

Emergent Equivariance in Deep Ensembles

Jan E. Gerken^{*}



UNIVERSITY OF
GOTHENBURG



in collaboration with



from



Prescient
Design
A Genentech Accelerator

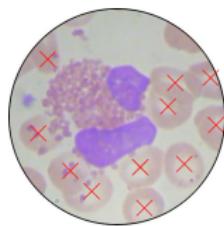
Pan Kessel^{*}

^{*} Equal Contribution

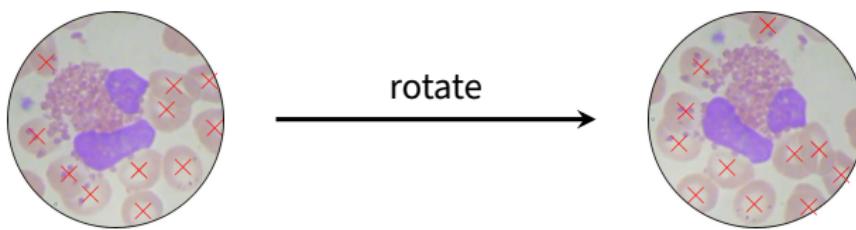


Equivariance

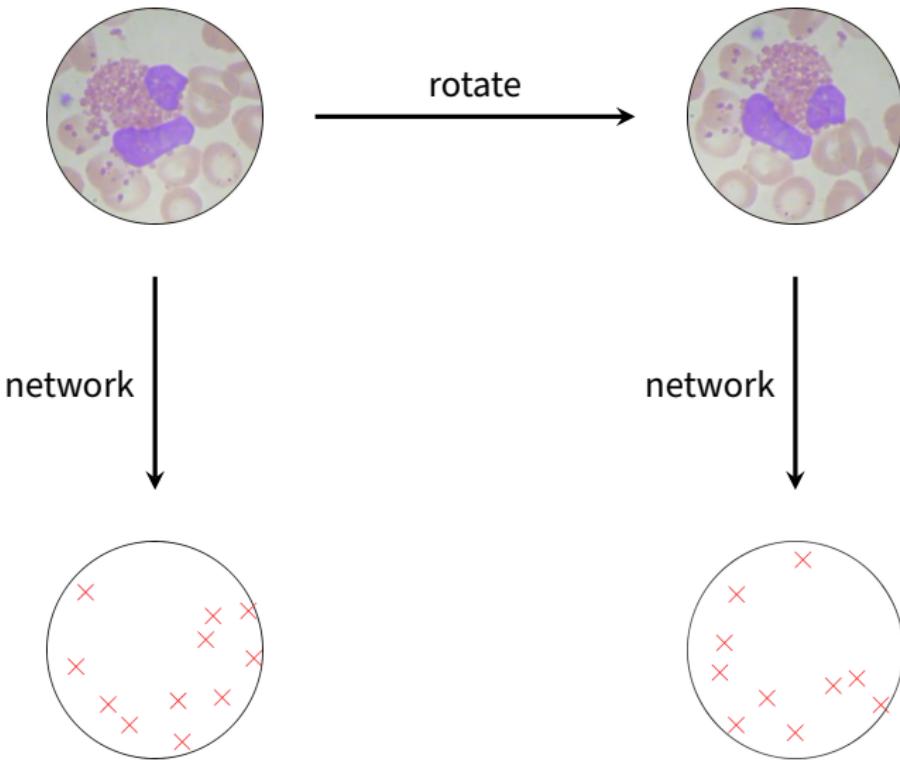
Equivariance



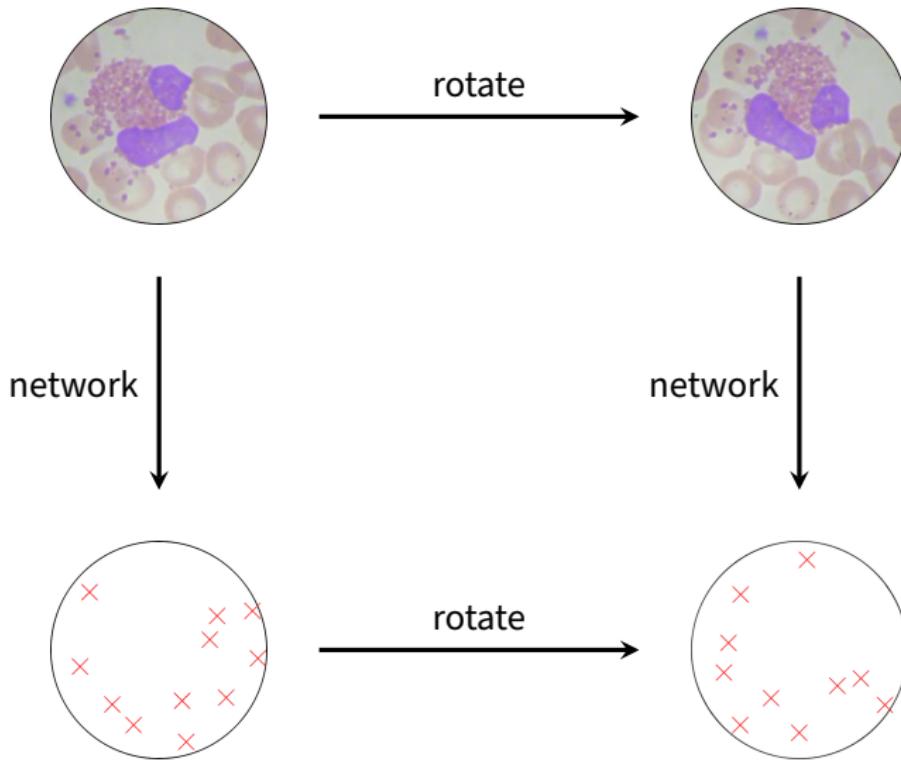
Equivariance



Equivariance

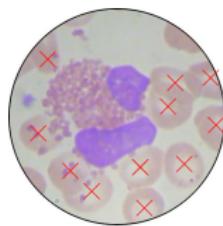


Equivariance

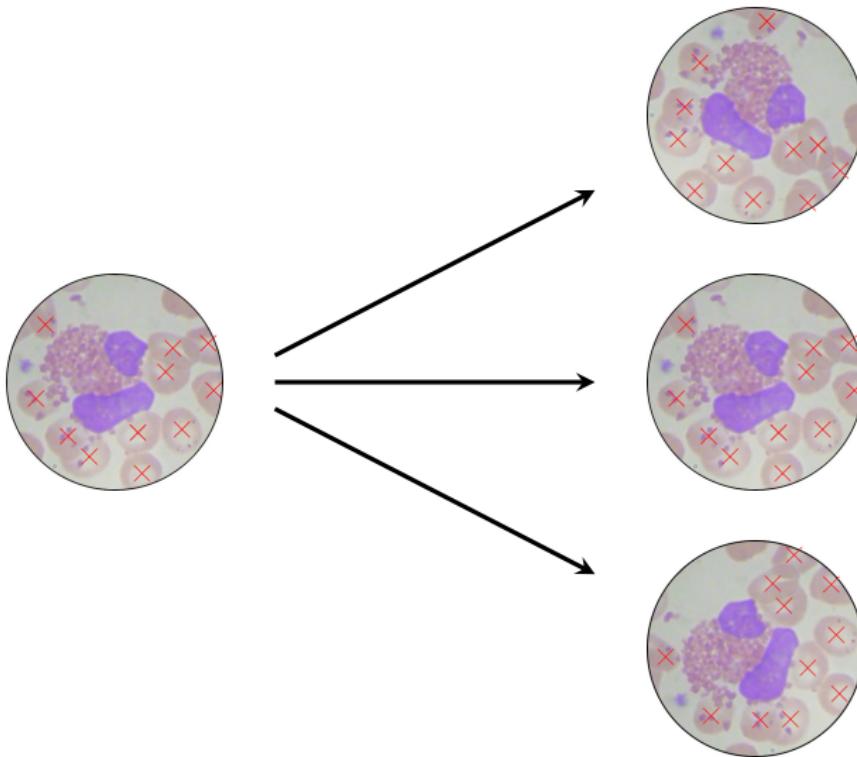


Data augmentation

Data augmentation



Data augmentation



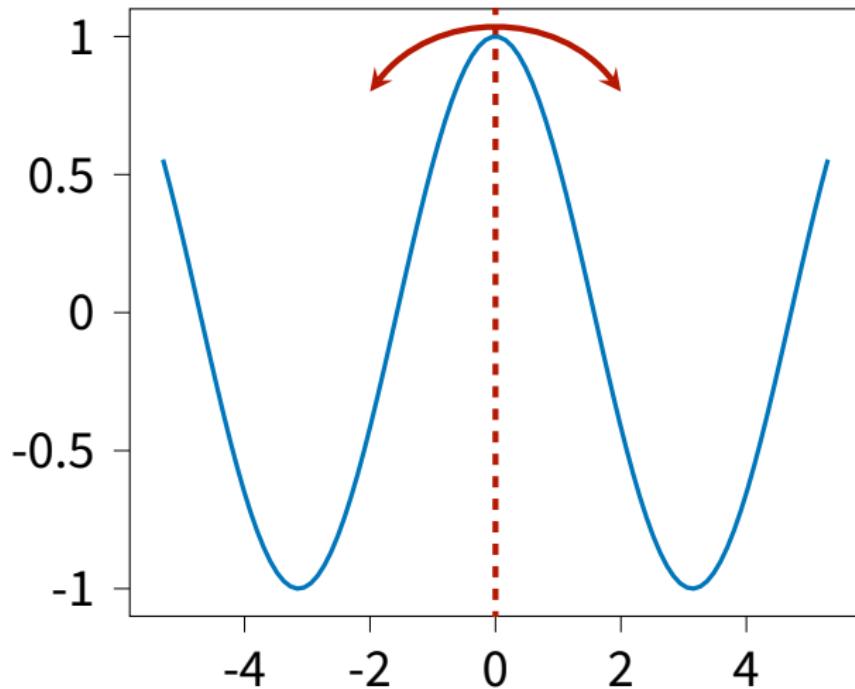
Data augmentation

- thumb-up Easy to implement
- thumb-up No specialized architecture necessary

Data augmentation

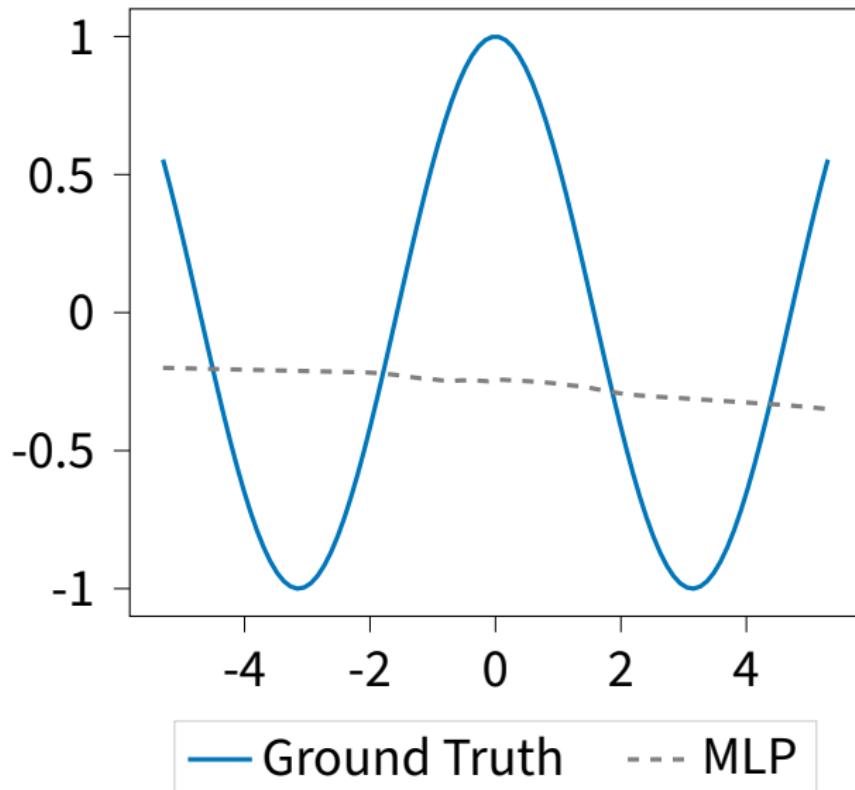
- 👍 Easy to implement
- 👍 No specialized architecture necessary
- 👎 No exact equivariance

