Agreement-on-the-Line: Predicting the Performance of Neural Networks under Distribution Shift

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recently, Miller et al. [38] showed that a model's in-distribution (ID) accuracy has a strong linear correlation with its out-of-distribution (OOD) accuracy, on several OOD benchmarks, a phenomenon they dubbed "accuracy-on-the-line". While a useful tool for model selection (i.e., the model most likely to perform the best OOD is the one with highest ID accuracy), this fact does not help to estimate the *actual* OOD performance of models without access to a labeled OOD validation set. In this paper, we show a similar surprising phenomena also holds for the *agreement* between pairs of neural network classifiers: whenever accuracy-on-the-line holds, we observe that the OOD agreement between the predictions of *any* two pairs of neural networks (with potentially different architectures) also observes a strong linear correlation with their ID agreement. Furthermore, we observe that the slope and bias of OOD vs ID agreement closely matches that of OOD vs ID accuracy. This phenomenon which we call *agreement-on-the-line*, has important practical applications: without any labeled data, we can *predict the OOD accuracy of classifiers*, since OOD agreement can be estimated with just unlabeled data. Our prediction algorithm outperforms previous methods both in shifts where agreement-on-the-line holds and, surprisingly, when accuracy is not on the line. This phenomenon also provides new insights into neural networks: unlike accuracy-on-the-line, agreement-on-the-line only appears to hold for neural network classifiers.

1 Introduction

Machine learning operates well when models observe and make decisions on inputs coming from the same distribution as the training data. Yet in the real world, this assumption rarely holds. Environments are never fully controlled. Robots interact with its surroundings, effectively changing what it sees in the future. Self-driving cars face constant distribution shift when driving to new cities under changing weather conditions. Models trained on clinical data from one hospital face challenges when deployed for a different hospital with different subpopulations. Under these premises, practitioners face the problem of estimating a model's performance on new data distributions (*out-of-distribution*, OOD) that are related to but different from the data distribution that the model was trained on (*in-distribution*, ID). Assessing OOD performance is difficult because in reality, labeled OOD data is usually close to nonexistent. On the other hand, *unlabeled* OOD data is much easier to obtain. A natural question is whether we can leverage unlabeled OOD data for estimating the OOD performance. This paradigm of using unlabeled data to predict the OOD generalization performance has received much attention recently [22, 7, 49, 15, 16, 8, 23]. Though a flurry of different metrics have been proposed, their success varies widely depending on the shift and the ID performance of the model. While it is in fact impossible for a method to always work with no assumptions [22], a major hurdle in using these methods is that there is currently no understanding of when they work or a recipe to detect when their predictions might be unreliable.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

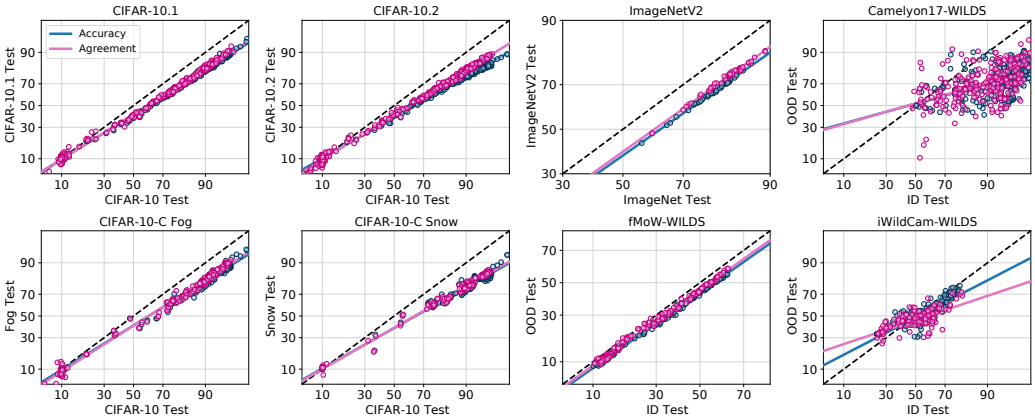

Figure 1: We see that when ID and OOD accuracy is linearly correlated, the ID and OOD agreement is also linearly correlated. Additionally, when ID and OOD accuracy is not linearly correlated, agreement is also not linearly correlated. Each blue point in the scatter plot represents the accuracy of a single model. Each pink point represents the agreement between a pair of models. To avoid cluttering the figure, given n models of interest, we randomly pair the models and plot the agreement of $n/2$ pairs of models. The axes are probit scaled as described in the experimental setup.

Recently, Miller et al. [38] demonstrated that in a wide variety of common OOD prediction benchmarks such as CIFAR-10.1 [44], ImageNetV2 [44], CIFAR-10C [26], FMoW-WILDS [9], there exists an almost perfect positive linear correlation between the ID test vs OOD accuracy of models. When the phenomena, called *accuracy-on-the-line*, occurs, improving performance on the ID test data directly leads to improvements in OOD performance. However, not all distribution shifts observed accuracy-on-the-line. In some datasets, such as Camelyon17-WILDS [1], models with the same ID test performance had OOD performance that varied largely. Miller et al. [38] conducted some preliminary analysis to understand when accuracy-on-the-line occurs, however characterizing this phenomena largely remains an open question. Thus, while the accuracy-on-the-line phenomena is extremely interesting, its practical use is somewhat limited. We cannot verify whether accuracy-on-the-line happens, nor can we estimate the OOD performance for classifiers using the slope of this linear correlation without access to labeled OOD data, which is precisely what we often do not have.

