Assessing Generalization of SGD via Disagreement

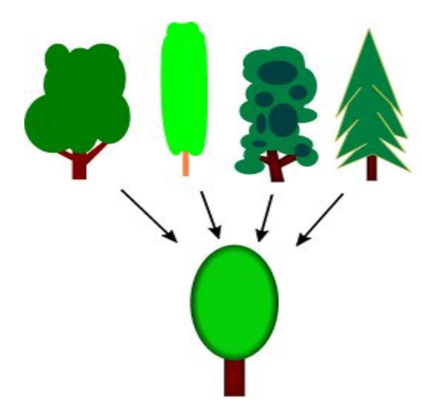
Yiding Jiang*, Vaishnavh Nagarajan*, Christina Baek, J. Zico Kolter

Carnegie Mellon University

Overview

We find a surprisingly simple technique to estimate test error of a deep network. We shed theoretical insight into why this works remarkably well.

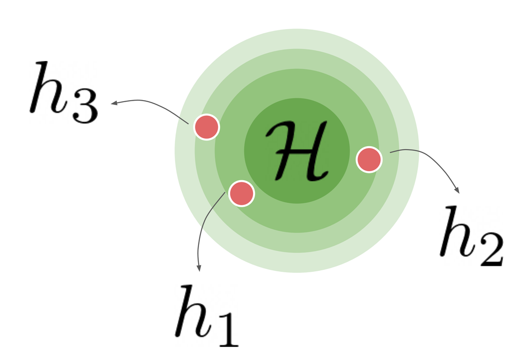
Predicting test performance remains a fundamental & challenging problem in deep learning.



- We demonstrate that **test error** can be accurately predicted by running two random seeds of SGD on the *same data* and measuring their **disagreement** on **unlabeled test** data.
- We prove that disagreement equals generalization error because deep SGD ensembles are (naturally) well-calibrated.
- Overall, we show a simple, yet new connection between generalization and calibration

Background

Let h_1 and h_2 be two hypotheses sampled from from the **distribution** of random SGD runs.



$$Y = \{\boxed{0}, 1, 1, 0, \dots\}$$
$$h_1(X) = \{\boxed{1}, 1, 1, 0, \dots\}$$

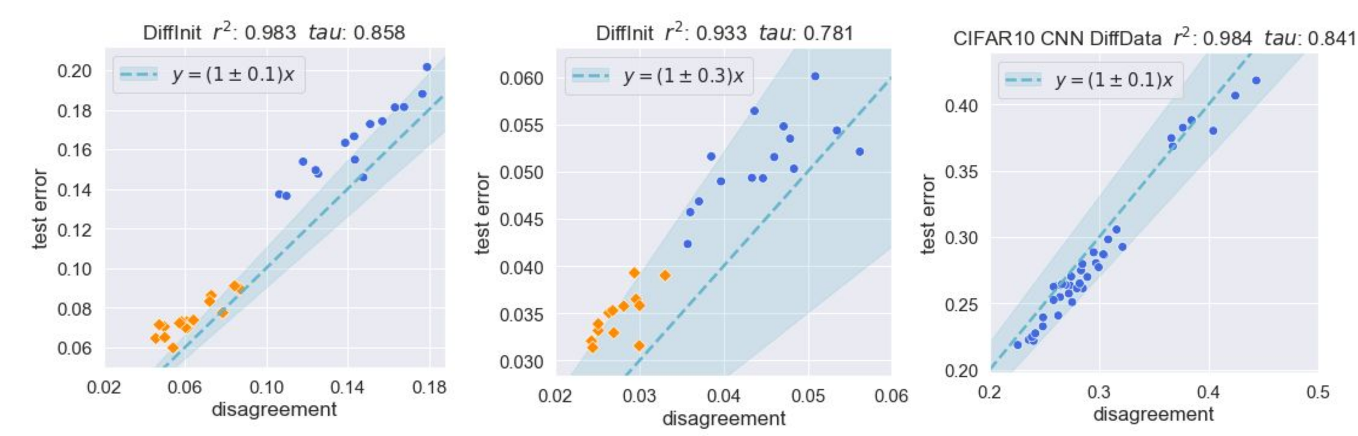
Test error measures the difference between prediction and the ground truth.

$$h_1(X) = \{1, \boxed{1}, 1, 0, \dots\}$$
$$h_2(X) = \{1, \boxed{0}, 1, 0, \dots\}$$

Disagreement measures the difference between predictions of two models (*no ground truth reqd.*)

An Intriguing Observation

Disagreement (x-axis) tracks test error (y-axis) extremely well across many architectures & datasets!

