Spurious Correlations and Where to Find Them

Gautam Sreekumar [1] **Vishnu Naresh Boddeti** [1]

Abstract

Spurious correlations occur when a model learns unreliable features from the data and are a well-known drawback of data-driven learning. Although there are several algorithms proposed to mitigate it, we are yet to jointly derive the indicators of spurious correlations. As a result, the solutions built upon standalone hypotheses fail to beat simple ERM baselines. We collect some of the commonly studied hypotheses behind the occurrence of spurious correlations and investigate their influence on standard ERM baselines using synthetic datasets generated from causal graphs. Subsequently, we observe patterns connecting these hypotheses and model design choices.

1. Introduction

Spurious correlation is a well-studied problem in machine learning literature and several solutions have been proposed to mitigate it (Arjovsky et al., 2019). Despite these best attempts, empirical risk minimization (ERM) (Vapnik, 1999) remains a strong baseline (Gulrajani & Lopez-Paz, 2020). We believe that the first step in developing robust solutions against spurious correlations is recognizing *when* the models trained using ERM succumb to spurious correlations.

Several factors have been proposed as indicators of spurious correlations. These include overparameterization, partial predictiveness of invariant features, and the amount of data from different environments. However, existing studies about the occurrence of spurious correlation limit their scope to one or a few of these factors. For example, Sagawa et al. (2020) study the effect of overparameterization on underrepresented groups in the training data but do not investigate this phenomenon on easy-to-learn tasks. This was theoretically analyzed by Nagarajan et al. (2020), although they used maximum margin models.

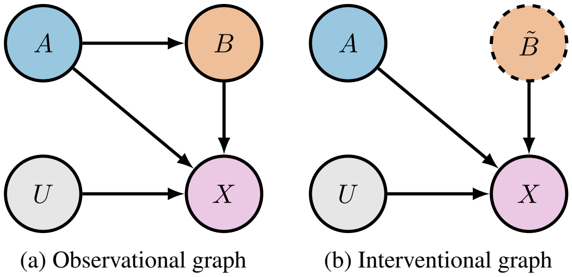


Figure 1. Our objective is to predict the binary variables A and B from the observed data X. Causal modeling of the data generation graph helps to study spurious correlation. (Section 3.1)

However, in practice, these factors may not appear separately. For example, Hill et al. (2021) noted that SARS-CoV-2 datasets were dominated by samples from the US and the UK due to sampling bias, which “may lead to false conclusions about true transmission pathways of virus lineages”. Additionally, the connectedness of air networks is a larger factor in the spread of air-borne diseases than geographical distance (Lemey et al., 2014). Thus, developed countries can introduce both sampling bias and additional confounders to the dataset.

Therefore, the interaction of these factors with the model must be jointly analyzed. Our goal is to investigate how these factors affect the models trained using ERM. To that end, we develop synthetic datasets using causal modeling that allow us to finely adjust the potential factors causing spurious correlation. Knowing the causal model also facilitates relating spurious correlations to dependence relations in the causal graph. We hope that the findings of this paper serve as a guiding principle for the future development of solutions to mitigate spurious correlations.

2. Background

Spurious correlations occur when a model learns correlations from the observed data that do not hold under natural distribution shifts[1]. A robust model is expected to use only invariant features that are reliable during testing. Thus, observed data X is assumed to consist of invariant (or core)

[1] Department of CSE, Michigan State University, East Lansing, MI-48823, USA. Correspondence to: Gautam Sreekumar <sreekum1@msu.edu>.

Published at the ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability. Honolulu, Hawaii, USA. Copyright 2023 by the author(s).

[1] “Natural” distribution shifts are found in passively collected data. For instance, a “STOP” sign with chipped edges. An example of an “unnatural” distribution shift is a green “STOP” sign.

features X_{inv} and spurious features X_{sp}. In practice, spurious features could occur due to biases during data-collection or measurement errors (Fan et al., 2014). Several hypotheses exist about the origin of spurious correlations from a learning perspective.

