

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

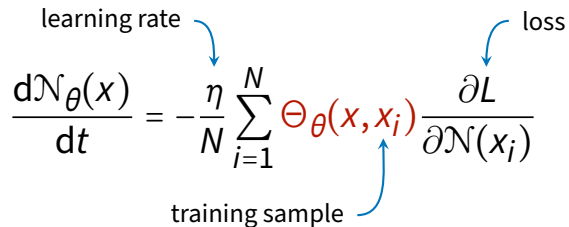
Data augmentation

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

Can we understand data augmentation theoretically?

Empirical NTK

Training dynamics under continuous gradient descent:



The diagram shows the equation for the time derivative of the Neural Tangent Kernel (NTK) under continuous gradient descent. The equation is
$$\frac{d\mathcal{N}_{\theta}(x)}{dt} = -\frac{\eta}{N} \sum_{i=1}^N \Theta_{\theta}(x, x_i) \frac{\partial L}{\partial \mathcal{N}(x_i)}$$
. Annotations with blue arrows point to specific parts: 'learning rate' points to η , 'loss' points to L , and 'training sample' points to x_i . The term $\Theta_{\theta}(x, x_i)$ is highlighted in red.

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

Empirical NTK

Training dynamics under continuous gradient descent:

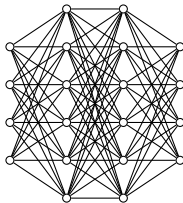
$$\frac{d\mathcal{N}_{\theta}(x)}{dt} = -\frac{\eta}{N} \sum_{i=1}^N \Theta_{\theta}(x, x_i) \frac{\partial L}{\partial \mathcal{N}(x_i)}$$

with the **empirical neural tangent kernel (NTK)**

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

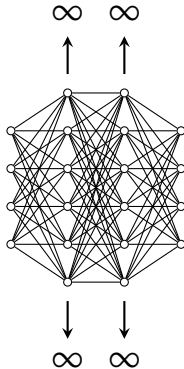
Infinite width limit

[Jacot et al. 2018]



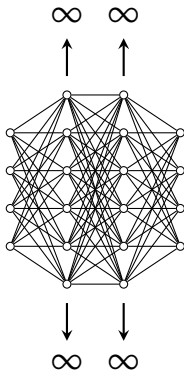
Infinite width limit

[Jacot et al. 2018]



Infinite width limit

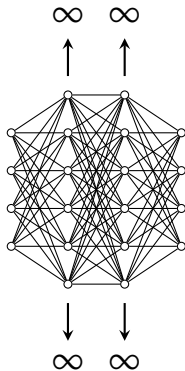
[Jacot et al. 2018]



👍 NTK becomes independent of initialization

Infinite width limit

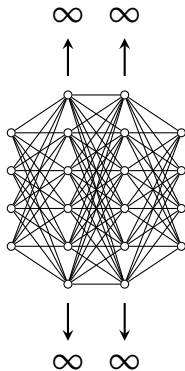
[Jacot et al. 2018]



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

Infinite width limit

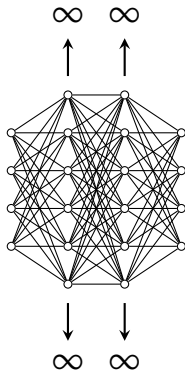
[Jacot et al. 2018]



- 👍 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]


① At infinite width, the mean prediction is given by

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

Mean prediction from NTK

[Jacot et al. 2018]

① At infinite width, the mean prediction is given by



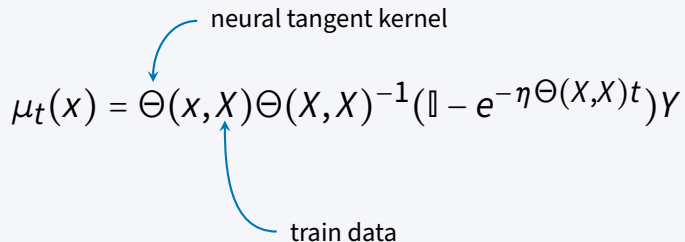
neural tangent kernel

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

Mean prediction from NTK

[Jacot et al. 2018]

① At infinite width, the mean prediction is given by



The diagram shows the equation $\mu_t(x) = \Theta(x, X) \Theta(X, X)^{-1} (\mathbb{I} - e^{-\eta \Theta(X, X) t}) \gamma$. A blue curved arrow points from the text "neural tangent kernel" to the $\Theta(x, X)$ term. Another blue curved arrow points from the text "train data" to the X in the $\Theta(X, X)$ term.

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

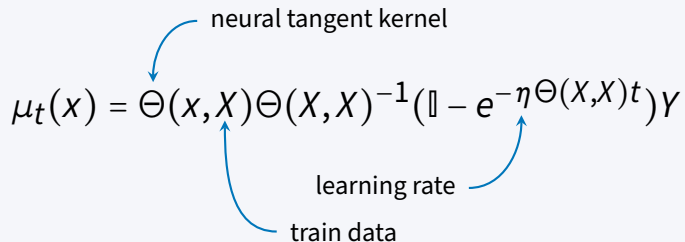
neural tangent kernel

train data

Mean prediction from NTK

[Jacot et al. 2018]

① At infinite width, the mean prediction is given by



The diagram shows the equation $\mu_t(x) = \Theta(x, X) \Theta(X, X)^{-1} (\mathbb{I} - e^{-\eta \Theta(X, X) t}) \gamma$. Three blue arrows point from text labels to parts of the equation: 'neural tangent kernel' points to $\Theta(x, X)$, 'train data' points to X in $\Theta(X, X)$, and 'learning rate' points to η .

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

neural tangent kernel

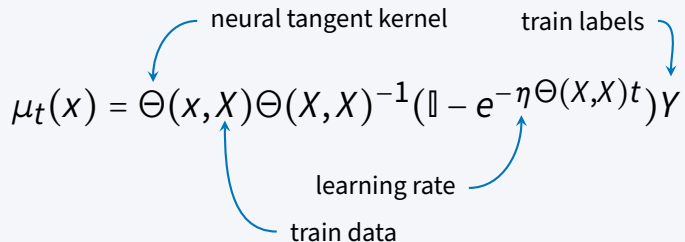
learning rate

train data

Mean prediction from NTK

[Jacot et al. 2018]

① At infinite width, the mean prediction is given by



The diagram shows the formula $\mu_t(x) = \Theta(x, X) \Theta(X, X)^{-1} (\mathbb{I} - e^{-\eta \Theta(X, X) t}) Y$ with four blue arrows pointing to its components: 'neural tangent kernel' points to $\Theta(x, X)$, 'train labels' points to Y , 'learning rate' points to η , and 'train data' points to X in the $\Theta(X, X)$ term.

