Accuracy on the Curve: On the Nonlinear Correlation of ML Performance Between Data Subpopulations

Weixin Liang [*1,2] **Yining Mao** [*2] **Yongchan Kwon** [*3] **Xinyu Yang** [4,5] **James Zou** [1,2,6]

Abstract

Understanding the performance of machine learning (ML) models across diverse data distributions is critically important for reliable applications. Despite recent empirical studies positing a near-perfect linear correlation between in-distribution (ID) and out-of-distribution (OOD) accuracies, we empirically demonstrate that this correlation is more nuanced under subpopulation shifts. Through rigorous experimentation and analysis across a variety of datasets, models, and training epochs, we demonstrate that OOD performance often has a nonlinear correlation with ID performance in subpopulation shifts. Our findings, which contrast previous studies that have posited a linear correlation in model performance during distribution shifts, reveal a "moon shape" correlation (parabolic uptrend curve) between the test performance on the majority subpopulation and the minority subpopulation. This non-trivial nonlinear correlation holds across model architectures, hyperparameters, training durations, and the imbalance between subpopulations. Furthermore, we found that the nonlinearity of this "moon shape" is causally influenced by the degree of spurious correlations in the training data. Our controlled experiments show that stronger spurious correlation in the training data creates more nonlinear performance correlation. We provide complementary experimental and theoretical analyses for this phenomenon, and discuss its implications for ML reliability and fairness. Our work highlights the importance of understanding the nonlinear effects of model improvement on performance in different subpopulations, and has the potential to inform the development of more equitable and responsible machine learning models.

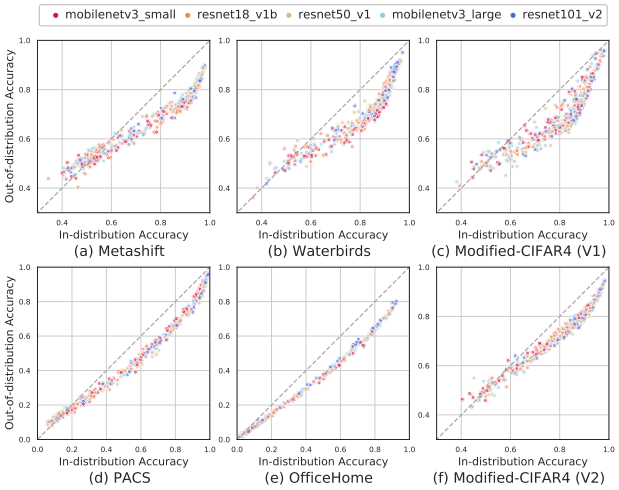

Figure 1: **A striking nonlinear correlation between out-of-distribution and in-distribution performance under subpopulation shifts.** Each dot represents a trained model and each panel represents a dataset. Our comprehensive experimentation, utilizing a variety of model architectures and hyperparameters, reveals a precise correlation between OOD and ID performance. The top panels (a-c) depict datasets constructed with spurious correlations, where the correlation is notably nonlinear. The bottom panels (d-f) depict datasets with rare subpopulations (absent of spurious correlations), where the correlation is more subtle, but as our analysis in Figure 2, illustrates, still nonlinear. Our findings have significant implications for understanding and improving the reliability and fairness of machine learning models.

1. Introduction

Machine learning (ML) models often exhibit vastly different performance and behaviors when applied to different data distributions. This can be a significant challenge in ML, as even the best efforts to create data that closely represents the real-world may not fully capture the dynamic,

[*]Equal contribution [1]Department of Computer Science, Stanford University, Stanford, CA, USA [2]Department of Electrical Engineering, Stanford University, Stanford, CA, USA [3]Department of Statistics, Columbia University, New York, NY, USA [4]Department of Computer Science and Engineering, Zhejiang University, Hangzhou, P.R.China [5]Department of Information Science, Cornell University, Ithaca, NY, USA [6]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. Correspondence to: Weixin Liang <wxliang@stanford.edu>, James Zou <jamesz@stanford.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

high-dimensional, and combinatoric complexity of many tasks. As a result, AI models deployed in the wild are likely to encounter out-of-distribution (OOD) data, raising fundamental questions such as, "Can you trust your model on out-of-distribution data? How does an AI model's in-distribution (ID) performance relate to its out-of-distribution (OOD) performance?"

Exciting progress has been made in addressing these questions on OOD performance, but classical domain adaptation theory only provides a partial answer. Built on the uniform-convergence framework, these classical works resort to bounding OOD performance by quantifying the distance between the ID and OOD (e.g., via the $\mathcal{H}\Delta\mathcal{H}$ divergence (Ben-David et al., 2010)), thereby producing an upper bound on OOD performance that becomes increasingly loose for larger distribution shifts (Redko et al., 2020).

Pioneering recent research, such as that by Taori et al. (2020); Miller et al. (2021); Kaplun et al. (2022), has uncovered an exciting phenomenon that is not captured by classical theory: an almost perfect linear correlation in probit scale between ID and OOD performance, which has been repeatedly found across a wide spectrum of OOD benchmarks. This phenomenon, known as "accuracy-on-the-line," holds not only for dataset reconstruction shifts (e.g., ImageNet-V2 (Recht et al., 2019), CIFAR-10.1 (Recht et al., 2018)), but also for more complex distribution shift benchmarks such as WILDS (Koh et al., 2021) and BREEDS (Santurkar et al., 2021).

The "accuracy-on-the-line" phenomenon has garnered significant interest and excitement, as it suggests that, given access to the slope and bias of the linear correlation, predicting OOD accuracy becomes straightforward. For example, Baek et al. (2022) have shown a related "agreement-on-the-line" phenomenon, observing that the OOD agreement between the predictions of any two pairs of neural networks also exhibits a strong linear correlation with their ID agreement. Furthermore, the slope and bias of OOD vs ID agreement closely match that of OOD vs ID accuracy, thereby enabling the prediction of OOD accuracy with just unlabeled data. This has been a long-standing research problem, highlighting the significance of "accuracy-on-the-line".

In this work, we investigate a common and challenging type of distribution shift known as *subpopulation shifts*. This phenomenon is frequently observed in real-world applications, such as medical AI models that perform differently when deployed on different sites with different demographics (Wu et al., 2021; Daneshjou et al., 2021).

Our research reveals a *nonlinear* correlation between in-distribution (ID) and out-of-distribution (OOD) performance under subpopulation shifts (as demonstrated in Figure 1). To gain insight into this phenomenon, we decompose the model's performance into performance on each subpopulation (as shown in Figure 2). We observe a consistent *"moon shape" correlation* (parabolic uptrend curve) between the test performance on the majority subpopulation and the minority subpopulation. These nonlinear correlations are observed across a variety of datasets, models, and training epochs (as depicted in Figure 4), and are also present in multi-subpopulation data (as shown in Figure 6) and under different distribution shift algorithms (Figure 7).

To understand the underlying causes of this nonlinear performance correlation, we conducted an extensive empirical analysis. Our results indicate that the degree of *spurious correlations* in training data plays a significant role. Spurious correlations refer to connections between variables that appear to be causal but are not (Pearl et al., 2000). We find that datasets with spurious correlations (as shown in Figure 2 top) exhibit more nonlinear performance correlations than datasets without spurious correlations (as shown in Figure 2 bottom). Our controlled study confirms that stronger spurious correlation leads to more nonlinear performance correlation (as depicted in Figure 8).

This research highlights an important issue with state-of-the-art AI models, which often pick up spurious correlations and biases in training data (Liang et al., 2022). These correlations may initially improve performance, but can fail catastrophically when deployed in slightly different environments. Furthermore, our findings indicate that current agreement-based approaches for predicting OOD performance *systematically overestimate* performance under the presence of spurious correlations (as shown in Figure 10), suggesting a need for new methods to address this problem.

It is worth emphasizing that this work does not contradict, but rather complements and extends previous work. Our work confirms the existence of *strong* and *precise* correlations between ID and OOD performance, which have not been fully captured by classical domain adaptation theory. Additionally, we have identified the presence of spurious correlation as a contributing factor to this broader performance correlation phenomenon. Although the correlations exhibit nonlinearity, they remain geometrically simple, opening up opportunities for the development of new methods for predicting OOD performance.