Partially-predictive invariant features: The most common hypothesis about spurious correlations is that a model relies on spurious features when the invariant features are only partially-predictive of the downstream task or are less useful compared to spurious features (Sagawa et al., 2020). This hypothesis has been further used to show how adversarial attacks exploit a model's dependence on spurious features (Ilyas et al., 2019; Zhang et al., 2021).

Simplicity bias: A related hypothesis is that SGD-trained neural networks prefer to learn simple features (Shah et al., 2020; Valle-Perez et al., 2018). As a result, these models may rely on simpler, but less predictive spurious features instead of sophisticated, yet fully predictive invariant features.

Majority advantage: Another factor that may result in spurious correlations is statistical bias (Nagarajan et al., 2020). If spurious features are present in the majority of the data samples, the model would rely on them to minimize the training error. Mitigation tools derived from this hypothesis usually exploit the diversity in training data to ensure that the model sees enough examples without these spurious features during training. (Arjovsky et al., 2019; Wang et al., 2021; Idrissi et al., 2022).

Other hypotheses that have been studied include noisy invariant features (Khani & Liang, 2020), imperfect partitions of the training data that allow group-specific spurious correlations (Zhou et al., 2021) and invariant features given less weight by the final layers (Kirichenko et al., 2022; Izmailov et al., 2022).

In our experiments, we consider datasets in which the invariant features vary in their predictive power, proportion in the training data, and complexity.

3. Setup

3.1. Causal interpretation of spurious correlation

Spurious correlations are often due to confounders and unobserved correlated variables, and not due to true causal relations. As a result, they are affected by natural distribution shifts. Therefore, causal graphs are a suitable choice to model the occurrence of spurious correlations. By treating prediction tasks as anti-causal learning, we can model the observed data X as $X = f_X(U_X, Y_1, Y_2, \ldots, Y_n)$ where $Y_1, Y_2, \ldots, Y_n$ are label variables and U_X is an exogenous variable denoting unobserved factors that affect X. The label variables may be causally related to each other as $Y_i = f_{Y_i}(\mathbf{Pa}(Y_i), U_{Y_i})$ where $\mathbf{Pa}(Y_i)$ denote the parent variables of Y_i in the causal graph and U_{Y_i} denote the unobserved factors affecting Y_i.

For the model to distinguish spurious features from invariant features, the training set must contain samples where spurious correlations do not hold. In (Arjovsky et al., 2019), samples were collected from different environments to break spurious correlations. Since we model our data-generating process as a causal graph, different environments correspond to different distributions of the label variables. One way to induce different distributions is through interventions on label variables. By intervening on, say, Y_i, it becomes independent of its parent variables $\mathbf{Pa}(Y_i)$. Let F_{Y_i} be the feature learned by a model to predict Y_i. If F_{Y_i} has spurious information, it may have non-zero dependence with some variable $Y_j \in \mathbf{Pa}(Y_i)$ during interventions.

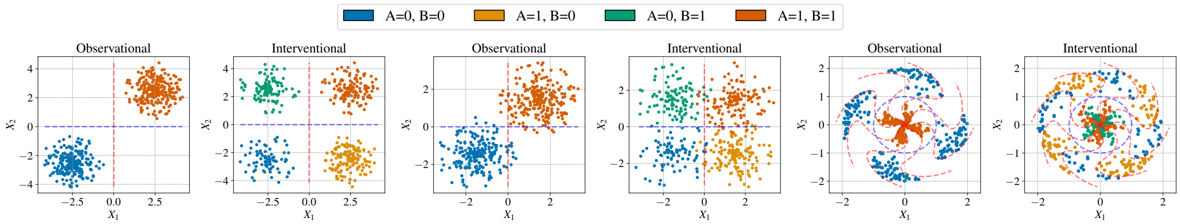
3.2. Datasets

For our study, we consider the simplest causal graph with two binary variables A and B. The causal model of the data-generation process is shown in Figure 1a. Our task is to predict the binary labels from the observed data X. Note that X embodies both the invariant feature X_{inv} and the spurious feature X_{sp}. Although invariant and spurious features are treated as separate casual variables in existing works (Khani & Liang, 2021; Arjovsky et al., 2019), we model them jointly since they may be entangled in practice. The unobserved factors of variation are collectively denoted by U. Since A is a parent of B, a change in distribution of A affects that of B. However, vice versa does not hold. In our experiments, the potential spurious correlation that a model may learn is to use features corresponding to B to predict A.

