Emergent Equivariance in Deep Ensembles

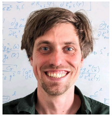
Jan E. Gerken*

CHALMERS UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

in collaboration with



Pan Kessel*

from Prescient Design

A Genentech Accelerator

Data augmentation

- 👍 Easy to implement
- 👍 No specialized architecture necessary

Data augmentation

- 👍 Easy to implement
- 👍 No specialized architecture necessary
- 👎 No exact equivariance

Data augmentation

- 👍 Easy to implement
- 👍 No specialized architecture necessary
- 👎 No exact equivariance

Can we understand data augmentation theoretically?

Empirical NTK

Training dynamics under continuous gradient descent:

learning rate

loss

$$\frac{\mathrm{d}\mathcal{N}_\theta(x)}{\mathrm{d}t} = -\frac{\eta}{N}\sum_{i=1}^{N} \Theta_\theta(x, x_i) \frac{\partial L}{\partial \mathcal{N}(x_i)}$$

training sample

Empirical NTK

Training dynamics under continuous gradient descent:

learning rate loss

$$\frac{\mathrm{d}\mathcal{N}_\theta(x)}{\mathrm{d}t} = -\frac{\eta}{N}\sum_{i=1}^{N} \Theta_\theta(x, x_i)\frac{\partial L}{\partial \mathcal{N}(x_i)}$$

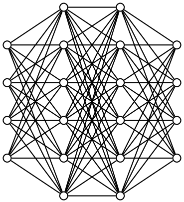
training sample

with the empirical neural tangent kernel (NTK)

$$\Theta_\theta(x, x') = \sum_\mu \frac{\partial \mathcal{N}(x)}{\partial \theta_\mu}\frac{\partial \mathcal{N}(x')}{\partial \theta_\mu}$$

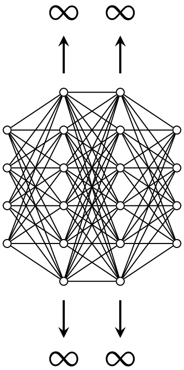
Infinite width limit

[Jacot et al. 2018]



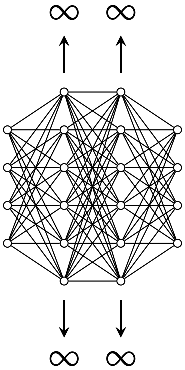
Infinite width limit

[Jacot et al. 2018]



Infinite width limit

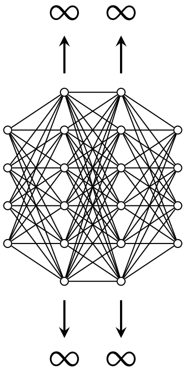
[Jacot et al. 2018]



NTK becomes independent of initialization

Infinite width limit

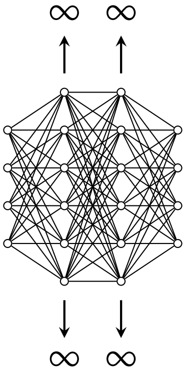
[Jacot et al. 2018]



- 👍 NTK becomes independent of initialization
- 👍 NTK becomes constant in training

Infinite width limit

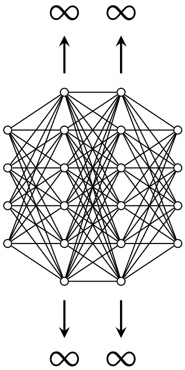
[Jacot et al. 2018]



- NTK becomes independent of initialization
- NTK becomes constant in training
- NTK can be computed for most networks

Infinite width limit

[Jacot et al. 2018]



- NTK becomes independent of initialization
- NTK becomes constant in training
- NTK can be computed for most networks
- ✓ Training dynamics can be solved

Mean prediction from NTK

[Jacot et al. 2018]

At infinite width, the mean prediction is given by

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

Mean prediction from NTK

[Jacot et al. 2018]

At infinite width, the mean prediction is given by

neural tangent kernel

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

Mean prediction from NTK

[Jacot et al. 2018]

ⓘ At infinite width, the mean prediction is given by

neural tangent kernel

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(I - e^{-\eta\Theta(X,X)t})Y$$

train data

Mean prediction from NTK

[Jacot et al. 2018]

- At infinite width, the mean prediction is given by

neural tangent kernel

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

learning rate

train data

Mean prediction from NTK

[Jacot et al. 2018]

ⓘ At infinite width, the mean prediction is given by

neural tangent kernel train labels

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

learning rate

train data

Data augmentation

Data augmentation at infinite width

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

Data augmentation at infinite width

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

augmented data

augmented labels

Data augmentation at infinite width

group transformation

$$\mu_t(\rho(g)x) = \Theta(\rho(g)x, X)\,\Theta(X, X)^{-1}\left(\mathbb{I} - e^{-\eta\Theta(X,X)t}\right)Y$$

augmented data

augmented labels

Data augmentation at infinite width

group transformation

for augmented data

$$\mu_t(\rho(g)x) = \Theta(\rho(g)x, X)\,\Theta(X, X)^{-1}\left(\mathbb{I} - e^{-\eta\Theta(X,X)t}\right)Y$$

augmented data

augmented labels

Data augmentation at infinite width

group transformation

$$\mu_t(\rho(g)x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})\rho(g)Y$$

augmented data

augmented labels

Data augmentation at infinite width

group transformation

augmented labels

$$\mu_t(\rho(g)x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})\underbrace{\rho(g)Y}_{=Y}$$

for invariance

Data augmentation at infinite width

group transformation

$$\mu_t(\rho(g)x) = \Theta(x,X)\Theta(X,X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})\underbrace{\rho(g)Y}_{=Y}$$
$$= \mu_t(x)$$

for invariance

Mean prediction

$$\mu_t(x)$$

Mean prediction

$$\mu_t(x) = \mathbb{E}_{\theta_0 \sim \text{initializations}}[\mathcal{N}_{\theta_t}(x)]$$

Mean prediction

$$\mu_t(x) = \mathbb{E}_{\theta_0 \sim \text{initializations}}[\mathcal{N}_{\theta_t}(x)] = \lim_{n\to\infty} \frac{1}{n} \sum_{\theta_0=\text{init}_1}^{\text{init}_n} \mathcal{N}_{\theta_t}(x)$$

Mean prediction

$$\mu_t(x) = \mathbb{E}_{\theta_0 \sim \text{initializations}}\left[\mathcal{N}_{\theta_t}(x)\right] = \lim_{n\to\infty} \underbrace{\frac{1}{n} \sum_{\theta_0=\text{init}_1}^{\text{init}_n} \mathcal{N}_{\theta_t}(x)}_{\text{mean prediction of deep ensemble}}$$

Main conclusion

Deep ensembles trained with data augmentation are equivariant.