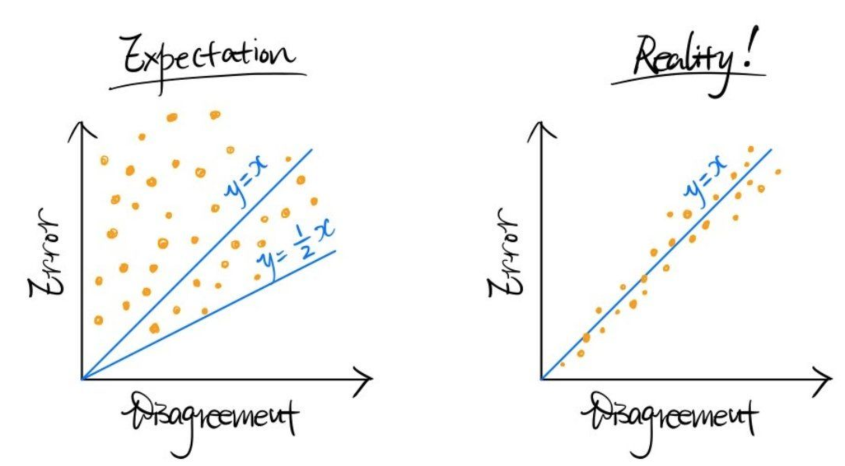


ResNet18 on CIFAR10 | ResNet18 on SVHN | CNN on CIFAR10

[2] showed this when h_2 was learned on an *independent* dataset. But we show it is enough to just retrain w/ different random seed (i.e., reorder/reinitialize).

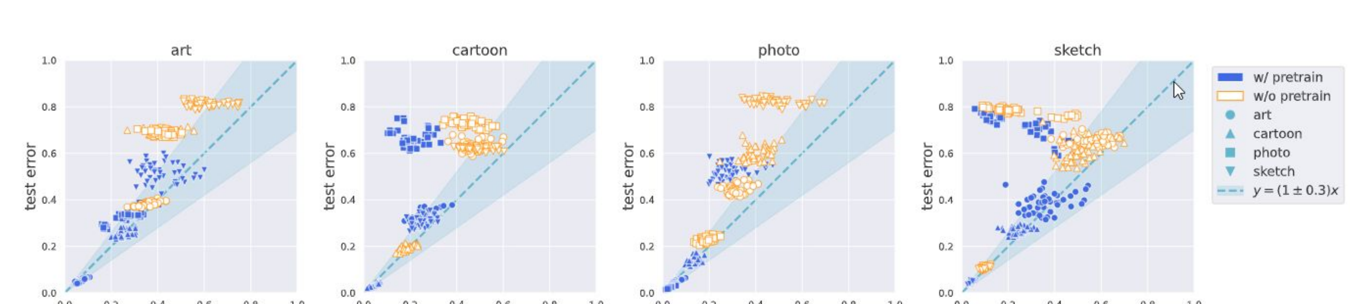
Why is this surprising?

The points could lie anywhere between x = 0 and y=0.5x but they are concentrated around y=x.



May work in some Out-of-Distribution scenarios!

Our technique works well for pre-trained models under 9 of 12 different domain shifts on the PACS dataset [1].



Generalization Disagreement Equality

Theorem

If the *ensemble* of models found by SGD is *well-calibrated*, then:

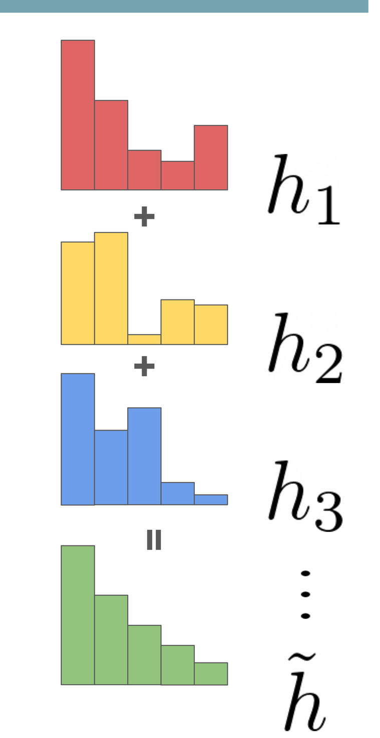
$$\underbrace{\mathbb{E}_{h\sim\mathcal{H}}\left[\texttt{TestErr}(h)\right]}_{\text{Expected Test Error over models sampled from SGD}} = \underbrace{\mathbb{E}_{h',h\sim\mathcal{H}}\left[\texttt{Dis}(h,h')\right]}_{\text{Expected Disagreement over pairs of models sampled from SGD}}$$

- Proves the observation **in expectation** rather than over a single draw of two models.
- Applies to any data distribution, model & algorithm!