In this work, we begin by observing an analogous phenomenon based upon *agreement* rather than accuracy. Specifically, if we consider pairs of neural network of classifiers, and look at the agreement of their predictions (the proportion of cases where they make the same prediction, which requires no labeled data to compute), we find that there *also* often exists a linear correlation between ID vs OOD agreement. We call this phenomenon "agreement-on-the-line". Importantly, however, this phenomenon appears to be tightly coupled with accuracy-on-the-line: when agreement-on-the-line holds, accuracy-on-the-line also holds; and when agreement-on-the-line *does not* hold, neither does accuracy-on-the-line. Furthermore, in the case of neural network classifiers, when these properties hold, the linear correlations of both accuracy-on-the-line and agreement-on-the-line appear to have roughly *the same slope and bias*. Interestingly, unlike accuracy-on-the-line, which appears to be a general phenomenon, agreement-on-the-line, especially the fact that the slope and bias of the linear correlation agree across accuracy and agreement, appears to occur only for neural networks. Indeed the phenomenon is quite unintuitive, given that there is no a-piori reason to believe that agreement and accuracy would be connected in such a manner; nonetheless, we find this phenomenon occurs repeatedly across multiple datasets and vastly different deep network architectures.

This phenomenon is of immediate practical interest. Since agreement-on-the-line can be validated without *any* labeled OOD data, we can immediately use it as a proxy to assess whether accuracy-on-the-line holds, and thus whether it is reasonable to use ID accuracy as model selection criteria for picking a model on OOD data. Furthermore, since the slope and bias of the agreement-on-the-line fit can *also* be estimated without labeled OOD data, (for the cases where agreement-on-the-line holds) we can use this approach to derive a very simple algorithm for estimating the OOD generalization of classifiers, without *any* access to labeled OOD data. The approach outperforms competing methods, and predicts OOD test error with a mean absolute estimation error of $\leq 2\%$ on datasets where agreement-on-the-line holds. On datasets where agreement-on-the-line does not hold, the method, as

expected does not perform as well, but somewhat surprisingly *still* outperforms competing methods in terms of predicting OOD performance.

To summarize, our contributions are as follows:

1. We discover and empirically analyze the agreement-on-the-line phenomenon: that ID and OOD *agreement* for pairs of classifier lies on a line precisely when the corresponding ID and OOD *accuracy* also lies on a line. Furthermore, for the case of neural network classifiers, the slope and bias of these two lines are approximately equal.
2. Exploiting this phenomenon, we develop a simple method for estimating the OOD performance of classifiers without *any* access to labeled OOD data (and by observing whether agreement-on-the-line holds, the method also provides a "sanity check" that these estimates are reasonable). The method outperforms all competing baselines that we've tried for this task.

2 Related Works

Agreement in ID generalization. Departing from approaches based on uniform convergence [42, 19, 2, 39], several recent works [30, 41, 50, 21] derive unconventional approaches for estimating generalization error or comparing different models. In particular, this work is closely related to Jiang et al. [30], which shows that the disagreement between two models trained with different random seeds closely tracks the generalization error of the models if the ensembles of the models are well-calibrated. Predicting ID generalization via disagreement has also previously been proposed by Madani et al. [36] and Nakkiran and Bansal [40]. Our method also uses disagreement but focuses on out-of-distribution generalization.

OOD generalization. Compared to the ID setting, the problem of characterizing generalization in OOD setting is relatively understudied. Ben-David et al. [4] provides one of the first uniform-convergence-based bounds for *domain adaptation*. Several works [37, 13, 33] build on this approach and extend it to other learning scenarios. Most of these works attempt to bound the difference between ID performance and OOD performance via a certain notion of closeness (e.g., the total variation distance and the $\mathcal{H}\Delta\mathcal{H}$ divergence which is related to agreement) between the original distribution and shifted distribution and build on the uniform-convergence framework [45]. As pointed out by Miller et al. [38], these approaches provide an upper bound on the OOD performance that grows looser as the distribution shift becomes larger, and these upper bounds do not capture the precise trends observed in practice. Predicting the actual OOD performance to the dot using unlabeled data has gained interest in the past decade. These methods can roughly be divided into three categories:

1. Placing assumptions on the distribution shift. Donmez et al. [17] assume knowledge of the marginal of the shifted distribution $P(y)$ and show that OOD accuracy can be predicted if the shifted distribution satisfies several properties. Steinhardt and Liang [47] works under the assumption that the data x can be separated into "views" that are conditionally independent given label y. Chen et al. [8] assumes prior knowledge about the shift and uses an importance weighting procedure.

2. Utilizing multiple classifiers. Given multiple classifiers of interest, Platanios et al. [43, 43] uses logical constraints to identify each one's error. For example, they assume that if two classifiers agree on some data point, their prediction is more likely to be correct, and that if two classifiers disagree, at least one must be incorrect. On the other hand, Jaffe et al. [29] uses a spectral-based approach under the assumption that classifiers make independent errors.