Toy example

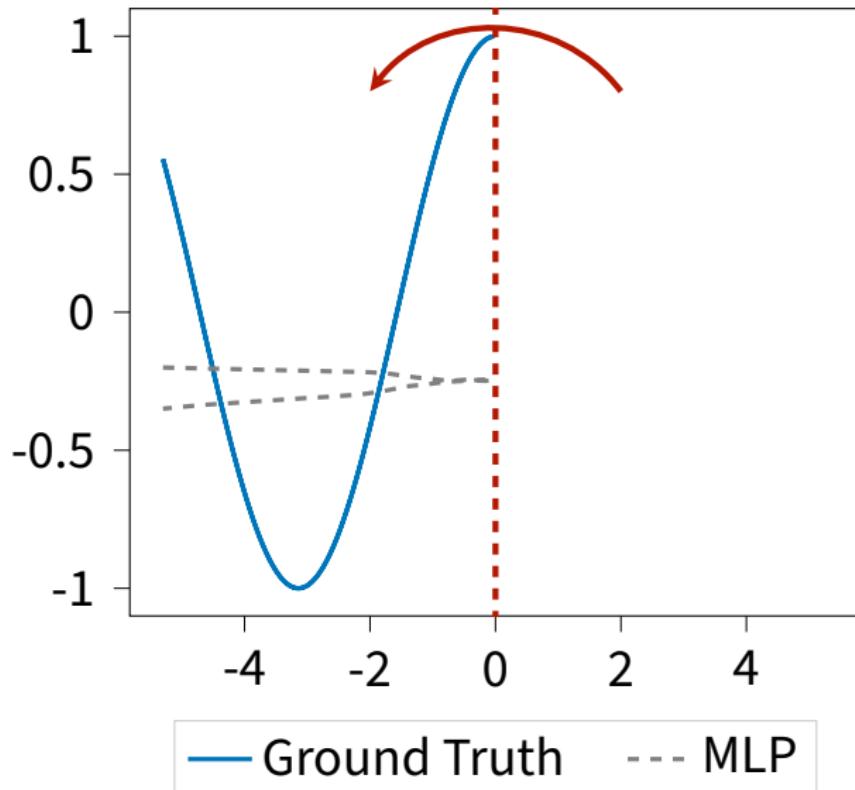


— Ground Truth

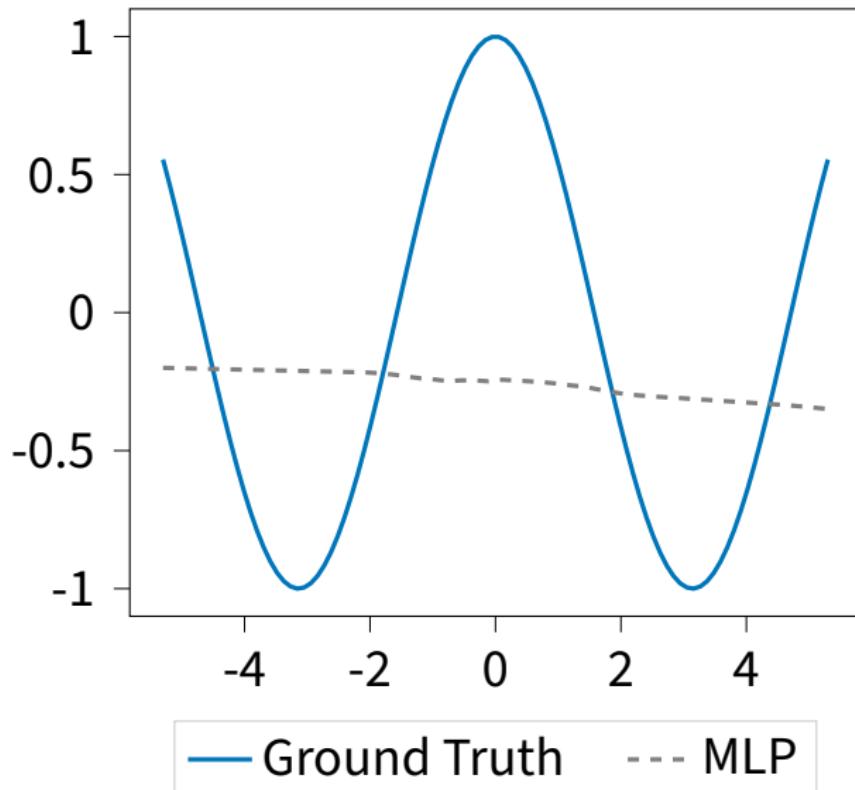
Initialization



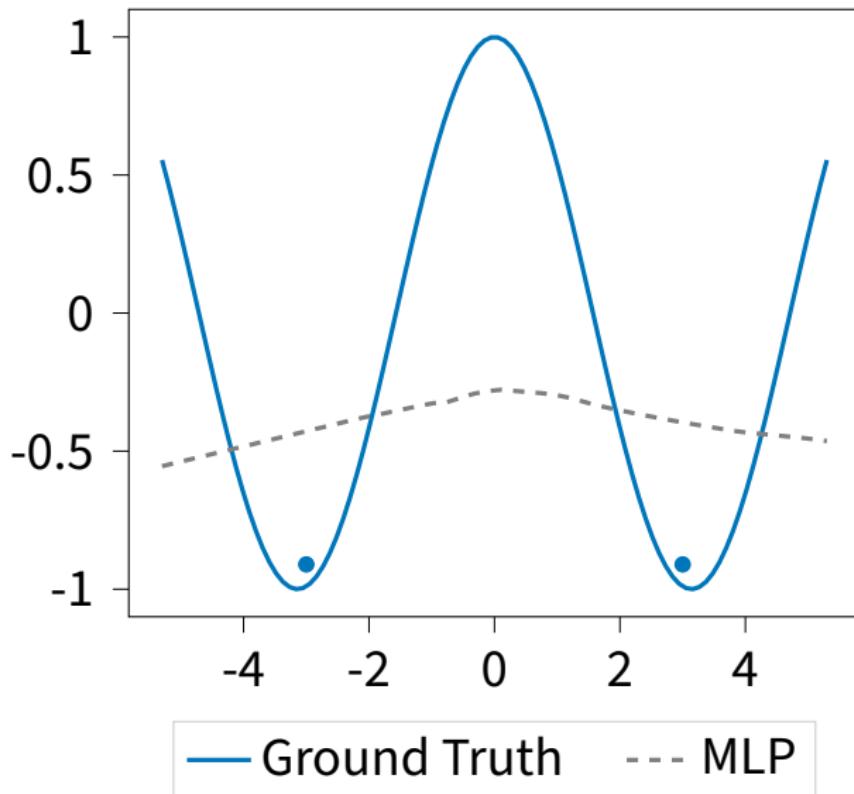
Initialization



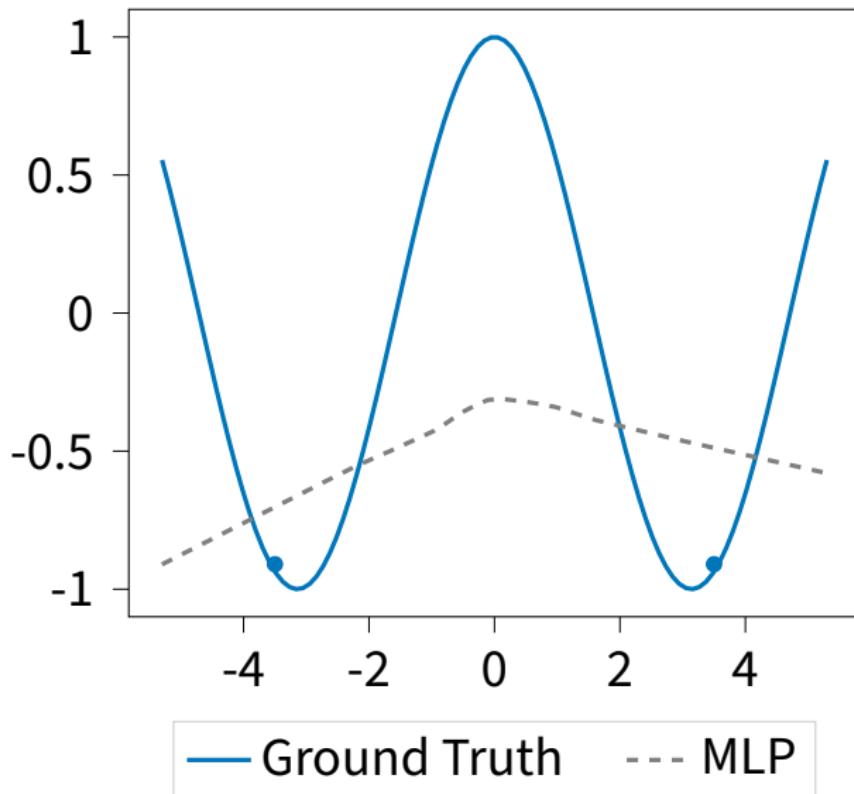
Initialization



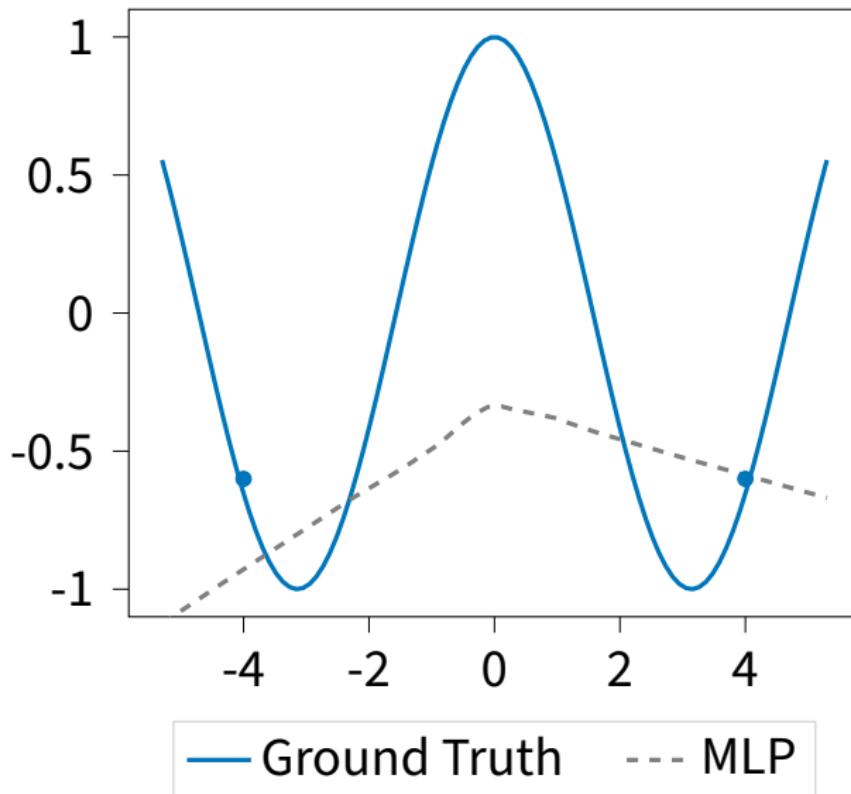
After 1 Training Step



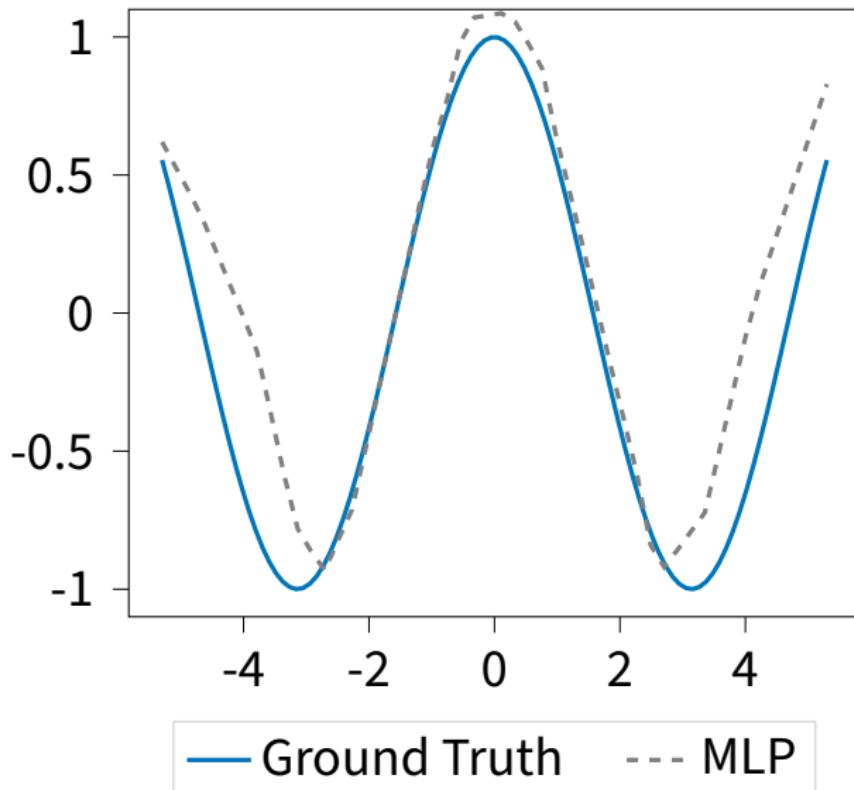
After 2 Training Steps



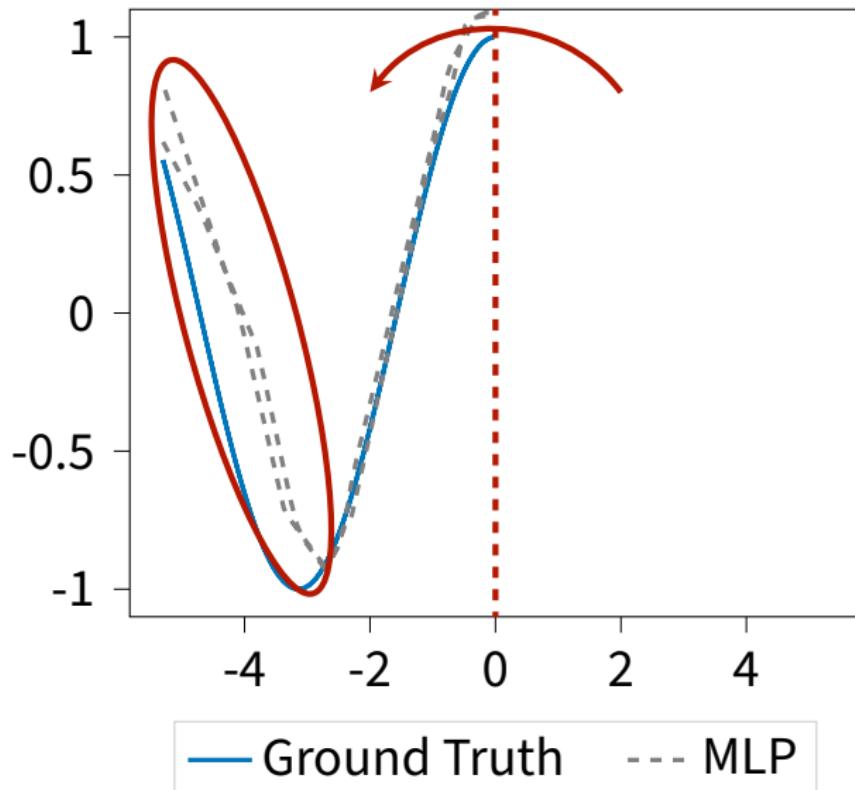
After 3 Training Steps



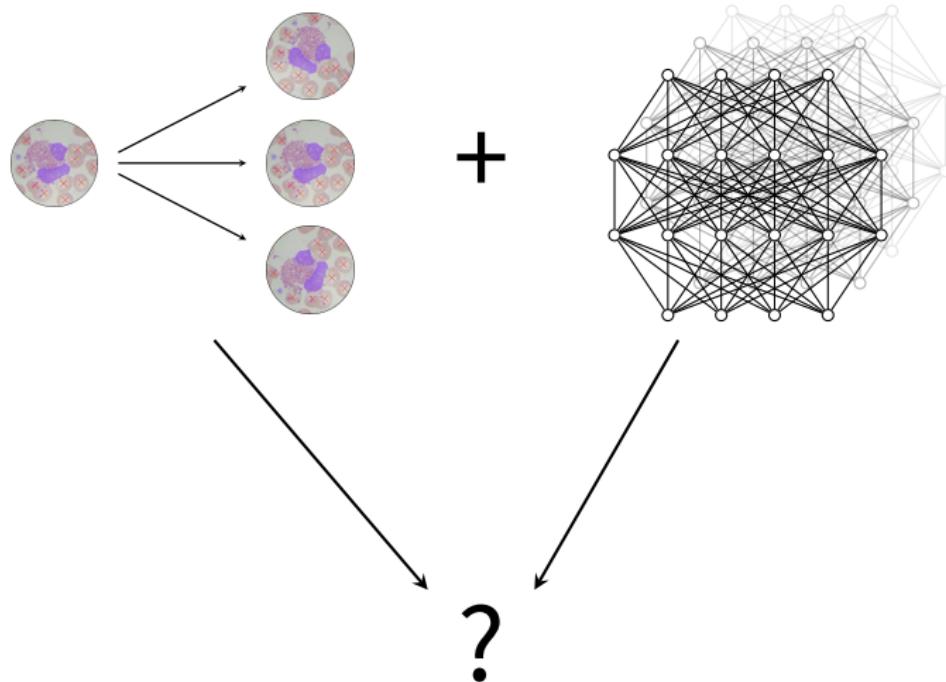
After 2000 Training Steps



After 2000 Training Steps



Can ensembles help?