Beyond distribution shifts, the significance of our findings can also be viewed in the context of machine learning (ML) reliability and fairness. It is well-documented that a model can have disparate performances even within different subsets of its training and evaluation data (Eyuboglu et al., 2022b; Liang et al., 2023). Furthermore, these performance disparities can have a cascading effect, leading to decreased user retention and further amplifying the performance gap over time (Hashimoto et al., 2018; Fuster et al., 2017). For example, computer-vision AI models for diagnosing malig-

nant skin lesions performed substantially worse on lesions appearing on dark skin compared to light skin, with the area under the receiver operating curves (AUROC) dropping by 10-15% across skin tones (Daneshjou et al., 2021). Our work shows that ML performances between data subpopulations, albeit disparate, can have much more precise correlations than previously expected from the literature. Furthermore, we also identify certain situations in the presence of spurious correlations where performance improvement for the majority subpopulation leads to *alarmingly consistent performance deterioration* for the minority subpopulation. As the existence of ML performance disparities across subpopulations sternly undermines the trustworthiness, reliability, and fairness of ML models, our work makes a critical step towards the empirical understanding of *how ML performances between data subpopulations are correlated*. Concretely, this paper makes the following main **contributions**:

- To the best of our knowledge, we present the first systematic study on the performance correlation between data subpopulations. We found a nonlinear, *"moon shape"* correlation between the test performance on the *majority* subpopulation and the *minority* subpopulation (Figure 2). This indicates that ML performances between data subpopulations, albeit disparate, can have much more precise correlations than previously expected from the literature.
- In contrast to recent works reporting a near-perfect linear correlation, we found a nonlinear correlation under subpopulation shifts between ID and OOD accuracies (Figure 1). We empirically show that this non-trivial *nonlinear* correlation holds across model architectures, hyperparameters, training durations, and the imbalance between subpopulations (Figure 4). In addition, our experiments on multi-subpopulation datasets and different distribution shift algorithms beyond ERM further highlights the generality of this phenomenon. We also demonstrate how our findings complement and contrast previous empirical studies under probit-transformed axes (Figure 9). Our findings significantly broaden the scope of the broader correlation phenomenon between ID performance and OOD performance.
- Supported by extensive empirical and theoretical analysis (see Appendix E for detailed theoretical analysis), we identify the degree of *spurious correlations* in training data as an important cause for this nonlinearity. We demonstrate that datasets with spurious correlations (Figure 2 top) show *more nonlinear* correlations than datasets without spurious correlations (Figure 2 bottom). We conduct rigorous controlled studies confirming that stronger spurious correlations create more nonlinear performance correlations (Figure 8).
- We demonstrate that current agreement-based approach for predicting OOD performance would *systematically overestimate* under the existence of spurious correlations (Figure 10). We also identify certain regimes where performance improvement for the majority subpopulation leads *alarmingly* to *consistent performance deterioration* for the minority subpopulation, thereby highlighting the significance and implications of our findings for ML reliability and fairness.

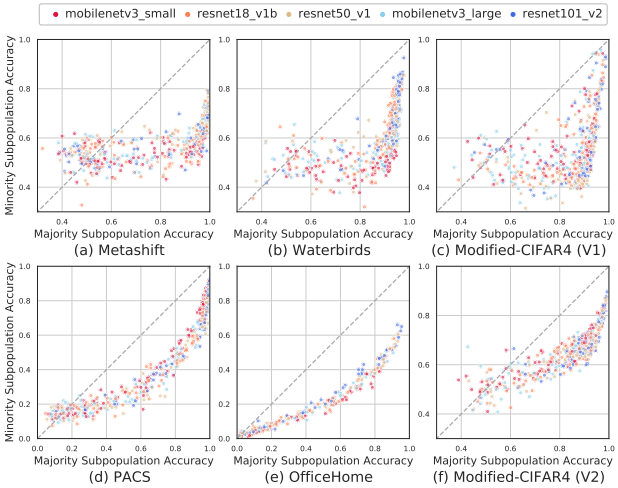

Figure 2: **A performance breakdown of Figure 1: majority subpopulation vs. minority subpopulation.** To gain a deeper understanding of the striking nonlinear correlation between out-of-distribution (OOD) and in-distribution (ID) accuracies observed in Figure 1, we decompose the model's performance into separate evaluations on the majority and minority subpopulations of the OOD test set. **The results reveal a clear and striking nonlinear correlation, which we term the "*moon shape*" correlation, between the majority subpopulation performance and the minority subpopulation performance.** This nonlinearity is particularly pronounced in datasets constructed with spurious correlations, as seen in the top panels (a-c), while datasets without such correlations, shown in the bottom panels (d-f), exhibit more subtle nonlinearity.

2. Experimental Setup

Preliminaries: Machine Learning with Diverse Subpopulations In our study, we investigate the performance of various machine learning (ML) models in the presence of diverse subpopulations in the data distribution. Specifically, the overall data distribution is denoted as $\mathcal{D} = 1, \ldots, D$, where each subpopulation $d \in \mathcal{D}$ corresponds to a fixed data distribution P_d. In our main experiments, we compare the performance of ML models on two different data distributions: (1) the in-distribution (ID), or the training distribution, $P^{tr} = \sum_{d\in\mathcal{D}} r_d^{tr} P_d$, where r_d^{tr} denotes the mixture probabilities in the training set, and (2) the out-of-distribution (OOD), which is also a mixture of the D subpopulations, $P^{ts} = \sum_{d\in\mathcal{D}} r_d^{ts} P_d$, where r_d^{ts} is the mixture probabilities

in the test distribution, but with a different proportion from the training distribution, i.e., $r_d^{ts} \neq r_d^{tr}$ for some $d \in \mathcal{D}$. This setting is known as subpopulation shifts in the literature (Yao et al., 2022; Koh et al., 2021).

Experimental Procedure We consider $D = 2$ subpopulations for simplicity. For the in-distribution (training distribution), we consider a dominating *majority subpopulation* (e.g., $r_d^{tr} = 90\%$) and an underrepresented *minority subpopulation* (e.g., $r_d^{tr} = 10\%$). As for the out-of-distribution, the majority subpopulation and minority subpopulation are equally representative (e.g., $r_d^{ts} = 50\%$). The goal of our paper is to compare the ID performance and the OOD performance across a wide spectrum of ML models. Therefore, for each subpopulation shifts dataset, our experimental procedure follows two steps:

1. Train *500* different ML models independently on the same training set drawn from the training distribution P^{tr} using the empirical risk minimization (ERM) method by varying the model architectures, training durations, and hyperparameters following the search space of commercial AutoML (AutoGluon). Details of the training process can be found in the appendix.
2. For each trained ML model, evaluate the *ID performance* on a test set of held-out samples from the training distribution P^{tr}, and the *OOD performance* on a test set drawn from the out-of-distribution P^{ts}. We then visualize the correlation of ID and OOD performance on a scatter plot.

Subpopulation Shift Datasets Based on our survey on the reported cases of ML performance disparity on the minority subpopulation in the wild and prior work (Eyuboglu et al., 2022a; Oakden-Rayner et al., 2020), we identified and evaluated on two important categories of subpopulation shift datasets (Figure 3):

- **Spurious correlation.** In statistics, a spurious correlation refers to a connection between two variables that appear to be causal, but are not. For example, Figure 3 (a) illustrates a scenario where cat images are mostly indoor and dog images are mostly outdoor (as indicated by the red boxes). A spurious correlation exists between the class labels and the indoor/outdoor contexts. To explore this scenario, we have experimented with three existing datasets in the community: MetaShift (Liang & Zou, 2022), Waterbirds (Sagawa et al., 2020b), and Modified-CIFAR4 V1 (Rolf et al., 2021).
- **Rare subpopulation.** ML models can still underperform on subpopulations that occur infrequently in the training set, even without the presence of obvious spurious correlation. To explore this scenario, we have adopted PACS (Li et al., 2017), OfficeHome (Venkateswara et al., 2017), and Modified-CIFAR4 V2 (Rolf et al., 2021).

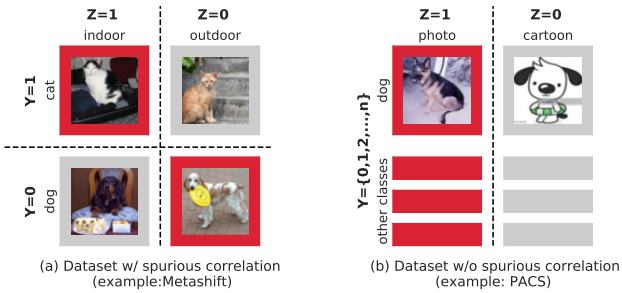

Figure 3: **Evaluation of subpopulation shift in two dataset configurations: presence or absence of spurious correlations.** As depicted, Y represents the class label, with $Z = 1$ and $Z = 0$ indicating the majority and minority subpopulations, respectively. **Left:** In the presence of spurious correlation, there exists a correlation between target label Y and Z (e.g. as exemplified by the red boxes, where cat images tend to be predominantly indoor and dog images outdoor); **Right:** Conversely, in the absence of spurious correlation, Y is statistically independent of Z.

3. The Moon Shape Phenomenon

In this section, we empirically show the *nonlinear* correlation between the ID and OOD performance across multiple subpopulation shifts datasets. To understand this phenomenon, we decompose the model's performance into performance on each subpopulation. We also found *nonlinear* correlation between the test performance on the *majority subpopulation* and the *minority subpopulation*. Moreover, this nonlinearity holds across model architectures, training durations and hyperparameters, and the imbalance between subpopulations.

3.1. Nonlinear Correlation of ML Performance Across Data Subpopulations

Out-of-Distribution vs. In-Distribution Prior research reports a near perfect *linear* correlation between the OOD accuracies and the ID accuracies. In contrast, Figure 1 shows the *nonlinear* correlation between the ID and OOD performance across multiple subpopulation shifts datasets. Moreover, datasets constructed with spurious correlations (Figure 1 Top) occurs more *nonlinear* compared to the datasets with only rare subpopulations (Figure 1 Bottom), which we further confirm and analyze below.

Majority vs. Minority To understand this phenomenon, we decompose the model's performance into performance on each subpopulation. As shown in Figure 2, there is a

"moon shape" correlation (parabolic uptrend curve) between the test performance on the *majority subpopulation* and the *minority subpopulation*. Since we have decomposed by subpopulations, the *nonlinearity* becomes much more visually apparent. Figure 2 also confirms that datasets with *spurious correlations* (Figure 2 Top) show more nonlinearity compared to those without (Figure 2 Bottom), which motivates our further analysis on how *spurious correlation* affect the correlation *nonlinearity* (§ 4).

