An Information-theoretic Approach to Distribution Shifts

Anonymous Author(s)
Affiliation
Address
email

Abstract

Safely deploying machine learning models to the real world is often a challenging process. Models trained with data obtained from a specific geographic location tend to fail when queried with data obtained elsewhere, agents trained in a simulation can struggle to adapt when deployed in the real world or novel environments, and neural networks that are fit to a subset of the population might carry some selection bias into their decision process. In this work, we describe the problem of data shift from an novel information-theoretic perspective by (i) identifying and describing the different sources of error, (ii) comparing some of the most promising objectives explored in the recent domain generalization and fair classification literature. From our theoretical analysis and empirical evaluation, we conclude that the model selection procedure needs to be guided by careful considerations regarding the observed data, the factors used for correction and the structure of the data-generating process.

1 Introduction

One of the most common assumptions for machine learning models is that the training and test data are independently and identically sampled (IID) from the same distribution. In practice, this assumption does not hold in many practical scenarios (Bengio et al., 2020). A machine learning model trained to recognize land usage from satellite images using pictures from the early 2000s may struggle to recognize the style of modern architectures (Christie et al., 2018), data collected on a limited set of hospitals may not be representative of the variation introduced by the use of different machines or procedures (Zech et al., 2018; Beede et al., 2020). Other kinds of distribution shifts are more subtle and difficult to recognize despite having a noticeable impact on the model's predictive performance. Examples include under-represented or over-represented population groups (Popejoy & Fullerton, 2016; Buolamwini & Gebru, 2018) or biased annotations collected from crowd-sourcing services (Zhang et al., 2017; Xia et al., 2020).

Different approaches in literature address these issues by using some external source of knowledge such as domain or environment annotations (Wang & Deng, 2018), protected attributes (Mehrabi et al., 2019) or sub-population groups (Santurkar et al., 2020) to reduce bias and minimize the model error outside of the training distribution. Despite the progress in the field of domain generalization literature, Gulrajani & Lopez-Paz (2020) has shown that the effectiveness of some of the most common algorithms heavily relies on the hyper-parameter tuning strategy, revealing limitations of examined models when compared to a more traditional empirical risk minimization strategy. One of the major issues behind this observed behavior is a lack of clarity and applicability for the underlying assumptions regarding the problem statement and the data-generating process (Zhao et al., 2019; Rosenfeld et al., 2020; Mahajan et al., 2020).

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

With this work, we aim to present a new perspective on the problem to analyze and clarify the fundamental differences between some of the most common approaches by:

1. Introducing a novel information-theoretical framework to describe the problem of distribution shift and connecting it to the out-of-distribution error and its components (section 2).
2. Analyzing three main families of objectives and describing some of their guarantees and assumptions (section 3).
3. Showing that the effectiveness of different criteria is determined by the structure of the underlying data-generating process other than the chosen approximations (section 5).

The analysis of the *independence*, *sufficiency*, and *separation* criteria reveals that some of the most popular models can systematically fail to reduce the out-of-distribution error even for simple datasets. No unique objective is simultaneously optimal for every problem, but additional insights and knowledge about the selection procedure can help to mitigate the bias.

2 Framework and Problem Statement

Consider $\mathbf{x}$ and $\mathbf{y}$ as the features and targets respectively with joint density $p(\mathbf{x}, \mathbf{y})$ for the predictive problem of interest. Let t be a binary random variable representing which data is selected for training ($\mathrm{t} = 1$) and which is not ($\mathrm{t} = 0$). We will refer to $p(\mathbf{x}, \mathbf{y}|\mathrm{t} = 1)$ as the joint *Training distribution* and $p(\mathbf{x}, \mathbf{y}|\mathrm{t} = 0)$ as the *Test distribution* that are induced by the selection t. Throughout the paper, we will consider selections t that can be expressed as a function of the features $\mathbf{x}$, targets $\mathbf{y}$, independent noise ϵ and other variables $\mathbf{e}$, which represent other factors that are affecting the data collection procedure (e.g. geographic location, time intervals, population groups).

Let $q(\mathbf{y}|\mathbf{x})$ represent a learnable model for the predictive distribution $p(\mathbf{y}|\mathbf{x})$. We use the Kullback-Leibler divergence to express the *Training* and *Out-Of-Distribution* (OOD) error respectively[1]:

$$D_{\mathrm{KL}}(p^{\mathrm{t}=1}_{\mathbf{y}|\mathbf{x}}||q_{\mathbf{y}|\mathbf{x}}) := D_{\mathrm{KL}}(p(\mathbf{y}|\mathbf{x}, \mathrm{t} = 1)||q(\mathbf{y}|\mathbf{x})), \tag{1}$$

$$D_{\mathrm{KL}}(p^{\mathrm{t}=0}_{\mathbf{y}|\mathbf{x}}||q_{\mathbf{y}|\mathbf{x}}) := D_{\mathrm{KL}}(p(\mathbf{y}|\mathbf{x}, \mathrm{t} = 0)||q(\mathbf{y}|\mathbf{x})). \tag{2}$$

2.1 Characterizing Distribution Shift

The effect of the selection can be quantified by considering the mutual information between the features-target pair $\mathbf{xy}$ and the selection variable t, which can be also expressed as a Kullback-Leibler divergence: $I(\mathbf{xy}; \mathrm{t}) = D_{\mathrm{KL}}(p(\mathbf{x}, \mathbf{y}|\mathrm{t})|p(\mathbf{x}, \mathbf{y}))$. This measure of distribution shift intuitively represents how many bits the selection variable carries about the joint distribution of targets and features or, equivalently, how much the joint density $p(\mathbf{x}, \mathbf{y})$ has changed as a result of the selection.

Using the chain rule of mutual information, one can express the distribution shift as the sum of two separate components:

$$\underbrace{I(\mathbf{xy}; \mathrm{t})}_{\text{Distribution shift}} = \underbrace{I(\mathbf{x}; \mathrm{t})}_{\text{Covariate shift}} + \underbrace{I(\mathbf{y}; \mathrm{t}|\mathbf{x})}_{\text{Concept shift}}, \tag{3}$$

which can be interpreted as the amount of *covariate shift* (Shimodaira, 2000) and *concept shift* (Widmer & Kubat, 1996; Moreno-Torres et al., 2012), respectively. These two quantities refer to the changes in the predictive distribution $p(\mathbf{y}|\mathbf{x})$ and the marginal features distribution $p(\mathbf{x})$ respectively, which add up to represent the changes in the joint distribution.

