LIMITS OF ALGORITHMIC STABILITY FOR DISTRIBUTIONAL GENERALIZATION

Anonymous authors

Paper under double-blind review

ABSTRACT

As machine learning models become widely considered in safety critical settings, it is important to understand when models may fail after deployment. One cause of model failure is distribution shift, where the training and test data distributions differ. In this paper we investigate the benefits of training models using methods which are algorithmically stable towards improving model robustness, motivated by recent theoretical developments which show a connection between the two. We use techniques from differentially private stochastic gradient descent (DP-SGD) to control the level of algorithmic stability during training. We compare the performance of algorithmically stable training procedures to stochastic gradient descent (SGD) across a variety of possible distribution shifts - specifically covariate, label, and subpopulation shifts. We find that models trained with algorithmically stable procedures result in models with consistently lower generalization gap across various types of shifts and shift severities. as well as a higher absolute test performance in label shift. Finally, we demonstrate that there is there is a tradeoff between distributional robustness, stability, and performance.

1 INTRODUCTION

As machine learning (ML) is applied in several high-stakes decision making situations such as healthcare (Ghassemi et al., 2017; Rajkomar et al., 2018; Zhang et al., 2021a) and lending (Liu et al., 2018; Weber et al., 2020), it is important to consider scenarios when models fail. Typically, models are trained with empirical risk minimization (ERM), which assumes that the training and test data are sampled i.i.d from the same underlying distribution (Vapnik, 1999). Unfortunately, this assumption means that ERM is susceptible to performance degradation under distribution shift (Nagarajan et al., 2021). Distribution shift occurs when the data distribution encountered during deployment is different, or changes over time while the model is used. In practice, even subtle shifts can significantly affect model performance (Rabanser et al., 2019). Given that distribution shift is a significant source of model failure, there has been much work directed toward improving model robustness to distribution shifts (Taori et al., 2020; Cohen et al., 2019; Engstrom et al., 2019; Geirhos et al., 2018; Zhang et al., 2019; Zhang, 2019).

One concept recently introduced to improve model robustness is *distributional generalization* (Kulynych et al., 2022; Nakkiran and Bansal, 2020; Kulynych et al., 2020). Distributional generalization (DG) extends classical generalization to encompass any evaluation function (instead of just the loss objective) and allows the train and test distributions to differ. Kulynych et al. (2022) prove that algorithms which satisfy total variation stability (TV stability) bound the gap between train and test metrics when distribution shift is present, i.e., algorithms which satisfy TV stability are also satisfy DG. This motivates the use of techniques from differentially private (DP) learning to satisfy DG, since DP implies TV stability (Kulynych et al., 2022). We know from other works that DP learning often comes at a cost to accuracy (Bagdasaryan et al., 2019; Suriyakumar et al., 2021; Jayaraman and Evans, 2019). Unfortunately these works don't provide any guidance on practical tradeoffs with robustness, except for a small positive result in Suriyakumar et al. (2021). Because robustness to new settings is an important question for real deployments of models, it is important to understand how distributional robustness theory will work practically in different types and severities of shifts. Furthermore, current theory provides no insight into the relationship between the level of stability and other model performance metrics, such as accuracy.

In this paper we conduct an extensive empirical study on the impact of using algorithmically stable learning strategies for robustness when facing distribution shift. Stable learning (SL) refers to approaches that constrain the model optimization objective or learning algorithm to improve model stability. We focus on three questions regarding the use of SL for DG in practice: **(i)** How much stability is required to see improvements over ERM? **(ii)** Are SL trained models robust across all hyperparameter settings and as the severity of the shift increases? **(iii)** What tradeoffs between distributional robustness and model performance exist? We target three common examples of shift: covariate (Shimodaira, 2000), label (Lipton et al., 2018; Storkey, 2009), and subpopulation shifts (Duchi and Namkoong, 2021; Koh et al., 2021). We test both synthetic and natural shifts using state of the art models and large benchmark datasets focusing on realistic prediction tasks in object recognition, satellite imaging, biomedical imaging, and clinical notes (see Table 1, with details in Section 4.2). The primary comparison we make is through the generalization gap, defined as the difference in model performance between training and testing (Zhang et al., 2021b). Under extensive experimentation–incorporating 32 distinct types of distribution shift and 5 severity levels–we find:

1. Stable learning consistently satisfies distribution generalization, with lower generalization gaps compared to ERM.
2. This distribution generalization holds true for different levels of stability across common distribution shifts of varying type and severity, as well as across hyperparameters.
3. Using stability for robustness does not guarantee high performance for covariate shift.
4. In label shift, with the right level of stability, SL results in improvements to both accuracy and robustness.
5. In our natural subpopulation shifts, we find that SL also results in improvements to both accuracy and robustness.

We conclude our paper in Section 7 where we elaborate on the insights our empirical investigation provides and the potential implications for future studies on designing algorithms which are distributionally robust and accurate.

2 Related Work

Many approaches have been developed in pursuit of robustness to distribution shift, including: domain adaption (Wang and Deng, 2018), out-of-distribution detection (Yang et al., 2021), adversarial training (Madry et al., 2018; Ilyas et al., 2019), as well as through algorithmic improvements (Sagawa et al., 2019). However, many of these methods do not perform better than ERM (Pfohl et al., 2022) and involve complex implementations, making them difficult to use practically.