Main conclusion

Deep ensembles trained with data augmentation are equivariant.

✓ Proof of exact equivariance for

- full data augmentation
- infinite ensembles
- at infinite width

Main conclusion

Deep ensembles trained with data augmentation are equivariant.

- ✓ Proof of exact equivariance for
    - full data augmentation
    - infinite ensembles
    - at infinite width
- ✓ Equivariance holds for all training times

Main conclusion

Deep ensembles trained with data augmentation are equivariant.

- ✓ Proof of exact equivariance for
    - full data augmentation
    - infinite ensembles
    - at infinite width
- ✓ Equivariance holds for all training times
- ✓ Equivariance holds away from the training data

Intuitive explanation

- ✓ Equivariance holds for all training times
- ✓ Equivariance holds away from the training data

Intuitive explanation

- ✓ Equivariance holds for all training times
- ✓ Equivariance holds away from the training data

ⓘ At infinite width, the mean output at initialization is zero everywhere.

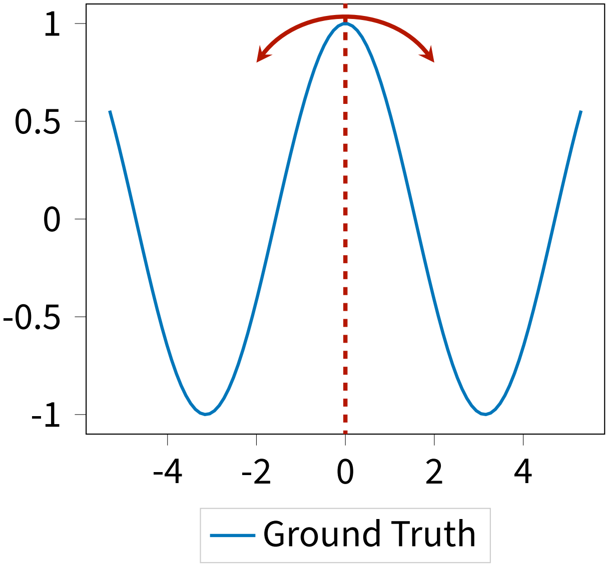
Intuitive explanation

- ✓ Equivariance holds for all training times
- ✓ Equivariance holds away from the training data

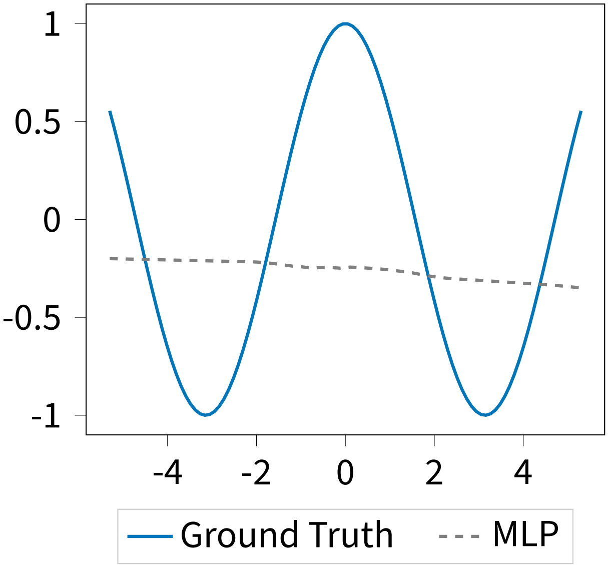
ⓘ At infinite width, the mean output at initialization is zero everywhere.

⇨ Training with full data augmentation leads to an equivariant function.