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

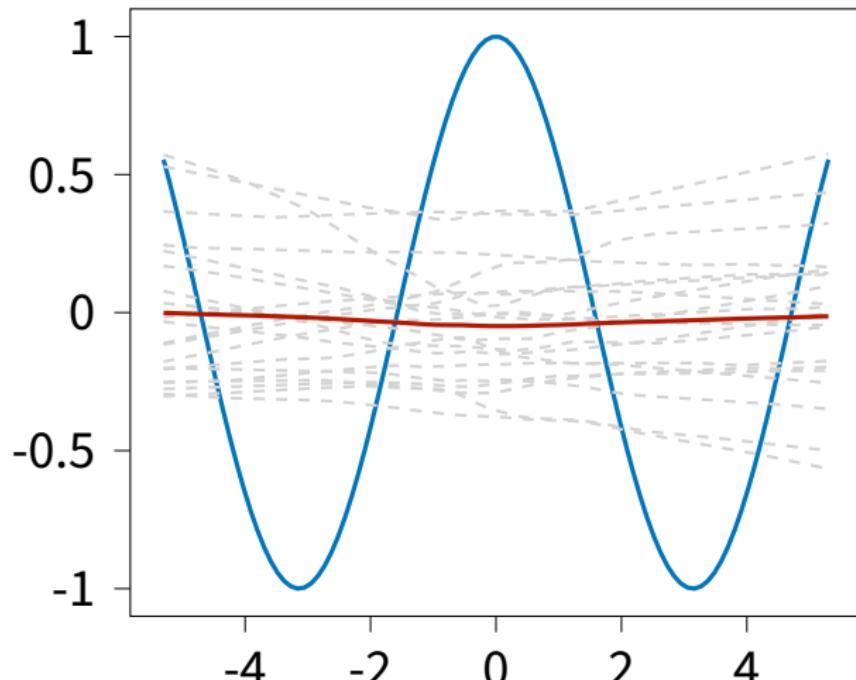
- ✓ Equivariance holds for all training times
 - ✓ Equivariance holds away from the training data
-
- ⌚ At infinite width, the mean output at initialization is zero everywhere.

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.

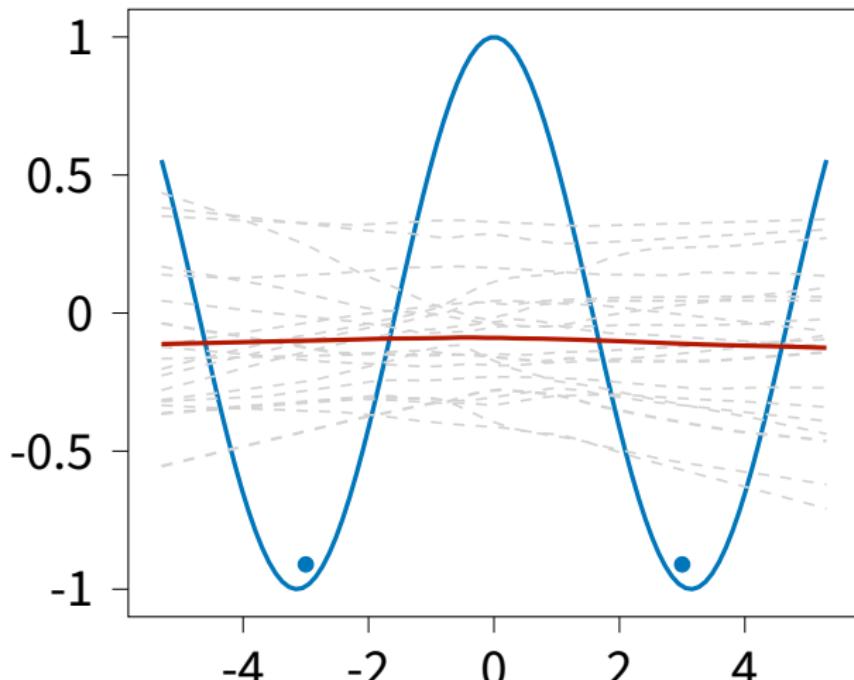
Toy example

Initialization



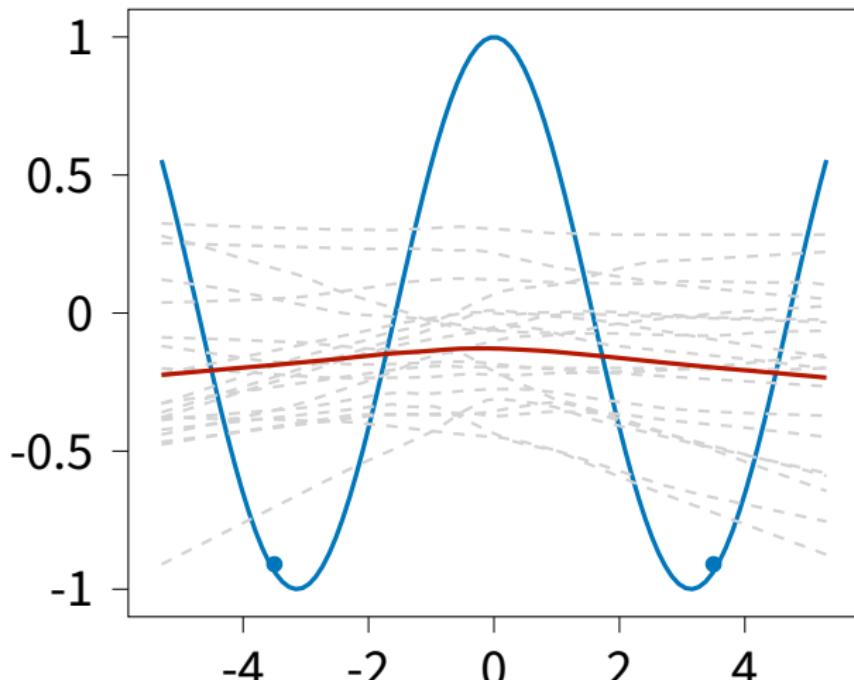
— Ground Truth - - - MLP — Ensemble Mean

After 1 Training Step



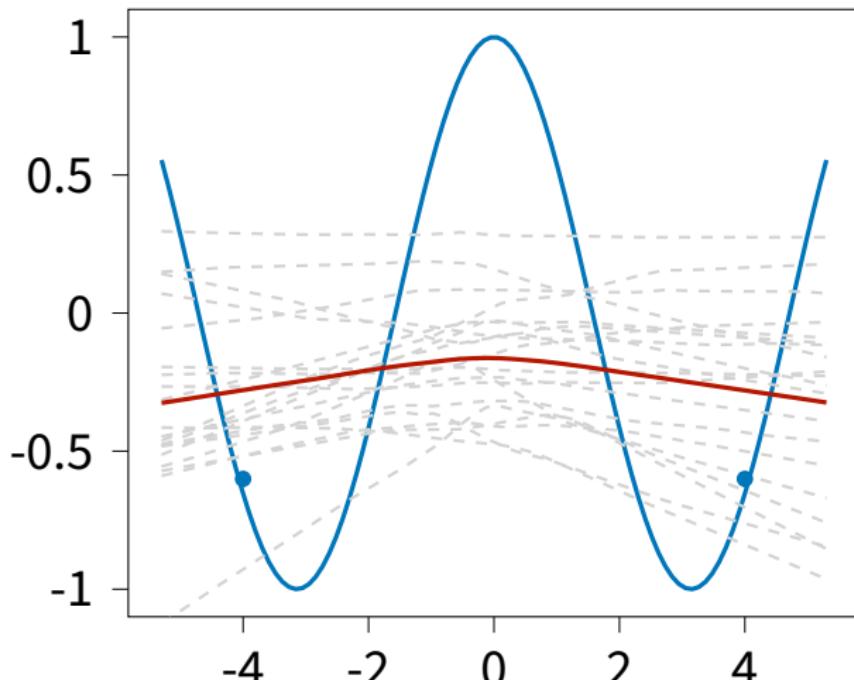
— Ground Truth - - - MLP — Ensemble Mean