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

Data augmentation

Data augmentation at infinite width

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

Data augmentation at infinite width

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

The diagram illustrates the components of the equation $\mu_t(x) = \Theta(x, X) \Theta(X, X)^{-1} (\mathbb{I} - e^{-\eta \Theta(X, X) t}) \gamma$. It features two red labels at the bottom: "augmented data" on the left and "augmented labels" on the right. From "augmented data", three blue arrows point upwards to the terms $\Theta(x, X)$, $\Theta(X, X)^{-1}$, and $\Theta(X, X)$ within the equation. From "augmented labels", two blue arrows point upwards to the terms $\Theta(X, X)$ and γ within the equation.

Data augmentation at infinite width

group transformation

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

augmented data

augmented labels

The diagram illustrates the equation for data augmentation at infinite width. A blue arrow points from the text 'group transformation' to the term $\rho(g)$ in the equation. Below the equation, the text 'augmented data' has four blue arrows pointing to the terms $\rho(g)x$, X , X , and X in the expression $\Theta(\rho(g)x, X) \Theta(X, X)^{-1}$. The text 'augmented labels' has two blue arrows pointing to the terms X and X in the expression $\Theta(X, X)^{-1}$. A final blue arrow points from 'augmented labels' to the term Y at the end of the equation.

Data augmentation at infinite width

group transformation for augmented data

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

augmented data augmented labels

The diagram illustrates the equation for data augmentation at infinite width. The equation is $\mu_t(\rho(g)x) = \Theta(\rho(g)x, X) \Theta(X, X)^{-1} (\mathbb{I} - e^{-\eta \Theta(X, X)t}) Y$. Annotations include: 'group transformation' pointing to $\rho(g)$; 'for augmented data' pointing to the entire equation; 'augmented data' pointing to $\rho(g)x$; 'augmented labels' pointing to Y ; and a curved arrow labeled 'for augmented data' connecting the input $\rho(g)x$ to the output Y . There are also three arrows pointing from the 'augmented data' label to the arguments of the first Θ function, and two arrows pointing from the 'augmented labels' label to the arguments of the second Θ function.

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

The diagram illustrates the equation for data augmentation at infinite width. A blue arrow points from the text 'group transformation' to the $\rho(g)$ term in the equation. Another blue arrow points from the text 'augmented data' to the x term. A third blue arrow points from the text 'augmented labels' to the Y term. There are also blue arrows pointing from the 'augmented data' label to the $\Theta(x, X)$ and $\Theta(X, X)^{-1}$ terms, and from the 'augmented labels' label to the $\Theta(X, X)$ term in the exponent of the exponential function.

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

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

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

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

Mean prediction

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

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.

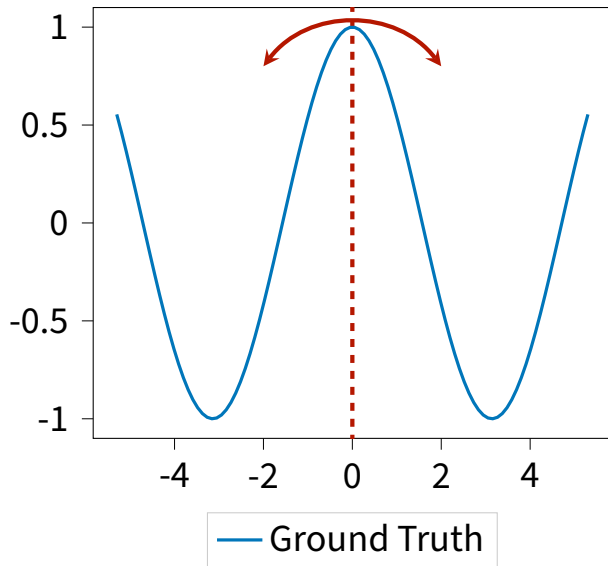
Intuitive explanation

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

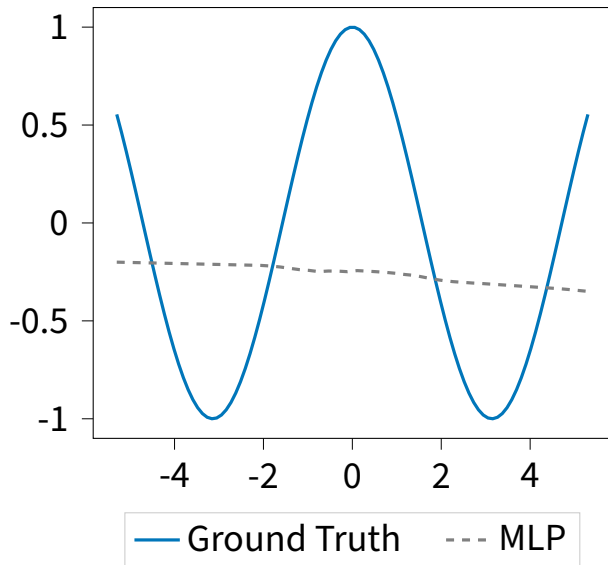
ⓘ At infinite width, the mean output at initialization is zero everywhere.