Converting from Figure 1 to Figure 2: We clarify that the test set is always balanced, with a 50/50 majority to minority ratio when $D = 2$. In our subpopulation shift setting, the correlation between "Majority Subpopulation Accuracy vs. Minority Subpopulation Accuracy" can be directly connected to "In-distribution Accuracy vs. Out-of-distribution Accuracy". This is because, in our setting, the in-distribution consists of a mixture of 90% majority subpopulation and 10% minority subpopulation, while the test distribution (out-of-distribution) is composed of a 50% majority subpopulation and 50% minority subpopulation. Consequently, we have:

- In-distribution Accuracy = 0.9 × Majority Subpopulation Accuracy + 0.1 × Minority Subpopulation Accuracy
- Out-of-distribution Accuracy = 0.5 × Majority Subpopulation Accuracy + 0.5 × Minority Subpopulation Accuracy

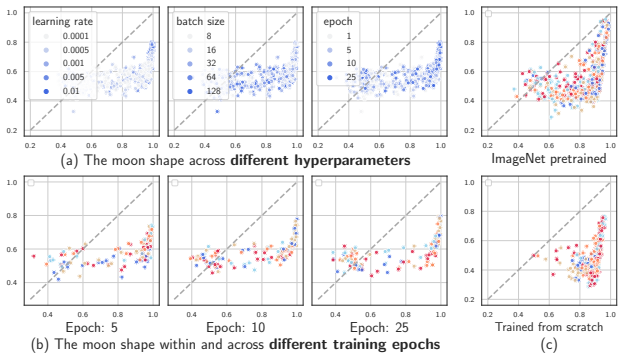

Figure 4: **Demonstrating the Consistency of the Moon Shape Phenomenon Across Various Factors.** The x-axis of each panel represents the performance of the majority subpopulation, while the y-axis represents the performance of the minority subpopulation. **(a)** (Metashift) The moon shape is present regardless of the model architecture, hyperparameters, or training duration utilized. **(b)** (Metashift) The moon shape is evident at each snapshot and persists across different training epochs. **(c)** (Modified-CIFAR4 (V1)) The moon shape is observed in both models that are pretrained on ImageNet and models that are trained from scratch.

Furthermore, we demonstrate that the moon shape phenomenon holds across various factors including (Figure 4):

- **Model architectures, training durations and hyperparameters (Figure 4(a)).** Following the search space of a commercial AutoML library (AutoGluon), we varied (1) pretrained model architectures, (2) training durations (i.e., training epochs), (3) hyperparameters such as learning rates and batch size. We found that models lies consistently on the same *"moon shape"*.
- **Training dynamics (Figure 4 (b)).** We further stratify the dots in Figure 2 (a) (i.e. the trained models) by the number of training epochs. We still find a clear moon shape for each fixed training epoch. Moreover, similar moon shapes persist across different training epochs. Results on other datasets are similar (Supp. Figure 11). This finding motivates us to focus our analysis on comparing across *different models* rather than comparing the subpopulation performance of a single model across training epochs (which is an interesting direction complementary to our scope).
- **Pretrained vs. trained from scratch (Figure 4(c)).** Although fine-tuning pretrained models has become a modern paradigm of ML, we also add an experiment of training from scratch, confirming that *moon shape* persists even when training from scratch. This shows that the moon shape is not an artifact of ImageNet pre-trained models, but a much broader phenomenon.

3.2. Discussion: Why the Moon Shape is not Obvious

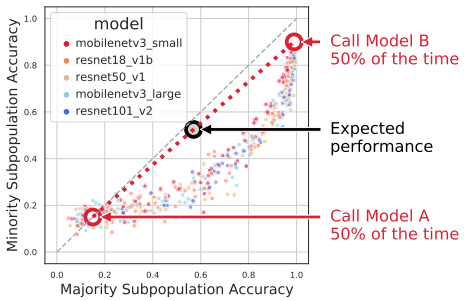

Figure 5: Why the moon shape is not obvious. Mixture of models can fill in the moon shape.

Figure 5 illustrates one reason why the non-linear correlation structure, represented by the moon shape, is not obvious. A thought experiment is presented in which two models, A and B (indicated by red circles), are interpolated by flipping a biased coin with probability p. If the coin lands heads, classification is performed with model A, and if it lands tails, classification is performed with model B. Varying p in the range $[0, 1]$ generates a line between the two models. This interpolation line represents an achievable region for machine learning models, yet our results demonstrate that all models deviate substantially from this line, resulting in the unexpected moon shape.

In summary, the moon shape is intriguing because it contradicts and extends the near-perfect linear correlation reported

by previous studies, revealing that performance correlation is more nuanced under subpopulation shifts. Additionally, while it would be expected that the dots of a scatter plot would reside in the *lower triangular area* since machine learning models generally perform worse on the minority subpopulation, our results show that the performance correlation is more concentrated than anticipated, with the dots *concentrated on one curve* rather than spreading out in the lower triangular area.

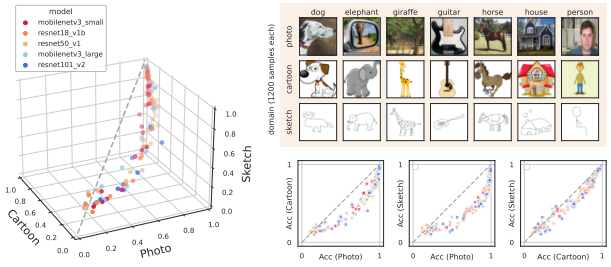

Figure 6: **Empirical demonstration of the 3D moon shape phenomenon.** Results of experiments conducted on the PACS dataset, comprising three subpopulations of images (Cartoon, Photo, and Sketch) are depicted. The figure illustrates that the moon shape phenomenon is not limited to just two subpopulations, but extends to three subpopulations as well. Importantly, it is demonstrated that the distribution of the number of training samples among the subpopulations is *not* a necessary condition for the emergence of the moon shape phenomenon.

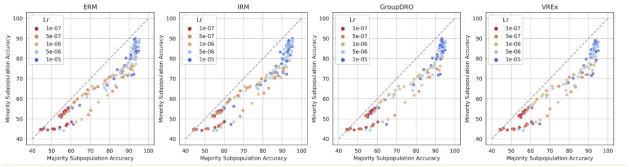

Figure 7: **Generalizability of the moon shape phenomenon beyond Empirical Risk Minimization (ERM).** Our experiments on MetaShift demonstrate that the moon shape phenomenon, previously observed in models trained via ERM, also holds for models trained with various distribution shift algorithms such as Invariant Risk Minimization, GroupDRO, and VREx. This suggests that the "moon shape" is a general phenomenon in machine learning, rather than being specific to ERM.

3.3. Multi-Subpopulation: 3D Moon Shape

To establish that the moon shape phenomenon is not limited to just two subpopulations, we conducted an additional experiment on the PACS dataset, which comprises three subpopulations of images (Cartoon, Photo, and Sketch). The results, depicted in Figure 6, clearly exhibit a 3D moon shape, providing empirical evidence for the generality of the moon shape phenomenon beyond two subpopulations. Furthermore, it is noteworthy that the three subpopulations in our experiment possess an equal number of samples, thereby demonstrating that the *imbalance* of training datasets is *not* a necessary condition for the emergence of nonlinear correlation as represented by the moon shape phenomenon.

3.4. Generalizability of the Moon Shape Phenomenon to Distribution Shift Algorithms

Distribution shift remains a significant challenge in machine learning, leading to the development of various algorithms to address it (Liu et al., 2021). To explore the generality of the "moon shape" phenomenon, we conducted additional experiments on distribution shift algorithms beyond Empirical Risk Minimization (ERM), including Invariant Risk Minimization (IRM)(Arjovsky et al., 2019), GroupDRO(Sagawa et al., 2020a), and VREx (Krueger et al., 2021). We employed the implementation and hyper-parameter range from DomainBed (Gulrajani & Lopez-Paz, 2020) to ensure rigorous and controlled experimental conditions. The results in Figure 7 demonstrate that the moon shape phenomenon also holds for IRM, GroupDRO, and VREx, indicating that it is a general phenomenon across different distribution shift algorithms.

4. The Impact of Spurious Correlation on the Moon Shape

4.1. Controlled Experiments: Spurious Correlation Makes the Moon Shape more Nonlinear

In the previous section (§ 3), we observed that datasets with spurious correlations exhibit increased nonlinearity compared to those without (as shown in Figure 1 and Figure 2, Top vs Bottom). In this subsection, we conduct well-controlled experiments to quantify the effect of spurious correlation on nonlinearity.

Experiment Design We use Modified-CIFAR4 (V1), a subset of the bird, car, horse, and plane classes from CIFAR-10 created by (Rolf et al., 2021), as our fixed dataset. We manipulate the degree of spurious correlation between the classification target label (air/land) and the spurious feature (vehicle/animal) in the training data by altering the mixture weights, while maintaining a fixed number of training data points in total (10,000), for each class (5,000), and for each spurious feature (6,000 for vehicle, 4,000 for animal). Additional details can be found in the appendix.

Results and Analysis Figure 8 illustrates the results of our experiments with increasingly stronger levels of spurious correlation. As predicted, nonlinearity in performance correlation increases with stronger levels of spurious correlation in the training data. These findings demonstrate that the presence of spurious correlation plays a significant role in shaping the relationship between out-of-distribution and in-distribution performance, an aspect that has been

previously overlooked in the literature.

4.2. Nonlinearity under Probit-Transformed Axes

Previous research has reported a near-perfect linear correlation between in-distribution and out-of-distribution accuracies, with some studies utilizing probit-transformed axes to enhance the linear trend (Miller et al., 2021; Taori et al., 2020). In this work, we use the probit transform, which is the inverse of the cumulative density function of the standard Gaussian distribution, to make our results directly comparable to previous studies.