3. Measuring the distribution shift. Elsahar and Gallé [20] trains a regression model over several metrics that measure the severity of the distribution shift ($\mathcal{H}$-divergence, confidence, reverse classification accuracy). However, their framework assumes access to other labeled OOD datasets. Similarly, Schelter et al. [46], Deng and Zheng [15] assume knowledge of typical augmentations or corruptions and similarly uses regression. In particular, Deng et al. [16] observes that simply looking at different rotations is often sufficient. Inspired by the observation that the maximum softmax probability (or confidence) for OOD points is typically lower [27, 25], Guillory et al. [23] and more recently Garg et al. [22] utilize model confidence to predict accuracy. Chuang et al. [10] provides an upper bound of the OOD error which they approximate using agreement between the model of

Dataset	Accuracy			Agreement		
	Slope	Bias	R^2	Slope	Bias	R^2
CIFAR-10.1v6	0.842	-0.216	0.999	0.857	-0.205	0.997
CIFAR-10.2	0.768	-0.287	0.999	0.839	-0.226	0.996
ImageNetv2	0.946	-0.309	0.997	0.972	-0.274	0.993
CIFAR-10C-Fog	0.834	-0.228	0.995	0.870	-0.239	0.996
CIFAR-10C-Snow	0.762	-0.289	0.974	0.766	-0.266	0.974
FMOW	0.952	-0.163	0.998	0.954	-0.121	0.995
Camelyon17	0.373	0.046	0.263	0.381	0.075	0.226
iWildCam	0.700	-0.037	0.738	0.411	-0.094	0.424

Table 1: Slope, bias, and coefficients of determination (R^2) values of linear correlations between ID vs OOD accuracy and ID vs OOD agreement. The slope/bias of these linear correlations match when the R^2 value is high (i.e. strong linear correlation).

interest and a set of domain-invariant classifiers. This method was extended upon by Chen et al. [7]. Yu et al. [49] observed that the distance between the model of interest f and a reference model trained on the pseudolabels of f showed strong linear correlation with the OOD accuracy.

Though a large number of methods have been proposed, for the large majority, it is not well-understood when they will work. Intuitively, no method will work on all shifts without additional assumptions [22]. But is there some *simple general structure* to shifts in the real world that allows us to reliably predict OOD accuracy? Even if the structure is not universal, can we easily *inspect* if this structure holds? What is a plausible assumption we can make about the OOD *behaviour of classifiers*? The novelty and significance of our work comes from trying to better understand and address these questions, specifically for neural networks. In this work, we observe a phenomenon related to, but stronger than accuracy-on-the-line that allows us to reliably predict the OOD accuracy of neural networks.

3 The agreement-on-the-line phenomenon

3.1 Notation and setup

Let $\mathcal{H}$ denote a set of neural networks trained on $(X_{\mathsf{train}}, \boldsymbol{y}_{\mathsf{train}}) = \{(x_i, y_i)\}_{i=1}^{m_{\mathsf{train}}}$ sampled from $\mathcal{D}_{\mathsf{ID}}$.

Given any pair of models $h, h' \in \mathcal{H}$, for a distribution $\mathcal{D}$, the expected accuracy and agreement are defined as:

$$\mathsf{Acc}(h) = \mathbb{E}_{x,y\sim\mathcal{D}}[\mathbb{1}\{h(x) = y\}], \quad \mathsf{Agr}(h, h') = \mathbb{E}_{x\sim\mathcal{D}}[\mathbb{1}\{h(x) = h'(x)\}]. \tag{1}$$

We assume access to labeled validation set $(X_{\mathsf{val}}, \boldsymbol{y}_{\mathsf{val}}) = \{(x_i, y_i)\}_{i=1}^{m_{\mathsf{val}}}$ sampled from $\mathcal{D}_{\mathsf{ID}}$ that allows us to estimate the ID accuracy $\widehat{\mathsf{Acc}}_{\mathsf{ID}}(h)$ as the sample average of $\mathbb{1}\{h(x) = y\}$ over the validation set. We do not assume access to a labeled OOD validation set, as this is often impractical to obtain, and thereby cannot directly estimate $\widehat{\mathsf{Acc}}_{\mathsf{OOD}}(h)$ in a similar manner.

Agreement, on the other hand, only requires access to unlabeled data. We assume access to *unlabeled* samples $X_{\mathsf{OOD}} = \{x_i\}_{i=1}^{m_{\mathsf{OOD}}}$ from the shifted distribution of interest as $\mathcal{D}_{\mathsf{OOD}}$. Hence, we can estimate both the ID and OOD agreement as follows:

$$\widehat{\mathsf{Agr}}_{\mathsf{ID}}(h, h') = \frac{1}{m_{\mathsf{val}}} \sum_{x\in X_{\mathsf{val}}} \mathbb{1}\{h(x) = h'(x)\}, \quad \widehat{\mathsf{Agr}}_{\mathsf{OOD}}(h, h') = \frac{1}{m_{\mathsf{OOD}}} \sum_{x\in X_{\mathsf{OOD}}} \mathbb{1}\{h(x) = h'(x)\} \tag{2}$$

3.2 Experimental setup

We study the ID vs OOD accuracy and agreement between pairs of models across 20+ common OOD benchmarks and hundreds of neural networks.

Datasets. We present results on 8 dataset shifts in the main paper, and include results for other distribution shifts in the Appendix B. These 8 datasets span

1. Dataset reproductions: CIFAR-10.1 [44], CIFAR-10.2 [35] reproductions of CIFAR-10 [32] and ImageNetV2 [44] reproduction of ImageNet [14]

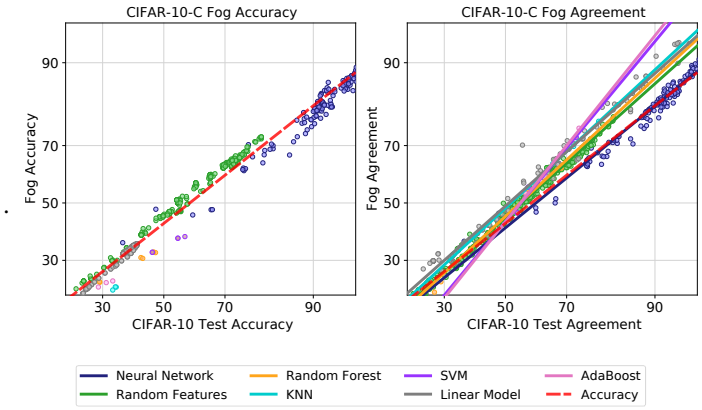

Figure 2: We observe whether the agreement-on-the-line phenomena happens across various model classes on the CIFAR-10 Fog dataset. As shown on the left, the ID vs OOD accuracy of all model classes lie on the same line. We plot ID vs OOD agreement between pairs of models from the same model class and observe that only the linear correlation between ID vs OOD agreement of neural networks match that of ID vs OOD accuracy (in red).

