Universität Tübingen

Masterarbeit im Fach Kognitionswissenschaft

Domain Adaptation under Causal Assumptions

Tosca Lechner

supervised by

Prof. Ulrike von Luxburg and

Prof. Ruth Urner

29. Oktober 2018

Erklärung

Hiermit erkläre ich, dass ich diese schriftliche Abschlussarbeit selbstständig verfasst habe, keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe und alle wörtlich oder sinngemäß aus anderen Werken übernommenen Aussagen als solche gekennzeichnet habe.

Ort, Datum Unterschrift

Acknowledgements

First and foremost, I would like to thank Ruth Urner, the main supervisor of my thesis, for many helpful discussions, providing support, encouragement and sometimes caffeinated drinks throughout the process. I was lucky to have an advisor who always encouraged me to ask questions and was willing to discuss them with me. I would also like to thank her for inviting me to Toronto two times to work on my thesis and helping to make my life there easier and more enjoyable.

I would also like to thank Ulrike von Luxburg for agreeing to be my supervisor at the University of Tübingen. In particular, I would like to thank her for her flexibility and being helpful, whenever I needed it, as well as motivating me to be part of her reading group.

I would also like to thank Bernhard Schölkopf for providing the topic of this thesis, helpful discussions and supporting me by employing me as a research assistant at the MPI for Intelligent Systems in Tübingen.

I would also like to thank Shai Ben-David for helpful discussions.

I would like to thank David Lorch for in general being awesome and supportive. In particular, I am thankful for his support throughout writing my thesis, be it by answering my LaTeX-related questions or applying his cooking skills. I would also like to thank him for proofreading my thesis and finding better homes for many of my commas.

Last but not least I would like to thank my friends and family for being supportive and making my life more enjoyable. In particular, I would like to thank my parents, who have always supported me in my endeavors.

Contents

1. Introduction

Machine Learning is a field in computer science where statistical techniques are used to learn from training data in order to perform well on a task. Due to more availability of data and faster processing tools, this research discipline has gotten more and more attention in recent years. From the first image classifiers that are comparable to human level to text processing systems that are able to extract main sentiments of a tweet, the development in this area has been quite astonishing. Outside of research, more of these systems are applied now than ever – be it algorithms automatically translating between two given natural languages, as Google Translate does, or the development of the first self-driving cars. This warrants questions about the security and reliability of these systems. To give performance guarantees about Machine Learning algorithms, we turn to a theoretical, mathematical rigorous approach.

The discipline in Machine Learning that is possibly best understood in theory is supervised learning. In supervised learning, the task is to predict correct labels for feature vectors, after having seen training data consisting of features and labels. One example for this task is labeling hand written figures correctly to the numbers from 0 to 9, after having seen several correctly labeled instances. A learner $\mathcal{A}$ will choose a hypothesis, that is, a function mapping a feature vector to a label. This can be interpreted as a prediction for the correct label, given the feature vector. We usually assume our training data and test data to come from the same distribution. Under this assumption, there have been results that relate the *empirical risk* of a hypothesis, i.e., a measure for the difference between the correct labeling and the labeling predicted by the hypothesis averaged over all elements of the training data, and the *true risk*, that is, the expected value of this difference on new data (e.g., the test data). Statistical learning theory introduces measures for the complexity of a hypothesis class $\mathcal{H}$ like the *VC-dimension* and the *Rademacher Complexity*. It then shows that for low complexities of $\mathcal{H}$ the difference between the empirical and the true risk will *uniformly converge* to zero with the number of training data points approaching infinity. Under the realizability assumption, which states that the true labeling is an element of $\mathcal{H}$, this implies that for a given ε one can bound the data points needed for the learner $\mathcal{A}$ to output a classifier, in such a way that the true risk of that classifier with high probability does not exceed ε. In particular, this upper bound for the sample complexity is independent of the marginal distribution of the feature vectors, as long as the marginal distribution of training and test data stays the same. Furthermore, the *No-Free-Lunch Theorem* tells us that without restricting the complexity of a hypothesis class we do not get finite sample bounds. In other words,

we need prior knowledge of our labeling data, in order to get meaningful finite sample bounds.

However, in reality we often find that this assumption about training and test data being distributed identically is violated. For example one can imagine an image recognition algorithm being fed training data from one photographer, who might use certain camera setup for all their pictures, while the test data consists of photos taken with a greater variety of setups. Another, possibly more relevant, example would come from different lightings or weather conditions of a scene. Imagine a self driving-car that only learned to drive by daylight on unclouded summer days. We would like this car to be able to drive without accidents in rain as well.

Therefore, we are interested in guarantees about learnability for settings where a domain shift happens, i.e., a shift in distributions between training data and test data. This problem is known in Machine Learning as the *Domain Adaptation* problem, where we assume that our training data is generated by some source distribution and the test data (or rather the data our learned hypothesis will be used for) comes from some (different) target domain.

There has been some theoretical work examining what kinds of assumptions about the distribution shift help to get learning guarantees in this setting (e.g.,Ben-David et al. (2006, 2012, 2010b), Mansour et al. (2009), Ben-David and Urner (2012)). Some of these works have introduced concepts like the $\mathcal{H}$-*divergence* and the *discrepancy distance*, which measure the similarity between source and target distributions with respect to the supervised learning problem one tries to solve. While these concepts lead to some positive results – that is if source and target distributions are similar in these measures, we obtain learning guarantees – we often do not have a good intuition about these quantities in practice, as they often do not reflect our prior knowledge of the learning problem.

Other concepts like *covariate shift*, i.e., only the marginal distribution of the features changes while the labeling/regression function stays the same in both source and target domains, are often better motivated in practice: Going back to our image recognition example, we can assume that the (true) labeling of an object should not change dependent on the light conditions in which the photograph was taken. Unfortunately, there are some theoretical results Ben-David et al. (2010b), Ben-David and Urner (2012) that show that covariate shift alone is not sufficient to get guarantees about domain adaptation learnability. However, the counter examples given in these papers seem quite artificial and the resulting impossibility results do not seem to reflect the success of domain adaptation algorithms in reality. Therefore, it is likely that there still exist other criteria – criteria that are often fulfilled in practice, but not by these counter examples and that make domain adaptation learning possible. One such criterion might consist of an assumption about how the label space and the feature space relate in terms of causality.

Causality and its impact on Machine Learning has gotten increasing attention in recent years. There has been an empirical study (Schölkopf et al. (2013)), dividing data sets by causal direction, that is if the features caused the labels, then a data set is said to be

causal and if the labels caused the features then they are called *anti-causal*. Data-sets for which the authors believed, that the cause of the correlation between features and labels was neither the features nor the labels, but a third variable, were called *confounded*. The study used several semi-supervised learning algorithms and compared them to supervised methods. The results of this study suggested that semi-supervised learning only improves corresponding supervised learning methods for anti-causal and confounded data sets. Some theoretical results (Janzing and Schölkopf (2015)) have provided proofs for this hypothesis in certain causal scenarios, like the *Information Geometric Causal Inference* (IGCI) model. There has also been some work that suggests, that assumptions about the causal structures also help for transfer learning and domain adaptation, giving a generalization of covariate shift(Rojas-Carulla et al. (2018)). However, this work assumes there to be several source domains instead of only one source domain. In contrast to this work, this thesis will mainly focus on the original domain adaptation setting, with only one source domain.

The main question this thesis tries to answer is whether assumptions about the underlying causal structure of the data can help to overcome existing lower bounds for DA learning. In particular, we will examine the *Principle of Independence of Cause and Mechanism* as a criterion for causality.

In practice we often know more about the underlying causal structure of a problem, i.e., whether the labels cause the features or vice versa, than we know about the $\mathcal{H}$-divergence of a distribution shift. Furthermore, if a change in distribution happens from source to target, it is likely that the underlying causal structure stays the same. Additionally, in the counter examples from Ben-David et al. (2010b), Ben-David and Urner (2012), the labeling and the distribution of the features seem to be constructed dependently. Therefore this construction might violate the Principle of Independence of Cause and Mechanism.

We will look at several formalizations of the Principle of Independence of Cause and Mechanism and their use for domain adaptation. We will mostly focus on the Information Geometric Causal Inference (IGCI) model as introduced in Janzing et al. (2012). We will give several attempts to adapt this criterion for binary classification and investigate its use for domain adaptation. Indeed, we can show that these criteria are violated by the counter example for Domain Adaptation (DA) learnability given in Ben-David and Urner (2012). However we will also show that there are similar problem sets that do fulfill our IGCI-model inspired criteria and that also serve as counter examples for DA-learnability.

In Chapter 2 we will briefly introduce the main results from Learning Theory for supervised learning. In Chapter 3 we will give a short summary of several formalizations of causality, in particular we will discuss formalizations of the Principle of Independence of Cause and Mechanism. In Chapter 4 we will then give a summary and discussion of existing results for domain adaptation. In particular we will discuss a lower bound from Ben-David and Urner (2012). In Chapter 5 we will present our results about DA

learnability under causal assumptions. Finally Chapter 6 gives a short discussion of these results.

2. Foundations of Learning Theory

In this chapter we will introduce the basic concepts of learning theory with respect to supervised learning. This includes a formal introduction of the setting, of the definitions of PAC learnability and uniform convergence. In the end of this chapter we will state the main theorem of the Vapnik-Chervonkevis theory. The definitions and theorems given in this chapter will mostly be taken from Shalev-Shwartz and Ben-David (2014). This chapter is meant as a brief summery of these concepts. For a more extensive introduction as well as for proofs we would refer the reader to Shalev-Shwartz and Ben-David (2014).

In supervised learning, a learner receives a sample S of pairs (x_i, y_i) of feature vectors (or instances) x_i and labels y_i and tries to predict the label of new unlabeled instances x_j.
To make this definition more formal, let $\mathcal{X}$ be the feature domain and $\mathcal{Y}$ be the label domain. A learner $\mathcal{A} : \bigcup_{n=1}^{\infty}(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}$ is a function taking a finite sample S of pairs (x_i, y_i) with $x_i \in \mathcal{X}_i$ and $y_i \in \mathcal{Y}$ as input and outputting functions from $\mathcal{X}$ to $\mathcal{Y}$.

For supervised learning, we will assume that the pairs (x_i, y_i) are identically and independently drawn from some distribution P over $\mathcal{X} \times \mathcal{Y}$. We will refer to the marginal distribution over the feature vectors as $\mathcal{D}$, i.e., $\mathcal{D}(A) = P(A, \{0\}) + P(A, \{1\})$ for all $A \subset \mathcal{X}$. Furthermore, for the case of binary classification, i.e. $\mathcal{Y} = \{0, 1\}$, the *labeling function* $f : \mathcal{X} \to [0, 1]$ for a joint distribution P is defined by $f(x) = P(y_i = 1 | x_i = x)$, where $P(y_i = 1 | x_i = x)$ denotes the conditional probability of the label 1 given the feature x. A labeling function is said to be *deterministic* if it only takes values in $\{0, 1\}$.

A good learner will be a learner that outputs a function $h \in \mathcal{Y}^{\mathcal{X}}$ such that for a new unlabeled data point x_j, the value $h(x_j)$ will likely be a "good prediction" for the corresponding (unknown) label y_j. In order to judge whether a prediction is good or not, we will need the concept of a *loss function.*

Definition 1 (Loss function)**.** *A loss function is a function $l : \mathcal{Y}^{\mathcal{X}} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ that takes as input a function h from the feature space $\mathcal{X}$ to the label space $\mathcal{Y}$ and a data point (x, y) and outputs a positive real value, rating the prediction $h(x)$ for the label y. The smaller the value $l(h, x, y)$, the better the prediction.*

Dependent on the label space $\mathcal{Y}$ and the underlying problem some loss functions are preferable to others. For binary classification, the common choice of loss function is the *0-1-loss*:

Definition 2 (0-1 loss)**.** *The* 0-1-loss *is the loss function* $l : \mathcal{Y}^{\mathcal{X}} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$*, defined by* $l(h, x, y) = \mathbb{1}[h(x) \neq y]$*, where* $\mathbb{1}$ *is the indicator function.*

For regression (i.e., if we have $\mathcal{Y} = \mathbb{R}$) we can still use the 0-1-loss, but then all false predictions will be judged to be equally bad. However, in most use-cases, a (false) prediction $h(x_i)$ that is close to y_i, will be viewed as better than a prediction $h'(x_i)$ that is further away from y_i. This judgment is reflected in the ℓ^2-loss which is most commonly used in regression.

Definition 3 (ℓ^2-loss)**.** *Let* $\mathcal{Y}$ *be a space, where the* ℓ^2*-norm* $\|\cdot\|_2$ *is defined. Then the* ℓ^2*-loss is defined as loss function* $l : \mathcal{Y}^{\mathcal{X}} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ *with* $l(h, x, y) = \|h(x) - y\|_2^2$*.*

Note that for $\mathcal{Y} = \{0, 1\}$,the ℓ^2-loss is defined, but equivalent to the 0-1-loss. However, for classification with multiple labels the ℓ^2-loss is not necessarily defined.
Now if we wish to select a good hypothesis, we would like to take one that has low expected loss for the underlying probability distribution. This leads us to the next definition.