We construct two synthetic datasets – (1) CIRCLES and (2) WINDMILL. For each dataset, we collect observational and interventional data points following Figure 1a and Figure 1b respectively. The exact functional formulations of the datasets are provided in Appendix A.

CIRCLES dataset: The CIRCLES dataset consists of vectors sampled from four circular regions in the $\mathbb{R}^2$-space (Figure 2a). Each cluster corresponds to $(A = a, B = b)$ for some $a, b \in \{0, 1\}$. A and B decide X_1 and X_2 in Figure 2a respectively. When the clusters are well-separated, a linear model can easily achieve zero test error. However, when they overlap, it is impossible to find a zero-error decision boundary (Figure 2b). We use the CIRCLES dataset to analyze spurious correlations in easy-to-learn and impossible-to-learn tasks.

WINDMILL dataset: Our second dataset is designed to explicitly prompt spurious correlations. The effects of the variables A and B on the observed data X are entangled and the true decision boundary for A is more difficult to



(a) CIRCLES dataset (b) Impossible CIRCLES dataset (c) WINDMILL dataset

Figure 2. Synthetic datasets constructed to study the occurrence of spurious correlations. Red and blue dotted lines indicate the optimal decision boundaries of A and B respectively.

learn than that of B[2]. The complexity of the true decision boundary of A is adjusted through a parameter λ_{off}. Higher the value of λ_{off}, the higher the complexity is. As predicting A using its invariant features alone becomes more difficult, the model tends to use spurious features. Observational and interventional data samples from WINDMILL dataset are illustrated in Figure 2c.

3.3. Why synthetic data?

The key advantage of synthetic datasets over natural ones is the ability to adjust their properties precisely. Specifically, the WINDMILL dataset allows us to change the complexity of its decision boundaries, and the CIRCLES dataset allows us to alter the geometric coordinates and dimensions of the clusters. Synthetic datasets have been used in prior works to analyze the evolution of features learned by models (Hermann & Lampinen, 2020).

4. Experiments

We design experiments to study the impact of task difficulty, statistical bias, and predictive power of invariant features. We first distinguish tasks based on their difficulty – easy-to-learn, difficult-to-learn, and impossible-to-learn. We then vary the amount of interventional data in each dataset. Then for each such dataset, we vary the capacities of the models by changing the depth and the width of the MLPs.

Method: We do not propose any novel method to mitigate spurious correlation. Our objective is to unify possible factors that contribute to spurious correlation in models trained using ERM. We consider two commonly followed sub-paradigms under ERM – (1) standard ERM (simply referred to as ERM), (2) ERM with resampling (ERM-Resampled). A single training batch in ERM comprises both observational and interventional samples. In contrast, observational and interventional samples never appear in the same batch in ERM-Resampled. Existing works (Idrissi et al., 2022; Gulrajani & Lopez-Paz, 2020) indicate that ERM-Resampled is a strong baseline against spurious correlation and in domain generalization. Throughout our experiments, we use MLPs with ReLU as the activation function.