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

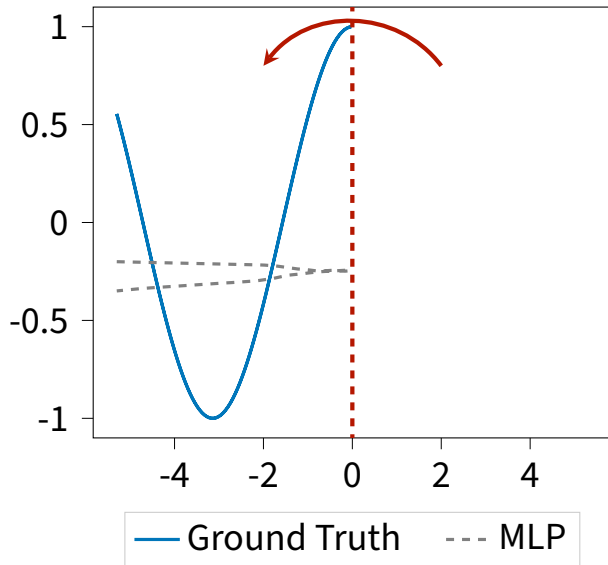
Toy example



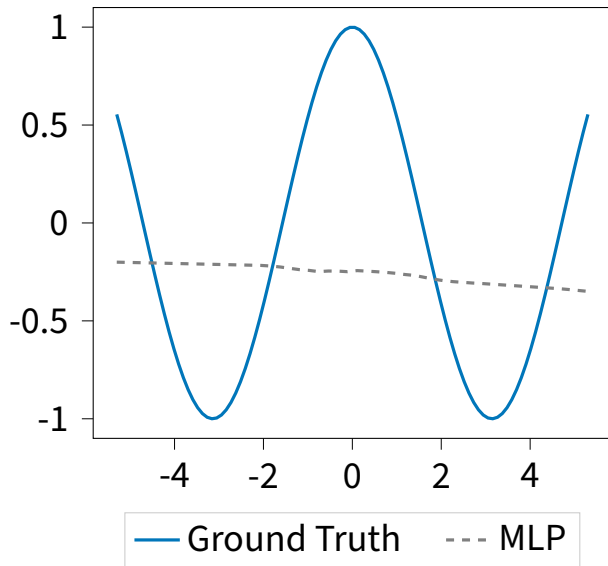
Initialization



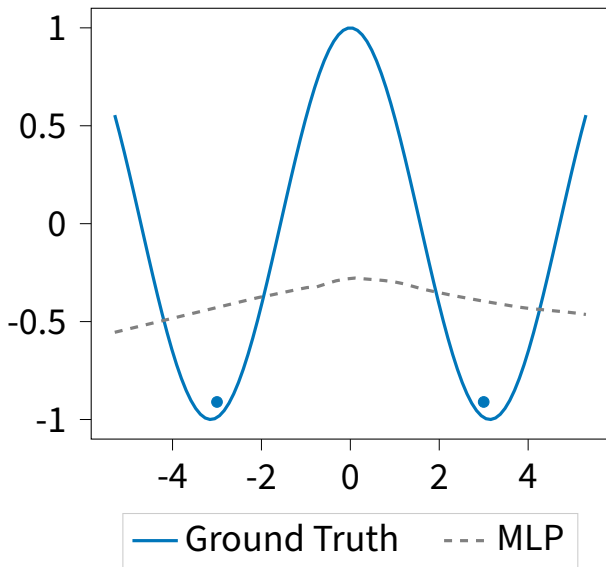
Initialization



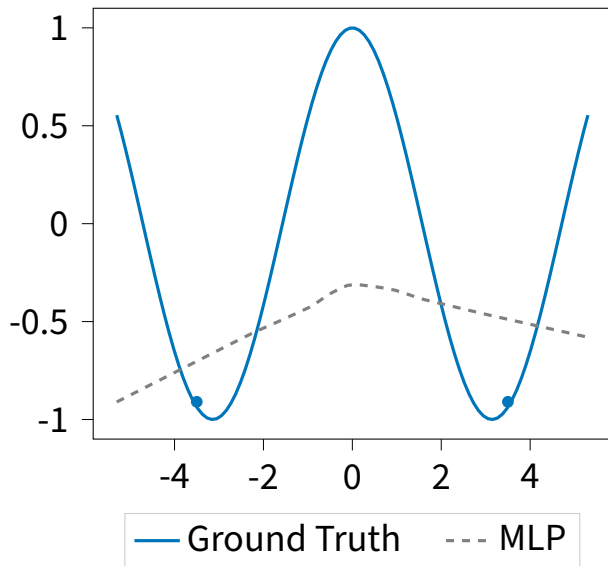
Initialization



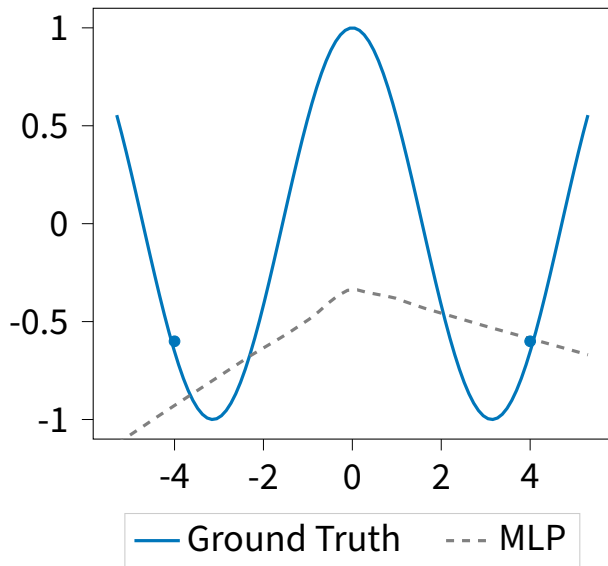
After 1 Training Step



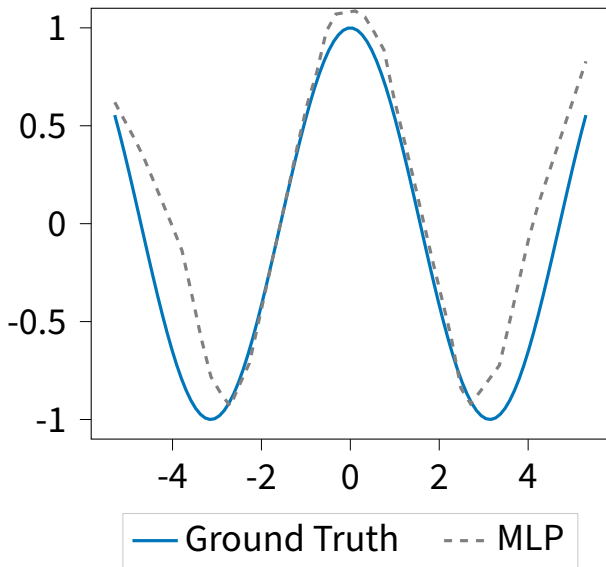
After 2 Training Steps