Definition 4 (Risk function)**.** *For a given loss-function* l *let* $\mathcal{L}$ *be the risk function with respect to a probability distribution* P*, defined as the expected value of* l *over pairs* (x, y)*, i.e.,* $\mathcal{L}_P(h) = \mathbb{E}_{(x,y)\sim P}[l(h, x, y)]$*. Sometimes instead of regarding the data-generating process as given by* P*, we will view it as given by the marginal distribution* $\mathcal{D}$ *and the labeling function* f*. In that case we will sometimes refer to* $\mathcal{L}_P(h)$ *by* $\mathcal{L}_{(\mathcal{D},f)}$*.*

Accordingly, we will define the *empirical risk* for a hypothesis h as the average of all loss terms $l(h, x_i, y_i)$ with sample points (x_i, y_i).

Definition 5 (Empirical risk)**.** *For a given loss function* l *and a sample* $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$*, the empirical loss* $\mathcal{L}_S(h)$ *of a hypothesis* h *is defined as*

$$\mathcal{L}_S(h) = \frac{1}{n} \sum_{i=1}^{n} l(h, x_i, y_i)$$

Our hope will be that the empirical risk $\mathcal{L}_S(h)$ is a good estimate for the true risk $\mathcal{L}_P(h)$. Furthermore, a learner $\mathcal{A}$ is called an *empirical risk minimizer* for the hypothesis class $\mathcal{H}$, if $\mathcal{A}(S) \in \arg\min_{h\in\mathcal{H}} \mathcal{L}_S(h)$.

The next theorem will show us that for large domains $\mathcal{X}$ (in particular infinite ones) any learner $\mathcal{A}$ (empirical risk minimizers included) the true risk can still be high (if the size of the training set is not proportional to the domain size), if we do not restrict the hypothesis class $\mathcal{H}$ from which $\mathcal{A}$ selects, i.e. $\mathcal{H} = 2^{\mathcal{X}}$.

Theorem 1 (No-Free-Lunch Theorem, Theorem 5.1 from Shalev-Shwartz and Ben-David (2014))**.** *Let $\mathcal{A}$ be any learning algorithm for the task of binary classification with respect to the* 0-1 *loss over a domain $\mathcal{X}$. Let m be any number smaller than $\frac{|\mathcal{X}|}{2}$ representing a training set size. Then there exists a distribution P over $\mathcal{X} \times \{0,1\}$ such that:*

1. *There exists a function $g : \mathcal{X} \to \{0,1\}$ with $\mathcal{L}_P(g) = 0$.*
2. *With probability of at least $\frac{1}{7}$ over the choice of $S \sim P^m$ we have $\mathcal{L}_P(\mathcal{A}(S)) \geq \frac{1}{8}$.*

We therefore see, that in general – i.e., for infinite domains $\mathcal{X}$ – an empirical risk minimizer (or any other learner) will not be able to approximate the true labeling function correctly, unless we make further assumptions about $\mathcal{H}$. We will now characterize the learnability of a problem by the hypothesis class $\mathcal{H}$ of possible outputs of the learner $\mathcal{A}$. We will first look at (PAC) learnability in the *realizable* case.

Definition 6 (Realizability)**.** *Given a hypothesis class $\mathcal{H}$ and a distribution P, we will denote the* optimal hypothesis *of $\mathcal{H}$ with respect to P as $opt_P(\mathcal{H}) = \inf_{h \in \mathcal{H}} \mathcal{L}_P(h)$. We say that a learning problem is* realizable*, if $opt_P(\mathcal{H}) = 0$. In particular, if there is a $h \in \mathcal{H}$, with $h(x) = f(x)$ for all $x \in \mathcal{X}$, we do have realizability.*

Definition 7 (Probably Approximately Correct (PAC) Learnability, Definition 3.1 from Shalev-Shwartz and Ben-David (2014))**.** *A hypothesis class $\mathcal{H}$ is* Probably Approximately Correct *(PAC) learnable if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\varepsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \to \{0,1\}$, if the realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$ then when running the learning algorithm on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d. samples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of of samples), $\mathcal{L}_{(\mathcal{D},f)}(h) \leq \varepsilon$.*

With this definition, we obtain a corollary about PAC learnability from the No-Free-Lunch Theorem.

Corollary 1 (Corollary 5.2 from Shalev-Shwartz and Ben-David (2014))**.** *Let $\mathcal{X}$ be an infinite domain set and let $\mathcal{H}$ be the set of all functions form $\mathcal{X}$ to $\{0,1\}$. Then $\mathcal{H}$ is not PAC learnable.*

We therefore see that, in order for $\mathcal{H}$ to be PAC learnable, its complexity needs to be restricted. We will now introduce the concept of *Vapnik-Chervonenkis dimension*, which quantifies this complexity and, as we will later see, fully characterizes the PAC-learnability of a hypothesis class.

To that aim, we first need to define shattering. For the following definition let $\mathcal{H}_C$ be the restriction of a hypothesis class $\mathcal{H}$ to a subset $C \subset \mathcal{X}$.

Definition 8 (Shattering, from Shalev-Shwartz and Ben-David (2014))**.** *A hypothesis class $\mathcal{H}$* shatters *a finite set $C \subset \mathcal{X}$ if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0,1\}$. That is, $|\mathcal{H}_C| = 2^{|C|}$.*

Now we can define the VC-dimension of a hypothesis class $\mathcal{H}$ as the maximal size of a subset $C \subset \mathcal{X}$ that can be shattered.

Definition 9 (VC-Dimension, from Shalev-Shwartz and Ben-David (2014))**.** *The VC-dimension of a hypothesis class $\mathcal{H}$, denoted $VC(\mathcal{H})$, is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathcal{H}$ has infinite VC-dimension.*

As we will see, having a finite VC-dimension will imply PAC learnability. But before we state the corresponding theorem, we will first introduce the concept of agnostic PAC learnability — a version of PAC learnability, that does not require realizability.

Definition 10 (Agnostic PAC Learnability, Definition 3.3 from Shalev-Shwartz and Ben-David (2014))**.** *A hypothesis class $\mathcal{H}$ is* agnostic PAC learnable *if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\varepsilon, \delta \in (0,1)$ and for every distribution P over $\mathcal{X} \times \mathcal{Y}$, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d. samples generated by P, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the m training samples),*

$$\mathcal{L}_P(h) \leq \min_{h' \in \mathcal{H}} \mathcal{L}_P(h') + \varepsilon$$

Finally, we will introduce the concepts of uniform convergence of a hypothesis class, which will link the empirical risk $\mathcal{L}_S$ with the true risk $\mathcal{L}_P$. If we know these two terms to be close, this gives us a result about the true risk of empirical risk minimizers.

Definition 11 (Uniform convergence, Definition 4.3 from Shalev-Shwartz and Ben-David (2014))**.** *We say that a hypothesis class $\mathcal{H}$ has the uniform convergence property (w.r.t. a domain Z and a loss function l) if there exists a function $m_{\mathcal{H}}^{UC} : (0,1) \to \mathbb{N}$ such that for every $\varepsilon, \delta \in (0,1)$ and for every probability distribution P over $\mathcal{X} \times \{0,1\}$, if S is a sample of $m \geq m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ examples drawn i.i.d. according to P, then, with probability of at least $1 - \delta$ we get*

$$\text{for all } h \in \mathcal{H}, |\mathcal{L}_S(h) - \mathcal{L}_P(h)| \leq \varepsilon$$

Finally we can state the fundamental theorem of statistical learning as provided in Shalev-Shwartz and Ben-David (2014), which shows the equivalence of PAC-learnability, uniform convergence and finite VC-dimension of a hypothesis class.

Theorem 2 (The Fundamental Theorem of Statistical Learning, Theorem 6.7 from Shalev-Shwartz and Ben-David (2014))**.** *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the 0-1 loss. Then the following are equivalent:*

1. *$\mathcal{H}$ has the uniform convergence property.*
2. *Any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$.*

3. *$\mathcal{H}$ is agnostic PAC learnable.*

4. *$\mathcal{H}$ is PAC learnable.*

5. *Any ERM rule is a successful PAC learner for $\mathcal{H}$.*

6. *$\mathcal{H}$ has a finite VC-dimension.*

Another version of this theorem (also taken from Shalev-Shwartz and Ben-David (2014)) gives explicit bounds for the sample complexities (i.e. a bound for the number of training data one needs to see in order to have learned a task up to a given error).

Theorem 3 (The Fundamental Theorem of Statistical Learning – Quantitative Version, Theorem 6.8 from Shalev-Shwartz and Ben-David (2014))**.** *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the 0-1 loss. Assume that $VC(\mathcal{H}) = d < \infty$. Then there are absolute constants C_1, C_2 such that:*

1. *$\mathcal{H}$ has the uniform convergence property with sample complexity*

$$C_1 \frac{d + \log(\frac{1}{\delta})}{\varepsilon^2} \leq m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \leq C_2 \frac{d + \log(\frac{1}{\delta})}{\varepsilon^2}$$

2. *$\mathcal{H}$ is agnostic PAC learnable with sample complexity*

$$C_1 \frac{d + \log(\frac{1}{\delta})}{\varepsilon^2} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d + \log(\frac{1}{\delta})}{\varepsilon^2}$$

3. *$\mathcal{H}$ is PAC learnable with sample complexity*

$$C_1 \frac{d + \log(\frac{1}{\delta})}{\varepsilon} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d \log(\frac{1}{\varepsilon}) + \log(\frac{1}{\delta})}{\varepsilon}$$

3. Causality

In this chapter we will introduce some ideas that have been proposed for the formalization of causality. In practice, we are often not only interested in the correlation between two variables but in the underlying causal structure – which often gives us a better idea of how an action might influence our situation. For example, a doctor would be interested in knowing if prescribing a drug actually causes patients to get healthy rather than just knowing if taking the drug and getting healthy just happen to co-occur. However, most statistical analysis has focused on correlational statements. The focus on correlational statements is likely due to the fact that correlational statements are easier to evaluate. If we only have observational data our statistical test can only infer statements about the correlation of two variables. But, as one often hears in introductory statistics courses, correlation does not imply causation. Indeed in our previous example the events "taking prescription drug" and "getting healthy" might have a joint probability distribution with high correlation, if some (maybe more health-conscious) people are more likely than others to get healthy soon, but also more likely to take a prescription drug. From just observational data, one might even get the same probability distribution as in the case where taking the drug causes the person to get healthy again. Thus, if we were to infer statements about the underlying causality, we would need to intervene on some variables and observe the change in other variables. A complete model of causality therefore needs to not only make statements about the statistical properties of its variables, but also about possible changes of variables under interventions.

In this section we will introduce several models of causality. First we will give a short introduction to *Structural Equation Models (SCMs*, since they do account for interventions. We will introduce the concept of the *identifiability* of the causal direction (for observational data. We will then introduce *Additive Noise Models (ANMs)* as an example of a restricted SCMs, were we sometimes do get identifiability. We then go on to introduce the *Principle of Independence of Cause and Mechanism* and discuss several formalizations of this principle, in particular, *algorithmically independent conditionals* and the *Information Geometric Causal Inference (IGCI) model*. This chapter closely follows Peters et al. (2017).

3.1. Structural Causal Models

One type of model that reflects the effect of interventions well, are the so-called structural equation models. They consist of a directed graph with random variables as vertices. An

arrow between two vertices indicates that the tail of this arrow is a direct cause for the head of the arrow. In the case of observational data, this graph can be interpreted similarly to a Bayes net, i.e., as a representation of factorization of the underlying joint probability distribution. Furthermore, the model offers a description of how the distribution would change if the distribution of one or several of the model's vertices changed due to some intervention.

We will now give an abbreviated definition of structural equation models that only takes into account the causal direction between two variables (ignoring the possibility of confounders). This will suffice for our purposes, since we will only regard simple cause-effect relations between feature vectors and labels in this thesis[1]. For a more complete introduction to SCMs, we refer the reader to Peters et al. (2017) or Pearl (2009).

Definition 12 (Structural Causal Models)**.** *Let C and E random variables, where C is the cause and E the effect. A structural causal model for C and E consists of a directed graph $\mathcal{G} = (\{C, E\}, \{A_{C\to E}\})$, where $A_{C\to E}$ is a directed edge between vertices C and E, and of two assignments*

$$C := N_C$$

$$E := f_E(C, N_E),$$

where N_E and N_C are two independently distributed random variables and f_E is a function from the domains of C and N_E to the domain of E. Thus, the SCM defines a joint probability distribution for (C, E). Furthermore, the SCM tells us that under an intervention (i.e., a direct change of one of the graphs variables), the other variables still need to be consistent with the assignments. More precisely, if an intervention consists of a replacement of C by some C', E becomes $E' := f_E(C', N_E)$. Whereas, if an intervention consists of a replacement of E by some E'', the corresponding C'' is still defined by $C'' = N_C$. In a probability statement, we will write "$X = x$" for an observed event x and "do $X = x$" for an instance, were we intervened on X in order to be x.

An intervention in a structural causal model works by replacing a random variable X in the causal model by a random variable X' and replacing the term X in the calculation of all descendants of X by X'. This way, if X is a cause of Y, the change of X will affect Y, but if it is not a cause of Y, it will be unaffected by the change in X. The probability of $Y = y$ conditioned on actively setting the value for a variable in the causal model X to a value x will be denoted by $P(Y = y|\text{ do } X = x)$. If Y is a cause of X, then this conditional probability will usually be different from the conditional probability we get by observational data, i.e. $P(Y = y|X = x) \neq P(Y = y|\text{ do } X = x)$. This is due to the fact, that the distribution of Y will be independent of X after an intervention.