After 2 Training Steps



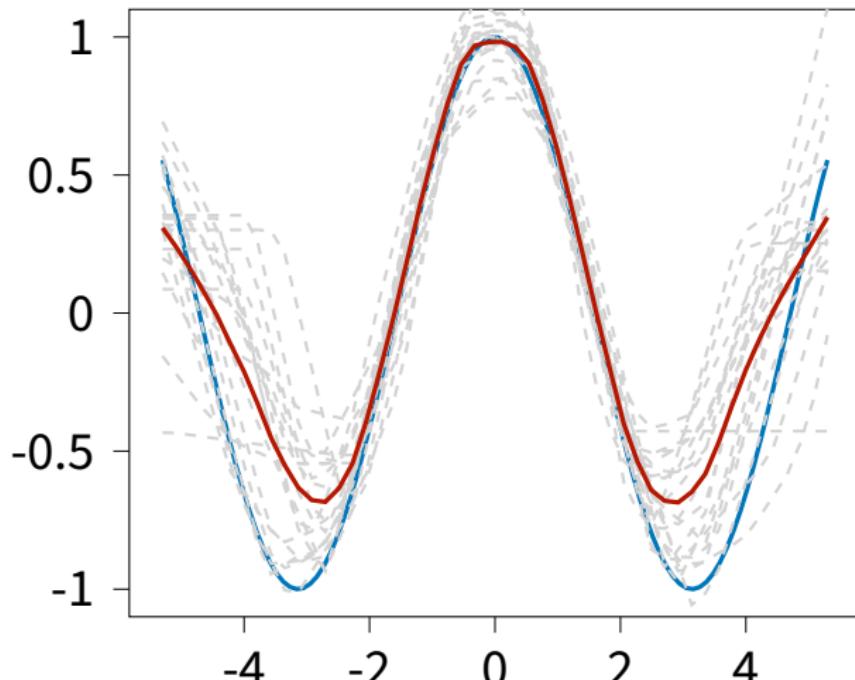
— Ground Truth - - - MLP — Ensemble Mean

After 3 Training Steps



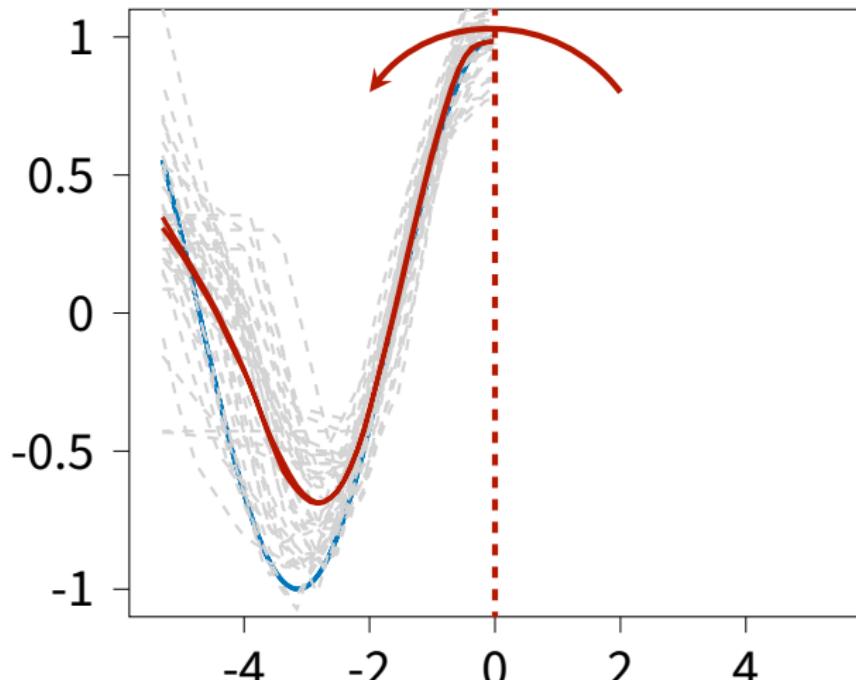
— Ground Truth - - - MLP — Ensemble Mean

After 2000 Training Steps



— Ground Truth - - - MLP — Ensemble Mean

After 2000 Training Steps



— Ground Truth - - - MLP — Ensemble Mean

Proof idea

Mean prediction

$$\underbrace{\frac{1}{n} \sum_{\theta_0=\text{init}_1}^{\text{init}_n} f_{\theta_t}(x)}_{\text{mean prediction of deep ensemble}}$$

test point

Mean prediction

$$\lim_{n \rightarrow \infty} \underbrace{\frac{1}{n} \sum_{\theta_0=\text{init}_1}^{\text{init}_n} f_{\theta_t}(x)}_{\text{mean prediction of deep ensemble}}$$

test point

Mean prediction

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

test point

Mean prediction

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

test point

Mean prediction from NTK

[Jacot et al. 2018]

- ① At infinite width, the mean prediction is given in terms of the neural tangent kernel (NTK)

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

Mean prediction from NTK

[Jacot et al. 2018]

- At infinite width, the mean prediction is given in terms of the neural tangent kernel (NTK)

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

neural tangent kernel



Mean prediction from NTK

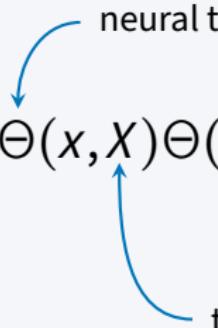
[Jacot et al. 2018]

- At infinite width, the mean prediction is given in terms of the neural tangent kernel (NTK)

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

neural tangent kernel

train data



Mean prediction from NTK

[Jacot et al. 2018]

- At infinite width, the mean prediction is given in terms of the neural tangent kernel (NTK)

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

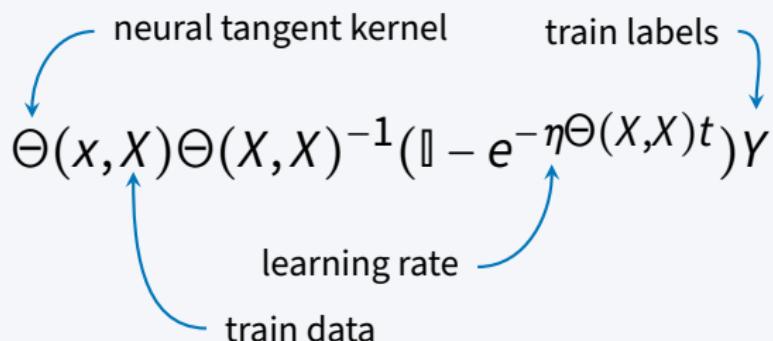
Diagram illustrating the components of the mean prediction formula:

- neural tangent kernel**: Points to the term $\Theta(X, X)^{-1}$.
- learning rate**: Points to the term $e^{-\eta \Theta(X, X)t}$.
- train data**: Points to the term Y .

Mean prediction from NTK

[Jacot et al. 2018]

- At infinite width, the mean prediction is given in terms of the neural tangent kernel (NTK)

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


The diagram illustrates the components of the mean prediction formula. Blue arrows indicate the inputs to the formula:

- A blue arrow labeled "train data" points to the matrix $\Theta(X, X)$.
- A blue arrow labeled "learning rate" points to the scalar $e^{-\eta \Theta(X, X)t}$.
- A blue arrow labeled "train labels" points to the vector Y .
- A blue arrow labeled "neural tangent kernel" points to the term $\Theta(x, X)$.

Proof idea

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

Proof idea

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

augmented data

augmented labels

Proof idea

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 components of the group transformation equation. It shows the equation $\mu_t(\rho(g)x) = \Theta(\rho(g)x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X, X)t})Y$ with four main terms. A blue arrow points from the label "group transformation" to the first term $\Theta(\rho(g)x, X)$. Another blue arrow points from the label "augmented data" to the second term $\Theta(X, X)^{-1}$. A third blue arrow points from the label "augmented labels" to the third term $(\mathbb{I} - e^{-\eta\Theta(X, X)t})Y$. The fourth term $e^{-\eta\Theta(X, X)t}$ is shown without an arrow, indicating it is a scalar factor.

Proof idea

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

group transformation

for augmented data

augmented data

augmented labels

The diagram illustrates the components of the equation. A blue bracket labeled "group transformation" points to the term $\rho(g)x$. Another blue bracket labeled "for augmented data" points to the entire right side of the equation. A blue arrow labeled "augmented data" points to the term $\Theta(X, X)^{-1}$. A blue arrow labeled "augmented labels" points to the term Y .

Proof idea

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

group transformation

augmented data

augmented labels

Proof idea

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

group transformation

augmented labels

for invariance

The diagram illustrates the proof idea for group invariance. It shows the equation $\mu_t(\rho(g)x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X, X)t})\rho(g)Y$. Two blue arrows point to specific terms in the equation: one arrow points from $\rho(g)x$ to the text "group transformation", and another arrow points from $\rho(g)Y$ to the text "augmented labels" and "for invariance".