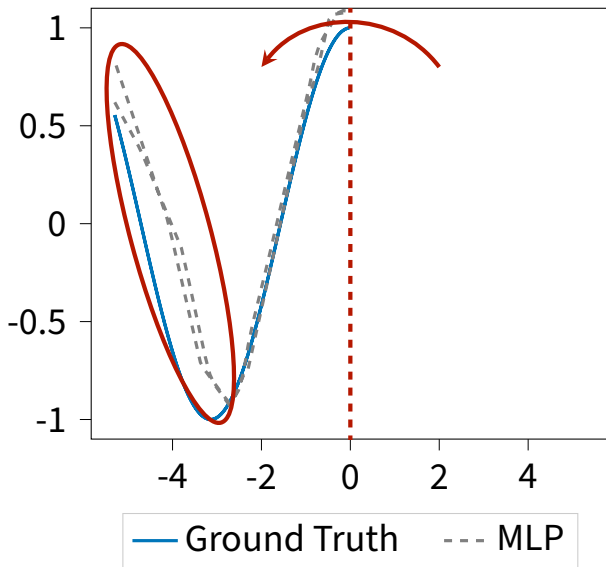
After 3 Training Steps



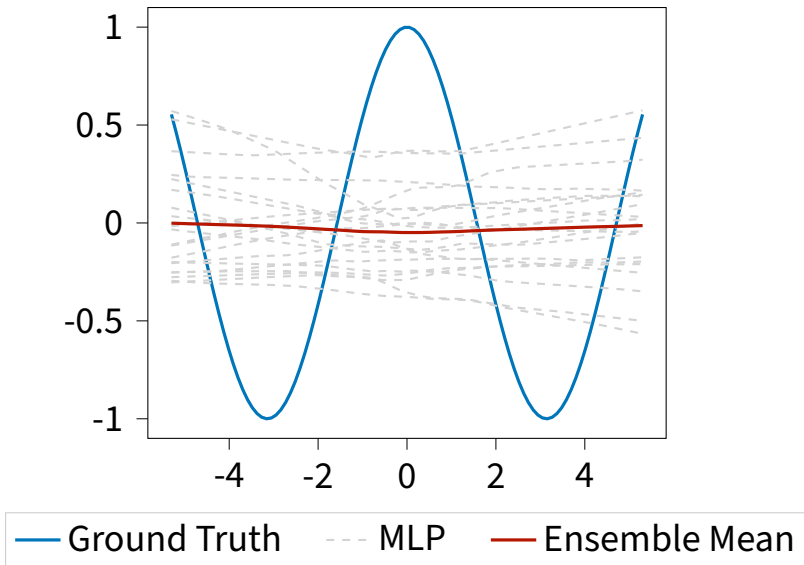
After 2000 Training Steps



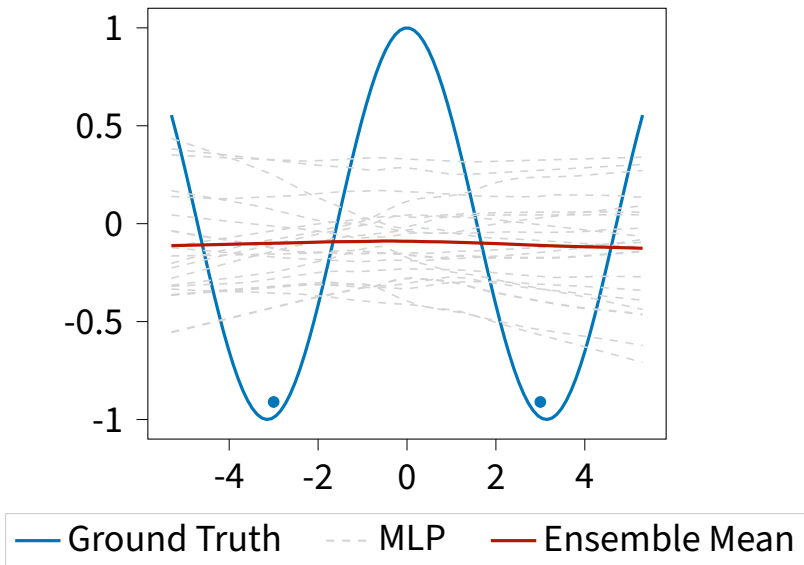
After 2000 Training Steps



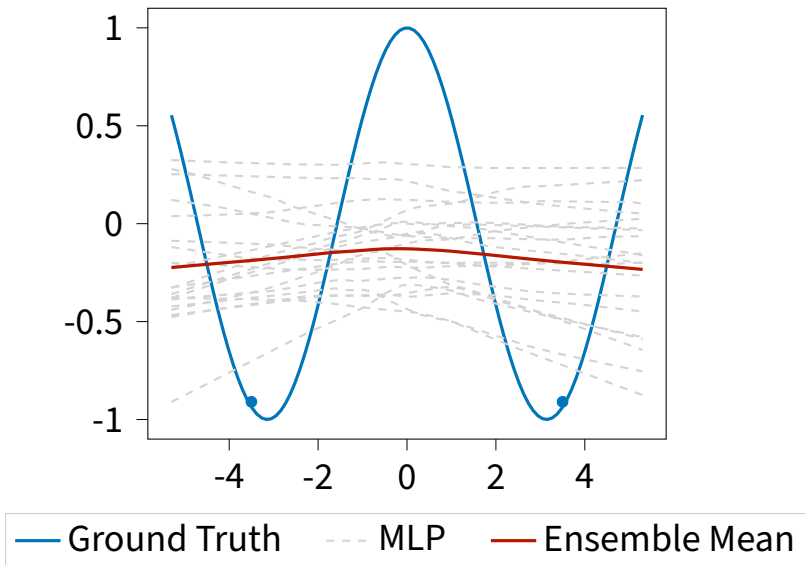
Initialization



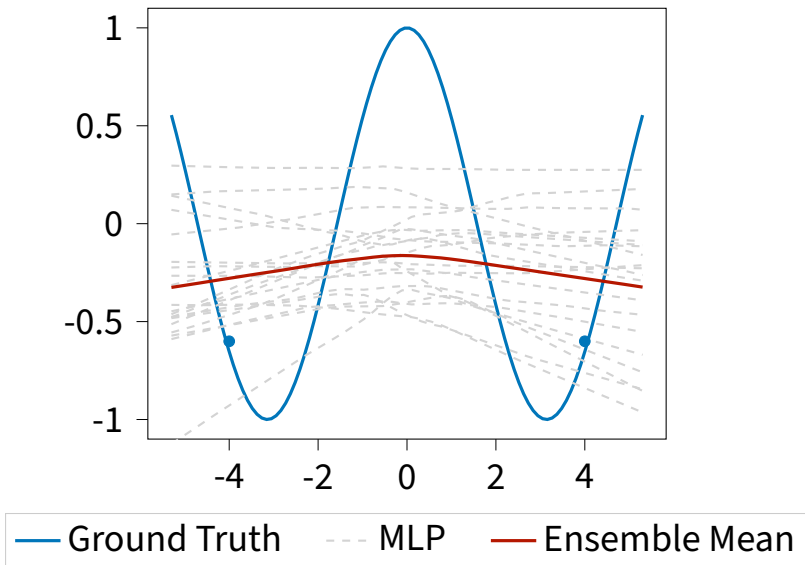
After 1 Training Step



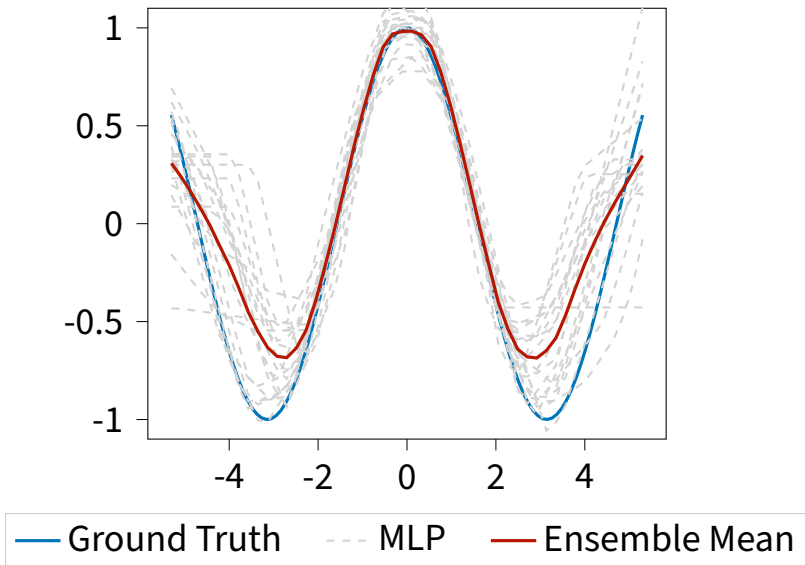
After 2 Training Steps