Algorithmic stability has also been explored to improve distributional robustness. It is often easier to implement, with simpler methods such as: ℓ_2 regularization (Wibisono et al., 2009), early stopping (Hardt et al., 2016), and differentially private stochastic gradient descent (DP-SGD) (Abadi et al., 2016). Early stopping and ℓ_2 regularization have already been studied for their potential to improve distributional robustness (Sagawa et al., 2019). However, it's difficult to conduct fine-grained analyses into improved robustness with these methods because their stability is not directly controllable. This motivates our use of DP-SGD to investigate the limits of stability for DG since we can control the level of stability by adjusting the noise multiplier σ in DP-SGD. While the algorithmic stability has been explored theoretically in previous works (see Section 3), we explore it empirically in this paper across various synthetic and natural distribution shifts.

3 Background and Notation

We provide an overview of the connections between algorithmic stability, DP, and different forms of generalization in this section. It is well-established that algorithmic stability implies generalization in the traditional ERM setup (Bousquet and Elisseeff, 2002). Additional work has proven that DP implies stability and thus, implies generalization (Bassily et al., 2016; Dwork et al., 2015). In this section we define these concepts and draw connections between them. This is done to clarify the theoretical implication that DP leads to improved distributional robustness.

Notation We assume there is a training dataset $D_{train} = \{(x_i, y_i)\}_{i=1}^n$ of labeled examples such that $D_{train} \sim \mathcal{D}$ and a testing dataset $D_{test} = \{(x_i, y_i)\}_{i=1}^m$ of labeled examples such that $D_{test} \sim \mathcal{D}'$. Given D_{train}, we use a randomized learning algorithm $\mathcal{M}(D_{train})$ to learn parameters $\theta \in \Theta$ of a model relating the datapoints $\{x_i\}$ to their corresponding label $\{y_i\}$.

Now we will describe differential privacy and its links to stability.

Definition 1 (Differential Privacy (Dwork et al., 2006))**.** *Suppose we have two datasets D, D' which have a Hamming distance of 1, then an algorithm $M(D)$ is (ϵ, δ)-differentially private if:*

$$Pr[\mathcal{M}(D) \in \Theta] \leq exp(\varepsilon) Pr[\mathcal{M}(D') \in \Theta] + \delta \tag{1}$$

By definition, DP guarantees privacy by bounding the effect that any individual datapoint has on the output of $\mathcal{M}$. This leads to DP implying strong forms of stability such as TV stability (Bassily et al., 2016) and uniform stability (Wang et al., 2016). Next, we will define these notions of algorithmic stability and show how DP implies them both.

Definition 2 (TV Stability)**.** *Suppose we have two datasets D, D' which have a Hamming distance of 1, then an algorithm $M(D)$ is (δ)-TV stable if:*

$$Pr[\mathcal{M}(D) \in \Theta] \leq Pr[\mathcal{M}(D') \in \Theta] + \delta \tag{2}$$

We can also write this as the an upper bound on the total variation distance between $\mathcal{M}(D)$ and $\mathcal{M}(D')$. Where $d_{TV}(P, Q) = \sup_T |P(T) - Q(T)|$, then TV stability is $d_{TV}(\mathcal{D}, \mathcal{D}') \leq \delta$.

Definition 3 (Uniform Stability (Bousquet and Elisseeff, 2002))**.** *Suppose we have two datasets D, D' which have a Hamming distance of 1, then an algorithm $M(D)$ is (δ)-uniformly stable if:*

$$\underset{s.\,t. \|D - D'\| = 1}{\forall D, D'} |\mathbb{E}\ell(D; \mathcal{M}(D)) - \mathbb{E}\ell(D; \mathcal{M}(D'))| \leq \delta \tag{3}$$

If the loss function ℓ is bounded between [0,1] then TV stability implies δ-uniform stability (Kulynych et al., 2022). All of these definitions assume we are sampling D and D' from the same underlying distribution. When distribution shift ocurrs, this is no longer true. Thus, we will present the results of Kulynych et al. (2022) who demonstrate that TV-stable algorithms satisfy a notion of generalization that captures distribution shift known as *distributional generalization*.

Definition 4 (Distributional Generalization (Kulynych et al., 2022; Nakkiran and Bansal, 2020; Kulynych et al., 2020))**.** *Given two datasets D and D' sampled from two different distributions P and Q, an algorithm $\mathcal{M}(D)$ satisfies δ-distributional generalization (DG) if for all ϕ: $D \times \Theta \rightarrow [0, 1]$*

$$|\mathop{\mathbb{E}}_{D \sim P} \phi(D; \mathcal{M}(D)) - \mathop{\mathbb{E}}_{D \sim P, D' \sim Q} \phi(D'; \mathcal{M}(D))| \leq \delta \tag{4}$$

Kulynych et al. (2022) prove that any algorithm which is δ-TV stable is δ-DG. Thus implying that algorithmic stability improves robustness to distribution shift. The level of TV stability and DG which are parameterized by δ are directly correlated with level of noise σ we use in DP-SGD. Throughout the rest of the paper, the larger σ is the more stable the algoirhtm is (i.e. the lower δ is).