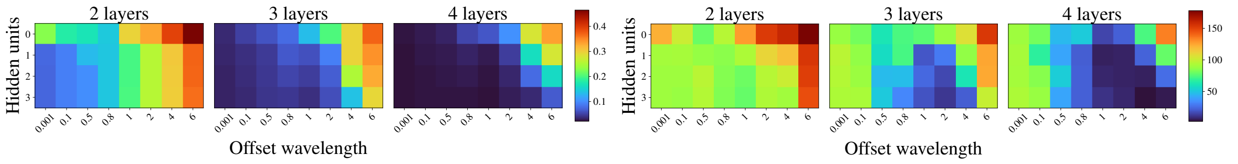
Metrics: The standard sign of spurious correlations is a drop in test accuracy due to a change in distribution. Therefore, we quantify spurious correlations in the model using the relative drop in test accuracy between observational and interventional samples. Additionally, since we know that the intervened variable must be independent of its parents, we evaluate the robustness of the features by measuring the dependence between the features on interventional samples.

Measuring dependence: Several methods have been proposed to measure dependence between high-dimensional vectors, of which kernel-based methods are popular (Gretton et al., 2005; Bach & Jordan, 2002). However, these are difficult to interpret from their absolute values. Therefore, we devise a new independence measure based on statistical independence testing called "ratio over independent samples" (RoIS). Given two features F_A and F_B with N samples, RoIS is measured as $\text{RoIS}(F_A, F_B) = \frac{\text{dep}(F_A,F_B)}{\frac{1}{K}\sum_{i=1}^{K}\text{dep}\left(F_A^{(\pi_i)},F_B\right)}$, where π_i is some permutation of N samples and dep is our choice of measure of dependence. In our experiments, we use a normalized version of HSIC (Gretton et al., 2005) as dep. Refer to Appendix B for a detailed description.

4.1. Easy-to-learn tasks

We construct easy-to-learn tasks similar to those proposed in (Nagarajan et al., 2020) using our CIRCLES dataset. The true decision boundaries for A and B are linear in all cases, *i.e.*, a single non-trivial parameter is sufficient to learn the true boundary for each label. We train MLPs for each task through ERM-Resampled and measure the relative drop accuracy between observational and interventional data during testing.

[2] "Difficulty to learn" is measured in terms of the minimum degree required by a polynomial to approximate it with zero test error.



(a) Relative drop in test accuracy

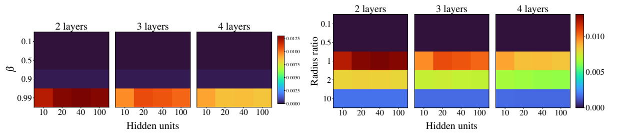
(b) Dependence between features RoIS(F_A, F_B)

Figure 3. Effect of task complexity on ERM-Resampled models for difficult-to-learn tasks

4.1.1. Effect of amount of interventional data

Each dataset contains N samples: βN observational and $(1-\beta)N$ interventional, where $0 < \beta < 1$. Varying β creates what is referred to as "statistical skew" in (Nagarajan et al., 2020).

Analysis: We trained models using the ERM-Resampled scheme. Surprisingly, we found that there was no drop in accuracy due to a variation of β. Even with just 1% interventional data, the model learned the true decision boundary. We attribute this to ERM-Resampled being competitively robust to spurious correlations. We do not report these results since they are trivial. Instead, we consider standard ERM, which is a weaker baseline. We observe that even a weak baseline like ERM is robust to spurious correlations until β reaches around 0.99. Figure 4a visualizes the relative accuracy drop for models trained using standard ERM.



(a) Due to change in β

(b) Due to change in radius ratio

Figure 4. Relative drop in test accuracy of ERM models in easy-to-learn tasks

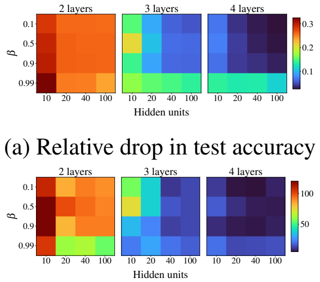
4.1.2. Effect of geometric distortion

We now consider the effect of the cluster shapes on spurious correlation. We adjust the ratio between the radii along the horizontal and vertical dimensions axis in our CIRCLES dataset and observe the occurrence of spurious correlation. The ratio can be written as $\frac{r_B}{r_A}$ where r_A and r_B are the radii along horizontal and vertical directions respectively. Changing the radius ratio essentially shears each circular cluster into an ellipse. Since we found ERM susceptible to spurious correlation only at larger values of β, we set $\beta = 0.99$ during this experiment.