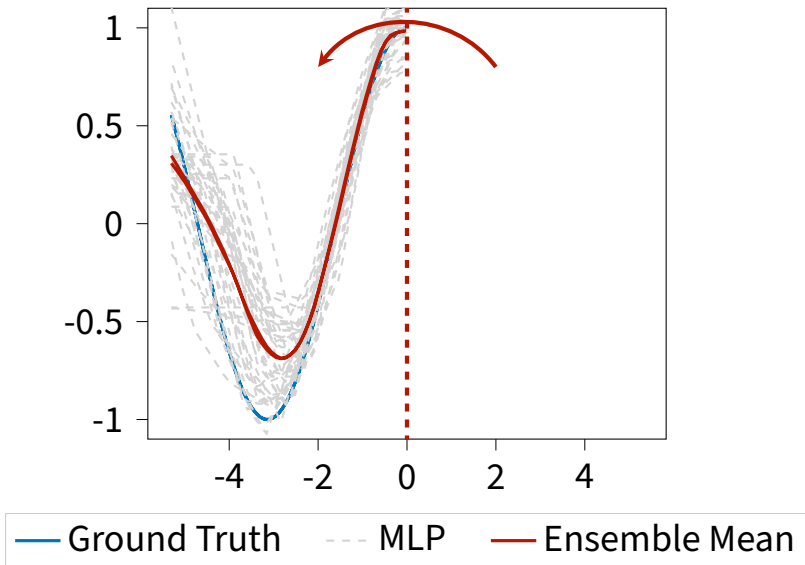
After 3 Training Steps



After 2000 Training Steps

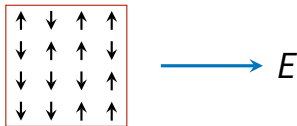


After 2000 Training Steps

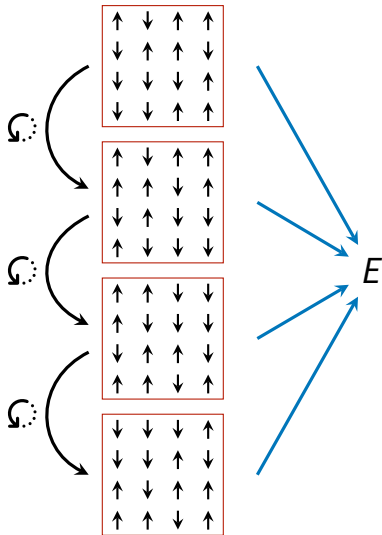


Experiments

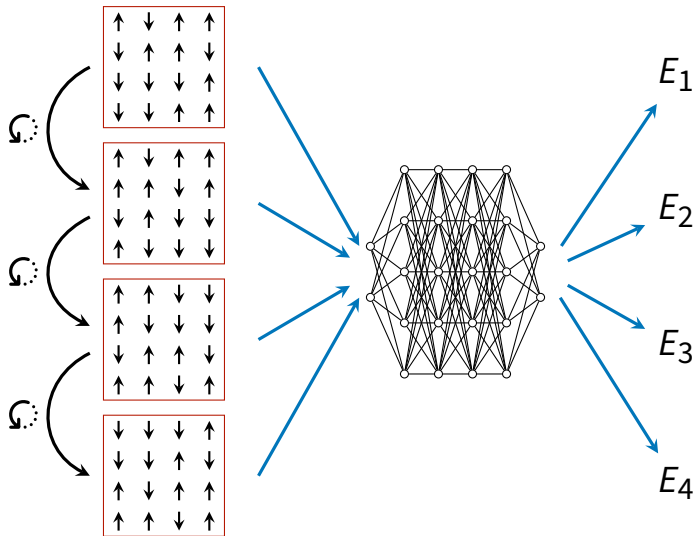
Ising model



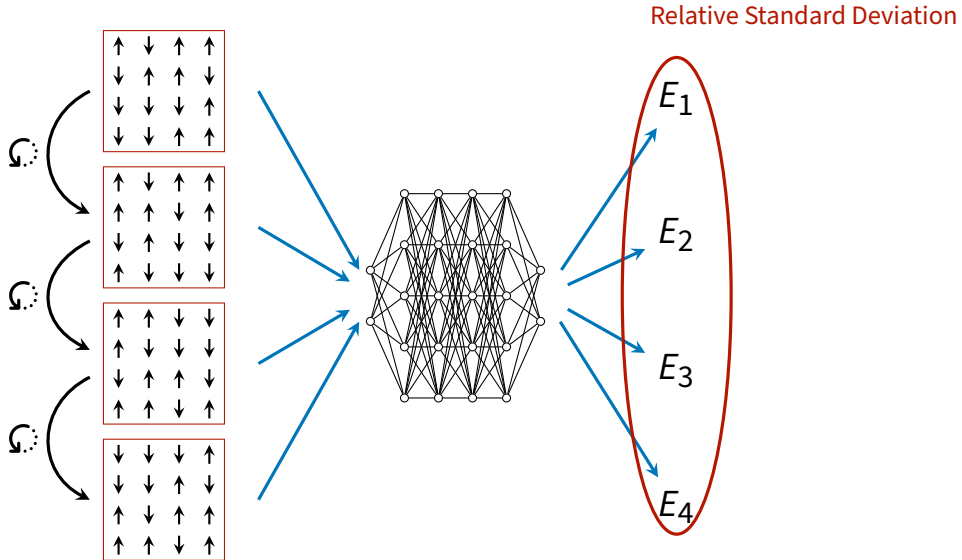
Ising model

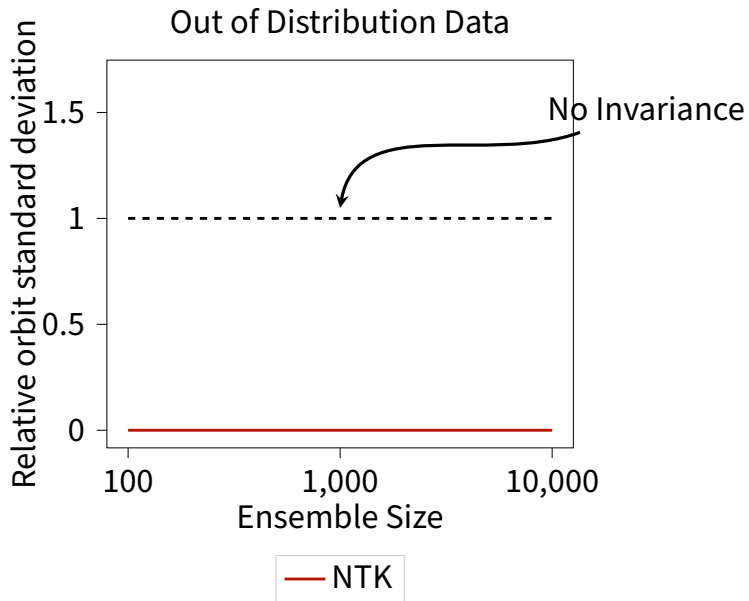


Ising model

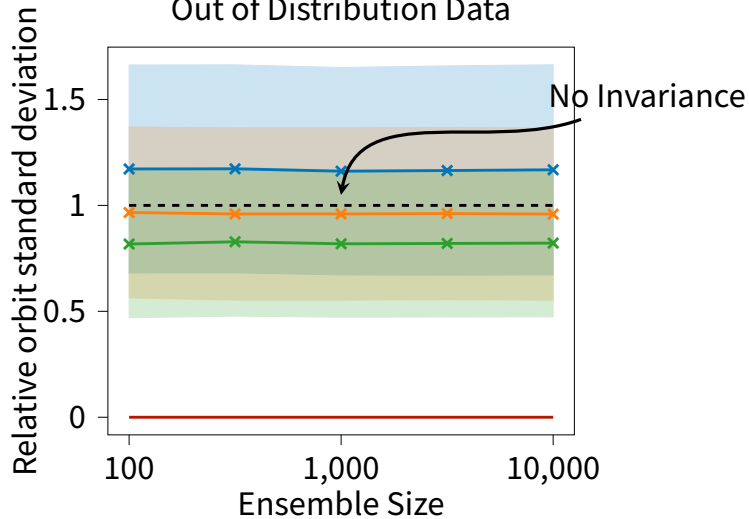