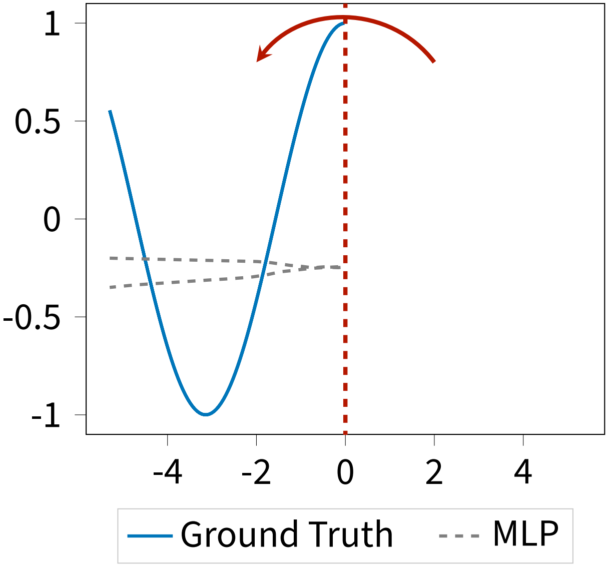
Toy example


1
0.5
0
-0.5
-1
-4
-2
0
2
4
Ground Truth

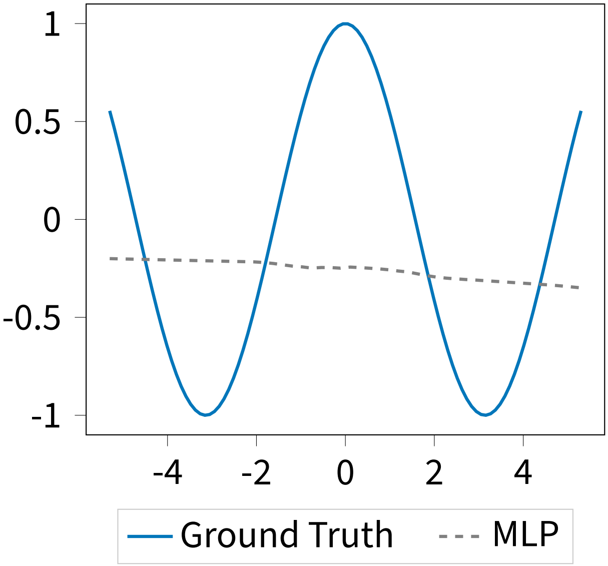
Initialization



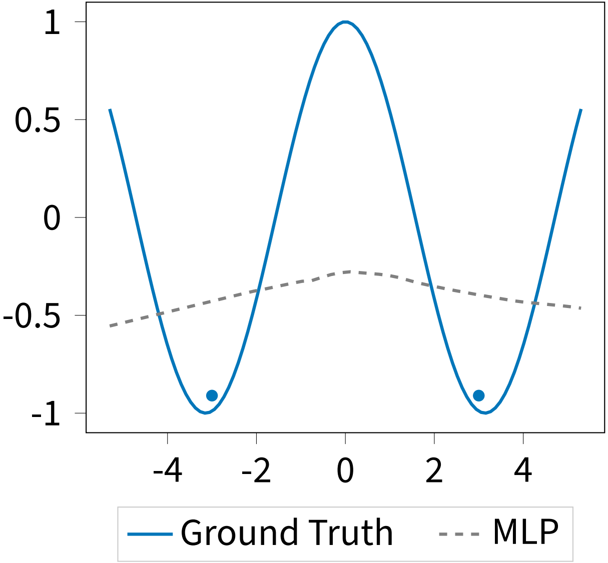
Initialization



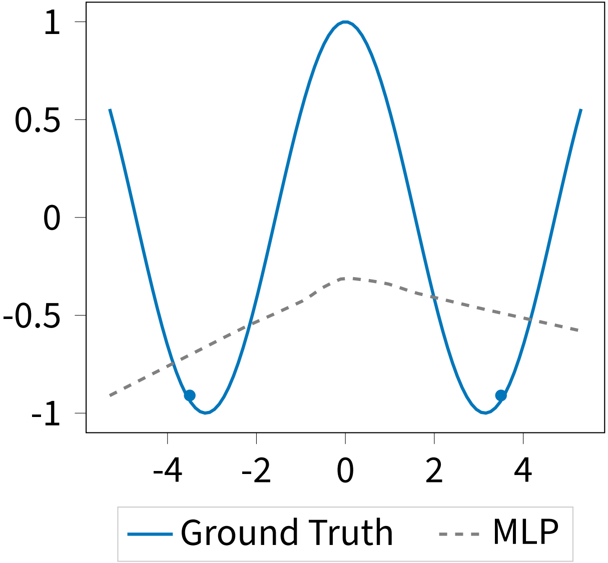
Initialization



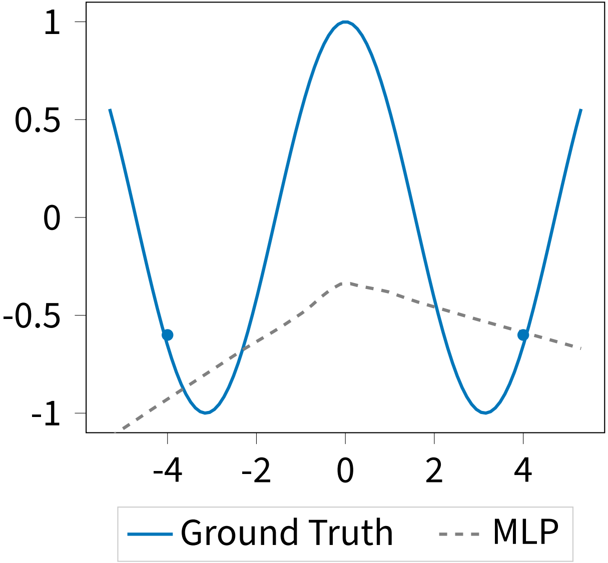
After 1 Training Step



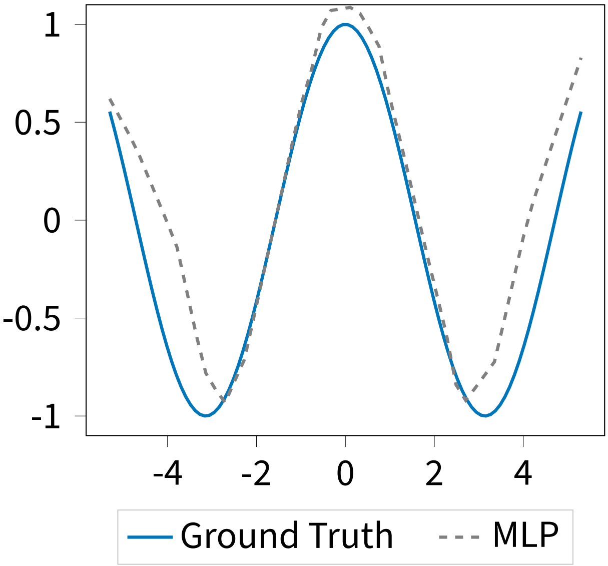
After 2 Training Steps



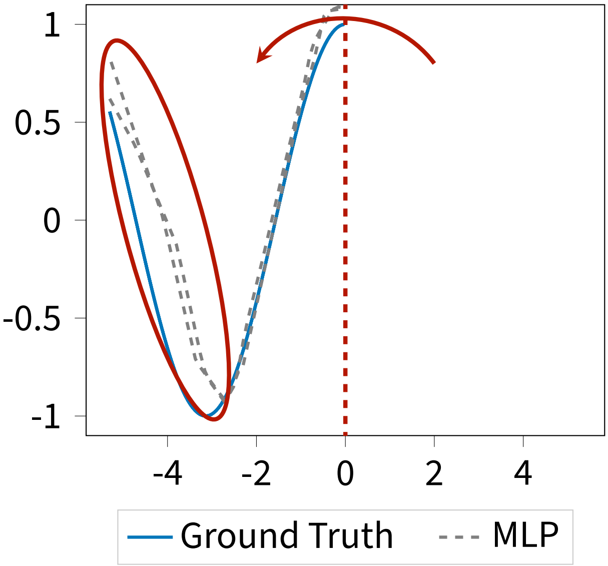
After 3 Training Steps



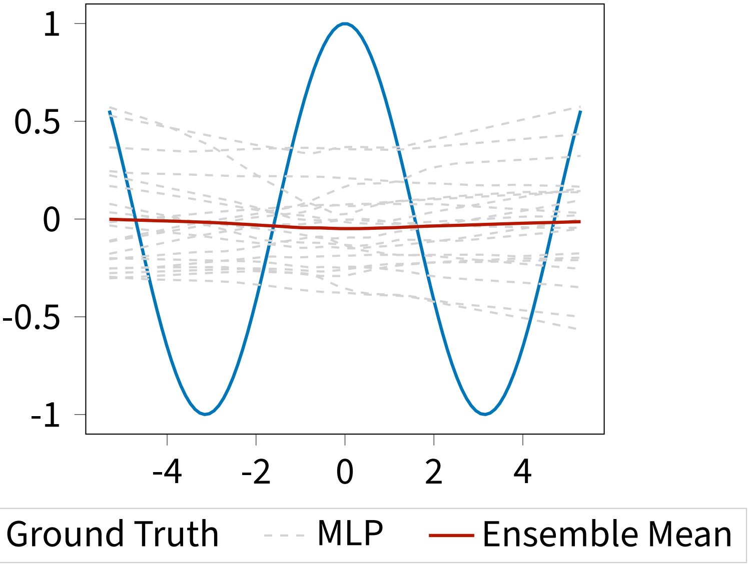
After 2000 Training Steps



