Assessing Generalization via Disagreement

Yiding Jiang*, Vaishnavh Nagarajan*, Christina Baek,
J. Zico Kolter

Carnegie Mellon University

* equal contribution

Estimating generalization

- Labeled test data is expensive

- Bounds based on Occam's Razor often vacuous

$$\text{train err}(f) - \text{test err}(f) \leq \mathcal{O}\left(\sqrt{\frac{\mu(f)}{\#\text{ of data}}}\right)$$

Complexity measure / Size of hypothesis space

[1] A theory of the learnable. Valiant et al., 84.

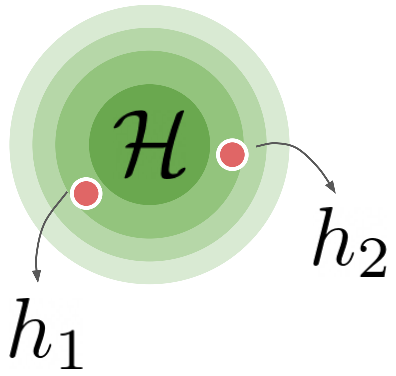
Overview

- We identify an empirical quantity that accurately estimates generalization error using **unlabeled test data**

- We prove that this is effective due to the fact that the **ensemble** of neural networks is **well-calibrated**

Disagreement & Test Error

Run SGD with **different random seeds** on the **same dataset** to get different hypotheses



Test Input $X = (x_1, x_2, x_3, x_4, x_5 \dots)$

Test Label $Y = (0, 1, 1, 2, 0, \dots)$

Predictions 1 $h_1 \circ X = (1, 1, 1, 2, 0, \dots)$

Predictions 2 $h_2 \circ X = (0, 1, 1, 1, 0, \dots)$

Test Error

Difference between predictions & ground truth

$$h_1 \circ X = (1, 1, 1, 2, 0, \dots)$$
$$Y = (0, 1, 1, 2, 0, \dots)$$

Disagreement

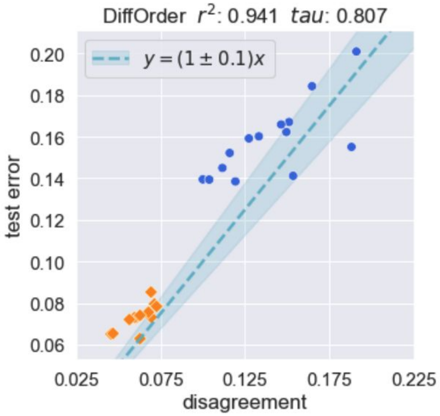
Difference between predictions of the two hypotheses; **does not need labels**

$$h_1 \circ X = (1, 1, 1, 2, 0, \dots)$$
$$h_2 \circ X = (0, 1, 1, 1, 0, \dots)$$

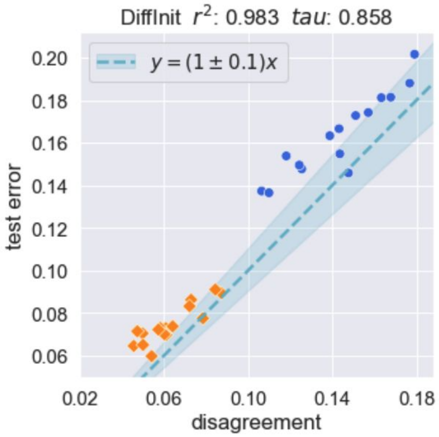
An intriguing observation

(that builds on Nakkiran and Bansal ‘20)

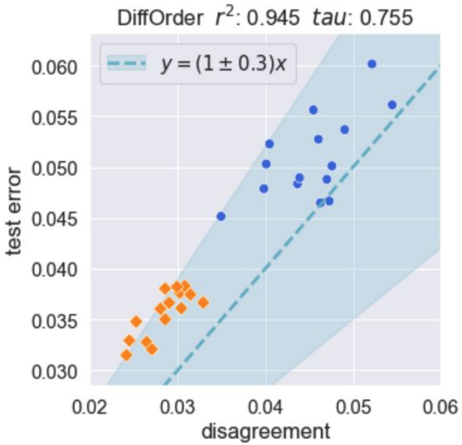
For a network trained twice with **same data but different random seeds**, disagreement (x-axis) tracks test error (y-axis) very well.



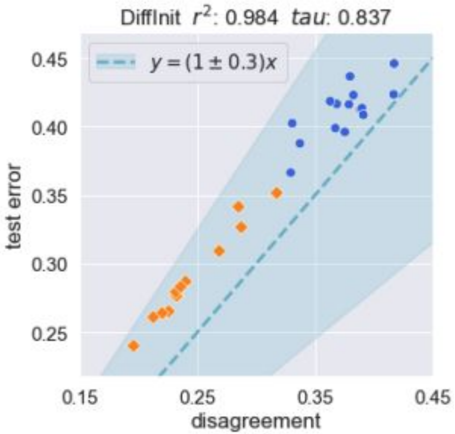
CIFAR10 + ResNet18
(different data order between the pair)



CIFAR10 + ResNet18
(different initializations between the pair)