Analysis: Figure 4b shows the drop in test accuracy due to the change in radius ratio. We observe the drop in test accuracy is (1) very small ($< 1\%$), and (2) limited to radius ratio ≥ 1, *i.e.*, only when the circular clusters are vertically sheared. This could be due to horizontal shearing limiting the number of decision boundaries that achieve zero training error, especially when the number of interventional samples is low. Furthermore, as the shearing increased along the vertical dimension, the drop in test accuracy decreased. Refer to Appendix D.1 for the visualization of corresponding decision boundaries.

4.2. Difficult-to-learn tasks

We now design a task where spurious features are easier to be learned compared to invariant features. The phenomenon is commonly referred to as "simplicity bias". To investigate the occurrence of spurious correlation due to simplicity bias, we design tasks where the model has to learn a complex, but zero-test error decision boundary from the training data. Here, the "complexity" of a decision boundary can be roughly defined as the minimum degree required by a polynomial to fully approximate it. Figure 7 illustrates different complexity levels of WINDMILL dataset due to change in λ_{off} in Appendix A.2.



(a) Relative drop in test accuracy

(b) Dependence between features RoIS(F_A, F_B)

Figure 5. Effect of amount of interventional data on ERM-Resampled models for difficult-to-learn tasks

4.2.1. Effect of variation of task difficulty

Our WINDMILL dataset allows us to adjust the complexity of the decision boundary of A by varying λ_{off}. We hypothesize that a model with fixed capacity will increasingly depend on spurious features to make predictions as the complexity of the task increases.

Analysis: In Figure 3a, we vary the complexity of the task via λ_{off}, and calculate the relative drop in accuracy between observational and interventional data for a model trained using ERM-Resampled. We make the following observations: (1) For a fixed task difficulty, the model becomes more robust as the capacity of the model increases, (2) With a fixed capacity, the model tends to learn spurious correlations as the task complexity increases.

For additional analysis, we measure the dependency between the features using our proposed RoIS score. A lower RoIS score indicates that learned representations contain fewer spurious features. Plotting RoIS values in Figure 3b allows us to make two observations: (1) Spurious correlation is always accompanied by a strong dependency between the features, (2) However, strong dependency between the features does not imply spurious correlations, as evident when the task complexity is low.

4.2.2. Effect of amount of interventional data

Similar to Section 4.1.1, we vary the amount of interventional data in difficult-to-learn tasks. Following the same convention, β denotes the proportion of observational data points. We fix the task complexity by setting $\lambda_{\text{off}} = 2$.

Analysis: Figure 5 visualizes the relative drop in test accuracy and RoIS between the features due to change in β. We make the following observations: (1) As expected, spurious correlations reduce with an increase in the amount of interventional data, irrespective of the model capacity, (2) For a given amount of interventional data, larger models seem to be robust against spurious correlations, (3) The variation in dependency seems to be indicative of spurious correlation, unlike in the previous case of varying task complexity.

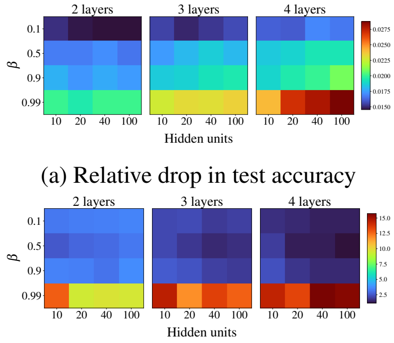
4.3. Impossible-to-learn

Our final experiment tests the hypothesis about the predictive power of invariant features. We modify the Circles dataset such that the circular regions partially overlap (Figure 2b). Due to this overlap, invariant features will no longer be able to provide a zero-error decision boundary, while spurious features can during training. Such situations may occur due to high noise in the invariant features.