Proof idea

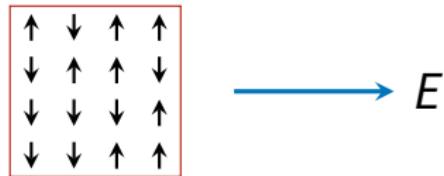
group transformation


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

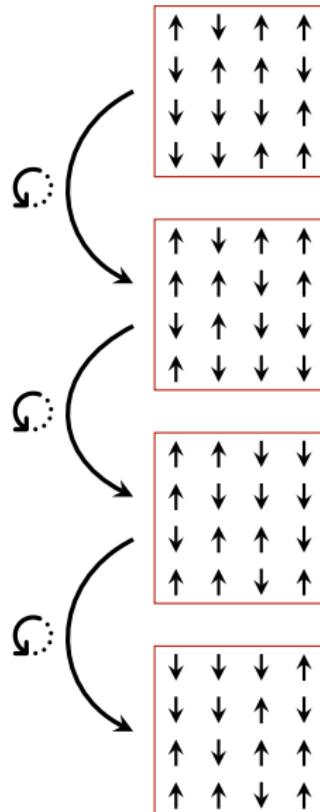
for invariance

Experiments

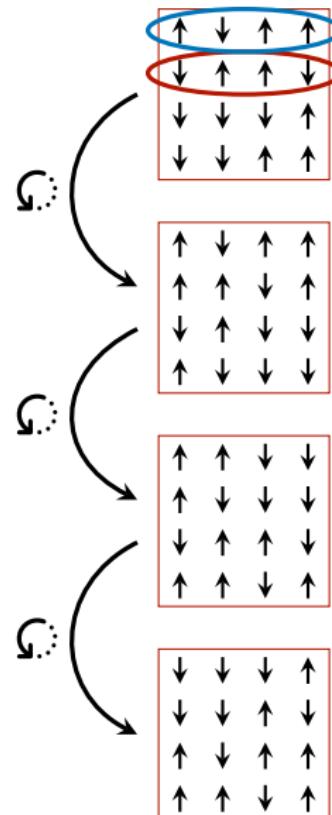
Ising model



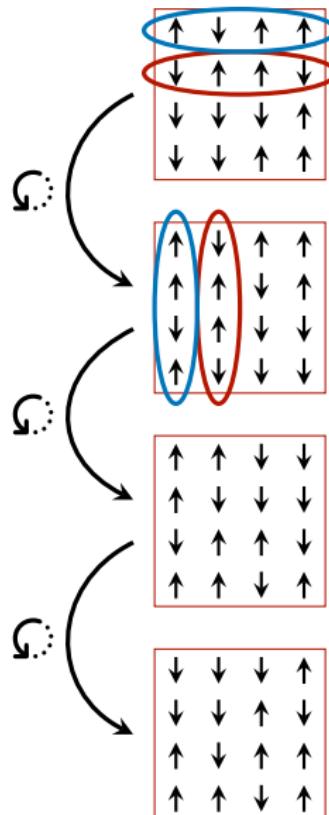
Ising model



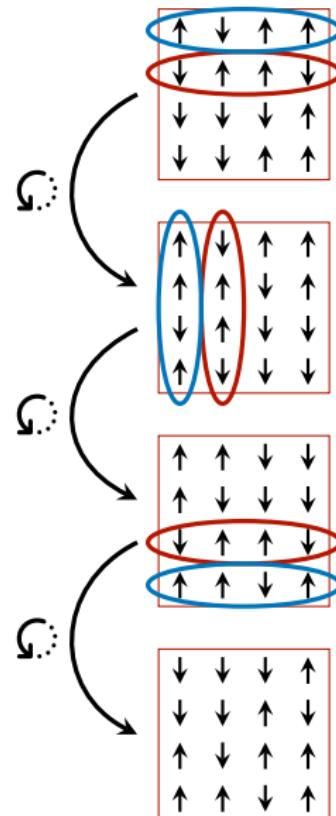
Ising model



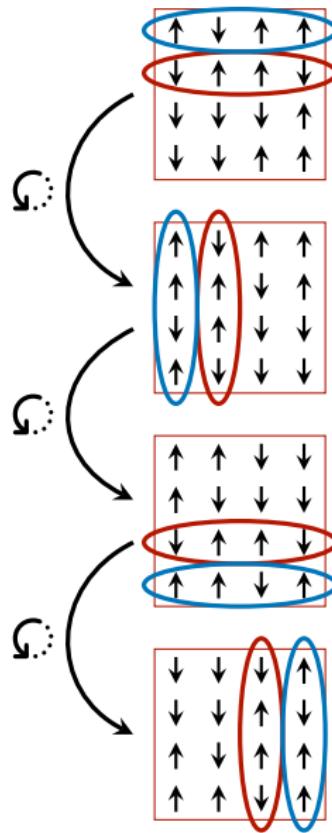
Ising model



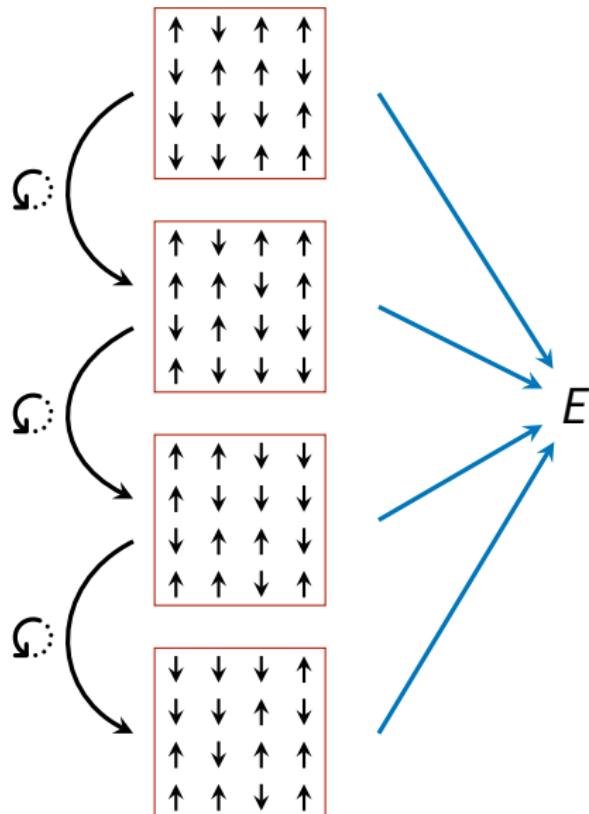
Ising model



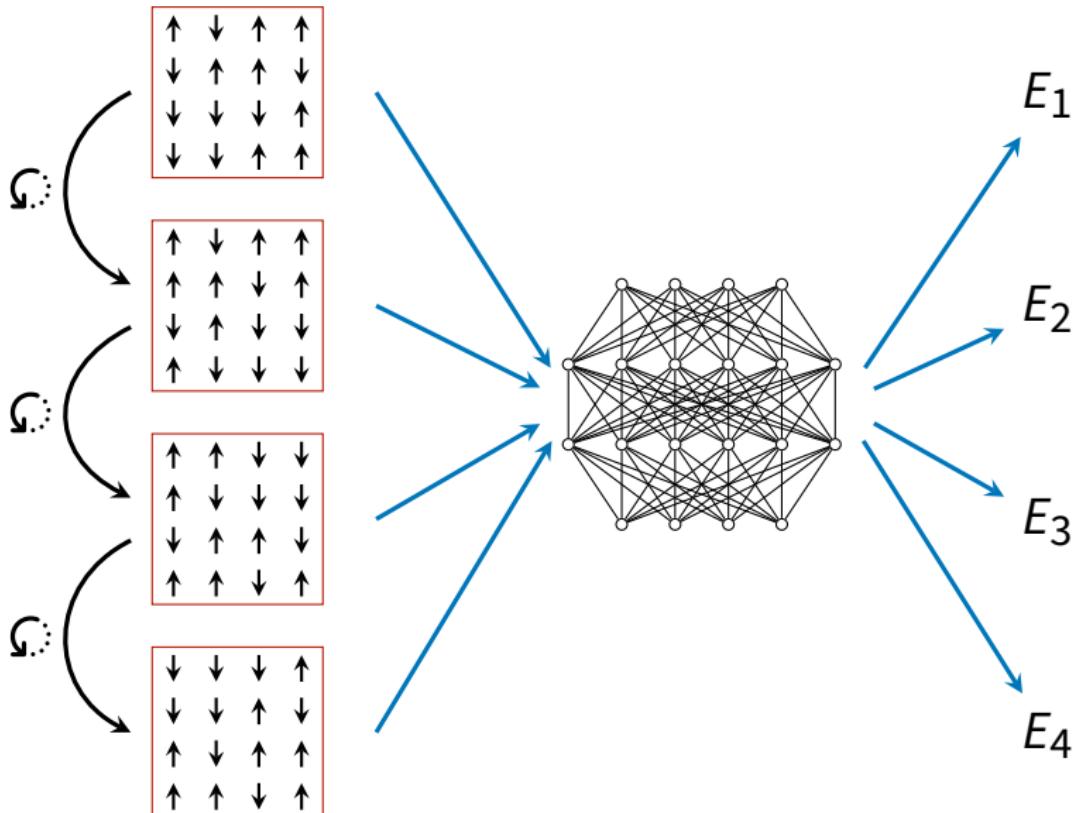
Ising model



Ising model

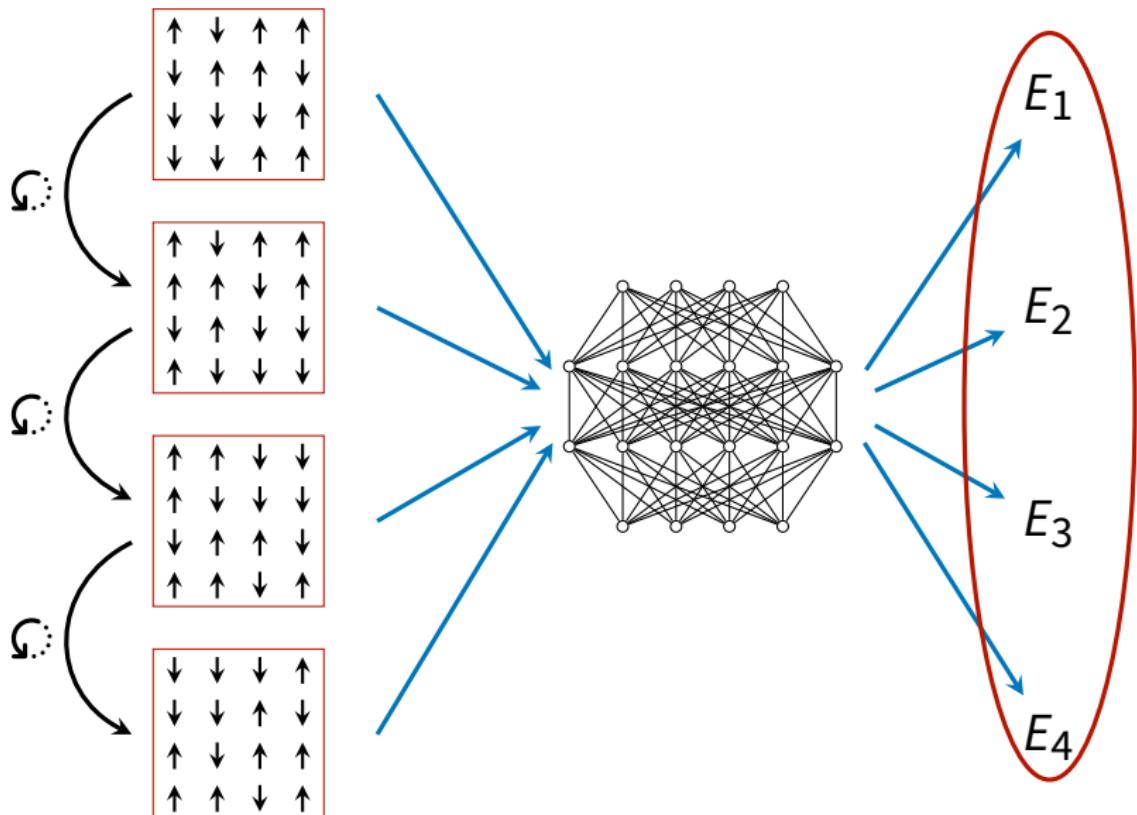


Ising model