[1] Which is not to say that more complex causal models between feature vectors and labels are not realistic or possibly relevant for domain adaptation learning.

3.2. Causal and Anti-Causal Directions

In subsequent chapters we will distinguish between the causal (i.e. the features X cause the label Y) and the anti-causal (i.e. the label Y causes the features X) direction, to investigate whether they make a difference for the learnability of a problem. Therefore we have to make this distinction formally. However, as we already mentioned and as we will show now, this distinction cannot always be made if we only observe the joint distribution $P_{C,E}$.

Proposition 1 (Non-uniqueness of graph structures, Proposition 4.1 from Peters et al. (2017))**.** *For every joint distribution $P_{X,Y}$ of two real-valued variables there is an SCM*

$$Y = f_Y(X, N_Y), \qquad X \perp\!\!\!\perp N_Y,$$

where f_Y is a measurable function and N_Y is a real-valued noise variable.

Proof. Analogous to Peters et al. (2017) for a given joint distribution $P_{X,Y}$, define the conditional cumulative distribution function

$$F_{Y|x}(y) := P(Y \leq y | X = x)$$

and then

$$f_Y(x, n_Y) := \inf_y \{F_{Y|x}(y) \mid F_{Y|x}(y) \geq n_Y\}$$

Furthermore let N_Y be uniformly distributed on $[0, 1]$ and independent of X. We can see that the resulting SCM is consistent with $P_{X,Y}$. □

For a given joint distribution $P_{X,Y}$, we can therefore define both an SCM, where X looks like the cause and Y like the effect and an SCM, where Y looks like the cause and X like the effect. Thus we see that the notation of SCMs itself is not sufficient to distinguish causal directions (without the possibility of interventions). Therefore we will need to make more assumptions about our causal model to be able to distinguish causal directions. One of these assumptions can be restrictions on the class of possible mechanisms $f_Y(X, N_Y)$ and noise-models.

3.2.1. Additive Noise Models

A typical way to restrict SCMs is by allowing only additive noise. This gives rise to so-called additive noise models (ANMs).

Definition 13 (Additive noise models, Definition 4.4 from Peters et al. (2017))**.** *The joint distribution $P_{X,Y}$ is said to admit an ANM from X to Y if there is a measurable function f_Y and a noise variable N_Y such that*

$$Y = f_Y(X) + N_Y, \qquad N_Y \perp\!\!\!\perp X \ .$$

These models are often realistic for physical data, since measurements are often subject to sensory noise, which can be seen as influencing the result of Y independent of the result of $f_Y(X)$, and therefore can be seen as additive. Often the noise variable N_Y will be further restricted to be gaussian (e.g. in Rojas-Carulla et al. (2018)).[2] However this might still not lead to an identifiable causal structure, as the next proposition will show. The next proposition will show an example of an ANM that is not identifiable. The claim of this proposition is a special case of Theorem 4.2 of Peters et al. (2017), which we will later introduce as Theorem 4.

Proposition 2 (Non-identifiability of linear ANMs with Gaussian noise)**.** *Let X and Y be real-valued random variables, such that $P_{X,Y}$ admits a linear ANM with gaussian noise from X to Y, i.e., there are $\alpha, \beta \in \mathbb{R}$ such that*

$$Y = \alpha X + \beta + N_Y,$$

with $X \perp\!\!\!\perp N_Y$ and $N_Y \sim \mathcal{N}(0, \sigma_N^2)$ for some $\sigma \in \mathbb{R}$. Furthermore let X be normally distributed.
Then $P_{X,Y}$ also admits a linear ANM with gaussian noise from Y to X

Proof. If $X \sim \mathcal{N}(\mu_X, \sigma_X)$ and $N_Y \sim \mathcal{N}(0, \sim_N^2)$, then $Y = \alpha_1 X + \beta_1 + N_Y$ is also normally distributed with $Y \sim \mathcal{N}(\alpha_1 \mu_X + \beta_1, \alpha_1^2 \sigma_X^2 + \sigma_N^2)$. Now if

$$\frac{1}{(1 - \frac{\sigma_X^2}{\sigma_{N_Y}^2})^2 (1 + \frac{\sigma_X}{\sigma_{N_Y}})} > \alpha_1^2,$$

we can define the random variable Z as follows

$$Z := \alpha_2 X + \beta_2 + N_Z$$

with

$$\alpha_2 := \frac{\alpha_1 \sigma_X^2 - \sigma_{N_Y}^2}{\alpha_1^2 \sigma_X^2},$$

$$\beta_2 := -\alpha_2 \alpha_1 \mu_X - \alpha_2 \beta_1$$

and $N_Z \sim \mathcal{N}(0, \sigma_{N_Z}^2)$ with $N_Z \perp\!\!\!\perp Y$ and $\sigma_{N_Z}^2 = (1 - \alpha_2 \alpha_1^2)\sigma_X^2 - \alpha_2 \sigma_{N_Y}$.
With this definition we obtain $Z \sim \mathcal{N}(\mu_X, \sigma_X)$ and $Cov[X, Y] = Cov[Z, Y]$. Therefore the joint distributions $P(X, Y)$ and $P(Z, Y)$ are the same. Since $P(Z, Y)$ admits a linear ANM with gaussian noise from Y to Z, $P(X, Y)$ also admits a linear ANM with gaussian noise from Y to X.

□

[2] This can be motivated by the fact that sensory noise is often caused by many small independent fluctuations influencing Y in an additive and independent way. Therefore N_Y can be seen as a sum of many independently distributed random variables, which by the Central Limit Theorem is approximately gaussian distributed.

For general linear ANMs, however, we can achieve identifiability, if N_Y is not Gaussian, as is stated in the next theorem.

Theorem 4 (Identifiability of linear non-Gaussian models, Theorem 4.2. from Peters et al. (2017)). *Assume that $P_{X,Y}$ admits the linear model*

$$Y = \alpha X + N_Y, N_Y \perp\!\!\!\perp X$$

with continuous random variables X, N_Y, and Y. Then there exists $\beta \in \mathbb{R}$ and a random variable N_X such that

$$X = \beta Y + N_X, N_X \perp\!\!\!\perp Y$$

if and only if N_Y and X are Gaussian.

For more general ANMs we get the following theorem.

Theorem 5 (Identifyability of ANMs, Theorem 4.5 from Peters et al. (2017)). *For the purpose of this theorem, let us call an ANM* smooth *if N_Y and X have strictly positive densities p_{N_Y} and p_X and if f_Y, p_{N_Y} and p_Y are three times differentiable. Furthermore let $\mathcal{X}$ be the domain of X and $\mathcal{Y}$ be the domain of Y.*
Assume that $P_{Y|X}$ admits a smooth ANM from X to Y, and that there exists a $y \in \mathbb{R}$ such that

$$(\log p_{N_Y})''(y - f_Y(x)) f_Y'(x) \neq 0$$

for all but countably many values x. Then, the set of log densities $\log p_X$ for which the obtained joint distribution $P_{X,Y}$ admits a smooth ANM from Y to X is contained in a 3-dimensional affine space of the otherwise infinitely dimensional solution space $\mathbb{R}_+^{\mathcal{X}}$ for log densities $\log p_X$.

This theorem implies that in the infinite dimensional space $\mathbb{R}_+^{\mathcal{X}}$ in which log densities $\log p_X$ for random variables in $\mathcal{X}$ lie, only log densities $\log p_X$ in a three dimensional subset leads to a joint distribution $P_{X,Y}$ that permit ANMs in both directions. We therefore see that in most cases ANMs are a good and realistic restriction for a causal model, and can make the causal direction identifiable. Therefore having a learning theory based on these models might lead to results. However, we have also seen cases where ANMs cannot distinguish between the two causal directions. A causal learning theory based on ANMs must therefore be subject to further restrictions or can only make one-directional conclusions of one of the two forms:

- Learning works in causal/anti-causal direction (but that implies that learning also works in some anti-causal/causal cases)
- Learning does not work in causal/anti-causal direction (but that implies that learning also does not work in some anti-causal/causal cases)

3.2.2. Independence of Cause and Mechanism

Another approach for determining the causal direction of a joint distribution (without using interventions) is known as the "Independence of Cause and Mechanism", which states that the distribution of the cause P_C should be independent of the mechanism f_E that produces the cause from the effect. The intuitive argument for this is that if we were to intervene on the cause, the mechanism should stay the same (or change independently). Note, that "independence" here does not necessarily mean statistical independence, since the argument is also supposed to work for a deterministic mechanism. By knowing the mechanism, we should not be able to infer anything about the distribution of the cause – and vice versa. In Peters et al. (2017) this principle was framed in the following way.

Principle 1 (Independence of Cause and Mechanism, Principle 2.1 from Peters et al. (2017))**.** *The causal generative process of a system's variable is composed of autonomous modules, that do not inform or influence each other.*
In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanisms) does not inform other conditional distributions. Whenever there are only two variables, this reduces to an independence between the cause distribution and the mechanism producing the effect distribution.

There have been several attempts to formalize this intuition, however, there has not yet been an agreement about the correct formalization in the scientific community. Below, we will review some of these approaches.

Minimal description length

One approach, inspired by algorithmic information theory, is to say that the minimal description length of P_C should not change with the knowledge of f_E. To make this definition formal, we need to introduce Kolmogorov complexity.

Definition 14 (Kolmogorov complexity)**.** *Let T be a universal turing machine. For any binary string s, we define the* Kolmogorov *complexity $K_T(s)$ as the length of the shortest program, denoted by s^*, for which T outputs s and then stops. Furthermore s^* is called the shortest compression of s. Let $|\cdot|$ denote the number of digits of a binary word. Then,*

$$K_T(s) := |s^*|$$

Note that the Kolmogorov complexity for a given string is only defined with respect to a given Turing machine or description language. Dependent on the specifics of the Turing machine, the Kolmogorov complexity of a given string can vary a lot, i.e., for Turing machines T_1 and T_2 we can have $K_{T_1}(s) \neq K_{T_2}(s)$. However, according to the Invariance Theorem, for two given Turing Machines T_1 and T_2, there exists a constant c (only dependent on T_1 and T_2), such that $-c < K_{T_1}(s) - K_{T_2}(s) < c$ for all strings s. Therefore we can speak of the Kolmogorov complexity of s without specifying the Turing

machine (up to an additive constant). Another problem of the Kolmogorov complexity is that it is not computable.
Furthermore, we can define the *conditional Kolmogorov complexity* $K_T(s|t)$ of s given t, as the length of the shortest program for T, when run on the input string t, outputs s and then stops. We will now fix a Turing machine T and regard all Kolmogorov complexities with regard to T. Some of the following equations will only hold up to an additive constant (which is not dependent on the input arguments of the K and I, but only depend on T). This will be denoted by "$\stackrel{+}{=}$". With this we can now define the *mutual information* $I(s:t)$ of s and t:

$$I(s:t) := K(s) - K(s|t^*)$$

Note that we have conditioned over t^* instead of over t, since t^* is more valuable –after all t^* contains all information of t, but t does not necessarily contain all information of t^*. According to Peters et al. (2017), $K(s|t^*)$ shows closer analogies to the Shannon Entropy than $K(s|t)$.

If we now identify probability distributions P_C and $P_{E|C}$ by their density functions p_C and $p_{E|C}$, we can interpret $K(P_C)$ and $K(P_{E|C})$ as the minimal length of a program that encodes the functions p_C and $p_{C|E}$, respectively. Now we can formalize the statement of independence of cause and mechanism with respect to algorithmic independence.

Principle 2 (Algorithmically independent conditionals, Principle 4.13 from Peters et al. (2017))**.** *P_C and $P_{E|C}$ are algorithmically independent, that is,*

$$I(P_C : P_{E|C}) \stackrel{+}{=} 0$$

or equivalently,

$$K(P_{C,E}) \stackrel{+}{=} K(P_C) + K(P_{E|C}).$$

Since the Kolmogorov complexity is not computable and these equations only hold up to a constant, these equations and therefore the principle itself is hard to check.

3.2.3. Information Geometric Independence of Cause and Mechanism

Another idea of formalizing independence of cause and mechanism is given in Peters et al. (2017) as a correlational statement between the distribution of the cause and the mechanism.

Definition 15 (IGCI model, Definition 4.9 from Peters et al. (2017))**.** *Here P_{XY} is said to satisfy an IGCI model from X to Y if the following conditions hold: $Y = f(X)$ for some diffeomorphism f of $[0,1]$ that is strictly monotonic and satisfies $f(0) = 0$ and $f(1) = 1$. Moreover, P_X has the strictly positive continuous density p_X, such that the following "independence condition" holds:*

$$\mathrm{Cov}[\log f', p_X] = 0,$$

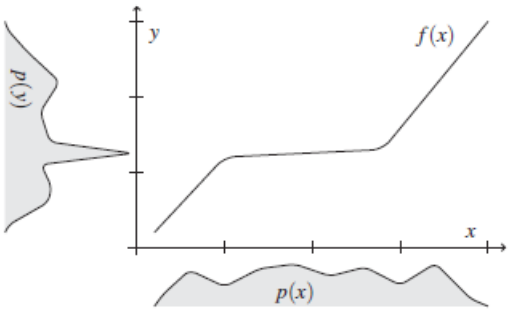


Figure 3.1.: (Figure 4.4 from Peters et al. (2017)) Visualization of the idea of the IGCI model: Peaks of p_Y tend to occur in regions where f has small slope and f^{-1} has large slope (provided that f has been chosen independently of p_X. Thus p_Y contains information about f^{-1}

where $\log f'$ and p_X are viewed as random variables on the probability space $[0,1]$ endowed with the uniform distribution.