Calibration & Ensembles

Ensemble predicts average of one-hot predictions across different SGD runs:

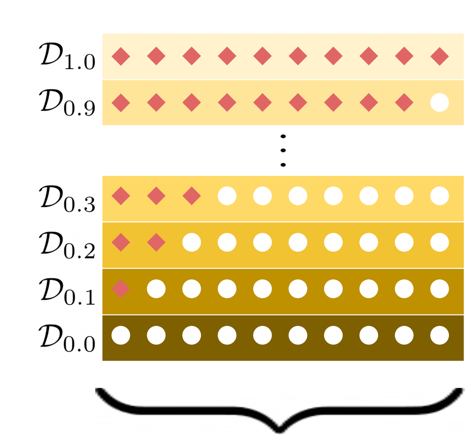
$$\tilde{h}(X) = \mathbb{E}_{h\sim\mathcal{H}}\left[h(X)\right]$$



What is a **well-calibrated model?**

Partition the distribution based on model's **confidence level.**

$$\mathcal{D}_q := (X, Y) \mid \tilde{h}(X) = q$$



Data Distribution

A well-calibrated model has accuracy q on D_q.

$$P(Y = k \mid \tilde{h}_k(X) = q) = q$$

I.e., it is neither over- nor under-confident.

Key proof idea for theorem: *On D_q*

Disagreement = Test error = 2q(1-q).

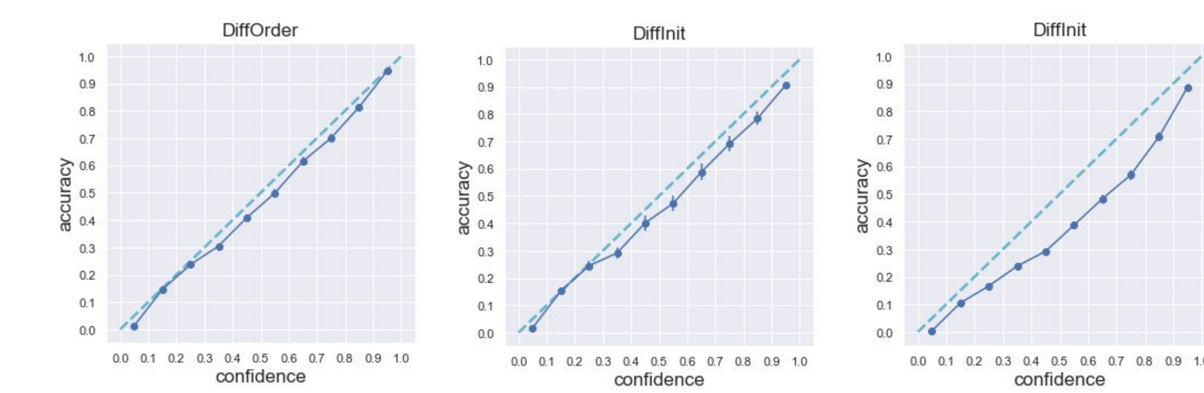
Empirical Verification

***Soft-max* ensembles are known to naturally have well-calibrated *top-class* predictions [3].**

We demonstrate that even ***one-hot* ensembles** are well-calibrated *on* ***average across all predictions***.

x-axis: the true probability of the data

y-axis: the confidence of the model



ResNet18 on CIFAR10 | ResNet18 on CIFAR10 2k | ResNet18 on CIFAR100

Future works

- In practice, GDE surprisingly holds even for a *single* (h1, h2) pair even though 2-ensembles are *not* calibrated! Why?
- Why are deep SGD ensembles well-calibrated? More generally, under what conditions?
- How else can unlabeled data be leveraged to estimate generalization in & out of distribution?

Reference

[1] Deeper, broader and artier domain generalization. Li et al.
[2] Distribution Generalization: A New Kind of Generalization. Nakkiran & Bansal.
[3] Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. Lakshminarayanan et al.