2. Synthetic corruptions: CIFAR-10C Fog and CIFAR10C Snow [26]
3. Real-world shifts from [31]: satellite images (FMoW-WILDS), images from camera traps in the wildlife (iWildCam-WILDS [3]), and images of cancer tissue (Camelyon17-WILDS [1])

Models. For ImageNetV2, we evaluate 50 ImageNet pretrained models from the pretrainedmodels [6] package. On all other shifts, we evaluate on all models in the testbed created and utilized by [38] consisting of ≥ 150 models for each shift. The evaluated models span a variety of convolutional neural networks (e.g. ResNet [24], DenseNet [28], EfficientNet [48], VGG [34]) as well as various Vision Transformers [18]. All architectures and models are listed in the Appendix C.

Probit scaling. Miller et al. [38] report their results after probit scaling [1] the ID vs OOD accuracies due to a better linear fit. We apply the same probit transform to both accuracy and agreement in our experiments.

3.3 Observations

We empirically observe a peculiar phenomena in deep neural networks, that we call *agreement-on-the-line* characterized by the following three properties:

Prop(i) When ID vs OOD accuracy observes a strong linear correlation (≥ 0.95 R^2 values), we see that ID vs OOD *agreement is also strongly linearly correlated*.

Prop(ii) When both accuracy and agreement observe strong linear correlation, we see that these linear correlations have almost the *same slope and bias*.

Prop(iii) When the linear correlation of ID vs OOD accuracy is weak (≤ 0.75 R^2 values), the linear correlation between ID and OOD agreement is similarly weak.

We show the agreement-on-the-line phenomenon on 8 datasets in Figure 1 and Table 3 (See Appendix B for other datasets). On CIFAR10.1, CIFAR10.2, ImageNetV2, CIFAR-10C Fog/Snow, and fMoW-WILDS, we observe that both ID vs OOD accuracy and agreement observe strong linear correlations, and the linear fit has the same slope and bias (Prop(i), Prop(ii)). On the other hand, on datasets Camelyon17-WILDS and iWildCam-WILDS where accuracy is not linearly correlated, agreement also not linearly correlated (Prop(iii)).

3.4 What makes agreement-on-the-line interesting?

First, *agreement can be estimated with just unlabeled data*. Hence, the agreement-on-the-line phenomenon has important practical implications for both *checking* whether the distribution shift observes accuracy-on-the-line and *predicting* the actual value of OOD accuracy without any OOD labels. We present a method to estimate OOD error using this phenomenon in Section 4.

Second, agreement-on-the-line does *not* directly follow from accuracy-on-the-line. It is not the case that the average ID/OOD accuracy between pairs of models is equal to their ID/OOD agreement.

[1] the probit transform is the inverse of the cumulative density function of the standard Gaussian distribution.

Prior work has observed that expected ID accuracy often equals ID agreement over pairs of models with the same architecture, trained on the same dataset but with different random seeds [30]. However, agreement-on-the-line goes beyond these results in two ways: (i) agreement between models with *different architectures* (Fig. 2) and (ii) agreement between different checkpoints on the *same training run* (Fig. 4) is also on the ID vs OOD agreement line. These ID/OOD agreements do not equal the expected ID/OOD accuracy. Indeed, understanding why agreement-on-the-line holds requires going beyond the theoretical conditions presented in the prior work [30] which do not hold for this expanded set of models (See Appendix E for further discussion).

Finally, we note that there is something special about neural networks that makes the ID vs OOD agreement trend identical to the ID vs OOD accuracy trend. This is unlike accuracy-on-the-line that holds across a wide range of models including neural networks and classical approaches. Figure 2 shows CIFAR-10 Test vs CIFAR-10C Fog accuracy and agreement of linear models (e.g. logistic and ridge regression) and various non-linear models (e.g. Kernel SVM [12], k-Nearest Neighbors, Random Forests [5], Random Features [11], AdaBoost [51]) (See plots for other datasets in Appendix B). We look at agreement between pairs of models from the same model family. While Prop(i) seems to hold for several other model families on several shifts, Prop(ii) only holds for neural networks, i.e. the slope and bias of the agreement line do not match the slope and bias of the accuracy line.

4 A method for estimating OOD accuracy

In this section, we describe how the phenomenon of agreement-on-the-line (described in Section 3) offers a simple practical method to perform model selection and estimate accuracy under distribution shifts. Recall from Section 3.1 that we have labeled ID validation data $(X_{\mathsf{val}}, y_{\mathsf{val}})$ and unlabeled OOD data X_{OOD}.

Model selection. Without OOD labeled data, can we determine which model is likely to achieve the best OOD performance? When accuracy-on-the-line holds and ID vs OOD accuracy is linearly correlated, we can simply pick the model with highest ID accuracy. However, this only works *if* accuracy-on-the-line holds. In practice, how does one determine if accuracy-on-the-line holds without labeled OOD data? Agreement-on-the-line provides an answer here. By Prop(i) and Prop(iii), agreement-on-the-line implies accuracy-on-the-line. Hence, we simply need to check if ID and OOD agreement (which can be estimated as in (2)) are linearly correlated, in order to know if our model selection criterion based on ID accuracy is valid.