The statement $\mathrm{Cov}[\log f', p_X]$ can be rewritten as $\mathrm{Cov}_{Z\sim\mathrm{Uni}([0,1])}[\log f'(Z), p_X(Z)]$. The idea behind this model is that if the distribution of the cause is approximately uniform and f' is approximately constant, then their variation away from the uniform distribution or the constant function, respectively, should be independent from each other, i.e. their covariance should be 0. However, if an effect Y is calculated by the cause as $Y := f(X)$, the corresponding covariance $Cov_{Z\sim Uni([0,1])}[\log g'(Z), p_Y(Z)]$ should be greater than 0, where $g = f^{-1}$, as will be stated in Theorem 6 . Figure 3.2.3 is an illustration from Peters et al. (2017) that shows an example of the distributions of X and Y and a mechanism for this model.

Theorem 6 (Identifiability of IGCI model, Theorem 4.10 from Peters et al. (2017))**.** *Assume the distribution $P_{X,Y}$ admits an IGCI model from X to Y. Then the inverse function f^{-1} satisfies*

$$Cov[\log f^{-1'}, p_Y] \geq 0$$

with equality if and only if f is the identity.

In Janzing and Schölkopf (2015) the same statement is made with $Cov_{Z\sim Uni([0,1])}[f'(Z), p_X(Z)]$ instead of $Cov_{Z\sim Uni([0,1])}[\log f'(Z), p_X(Z)]$. Later, we will need to formulate a similar covariance statement in cases where the derivative is not defined and the original IGCI model is therefore not applicable. We see the fact, that several versions of this statement are used in the literature as motivation to explore the general idea rather than using this specific formulation.
An attempt to give a more formal reasoning for this kind of model is given in Janzing et al. (2012). There, the authors justify this model by a particular kind of generating

process, where either f or some variation of f (e.g. f' or $\log f'$) is a piecewise constant function, with the value for each interval sampled independently from the same distribution. They then observe that

$$\begin{aligned} & \int h(x)P(x)dx - \int h(x)U(x)dx \\ = & \int h(x)U(x)\frac{P(x)}{U(x)}dx - \int h(x)U(x)dx \int U(x)\frac{P(x)}{U(x)}dx \\ = & Cov_{X\sim U}[h(X), \frac{U(X)}{P(X)}], \end{aligned}$$

where $U(X)$ is some reference distribution[3] for P. They then show that this expression is small, if we assume there is a generating process for h as described above.

Lemma 1 (Lemma 1 from Janzing et al. (2012))**.** *Let X, Y be real-valued. Let $r_j > 0$ with $j \in \mathbb{Z}$ be random numbers i.i.d drawn from a distribution $Q(r)$ with standard deviation σ_r. We then define a piecewise constant function h via $h(x) := r_j$ for $X \in [j, j+1)$. We then have for every $c > 0$:*

$$|\int h(x)P(x)dx - \int h(x)U(x)dx| \leq c\sigma_r \sqrt{\sum_j \left(\int_j^{j+1} P(x) - U(x)dx\right)^2}$$

with probability $1 - \frac{1}{c^2}$ or higher.

This implies that if $Uni([0,1])$ is a good reference distribution for P_X and $\log f'$ (or respectively, f') was generated as described above, $Cov_{Z\sim Uni([0,1])}[\log f'(Z), p_X(Z)]$ (or $Cov_{Z\sim Uni([0,1])}[\log f'(Z), p_X(Z)]$,respectively) is small with high probability in terms of the generation of $\log f'$ (or f' respectively).
Certainly most monotone causal process between $[0,1]$ will not follow this construction. For example, it could be the case that higher values for x also lead to higher values for $f'(x)$, as would be the case for the family of quadratic functions with positive leading coefficient. In this case one cannot assume f' to be created by a piecewise constant function whose values are i.i.d sampled from some distribution. First of all, quadratic functions are not piecewise constant, nor is their derivative or log derivative. But even if we assume that for every quadratic function f there is some f_{approx} serving as a good approximation for f, such that either f'_{approx} or $\log f'_{\text{approx}}$ are piecewise constant, the values $f'_{\text{approx}}(x) = r_j$ (or respectively $f'_{\text{approx}} = r_j$) for intervals $[\frac{j}{n}, \frac{j+1}{n})$ would not be independent of each other. Furthermore, they would not come from the same distribution, since r_j would be higher for higher j. It can, however still be argued that the covariance statement in the causal direction is likely smaller than in the anti-causal direction.
We might want to keep this generating process in mind, whenever we use the IGCI model as a criterion for causality or formulate our own criterion based on the IGCI model, since a different generating process would lead to a different covariance statement.

[3] A reference distribution can be thought of as a distribution that takes into account our prior knowledge of the problem. The later argument will use the fact that they believe $\int |P(x) - U(x)|dx$ to be small.

4. Domain Adaptation

Usually, in classification the assumption is that data from which we learn (i.e. training data) and data on which the algorithm should perform (i.e. test data) are i.i.d from the same distribution P and that the labeling function f^* is the same for both training and test data. This assumption is often violated in reality, since the data sets might be generated under different conditions. In image recognition for example we could get different marginals for data sets if the camera settings used differs between data sets, while we might still have the same labeling rule for both data sets independent of that change in settings.
In domain adaptation we will therefore not make this assumption. Instead, we will assume that the training data is identically and independently distributed according to some source distribution P_S over $\mathcal{X} \times \mathcal{Y}$, while the test data is i.i.d. according to some target distribution P_T over $\mathcal{X} \times \mathcal{Y}$. Furthermore, let $\mathcal{D}_S$ and $\mathcal{D}_T$ be the marginal distributions on $\mathcal{X}$ for P_S and P_T respectively. Furthermore, we will refer to the source labeling function as f_S^* and to the target labeling function as f_T^*.
In this chapter we will investigate which assumptions yield learnability guarantees in the domain adaptation setting.

First, we will give a formal definition of a *Domain Adaptation (DA) learner* and of *Domain Adaptation (DA) learnability*. We will then introduce some properties, like *covariate shift* and (small) $\mathcal{H}$*-divergence* that serve as common assumptions in domain adaptation scenarios. We will then state a positive result of DA-learnablity, that was given in Ben-David et al. (2010a), but only works for a very restricted set of assumptions. We will then provide an example of a way to adapt to a new domain by briefly introducing the *reweighting technique* from Mansour et al. (2009). We will then provide theorems from Ben-David et al. (2010b) and Ben-David and Urner (2012) that show the shortcomings of this and other techniques, for cases were covariate shift is fulfilled, but other criteria, like small $\mathcal{H}$-divergence are not met. The definitions and results presented in this chapter are primarily taken from Ben-David et al. (2010a, 2006, 2010b), Ben-David and Urner (2012), Mansour et al. (2009) and Urner (2013).

We will formalize the domain adaptation scenario as described above, by providing a definition for a *Domain Adaptation (DA) learner* and for *Domain Adaptation (DA) learnability*. In the following, if nothing else is stated, we will examine binary classification scenarios. We will look at the scenario, where our learner gets labeled data from *one* source domain as well as unlabeled data from the target domain as input and outputs a hypothesis that is supposed to perform well on the target domain.

Definition 16 (Domain Adaptation Learner, Definition 1 from Ben-David et al. (2010b))**.** *A* domain adaptation (DA) learner *is a function*

$$\mathcal{A} : \bigcup_{m=1}^{\infty} \bigcup_{n=1}^{\infty} (\mathcal{X} \times \{0,1\})^m \times \mathcal{X}^n \to \{0,1\}^{\mathcal{X}}$$

The performance of a DA-learner will again be measured with respect to the target distribution.

Definition 17 (Learnability of DA-learner, Definition 2 from Ben-David et al. (2010b))**.** *Let P_S, P_T be distributions over $\mathcal{X} \times \{0,1\}$, let $\mathcal{D}_S, \mathcal{D}_T$ their marginals on $\mathcal{X}$, $\mathcal{H}$ a hypothesis class, $\mathcal{A}$ a DA learner, $\varepsilon, \delta > 0$ and m, n positive integers. The learner $\mathcal{A}$ will be said to $(\varepsilon, \delta, m, n)$-learn P_T from P_S relative to $\mathcal{H}$, if when given access to a labeled sample L of size m, generated i.i.d. by P_S, and an unlabeled sample U of size n, generated i.i.d by $\mathcal{D}_T$, with probability at least $1 - \delta$ (over the choice of samples L and U), the learned classifier does not exceed the P_T-error of the best classifier in $\mathcal{H}$ by more than ε. In other words,*

$$\Pr_{L \sim P_S^m, U \sim \mathcal{D}_T^n} [\mathcal{L}_{P_T}(\mathcal{A}(L,U)) \leq \inf_{h \in \mathcal{H}} \mathcal{L}_{P_T}(h) + \varepsilon] \geq 1 - \delta$$

We will also say the learner $\mathcal{A}$ $(\varepsilon, \delta, m, n)$-solves the *DA-problem* for a class $\mathcal{W}$ of pairs (P_S, P_T) for hypothesis class $\mathcal{H}$, if $\mathcal{A}$ can $(\varepsilon, \delta, m, n)$-learn P_T from P_S relative to $\mathcal{H}$ for any pair $(P_S, P_T) \in \mathcal{W}$. Furthermore we will call a problem set $\mathcal{W}$, *DA-learnable* if there is a DA-learner $\mathcal{A}$ such that for every $\varepsilon, \delta > 0$ there exist $m, n \in \mathbb{N}$, such that $\mathcal{A}$ $(\varepsilon, \delta, m, n)$-solves the DA problem for $\mathcal{W}$ and $\mathcal{H}$.

In the following, we will often use L and U to denote the labeled data from the source domain and the unlabeled data from the target domain, respectively. The features (without the labels) of the source data will be denoted as $L_{\mathcal{X}}$.

4.1. Common Assumptions in Domain Adaptation

In this section we are going to introduce some assumptions that are often made in the domain adaptation setting. Some of these assumptions give learnability guarantees, as we will later see. First and foremost, all relevant assumptions we could make for a normal supervised learning problem – e.g. VC-dimension, realizability, deterministic labeling function – are still relevant (or become even more relevant) for domain adaptation. Additionally, the relation between target and source domain and the hypothesis class will be relevant.

One assumption for the labeling function that is often made in domain adaptation (Ben-David et al. (2010a),Ben-David et al. (2006),Ben-David and Urner (2012),Rojas-Carulla et al. (2018)) is covariate shift.

Definition 18 (Covariate shift, from Ben-David et al. (2010a))**.** *We say that the* covariate shift *assumption holds if the conditional labeling functions are the same in target and source domain, i.e.*

$$f_S^* = f_T^*$$

Another assumption that is often made (Ben-David et al. (2010a),Ben-David et al. (2006),Ben-David et al. (2010b)) concerns the existence of a hypothesis $h^* \in \mathcal{H}$ that has low risk in both source and target domain. For this so-called *low-error joint predictions assumption*, we will denote the source risk of a hypothesis h with $\mathcal{L}_{P_S}(h)$ and the target risk of a hypothesis with $\mathcal{L}_{P_T}$

Definition 19 (Optimal joint hypothesis, from Ben-David et al. (2006))**.** *The optimal joint hypothesis is the hypothesis which minimizes the combined error*

$$h^* = \arg\min_{h \in \mathcal{H}} \mathcal{L}_{P_S}(h) + \mathcal{L}_{P_T}(h).$$

We denote the combined error of the optimal hypothesis for a hypothesis class $\mathcal{H}$ by

$$\lambda_{\mathcal{H}}(P_S, P_T) = \mathcal{L}_{P_S}(h^*) + \mathcal{L}_{P_T}(h^*).$$

Furthermore, guarantees for DA-learnability must obviously depend on how similar the source and target distribution are to each other. Thus we need a formal way to measure the similarity between two distributions. While there are several notions of distances for probability distributions to choose from, one that is useful for obtaining worst-case bounds, is the $\mathcal{A}$-*distance*. [1]

Definition 20 ($\mathcal{A}$-distance, from Ben-David et al. (2006))**.** *Let $\mathcal{A} \subset 2^{\mathcal{X}}$ and let P and Q be two probability measures over the set $\mathcal{X}$, such that every set $A \in \mathcal{A}$ is measurable with respect to both distributions. Then the $\mathcal{A}$-distance $d_{\mathcal{A}}$ is defined as:*

$$d_{\mathcal{A}}(P, Q) := 2 \sup_{A \in \mathcal{A}} |P(A) - Q(A)|.$$

Since we are interested in worst-case scenarios, this choice of similarity-measure makes sense: it only considers the largest occurring difference over all sets for some sets $A \in \mathcal{X}$. The remaining question is: what $\mathcal{A}$ to choose? An obvious choice of $\mathcal{A}$ would be all measurable subsets of $\mathcal{X}$. This choice results in the *total variation distance* between distributions. However, this distance measure cannot be estimated from finite samples for distributions over an uncountable domain. Furthermore we are only interested in differences in distributions, that are relevant to our (with respect to distribution optimal) choice of a hypothesis $h \in \mathcal{H}$. By slight abuse of notation, we will now define the $\mathcal{H}$-divergence (cf. Ben-David et al. (2006)) for two distributions P and Q over some domain $\mathcal{X}$:

$$d_{\mathcal{H}}(P, Q) := 2 \sup_{A \in \mathcal{X} : \mathbb{1}(A) \in \mathcal{H}} |P(A) - Q(A)|,$$

[1] Please note, that $\mathcal{A}$ does *not* denote a learner in the next two paragraphs, but a collection of subsets of $\mathcal{X}$.

where $\mathbb{1}(A)$ denotes the indicator function of the set A. For two functions $a, b \in \{0,1\}^{\mathcal{X}}$ let $a \oplus b(x) := \mathbb{1}[a(x) \neq b(x)]$ for all $x \in \mathcal{X}$ denote the XOR function. We define the symmetric difference hypothesis space as

$$\mathcal{H}\Delta\mathcal{H} := \{h\Delta h' | h, h' \in \mathcal{H}\}.$$

Since $h(x) \in \{0,1\}$ for all $h \in \mathcal{H}$, we can interpret every function of $\mathcal{H}$ as and indicator functions for some set $A \in \mathcal{X}$. Now, we can define the $\mathcal{H}\Delta\mathcal{H}$-divergence analogous to Ben-David et al. (2010a).