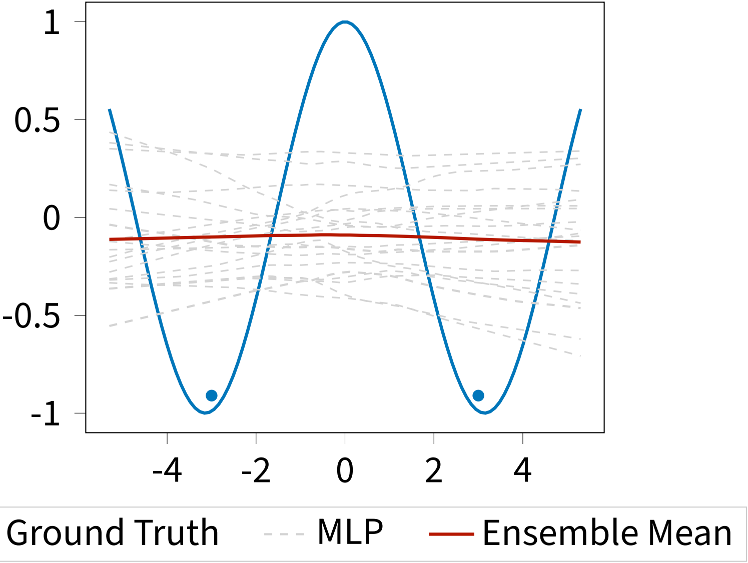
After 2000 Training Steps



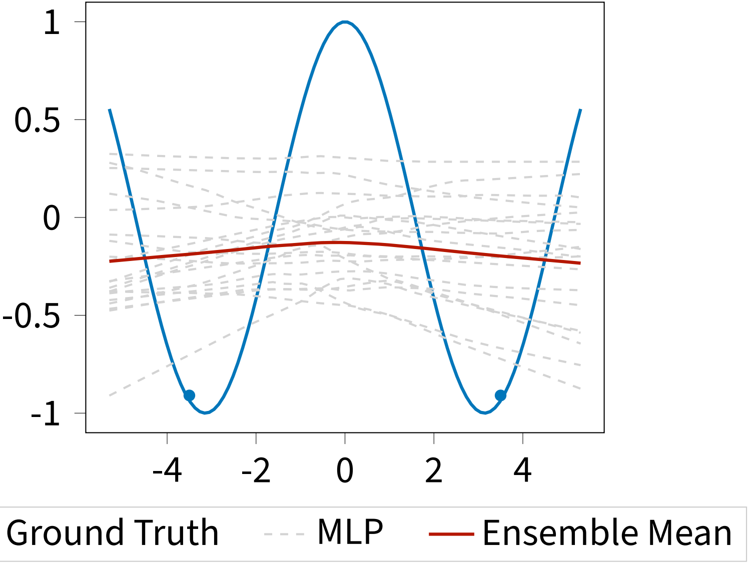
Initialization



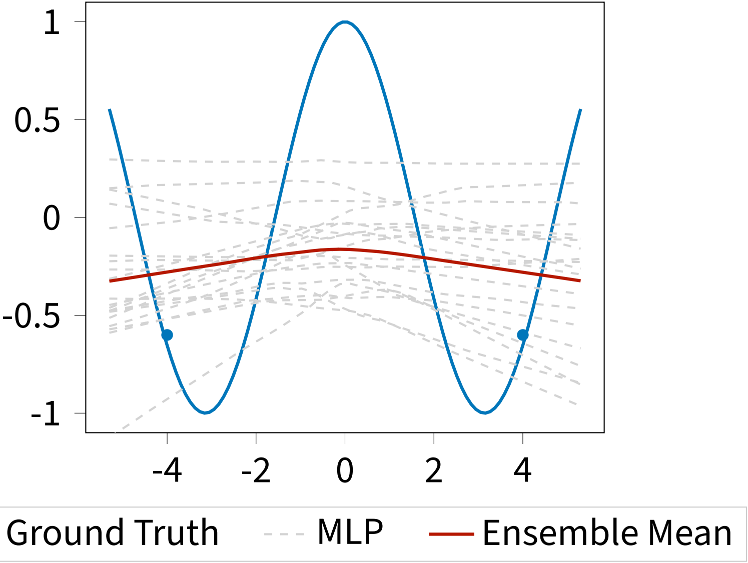
After 1 Training Step



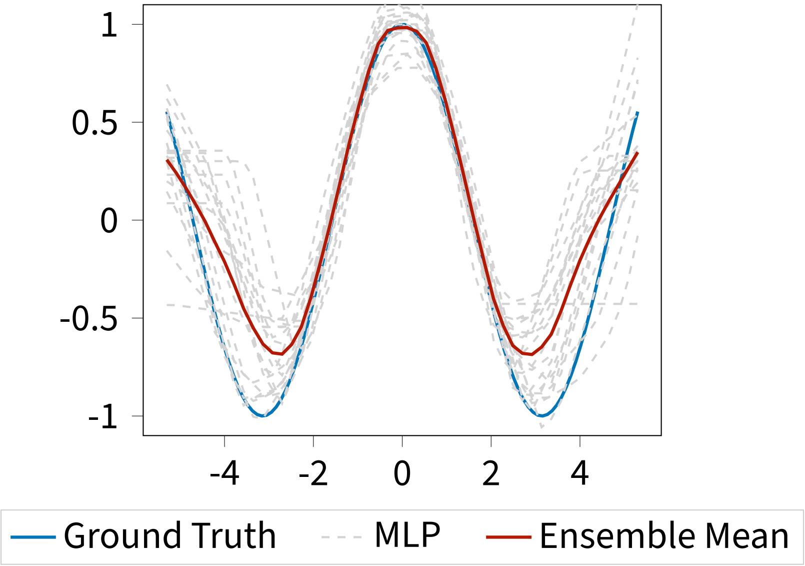
After 2 Training Steps



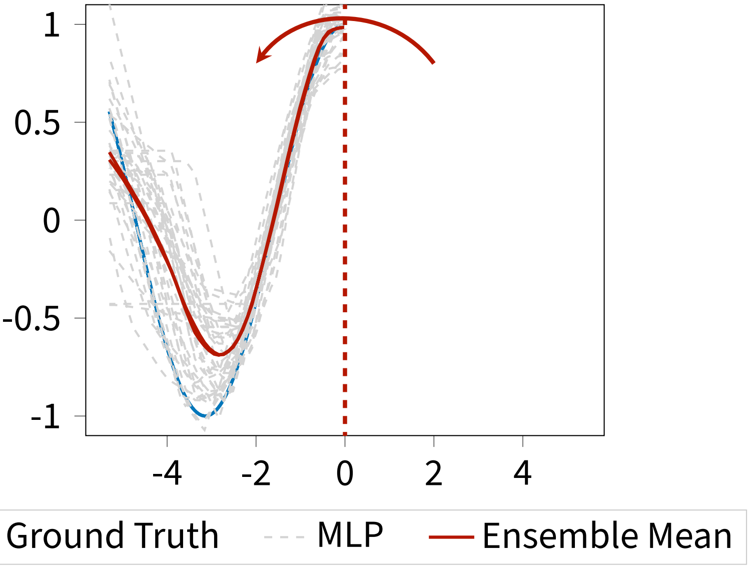
After 3 Training Steps



After 2000 Training Steps

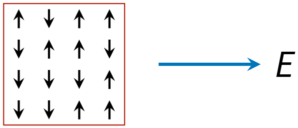


After 2000 Training Steps

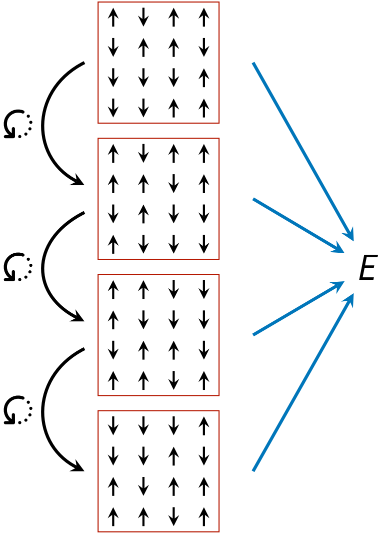


Experiments

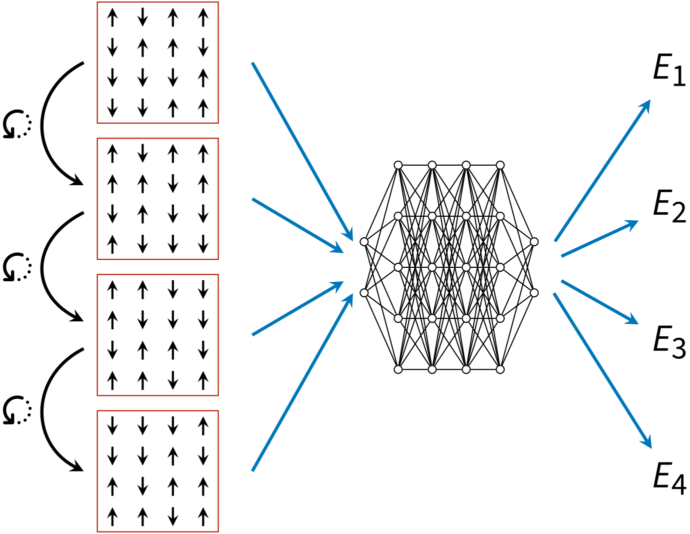