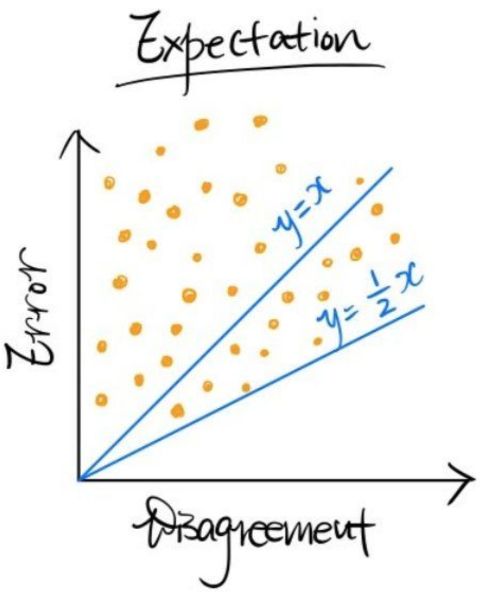


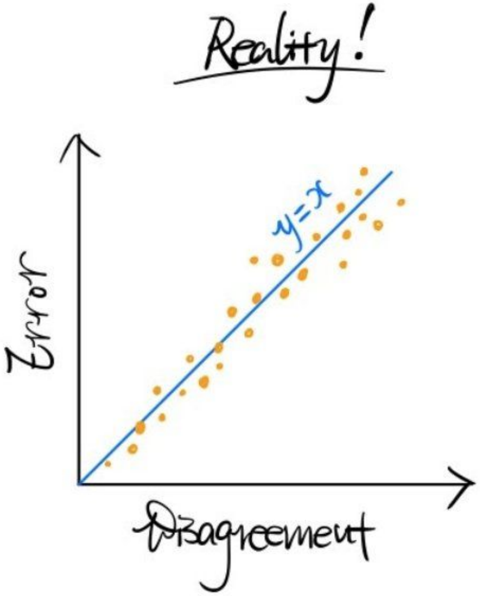
SVHN + ResNet18
(different data order between the pair)



CIFAR100 + ResNet18
(different data order between the pair)

Why is this surprising?

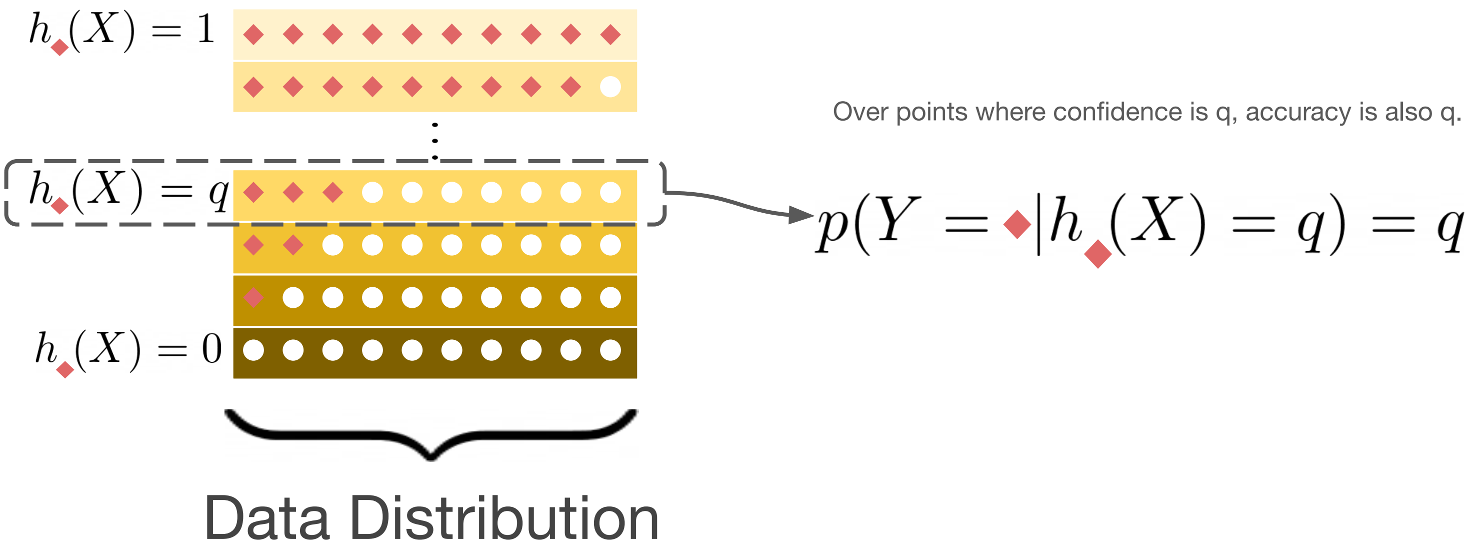




Why does disagreement = test error?

Key concept: Calibration

- For a calibrated classifier, the classifier's confidence matches its accuracy:



(Deep) Ensembles are well-calibrated

- Consider the ensemble of networks trained with different random seeds

$$\tilde{h}(X) = \mathbb{E}_{h \sim \mathcal{H}} \left[h(X) \right]$$

- While each member is not well-calibrated, the ensemble is known to be well-calibrated [3]!

[3] Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. Lakshminarayanan et al., 17.

Our result:

Calibration implies Generalization Disagreement Equality (GDE)

Theorem

If the ensemble satisfies class-wise calibration, then

$$\mathbb{E}_{h\sim\mathcal{H}}\left[\mathtt{TestErr}(h)\right] = \mathbb{E}_{h',h\sim\mathcal{H}}\left[\mathtt{Dis}(h,h')\right]$$

- This proves the 2 model observation **in expectation** over the stochasticity of SGD

Carnegie Mellon University

Open questions:

- Why does GDE hold over a single pair of models and not just in expectation?
- Why are deep ensembles well-calibrated?
- How do we leverage these insights out-of-distribution?

Carnegie Mellon University

Thanks for listening!