4 METHODS

To better understand the potential and limits for using algorithmic stability to improve model robustness we conduct a thorough empirical study across several datasets, types of shifts, and shift severities. We explore both synthetic and natural distribution shifts that arise due to differences in the covariate, label, and subpopulation distributions between training and testing (Table 1).

In each of our experiments—given a training dataset $\mathcal{D}_{\text{train}}$ and test dataset $\mathcal{D}_{\text{test}}$ with known distribution shift—we compare the difference in generalization gap (Definition 5) of models with and without algorithmic stability("stable learning" (SL)) and ERM respectively. Characterizing this gap is an important step to determine how well the theoretical guarantees of DG hold in practice.

We aim to provide answers to the following critical questions about practical use of SL for DG, motivating its use beyond theory: **(i)** How much stability (i.e. how much noise σ in DP-SGD) is required to see improvements over ERM? **(ii)** Are SL trained models still robust as the severity of the shift increases and across all hyperparameters? **(iii)** What tradeoffs between DG / distributional robustness and model performance exist?

Dataset	Models	Prediction Task	Performance Metric	Data Ref	Model Ref
CIFAR10-C	end-to-end CNN	Classification	Accuracy	Hendrycks and Dietterich (2019)	Tramèr and Boneh (2021)
Imbalanced CIFAR10	end-to-end CNN	Classification	Accuracy	Krizhevsky et al. (2009)	Tramèr and Boneh (2021)
Waterbirds	LR	Classification	Worst-Group Accuracy	Sagawa et al. (2019)	Cox (1958)
COOS	end-to-end ResNet18	Classification	Accuracy	Lu et al. (2019)	He et al. (2016)
PovertyMap	end-to-end ResNet18	Regression	Pearson Correlation Coefficient	Koh et al. (2021)	He et al. (2016)
MIMIC-III	LR	Mortality	Worst-Group AUC	Johnson et al. (2016)	Cox (1958)

Table 1: The datasets, prediction tasks, and model architectures used throughout the paper to evaluate the relationship between algorithmic stability and distribution shift.

4.1 Distribution Shift

We present and define the three different kinds of distribution shifts that we explore in our empirical study.

Consider a joint distribution $\mathcal{D}$ between features X and labels Y, $\mathcal{D} = X \times Y$, with training data $\mathcal{D}_{\text{train}}$ and testing data $\mathcal{D}_{\text{test}}$. We define *distribution shift* to be any change to the the marginal, conditional, or joint distributions of X and Y between $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, based on prior work (Quionero-Candela et al., 2009; Rabanser et al., 2019; Duchi and Namkoong, 2018). We explain the three types of shift below:

Covariate shift occurs when the features X change between training and testing, but the prediction given a feature stays the same, such that $P(X_{\text{train}}) \neq P(X_{\text{test}})$, $P(Y_{\text{train}}|X_{\text{train}}) = P(Y_{test}|X_{test})$. A natural example of covariate shift arises when datasets of MRI images are recorded by different machines, providing different features but not changing disease prevalence.

Label shift is when the distribution of labels changes, but the features themselves stay the same, such that $P(Y_{\text{train}}) \neq P(Y_{\text{test}})$, $P(X_{\text{train}}|Y_{\text{train}}) = P(X_{\text{test}}|Y_{\text{test}})$. This shift occurs when one class is overrepresented in the dataset, but the underlying feature distribution is conditionally equivalent.

Subpopulation shift is when the distribution of subpopulations (e.g. subgroups defined by the intersection of sex and race) changes. Define Q to be the random variable over pre-defined subpopulations. Subpopulation shift occurs when $P(Q_{train}) \neq P(Q_{test})$. For example, when we move between rural and urban medical centers the proportion of subpopulations changes where minority populations are less represented in rural settings.

4.2 Datasets

We investigate synthetic covariate, label, and subpopulation shifts, as well as naturally occurring instances of covariate and subpopulation shifts. We use the definition of natural and synthetic shift from Taori et al. (2020). Synthetic shifts are those where the data originally is all from the same distribution but is manipulated such that the training and test distributions are different. Meanwhile, natural shifts are those where the training and test distributions are already different. Based on these definitions, we decide not to investigate an explicit natural label shift example as it often manifests as sampling bias, imbalanced data, or subpopulation shift in practice. Specifically, the `MIMIC-III` mortality prediction dataset already has a label shift present as part of its subpopulation shift.

Synthetic datasets: For covariate shift, we use the `CIFAR10-C` dataset, created by shifting CIFAR10's test set with 19 types of algorithmically generated corruptions of 5 different severities for a total of 95 different synthetic covariate shifts (Hendrycks and Dietterich, 2019). For synthetic label shift, we use `Imbalanced-CIFAR10`, which we created by inducing a class imbalance in CIFAR10 to create a shift in $P(Y)$. These shifts were created randomly, where the percentage of of samples in the shifted dataset from the original test dataset was was chosen randomly from $10 - 100\%$. To explore synthetic subpopulation shifts, we use the `Waterbirds` dataset (Sagawa et al., 2019) made up of bird images with synthetic backgrounds.