Ising model



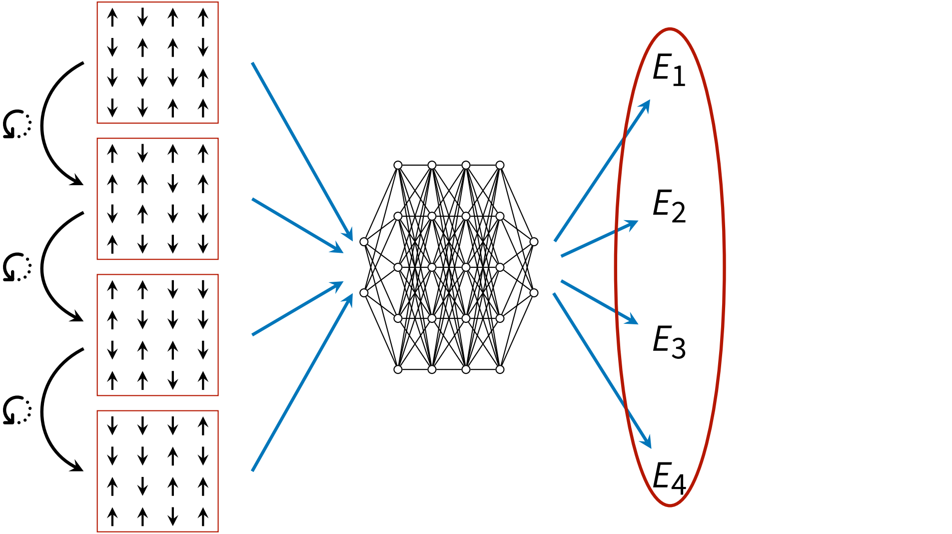
Ising model

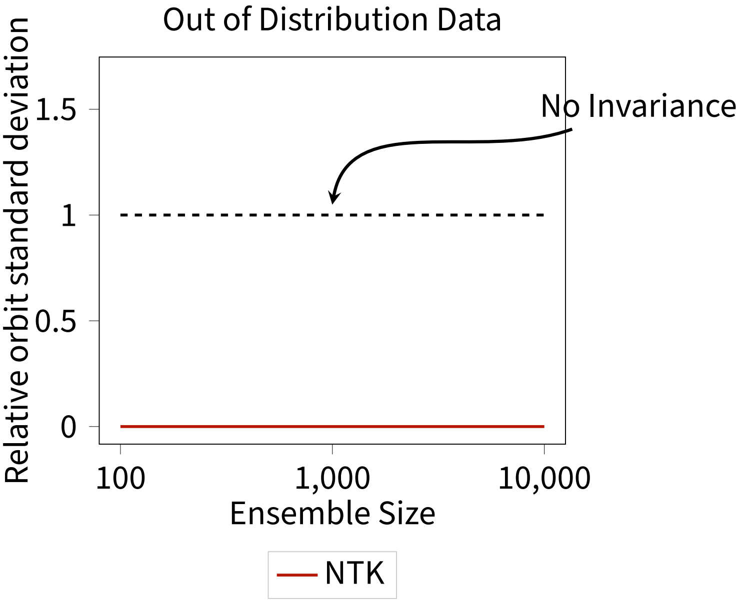


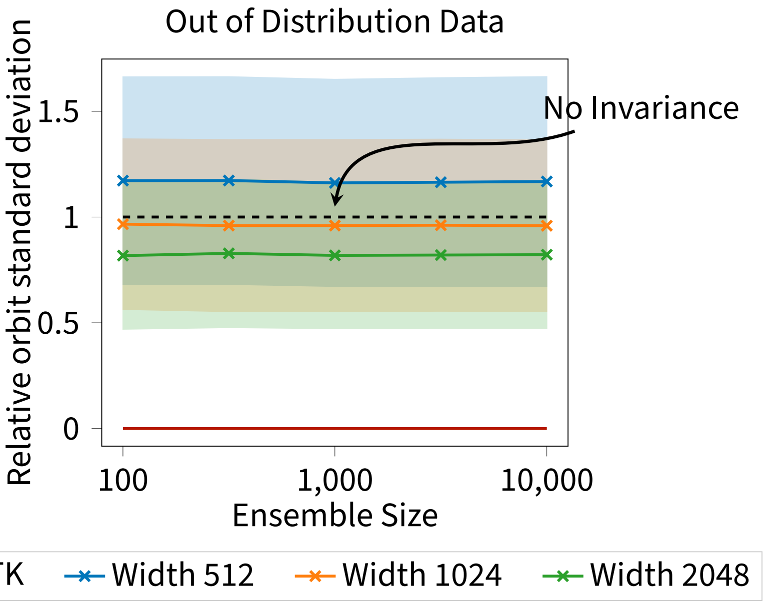
Ising model

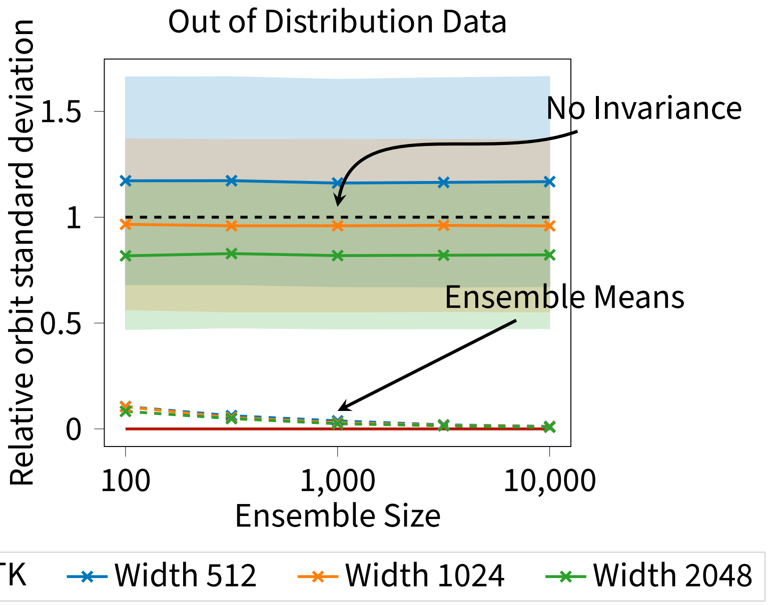


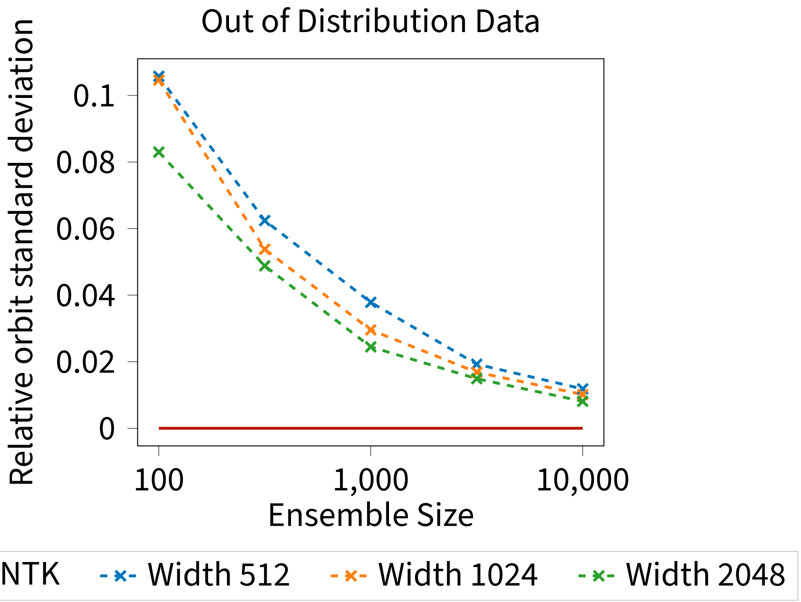
Ising model




Out of Distribution Data
No Invariance
Relative orbit standard deviation
1.5
1
0.5
0
100
1,000
10,000
Ensemble Size
NTK


Out of Distribution Data
Relative orbit standard deviation
1.5
1
0.5
0
No Invariance
100
1,000
10,000
Ensemble Size
NTK
Width 512
Width 1024
Width 2048