OOD error prediction. Agreement-on-the-line allows us to go beyond model selection and actually *predict OOD accuracy*. Intuitively, we can estimate the slope and bias of the agreement line with just unlabeled data. By Prop(ii), they match the slope and bias of the accuracy line and hence, we can estimate the OOD accuracy by linearly transforming the ID accuracy (with the appropriate probit scaling). We formalize this intuition below and provide an algorithm for OOD accuracy estimation in Algorithm 1.

Recall (Section 3.1) that given n distinct models of interest $\mathcal{H} = \{h_i\}_{i=1}^n$, we can estimate $\widehat{\mathsf{Acc}}_{\mathsf{ID}}(h)$, $\widehat{\mathsf{Agr}}_{\mathsf{ID}}(h, h')$ and $\widehat{\mathsf{Agr}}_{\mathsf{OOD}}(h, h')$ as sample averages over ID labeled validation data and OOD unlabeled data for all $h, h' \in \mathcal{H}$. We now describe an estimator $\widehat{\mathsf{Acc}}_{\mathsf{OOD}}(h)$ for the OOD accuracy of a model $h \in \mathcal{H}$.

From agreement-on-the-line, we know that when ID vs OOD agreement lies on a line for all $h, h' \in \mathcal{H}$, ID vs OOD accuracy for all $h \in \mathcal{H}$ would approximately also lie on the same line:

$$\Phi^{-1}(\mathsf{Acc}_{\mathsf{OOD}}(h)) = a \cdot \Phi^{-1}(\mathsf{Acc}_{\mathsf{ID}}(h)) + b \Leftrightarrow \Phi^{-1}(\mathsf{Agr}_{\mathsf{OOD}}(h, h')) = a \cdot \Phi^{-1}(\mathsf{Agr}_{\mathsf{ID}}(h, h')) + b \tag{3}$$

We estimate the slope and bias of the linear by performing linear regression after applying a probit transform on the disagreements as follows.

$$\hat{a}, \hat{b} = \arg\min_{a,b \in \mathbb{R}} \sum_{i,j \neq ix} (\Phi^{-1}(\widehat{\mathsf{Agr}}_{\mathsf{OOD}}(h_i, h_j)) - a \cdot \Phi^{-1}(\widehat{\mathsf{Agr}}_{\mathsf{ID}}(h_i, h_j)) - b)^2 \tag{4}$$

For each model $h \in \mathcal{H}$, given its ID validation accuracy, one could simply plug estimate $\hat{a}$ and bias $\hat{b}$ from (4), and $\widehat{\mathsf{Acc}}_{\mathsf{ID}}(h)$ (sample average over validation set) into (3) to get an estimate of the model's OOD accuracy. We call this *simple* algorithm ALine-S.

Notice that ALine-S does not directly use the OOD disagreement estimates concerning the model of interest—we only use disagreements indirectly via the estimates $\hat{a}$ and $\hat{b}$. We find that a better estimator can be obtained by *directly* using the model's OOD agreement estimates via simple algebra as follows.

First, note that for any pair of models $h, h' \in \mathcal{H}$, it directly follows from (3) that

$$\frac{\Phi^{-1}(\mathsf{Acc}_{\mathsf{OOD}}(h)) + \Phi^{-1}(\mathsf{Acc}_{\mathsf{OOD}}(h'))}{2} = a \cdot \frac{\Phi^{-1}(\mathsf{Acc}_{\mathsf{ID}}(h)) + \Phi^{-1}(\mathsf{Acc}_{\mathsf{ID}}(h'))}{2} + b \tag{5}$$

By subtracting $\Phi^{-1}(\mathsf{Agr}_{\mathsf{OOD}}(h, h')) = a \cdot \Phi^{-1}(\mathsf{Agr}_{\mathsf{ID}}(h, h')) + b$ from (5), we can get that average OOD accuracy of any pair of models $h, h' \in \mathcal{H}$ is

$$\begin{aligned} &\frac{1}{2}\underbrace{\Phi^{-1}(\mathsf{Acc}_{\mathsf{OOD}}(h))}_{\text{unknown}} + \frac{1}{2}\underbrace{\Phi^{-1}(\mathsf{Acc}_{\mathsf{OOD}}(h'))}_{\text{unknown}} \\ &= \underbrace{\Phi^{-1}(\mathsf{Agr}_{\mathsf{OOD}}(h, h')) + a \cdot \left(\frac{\Phi^{-1}(\mathsf{Acc}_{\mathsf{ID}}(h)) + \Phi^{-1}(\mathsf{Acc}_{\mathsf{ID}}(h'))}{2} - \Phi^{-1}(\mathsf{Agr}_{\mathsf{ID}}(h, h')) \right)}_{\text{known (can estimate via sample average over } X_{\mathsf{OOD}} \text{ and } (X_{\mathsf{val}}, y_{\mathsf{val}}))}. \end{aligned} \tag{6}$$

We can plug in estimates of the terms on the right hand side ($\hat{a}$ from linear regression (4)) and the rest from sample averages. In this way, we can construct a system of linear equations of the form (6) involving "unknown" estimates of the OOD accuracy of models and other "known" quantities which we solve via linear regression to obtain the unknown estimates. We call this procedure ALine-D, and it is described more explicitly in Algorithm 1. As one caveat, this method requires at least 3 models in the set of interest $\mathcal{H}$ for the system of linear equations to have a unique solution.