Natural datasets: Our natural covariate shifts are derived from the Cells-Out-of-Sample (`COOS`) dataset, which consists of mouse cell images of 7 biological classes, with 4 separate test sets of increasing degrees of covariate shift (Lu et al., 2019). We also explore natural subpopulation shift with the `PovertyMap` dataset predicting poverty levels from satellite imagery (Koh et al., 2021), and `MIMIC-III` clinical notes predicting mortality (Johnson et al., 2016).

For more detailed information about the datasets used in this paper, please refer to Appendix A.2.

4.3 Model Training

We train models using DP-SGD (with varying levels of noise and clipping as a way to modify the amount of stability) and ERM for each dataset and type of shift. Each model is trained with early stopping on the validation loss to prevent overfitting. The scale of the hyperparameter search and noise levels we used for determining the best performing models can be found in Appendix A.3. For all experiments, we use the Opacus package (Yousefpour et al., 2021) or Tensorflow Privacy package (McMahan and Andrew, 2018) to implement DP-SGD. Models and data references are given in Table 1. Note that the models used for the CIFAR dataset (Tramèr and Boneh, 2021) in covariate and label shift are small models created for differential privacy. Thus, the reported ERM test accuracy is lower than state-of-the-art performance using larger model architectures such as the ResNet (He et al., 2016) on the same dataset.

4.4 Experimental Procedure

Our empirical investigation covers more than 200 experiments, incorporating 32 distinct forms of distribution shift with varying levels of severity. For each pair of $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ in an experiment, we perform the following procedure:

- Train a set of models $\{M_1, M_2, ...\}$ (performing a hyperparameter search) on $\mathcal{D}_{\text{train}}$ with DP-SGD using different noise multiplier values $\{\sigma_1, \sigma_2, ...\}$ as varying levels of algorithmic stability. We also train ERM (not stable) models as a comparative baseline.
- Test each model in the sets $\{M_1, M_2, ...\}$ on shifted testing datasets $\{\mathcal{D}'_1, \mathcal{D}'_2 ...\}$.
- Measure the generalization gap G, the difference between training and testing accuracy for both SL and ERM models, G_{SL} and G_{ERM}. To measure the improvement SL has over ERM, we report the difference in generalization gap Δ_G as defined in Definition 5 below.

4.5 Evaluation

We design our empirical investigation to answer the three previous questions posed in Section 4 on the limits of using SL for distributional robustness. We address them as follows:

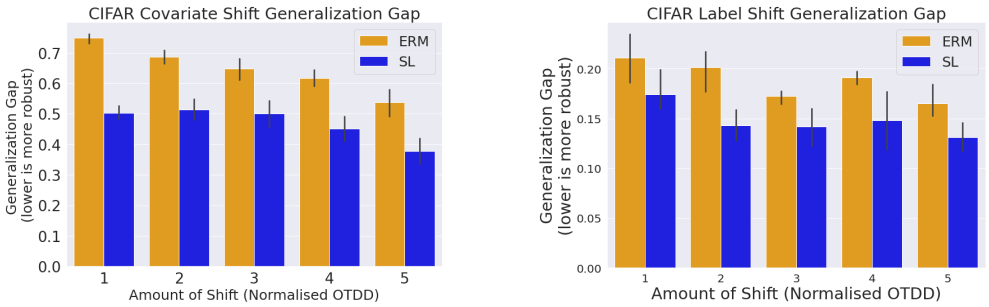
(i) The metric we use to primarily compare the robustness of models is the difference in generalization gap, Δ_G (Definition 5). If $\Delta_G > 0$, this indicates that the model trained with a stable learning algorithm has a lower generalization gap and is therefore more robust than ERM trained models. As mentioned in Section 4.4, we test across various levels of stability to find the optimal σ value.

Definition 5 (Difference in Generalization Gap)**.** *Given the training and D and D' sampled from two different distributions P and Q, an algorithm trained with ERM $\mathcal{M}_{ERM}(D)$ and stability $\mathcal{M}_{SL}(D)$ and metric ϕ: $D \times \Theta \to \mathcal{R}$ we define the generalization gap as:*

$$\Delta_G = \left| \mathop{\mathbb{E}}_{D \sim P} \phi(D; \mathcal{M}_{ERM}(D)) - \mathop{\mathbb{E}}_{D \sim P, D' \sim Q} \phi(D'; \mathcal{M}_{ERM}(D)) \right| - \left| \mathop{\mathbb{E}}_{D \sim P} \phi(D; \mathcal{M}_{SL}(D)) - \mathop{\mathbb{E}}_{D \sim P, D' \sim Q} \phi(D'; \mathcal{M}_{SL}(D)) \right| \tag{5}$$

(ii) We investigate if stability holds over different shift severities using the synthetic datasets by evaluating for which shifts is $\Delta_G > 0$. Shift severity is characterized as the distance between $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$. We use the Optimal Transport Dataset Distance (OTDD) (Alvarez-Melis and Fusi, 2020). We choose this metric as opposed to other dataset distances for its provable guarantees and that it allows for completely disjoint datasets to be compared. We categorize our synthetic datasets by shift severity by first normalizing the computed OTDD and sorting them into quintiles 1-5.