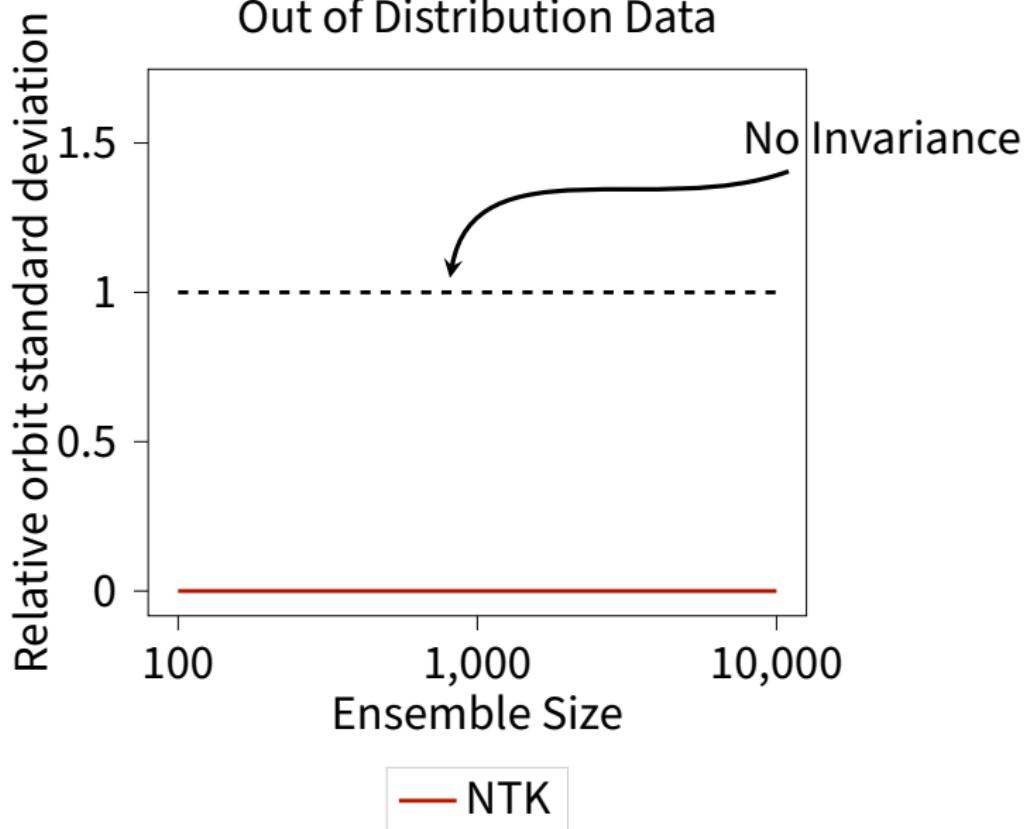


Ising model

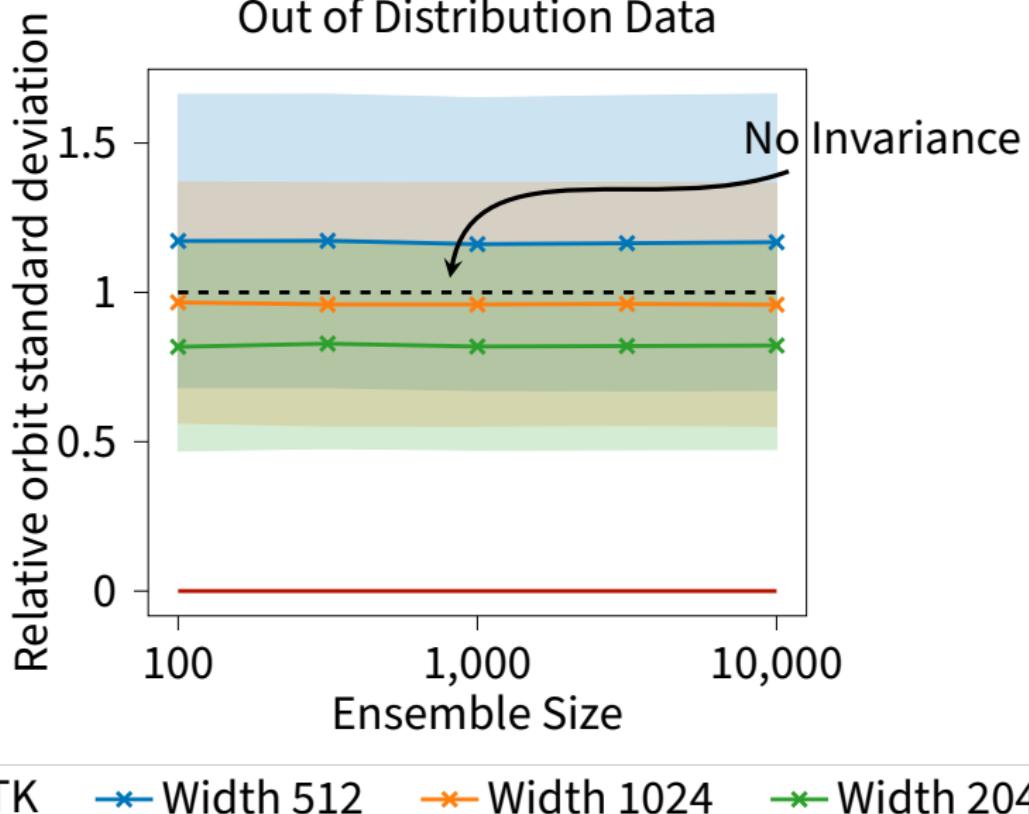
Relative Standard Deviation



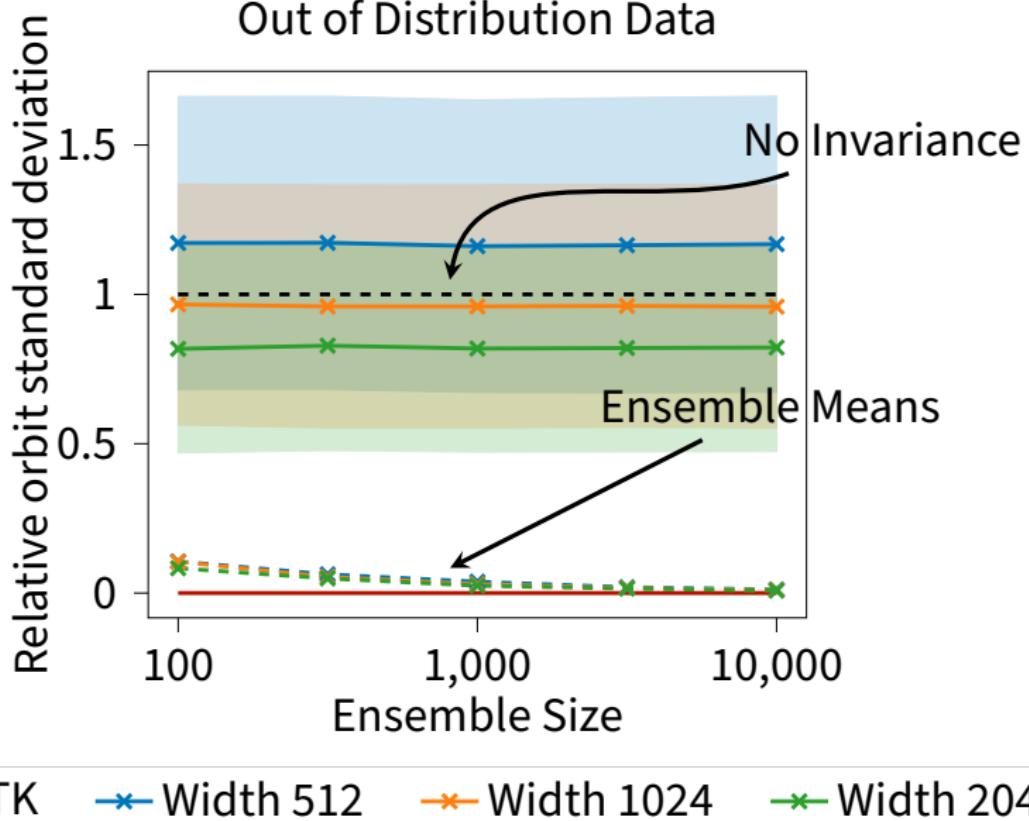
Out of Distribution Data



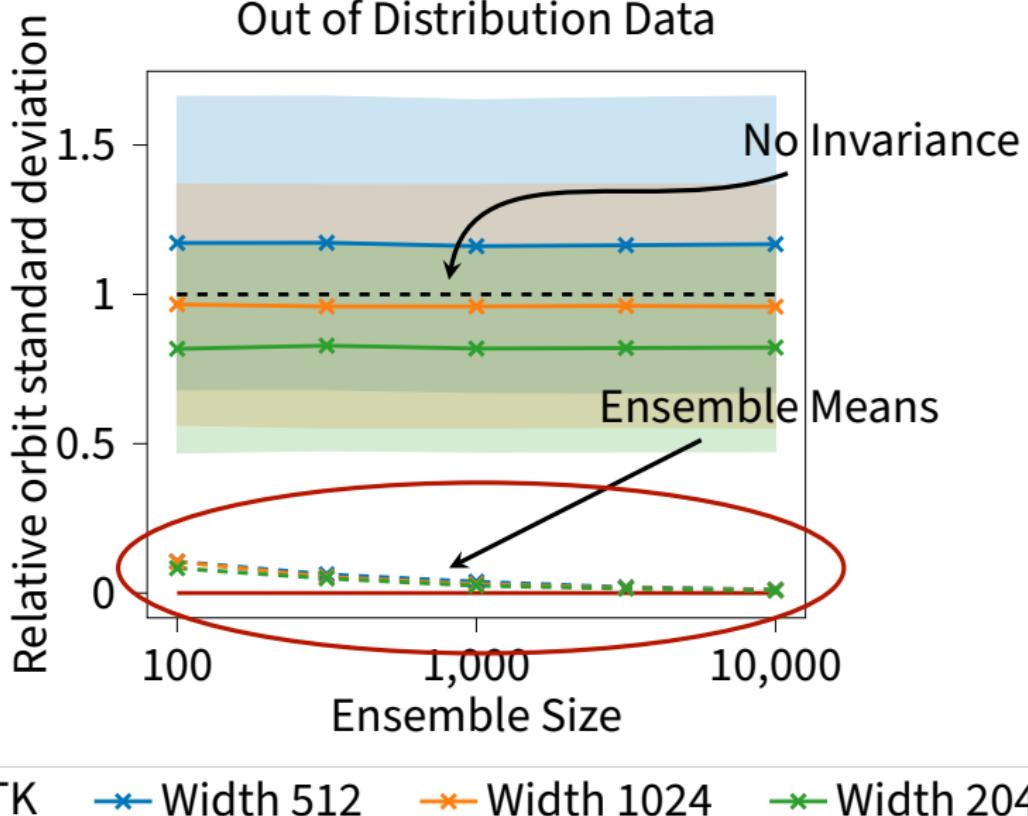
Out of Distribution Data



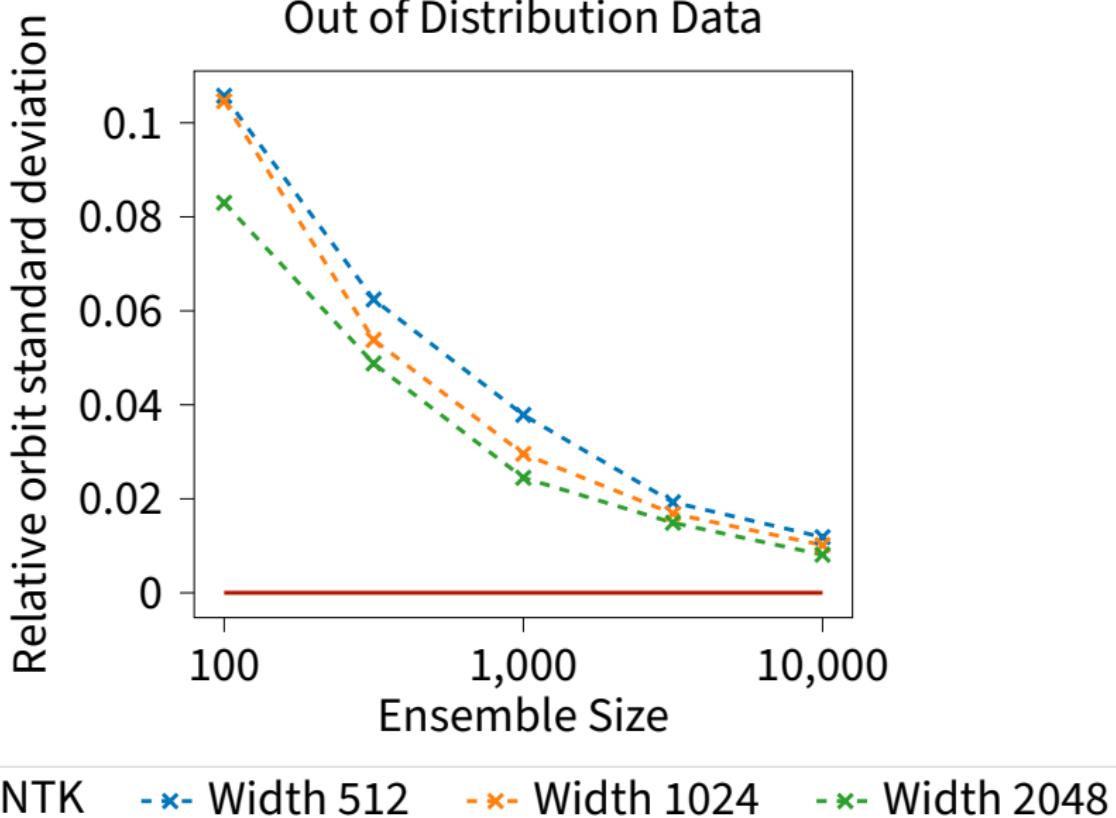
Out of Distribution Data



Out of Distribution Data

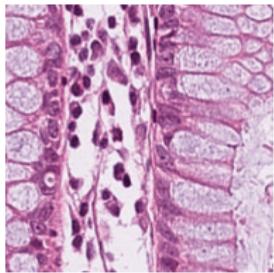


Out of Distribution Data



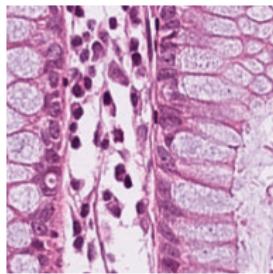
Histological slices

[Kather et al. 2018]



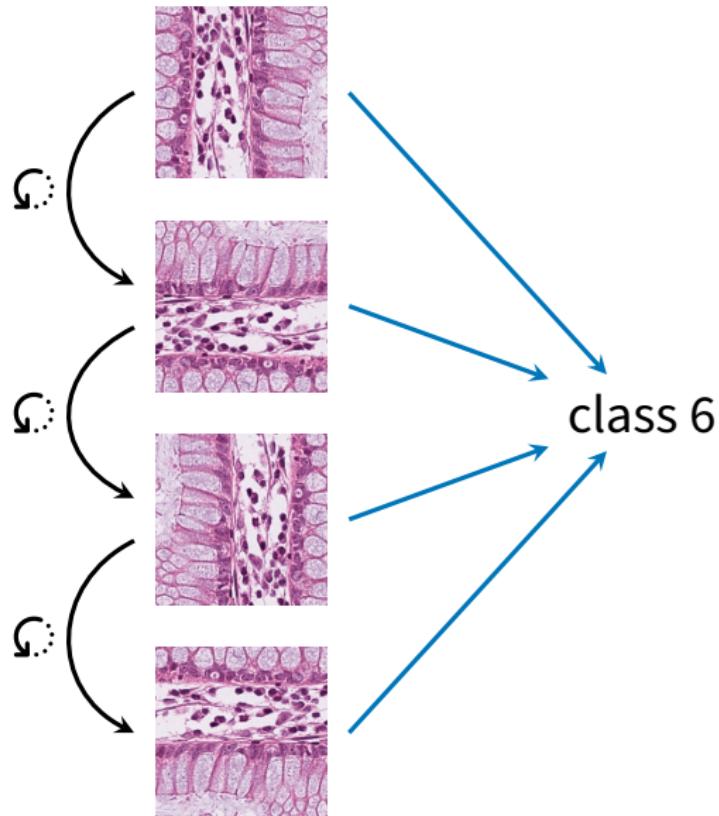
Histological slices

[Kather et al. 2018]