Ising model



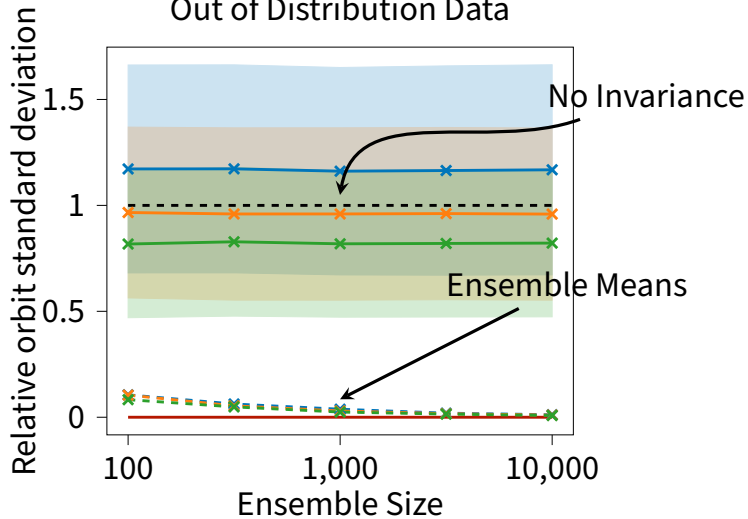


Out of Distribution Data

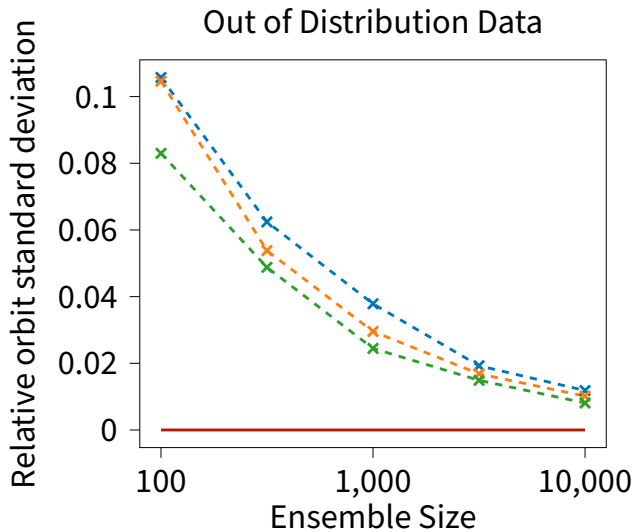


— NTK × Width 512 × Width 1024 × Width 2048

Out of Distribution Data



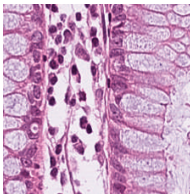
— NTK × Width 512 × Width 1024 × Width 2048



— NTK -x- Width 512 -x- Width 1024 -x- Width 2048

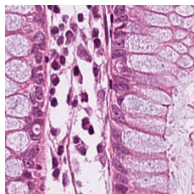
Histological slices

[Kather et al. 2018]