Our findings reveal that, for datasets with spurious correlations, the performance correlation remains nonlinear even under probit transformation (Figure 9 top). This confirms that the nonlinear correlations we found are not captured by previous research. However, for datasets without spurious correlations, the performance correlation becomes linear after probit-transformation (Figure 9 bottom), as demonstrated by the significant difference in R^2 values in the linear fit. This highlights the importance of spurious correlations and how our work complements previous studies.

4.3. Implications of Model Agreements for Out-of-Distribution Performance Prediction

Motivation The ability to accurately predict out-of-distribution (OOD) performance is a valuable application of the *accuracy-on-the-line* phenomenon, as outlined by Baek et al. (2022). They propose using model agreements, which are calculated by evaluating the consistency of predictions between pairs of models, as a means of achieving this. They observe a strong linear correlation between OOD and in-distribution (ID) performance for model agreements, referred to as the *agreement-on-the-line* phenomenon. Additionally, they find a precise match between model accuracy and model agreements, which allows for OOD performance prediction using unlabeled data.

The agreement-on-the-line approach proposed by Baek et al. (2022) reveals a fascinating phenomenon related to the agreement between pairs of neural network classifiers. Specifically, when accuracy-on-the-line holds, a strong linear correlation is observed between out-of-distribution (OOD) agreement and in-distribution (ID) agreement for any two pairs of neural networks, regardless of architectural differences. Notably, the linear trend of ID vs. OOD agreement exhibits the same empirical slope and intercept as the linear trend between ID and OOD accuracy. This discovery, referred to as "agreement-on-the-line," has significant practical implications.

The agreement-on-the-line phenomenon allows for the prediction of OOD accuracy for classifiers without the need for labeled data. OOD agreement can be estimated using only unlabeled data by assessing the disagreement on an unlabeled dataset between pairs of neural networks trained with different sources of randomness. This provides the empirical slope and intercept of ID vs. OOD agreement, which can then be used as the empirical slope and intercept of ID and OOD accuracy.

Inspired by the potential of agreement-based OOD prediction, we conduct experiments on subpopulation shift datasets and present the results using smoothing cubic spline interpolation (Figure 10).

Results Our results indicate that the presence of spurious correlation can significantly impact the effectiveness of the agreement-based approach. In the absence of spurious correlation, the agreement results align almost perfectly with accuracy results, consistent with previous work. However, with spurious correlation, the agreement results deviate from accuracy results, forming a distinct moon-shaped pattern that lies above the accuracy moon shape.

This deviation is alarming as it suggests that naively applying the current agreement-based approach for predicting OOD performance would *systematically overestimate* performance in the presence of spurious correlation. This can be understood intuitively as the presence of spurious correlation causes models to rely on a variable correlated with the class label for predictions, breaking the independence of each model's prediction error and leading to increased agreement between models.

5. Related Work

Linear Correlations Between ID and OOD Performances Existing research mostly reports *linear* correlations between ID and OOD performances. The linear correlations were first reported in recent dataset reconstruction settings including ImageNet-V2 (Recht et al., 2019), CIFAR-10.1 (Recht et al., 2018), CIFAR-10.2 (Lu et al., 2020), where new test sets of popular benchmarks are collected closely following the original dataset creation process. As there are subtle differences in the dataset creation pipeline, the test performance on the new test set is often lower, but appears to be linearly correlated with the performance on the original test set (Lu et al., 2020; Miller et al., 2020; Recht et al., 2018; 2019; Yadav & Bottou, 2019). Later researchers also found the linear trends in the context of cross-benchmark evaluation (Taori et al., 2020; Miller et al., 2021), and transfer learning (Kornblith et al., 2019; Andreassen et al., 2021), where a model's ImageNet test accuracy linearly correlates with the transfer learning accuracy. Similar linear trends are also observed in sub-type shifts (Hendrycks & Dietterich, 2019; Santurkar et al., 2021) (e.g., the training data for the "dog" class are all from a specific breed while the test data come from another breed). Different from these studies, we

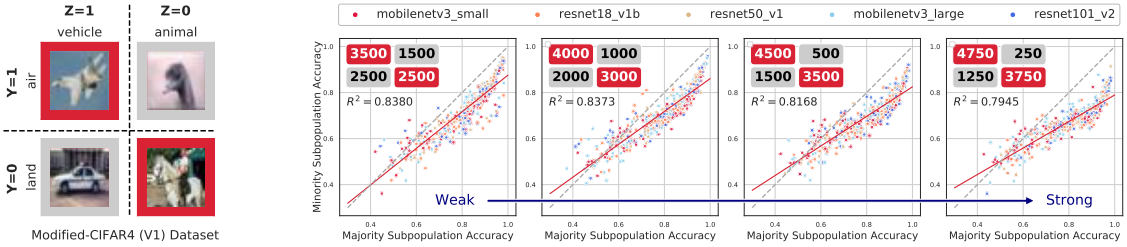

Figure 8: **Stronger spurious correlation creates more nonlinear performance correlation.** The left panel depicts binary classification on Modified-CIFAR4 (V1) with two subpopulations: a majority subpopulation (Z=1) and a minority subpopulation (Z=0). The right panel illustrates that as spurious correlation increases (i.e., more samples in the red boxes), nonlinear performance correlation also increases. The four panels represent different training data, with the number of training samples indicated by the 2x2 tables. The total number of training samples (10,000 images) and the majority:minority ratio (6,000:4,000) are held constant, with evaluation data also fixed.

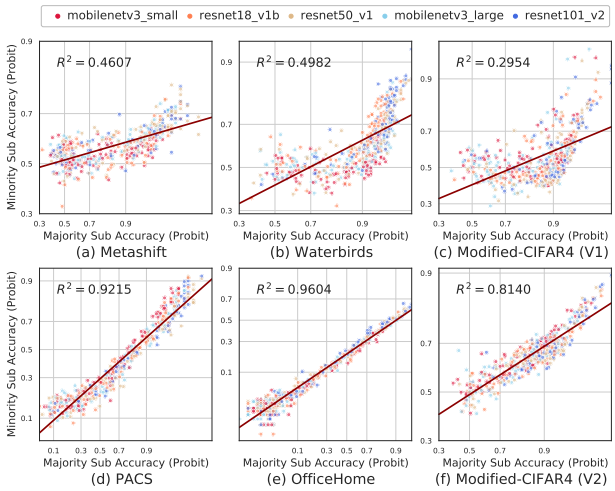

Figure 9: **Probit-scaled comparison of majority and minority subpopulation accuracy.** The top row (a-c) illustrates that, in datasets with spurious correlation, the performance correlation remains nonlinear even under probit transformation. In contrast, the bottom row (d-f) demonstrates that, in the absence of spurious correlation, a linear correlation emerges under probit scaling, as previously reported in literature. Linear fits and corresponding R^2 values are included for reference.

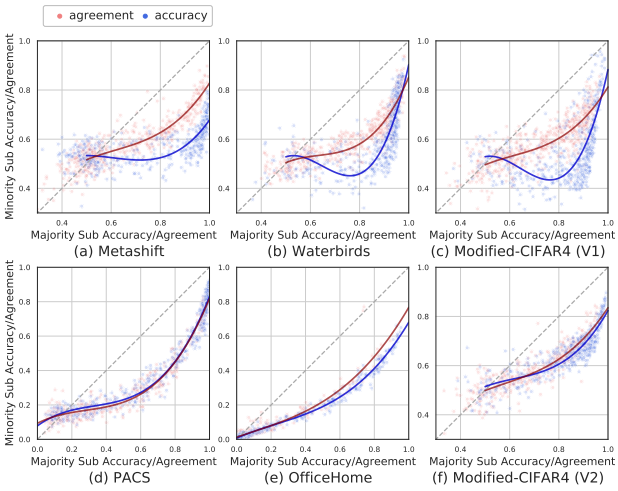

Figure 10: **Model agreements and accuracies in the presence and absence of spurious correlation.** Panels (a-c) demonstrate the significant disparity between the two measures when spurious correlation is present, while panels (d-f) reveal near-perfect alignment between them in the absence of such correlation. To prevent overcrowding, only 500 randomly chosen model agreement pairs with the 500 model accuracies are depicted and both sets of data are smoothed using cubic spline fitting.

(1) focus on subpopulation shifts, where we also present the first systematic study on the performance correlation between data subpopulations, and (2) find *nonlinear* correlations of ML performance across data subpopulations, which is not captured in previous work. Importantly, we show that for datasets with spurious correlations, the performance correlations still remain nonlinear under probit scale. This confirms that the nonlinear ("moon shape") correlation phenomenon is indeed not captured by previous work.

ML with Diverse Subpopulations A major challenge in ML is that a model can have very disparate performances even when it is applied to different subpopulations of its training and evaluation data. Models with low average error can still fail on particular groups of data points (Hashimoto et al., 2018; Buolamwini & Gebru, 2018; Blodgett et al., 2016). For example, (Pfohl et al., 2022) reported that predictive models for clinical outcomes which are accurate on average in a patient population underperform drastically for some subpopulations, potentially introducing or reinforcing inequities in care access and quality. Similar performance disparity has also been observed in radiograph classification (Badgeley et al., 2019; Zech et al., 2018; DeGrave et al., 2021), face recognition (Grother et al., 2011; Buolamwini & Gebru, 2018), speech recognition (Koenecke et al., 2020; Blodgett et al., 2016; Jurgens et al., 2017), academic recommender systems (Sapiezynski et al., 2017)), and automatic video captioning (Tatman, 2017), among others. As the

existence of ML performance disparity across subpopulations sternly undermines the trustworthiness, reliability, and fairness of ML models, our work makes a critical step towards the empirical understanding of how ML performances between data subpopulations are correlated.