Whenever the data selection is perfectly IID, the selection variable can be expressed as a function of some independent noise ($\mathrm{t} = f(\epsilon)$) and the corresponding distribution shift $I(\mathbf{xy}; \mathrm{t})$ is zero. On the other hand, if the data collection procedure has been influenced by other factors, we do not have such a guarantee, even when the selection does not depend on features and targets directly ($\mathrm{t} = f(\mathbf{e}, \epsilon)$).

Train and OOD error relate to the concept shift via the following expression:

Proposition 1. *For any model $q(\mathbf{y}|\mathbf{x})$ and $\alpha := \min\{p(\mathrm{t} = 0), p(\mathrm{t} = 1)\}$:*

$$D_{\mathrm{KL}}(p^{\mathrm{t}=1}_{\mathbf{y}|\mathbf{x}}||q_{\mathbf{y}|\mathbf{x}}) + D_{\mathrm{KL}}(p^{\mathrm{t}=0}_{\mathbf{y}|\mathbf{x}}||q_{\mathbf{y}|\mathbf{x}}) \geq \frac{1}{1-\alpha} I(\mathbf{y}; \mathrm{t}|\mathbf{x}), \tag{4}$$

[1]Further details regarding the convention used for conditional KL-divergence can be found in appendix A

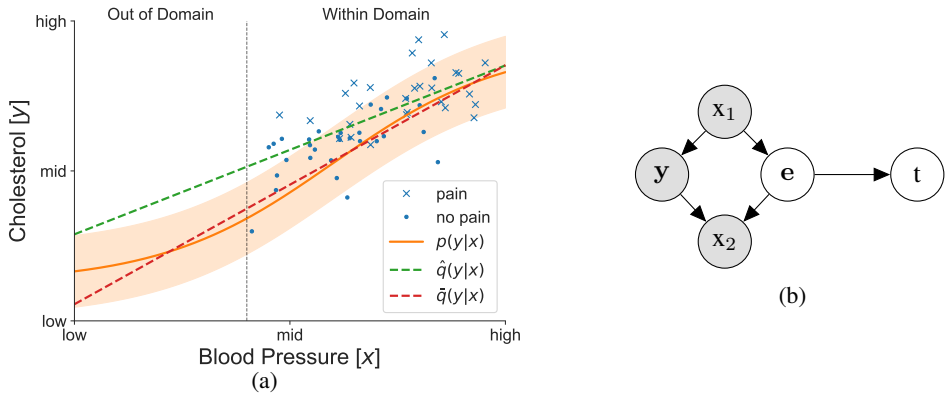

Figure 1: (a) Example of distribution shift for the prediction of cholesterol levels (y-axis) given blood pressure (x-axis). A maximum likelihood approach (green dashed line) results in a model that over-estimates y for any given x. Correcting for the concept shift (red dashed line) using the information about chest pain reduces the error within the observed region. (b) Example of a data generating process in which the features $\mathbf{x}$ are composed of two parts x_1 and x_2. When both feature components x_1 and x_2 are observed, concept shift can be strictly positive. Discarding information about x_2 will reduce the effect of the selection bias, while removing x_1 might increase it $0 = I(\mathbf{y}; \mathrm{t}|\mathrm{x}_1) \leq I(\mathbf{y}; \mathrm{t}|\mathrm{x}_1\mathrm{x}_2) \leq I(\mathbf{y}; \mathrm{t}|\mathrm{x}_2)$.

As a consequence, whenever the selection induces concept shift $I(\mathbf{y}; \mathrm{t}|\mathbf{x}) > 0$, any sufficiently flexible model $\hat{q}(\mathbf{y}|\mathbf{x})$ trained using a maximum likelihood approach must incur in strictly positive OOD error. Intuitively, whenever the selection induces a change in the predictive distribution ($p(\mathbf{y}|\mathbf{x}) \neq p(\mathbf{y}|\mathbf{x}, \mathrm{t} = 1)$) fitting the model to the training distribution will incorporate the selection bias into the model prediction, necessarily resulting in errors when evaluated on the test distribution, as shown in the example reported in figure 1a.

2.2 A Representation-Learning-based Approach

We first introduce a *latent representation* $\mathbf{z}$ and parametrize the model $q(\mathbf{y}|\mathbf{x})$ with an *encoder* $q(\mathbf{z}|\mathbf{x})$ and a *latent classifier*[2] $q(\mathbf{y}|\mathbf{z})$.

$$q(\mathbf{y}|\mathbf{x}) = \mathbb{E}_{\boldsymbol{z} \sim q(\mathbf{z}|\mathbf{x})} \left[q(\mathbf{y}|\mathbf{z} = \boldsymbol{z}) \right]. \tag{5}$$

The effect of the introduction of a latent representation can be observed by expressing the training and OOD error as a function of the encoder and the classifier[3].

Proposition 2. *For any encoder $q(\mathbf{z}|\mathbf{x})$ and classifier $q(\mathbf{y}|\mathbf{z})$:*

$$D_{\mathrm{KL}}(p^{\mathrm{t}=1}_{\mathbf{y}|\mathbf{x}} || q_{\mathbf{y}|\mathbf{x}}) \leq \underbrace{I_{\mathrm{t}=1}(\mathbf{x}; \mathbf{y}|\mathbf{z})}_{\textit{Train information loss}} + \underbrace{D_{\mathrm{KL}}(p^{\mathrm{t}=1}_{\mathbf{y}|\mathbf{z}} || q_{\mathbf{y}|\mathbf{z}})}_{\textit{Latent train error}} \tag{6}$$

$$D_{\mathrm{KL}}(p^{\mathrm{t}=0}_{\mathbf{y}|\mathbf{x}} || q_{\mathbf{y}|\mathbf{x}}) \leq \underbrace{I_{\mathrm{t}=0}(\mathbf{x}; \mathbf{y}|\mathbf{z})}_{\textit{Test Information loss}} + \underbrace{D_{\mathrm{KL}}(p^{\mathrm{t}=0}_{\mathbf{y}|\mathbf{z}} || q_{\mathbf{y}|\mathbf{z}})}_{\textit{Latent OOD error}}. \tag{7}$$