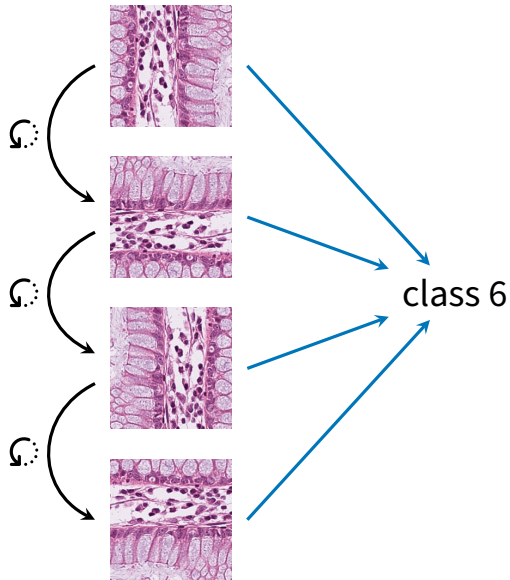
Histological slices

[Kather et al. 2018]

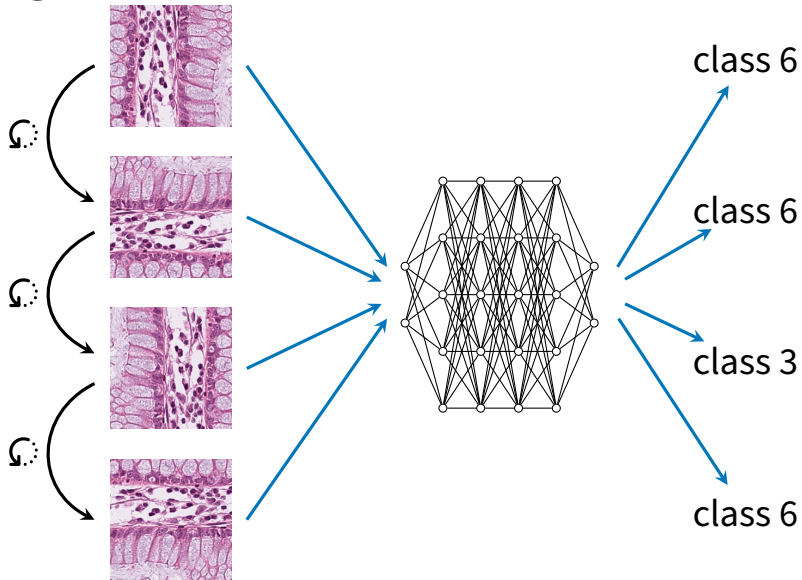


→ class 6

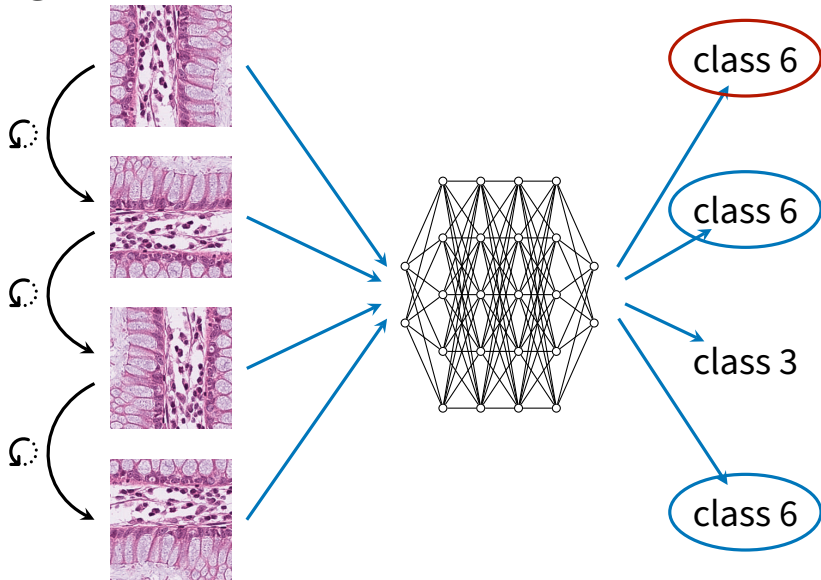
Histological slices



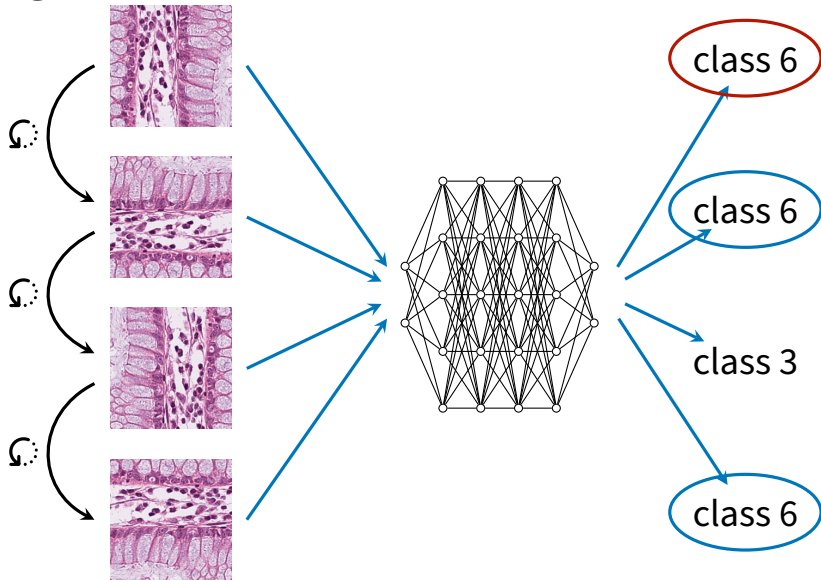
Histological slices



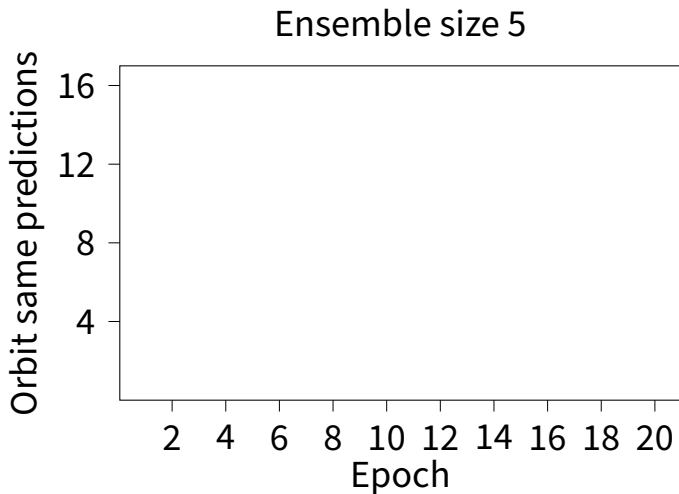
Histological slices