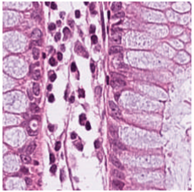

Out of Distribution Data
Relative orbit standard deviation
1.5
1
0.5
0
100
1,000
10,000
Ensemble Size
No Invariance
Ensemble Means
NTK
Width 512
Width 1024
Width 2048



Out of Distribution Data
Relative orbit standard deviation
0.1
0.08
0.06
0.04
0.02
0
100
1,000
10,000
Ensemble Size
NTK
Width 512
Width 1024
Width 2048

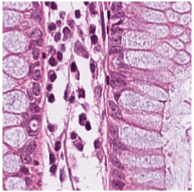
Histological slices

[Kather et al. 2018]



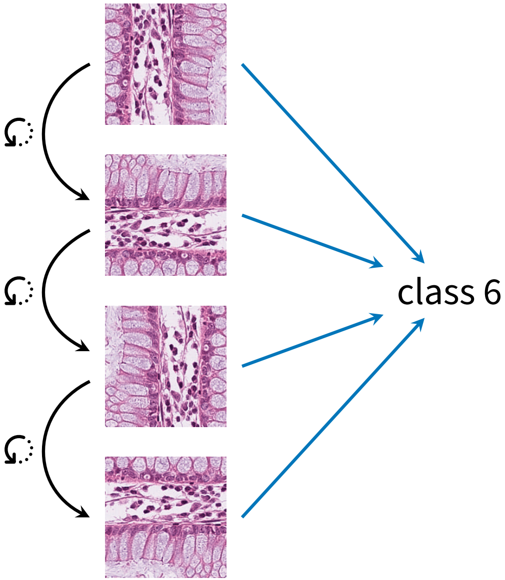
Histological slices

[Kather et al. 2018]