The two terms $I_{t=1}(\mathbf{x}; \mathbf{y}|\mathbf{z})$ and $I_{t=0}(\mathbf{x}; \mathbf{y}|\mathbf{z})$ represent the amount of predictive information that is lost by encoding the features $\mathbf{x}$ into $\mathbf{z}$ on train and test distribution respectively, while $D_{\mathrm{KL}}(p^{\mathrm{t}=1}_{\mathbf{y}|\mathbf{z}} || q_{\mathbf{y}|\mathbf{z}})$ and $D_{\mathrm{KL}}(p^{\mathrm{t}=0}_{\mathbf{y}|\mathbf{z}} || q_{\mathbf{y}|\mathbf{z}})$ refer to the train and OOD error when using $\mathbf{z}$ instead of the original observations as the predictive features, which will be referred to as *latent training error* and *latent out of distribution error* respectively. Latent OOD error and the test information loss capture two intrinsically different kinds of error. The former indicates the discrepancy between the model $q(\mathbf{y}|\mathbf{z})$ and the test predictive distribution $p(\mathbf{y}|\mathbf{z}, \mathrm{t} = 0)$, while the latter represents the increase in the prediction uncertainty as a result of the encoding procedure.

[2]The name classifier will be used for convenience to express both classification and regression problems.

[3]The expressions in equations 6 and 7 hold with equality for deterministic encoders as shown in appendix B.

2.3 Latent OOD error

Guaranteeing that $q(\mathbf{y}|\mathbf{z})$ is not confident about the wrong prediction requires minimizing the latent OOD error, which represents how far the classifier model is from the test latent predictive distribution. We can show that the Jensen-Shannon divergence between the two distributions is upper-bounded by a monotonic function of the latent training error and the amount of concept shift in the latent space (*latent concept-shift* $I(\mathbf{y};\mathrm{t}|\mathbf{z})$):

Proposition 3. *For any $q(\mathbf{z}|\mathbf{x})$, $q(\mathbf{y}|\mathbf{z})$ and any representation $\boldsymbol{z}$ that satisfies $p(\mathbf{z}=\boldsymbol{z}|\mathrm{t}=0)>0$ and $p(\mathbf{z}=\boldsymbol{z}|\mathrm{t}=1)>0$:*

$$\left(\sqrt{\frac{1}{2\alpha}I(\mathbf{y};\mathrm{t}|\mathbf{z}=\boldsymbol{z})}+\sqrt{\frac{1}{2}D_{\mathrm{KL}}(p^{\mathrm{t}=1}_{\mathbf{y}|\mathbf{z}=\boldsymbol{z}}||q_{\mathbf{y}|\mathbf{z}=\boldsymbol{z}})}\right)^2 \geq D_{\mathrm{JSD}}(p^{\mathrm{t}=0}_{\mathbf{y}|\mathbf{z}=\boldsymbol{z}}||q_{\mathbf{y}|\mathbf{z}=\boldsymbol{z}}). \tag{8}$$

Whenever the Jensen-Shannon divergence between the test predictive distribution $p(\mathbf{y}|\mathbf{z},\mathrm{t}=0)$ and the classifier $q(\mathbf{y}|\mathbf{z})$ is small, the Latent OOD error (measured in terms of KL-divergence) must also be small at least for the regions that have positive probability according to both train $p(\mathbf{x}|\mathrm{t}=1)$ and test $p(\mathbf{x}|\mathrm{t}=0)$ data distributions (Figure 1a). Since the train predictive distribution $p(\mathbf{y}|\mathbf{x},\mathrm{t}=1)$ is not defined for $\boldsymbol{x}$ that have zero probability on the train distribution, we have no guarantees regarding the model predictions in those regions unless other inductive biases are considered.

The left hand side of the expression in proposition 3 can be minimized by addressing $I(\mathbf{y};\mathrm{t}|\mathbf{z})$ and $D_{\mathrm{KL}}(p^{\mathrm{t}=1}_{\mathbf{y}|\mathbf{z}=\boldsymbol{z}}||q_{\mathbf{y}|\mathbf{z}=\boldsymbol{z}})$ with respect to the encoder and classifier respectively. In other words, we can minimize the distance between the classifier and the latent test predictive distribution by simultaneously fitting $q(\mathbf{y}|\mathbf{z})$ to the train predictive distribution $p(\mathbf{y}|\mathbf{z},\mathrm{t}=1)$ and minimizing the latent concept shift induced by the encoder $q(\mathbf{z}|\mathbf{x})$. When the latent concept shift and latent OOD error approach zero, the latent train predictive distribution $p(\mathbf{y}|\mathbf{z},\mathrm{t}=1)$ approaches the true (unselected) predictive distribution $p(\mathbf{y}|\mathbf{z})$, and so does the modeled classifier $q(\mathbf{y}|\mathbf{z})$. In this scenario, the only source of OOD error is the information that has been lost in the encoding procedure.

2.4 Minimizing the information loss

Combining the results from propositions 1 and 2 we can infer that reducing the latent OOD error necessarily requires the representation to discard some train predictive information ($I_{\mathrm{t}=1}(\mathbf{x};\mathbf{y}|\mathbf{z})$>0). This is because a lossless encoder $\hat{q}(\mathbf{z}|\mathbf{x})$ together with the optimal latent classifier $\hat{q}(\mathbf{y}|\mathbf{z})$ would result in an overall model $\hat{q}(\mathbf{y}|\mathbf{x})$ that achieves zero training error, and, therefore, positive latent OOD error (proposition 1). However, discarding information without any specific criteria can even increase the latent concept shift, as shown by the example in figure 1b.