Algorithm 1 ALine-D: Predicting OOD Accuracy

Input: m_{ID} validation samples $(X_{\text{ID}-\text{val}}, \mathbf{y}_{\text{ID}-\text{val}})$, m_{OOD} unlabeled samples X_{OOD}, a set containing n models of interest $\mathcal{H}$
Get $\widehat{\mathsf{Acc}}_{\mathsf{ID}}(h_i)\ \forall i \in [n]$
Get $\widehat{\mathsf{Agr}}_{\mathsf{ID}}(h_i, h_j)$ and $\widehat{\mathsf{Agr}}_{\mathsf{OOD}}(h_i, h_j)$ for all pairs of models $i \neq j$
Get $\hat{a}, \hat{b} = \arg\min_{a,b \in \mathbb{R}} \sum_{i \neq j} (\Phi^{-1}(\widehat{\mathsf{Agr}}_{\mathsf{OOD}}(h_i, h_j)) - a \cdot \Phi^{-1}(\widehat{\mathsf{Agr}}_{\mathsf{ID}}(h_i, h_j)) - b)^2$
Initialize $A = \mathbb{R}^{\frac{n(n-1)}{2} \times n}$, $\boldsymbol{b} = \mathbb{R}^{\frac{n(n-1)}{2}}$
$i = 0$
for $h_j, h_k \in \mathcal{H}$ **do**
  $A_{ij} = \frac{1}{2}, A_{ik} = \frac{1}{2}, A_{i\ell} = 0\ \forall l \notin \{j, k\}$
  $\boldsymbol{b}_i = \Phi^{-1}(\widehat{\mathsf{Agr}}_{\mathsf{OOD}}(h_j, h_k)) + \hat{a} \cdot \left(\frac{\Phi^{-1}(\widehat{\mathsf{Acc}}_{\mathsf{ID}}(h_j) + \Phi^{-1}(\widehat{\mathsf{Acc}}_{\mathsf{ID}}(h_k)))}{2} - \Phi^{-1}(\widehat{\mathsf{Agr}}_{\mathsf{ID}}(h_j, h_k)) \right)$
  $i = i + 1$
end for
Get $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w} \in \mathbb{R}^n} \|A\boldsymbol{w} - \boldsymbol{b}\|_2^2$
return $\Phi(w_i^*)\ \forall i \in [n]$

5 Experiments

Datasets and models. We evaluate our methods, the simple plug in of slope/bias estimate ALine-S and the more involved ALine-D, on the same models and datasets where we observed agreement-on-the-line in Section 3. Specifically, we look at CIFAR10.1, CIFAR10.2, ImageNetV2, CIFAR-10C, and FMoW-WILDS where we observed a strong correlation. We also look at the performance on datasets where we do not see a strong linear correlation, specifically Camelyon-WILDS and iWildCam-WILDS.

Baseline methods. We choose 4 existing methods for comparison: Average Threshold Confidence (ATC) by Garg et al. [22], DOC-Feat in Guillory et al. [23], Average Confidence (AC) in [27], and naive Agreement [36, 40, 30]. We implement the version of ATC that performed best in the paper i.e. with negative entropy as the score function and do temperature scaling to calibrate the models in-distribution. Although DOC was deemed the best method in Guillory et al. [23], we use DOC-Feat

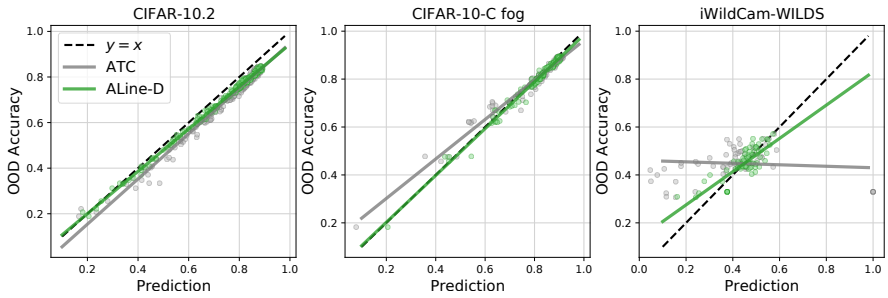

Figure 3: Prediction vs OOD accuracy. We observe the scatter plot of prediction vs OOD accuracy of ALine-D and ATC, the second best performing method from Table 2. We observe that our linear fit is closer to the diagonal, as ATC underperforms on models that have low OOD accuracy.

since DOC requires information from multiple OOD datasets. We also compare with the most recent, ProjNorm by Yu et al. [49]. Yu et al. [49] showed that ProjNorm has strong linear correlation with the OOD accuracy, more so than other methods such as [16] and [22]. We compare with this method separately in Section 5.1, as they do not provide a way to directly estimate the OOD accuracy.

5.1 Main results: comparison to other methods.

In Table 2, we observe that ALine-D outperforms ATC across all synthetic shift datasets and WILDS datasets. For the natural shift datasets, ALine-D outperformed other methods excluding the dataset shift from ImageNet to ImageNetV2 where ATC performed marginally better. As can be seen in Figure 3, ATC generally could not accurately predict the model's OOD performance for models that do not perform very well. This is consistent with experimental results in [22] and [49]. On the other hand, ALine performed equally well on "bad" models and "good" models. In some sense, given a collection of models where we're interested in the performance of each, ATC, AC, DOC-Feat, and Agreement only utilizes information from the model of interest, whereas ALine utilizes the collective information from all models for each individual prediction.

As expected, on datasets where we did not observe a linear correlation between ID and OOD agreement (and accuracy), ALine did not perform very well, with a mean absolute estimation error of around 5%. Interestingly, the other benchmarks also did not perform very well on these datasets, suggesting that perhaps the success of these prediction methods could also partially be attributed to accuracy-on-the-line. As proven in Garg et al. [22], there does not exist a single method that can successfully predict the OOD accuracy for any particular distribution shift. The advantage of ALine is that there is a concrete way to verify when our method will successfully predict the OOD accuracy (i.e. check whether agreement is on the line), but other benchmarks do not have any way of characterizing when they will be successful.