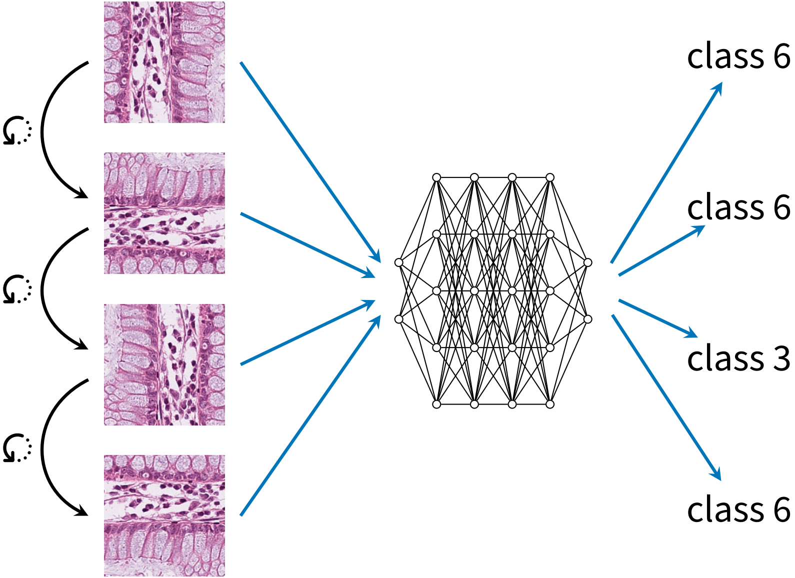


class 6

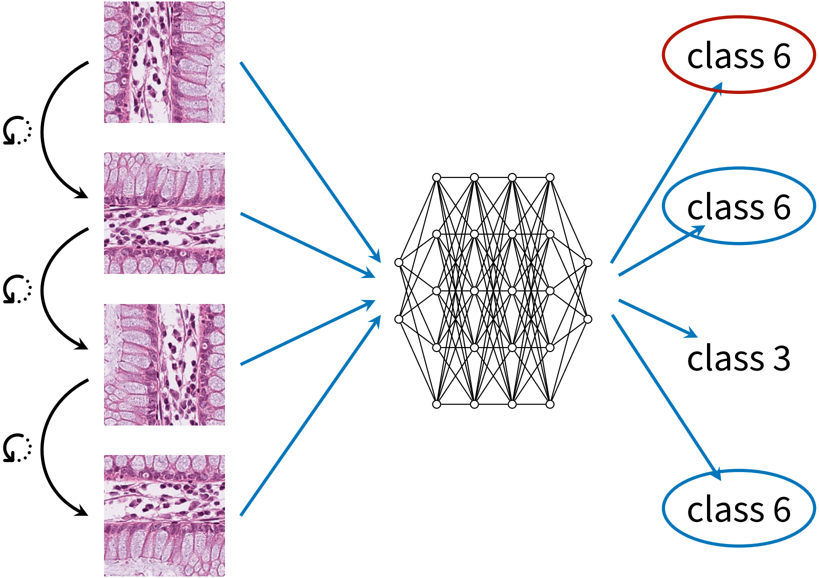
Histological slices



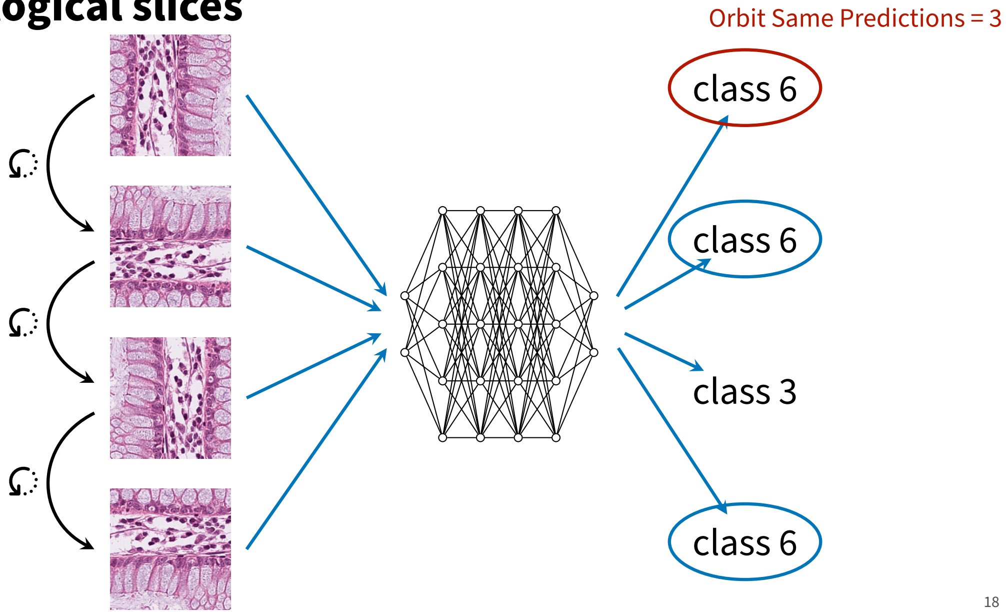
Histological slices



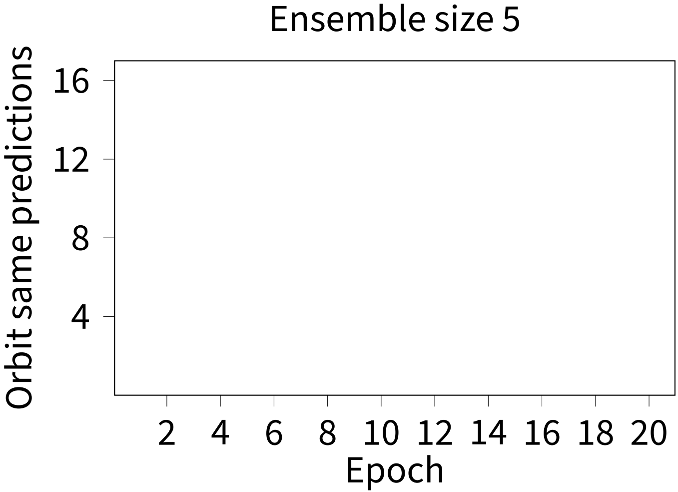
Histological slices



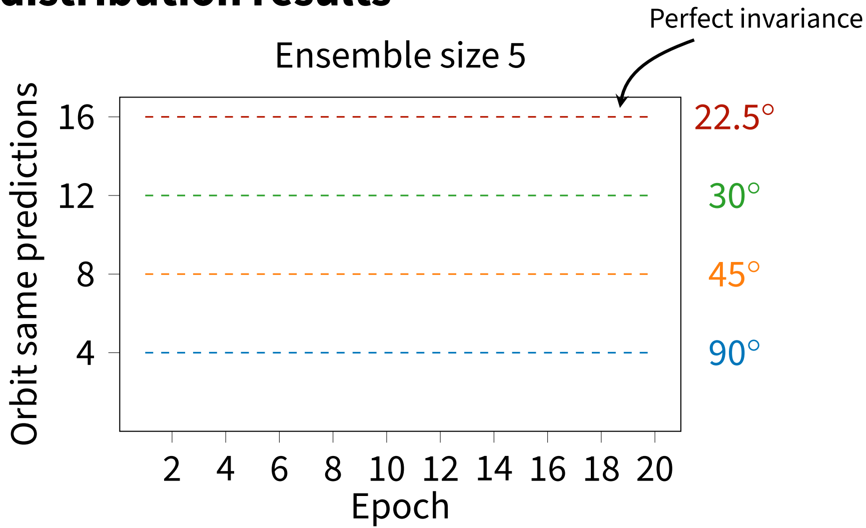
Histological slices



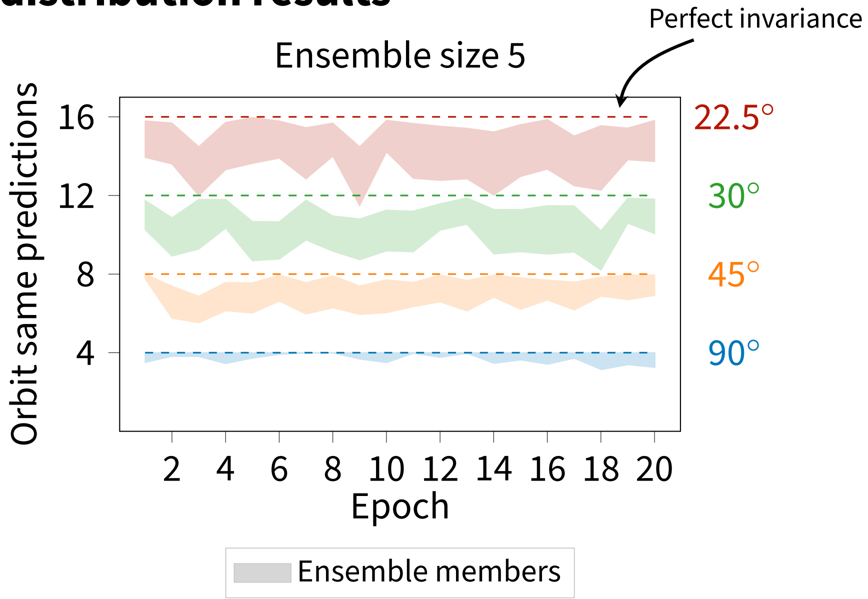
Out of distribution results



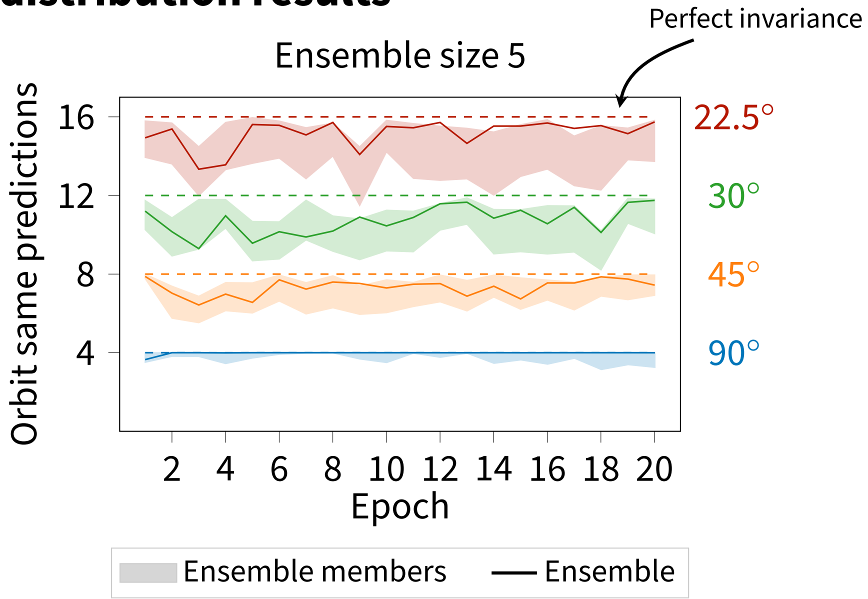
Out of distribution results



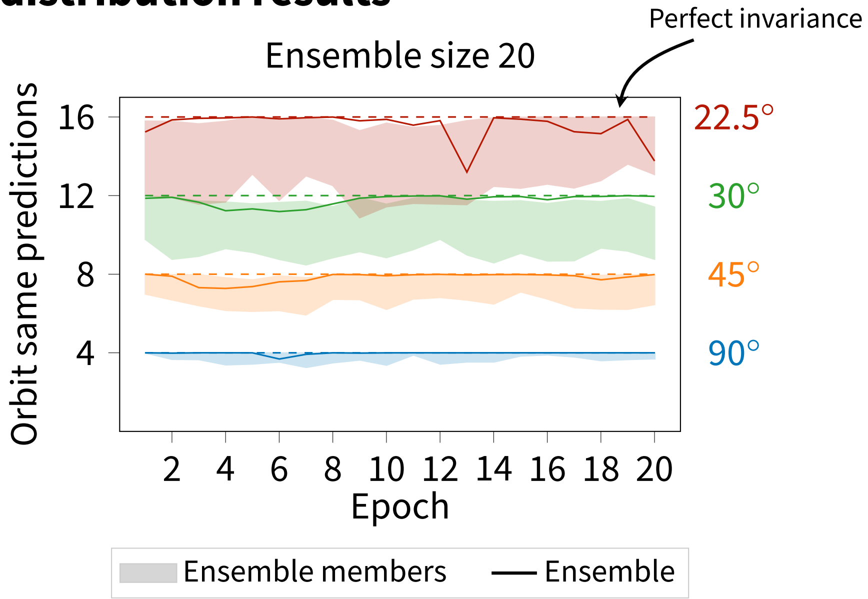
Out of distribution results



Out of distribution results



Out of distribution results



Further experimental results

Further experimental results

- ✓ Emergent invariance for rotated FashionMNIST

Further experimental results

- ✓ Emergent invariance for rotated FashionMNIST
- ✓ Partial augmentation for continuous symmetries

Further experimental results

- ✓ Emergent invariance for rotated FashionMNIST
- ✓ Partial augmentation for continuous symmetries
- ✓ Emergent equivariance (as opposed to invariance)

Comparison to other methods

Comparison to other methods