As the amount of information that is lost in the encoding procedure is one of the two components of the OOD error (equation 7), we will consider objectives that discard the minimal amount of information required to reduce the latent concept shift. Since only the selected training distribution is available at training time, we will consider objectives that maximize the amount of predictive training information $I_{\mathrm{t}=1}(\mathbf{y};\mathbf{z})$ or, equivalently, minimize the KL-divergence between the training predictive distribution $p(\mathbf{y}|\mathbf{x};\mathrm{t}=1)$ and the latent classifier model $q(\mathbf{y}|\mathbf{z})$:

$$\min_{q(\mathbf{z}|\mathbf{x})} I_{\mathrm{t}=1}(\mathbf{x};\mathbf{y}|\mathbf{z}) = I_{\mathrm{t}=1}(\mathbf{y};\mathbf{x}) - \max_{q(\mathbf{z}|\mathbf{x})} I_{\mathrm{t}=1}(\mathbf{y};\mathbf{z}) \leq \min_{q(\mathbf{z}|\mathbf{x}),q(\mathbf{y}|\mathbf{z})} D_{\mathrm{KL}}(p^{\mathrm{t}=1}_{\mathbf{y}|\mathbf{x}}||q_{\mathbf{y}|\mathbf{z}}) \tag{9}$$

2.5 A General Loss function

To summarize, the overall objective is finding the maximally informative representation that minimizes latent concept shift while fitting the latent classifier to the data. This is achieved by (i) matching the classifier $q(\mathbf{y}|\mathbf{z})$ and the training latent predictive distribution $p(\mathbf{y}|\mathbf{z},\mathrm{t}=1)$, (ii) maximizing the amount of predictive information in the representation, and (iii) minimizing the amount of latent concept shift induced by the representation. The three requirements can be enforced by considering the following loss function:

$$\mathcal{L}(q_{\mathbf{z}|\mathbf{x}}, q_{\mathbf{y}|\mathbf{z}};\lambda) = D_{\mathrm{KL}}(p^{\mathrm{t}=1}_{\mathbf{y}|\mathbf{x}}||q_{\mathbf{y}|\mathbf{z}}) + \lambda\mathcal{R}(q_{\mathbf{z}|\mathbf{x}}), \tag{10}$$

in which, the first term addresses (i) and (ii), while the second term represent a regularization term that acts on the encoder $q(\mathbf{z}|\mathbf{x})$ to minimize latent concept shift $I(\mathbf{y};\mathrm{t}|\mathbf{z})$, following (iii). The hyper-parameter λ defines the trade-off requirements (ii) and (iii).

3 Regularization Criteria

Since only selected data ($\mathrm{t} = 1$) is accessible at training time, the latent concept shift can not be computed or minimized directly. The different approaches considered in this analysis make use of a regularization $\mathcal{R}(q_{\mathbf{z}|\mathbf{x}})$ that is based on the observation of an additional variable $\mathbf{e}$ which relates to the selection criteria. This variable is usually referred to as *domain* or *environment*[4] in the domain adaptation and generalization literature, while the name *protected attribute* is used in the context of fair classification. We will refer to this variable $\mathbf{e}$ as *environmental factors* in the following sections.

We analyze three families of criteria proposed in the domain generalization (Koyama & Yamaguchi, 2020) and fair classification (Barocas et al., 2018) literature focusing on their underlying assumptions and theoretical guarantees. The different regularization strategies and models can be seen as specific instance of the loss function in equation 10. An empirical comparison between instances of the different criteria can be found in section 5, proofs are reported in appendix B, while the relation between the reported criteria is further discussed in appendix B.8.

3.1 Independence Criterion

Whenever the data selection t depends exclusively on some variable $\mathbf{e}$ ($\mathrm{t} = f(\mathbf{e}; \epsilon)$), the most intuitive approach to reduce the latent OOD error is to make the representation $\mathbf{z}$ independent of the environmental factors $\mathbf{e}$. This can be done by minimizing mutual information between $\mathbf{e}$ and $\mathbf{z}$: $\mathcal{R}(q_{\mathbf{z}|\mathbf{x}}) := I_{t=1}(\mathbf{e}; \mathbf{z})$. This criterion, known as *statistical parity* in the fair classification literature (Dwork et al., 2011; Corbett-Davies et al., 2017), aims to remove any environmental information from $\mathbf{z}$ resulting in a representation that satisfies $p(\mathbf{z}|\mathrm{t} = 1) = p(\mathbf{z}|\mathbf{e}, \mathrm{t} = 1)$.

Despite the usefulness of this criterion in the fairness and differential privacy literature, enforcing independence does not necessarily reduce the OOD error (Zhao et al., 2019; Johansson et al., 2019). This is because a consistent marginal across different environments does not imply a consistent predictive distribution ($p(\mathbf{y}|\mathbf{z}, \mathbf{e}, \mathrm{t} = 1) \neq p(\mathbf{y}|\mathbf{z}, \mathrm{t} = 1)$), and enforcing independence may even increase the latent concept shift and the OOD error (as shown in section 5.3).

3.2 Sufficiency Criterion

Instead of enforcing a property on the marginal feature distribution, one can consider stable properties of the joint distribution of features $\mathbf{z}$ and labels $\mathbf{y}$. The requirement of creating a representation that yields a stable classifier for different values of the environmental factor $\mathbf{e}$ can be captured by the following regularization: $\mathcal{R}(q_{\mathbf{z}|\mathbf{x}}) := I_{t=1}(\mathbf{e}; \mathbf{y}|\mathbf{z})$. Intuitively minimizing $I_{t=1}(\mathbf{e}; \mathbf{y}|\mathbf{z})$ corresponds to minimizing the distance between the predictive distribution $p(\mathbf{y}|\mathbf{e}, \mathbf{z}, \mathrm{t} = 1)$ for each one of the observed environmental conditions. As a result:

Proposition 4. *Whenever the selection* t *is a function of* $\mathbf{e}$ *and some independent noise* ϵ*, the latent concept shift induced by a representation* $\mathbf{z}$ *of* $\mathbf{x}$ *must be upper-bounded by* $I(\mathbf{y}; \mathbf{e}|\mathbf{z})$*:*

$$\exists f : \mathrm{t} = f(\mathbf{e}, \epsilon) \implies I(\mathbf{y}; \mathrm{t}|\mathbf{z}) \leq I(\mathbf{y}; \mathbf{e}|\mathbf{z}). \tag{11}$$

In other words, for a given selection t, it is possible to remove the effect of concept shift by enforcing the sufficiency constraint using the variable (or variables) $\mathbf{e}$ that are responsible for that selection.