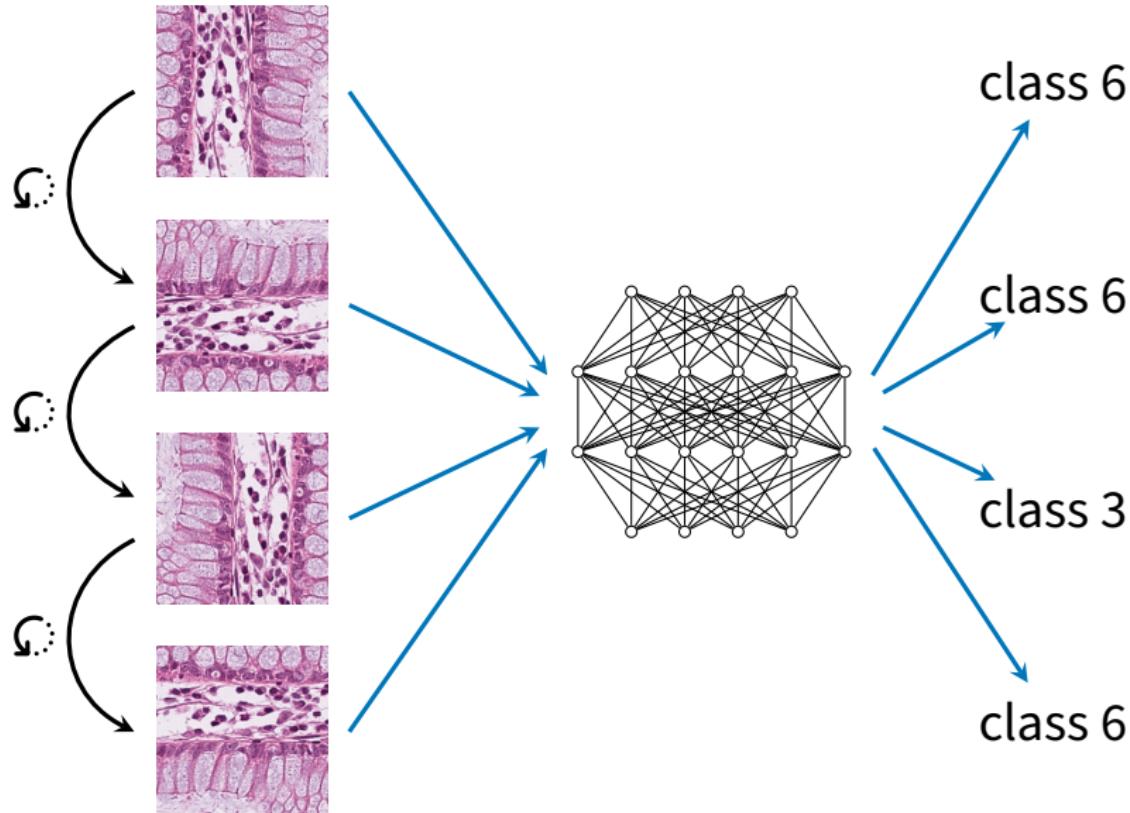


→ class 6

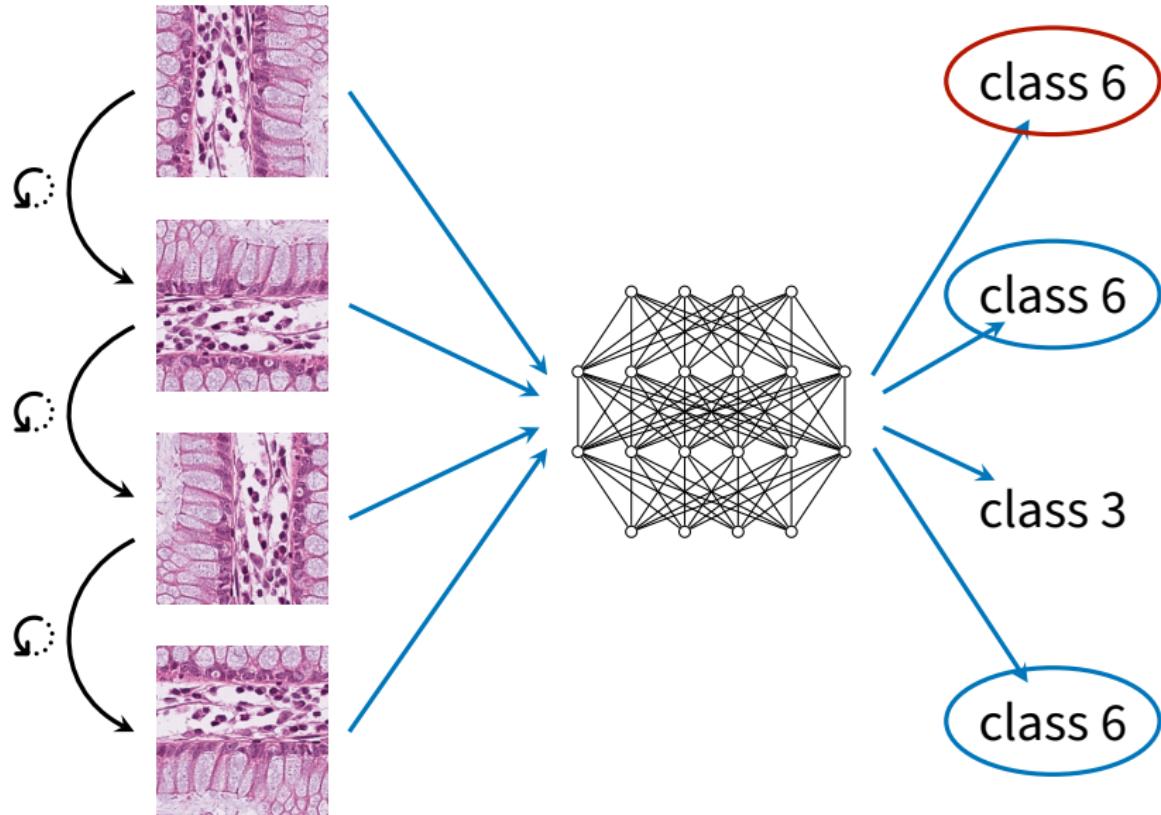
Histological slices



Histological slices

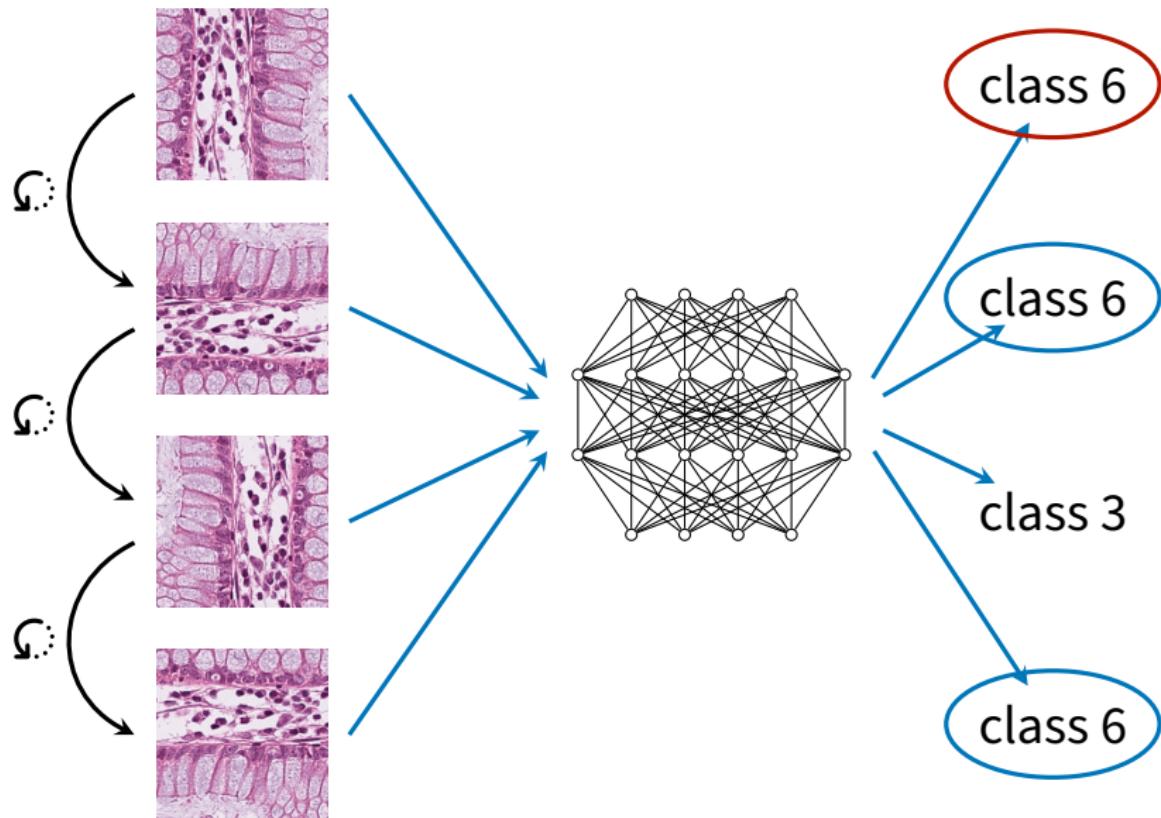


Histological slices



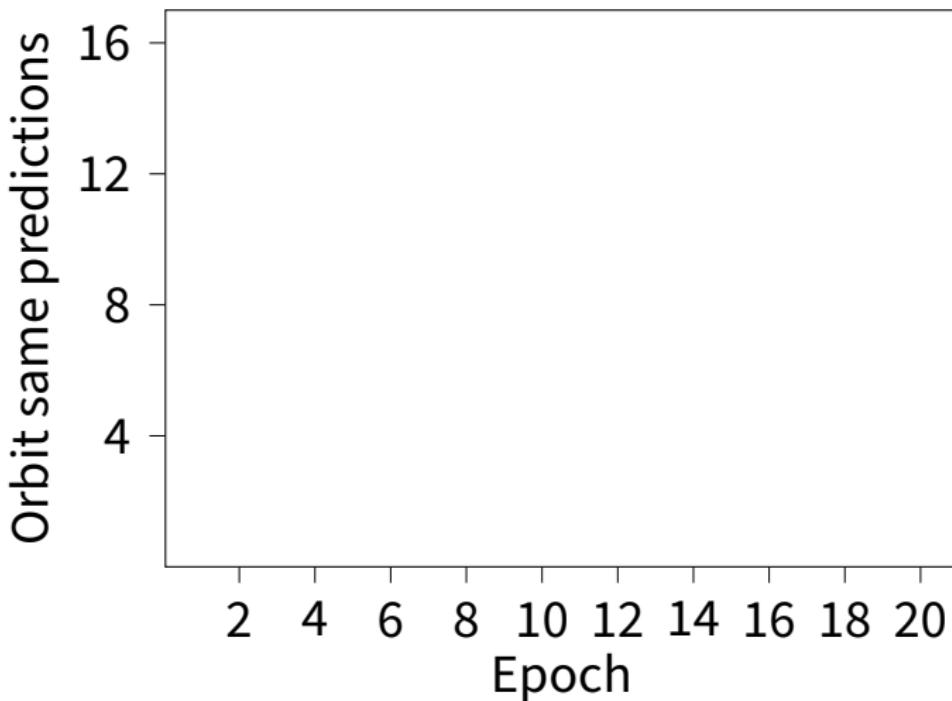
Histological slices

Orbit Same Predictions = 3

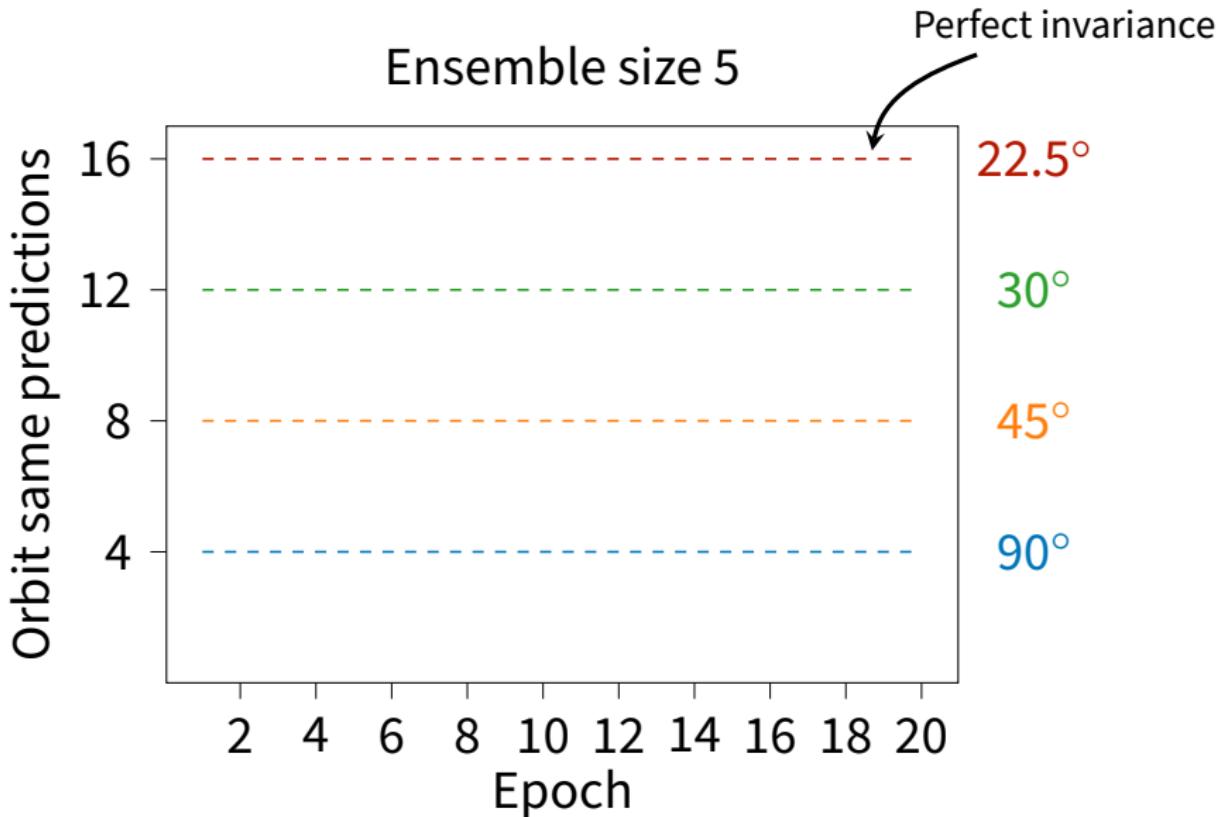


Out of distribution results