Additionally, we observe that ALine-D in particular does well, even when the approximation of the slope and bias are slightly off. In fact, ALine-D supercedes previous methods even when accuracy-on-the-line does not hold, suggesting that there exists some additional beneficial properties of ALine-D that requires further study.

Dataset	ALine-D*	ALine-S*	ATC	AC	DOC	Agreement
CIFAR10.1	**1.14**	1.20	1.55	1.09	5.82	5.98
CIFAR10.2	**4.07**	4.10	5.06	14.58	9.66	5.42
ImageNetV2	2.06	2.08	**1.12**	66.2	11.50	6.70
CIFAR10C-Fog	**1.45**	1.70	2.06	9.26	5.62	3.47
CIFAR10C-Snow	**1.42**	1.89	1.76	12.01	8.52	2.57
FMoW-WILDS	**1.78**	2.08	2.38	22.13	3.01	9.00
Camelyon17-WILDS	**5.83**	8.60	13.67	12.75	14.18	6.79
iWildCam-WILDS	**3.37**	4.05	13.24	21.66	4.53	4.81

Table 2: Mean Absolute Estimation Error (MAE) results for different datasets. We calulate the mean absolute difference between prediction vs OOD accuracy with $\%$ as units. ALine-D outperforms other methods on both shifts where we do and do not see accuracy-on-the-line. * denotes our methods.

5.2 Correlation analysis

Rather than looking at the error in OOD accuracy predictions, it could also be useful to have a metric that just strongly correlates with the OOD accuracy, particularly for model selection. Recently, Yu et al. [49] proposed ProjNorm, a measurement they show has a very strong linear correlation with OOD accuracy, moreso than other recent methods including Rotation [16] and ATC [22]. We compare the linear correlation between ALine-D's estimates of accuracy and the true OOD accuracy with that of ProjNorm and find that ALine-D achieves a high linear correlation and also outperforms ProjNorm. Specifically, we reproduce their experiment on ResNet18 across all corruptions and severity levels of CIFAR-10C (See their Table 1 in [49]). Since ALine-D is an algorithm that requires a set of models for prediction, we use the models from the CIFAR-10 testbed of Miller et al. [38], as we have across all experiments, as the other models in the model prediction set. (See Appendix D for experimental details.)

Dataset	ALine-D		ProjNorm	
	ρ	R^2	ρ	R^2
CIFAR-10C	**0.995**	**0.974**	0.98	0.973

Table 3: Correlation analysis. We compare the coefficients of determination (R^2) and rank correlations (ρ) between ALine-D and ProjNorm, a metric shown to have stronger correlation than ATC and Rotation.

5.3 Estimating performance along a training trajectory

So far in our experiments, ALine uses a large collection of models to get the linear fit and predict the true slope and bias of ID versus OOD accuracy. We assess whether ALine can be utilized even in situations where the practioner only cares about the performance of a few models. In such situations, one could efficiently gather many models by training a single model and saving checkpoints along the way. We analyze whether our phenomena is helpful for predicting such highly correlated hypotheses, instead of independently trained models. In Figure 4, we collect the logits of the ID validation set and OOD dataset every 5 epochs across the training of a single ResNet18 model trained on CIFAR-10 and compute the agreement between every pair. We see that even the agreement between the checkpoints of a model across training is enough to get a good linear fit that matches the slope and bias of CIFAR-10 versus CIFAR-10.1 accuracy. Thus, by applying ALine-D to these checkpoints, we can get a very good estimate of the OOD performance over the training. This suggests that given a model of interest, ALine does not require practitioners to train a large number of models, but just train one and save its predictions across training iterations. We do a more careful ablation study in the Appendix F, looking at the number of models required.

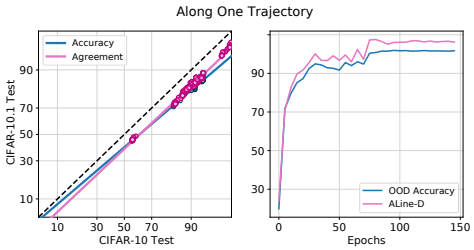

Figure 4: ALine-D tracks OOD accuracy across training with a MAE of 2.19%.

6 Conclusion

The contributions of this work are two-fold. First, we observe the agreement-on-the-line phenomena, and show that it correlates strongly with accuracy-on-the-line over a range of datasets and models. We also highlight that certain aspects of this phenomenon, namely the fact that the slope and bias of the linear fit is largely the *same* across agreement and accuracy, are specific to neural networks, and thus fundamentally seem connected to these classes of models. Second, using this empirical phenomenon, we propose a surprisingly simple but effective method for predicting OOD accuracy of classifiers, while only having access to unlabeled data from the new domain (and one that can be "sanity checked" via testing whether agreement-on-the line holds). Our method outperforms existing state-of-the-art approaches to this problem. Importantly, we do *not* claim that this phenomenon is universal, but we found it to be true across a wide range of neural networks and OOD benchmarks that we experimented on. In addition to its practical relevance, this observation itself reveals something very interesting about the way neural networks learn, which we leave for future study. A better understanding of what causes this phenomenon to hold and its failure modes could give novel insights about OOD generalization of neural networks.

References

[1] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018.

[2] Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 2017.

[3] Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020.

[4] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

[5] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

[6] Remi Cadene. Pretrained models for pytorch. https://github.com/cadene/pretrained-models.pytorch, 2018.