4.3.1. Effect of amount of interventional data

Similar to our previous experiments, we vary β and compare the relative drop in test accuracy for various models.

Analysis: When it is impossible to learn a zero-test error decision boundary, the behavior of the model can vary due to the data it sees, and is especially consequential when the interventional samples are too few. This is evident in Figure 6 and allows us to make two interesting observations: (1) The model becomes more robust as more interventional data is available, (2) In contrast to difficult-to-learn tasks, larger models are more prone to spurious correlation especially when the amount of interventional data is limited, (3) Dependency between features seems to be positively correlated with accuracy drop for lower amounts of interventional while being negatively correlated everywhere else.



(a) Relative drop in test accuracy

(b) Dependence between features RoIS(F_A, F_B)

Figure 6. Effect of amount of interventional data on ERM-Resampled models for impossible-to-learn tasks

5. Concluding Remarks

Our work attempted to unify some of the commonly pursued hypotheses behind spurious correlations – statistical bias, simplicity bias, and predictive power of the invariant features. We designed synthetic datasets that reflected these qualities following a simple causal graph. We measured the drop in test accuracy to quantify the spurious correlations learned by models trained using variants of ERM. Furthermore, our causal formulation allowed us to measure the dependence between features from interventional data.

Some of our findings were surprising. We found that SGD-based solutions were robust against spurious correlations in easy-to-learn tasks, unlike maximum margin solutions (Nagarajan et al., 2020). We also found that larger models were more robust than smaller models in difficult-to-learn tasks, while smaller models were more robust in impossible prediction tasks. However, other findings were similar to those reported previously. We noted that having more interventional data improved the robustness of the model. Additionally, we showed that models learn shortcuts from data (Geirhos et al., 2020) when the task is disproportionately difficult for the capacity of the model.

Our findings indicate that spurious correlations are a product of the dataset, the model, and the training scheme. Therefore, future algorithms that are proposed to mitigate spurious correlations must evaluate using models with different capacities with varying proportions of interventional points. In addition to drop in test accuracy, independence relations between known causal variables can provide a deeper understanding of the algorithms.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bach, F. R. and Jordan, M. I. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul): 1–48, 2002.

Fan, J., Han, F., and Liu, H. Challenges of big data analysis. *National science review*, 1(2):293–314, 2014.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory: 16th International Conference, ALT 2005, Singapore, October 8-11, 2005. Proceedings 16*, pp. 63–77. Springer, 2005.

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

Hermann, K. and Lampinen, A. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems*, 33: 9995–10006, 2020.

Hill, V., Ruis, C., Bajaj, S., Pybus, O. G., and Kraemer, M. U. Progress and challenges in virus genomic epidemiology. *Trends in Parasitology*, 37(12):1038–1049, 2021.

Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pp. 336–351. PMLR, 2022.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Izmailov, P., Kirichenko, P., Gruver, N., and Wilson, A. G. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532, 2022.

Khani, F. and Liang, P. Feature noise induces loss discrepancy across groups. In *International Conference on Machine Learning*, pp. 5209–5219. PMLR, 2020.

Khani, F. and Liang, P. Removing spurious features can hurt accuracy and affect groups disproportionately. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 196–205, 2021.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., Russell, C. A., Smith, D. J., Pybus, O. G., Brockmann, D., et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza h3n2. *PLoS pathogens*, 10(2): e1003932, 2014.

Nagarajan, V., Andreassen, A., and Neyshabur, B. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33: 9573–9585, 2020.

Valle-Perez, G., Camargo, C. Q., and Louis, A. A. Deep learning generalizes because the parameter-function map is biased towards simple functions. *arXiv preprint arXiv:1805.08522*, 2018.

Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 1999.

Wang, T., Sridhar, R., Yang, D., and Wang, X. Identifying and mitigating spurious correlations for improving robustness in nlp models. *arXiv preprint arXiv:2110.07736*, 2021.