Definition 21 ($\mathcal{H}\Delta\mathcal{H}$-divergence, from Ben-David et al. (2010a))**.**

$$d_{\mathcal{H}\Delta\mathcal{H}}(P,Q) := 2 \sup_{A\subset\mathcal{X}:\mathbb{1}(A)\in\mathcal{H}\Delta\mathcal{H}} |P(A) - Q(A)| = 2 \sup_{A,B\subset\mathcal{X}:\mathbb{1}(A),\mathbb{1}(B)\in\mathcal{H}} |P(A\Delta B) - Q(A\Delta B)|.$$

Using Lemma 1 from Ben-David et al. (2010a), we can estimate $d_{\mathcal{H}\Delta\mathcal{H}}(P,Q)$ empirically using finite samples from both distributions. For this let $\mathcal{U}$ and $\mathcal{U}'$ be samples from $\mathcal{D}$ and $\mathcal{D}'$, respectively. We define the empirical $\mathcal{H}$-divergence of $\mathcal{U}$ and $\mathcal{U}'$ as

$$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U},\mathcal{U}') := 2 \sup_{A\subset\mathcal{X},\mathbb{1}\in\mathcal{H}\Delta\mathcal{H}} \left|\frac{|\mathcal{U}\cap A|}{|\mathcal{U}|} - \frac{|\mathcal{U}'\cap A|}{|\mathcal{U}'|}\right|.$$

Building on this definition, we can now introduce Lemma 1 from Ben-David et al. (2010a).

Lemma 2 (Lemma 1 from Ben-David et al. (2010a))**.** *Let $\mathcal{H}$ be a [binary] hypothesis space on $\mathcal{X}$ with VC dimension d. If $\mathcal{U}$ and $\mathcal{U}'$ are samples of size m from $\mathcal{D}$ and $\mathcal{D}'$ respectively and $\hat{d}_{\mathcal{H}}(\mathcal{U},\mathcal{U}')$ is the empirical $\mathcal{H}$-divergence between samples, then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$d_{\mathcal{H}}(\mathcal{D},\mathcal{D}') \leq \hat{d}_{\mathcal{H}}(\mathcal{U},\mathcal{U}') + 4\sqrt{\frac{d\log(2m) + \log(\frac{2}{\delta})}{m}}.$$

To estimate the $\mathcal{H}\Delta\mathcal{H}$-divergence we need the VC-dimension of $\mathcal{H}\Delta\mathcal{H}$. It is often stated (Ben-David et al. (2006),Mansour et al. (2009),citeDAhard), that $\mathrm{VC}(\mathcal{H}\Delta\mathcal{H}) \leq \mathrm{VC}(\mathcal{H})$. However, we were not able to find a valid proof of this statement in the literature. While this inequality might still hold, we will only use the slightly worse bound of $\mathrm{VC}(\mathcal{H}\Delta\mathcal{H}) \leq 4\mathrm{VC}(\mathcal{H})\log(4\mathrm{VC}(\mathcal{H}))$ and will refer the reader to the appendix for a proof and further discussion.

Having a low $\mathcal{H}\Delta\mathcal{H}$-divergence between source and target distributions is another common assumption in domain adaptation.[2]

A more generalized version of this similarity measure, which works for general loss-functions and general label spaces $\mathcal{Y}$, is the *discrepancy distance*, as introduced in Mansour et al. (2009).

[2]For the rest of this thesis the $\mathcal{H}\Delta\mathcal{H}$-divergence will be the only $\mathcal{A}$-distance we will use. From this point on $\mathcal{A}$ will therefore only refer to learners.

Definition 22 (Discrepancy Distance, analogous to Definition 4 from Mansour et al. (2009))**.** *Let $\mathcal{H} : \mathcal{X} \to \mathcal{Y}$ be a hypothesis class and let $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ define a loss function over $\mathcal{Y}$. The discrepancy distance $disc_l$ between two distributions Q_1 and Q_2 over $\mathcal{X}$ is defined by*

$$disc_l(Q_1, Q_2) = \max_{h,h' \in \mathcal{H}} |\mathcal{L}_{Q_1}(h', h) - \mathcal{L}_{Q_2}(h', h)|$$

Up to the difference between taking the supremum or the maximum and a factor of 2, this definition coincides with the $\mathcal{H}\Delta\mathcal{H}$-divergence for binary classification under the 0-1-loss.

Another notion of similarity between distributions, that has been used in the literature, is the so-called *weight-ratio* assumption. This definition is motivated by the question, how well a learner can adapt to a target domain, that is a subset of the source domain.

Definition 23 (Weight-ratio, Definition 2 from Ben-David and Urner (2012))**.** *Let $\mathcal{B} \subset 2^{\mathcal{X}}$ be a collection of subsets of the domain $\mathcal{X}$ measurable with respect to both P_S and P_T. We define the* weight ratio *of the source distribution and the target distribution with respect to $\mathcal{B}$ as*

$$C_{\mathcal{B}}(P_S, P_T) = \inf_{b \in \mathcal{B}(\mathcal{X}), P_T(b) \neq 0} \frac{P_S(b)}{P_T(b)}.$$

We denote the weight ratio with respect to the collection of all sets that are P_S- and P_T-measurable by $C(P_S, P_T)$.

Note that this definition is not dependent on the hypothesis class used and is the first assumption we have used that is not symmetric in P_S and P_T. In practice, we might find situations where we could adapt from the domain distribution P_S to P_T, but *not* vice versa, e.g., if $\mathcal{D}_T$ is distributed over a subset of the source support $\text{supp}(\mathcal{D}_S)$. Bounds that only consider criteria that are symmetric in P_S and P_T will fail to give guarantees in these scenarios. In contrast, the weight ratio assumption might give an insight into these situations.

4.2. Existing results

4.2.1. Upper bounds for domain adaptation

One of the first formal guarantees for domain adaptation was given in Ben-David et al. (2006), providing a bound for the target error of a classifier learned on the source domain in terms of its $\mathcal{H}\Delta\mathcal{H}$-distance and its joint hypothesis error :

Theorem 7 (Theorem 2 from Ben-David et al. (2010a)). *Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a hypothesis space with finite VC-dimension d. If $\mathcal{U}_S, \mathcal{U}_T$ are unlabeled samples of size m each, drawn from $\mathcal{D}_S, \mathcal{D}_T$ respectively, then for any $\delta \in (0,1)$ with probability at least $1-\delta$ (over the choice of samples), for every $h \in \mathcal{H}$ we obtain:*

$$\mathcal{L}_{P_T}(h) \leq \mathcal{L}_{P_S}(h) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T) + 4\sqrt{\frac{2d\log(m) + \log(\frac{\delta}{2})}{m}} + \lambda_{\mathcal{H}}(P_T, P_S)$$

By this theorem, a problem is DA-learnable if the following two assumptions hold.

(1) $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_T, \mathcal{D}_S) = 0$ and

(2) $\lambda_{\mathcal{H}}(P_T, P_S) = 0$.

But this case is still quite restricted, since we are only looking at distributions $\mathcal{D}_S$ and $\mathcal{D}_T$ that can be considered equal with respect to $\mathcal{H}$. However, note that this guarantee also holds for the case, where we do not have covariate shift.

Another upper bound for domain adaptation was provided in Ben-David et al. (2012). The authors make assumptions about the weight ratio instead of an assumption about $\mathcal{H}$-divergence. Note that for this result, covariate shift is indeed needed.

Observation 1 ((Observation 5 from Ben-David et al. (2012))). *Let $\mathcal{X}$ be a domain and let P_S and P_T be a source and a target distribution over $\mathcal{X} \times \{0,1\}$ satisfying the covariate shift assumption, with $C_{\{\{x\}:x\in\mathcal{X}\}}(\mathcal{D}_S, \mathcal{D}_T) > 0$. Then we have $\mathcal{L}_{P_T}(h) \leq \frac{1}{C_{\{\{x\}:x\in\mathcal{X}\}}}\mathcal{L}_{P_S}(h)$ for all $h : \mathcal{X} \to \{0,1\}$.*

In both of these upper bounds, we do not have to actually "adapt" for the new domain, since we only rely on the performance guarantee we get for the classifier trained on the source. The only use we had for unlabeled target data so far was to estimate the $\mathcal{H}$-divergence for the bound from Ben-David et al. (2006).

4.2.2. Reweighting technique

Before exploring further learnability results for domain adaptation, we will briefly mention one common technique from Mansour et al. (2009), one that actually takes into account the unlabeled data from the source domain.
The idea behind this technique is to assign weights to the source data, in such a way that the reweighted source distribution and the target distribution have low discrepancy distance.
Let $\hat{\mathcal{D}}_S$ and $\hat{\mathcal{D}}_T$ be the empirical source and target distribution, respectively. The aim now is to construct a $\hat{\mathcal{D}}'_S$ such that $disc(\hat{\mathcal{D}}'_S, \hat{\mathcal{D}}_T)$ is small under the constraint $\text{supp}(\hat{\mathcal{D}}'_S) \subset \text{supp}(\hat{\mathcal{D}}_S)$. It is argued in Mansour et al. (2009) that for binary classification this results in the optimization problem

$$\arg\min_{\hat{\mathcal{D}}'_S} \max_{\mathbb{1}(A)\in\mathcal{H}\Delta\mathcal{H}} |\hat{\mathcal{D}}'_S(A) - \hat{\mathcal{D}}_T(A)|$$
$$\text{subject to } \forall x \in L_\mathcal{X}, \hat{\mathcal{D}}'_S(x) \geq 0 \wedge \sum_{x\in L_\mathcal{X}} \hat{\mathcal{D}}'_S(x) = 1$$

where $L_\mathcal{X}$ denotes the projections of the labeled source data L to the domain $\mathcal{X}$. This, the authors argue, they argue can be rewritten as

$$\begin{aligned}\min_{\hat{\mathcal{D}}'_S} \quad & \delta \\ \text{subject to } & \forall \mathbb{1}_A \in \mathcal{H}\Delta\mathcal{H}, \hat{\mathcal{D}}'_S(A) - \hat{\mathcal{D}}_T(A) \leq \delta \\ & \forall \mathbb{1}_A \in \mathcal{H}\Delta\mathcal{H}, \hat{\mathcal{D}}_T(A) - \hat{\mathcal{D}}'_S(A) \leq \delta \\ & \forall x \in L_\mathcal{X}, \hat{\mathcal{D}}'_S(x) \geq 0 \wedge \sum_{x\in L_\mathcal{X}} \hat{\mathcal{D}}'_S(x) = 1.\end{aligned}$$

They go on to argue that the number of constraints is proportional to $|\mathcal{H}\Delta\mathcal{H}|$. However for two $A, A' \subset \mathcal{X}$ with $\mathbb{1}_A, \mathbb{1}_{A'} \in \mathcal{H}\Delta\mathcal{H}$, two constraints coincide if they include the same points of $L_\mathcal{X}$ and U. Therefore, if $VC(\mathcal{H}\Delta\mathcal{H})$ is finite, then the number of constraints can be reduced as well. To be more precise, if we have m samples from $\hat{\mathcal{D}}_S$ and n samples from $\hat{\mathcal{D}}_T$, the number of constraints is bounded by $(n+m)^{VC(\mathcal{H}\Delta\mathcal{H})} \leq (n+m)^{4VC(\mathcal{H})\log(4VC(\mathcal{H}))}$ (see Appendix).
The paper goes on to argue that in cases where we can test efficiently whether there is a consistent hypothesis in $\mathcal{H}$, we can generate all consistent labelings of the same points by $\mathcal{H}$ in $O((m+n)^{VC(\mathcal{H}\Delta\mathcal{H})})$ time.