The result from proposition 4 is applicable only if the two following conditions are met: (i) a sufficient representation must exist (Koyama & Yamaguchi, 2020); (ii) enforcing sufficiency on the selected train distribution results in sufficiency for the overall joint distribution ($I_{\mathrm{t}=1}(\mathbf{y}; \mathbf{e}|\mathbf{z}) = 0 \implies I(\mathbf{y}; \mathbf{e}|\mathbf{z}) = 0$) (Rosenfeld et al., 2020). Even if assumptions (i) and (ii) are often acceptable in practice, lack of environment variety at training time or direct dependencies between environmental factors and targets (such as in the y-CMNIST dataset in section 5.3) can compromise the effectiveness of the sufficiency criterion. An in depth discussion with a simple example is reported in appendix F.

3.3 Separation Criterion

The last family of objectives includes approaches that aim to capture the stability in the latent feature distribution when the target is observed across different environmental conditions $p(\mathbf{z}|\mathbf{y}) = p(\mathbf{z}|\mathbf{y}, \mathbf{e})$.

[4] In contrast with the domain adaptation and generalization literature, we will consider the more general case in which $\mathbf{e}$ is represented by a vector.

This requirement can be enforced by minimizing the dependency between environmental factors and the representation when the label is observed: $\mathcal{R}(q_{\mathbf{z}|\mathbf{x}}) := I_{t=1}(\mathbf{e}; \mathbf{z}|\mathbf{y})$. The resulting separation criterion can be used to identify stable properties of the joint distributions even when the selection t depends on both targets $\mathbf{y}$ and environmental factors $\mathbf{e}$:

Proposition 5. *If the selection* t *is a function of* $\mathbf{e}$*,* $\mathbf{y}$ *and some independent noise* ϵ*, the latent concept shift of a representation* $\mathbf{z}$ *of* $\mathbf{x}$ *is upper-bounded by the sum of prior-shift* $I(\mathbf{y}; \mathrm{t})$ *and* $I(\mathbf{e}; \mathbf{z}|\mathbf{y})$*:*

$$\exists f : \mathrm{t} = f(\mathbf{e}, \mathbf{y}, \epsilon) \implies I(\mathbf{y}; \mathrm{t}|\mathbf{z}) \leq I(\mathbf{y}; \mathrm{t}) + I(\mathbf{e}; \mathbf{z}|\mathbf{y}). \tag{12}$$

When selection t and targets $\mathbf{y}$ are marginally independent, proposition 5 guarantees that the latent concept shift of a representation that enforces separation is zero. Furthermore, whenever the marginal distribution $p(\mathbf{y}|\mathrm{t} = 0)$ is known, it is possible to adjust the prediction of the latent classifier on the test distribution[5].

Similarly to the sufficiency criterion, one needs to assume that enforcing separation on the train suffices to guarantee $I(\mathbf{e}; \mathbf{z}|\mathbf{y}) = 0$. Although a representation that enforces separation always exists, the effectiveness of the separation criteria depends on the data-generating process, since the requirement could be exclusively satisfied by a constant representation.

4 Related work

The problem of distribution (or dataset) shift (Quionero-Candela et al., 2009; Koh et al., 2020) has been explored in different areas in the machine learning literature ranging domain adaptation and generalization (Wang & Deng, 2018) to sub-population shift and fair classification (Mehrabi et al., 2019; Santurkar et al., 2020). Although goals, availability of test features and environments at training time, and data selection criteria (Koh et al., 2020) may differ, most recent approaches focus on extracting features with some desired properties through three different families of objectives.

The independence criterion has been used to find the transformation which minimizes the distance between the distribution of the encoded features across different environments by using adversarial training (Ganin et al., 2016; Xie et al., 2017; Li et al., 2018a), features adjustment (Lum & Johndrow, 2016), kernel methods (Muandet et al., 2013), or variational approaches (Louizos et al., 2016; Moyer et al., 2018; Ilse et al., 2020). Despite the success of the independence criterion, Zhao et al. (2019); Johansson et al. (2019) have shown that enforcing stability of the marginal feature distribution is not sufficient to guarantee generalization.

Li et al. (2018b) extends the independence criterion using adversarial training to create representations that are independent of the environment when the label is observed. The corresponding separation criterion has been explored in the context of fair classification as *equalized odds* (Hardt et al., 2016) and some tractable relaxations (Darlington, 1971; Woodworth et al., 2017; Chouldechova, 2017).

The idea of considering features that lead to a consistent predictive distribution has been explored in the causality literature (Peters et al., 2016; Magliacane et al., 2018) for linear models as a feature selection criteria. Other work approached the problem in the non-linear case by considering gradients of the classifier (Arjovsky et al., 2019; Koyama & Yamaguchi, 2020; Parascandolo et al., 2020) or penalizing the variance of error (Krueger et al., 2020; Xie et al., 2020) across different environments. Other relaxations of the sufficiency criterion such as as *predictive parity* (Chouldechova, 2017) have been considered in the fairness literature.

Despite the promising directions of research, Gulrajani & Lopez-Paz (2020) have shown that the performance of the most popular approaches in the literature strongly depends on the hyper-parameter tuning strategy, underlying the problem of the lack of (i) standardized benchmark procedures and (ii) clarity regarding guarantees and hidden assumptions for the different algorithms. Although the problem of distribution shift has been widely studied in the classic domain generalization literature (Mansour et al., 2009; Schölkopf et al., 2012; Gong et al., 2016), this work aims to characterize the problem from a representation learning and information theoretical perspective, providing different insights and identifying potential issues and benefits of several popular models.