Zhang, Y., Gong, M., Liu, T., Niu, G., Tian, X., Han, B., Schölkopf, B., and Zhang, K. Causaladv: Adversarial robustness through the lens of causality. *arXiv preprint arXiv:2106.06196*, 2021.

Zhou, C., Ma, X., Michel, P., and Neubig, G. Examining and combating spurious features under distribution shift. In *International Conference on Machine Learning*, pp. 12857–12867. PMLR, 2021.

A. Dataset Generation

Notation: $\text{Bern}(p)$ denotes a Bernoulli distribution with parameter p. $\mathcal{U}(a, b)$ denote a uniform distribution between a and b. $\mathcal{C}(S)$ denote uniform categorical distribution over the elements of set S. $\mathcal{B}(\alpha, \beta)$ denotes a beta distribution with parameters α and β. Variables in uppercase denote random variables and those in lowercase denote scalar constants.

The structural causal model (SCM) (Peters et al., 2017) corresponding to the observational causal graph shown in Figure 1a is as follows. f_X and U vary with the dataset.

$$
\begin{aligned}
A &\sim \text{Bern}(p) \\
B &= A && \text{(During observation, or)} \\
B &= \tilde{B} && \text{(During intervention)} \\
X &= f_X(A, B, U)
\end{aligned}
$$

A.1. CIRCLES dataset generation

The parameters are:

Parameter	Description	Default value	
		Easy	Impossible
$r_1^{(\text{max})}$	Maximum radius along X_1 direction	2	2
$r_2^{(\text{max})}$	Maximum radius along X_2 direction	2	2
μ_1	Shift in center of ellipse along X_1 direction	2.5	1.5
μ_2	Shift in center of ellipse along X_2 direction	2.5	1.5

Table 1. Parameters used for generating the CIRCLES dataset, what they mean, and their default values if applicable.

$$
\begin{aligned}
\Theta &\sim \mathcal{U}(0, 2\pi) && \text{(Sample polar angle)} \\
R_1 &\sim \mathcal{U}(0, r_1^{(\text{max})}) && \text{(Sample polar distance along } X_1 \text{ direction)} \\
R_2 &\sim \mathcal{U}(0, r_2^{(\text{max})}) && \text{(Sample polar distance along } X_2 \text{ direction)} \\
R &= \frac{R_1 R_2}{\sqrt{R_1^2 \cos^2 \Theta + R_2^2 \sin^2 \Theta}} && \text{(Polar form of an ellipse)} \\
X_1 &= (2A - 1)\mu_1 + R \cos \Theta && \text{(Shift according to value of } A\text{)} \\
X_2 &= (2B - 1)\mu_2 + R \sin \Theta && \text{(Shift according to value of } B\text{)} \\
X &= \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}
\end{aligned}
$$

A.2. WINDMILL dataset generation

The parameters are:

Parameter	Description	Default value
n_{arms}	Number of ”arms” in WINDMILL dataset	4
r_{max}	Radius of the circular region spanned by the observed data	2
θ_{wid}	Angular width of each arm	$\frac{0.9\pi}{n_{\text{arms}}} = 0.7068$
λ_{off}	Offset wavelength. Determines the complexity of the dataset	-
$\theta_{\text{max-off}}$	Maximum offset for the angle	$\pi/6$

Table 2. Parameters used for generating WINDMILL dataset, what they mean, and their default values if applicable.