4.2.3. Impossibility results

Covariate shift is an assumption that is often easily justifiable. For example in image recognition the labeling rules between two data sets will rarely change, but the underlying marginal distribution of two data sets are often different: they might have different light conditions or certain motives might be present more often in one data set than in the other. Therefore it is a question of interest, whether or not this assumption leads to new domain adaptation guarantees. As shown in Ben-David et al. (2010b), however, covariate shift alone does not suffice to make any of the previously used assumptions – i.e., low $\mathcal{H}\Delta\mathcal{H}$-distance and low joint prediction error – obsolete.

Theorem 8 (Necessity of small $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_T, \mathcal{D}_S)$, Theorem 1 from Ben-David et al. (2010b))**.** *Let $\mathcal{X}$ be some domain set, and $\mathcal{H}$ a class of functions over $\mathcal{X}$. Assume that*

for some $A \subset \mathcal{X}$, $\{h^{-1}(1) \cap A : h \in \mathcal{H}\}$ contains more than two [non-empty][3] sets and is linearly ordered by inclusion. Then the conditions "covariate shift plus small $\lambda_{\mathcal{H}}$" do not suffice for DA-learnability. In particular, for every $\varepsilon > 0$ there exist probability distributions P_S over $\mathcal{X} \times \{0,1\}$, $\mathcal{D}_T$ over $\mathcal{X}$ such that for every *domain adaptation learner $\mathcal{A}$ with outputs in $\mathcal{H}$, and all integers $m, n > 0$, there exists a labeling function $f : \mathcal{X} \to \{0,1\}$ such that*

1. $\lambda_{\mathcal{H}}(P_T, P_S) \leq \varepsilon$,
2. *P_T and P_S satisfy the covariate shift assumption,*
3. $\underset{L \sim {P_S}^m, U \sim \mathcal{D}_T^n}{Pr}[\mathcal{L}_{P_T}(\mathcal{A}(L,U)) \geq \frac{1}{2}] \geq \frac{1}{2}$.

The following proof is different from the proof provided in Ben-David et al. (2010b). After the proof there will be a short discussion and comparison of the two proofs.

Proof. Since there is a set $A \subset \mathcal{X}$ such that the set $\{h^{-1}(1) \cap A | h \in \mathcal{H}\}$ is linearly ordered by inclusion and has at least two elements, we can choose $h_1, h_2 \in \mathcal{H}$, such that $\emptyset \subsetneq h_1^{-1}(1) \cap A \subsetneq h_2^{-1}(1) \cap A$. Let P_S be uniformly distributed[4] over $h_1^{-1}(1) \cap A \times \{1\}$. Furthermore, let P_{T_1} be uniformly distributed over $((h_2^{-1}(1) \cap A) \setminus (h_1^{-1}(1) \cap A)) \times \{0\}$ and P_{T_2} be uniformly distributed over $((h_2^{-1}(1) \cap A) \setminus (h_1^{-1}(1) \cap A)) \times \{1\}$. Note that $\mathcal{D}_{T_1} = \mathcal{D}_{T_2} =: \mathcal{D}_T$. Furthermore note that the labeling function $f_1 = h_1$ is consistent with both P_S and P_{T_2} and that the labeling function $f_2 = h_2$ is consistent with both P_S and P_{T_2}. Therefore (1) and (2) are fulfilled by the labeling functions f_1 and f_2 with their respective source and target distributions. Also note that $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) = 1$. Since $\mathcal{D}_{T_1} = \mathcal{D}_{T_2}$, the learning problems $(P_S, P_{T_1}, \mathcal{H})$ and $(P_S, P_{T_1}, \mathcal{H})$ are indistinguishable from a pair (L, U) of labeled data L from P_S and unlabeled data U from $\mathcal{D}_T$. Now let

[3] The assumption that there are at least two *non-empty* sets in $A \subset \mathcal{X}$, $\{h^{-1}(1) \cap A : h \in \mathcal{H}\}$ is an addition to the original theorem in Ben-David et al. (2010b). However, this addition is necessary, since otherwise, we could look at the function class $\mathcal{H}_{0,1} := \{h_0, h_1 : \mathcal{X} \to \{0,1\}\}$ with $h_0(x) = 0$ and $h_1(x) = 1$ for all $x \in \mathcal{X}$. In this case, $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) = 0$ is implied for all distributions P_T, P_S on $\mathcal{X}$. As seen in Theorem 7, this combined with the additional assumption $\lambda_{\mathcal{H}}(P_T, P_S) \leq \varepsilon$ implies $\mathcal{L}_{P_T}(h) \leq \varepsilon + \mathcal{L}_{P_S}(h)$ for every $h \in \mathcal{H}_{0,1}$. For the optimal joint hypothesis h^* this means h^* that $\mathcal{L}_{P_T}(h^*) \leq 2\varepsilon$. Since $\mathcal{H}_{0,1} = \{|h^*, |1 - h^*||\}$, $\lambda(P_S, P_T)_{\mathcal{H}} \leq \varepsilon$ implies h^* is the only optimal hypothesis for the source domain. VC($\mathcal{H}_{0,1}$)=1 then implies that any ERM-rule $\mathcal{A}$ on the source domain, with sufficiently large sample set L, will output h^* with high probability. Therefore, for every $\varepsilon' > \varepsilon$.

[4] In cases where the set A is not compact, one could argue that there is no uniform distribution over $h_1^{-1}(1) \cap A \times \{1\}$. However in that case, we can define a compact subset $A' \subset A$, such that $h_1^{-1}(1) \cap A' \subsetneq h_2^{-1}(1) \cap A'$. In this case it is possible to define a uniform distribution over A' (and thus over all measurable subsets of A'). In the following we will therefore assume A to be compact.

$\mathcal{A}$ be any DA-learner on $\mathcal{H}$. If $\underset{L\sim {P_S}^m, U\sim \mathcal{D}_T^n}{Pr}[\mathcal{L}_{P_{T_1}}(\mathcal{A}(L,U)) \geq \frac{1}{2}] < \frac{1}{2}$, then

$$\begin{aligned}
&\underset{L\sim {P_S}^m, U\sim \mathcal{D}_T^n}{Pr}[\mathcal{L}_{P_{T_2}}(\mathcal{A}(L,U)) \geq \frac{1}{2}] = \\
= &\underset{L\sim {P_S}^m, U\sim \mathcal{D}_T^n}{Pr}[1 - \mathcal{L}_{P_{T_1}}(\mathcal{A}(L,U)) \geq \frac{1}{2}] = \\
= &\underset{L\sim {P_S}^m, U\sim \mathcal{D}_T^n}{Pr}[\mathcal{L}_{P_{T_1}}(\mathcal{A}(L,U)) \leq \frac{1}{2}] \geq \frac{1}{2}.
\end{aligned}$$

Any learner $\mathcal{A}$ will therefore either fail to solve the DA-problem $(P_S, P_{T_1}, \mathcal{H})$ or fail to solve $(P_S, P_{T_2}, \mathcal{H})$. □

Note that in this proof, the distributions P_{T_1}, P_{T_2} and P_S were all realizable in $\mathcal{H}$. In the original proof from the paper, there was no realizability constraint, however $\mathcal{D}_S$ and $\mathcal{D}_T$ were constructed in such a way that $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) = 1 - \varepsilon$. In this way their construction allowed the use of the reweighting technique and could therefore demonstrate the method's short-comings. Hypothesis classes $\mathcal{H}$ that fulfill the requirements of this theorem include threshold functions and half-spaces. Furthermore, note that the assumption that there is a subset $A \subset \mathcal{X}$ such that $\{h^{-1}(1) \cap A : h \in \mathcal{H}\}$ is linearly ordered by inclusion and contains more than two sets, is sufficient, but not necessary for the proof to work. It is already sufficient that there exists a subset $\mathcal{H}' \subset \mathcal{H}$ and $A \subset \mathcal{X}$ such that $\{h^{-1}(1) \cap A : h \in \mathcal{H}'\}$ is linearly ordered by inclusion and contains at least two sets. Indeed this is the case for every function class with a VC-dimension of at least 2 or more precisely for every hypothesis class, which containing at least 3 hypotheses.[5]

Theorem 9 (Necessity of small $\lambda_{\mathcal{H}}(\mathcal{D}_T, \mathcal{D}_S)$ (Theorem 2 from Ben-David et al. (2010b)))**.** *Let $\mathcal{X}$ be some domain set, and $\mathcal{H}$ be a class of functions over $\mathcal{X}$ whose VC dimension is much smaller than $|\mathcal{X}|$(in particular, any $\mathcal{H}$ with a finite VC dimension over an infinite $\mathcal{X}$ will do). Then the conditions "covariate shift plus small $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_T, \mathcal{D}_S)$" do not suffice for DA-learnability. In particular, for every $\varepsilon > 0$ there exist probability distributions P_S over $\mathcal{X} \times \{0,1\}$, $\mathcal{D}_T$ over $\mathcal{X}$ such that for* every *DA learner $\mathcal{A}$ with outputs in $\mathcal{H}$, and for all integers $m, n > 0$, there exists a labeling function $f : \mathcal{X} \to \{0,1\}$ such that*

1. $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_T, \mathcal{D}_S) \leq \varepsilon$,
2. *The covariate shift assumption holds,*
3. $\underset{L\sim P_S^m, U\sim \mathcal{D}_T^n}{Pr}[\mathcal{L}_{P_T}(\mathcal{A}(L,U)) \geq \frac{1}{2}] \geq \frac{1}{2}$.

[5] To be more precise, it might be the case that if a hypothesis class $\mathcal{H}$ contains more than three hypotheses, but there is still no$A \subset \mathcal{X}$, such that $\{h^{-1}(1) \cap A : h \in \mathcal{H}\}$ is not linearly ordered. This is indeed the case for the class of singletons. However, in these cases there exists a set $\mathcal{H}' \subset \mathcal{H}$ and $A \subset \mathcal{X}$, such that $\{h^{-1}(0) \cap A : h \in \mathcal{H}'\}$ is linearly ordered by inclusion. In these cases we can make the same proof as before by exchanging the roles of the labels 0 and 1.

The proof of this lemma will follow the sketch of proof given in Ben-David et al. (2010b). But first let us introduce a lemma from Urner (2013).

Lemma 3 (Lemma 47 from Urner (2013))**.** *Let $\mathcal{X}$ be a finite domain of size m. For every $0 < \beta < 1$, with probability exceeding β, an i.i.d. sample of size at most $n \leq \min\{\sqrt{ln(2)m}, \sqrt{ln(\frac{1}{\beta})m}\}$ uniformly drawn over $\mathcal{X}$ contains no repeated elements.*

Furthermore we will need the concept of *ε-approximations*. A class $\mathcal{B}$ of subsets of $\mathcal{X}$ is said to have VC-dimension d, if the class $\mathcal{H}_{\mathcal{B}} = \{h : \mathcal{X} \to \{0,1\} | h = \mathbb{1}_B \text{ for some } B \in \mathcal{B}\}$ of indicator functions of elements of $\mathcal{B}$ has VC-dimension d. For a sample S, we denote the *empirical estimate* of the weight of a set $B \in \mathcal{B}$ by $\widehat{S(B)} = \frac{|S \cap B|}{|S|}$.

Definition 24 (Definition 29 from Urner (2013))**.** *Let $\mathcal{X}$ be some domain, $\mathcal{B} \subset 2^{\mathcal{X}}$ a collection of subsets of $\mathcal{X}$ and P a distribution over $\mathcal{X}$. An* ε-approximation *for $\mathcal{B}$ with respect to P is a finite subset $S \subset \mathcal{X}$ with*

$$|\widehat{S(B)} - P(B)| \leq \varepsilon$$

for all sets $B \in \mathcal{B}$. In this case, we will also call S an ε-approximation for the corresponding hypothesis class $\mathcal{H}_{\mathcal{B}}$ with respect to P.