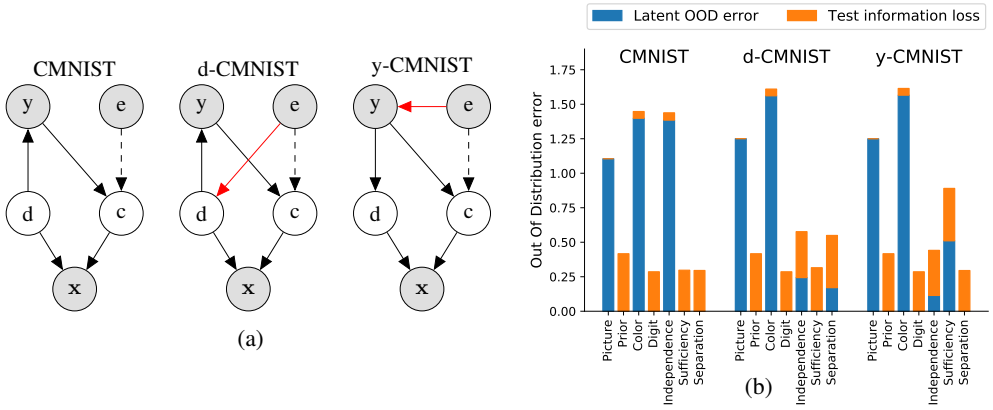

Figure 2: Graphical models (a) and error components (b) for the CMNIST, d-CMNIST and y-CMNIST data distributions. (a) Dashed lines are used to underline marginal independence between color c and environment e, while red arrows denote dependencies added to the original CMNIST distribution. (b) Models trained with strong regularization ($\lambda \approx 10^7$) for the three different criteria are compared against the classifiers trained using only color ($p(\mathrm{y}|\mathrm{c}, \mathrm{t}=1)$), digit ($p(\mathrm{y}|\mathrm{d}, \mathrm{t}=1)$), picture ($p(\mathrm{y}|\mathbf{x}, \mathrm{t}=1)$), or prior information ($p(\mathrm{y}|\mathrm{t}=1)$). The colors show the proportion of the OOD error (in nats) due to the predictive information loss ($I_{t=0}(\mathbf{x}; \mathbf{y}|\mathbf{z})$, in orange) and the latent OOD error ($D_{\mathrm{KL}}(p^{\mathrm{t}=0}_{\mathbf{y}|\mathbf{z}}||q_{\mathbf{y}|\mathbf{z}})$, in blue) according to the decomposition in equation 7.

5 Experiments

We evaluate the effectiveness of different criteria presented in section 3 and some of their most popular implementations on multiple versions of the CMNIST dataset (Arjovsky et al., 2019) produced by altering the data-generating process (figure 2a) to underline the shortcomings of different criteria.

The main advantage of using a synthetic dataset is the possibility to estimate the training and out-of-distribution error for the different criteria while directly optimizing their objective. This is possible only when the variables in the data-generating process are discrete and low-dimensional, since direct optimization requires mutual information terms to be computed. As a result, the effect of the different criteria can be observed in isolation, without the influence introduced by the approximations. Further details regarding the direct optimization of the analyzed criteria are reported in appendix D.

5.1 Models and Criteria

The theoretical performance for each criterion is compared against the results obtained by training different models in the literature using neural network architectures to parametrize the encoder $q(\mathrm{z}|\mathbf{x})$ and classifier $q(\mathrm{y}|\mathbf{z})$. For the comparison, we selected one representative model for each criterion defined in section 3 together with a baseline trained using the Information Bottleneck criterion (Tishby & Zaslavsky, 2015), which does not use any environmental information and blindly discards data-features depending on the regularization strength λ. The overall comparison includes the following models:

- Variational Information Bottleneck (VIB) (Alemi et al., 2017): a variational approximation of the **Information Bottleneck criterion**;
- Domain Adversarial Neural Network (DANN) (Ganin et al., 2016): an adversarial model based on a min-max game with discriminator $d(\mathbf{e}|\mathbf{z})$ that is optimized to predict environment information from the representation $\mathbf{z}$ to enforce the **Independence criterion**;
- Variance-based Risk Extrapolation (VREx) (Krueger et al., 2020): a relaxation of the **Sufficiency criterion** based on the minimization of the training error variance across the different environments[6];

[5]Further details on the re-weighting procedure can be found in appendix B.7

[6]Results obtained using the Invariant Risk Minimization model (Arjovsky et al., 2019) as an alternative to enforce sufficiency are also reported in appendix E.2.

- Conditional Domain Adversarial Neural Network (CDANN) (Li et al., 2018b): an adversarial approximation of the **Separation criterion** in which, analogously to DANN, a discriminator tries to predict the environment when the representation $\mathbf{z}$ and the true label $\mathbf{y}$ are given.

For a better comparison, all the models use the same encoder and classifier neural network architectures. Each model has been trained by slowly increasing the regularization strength after an initial pre-training with $\lambda = 0$. Further details regarding the neural network architectures, objectives, optimization, and specific hyper-parameters can be found in appendix E.

5.2 Evaluation metric

The measure of OOD error, information Loss and latent OOD error reported in figure 2b have been computed directly by considering a discrete version of the analyzed datasets. For neural network models, on the other hand, the training and OOD error defined in section 2 can be estimated up to a constant entropy by considering the expected negative log-likelihood (empirical cross-entropy):

$$D_{\mathrm{KL}}(p^{\mathrm{t}=1}_{\mathbf{y}|\mathbf{x}}||q_{\mathbf{y}|\mathbf{x}}) \approx -\frac{1}{N}\sum_{i=1}^{N} \log q(\mathbf{y} = \boldsymbol{y}_i|\mathbf{x} = \boldsymbol{x}_i) - \underbrace{H_{\mathrm{t}=1}(\mathbf{y}|\mathbf{x})}_{\text{constant}} \tag{13}$$

The samples $\boldsymbol{x}_i, \boldsymbol{y}_i$ are obtained from $p(\mathbf{x}, \mathbf{y}|\mathrm{t} = 1)$ and $p(\mathbf{x}, \mathbf{y}|\mathrm{t} = 0)$ for train and test cross-entropy respectively. By computing the training and OOD cross-entropy values for different regularization strength λ, each model defines a trajectory from a maximum likelihood solution ($\lambda = 0$) to the results obtained when the corresponding independence constraint is enforced (large λ), as shown in figure 3. Contrarily to accuracy measurements, the use of cross-entropy allows us to detect under-confident or over-confident predictive distributions.

5.3 Datasets

The two variants of the CMNIST dataset considered in this work have been designed by minimally changing the original distribution to underline the strengths and weaknesses of the different criteria. Across different experiments, each MNIST picture $\mathbf{x}$ with digit label d is associated with a color c which depends on a binary target y and an environment e. The d-CMNIST and y-CMNIST datasets are created by adding a dependency from the environment to digit d (d-CMNIST) and label y (y-CMNIST), which results in a correlation between targets and environment ($I(\mathrm{e}; \mathrm{y}) > 0$). Further details regarding the conditional distributions used to produce the different datasets can be found in appendix C.