(iii) To explore the tradeoff between robustness and accuracy, we also report model performance throughout the paper and compare it to Δ_G. Model performance is reported as accuracy except for `MIMIC-III` and `PovertyMap`, where area under the curve (AUC) and Pearson Correlation Coefficient are used, respectively (See Table 1).



(a) Covariate Shift

(b) Label Shift

Figure 1: On the `CIFAR10-C` and `Imbalanced-CIFAR` datasets, stable learning (SL) has a lower generalization gap than ERM across all shift types and severities. Furthermore, as the shift increases, the generalization gap drops and $\Delta_G > 0$ across different levels of shift and noise level, both of which indicate greater robustness to distribution shift than ERM.

We illustrate the results of queries (i) and (ii) in Section 5 and (iii) in Section 6. Further information detailing the specific datasets, experiments and, models used can be found in Appendix A.3.

5 Stability Improves Robustness to Distribution Shift

In this section, we focus on validating that the theory proposed by Kulynych et al. (2022) holds up against difficult real-world distribution shifts. We focus on the first two questions: **(i)** How much stability is required to see improvements over ERM? **(ii)** Are SL trained models still robust as the severity of the shift increases and across all hyperparameters? We summarize our findings across the different types of shifts we experiment with.

Synthetic Shifts For the synthetic datasets, we validate for covariate, label, and subopulation shift (Table 3) that stable learning algorithms have a better (i.e. smaller) generalization gap than ERM. We notice this across the range of noise multipliers we use to represent varying levels of stability. We find that even when training a model with small amounts of noise $\sigma = 0.1$ (or a very small amount of TV stability) models are more robust than ERM. It is only when we reduce the stability to be very small (i.e. $\sigma < 0.001$) do we see that ERM is more robust (Table 2 & 8). In Fig. 1, we observe that as the severity of shift increases, SL outperforms ERM at higher margins as evidenced by the increasing trend in Δ_G. Stability is one of the only methods which is robust to severe shifts — other methods, such as distributionally robust optimization, have been shown to perform poorly at these larger shifts (Koh et al., 2021).

Shift Severity	ERM train acc	ERM test acc	DP train acc	DP test acc	Accuracy Gain	$\mathbf{\Delta_G}$
1	0.906 ± 0.032	0.697 ± 0.052	0.896 ± 0.049	0.696 ± 0.061	-0.002 ± 0.018	0.008 ± 0.045
2	0.901 ± 0.046	0.681 ± 0.063	0.909 ± 0.051	0.685 ± 0.068	0.004 ± 0.018	-0.004 ± 0.064
3	0.888 ± 0.042	0.611 ± 0.084	0.899 ± 0.052	0.61 ± 0.093	-0.001 ± 0.02	-0.011 ± 0.08
4	0.871 ± 0.042	0.618 ± 0.075	0.893 ± 0.049	0.61 ± 0.093	-0.008 ± 0.035	-0.029 ± 0.08
5	0.857 ± 0.049	0.526 ± 0.12	0.88 ± 0.039	0.511 ± 0.128	-0.015 ± 0.04	-0.039 ± 0.078

Table 2: For $\sigma = 0.001$, SL has similar results to ERM to covariate shift for `CIFAR10-C`, with $\Delta_G > 0$ not always being true. We really have to reduce the amount of stability to a very small amount for ERM and SL to have similar robustness.

Natural Shifts In our analyses for both natural covariate and subpopulation shifts we validate again that stability improves robustness. In the `COOS` dataset (covariate shift) we observe that for shifts 1-3, SL is more robust, with a lower generalization gap than ERM (Table 4). However, for the most severe shift test dataset 4, ERM is more robust, with a lower generalization gap than SL. This may be because shift severity 4 is a domain generalization problem, with images taken under different

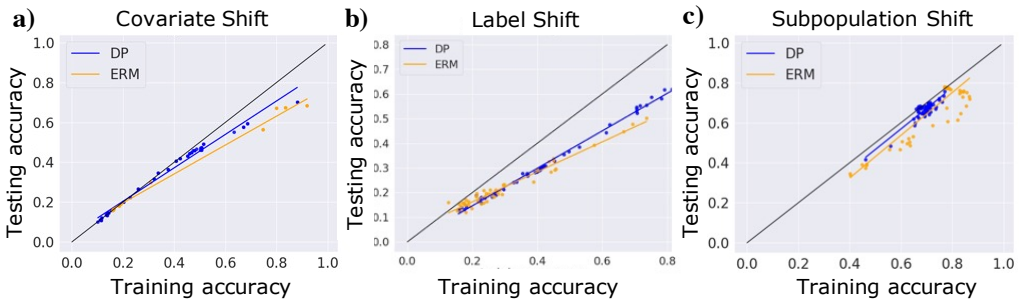


Figure 2: Training vs. testing accuracy across 3 representative examples of a) covariate, b) label, and c) subpopulation shift from the `CIFAR10-C`, `Imbalanced-CIFAR`, and `Waterbirds` datasets, respectively. SL follows the $y = x$ line more closely than ERM, indicating that the generalization gap of SL is lower than ERM and consistent across all hyperparameters.

instrumentation, indicating a high degree of distribution shift. Although the theory of stability is agnostic to distribution distance, this shows empirically that stability has a limit.