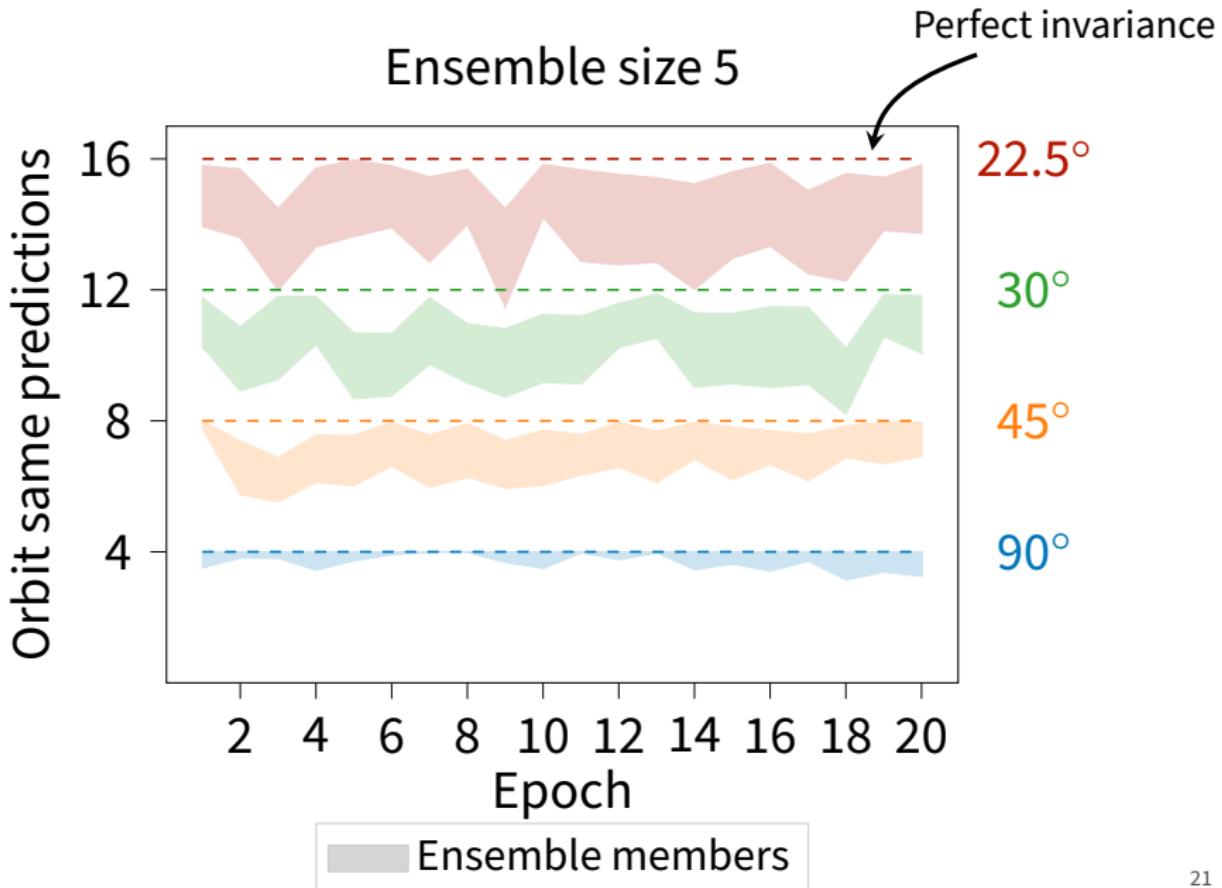
Ensemble size 5



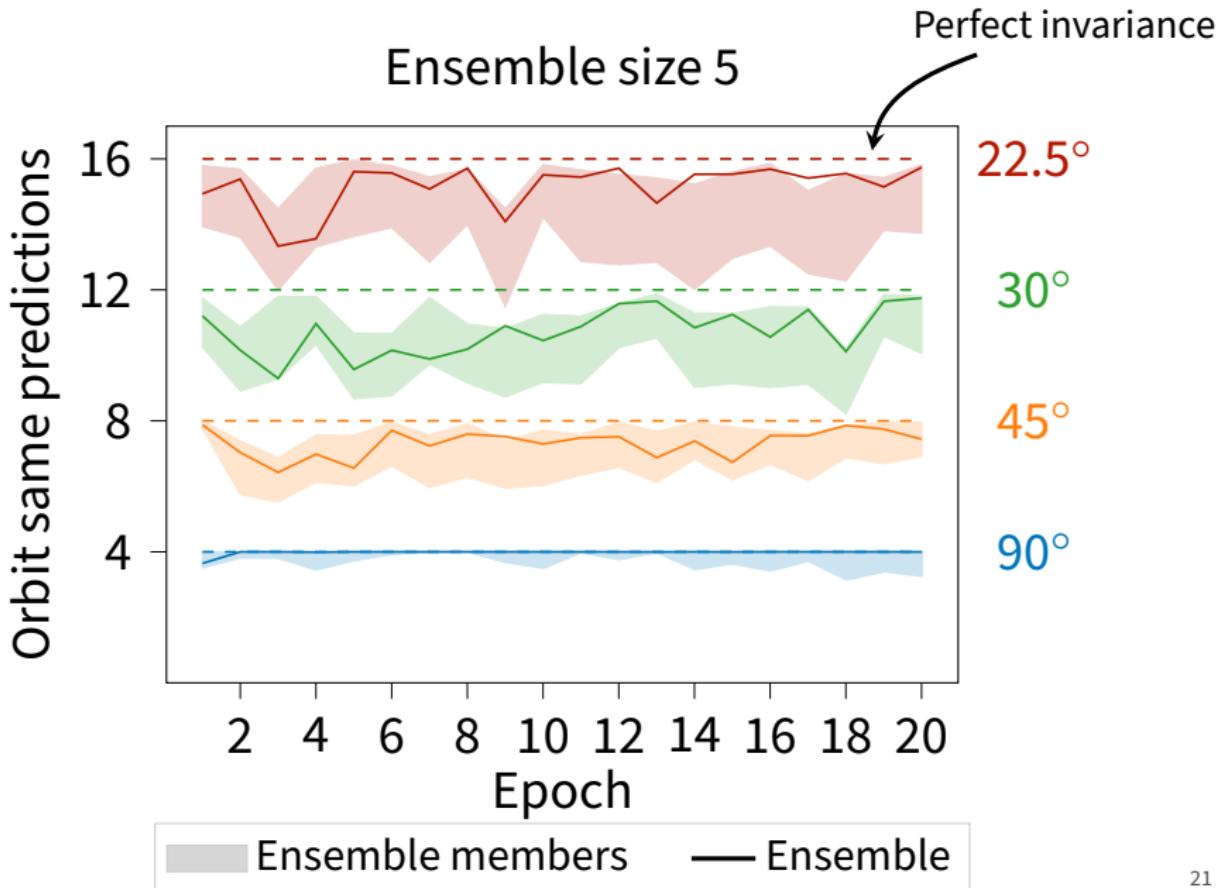
Out of distribution results



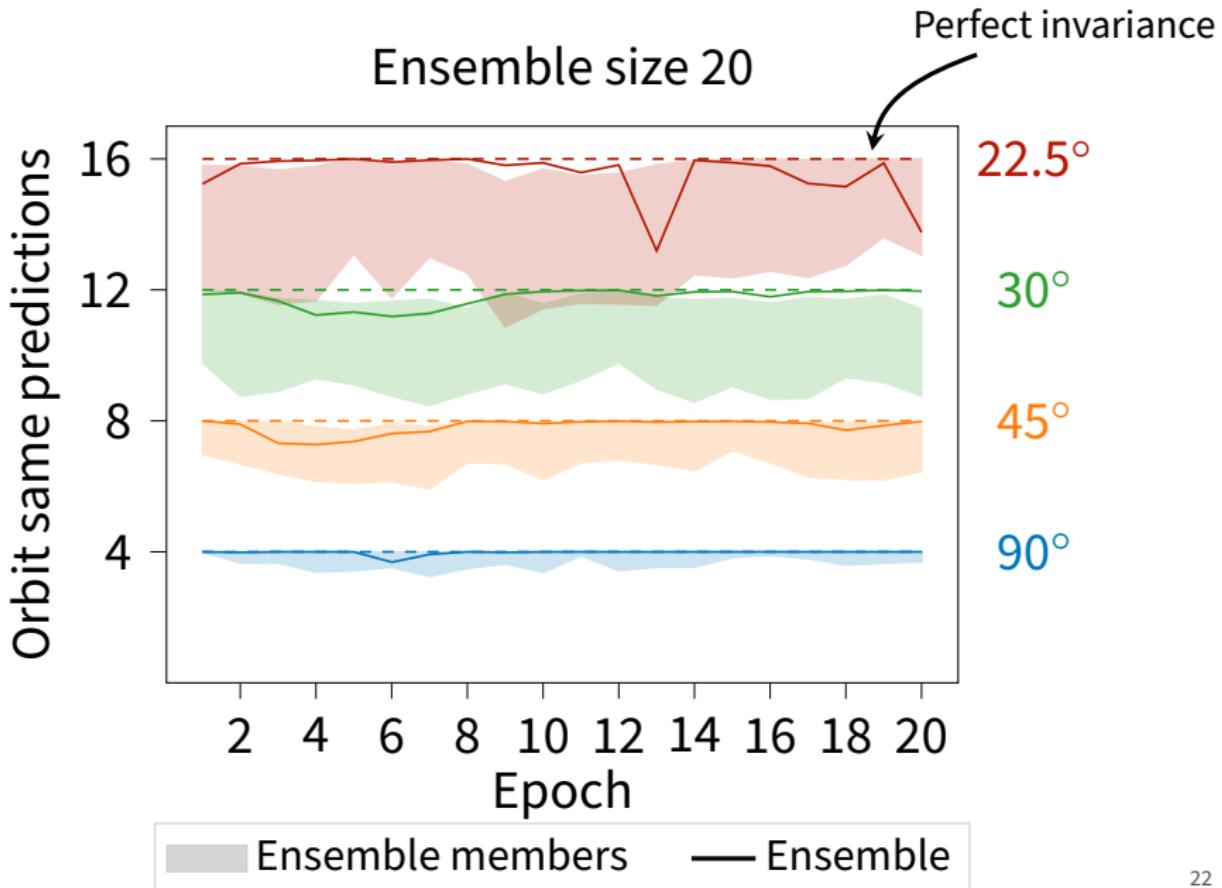
Out of distribution results



Out of distribution results



Out of distribution results



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.

Poster

Thursday, 25 July 2024

11.30am – 1.00pm

Hall C 4-9

Poster 817