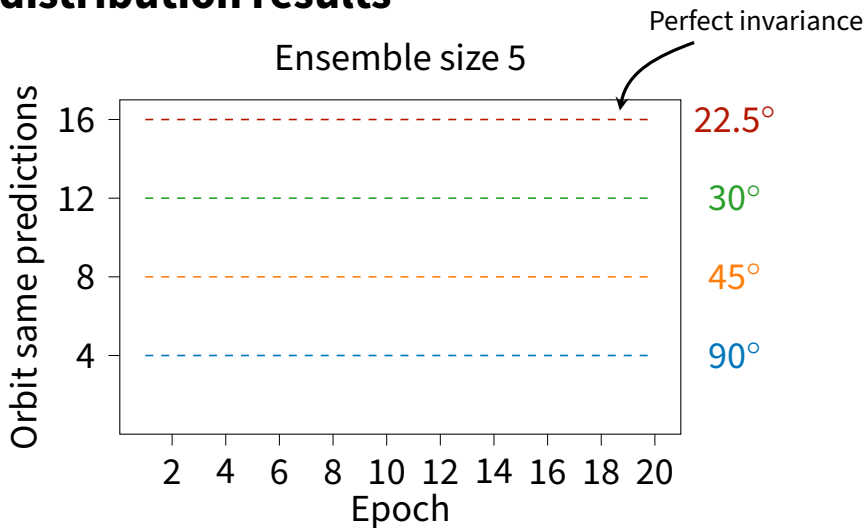
Histological slices



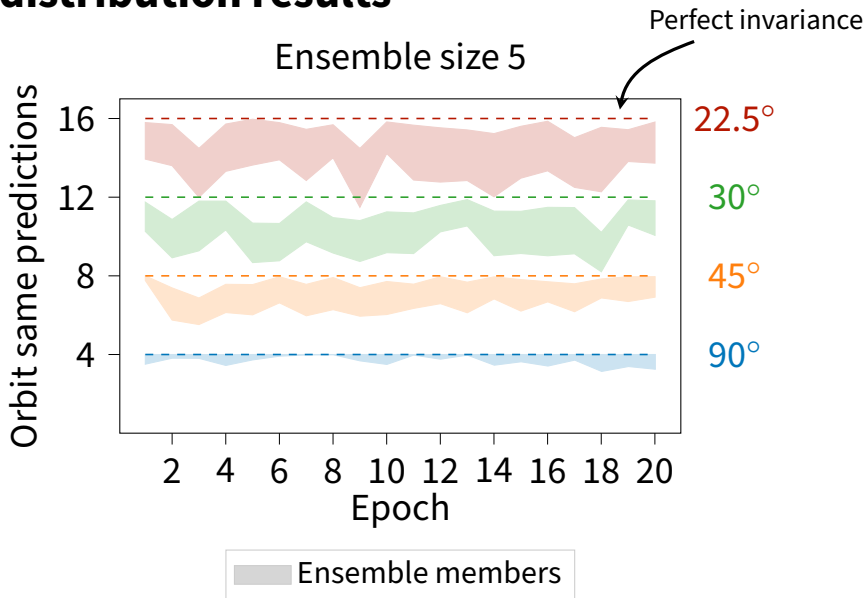
Out of distribution results



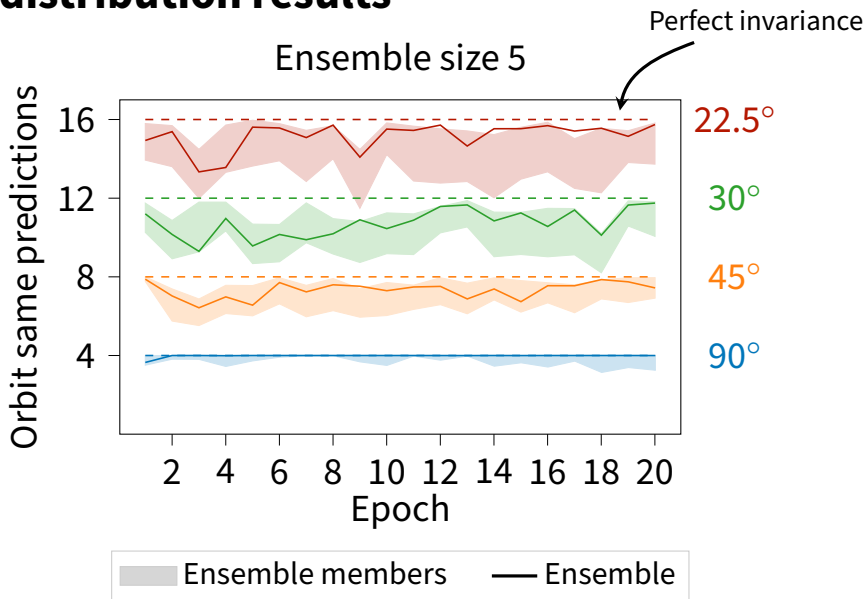
Out of distribution results



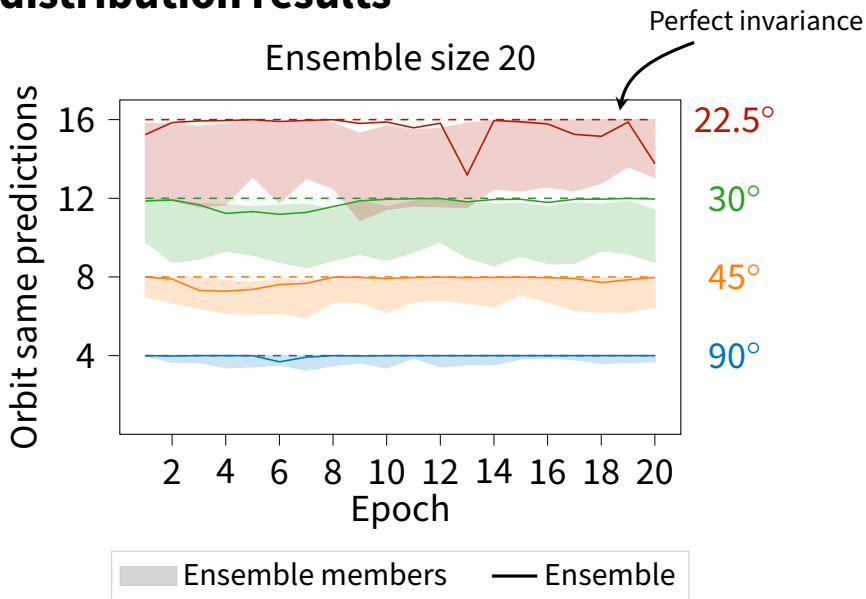
Out of distribution results



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