[7] Jiefeng Chen, Frederick Liu, Besim Avci, Xi Wu, Yingyu Liang, and Somesh Jha. Detecting errors and estimating accuracy on unlabeled data with self-training ensembles. *arXiv preprint arXiv:2106.15728*, 2021.

[8] Mayee Chen, Karan Goel, Nimit S Sohoni, Fait Poms, Kayvon Fatahalian, and Christopher Re. Mandoline: Model evaluation under distribution shift. *International Conference on Machine Learning*, page 1617–1629, 2021.

[9] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[10] Ching-Yao Chuang, Antonio Torralba 0001, and Stefanie Jegelka. Estimating generalization under distribution shifts via domain-invariant representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1984–1994. PMLR, 2020. URL http://proceedings.mlr.press/v119/chuang20a.html.

[11] Adam Coates and Andrew Y Ng. *Learning feature representations with k-means*. 2012. URL https://www-cs.stanford.edu/~acoates/papers/coatesng_nntot2012.pdf.

[12] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995. doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

[13] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[15] Weijian Deng and Liang Zheng. Are labels always necessary for classifier accuracy evaluation? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15064–15073, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society. doi: 10.1109/CVPR46437.2021.01482. URL https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.01482.

[16] Weijian Deng, Stephen Gould, and Liang Zheng. What does rotation prediction tell us about classifier accuracy under varying testing environments? *arXiv preprint arXiv:2106.05961*, 2021.

[17] Pinar Donmez, Guy Lebanon, and Krishnakumar Balasubramanian. Unsupervised supervised learning i: Estimating classification and regression errors without labels. *Journal of Machine Learning Research*, 11(44):1323–1351, 2010. URL http://jmlr.org/papers/v11/donmez10a.html.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[19] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[20] Hady Elsahar and Matthias Gallé. To annotate or not? predicting performance drop under domain shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2163–2173, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1222. URL https://aclanthology.org/D19-1222.

[21] S. Garg, Sivaraman Balakrishnan, J. Zico Kolter, and Zachary Chase Lipton. Ratt: Leveraging unlabeled data to guarantee generalization. In *International Conference of Machine Learning*, 2021.

[22] Saurabh Garg, Sivaraman Balakrishnan, Zachary Chase Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=o_HsiMPYh_x.

[23] Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1114–1124, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00117. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00117.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[25] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

[26] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

[27] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

[28] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

[29] Ariel Jaffe, Boaz Nadler, and Yuval Kluger. Estimating the accuracies of multiple classifiers without labeled data. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 407–415, San Diego, California, USA, 09–12 May 2015. PMLR. URL https://proceedings.mlr.press/v38/jaffe15.html.

[30] Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of SGD via disagreement. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=WvOGCEAQhxl.

[31] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

[32] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

[33] Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pages 942–950. PMLR, 2013.

[34] Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734, 2015. doi: 10.1109/ACPR.2015.7486599.

[35] Shangyun Lu, Bradley Nott, Aaron Olson, Alberto Todeschini, Hossein Vahabi, Yair Carmon, and Ludwig Schmidt. Harder or different? a closer look at distribution shift in dataset reproduction. *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020. URL www.gatsby.ucl.ac.uk/~balaji/udl2020/accepted-papers/UDL2020-paper-101.pdf.

[36] Omid Madani, David Pennock, and Gary Flake. Co-validation: Using model disagreement on unlabeled data to validate classification algorithms. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper/2004/file/92f54963fc39a9d87c2253186808ea61-Paper.pdf.

[37] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

[38] John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt§. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. 2021.

[39] Vaishnavh Nagarajan and J Zico Kolter. Deterministic pac-bayesian generalization bounds for deep networks via generalizing noise-resilience. *arXiv preprint arXiv:1905.13344*, 2019.

[40] Preetum Nakkiran and Yamini Bansal. Distributional generalization: A new kind of generalization, 2021. URL https://openreview.net/forum?id=iQxS0S9ir1a.

[41] Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel M. Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[42] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems 30, NeurIPS 2017*, 2017.

[43] Emmanouil Antonios Platanios, Avrim Blum, and Tom Mitchell. Estimating accuracy from unlabeled data. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, page 682–691, Arlington, Virginia, USA, 2014. AUAI Press. ISBN 9780974903910.

[44] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

[45] Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*, 2020.

[46] Sebastian Schelter, Tammo Rukat, and Felix Biessmann. Learning to validate the predictions of black box classifiers on unseen data. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, SIGMOD '20, page 1289–1299, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367356. doi: 10.1145/3318464.3380604. URL https://doi.org/10.1145/3318464.3380604.

[47] Jacob Steinhardt and Percy Liang. Unsupervised risk estimation using only conditional independence structure. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[48] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

[49] Yaodong Yu, Zitong Yang, Alexander Wei, Yi Ma, and Jacob Steinhardt. Predicting out-of-distribution error with the projection norm, 2022.

[50] Lijia Zhou, Danica J. Sutherland, and Nati Srebro. On uniform convergence and low-norm interpolation learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6867–6877. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/4cc5400e63624c44fadeda99f57588a6-Paper.pdf.

[51] Ji Zhu, Hui Zou, Saharon Rosset, and Trevor Hastie¶[. Multi-class adaboost. *Statistics and Its Interface*, 2, 2009.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provide code/instructions in the supplementary, but the testbed we use by Miller et. al. is proprietary that we received through communication with the authors.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We provided training details for any new models we've trained. The details for the CIFAR10 and WILDS testbed are provided in Miller et. al. Training details of models we simply evaluate are in the respective papers.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] Each point in the scatter plot represents one model or a pair of models.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Details are provided in appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] The license, PII, and consent details of datasets and models are in the respective papers, which we cite.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]