6. Discussion

In this study, we presented a novel and significant discovery of a nonlinear correlation, which we term the "moon shape" phenomenon, between the performance of models on majority and minority data subpopulations. Through meticulous experimentation and analysis across a variety of datasets, models, and training epochs, we demonstrated that this phenomenon is persistent and has far-reaching implications for model selection and performance evaluation. We emphasize the following key aspects regarding the clarification of implications of the moon shape phenomenon:

Rethinking Accuracy-on-the-Line: Our findings present counterexamples to the previously reported linear correlation between in-distribution (ID) and out-of-distribution (OOD) model performance during distribution shifts. We identify the moon shape phenomenon, which exhibits nonlinear yet precise correlations between ID and OOD accuracies, expanding the conventional understanding and offering new insights (Taori et al., 2020; Miller et al., 2021; Kaplun et al., 2022; Lu et al., 2020; Miller et al., 2020; Recht et al., 2018; 2019; Yadav & Bottou, 2019; Kornblith et al., 2019).

Implications for OOD Performance Estimation: Our results reveal that practitioners relying on a linear trend to predict OOD performance may systematically *overestimate* model performance under subpopulation shifts, particularly when spurious correlations are present. This finding suggests that alternative methods are needed to accurately estimate OOD performance. One plausible explanation for this observation is that different ML models, regardless of their performance levels, tend to capture similar spurious correlations in the training data to varying degrees. This breaks the independence of each model's prediction error, leading these models to make similar errors and resulting in higher agreement between them.

Impact on Machine Learning Reliability and Fairness: Our work highlights that ML performance between data subpopulations can display more intricate correlations than previously assumed. Furthermore, we demonstrate situations where, given the presence of spurious correlations, performance improvement for the majority subpopulation coincides with a consistent decline in performance for the minority subpopulation. This observation has significant implications for the reliability and fairness of ML models in real-world applications.

The increasing use of automated machine learning (AutoML) in model building makes selecting models that perform well across diverse subpopulations a paramount challenge. Our results indicate that when there is no spurious correlation, models with higher aggregate performance tend to also perform well on minority subpopulations. However, in the presence of spurious correlation, the situation becomes more complex, with a phase transition point between negative and positive correlation. In settings where subpopulations performance is important (e.g. fairness considerations), we recommend AutoML practitioners to use similar type of scatter plots as our Figure 2 to diagnose model selection.

Subpopulation shift is a ubiquitous phenomenon in machine learning applications, and our work highlights the importance of understanding the nonlinear effects of model improvement on performance in different subpopulations. Further research and analysis of this nonlinear pattern is a crucial direction for future work, as it has the potential to inform the development of more equitable and responsible machine learning models, and contribute to addressing the fairness considerations in ML.

Data and Code Availability

All original data and code has been deposited at Github under `https://github.com/yining-mao/Moon-Shape-ICML-2023` and is publicly available as of the date of publication.

Acknowledgements

We thank Jonas Mueller, Rasool Fakoor and Shirley Wu for discussions. J.Z. is supported by the National Science Foundation (CCF 1763191 and CAREER 1942926), the US National Institutes of Health (P30AG059307 and U01MH098953) and grants from the Silicon Valley Foundation and the Chan-Zuckerberg Initiative.

Declaration of Interests

The authors declare no conflict of interest.

References

Andreassen, A., Bahri, Y., Neyshabur, B., and Roelofs, R. The evolution of out-of-distribution robustness throughout fine-tuning. *CoRR*, abs/2106.15831, 2021.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019.

AutoGluon. AutoGluon: AutoML for Text, Image, and Tabular Data. `https://auto.gluon.ai/`, 2022.

Badgeley, M. A., Zech, J. R., Oakden-Rayner, L., Glicksberg, B. S., Liu, M., Gale, W., McConnell, M. V., Percha, B., Snyder, T. M., and Dudley, J. T. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*, 2(1):1–10, April 2019.

Baek, C., Jiang, Y., Raghunathan, A., and Kolter, Z. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *arXiv preprint arXiv:2206.13089*, 2022.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Blodgett, S. L., Green, L., and O'Connor, B. T. Demographic dialectal variation in social media: A case study of african-american english. In *EMNLP*, pp. 1119–1130. The Association for Computational Linguistics, 2016.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 2018.

Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., and Zhang, Z. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015.

Daneshjou, R., Vodrahalli, K., Liang, W., Novoa, R. A., Jenkins, M., Rotemberg, V. M., Ko, J. M., Swetter, S. M., Bailey, E. E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Zou, J., and Chiou, A. S. Disparities in dermatology ai: Assessments using diverse clinical images. *ArXiv*, abs/2111.08006, 2021.

DeGrave, A. J., Janizek, J., and Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, May 2021.

Eyuboglu, S., Varma, M., Saab, K. K., Delbrouck, J.-B., Lee-Messer, C., Dunnmon, J., Zou, J., and Re, C. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=FPCMqjI0jXN.

Eyuboglu, S., Varma, M., Saab, K. K., Delbrouck, J.-B., Lee-Messer, C., Dunnmon, J., Zou, J., and Re, C. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=FPCMqjI0jXN.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. Predictably unequal? the effects of machine learning on credit markets. *Regulation of Financial Institutions eJournal*, 2017.

Grother, P. J., Grother, P. J., Phillips, P. J., and Quinn, G. W. *Report on the evaluation of 2D still-image face recognition algorithms*. Citeseer, 2011.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1934–1943. PMLR, 2018.

Hendrycks, D. and Dietterich, T. G. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR (Poster)*. OpenReview.net, 2019.

Jurgens, D., Tsvetkov, Y., and Jurafsky, D. Incorporating dialectal variability for socially equitable language identification. In *ACL (2)*, pp. 51–57. Association for Computational Linguistics, 2017.

Kaplun, G., Ghosh, N., Garg, S., Barak, B., and Nakkiran, P. Deconstructing distributions: A pointwise framework of learning. *arXiv preprint arXiv:2202.09931*, 2022.

Koenecke, A., Nam, A. J. H., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117:7684 – 7689, 2020.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A benchmark of in-the-wild distribution shifts. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021.

Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *CVPR*, pp. 2661–2671. Computer Vision Foundation / IEEE, 2019.

Krueger, D., Caballero, E., Jacobsen, J., Zhang, A., Binas, J., Zhang, D., Priol, R. L., and Courville, A. C. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 2021.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. *2017 IEEE*

International Conference on Computer Vision (ICCV), pp. 5543–5551, 2017.

Liang, W. and Zou, J. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=MTex8qKavoS.

Liang, W., Tadesse, G. A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., and Zou, J. Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, 4(8):669–677, 2022.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. Gpt detectors are biased against non-native english writers. *arXiv preprint arXiv:2304.02819*, 2023.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, 2021.

Lu, S., Nott, B., Olson, A., Todeschini, A., Vahabi, H., Carmon, Y., and Schmidt, L. Harder or different? a closer look at distribution shift in dataset reproduction. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020. http://www.gatsby.ucl.ac.uk/~balaji/udl2020/accepted-papers/UDL2020-paper-101.pdf.

Miller, J., Krauth, K., Recht, B., and Schmidt, L. The effect of natural distribution shift on question answering models. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6905–6916. PMLR, 2020.

Miller, J., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7721–7735. PMLR, 2021.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Re, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, pp. 151–159, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370462. doi: 10.1145/3368555.3384468. URL https://doi.org/10.1145/3368555.3384468.

Pearl, J. et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000.

Pfohl, S. R., Zhang, H., Xu, Y., Foryciarz, A., Ghassemi, M., and Shah, N. H. A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. *Scientific reports*, 12(1):1–13, 2022.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do CIFAR-10 classifiers generalize to cifar-10? *CoRR*, abs/1806.00451, 2018.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 2019.

Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*, 2020.

Rolf, E., Worledge, T. T., Recht, B., and Jordan, M. I. Representation matters: Assessing the importance of subgroup allocations in training data. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9040–9051. PMLR, 2021.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *ICLR*. OpenReview.net, 2020a.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020b.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8346–8356. PMLR, 2020c.

Santurkar, S., Tsipras, D., and Madry, A. BREEDS: benchmarks for subpopulation shift. In *ICLR*. OpenReview.net, 2021.

Sapiezynski, P., Kassarnig, V., and Wilson, C. Academic performance prediction in a gender-imbalanced environment. In *FATREC Workshop on Responsible Recommendation*, 2017.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. In *NeurIPS*, 2020.

Tatman, R. Gender and dialect bias in youtube's automatic captions. In *EthNLP@EACL*, 2017.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017.

Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D. E., and Zou, J. How medical ai devices are evaluated: limitations and recommendations from an analysis of fda approvals. *Nature Medicine*, 27(4):582–584, 2021.

Yadav, C. and Bottou, L. Cold case: The lost MNIST digits. In *NeurIPS*, pp. 13443–13452, 2019.

Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., and Finn, C. Improving out-of-distribution robustness via selective augmentation. In *Proceeding of the Thirty-ninth International Conference on Machine Learning*, 2022.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

A. Extended Description of Experiment Setups

A.1. Subpopulation Shift Datasets

We categorize our subpopulation shift datasets based on the underlying reason that ML models exhibits degraded performance on the minority subpopulation. Based on our survey and prior work (Eyuboglu et al., 2022a; Oakden-Rayner et al., 2020), we identified two important scenarios and the corresponding datasets as follows:

- **Spurious correlation.** In statistics, a spurious correlation refers to a connection between two variables that appear to be causal but are not. Take the Cat vs. Dog task as an example: With cat images mostly indoor and dog images mostly outdoor, A spurious correlation exists between the class labels and the indoor/outdoor contexts (Liang & Zou, 2022). To explore the scenario of spurious correlation, we experiment with three existing datasets in the community: MetaShift (Liang & Zou, 2022), Waterbirds (Sagawa et al., 2020b), and Modified-CIFAR4 V1 (Rolf et al., 2021). The detailed setups are as follows:
  - **MetaShift-Cat-Dog** (Liang & Zou, 2022): A cat vs. dog binary classification task, where cat images are mostly indoor and dog images are mostly outdoor. We choose indoor context as the majority subpopulation and outdoor context as the minority subpopulation. In the training set, with spurious correlation, 88% of the cat images are *indoors*, and 88% of the dog images are *outdoors*. In the ID test set, the two subpopulations are in the same proportion as the train set; In the OOD test set, the two subpopulations are equally represented. Same for the following datasets.
  - **Waterbirds** (Sagawa et al., 2020b): A waterbird vs. landbird binary classification task, with waterbirds (landbirds) more frequently appearing against a water (land) background in the training distribution. Here the majority subpopulation is water background and the minority subpopulation is land background. We select 80% of the waterbird images with water background, and 80% of the landbird images with land background.
  - **Modified-CIFAR4 V1**: Created by (Rolf et al., 2021) based on CIFAR-10, a binary classification task to predict whether the image subject moves primarily by air (plane/bird) or land (car/horse), which is spuriously correlated with whether the image contains an animal (bird/horse) or vehicle (car/plane). Here the majority subpopulation is vehicle domain and the minority subpopulation is animal domain. We select 90% of the air images as "air-vehicle(airplane)", and 90% of the land images as "land-animal(horse)".
- **Rare subpopulation.** Without spurious correlation, ML models can still underperform on subpopulation that occur infrequently in the training set (e.g. patients with a darker skin tone, photos taken at night). Since the rare subpopulation will not significantly affect model loss during training, the model may fail to learn to classify examples within theis subpopulation. To explore the rare subpopulation scenario (without spurious correlation), we adopted PACS (Li et al., 2017), OfficeHome (Venkateswara et al., 2017), Modified-CIFAR4 V2 (Rolf et al., 2021) for the experiments. For each dataset, we select 2 domains independent of the target Y. For example, for PACS, we construct a training set where most of the training images (67%) come from the photo domain, and the rest (33%) come from the cartoon domain, but the domain does not correlate with the class label.
  - **PACS** (Li et al., 2017): An image classification task of 7 classes. Most of the training images (1200 images, 67%) come from the photo domain, and the rest (600 images, 33%) come from the cartoon domain. Here the majority subpopulation is the photo domain, and the minority subpopulation is the cartoon domain.
  - **OfficeHome** (Venkateswara et al., 2017): An object recognition task of 65 classes. Most of the training images (3965 images, 83%) come from the product domain, and the rest (800 images, 17%) come from the clipart domain. Here the majority subpopulation is the product domain, and the minority subpopulation is the clipart domain.
  - **Modified-CIFAR4 V2** (Rolf et al., 2021): A binary classification task on air vs. land. The majority subpopulation (4500 images, 90% in training) is "air-vehicle(airplane)", "land-vehicle(automobile)", and the minority subpopulation (500 images, 10% in training) is "air-animal(bird)", "land-animal(horse)" .

A.2. Configuration space of different ML models

The goal of our paper is to compare the ID and OOD performance accross a wide spectrum of ML models. Therefore, for each subpopulation shifts dataset, we vary the model architectures, training durations, and hyperparameters to explore the performances of different ML models.

Specifically, we follow the search space of AutoGluon, the state-of-the-art commercial AutoML library (AutoGluon), to train 500 different ML models with varying training settings. For each dataset, we implement with 5 model architectures, namely mobilenetv3_small, resnet18_v1b, resnet50_v1, mobilenetv3_large, and resnet101_v2. We set the learning rates to be in the search space of $\{0.0001, 0.0005, 0.001, 0.005, 0.01\}$ and batch sizes in $\{8, 16, 32, 64, 128\}$. The training durations are set to $\{1, 5, 10, 25\}$ epochs, as we did not find significant improvements from training for more epochs for any dataset. Together with the 5 model architectures, 5 learning rates, 5 batch sizes, and 4 training durations, we train 500 different ML models independently with all the combinations of different model architectures and hyperparameters.

A.3. Implementation and Reproducibility

Our implementation framework is based on MXNet (Chen et al., 2015) and AutoGluon (AutoGluon). Following the training configurations in AutoGluon, we adopt the Nesterov accelerated gradient optimizer with momentum=0.9. Other unspecified parameters are set to the AutoGluon defaults. There is no learning rate decay. All ML models are fine-tuned starting from its ImageNet pre-trained checkpoints. Code and data are available at `https://github.com/yining-mao/Moon-Shape-ICML-2023`

A.4. Controlled Experiments on Spurious Correlation

Setup Details The Modified-CIFAR4 (V1) in Figure 8 was created by (Rolf et al., 2021) based on CIFAR-10 by subsetting to the bird, car, horse, and plane classes. This is a binary classification task to predict whether the image subject moves primarily by air (airplane/bird) or land (automobile/horse), which is spuriously correlated with 2 domains where the image contains an animal (bird/horse) or vehicle (automobile/airplane). Here the majority subpopulation is vehicle domain as indicated by $Z = 1$ in Figure 8; the minority subpopulation is animal domain as indicated by $Z = 0$ in Figure 8.

We modulate the degree of *spurious correlations* between the classification target label (air/land) and spurious feature (vehicle/animal) in the training data by changing the *mixture weights* in the training data. The 2×2 table in each panel in Figure 8 indicates the dataset construction procedure. Specifically, we ensure that the dataset is class-balanced: i.e., 5,000 images for both $Y = 0$ and $Y = 1$, and fixed the ratio of spurious feature vehicle $Z = 1$ (60% in training, i.e., 6,000 images) and spurious feature animal $Z = 0$ (40% in training, i.e., 4,000 images). We increase the level of the spurious correlation between the classification target label (air/land) and spurious feature (vehicle/animal) by increasing the number of samples in air-vehicle(airplane)" and "land-animal(horse)" as indicated by the red boxes in Figure 8. Formally, since we fix (1) the total number of data points in the training set $(10,000)$, (2) the ratio of data point number in each class ($P(Y = 1) = 0.5, P(Y = 0) = 0.5$), and (3) the ratio of data point number in each spurious features ($P(Z = 1) = 0.6, P(Z = 0) = 0.4$), there is effectively only one degree of freedom left, which we vary to change the level of spurious correlation.

B. Additional Experimental Results

Training dynamics. We show that the moon shape persists both across and within different training epochs in **Figure 4 (b)**. Results on other datasets are similar as shown in **Supp. Figure 11**.

C. Extended Related Work

Linear correlations between ID and OOD performances Existing research mostly reports *linear* correlations between ID and OOD performances. The linear correlations were first reported in recent dataset reconstruction settings including ImageNet-V2 (Recht et al., 2019), CIFAR-10.1 (Recht et al., 2018), CIFAR-10.2 (Lu et al., 2020), where new test sets of popular benchmarks are collected closely following the original dataset creation process. As there are subtle differences in the dataset creation pipeline, the test performance on the new test set is often lower, but appears to be linearly correlated with the performance on the original test set (Lu et al., 2020; Miller et al., 2020; Recht et al., 2018; 2019; Yadav & Bottou, 2019). Later researchers also found the linear trends in the context of cross-benchmark evaluation (Taori et al., 2020; Miller et al., 2021), and transfer learning (Kornblith et al., 2019; Andreassen et al., 2021), where a model's ImageNet test accuracy linearly correlates with the transfer learning accuracy. Similar linear trends are also observed in sub-type shifts (Hendrycks & Dietterich, 2019; Santurkar et al., 2021) (e.g., the training data for the "dog" class are all from a specific breed while the test data come from another breed). Different from these studies, we (1) focus on subpopulation shifts, where we also present the first systematic study on the performance correlation between data subpopulations, and (2) find *nonlinear* correlations