Hyperparameter-Agnostic Robustness In our experiments we consistently observed lower generalization gaps across all hyperparameter settings of SL models. This indicates that the robustness improvements provided by SL hold across different model settings, and are not simply a result of well-chosen hyperparameters. In Fig. 2, we examine three representative covariate, label, and subpopulation shifts. We find that the SL models more closely follow the ideal generalization trendline (in black) where the train set performance is equal to the test set performance.

(a) Accuracy and Δ_G of covariate shift across stability levels

(b) Accuracy and Δ_G of label shift across stability levels

Figure 3: Robustness to distribution shift and accuracy are at odds for covariate shift and label shift when we use stable learning for models trained on `CIFAR-C` and `Imbalanced-CIFAR` respectively. This tradeoff worsens as the level of stability is increased.

6 Tradeoffs from Using Stability for Robustness

We show in Section 5 that stability (even small amounts) improves robustness to all three types of shifts and to increasing severity of shifts. We now focus on our third question – does using stability for robustness come at a cost to accuracy? In practice, deployed models must be both accurate and robust. We analyze the results from our experiments and see that the tradeoffs are highly dependent on the type of shift and varies between our synthetic and natural datasets.

6.1 Synthetic Shifts

First, we look at our synthetic experiments seen in Table 3. For covariate shift there is a tradeoff between robustness and accuracy, which gets worse as we increase the level of stability. Thus, suggesting that the reason for this tradeoff between robustness and accuracy is the use of stability to

Dataset	Shift Severity	ERM train acc	ERM test acc	DP train acc	DP test acc	Accuracy Gain	Δ_G
CIFAR-C	None	0.938 ± 0.001	0.771 ± 0.034	0.748 ± 0.165	0.700 ± 0.113	-0.071 ± 0.113	$\mathbf{0.119} \pm 0.057$
	1	0.923 ± 0.025	0.748 ± 0.042	0.731 ± 0.072	0.666 ± 0.029	-0.082 ± 0.037	$\mathbf{0.109} \pm 0.061$
	2	0.919 ± 0.000	0.686 ± 0.050	0.808 ± 0.089	0.646 ± 0.094	-0.041 ± 0.057	$\mathbf{0.071} \pm 0.047$
	3	0.892 ± 0.049	0.649 ± 0.097	0.780 ± 0.139	0.612 ± 0.124	-0.038 ± 0.046	$\mathbf{0.075} \pm 0.127$
	4	0.875 ± 0.043	0.618 ± 0.067	0.783 ± 0.120	0.556 ± 0.104	-0.060 ± 0.059	$\mathbf{0.0322} \pm 0.096$
	5	0.857 ± 0.046	0.538 ± 0.115	0.664 ± 0.144	0.468 ± 0.124	-0.070 ± 0.045	$\mathbf{0.122} \pm 0.111$
Imbalanced-CIFAR	None	0.938 ± 0.001	0.771 ± 0.034	0.748 ± 0.165	0.700 ± 0.113	-0.071 ± 0.113	$\mathbf{0.119} \pm 0.057$
	1	0.776 ± 0.242	0.577 ± 0.174	0.824 ± 0.098	0.63 ± 0.08	$\mathbf{0.086} \pm 0.091$	$\mathbf{0.013} \pm 0.052$
	2	0.863 ± 0.079	0.576 ± 0.066	0.824 ± 0.026	0.667 ± 0.006	$\mathbf{0.091} \pm 0.061$	$\mathbf{0.045} \pm 0.038$
	3	0.734 ± 0.121	0.570 ± 0.118	0.829 ± 0.200	0.665 ± 0.168	$\mathbf{0.095} \pm 0.168$	$\mathbf{0.001} \pm 0.054$
	4	0.682 ± 0.210	0.509 ± 128	0.819 ± 234	0.700 ± 0.217	$\mathbf{0.191} \pm 0.217$	$\mathbf{0.064} \pm 0.080$
	5	0.590 ± 0.057	0.417 ± 0.041	0.819 ± 0.014	0.670 ± 0.014	$\mathbf{0.252} \pm 0.047$	$\mathbf{0.024} \pm 0.020$
Waterbirds	None	0.913 ± 0.002	0.864 ± 0.020	0.852 ± 0.093	0.796 ± 0.123	-0.068 ± 0.089	$\mathbf{0.007} \pm 0.084$
	1	0.817 ± 0.202	0.7784 ± 0.232	0.790 ± 0.215	0.748 ± 0.223	-0.031 ± 0.009	$-\mathbf{0.004} \pm 0.022$
	2	0.900 ± 0.267	0.777 ± 0.282	0.812 ± 0.120	0.749 ± 0.161	-0.028 ± 0.121	$\mathbf{0.054} \pm 0.025$
	3	0.897 ± 0.250	0.773 ± 0.251	0.812 ± 0.109	0.742 ± 0.157	-0.031 ± -0.094	$\mathbf{0.054} \pm 0.048$
	4	0.870 ± 0.197	0.715 ± 0.207	0.813 ± 0.134	0.7331 ± 0.155	0.018 ± -0.052	$\mathbf{0.074} \pm 0.011$
	5	0.937 ± 0.265	0.765 ± 0.200	0.850 ± 0.000	0.731 ± 0.160	-0.034 ± -0.041	$\mathbf{0.053} \pm 0.154$