$$
\begin{aligned}
R_B &\sim \mathcal{B}(1, 2.5) && \text{(Sample radius)}\\
R &= \frac{r_{\text{max}}}{2}\left(BR_B + (1-B)(2-R_B)\right) && \text{(Modify sampled radius based on } B\text{)}\\
\Theta_A &\sim \mathcal{C}\left(\left\{2\pi\frac{i}{n_{\text{arms}}+1} : i = 0, \ldots, n_{\text{arms}}-1\right\}\right) && \text{(Choose an arm)}\\
U &\sim \mathcal{U}(0,1) && \text{(To choose a random angle)}\\
\Theta_{\text{off}} &= \theta_{\text{max-off}} \sin\left(\pi\lambda_{\text{off}}\frac{R}{r_{\text{max}}}\right) && \text{(Calculate radial offset for the angle)}\\
\Theta &= \theta_{\text{wid}}\left(U-0.5\right) + A\left(\Theta_A + \frac{\pi}{n_{\text{arms}}}\right) + (1-A)\Theta_A + \Theta_{\text{off}} && \text{(Angle is decided by } A \text{ and the radial offset)}\\
X_1 &= R\cos\Theta && \text{(Convert to Cartesian coordinates)}\\
X_2 &= R\sin\Theta\\
X &= \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}
\end{aligned}
$$

Figure 7 shows the various datasets generated by varying λ_{off}.

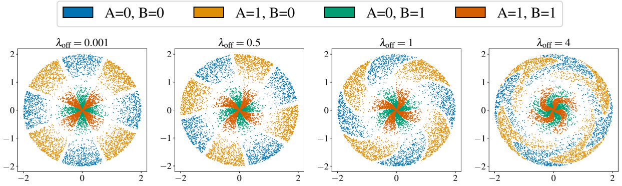


Figure 7. Difference in dataset generated by varying λ_{off}

B. Quantifying dependence between features

Suppose we wish to measure the dependence between two random vectors X and Y using some kernel-based measure of dependence $D(X, Y)$. Let X and Y consist of n samples and $\{\mathcal{P}_i : i = 1, \ldots m\}$ be m permutations of these samples. We can safely assume that $X \perp\!\!\!\perp Y^{(\mathcal{P}_i)}$ for all $\mathcal{P}_i$. Therefore, the average score from the measure of dependence for independent samples can be written as,

$$d^* = \frac{\sum_i^m D(X, Y^{\mathcal{P}_i})}{m} \tag{1}$$

d^* can be interpreted as the highest value of the measure of dependence D that it may give for any pair of independent random vectors. We use d^* to define our metric "ratio over independent samples" (RoIS) as,

$$\text{RoIS}(X, Y) = \frac{D(X, Y)}{d^* + \delta} \tag{2}$$

where δ adjusts the smoothness of our metric. For our experiments, we use a normalized version of HSIC (Gretton et al., 2005) in place of D, denoted by NHSIC.

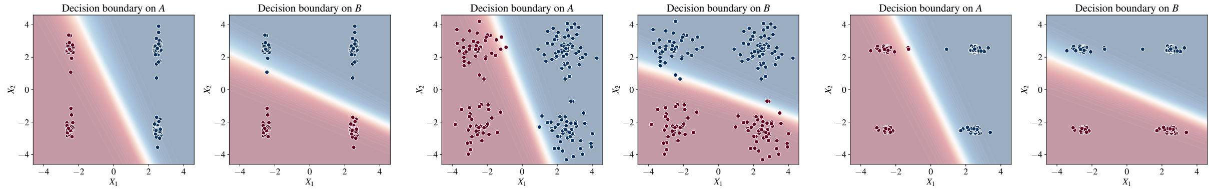
$$\text{NHSIC}(X, Y) = \frac{\text{HSIC}(X, Y)}{\sqrt{\text{HSIC}(X, X)\text{HSIC}(Y, Y)}} \tag{3}$$

C. Connecting spurious correlations through causal graphs

Consider the causal graph provided in Figure 1. Due to the intervention on B, it becomes independent of its parent – A. On the other hand, any change in A must affect B. Therefore, a robust model cannot use features corresponding to B to predict A. However, a bad model may be tempted to do so if, say, the features corresponding to A are difficult to learn. Hence, a bad model is prone to show a drop in validation accuracy in predicting A during interventions.

D. Additional Results

D.1. Decision boundaries for easy-to-learn tasks



(a)

(b)

(c)

Figure 8. Decision boundaries learned for different radius ratios