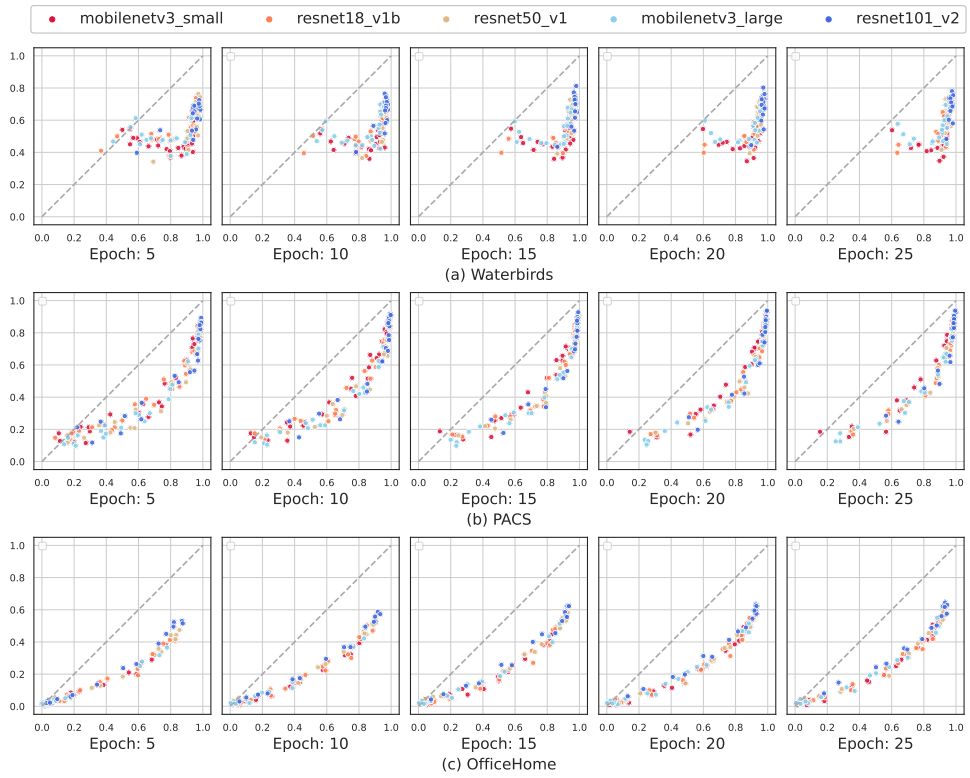

Figure 11: **The moon shape persists across different training epochs. Results on other datasets similar to Figure 4.** We stratify Figure 2 based on the number of training epochs. The x-axis indicates majority subpopulation performance. The y-axis indicates minority subpopulation performance. Most of the models have converged after 10 epochs. The moon shape is apparent in each snapshot and persists across training epochs.

of ML performance across data subpopulations, which is not captured in previous work. Importantly, we show that for datasets with spurious correlations, even with probit-transformed axes as used by several prior work (Miller et al., 2021; Recht et al., 2019; Taori et al., 2020), the performance correlations still remain nonlinear. This confirms that the nonlinear ("moon shape") correlation phenomenon is indeed not captured by previous work.

ML with diverse subpopulations A major challenge in ML is that a model can have very disparate performances even when it's applied to different subpopulations of its training and evaluation data. Models with low average error can still fail on particular groups of data points (Hashimoto et al., 2018; Buolamwini & Gebru, 2018; Blodgett et al., 2016). For example, predictive models for clinical outcomes that are accurate on average in a patient population are reported to underperform drastically for some subpopulations, potentially introducing or reinforcing inequities in care access and quality (Pfohl et al., 2022). Similar performance disparity, have also been observed in radiograph classification (Badgeley et al., 2019; Zech et al., 2018; DeGrave et al., 2021), face recognition (Grother et al., 2011; Buolamwini & Gebru, 2018), speech recognition (Koenecke et al., 2020; Blodgett et al., 2016; Jurgens et al., 2017), academic recommender systems (Sapiezynski et al., 2017)), and automatic video captioning (Tatman, 2017), among others. Worse, as model accuracy affects user retention, the minority group might shrink and thus even amplifies the performance disparity over time (Hashimoto et al., 2018; Fuster et al., 2017). These case studies highlight the importance of understanding the ML performance disparity across subpopulations.

D. Extended Analysis: Simulation Studies

TLDR: a simulation study where we concatenate the real features with spurious features to simulate spurious correlation.

Related to: Moon shape, training dynamics.

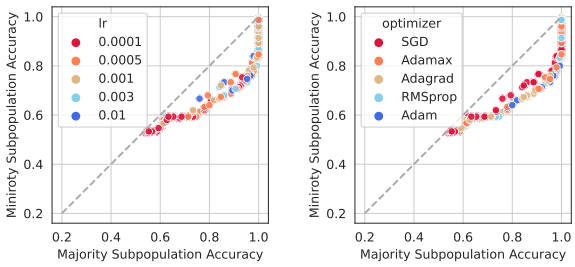

Figure 12: **We present a simple simulation study which successfully replicates the moon shape phenomenon.** Moon-shape trend holds in synthetic dataset and persists across hyper parameters.

Possible figures.

Following the simulation study in this figure: **An Investigation of Why Overparameterization Exacerbates Spurious Correlations** (Sagawa et al., 2020c) "overparameterization—increasing model size well beyond the point of zero training error—can hurt test error on minority groups despite improving average test error when there are spurious correlations in the data."

Simulation Data distribution. We construct a synthetic dataset that replicates the "moon shape" phenomenon in section 3. The label $y \in \{1, -1\}$ is spuriously correlated with a spurious attribute $a \in \{1, -1\}$.

Inspired by (Sagawa et al., 2020c), we divide our training data into two groups: majority group of size n_{maj} with $a = y$, and minority group of size n_{min} with $a = -y$. Both of majority and minority group have their own distribution over input features $x = [x_{\mathsf{core}}, x_{\mathsf{spu}}] \in \mathbb{R}^{d_{\mathsf{core}}+d_{\mathsf{spu}}}$ comprising core features $x_{\mathsf{core}} \in \mathbb{R}^{d_{\mathsf{core}}}$ generated from the label/core attribute y, and spurious features $x_{\mathsf{spu}} \in \mathbb{R}^{d_{\mathsf{spu}}}$ generated from the spurious attribute a, and the core and spurious features are both balanced in each group:

$$\begin{aligned} x_{\mathsf{core}} \mid y &\sim \mathcal{N}(y\mathbf{1}, \sigma_{\mathsf{core}}^2 I_d) \\ x_{\mathsf{spu}} \mid a &\sim \mathcal{N}(a\mathbf{1}, \sigma_{\mathsf{spu}}^2 I_d). \end{aligned} \tag{1}$$

We fix (1) the total number of training points n as 3000, (2) $\sigma_{\mathsf{core}} = 10$, (3) $\sigma_{\mathsf{spu}} = 1$, and (4) the training model as logistic regression model with different hyper parameters to train on the synthetic dataset. We test the performance on the majority group and the minority group separately, and conduct several controlled experiments to study the moon shape phenomenon and the effect of the spurious correlation on synthetic dataset.

Experiments and Results We find the moon shape trend as we observe on real datasets, which indicates the crucial role of the existence of spurious correlation in shaping the moon shape trend. And the moon shape correlation holds across hyperparameters. (Figure 12)

We draw the dots when the models have converged in Figure 13. Different from the real datasets, the dots do not perform the moon shape when the models have converged, but centralize at the top right corner under the synthetic dataset setting since both of majority and minority accuracy have converged to 1. And there are also some dots overlapping with the same majority and minority accuracy, which makes the dots fewer than the number of trials. This shows the difference between the synthetic datasets and the real datasets.
In Figure 14 , we conduct multiple controlled experiments on the synthetic dataset to explore the effect of the level of spurious correlation on moon shape, which is discussed in section 3.2. We define *spurious-core dimension ration* (SDR) as $d_{\mathsf{s:c}} = d_{\mathsf{spu}}/d_{\mathsf{core}}$, and define $n = n_{\mathsf{maj}} + n_{\mathsf{min}}$ as the total number of training points, with $p_{\mathsf{maj}} = n_{\mathsf{maj}}/n$ as the ratio of majority group. Both of SDR and p_{maj} can reflect the level of the spurious correlation. The higher it is, the stronger spurious correlation there is in the train dataset. The results show the increasing of spurious correlation results in the increasing curvature in the moon shape.

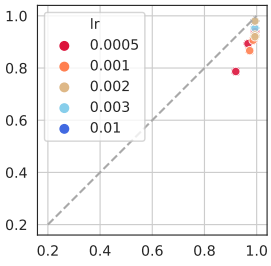

Figure 13: The dots centralize in the top right corner when the models have converged in the synthetic dataset.

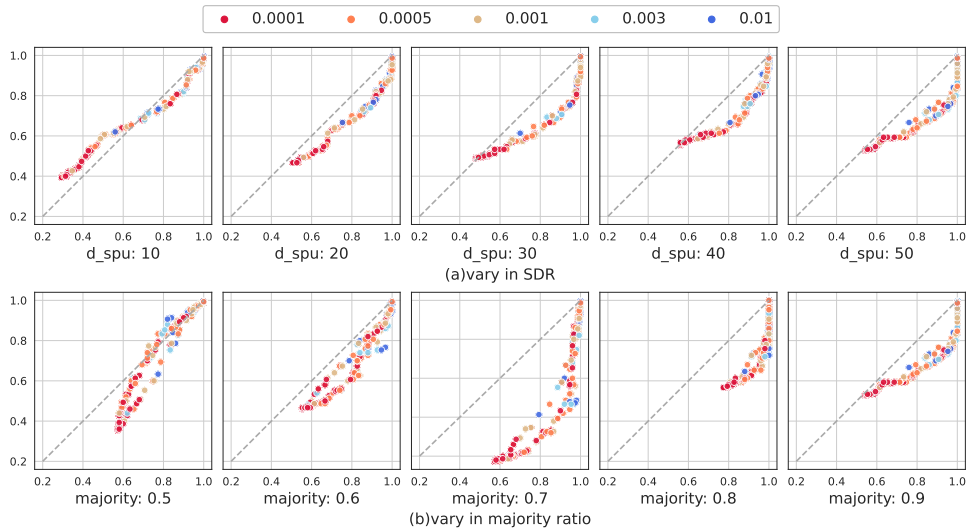

Figure 14: **Stronger spurious correlation creates increasing curvature in the moon shape. Series(a):** d_{core} is fixed as 100, and d_{spu} is varied from 10 to 50, simulating the increasing of SDR. **Series(b):** the ratio of the majority group is varied from 0.5 to 0.9. The figures plotted from left to right simulates the increasing of the spurious correlation.

