want a smarter neuron that:

- learns architecture, as well as weights,

Research Plan

What? We want a smarter neuron that:

- learns architecture, as well as weights,
- generalises better outside the training dataset,

Research Plan

What? We want a smarter neuron that:

- learns architecture, as well as weights,
- generalises better outside the training dataset,
- can better detect causal relations,

Research Plan

What? We want a smarter neuron that:

- learns architecture, as well as weights,
- generalises better outside the training dataset,
- can better detect causal relations,
- removes itself from computation, if not needed,

Research Plan

What? We want a smarter neuron that:

- learns architecture, as well as weights,
- generalises better outside the training dataset,
- can better detect causal relations,
- removes itself from computation, if not needed,
- fits better with physical constraints on computation.

How?

- Existing strong statistical principles (Bayes, compression)

Research Plan

What? We want a smarter neuron that:

- learns architecture, as well as weights,
- generalises better outside the training dataset,
- can better detect causal relations,
- removes itself from computation, if not needed,
- fits better with physical constraints on computation.

How?

- Existing strong statistical principles (Bayes, compression)
- Geometric deep learning

Research Plan

What? We want a smarter neuron that:

- learns architecture, as well as weights,
- generalises better outside the training dataset,
- can better detect causal relations,
- removes itself from computation, if not needed,
- fits better with physical constraints on computation.

How?

- Existing strong statistical principles (Bayes, compression)
- Geometric deep learning
- Potentially through connections to neural cellular automata

Research Plan

What? We want a smarter neuron that:

- learns architecture, as well as weights,
- generalises better outside the training dataset,
- can better detect causal relations,
- removes itself from computation, if not needed,
- fits better with physical constraints on computation.

How?

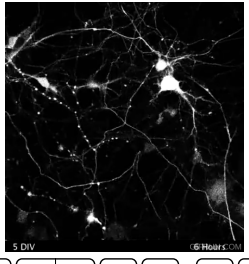
- Existing strong statistical principles (Bayes, compression)
- Geometric deep learning
- Potentially through connections to neural cellular automata

Why?

- Autonomous, energy-efficient training. New types of intelligence?

Projects & Questions?

- Understanding RKHS features, to link with NNs
- Perturbation theory for evaluating partition functions



YouTube, *Neuron time lapse video*, Leigh Needleman