The strong correlation between color and label across the different CMNIST versions can be seen as an artifact introduced by the selection $\mathrm{t} = f(\mathrm{e})$, and models that consider only digit information (green crosses in figure 3) outperform the ones that capture color information (red crosses) or both (orange crosses) in terms of OOD cross-entropy.

CMNIST The CMNIST dataset was originally designed to underline the weaknesses of maximum-likelihood and Empirical Risk Minimization strategies (Arjovsky et al., 2019). The results displayed in the first column of figure 3 confirm that both the sufficiency and separation criteria manage to effectively reduce the OOD cross entropy error for sufficiently strong regularization λ by minimizing the latent OOD error while retaining more predictive information than a constant representation (first column in figure 2b). On the CMNIST dataset, both the DANN and CDANN models closely follow the trajectory estimated by optimizing the corresponding criterion directly. Despite the theoretical effectiveness of the sufficiency criterion, the VREx model is unable to achieve optimal OOD performance. We believe that this is due to the relaxation of the sufficiency criterion that the model introduces for optimization rather than the neural network training since the same architecture performs well when trained with the CDANN objective.

d-CMNIST The d-CMNIST dataset adds a dependency between environment e and digit d, increasing the frequency of specific digits for some environments. Digit information makes environment and label conditionally independent (y and e are d-separated in figure 2a). As claimed in Arjovsky et al. (2019), models based on the sufficiency criterion manage to reduce the latent OOD error (second column of figure 2b), while the separation criterion fails because of the direct dependency between

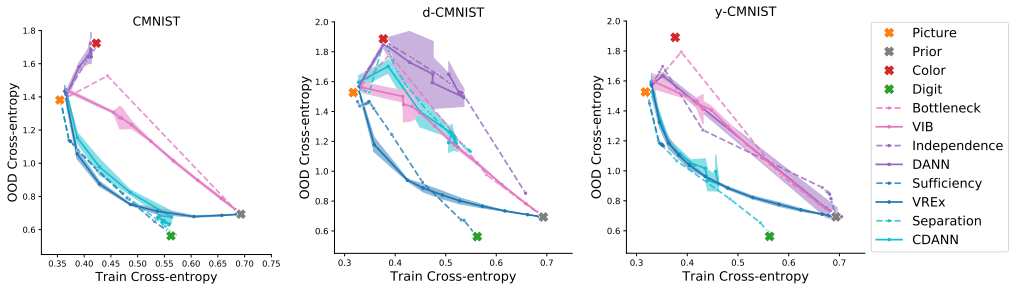

Figure 3: Comparison of training cross-entropy (x-axis) and OOD cross-entropy (y-axis) on the CMNIST, d-CMNIST and y-CMNIST datasets for models trained using different criteria (dashed lines) and corresponding approximations/relaxations (continuous lines with shaded area to represent the standard deviation obtained across three runs). Each trajectory describes the trade-off between the two errors for a wide range ($0 \rightarrow 10^7$) of regularization strength values λ. The four crosses in each plot represent the error of models that consider only color ($p(\mathrm{y}|\mathrm{c}, \mathrm{t}=1)$), digit ($p(\mathrm{y}|\mathrm{d}, \mathrm{t}=1)$), picture ($p(\mathrm{y}|\mathbf{x}, \mathrm{t}=1)$) or prior ($p(\mathrm{y}|\mathrm{t}=1)$) information. Models optimized with the VREx and CDANN objectives introduce errors due to relaxations and instabilities of the adversarial optimization.

environment and digits. The independence criterion does not improve the OOD performance since both color and digit correlate with the environment.

Both DANN and CDANN architectures exhibit the expected behavior across different runs while, despite the effectiveness of the sufficiency criterion, the model trained with the VREx objective is unable to recover the optimal solution. Note that enforcing separation or independence does not improve upon the result that can be obtained by blindly discarding information using VIB.

y-CMNIST Adding a dependency between environment e and label y simulates the scenario in which some labels are more frequent in some environments. The arrow between d and y flips when compared to the d-CMNIST dataset to represent stable $p(\mathrm{d}|\mathrm{y})$ across different environments. The corresponding y-CMNIST dataset includes a path from e to y that can not be blocked since no representation $\mathbf{z}$ can achieve sufficiency ($I(\mathrm{e}; \mathrm{y}|\mathbf{z}) = 0$). Despite the optimality of the separation criterion (dashed light blue line, third column of figure 3), the model trained with the DANN objective fails to minimize the OOD error due to instabilities of the adversarial training procedure[7]. Once again, the trajectory described by the VREx objective interpolates between the maximum likelihood solution and a prior classifier $p(\mathrm{y}|\mathrm{t}=0)$.

Figure 2b confirms that the separation criterion is the only one that manages to decrease the latent OOD error to zero without discarding the entirety of the information contained in $\mathbf{x}$ on the y-CMNIST dataset. This confirms that the effectiveness of both sufficiency and separation criteria depends on the structure of the underlying graphical model.

6 Discussion and Conclusion

In this work, we characterize an information-theoretic framework to analyze the distribution shift problem by relating it to the train and out of distribution error. We identify two sources of errors for models based on latent representations, and we show that different criteria explored in literature can be seen as different strategies to minimize concept shift in the latent space using extra information about the data selection procedure. We demonstrate both theoretically and empirically that their effectiveness depends on the structure of the underlying graphical model with respect to the observed variables, training data, and data selection criteria other than the chosen approximation and relaxations for optimization.

Although test information loss and latent out of distribution error are challenging to estimate for real-world datasets, we argue that the presented analysis can be useful to better understand and mitigate the effect of selection bias and distribution shift in machine learning models. An accurate estimation of the error decomposition could be further used to guide processes of causal discovery.

[7]Further details on the observes instabilities are discussed in appendix E.

References

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

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

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2018. `http://www.fairmlbook.org`.

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., and Vardoulakis, L. M. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In Bernhaupt, R., Mueller, F. F., Verweij, D., Andres, J., McGrenere, J., Cockburn, A., Avellino, I., Goguey, A., Bjøn, P., Zhao, S., Samson, B. P., and Kocielnik, R. (eds.), *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pp. 1–12. ACM, 2020.