Summary In this section, we present a simple simulation study which successfully replicates the moon shape phenomenon in section 3. In addition, stronger spurious correlation again leads to more non-linear performance correlations. This indicates that the moon shape phenomenon we found might be a very general phenomenon. However, one distinct difference between the simulation study and our results on the real world datasets is on the training dynamics: On real-world datasets, we found that the moon shape persists within and across different training epochs (Figure 4), while in the simulation study, models converges to the top-right corner (Supp. Figure 4). Our finding on the real-world datasets motivates us to focus our analysis on comparing across *different models* rather than comparing the subpopulation performance of a single model across training epochs. Meanwhile, our simulation study indicates that comparing the subpopulation performance of a single model across training epochs is an interesting direction of future work that is complementary to our scope of our main paper.

E. Theoretical analysis of the accuracy gap across subpopulations

We rigorously study the effect of the spurious correlation on the accuracy gap between the majority and minority subpopulations in a binary classification setting. Our theoretical result shows that the accuracy gap becomes larger when there is a strong spurious correlation between subpopulations and labels and explain why multiple models form the moon shape curve through the ROC curve analysis. This is aligned with the theoretical models in the previous literature that spurious correlations can lead to poor accuracy in minority groups (Sagawa et al., 2020c).

For $d_X \in \mathbb{N}$, we denote an input space by $\mathcal{X} \subseteq \mathbb{R}^{d_X}$ and denote an input and an output random variable by X and Y, respectively. We denote a random variable for a subpopulation by Z, where $Z = 1$ indicates the majority subpopulation[1], otherwise $Z = 0$. We assume that the underlying data generating mechanism is $Z \to Y \to X$. This implies that X and Z are conditionally independent given Y, *i.e.*, $X \perp Z \mid Y$. We suppose a conditional distribution of X given Y is given as follows.

$$X \mid Y = 0 \sim F_0, \quad X \mid Y = 1 \sim F_1,$$

for some arbitrary distributions F_0 and F_1, and we assume

$$\mathbb{P}(Z = 1 \mid Y = 1) = \pi_1, \quad \mathbb{P}(Z = 1 \mid Y = 0) = \pi_0.$$

It is noteworthy that $\pi_1 = \pi_0$ is a necessary and sufficient condition for $Y \perp Z$ because Y and Z are Bernoulli random variables. In this respect, the level of spurious correlation can be expressed as $|\pi_1 - \pi_0|$. With these notations, the subpopulation accuracy gap for a model $g : \mathcal{X} \to \{0, 1\}$ is expressed as follows:

$$\left|\mathbb{E}[\mathbb{1}(Y = g(X)) \mid Z = 1] - \mathbb{E}[\mathbb{1}(Y = g(X)) \mid Z = 0]\right|.$$

In the following theorem, we explicitly show that the subpopulation accuracy gap is proportional to the level of spurious correlation $|\pi_1 - \pi_0|$.

Theorem E.1 (The higher the level of spurious correlation, the larger the accuracy gap)**.** *The subclass accuracy gap for a classifier $g : \mathcal{X} \to \{0, 1\}$ is expressed as follows.*

$$\textit{Accuracy Gap} = \frac{\mathbb{P}(Y = 1)\mathbb{P}(Y = 0)}{\mathbb{P}(Z = 1)\mathbb{P}(Z = 0)} |\pi_1 - \pi_0| \, |\mathrm{TPR} - \mathrm{TNR}|,$$

where TPR *and* TNR *denote the true positive rate $\mathbb{E}(g(X) = 1 \mid Y = 1)$ under F_1 and the true negative rate $\mathbb{E}(g(X) = 0 \mid Y = 0)$ under F_0, respectively.*

Theorem E.1 shows that the subpopulation accuracy gap is expressed as a function of $|\pi_1 - \pi_0|$ and $|\mathrm{TPR} - \mathrm{TNR}|$. A direct consequence is that the accuracy gap gets larger when the level of spurious correlation $|\pi_1 - \pi_0|$ increases. It is possible to keep $\mathbb{P}(Z = 1)$ and $\mathbb{P}(Y = 1)$ as constants while the spurious correlation $|\pi_1 - \pi_0|$ changes. In particular, it occurs when π_1 and π_0 are related as $\pi_1 = (\mathbb{P}(Z = 1) - \mathbb{P}(Y = 0)\pi_0)/\mathbb{P}(Y = 1)$, which captures the setting of Figure 8. Specifically, $\mathbb{P}(Y = 1)$ and $\mathbb{P}(Z = 1)$ are fixed to 0.5 and 0.6 respectively, yet the experimental result shows that the accuracy gap increases once $|\pi_1 - \pi_0|$ increases, which is supported by our theoretical result.

Remark E.2 (Models on the similar ROC curve). Suppose that there is a trained binary classification model and its ROC curve is not a straight line, which is typically the case. We can think of different points on the ROC curve as different models whose predicted probability outputs are only different by constant shifts. Given that a point on the ROC curve is described as $(1 - \mathrm{TNR}, \mathrm{TPR})$, TPR changes nonlinearly with respect to TNR. Hence, the $|\mathrm{TPR} - \mathrm{TNR}|$ changes nonlinearly, and so does the accuracy gap by Theorem E.1. This can provide one explanation for our experimental observations that different models form the moon shape curve.

The setting considered in this remark is admittedly simplified to provide some intuition. In practice, different models (with different architectures and hyperparameters) may not correspond to different points on one ROC curve. However, if the different models do approximately trace out an ROC curve, then the intuition here can apply.

[1]Whether $\mathbb{P}(Z = 1) \geq 1/2$ or not is not critical in our theoretical analysis, but in terms of our notations, one sufficient condition for $\mathbb{P}(Z = 1) \geq 1/2$ is $\pi_1 \geq 1/2$ and $\pi_0 \geq 1/2$.

E.1. Proof of Theorem E.1

Proof of Theorem E.1. For any $z \in \{0, 1\}$, we have

$$\begin{aligned}\mathbb{E}[\mathbb{1}(Y = g(X)) \mid Z = z] &= \sum_{y=0}^{1} \mathbb{E}[\mathbb{1}(y = g(X)) \mid Z = z, Y = y]\mathbb{P}(Y = y \mid Z = z) \\ &= \sum_{y=0}^{1} \mathbb{E}[\mathbb{1}(y = g(X)) \mid Y = y]\mathbb{P}(Y = y \mid Z = z) \\ &= \mathrm{TPR} \times \mathbb{P}(Y = 1 \mid Z = z) + \mathrm{TNR} \times \mathbb{P}(Y = 0 \mid Z = z).\end{aligned}$$

Here, the second equality is due to $X \perp Z \mid Y$. Therefore, the accuracy gap between the two subpopulations is expressed as follows.

$$\begin{aligned}\text{Accuracy Gap} &= |\mathbb{E}[\mathbb{1}(Y = g(X)) \mid Z = 1] - \mathbb{E}[\mathbb{1}(Y = g(X)) \mid Z = 0]| \\ &= \Big|\mathrm{TPR} \times (\mathbb{P}(Y = 1 \mid Z = 1) - \mathbb{P}(Y = 1 \mid Z = 0)) \\ &\quad + \mathrm{TNR} \times (\mathbb{P}(Y = 0 \mid Z = 1) - \mathbb{P}(Y = 0 \mid Z = 0))\Big| \\ &= |\mathbb{P}(Y = 1 \mid Z = 1) - \mathbb{P}(Y = 1 \mid Z = 0)| \times |\mathrm{TPR} - \mathrm{TNR}|\,.\end{aligned}$$

By the Bayes' theorem

$$\mathbb{P}(Y = 1 \mid Z = 1) = \frac{\pi_1 \mathbb{P}(Y = 1)}{\mathbb{P}(Z = 1)}, \quad \mathbb{P}(Y = 1 \mid Z = 0) = \frac{(1 - \pi_1)\mathbb{P}(Y = 1)}{\mathbb{P}(Z = 0)},$$

we have

$$\begin{aligned}\mathbb{P}(Y = 1 \mid Z = 1) - \mathbb{P}(Y = 1 \mid Z = 0) &= \frac{\pi_1 \mathbb{P}(Y = 1)}{\mathbb{P}(Z = 1)} - \frac{(1 - \pi_1)\mathbb{P}(Y = 1)}{\mathbb{P}(Z = 0)} \\ &= \frac{\pi_1 \mathbb{P}(Y = 1)\mathbb{P}(Z = 0) - (1 - \pi_1)\mathbb{P}(Y = 1)\mathbb{P}(Z = 1)}{\mathbb{P}(Z = 1)\mathbb{P}(Z = 0)} \\ &= \frac{\{\mathbb{P}(Z = 0) - (1 - \pi_1)\}\,\mathbb{P}(Y = 1)}{\mathbb{P}(Z = 1)\mathbb{P}(Z = 0)}.\end{aligned}$$

Since $\mathbb{P}(Z = 0) = 1 - (\pi_1 \mathbb{P}(Y = 1) + \pi_0 \mathbb{P}(Y = 0)) = 1 - \pi_1 + (\pi_1 - \pi_0)\mathbb{P}(Y = 0)$, we have

$$\text{Accuracy Gap} = \frac{\mathbb{P}(Y = 1)\mathbb{P}(Y = 0)}{\mathbb{P}(Z = 1)\mathbb{P}(Z = 0)} |\pi_1 - \pi_0| \times |\mathrm{TPR} - \mathrm{TNR}|\,.$$

It concludes a proof. □