➪ Models trained on rotated FashionMNIST

Comparison to other methods

➪ Models trained on rotated FashionMNIST

Orbit same predictions out of distribution:

	C_4	C_8	C_{16}
DeepEns+DA	3.85±0.12	**7.72±0.34**	**15.24±0.69**
only DA	3.41±0.18	6.73±0.24	12.77±0.71
E2CNN[1]	**4±0.0**	**7.71±0.21**	**15.08±0.34**
Canon[2]	**4±0.0**	**7.45±0.14**	12.41±0.85

[1] [Weiler et al. 2019], [2] [Kaba et al. 2022]

Key takeaways

Key takeaways

If you need ensembles

👍 use data augmentation to obtain an equivariant model.

Key takeaways

If you need ensembles

👍 use data augmentation to obtain an equivariant model.

If you need data augmentation

👍 use an ensemble to boost the equivariance.

Key takeaways

If you need ensembles
- 👍 use data augmentation to obtain an equivariant model.

If you need data augmentation
- 👍 use an ensemble to boost the equivariance.

Analysis of neural tangent kernel can lead to powerful practical insights!

Paper

Emergent Equivariance in Deep Ensembles

Jan E. Gerken*, Pan Kessel*

ICML 2024 (Oral)

* Equal contribution

Thank you