Bengio, Y., Deleu, T., Rahaman, N., Ke, N. R., Lachapelle, S., Bilaniuk, O., Goyal, A., and Pal, C. J. A meta-transfer objective for learning to disentangle causal mechanisms. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C. (eds.), *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 2018.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

Christie, G., Fendley, N., Wilson, J., and Mukherjee, R. Functional map of the world. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 6172–6180. IEEE Computer Society, 2018.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pp. 797–806. ACM, 2017.

Darlington, R. B. Another look at "cultural fairness". *Journal of Educational Measurement*, 8:71–82, 1971. ISSN 00220655, 17453984.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. Fairness through awareness. *CoRR*, abs/1104.3913, 2011.

Endres, D. M. and Schindelin, J. E. A new metric for probability distributions. *IEEE Trans. Inf. Theory*, 49(7):1858–1860, 2003.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. S. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17: 59:1–59:35, 2016.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., and Schölkopf, B. Domain adaptation with conditional transferable components. In Balcan, M. and Weinberger, K. Q. (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 2839–2848. JMLR.org, 2016.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *CoRR*, abs/2007.01434, 2020.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3315–3323, 2016.

Ilse, M., Tomczak, J. M., Louizos, C., and Welling, M. DIVA: domain invariant variational autoencoders. In Arbel, T., Ayed, I. B., de Bruijne, M., Descoteaux, M., Lombaert, H., and Pal, C. (eds.), *International Conference on Medical Imaging with Deep Learning, MIDL 2020, 6-8 July 2020, Montréal, QC, Canada*, volume 121 of *Proceedings of Machine Learning Research*, pp. 322–348. PMLR, 2020.

Johansson, F. D., Sontag, D. A., and Ranganath, R. Support and invertibility in domain-invariant representations. In Chaudhuri, K. and Sugiyama, M. (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 527–536. PMLR, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A benchmark of in-the-wild distribution shifts. *CoRR*, 2020.

Koyama, M. and Yamaguchi, S. Out-of-distribution generalization with maximal invariant predictor. *CoRR*, 2020.

Krueger, D., Caballero, E., Jacobsen, J., Zhang, A., Binas, J., Priol, R. L., and Courville, A. C. Out-of-distribution generalization via risk extrapolation (rex). *CoRR*, 2020.

Li, H., Pan, S. J., Wang, S., and Kot, A. C. Domain generalization with adversarial feature learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 5400–5409. IEEE Computer Society, 2018a.

Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., and Tao, D. Deep domain generalization via conditional invariant adversarial networks. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y. (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, volume 11219 of *Lecture Notes in Computer Science*, pp. 647–663. Springer, 2018b.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. S. The variational fair autoencoder. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Lum, K. and Johndrow, J. E. A statistical framework for fair predictive algorithms. *CoRR*, abs/1610.08077, 2016.

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. Domain adaptation by using causal inference to predict invariant conditional distributions. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 10869–10879, 2018.

Mahajan, D., Tople, S., and Sharma, A. Domain generalization using causal matching. *CoRR*, abs/2006.07500, 2020.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *CoRR*, 2019.

Moreno-Torres, J. G., Raeder, T., Alaíz-Rodríguez, R., Chawla, N. V., and Herrera, F. A unifying view on dataset shift in classification. *Pattern Recognit.*, 45(1):521–530, 2012.

Moyer, D., Gao, S., Brekelmans, R., Galstyan, A., and Steeg, G. V. Invariant representations without adversarial training. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 9102–9111, 2018.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 10–18. JMLR.org, 2013.

Parascandolo, G., Neitz, A., Orvieto, A., Gresele, L., and Schölkopf, B. Learning explanations that are hard to vary. *CoRR*, abs/2009.00329, 2020.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 947–1012, 2016.

Popejoy, A. and Fullerton, S. Genomics is failing on diversity. *Nature*, 538:161–164, 2016.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset Shift in Machine Learning*. The MIT Press, 2009. ISBN 0262170051.

Rosenfeld, E., Ravikumar, P., and Risteski, A. The risks of invariant risk minimization. *CoRR*, abs/2010.05761, 2020.

Santurkar, S., Tsipras, D., and Madry, A. BREEDS: benchmarks for subpopulation shift. *CoRR*, abs/2008.04859, 2020.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. M. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 2000.

Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop, ITW 2015, Jerusalem, Israel, April 26 - May 1, 2015*, pp. 1–5. IEEE, 2015.

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

Widmer, G. and Kubat, M. Learning in the presence of concept drift and hidden contexts. *Mach. Learn.*, 23(1):69–101, 1996.

Woodworth, B. E., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In Kale, S. and Shamir, O. (eds.), *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1920–1953. PMLR, 2017.

Xia, M., Field, A., and Tsvetkov, Y. Demoting racial bias in hate speech detection. In Ku, L. and Li, C. (eds.), *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, SocialNLP@ACL 2020, Online, July 10, 2020*, pp. 7–14. Association for Computational Linguistics, 2020.

Xie, C., Chen, F., Liu, Y., and Li, Z. Risk variance penalization: From distributional robustness to causality. *CoRR*, abs/2006.07544, 2020.

Xie, Q., Dai, Z., Du, Y., Hovy, E. H., and Neubig, G. Controllable invariance through adversarial feature learning. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 585–596, 2017.

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*, 2018.

Zhang, J., Sheng, V. S., Li, Q., Wu, J., and Wu, X. Consensus algorithms for biased labeling in crowdsourcing. *Inf. Sci.*, 382-383:254–273, 2017.

Zhao, H., des Combes, R. T., Zhang, K., and Gordon, G. J. On learning invariant representations for domain adaptation. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7523–7532. PMLR, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The claims point directly to the part of the papers in which they are supported
   (b) Did you describe the limitations of your work? [Yes] Discussion regarding the assumptions is reported in appendix F and in the conclusions.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] The whole paper aims to better understand and address the problems of distribution shift and selection bias, which can have negative social impact
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Precise statements are reported in appendix B
   (b) Did you include complete proofs of all theoretical results? [Yes] Proofs are reported in appendix B
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code used to compute the results is either publicly available or included in the anonymized URL that is reported in the appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See appendix E
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Error bars are reported for neural network experiments. The standard deviation for the results obtained with direct optimization is negligible.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See appendix E
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] The Code License is included in the repository.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The procedure used to produce the different versions of the CMNIST dataset is accessible from the released code.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] The MNIST dataset is publicly available.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]