Proof of Theorem 9. The idea behind this proof is to choose P_S and P_T in a way that $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) < \varepsilon$, but such that the supports $\mathcal{D}_S$ and $\mathcal{D}_T$ are disjoint. If this is the case, we can construct a labeling function f that is both source- and target realizable, but where the joint prediction error is still large and such that DA-learnability is not possible.
But first we have to show that this construction of P_S and P_T is possible. Note that if there are finite, disjoint sets S_1, S_2 such that

$$|\widehat{S_1(B)} - \widehat{S_2(B)}| \leq \varepsilon$$

for all B with $\mathbb{1}_B \in \mathcal{H}\Delta\mathcal{H}$, defining $\mathcal{D}_S$ as the uniform distribution over S_1 and defining $\mathcal{D}_T$ as the uniform distribution over S_2 will fulfill $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) < \varepsilon$. Now let P be some uniform distribution over a compact subset $\mathcal{X}' \subset \mathcal{X}$. According to Urner (2013) two i.i.d samples S_1 and S_2 of size

$$\frac{16}{(2\varepsilon)^2}\left(VC\left(\mathcal{H}\Delta\mathcal{H}\right)\ln\left(\frac{16VC(\mathcal{H}\Delta\mathcal{H})}{(2\varepsilon)^2}\right) + \ln\left(\frac{4}{\delta'}\right)\right)$$

each are both $\frac{\varepsilon}{2}$-approximations of $\mathcal{H}\Delta\mathcal{H}$ with respect to P with probability at least $(1-\delta)$, where $\delta = 2\delta'$. This implies

$$|\widehat{S_1(B)} - \widehat{S_2(B)}| \leq |\widehat{S_1(B)} - P(B)| + |\widehat{S_2(B)} - P(B)| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \leq \varepsilon$$

with probability $(1-\delta)$. Furthermore, Lemma (47 from Urner (2013)) tells us that both samples will be disjoint with probability at least $\frac{1}{2}$, if

$$\frac{32}{(2\varepsilon)^2}\left(VC\left(\mathcal{H}\Delta\mathcal{H}\right)\ln\left(\frac{16VC(\mathcal{H}\Delta\mathcal{H})}{(2\varepsilon)^2}\right)+\ln\left(\frac{4}{\delta}\right)\right)\leq\sqrt{|\mathcal{X}'|\ln(2)}$$

Therefore if $\mathcal{X}'$ is infinite or if $\mathcal{X}'$ is finite with

$$\frac{1}{\ln(2)}\sqrt{\frac{32}{(2\varepsilon)^2}\left(VC\left(\mathcal{H}\Delta\mathcal{H}\right)\ln\left(\frac{16VC(\mathcal{H}\Delta\mathcal{H})}{(2\varepsilon)^2}\right)+\ln\left(\frac{4}{\delta}\right)\right)}\leq|\mathcal{X}'|$$

there exist $\mathcal{D}_S$ and $\mathcal{D}_T$ such that their support is disjoint and $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T)\leq\varepsilon$.
Now let $h\in\mathcal{H}$. We choose P_S in a way that the labeling is consistent with h, i.e., $\lambda_{P_S}(h)=0$. Now let us define two labeling functions f_1 and f_2. Let $f_1(x)=h(x)$ for all $x\in\mathcal{X}$. Furthermore let $f_2(x)=h(x)$ for all $x\in\mathcal{X}$ and $x\in\text{supp}(\mathcal{D}_S)$ and $f_2(x)=|1-h(x)|$ for all $x\in\text{supp}(\mathcal{D}_T)$. Let P_{T_1} be the target distribution consistent with f_1 and P_{T_2} be the target distribution consistent with f_2. Any learner $\mathcal{A}$ that succeeds on the DA-problem (P_S,P_{T_1}) will fail on (P_S,P_{T_2}) and vice versa. □

Note that this theorem does not make any assumptions about source and target realizability. There are cases where we have source and target realizability, but still get $\lambda_{\mathcal{H}}(P_T,P_S)\geq 1-\varepsilon$. For example, this is the case if there exists an $h\in\mathcal{H}$, such that $h':x\to|h(x)-1|$ is in $\mathcal{H}$, source and target supports are disjoint and the source is labeled by h, while the target is labeled by h'. In the case where $\mathcal{H}=\{h,h'\}$, we additionally get $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T)=0$ for every pair of distributions $(\mathcal{D}_S,\mathcal{D}_T)$. Thus, the theorem also holds for some cases where we have source and target realizability.
However this is not the case for certain function classes, where source and target realizability combined with small $d_{\mathcal{H}\Delta\mathcal{H}}(P_S,P_T)$ implies small $\lambda_{\mathcal{H}}(\mathcal{D}_T,\mathcal{D}_S)$.

Another set of assumptions was explored in Ben-David and Urner (2012). In this case, the target domain was assumed to be a subset of the source domain, with a weight-ratio of $\frac{1}{2}$. We furthermore have covariate shift and a small $\mathcal{H}\Delta\mathcal{H}$-distance in this scenario, while the joint prediction error is not bounded. It was shown that this set of assumptions does not suffice to give a domain-adaptation bound by giving the following counter example:

In the following let $\mathcal{H}_{(1,0)}:=\{h_0,h_1\in\{0,1\}^{\mathcal{X}}\}$, where $h_0(x)=0$ and $h_1(x)=1$ for all $x\in\mathcal{X}$.

While a version of this bound is first given in Ben-David and Urner (2012), we will now provide a slightly better version from Urner (2013).

Theorem 10 (Theorem 41 from Urner (2013))**.** *For every finite domain $\mathcal{X}$, for every ε and δ with $\varepsilon+\delta<\frac{1}{2}$, no algorithm can (ε,δ,s,t)- solve the DA problem for the class $\mathcal{W}_n$ of pairs (P_S,P_T) satisfying the covariate shift with $C(\mathcal{D}_S,\mathcal{D}_T)\geq\frac{1}{2}$, $d_{\mathcal{H}_{1,0}\Delta\mathcal{H}_{1,0}}(\mathcal{D}_S,\mathcal{D}_T)=0$ and $opt_{P_T}(\mathcal{H}_{1,0})=0$ if*

$$s+t\leq\min\left\{\sqrt{\ln(2)|\mathcal{X}|},\sqrt{\ln\left(\frac{1}{2(\varepsilon+\delta)}\right)|\mathcal{X}|}\right\}-1.$$

We will now give a description of the proof of this theorem as it is done in Urner (2013). For the proof of this theorem the problem is reduced to the so-called Left/Right Problem (first introduced in Kelly et al. (2010)), which takes as input three finite samples L, R and M over the finite domain $\mathcal{X}$. It is assumed that L is an i.i.d. sample of some distribution P_1 over $\mathcal{X}$, R is an i.i.d. sample of some distribution P_2 over $\mathcal{X}$ and M is an i.i.d. sample of either P_1 or P_2. The Left/Right Problem is to find out whether M is generated by P_1 or by P_2. More formally the solvability of this problem can be defined as follows:

Definition 25 (Left/Right problem solvability (Definition 43 from Urner (2013))). *We say that a (randomized) algorithm (δ, l, r, m)- solves the* Left/Right problem *with respect to a class $\mathcal{W}$ of triplet (P_1, P_2, P_3) of distributions (where $P_3 = P_1$ or $P_3 = P_2$), if given sample L i.i.d. form P_1, R i.i.d. from P_2 and M i.i.d. from P_3 of sizes l, r and m respectively, it correctly decides whether $P_3 = P_1$ or $P_3 = P_2$ with probability at least $1 - \delta$.*

For the construction of the learning problem that gives rise to the lower bound given in Theorem 10, the particular problem class $\mathcal{W}_n^{uni} : \{(U_A, U_B, U_C) | A \cup B = \{1, \ldots, n\}, A \cap B = \emptyset, |A| = |B|$, and $C = A$ or $C = B\}$ of Left/Right problems, where P_1 and P_2 are uniform distributions over disjoint sets A and B, which partition the set $\{1, \ldots, n\}$, is introduced. Since no additional structure for A and B is given, the problem of (δ, l, r, m)-solvability of $\mathcal{W}_n^{uni}$ can be reduced to the question of how likely it is that a sample of M of size m of either P_1 or P_2 coincides in at least one point with a sample L from P_1 of size l or a sample R from P_2 of size r respectively. Using Lemma 47 from Urner (2013), we can give the following lower bound for this problem:

Lemma 4 (Lemma 44 from Urner (2013)). *For any given example sizes l for L, r for R and m for M and any $0 < \gamma < \frac{1}{2}$, if $k = \max\{l, r\} + m$, then for*

$$n > \max\left\{\frac{k^2}{\ln(2)}, \frac{k^2}{\ln(\frac{1}{2\gamma})}\right\}$$

no algorithm has probability of success greater than $1 - \gamma$ over the class $\mathcal{W}_n^{uni}$.

Let us now consider the class of DA-problems $\mathcal{W}_n$ of pairs (P_S, P_T) that are constructed in the following way. Let the corresponding target marginal distribution $\mathcal{D}_T$ be uniform over some subset T of $\mathcal{X}$, with $|U| = |\mathcal{X}|$ and the corresponding source marginal distribution $\mathcal{D}_S$ be uniform over $\mathcal{X}$. Furthermore, let there be a labeling function that is consistent with both P_S and P_T that labels all elements of T with "1" and all elements of $\mathcal{X} \setminus T$ with "0" (or vice versa). An illustration of this problem set can be seen in Figure 4.1.

It is easy to see that we have $opt_{P_T}(\mathcal{H}_{1,0}) = 0$ in this case. Furthermore we also get $C(P_S, P_T) = \frac{1}{2}$ and $d_{\mathcal{H}_{1,0}\Delta\mathcal{H}_{1,0}}(P_T, P_S) = 0$. If we now regard all data from the source distribution with label "1" to be generated by some uniform distribution U_A (with either $A = T$ or $A = \mathcal{X} \setminus T$) and all data with label "0" to be generated by some uniform

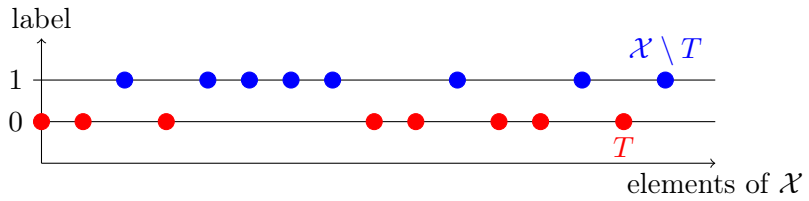


Figure 4.1.: An illustration of one pair $(P_S, P_T) \in \mathcal{W}_n$. The elements of $\mathcal{X}$ are labeled "0" for the target support (marked red) and "1" everywhere else.

distribution U_B (with $B = \mathcal{X} \setminus A$), deciding on the correct labeling function $h \in \mathcal{H}_{1,0}$ for the target domain is equivalent to the problem of deciding whether the (unlabeled) target data is generated by U_A or by U_B. Or to put it more formally: With this construction we can finally reduce the solvability of the Left/Right problem $\mathcal{W}_n^{uni}$ to the solvability of the DA problem $\mathcal{W}_n^{DA}$:

Lemma 5 (Lemma 48 from Urner (2013))**.** *The Left/Right problem reduces to domain adaptation. More precisely, given a number n and an algorithm $\mathcal{A}$ that, given the promise that the target task is realizable by the class $\mathcal{H}_{1,0}$, can $(\varepsilon, \delta, s, t)$-solve DA for a class $\mathcal{W}$ that includes $\mathcal{W}_n$, we can construct an algorithm that $(\varepsilon+\delta, s, s, t+1)$-solves the Left/Right problem on $\mathcal{W}_n^{uni}$.*

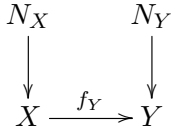
This means that we can derive a lower bound for the DA problem $\mathcal{W}_n$ from the lower bound for the Left/Right problem introduced in Lemma 25. This results in Theorem 10.

5. Domain Adaptation under Causal Assumptions

We will now look at what kind of assumptions for domain adaptation result from causal models, as introduced in Chapter 3. We will mainly focus on the causal direction, i.e., the feature vectors cause the labels (and not vice versa). We will first look at SCMs and their implications for domain adaptation. We will then go on to look at several formalizations of the Principle of Independence of Cause and Mechanism. In particular, we will consider the IGCI model and attempt to adapt it for binary classification. We will show that for all of these attempts, we are still able to construct lower bounds similar to the one given in Ben-David and Urner (2012) for the sample complexity needed to solve the DA problem. Since these bounds are all dependent on the cardinality of the feature domain $\mathcal{X}$, they translate to impossibility results for scenarios with infinite domain spaces $\mathcal{X}$.

5.1. Domain Adaptation assumptions resulting from Structural Causal Models

If we assume an SCM as underlying model for our data, the particulars of this model can give us an insight on how this distribution shift happens. In practice, the assumption that even if a shift in distribution happens between source and target domains, both models allow for the same SCM structure. A distribution change would therefore most likely happen, due to the change of one marginal. We will assume a simple causal model to distinguish between several cases of distribution shifts. Let X be the only cause Y, i.e., the following SCM holds:



where N_X, X and Y are random variables, X denotes the features and Y the labels we would like to learn. In this scenario the most likely distribution shift would happen if the distribution of X changes from source to target. This implies that the mechanism f_Y would either

1. be the same for both source and target, or

2. change independently of the shift between P_S and P_T.

The first case, where only the distribution of X changes, implies covariate shift, as defined before. However, we have also seen in the previous section that Ben-David et al. (2010b) and Ben-David and Urner (2012) shows that covariate shift alone is not sufficient to give guarantees for domain adaptation in the general case. The second case is less clearly defined as covariate shift, since it is unclear how this independent change would look. However, we can assume that the covariate shift is a special case of the second case. It therefore seems unlikely that it would yield a meaningful criterion for domain adaptation.

Another kind of distribution change could happen, due to a change in N_Y. In the causal scenario it is obvious, that unlabeled data of X will not help to adapt for this change. We therefore do not see how this kind of distribution shift could be accounted for by DA algorithms.

Aside from covariate shift and its generalization, we do not see any further insight SCMs give us into the distribution change in the causal direction.

But the use of SCMs might not be limited to an insight into the distribution shift. They might also give us other criteria that hold for source and target domain.

One possible use of a SCM might be, that it gives us a restricted class of possible mechanisms. If we can assume this model to be true, this might lead to a hypothesis class $\mathcal{H}$ with finite VC-dimension for which the realizability assumption holds. However, as we have seen the impossibility results of Ben-David et al. (2010b) given here by Theorem 8 and Theorem 9 still hold if we have source and target realizability.