Table 3: SL improves robustness of models to distribution shift in the CIFAR-C, Imbalanced-CIFAR, and Waterbirds datasets, representing synthetic covariate, label, and subpopulation shifts, respectively. We present results across five shift severities for $\sigma = 0.1$. We observe that for all shifts, $\Delta_G > 0$, indicating that each SL model is more robust than ERM. However, a tradeoff exists between accuracy and robustness among covariate and subpopulation shifts.

achieve robustness. For example, for the most severe shift we see that robustness increases by 5.3% but the performance drops by 3.4%. We note a similar tradeoff for our synthetic subpopulation shift experiments. This tradeoff is not captured by the current theory which connects stability and robustness. We leave this as an important future direction for understanding why satisfying robustness via stability results in much worse accuracy for covariate and subopulation shift.

Next, we look examine the tradeoffs on our synthetic label shift experiments. This tradeoff is much more nuanced than the other ones we will discuss. We find that with only a small amount of stability ($\sigma = 0.1$), both accuracy and robustness are positively correlated Table 3. Once we start to increase the amount of stability (Fig. 3b), we see that they become inversely correlated indicating a tradeoff similar to covariate shift (Fig. 3a).

6.2 Natural Shifts

The tradeoffs across the shifts is mixed when we look at our natural datasets. In COOS (Table 4), there is a tradeoff between accuracy and robustness for the first two shifts, and the tradeoff is worse as the severity of shift increases with a drop in accuracy of around 1-1.2% for an increase robustness of 2.2-2.4% respectively. In the last two shifts we observe two different behaviours. For the shift severity 3, we find that both accuracy and robustness are positively impacted, with positive accuracy gains and Δ_G. Meanwhile, accuracy and robustness are both negatively impacted in shift severity 4.

Shift Severity	SL train acc	SL test acc	ERM train acc	ERM test acc	Accuracy Gain	Δ_G
1	0.872	0.840	0.907	0.852	-0. 012	**0.022**
2	0.872	0.843	0.907	0.853	-0.010	**0.024**
3	0.872	0.810	0.907	0.706	**0.166**	**0.033**
4	0.872	0.467	0.907	0.559	-0.092	**-0.056**

Table 4: SL is more robust to most natural shifts found in COOS except for the most severe. This is most likely due to lower model accuracy. These results are for $\sigma = 0.1$. Standard deviations are not provided because of computational constraints.

Dataset	ERM train perf	ERM test perf	SL train perf	SL test perf	Performance Gain	$\mathbf{\Delta_G}$
MIMIC-III Notes	0.840 ± 0.008	0.814 ± 0.039	0.827 ± 0.000	0.843 ± 0.035	$\mathbf{0.029 \pm 0.004}$	$\mathbf{0.042 \pm 0.008}$
PovertyMap	0.626 ± 0.152	0.489 ± 0.138	0.727 ± 0.085	0.602 ± 0.090	$\mathbf{0.113 \pm 0.046}$	$\mathbf{0.008 \pm 0.004}$

Table 5: SL improves robustness and performance of models in the presence of natural subpopulation shifts. Here we use $\sigma = 0.1$ for `MIMIC-III` and $\sigma = 0.001$ for `PovertyMap`.

Overall, we find that the tradeoff is quite shift and dataset dependent. This indicates again that using stability for robustness does not necessarily guarantee high performing models for covariate shift. Looking at the results for the natural subpopulation shifts we find much different ones from the synthetic subpopulation tradeoffs. For both of our natural experiments we find that both accuracy and robustness are improved when using stability (Table 5).

7 Discussion & Conclusion

In practice it important to understand how model performance degrades when facing arbitrary distribution shifts. In Fig. 4 (Appendix A.4) we present three of the 19 types of covariate shift of increasing severity and demonstrate that the testing accuracy of models trained with SL is more stable as the severity of the shift increases. This graceful degradation shows that algorithm designers may want to incorporate techniques from SL as a starting point so they are confident about their model performance staying stable throughout deployment.

A limitation of our study is that we we only consider datasets with isolated shifts (covariate, label, or subpopulation) except for `PovertyMap` (which is both domain generalization and subpopulation shift). Currently, the theory does not also describe what we should expect when we compose different types of shifts. We believe that future work should explore the impact of SL on robustness to settings where multiple shifts occur together.

While we show that SL adds increased robustness to distribution shift, it is clear that there is a tradeoff between robustness and accuracy – depending on the type of distribution shift we are considering. For covariate shift there is a clear tradeoff. For subpopulation shift the tradeoff is much more mixed, while for label shift, we show that accuracy and robustness are aligned. As such, for a practitioner, it might be useful to use SL as a robustness measure when the type of shift is known, to avoid a loss of performance.

From Fig. 3a and Fig. 3b, we observe that while increasing stability leads to a lower generalization gap, it decreases performance. Furthermore, the amount of stability needed to balance this tradeoff is model and dataset dependent. As such, stability can be treated as a hyperparameter to be tuned for robustness (similar to regularization), rather than a guaranteed solution.

In this work, we perform an empirical evaluation of stability as a technique to encourage robustness to distribution shift. We find this to be especially true where high-capacity models are being trained on large, potentially skewed, datasets. Importantly, a reduction in model generalization gap means that performance numbers reported at training time are more likely to reflect those seen in a deployment. We hope that our work will motivate further research exploring the connection between SL and distribution shift.

REFERENCES

Marzyeh Ghassemi, Mike Wu, Michael C Hughes, Peter Szolovits, and Finale Doshi-Velez. Predicting intervention onset in the icu with switching state space models. *AMIA Summits on Translational Science Proceedings*, 2017:82, 2017.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

Haoran Zhang, Natalie Dullerud, Laleh Seyyed-Kalantari, Quaid Morris, Shalmali Joshi, and Marzyeh Ghassemi. An empirical framework for domain generalization in clinical settings. *Proceedings of the Conference on Health, Inference, and Learning*, 2021a.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

Mark Weber, Mikhail Yurochkin, Sherif Botros, and Vanio Markov. Black loans matter: Distributionally robust fairness for fighting subgroup discrimination. *arXiv preprint arXiv:2012.01193*, 2020.

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

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=fSTD6NFIW_b`.

Stephan Rabanser, Stephan Günnemann, and Zachary Chase Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *NeurIPS*, 2019.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*, pages 1802–1811. PMLR, 2019.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2018.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019.

Bogdan Kulynych, Yao-Yuan Yang, Yaodong Yu, Jarosław Błasiok, and Preetum Nakkiran. What you see is what you get: Distributional generalization for algorithm design in deep learning. *arXiv preprint arXiv:2204.03230*, 2022.

Preetum Nakkiran and Yamini Bansal. Distributional generalization: A new kind of generalization. *arXiv preprint arXiv:2009.08092*, 2020.

Bogdan Kulynych, Mohammad Yaghini, Giovanni Cherubin, and Carmela Troncoso. Disparate vulnerability: on the unfairness of privacy attacks against machine learning. In *22nd Privacy Enhancing Technologies Symposium (PETS 2022)*, volume abs/1906.00389, 2020. URL `http://arxiv.org/abs/1906.00389`.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 32, 2019.

Vinith M Suriyakumar, Nicolas Papernot, Anna Goldenberg, and Marzyeh Ghassemi. Chasing your long tails: Differentially private prediction in health care settings. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 723–734, 2021.

Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 1895–1912, 2019.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

Amos Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28, 2009.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

Shufei Zhang, Zhuang Qian, Kaizhu Huang, Qiufeng Wang, Rui Zhang, and Xinping Yi. Towards better robust generalization with shift consistency regularization. In *ICML*, 2021b.

Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312: 135–153, 2018.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *ArXiv*, abs/2110.11334, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136, 2019.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Stephen R Pfohl, Haoran Zhang, Yizhe Xu, Agata Foryciarz, Marzyeh Ghassemi, and Nigam H Shah. A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. *Scientific reports*, 12(1):1–13, 2022.

Andre Wibisono, Lorenzo Rosasco, and Tomaso Poggio. Sufficient conditions for uniform stability of regularization algorithms. *Computer Science and Artificial Intelligence Laboratory Technical Report, MIT-CSAIL-TR-2009-060*, 2009.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/hardt16.html`.

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1046–1059, 2016.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *Advances in Neural Information Processing Systems*, 28, 2015.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Yu-Xiang Wang, Jing Lei, and Stephen E Fienberg. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of erm principle. *The Journal of Machine Learning Research*, 17(1):6353–6392, 2016.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *CoRR*, abs/1903.12261, 2019. URL `http://arxiv.org/abs/1903.12261`.

Florian Tramèr and Dan Boneh. Differentially private learning needs better features (or much more data). *ArXiv*, abs/2011.11660, 2021.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

Alex X. Lu, Amy X. Lu, Wiebke Schormann, David W. Andrews, and Alan M. Moses. The cells out of sample (COOS) dataset and benchmarks for measuring out-of-sample generalization of image classifiers. *CoRR*, abs/1906.07282, 2019. URL `http://arxiv.org/abs/1906.07282`.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, May 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL `https://www.nature.com/articles/sdata201635`.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Gosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in pytorch. *ArXiv*, abs/2109.12298, 2021.

H. B. McMahan and Galen Andrew. A general approach to adding differential privacy to iterative training procedures. *ArXiv*, abs/1812.06210, 2018.

David Alvarez-Melis and Nicoló Fusi. Geometric dataset distances via optimal transport. *ArXiv*, abs/2002.02923, 2020.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Michael Xie, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. Transfer learning from deep features for remote sensing and poverty mapping. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Barak Oshri, Annie Hu, Peter Adelson, Xiao Chen, Pascaline Dupas, Jeremy Weinstein, Marshall Burke, David Lobell, and Stefano Ermon. Infrastructure quality assessment in africa using satellite imagery and deep learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 616–625, 2018.

Agatha CH de Mattos, Gavin McArdle, and Michela Bertolotto. Mapping slums with medium resolution satellite imagery: a comparative analysis of multi-spectral data and grey-level co-occurrence matrix techniques. *arXiv preprint arXiv:2106.11395*, 2021.

Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, pages 3539–3550, 2018.