Without accounting for interventions, the only other assumption we get from SCMs is the statistical independence of N_Y and X *within* one domain. This might also restrict our class of possible labeling functions if we allow for a restricted noise model like ANMs. However, for deterministic labeling functions – as were used for the results of Ben-David et al. (2010b) and Ben-David and Urner (2012) – this statistical independence always holds (since we can model the labeling function with an SCM with $P(N_Y = 0) = 1$). Any criterion that results from this independence therefore already holds for the counterexamples from Ben-David et al. (2010b) and Ben-David and Urner (2012).

Therefore, we conclude that in the causal direction SCMs do not give us additional assumptions that would help for domain adaptation in the scenarios discussed in Ben-David and Urner (2012) and Ben-David et al. (2010b).

5.2. Independence of Cause and Mechanism

In this section we will look at scenarios that are inspired by the Principle of Independence of Cause and Mechanism and its implications for domain adaptation. In particular, we

will look at the lower bound given in Ben-David and Urner (2012) that we introduced in Theorem 10 of Chapter 4.

We hope that an additional causal assumption is sufficient to obtain an upper bound on the sample complexity in DA-learning. If this is the case, then this additional assumption cannot be satisfied in the counterexample given above. Therefore this counterexample is a good example to test candidates for causal assumptions we could make when hoping to facilitate domain adaptation.

5.2.1. Information geometric criterion

The first assumption we look at is inspired by the IGCI model as introduced in Chapter 3. Here it was assumed that there is a cause C and an effect E, both of which are random variables distributed in the interval $[0, 1]$ and which have probability density functions p_C and p_E respectively. Furthermore there exists a function $g : [0, 1] \to [0, 1]$, determining the effect from the cause, i.e. $g(C) = E$. In the information geometric model of independence of cause and mechanism, it was furthermore assumed that g is a monotonic diffeomorphism. The independence of mechanism is then formalized by setting the covariance between some function of $g(X)$ and $p_C(X)$ to 0, where X is a uniformly distributed random variable on $[0, 1]$. In particular the covariances $\mathrm{Cov}[g'(X), p_C(X)]$ and $\mathrm{Cov}[\log(g'(X)), p_C(X)]$ were considered to be 0 in Janzing and Schölkopf (2015). The main argument behind these choices is that one can prove the identifiability of the causal direction under these assumptions as we have seen in Theorem 6. However, since we are in a binary classification scenario g' is not well defined. Therefore we will need to change the setting and hope that our new definition can still keep the parts that are important for causality.

If the labeling function f is differentiable, one could look at the covariances $\mathrm{Cov}[f'(X), p_C(X)]$ or $\mathrm{Cov}[\log f'(X), p_C(X)]$. However, since f maps to the probability of a feature vector x being mapped to the label 1 and the mechanism function g in the IGCI model maps to the true real-valued label, the mechanisms g and f are still different objects. Furthermore we would like to introduce as few additional assumptions as possible to define causality, since our aim is to find assumptions that imply DA-learnability and that are as general as possible. Therefore we would like to avoid the assumption that f' is differentiable, if possible, because this assumption introduces additional structure in our feature space. In particular, we try to avoid introducing additional structure that by itself helps for DA learnability. For example, we know that assuming the labeling function f to be Lipschitz will resolve the hardness result given in Ben-David and Urner (2012).
For these reasons, we will use a different definition, one that captures – to the best of our knowledge – the intuition behind the original IGCI-model. This definition has the advantage of not needing more additional structure, but has – as a model of causality –

the disadvantage of not being identifiable[1]. However, as mentioned in Chapter 3, Proposition 1 implies that a causal direction does not necessarily lead to its identifiability. Therefore, if we try to prove a general statement of the form "Features causing the label implies learnability.", we cannot make identifiability a requirement, because then we will not cover all causal scenarios. Moreover, if the additional assumptions that we need to be able to distinguish between a causal and an anti-causal case are too strong, they might imply learnability by themselves – without actually using the causal direction of the setting.
As discussed when introducing the IGCI model, it is motivated by the fact that for a particular kind of data generating process the covariance statement is small with high probability in terms of the data generating process. In our examples, we will discuss whether a similar data generating process would be likely. However, we try to avoid making probability statements with respect to some data generating process (we likely would not know in reality). The hope here is that if our IGCI criteria for causality hold, then they imply DA learnability. We can then independently discuss how likely this criterion holds in a causal setting. The first causal assumption we propose is designed to fit the scenario in Ben-David and Urner (2012). It postulates an independence of the labeling function f and the target distribution in the following sense:

Assumption 1 (First IGCI criterion for binary classification)**.** *Let $\mathcal{X}$ be a finite domain. If we have a causal labeling $X \to Y$, that is the features X cause the labels Y, then we have the following relation between the labeling function f and the cause (target) distribution p_T:*

$$\mathrm{Cov}_{Z \sim Uni(\mathcal{X})}[f(Z), p_T(Z)] = 0$$

where Z is a random variable that is uniformly distributed over the source domain $\mathcal{X}$.[2]

We will see that this assumption is indeed violated by the counterexample given in Theorem 10. This violation is implied in this setting by the fact that the labeling function f and the target distribution p_T are related, due to the assumption of target realizability. We will now show that the first IGCI criterion for binary classification is violated, by the problem class given in Theorem 10. Remember that p_S is uniformly distributed over the (finite) set $\mathcal{X}$ and that p_T is a uniform distribution over some (unknown) set T, with $|T| = \frac{|X|}{2}$. Furthermore the labeling function is given by either

$$f(x) = \begin{cases} 0 & \text{, if } x \in T \\ 1 & \text{, if } x \in \mathcal{X} \setminus T \end{cases}$$

[1] It is a good question how we would define identifiability in binary classification. Most arguments about identifiability detect an asymmetry between the cause and the effect in their joint distribution. But in the case of binary classification, there is already an asymmetry between the feature space $\mathcal{X}$ and the label space $\mathcal{Y}$, since they are of different sizes in most cases (that is, if $|\mathcal{X}| > 2$). Therefore there will be an an inherent difference between functions $f_1 : \mathcal{X} \to \mathcal{Y}$ and functions $f_2 : \mathcal{Y} \to \mathcal{X}$. This difference does not depend on the causal direction.

[2] In the case of the counterexample from Ben-David and Urner (2012), this uniform distribution is also the source distribution. Maybe it would in general be better to assume X to be distributed according to the source distribution, but since it does not matter here, we will not discuss it here.

or

$$f(x) = \begin{cases} 1 & \text{, if } x \in T \\ 0 & \text{, if } x \in \mathcal{X} \setminus T \end{cases}.$$

Let us also remember that the hypothesis class considered is $\mathcal{H}_{(1,0)}$. The labeling function is by construction either 0 in the whole target domain or it is 1 in the target domain. So we have two cases to distinguish:
Case 1: The labeling function is 0 in the whole target domain:

$$\begin{aligned} \mathrm{Cov}[f(Z), p_T(Z)] =& \sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} \cdot f(x) \cdot p_T(x) - (\sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} \cdot f(x))(\sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} p_T(x)) \\ =& 0 - \frac{|T|}{|\mathcal{X}|} \cdot \frac{1}{|\mathcal{X}|} = -\frac{1}{2|\mathcal{X}|} \end{aligned}$$

Case 2: The labeling function is 1 in the whole target domain:

$$\begin{aligned} \mathrm{Cov}[f(Z), p_T(Z)] =& \sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} f(x) \cdot p_T(x) - (\sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} \cdot f(x))(\sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} \cdot p_T(x)) \\ =& \frac{1}{|\mathcal{X}|} - \frac{1}{|\mathcal{X}|} \cdot \frac{|T|}{\mathcal{X}} = \frac{1}{2|\mathcal{X}|} \end{aligned}$$

In both cases we see that the covariance between the labeling function and the target distribution is not 0 and there is a dependence between the two.

However we can slightly tweak the counterexample from Ben-David and Urner (2012) for it to fulfill the proposed condition, while still providing the same lower bound[3] for DA-learnability. For this we will introduce a function class $\mathcal{H}_C$ and a problem class $\mathcal{W}_{C,n}$ simultaneously, in such a way that $\mathcal{H}_C$ has two elements h_c and $|1 - h_c|$ and for each problem $(P_S, P_T) \in \mathcal{W}_{C,n}$ such that covariate shift holds and the labeling function f agrees with one element of $\mathcal{H}_C$ on the target support T agrees and with the other element on $\mathcal{H}_C$ on the rest of $\mathcal{X}$. For this function class and problem set, the following theorem holds.

Theorem 11 (Lower bound with first IGCI causal assumption)**.** *For every finite domain $\mathcal{X}$, for every ε and δ with $\varepsilon + \delta < \frac{1}{2}$, no algorithm can $(\varepsilon, \delta, s, t)$-solve the DA problem for the class $\mathcal{W}_{C,n}$ of pairs (P_S, P_T) satisfying the covariate shift with $C(\mathcal{D}_S, \mathcal{D}_T) \geq \frac{1}{2}$, $d_{\mathcal{H}_C \Delta \mathcal{H}_C}(\mathcal{D}_S, \mathcal{D}_T) = 0$, $opt_{P_T}(\mathcal{H}_C) = 0$ and the first IGCI criterion for binary classification, if*

$$s + t \leq \min \left\{ \sqrt{\ln(2) \frac{|\mathcal{X}|}{2}}, \sqrt{\ln\left(\frac{1}{2(\varepsilon + \delta)}\right) \frac{|\mathcal{X}|}{2}} \right\} - 1.$$

First – in order for this to be a meaningful theorem – we need to show that the construction of $\mathcal{H}_C$ and $\mathcal{W}_{C,n}$ is indeed possible. In the following we assume $n = |\mathcal{X}|$ to be divisible

[3] up to a factor of $\sqrt{\frac{1}{2}}$

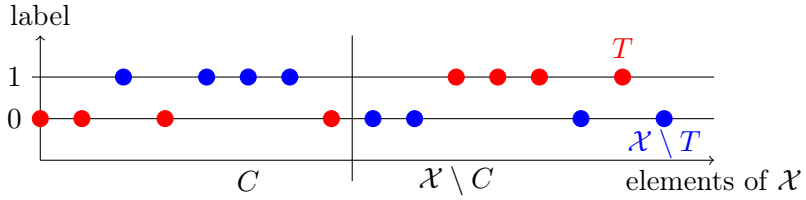


Figure 5.1.: An illustration of a pair $(P_S, P_T) \in \mathcal{W}_{C,n}$. The label of the target (marked red) support is "0" in C and "1" in $\mathcal{X} \setminus C$.

by 4. We can therefore partition $\mathcal{X}$ in the following way: $\mathcal{X} = \mathcal{X}_{a,t} \cup \mathcal{X}_{a,u} \cup \mathcal{X}_{b,t} \cup \mathcal{X}_{b,u}$, with pairwise disjoint $\mathcal{X}_{a,t}, \mathcal{X}_{a,u}, \mathcal{X}_{b,t}, \mathcal{X}_{b,u}$ and $|\mathcal{X}_{a,t}| = |\mathcal{X}_{a,u}| = |\mathcal{X}_{b,t}| = |\mathcal{X}_{b,u}|$. Let the target domain be $\mathcal{X}_t = \mathcal{X}_{a,t} \cup \mathcal{X}_{b,t}$ and the target distribution P_T be uniform over $\mathcal{X}_T$. Furthermore let $\mathcal{X}_a = \mathcal{X}_{a,t} \cup \mathcal{X}_{a,u}$ and $\mathcal{X}_b = \mathcal{X}_{b,t} \cup \mathcal{X}_{b,u}$. Now let the labeling function f be chosen in such a way that we have either $f(a) = 0$ and $f(b) = 1$ for every $a \in \mathcal{X}_a$ and $b \in \mathcal{X}_b$, or $f(a) = 1$ and $f(b) = 0$ for every $a \in \mathcal{X}_a$ and $b \in \mathcal{X}_b$. In the following we will now also refer to $\mathcal{X}_{a,t} \cup \mathcal{X}_{b,u}$ as C and to its complement $\mathcal{X}_{a,u} \cup \mathcal{X}_{b,t}$ as D. We can now construct our hypothesis class $\mathcal{H}_C = \{h_c, h_d\}$, with h_c chosen in such a way that $h_c(x) = 1$ for every $x \in C = \mathcal{X}_{a,t} \cup \mathcal{X}_{b,u}$ and $h_c(x) = 0$ for every $x \in D = \mathcal{X}_{a,u} \cup \mathcal{X}_{b,t}$ and $h_d(x)$ chosen in such a way that $h_d(x) = 0$ for every $x \in C = \mathcal{X}_{a,t} \cup \mathcal{X}_{b,u}$, and $h_d(x) = 1$ for every $x \in D = \mathcal{X}_{a,u} \cup \mathcal{X}_{b,t}$.
Now, we have constructed $\mathcal{W}_{C,n}$ and $\mathcal{H}_C$ properly and can continue to prove the theorem. An illustration of $\mathcal{W}_{C,n}$ can be seen in Figure 5.1.

Proof. First we need to show that all conditions of the theorem are fulfilled by $\mathcal{H}_C$ and $\mathcal{W}_{C,n}$. It is easy to see that we have target realizability for $\mathcal{H}_C$. Furthermore the VC-dimension of $\mathcal{H}_C$ is 1, since it only contains two functions. Since these two functions disagree in every point, we have $\mathcal{H}_C \Delta \mathcal{H}_C = \{h_0, h_1\}$, where h_0 is the function that is 0 everywhere and h_1 is the function that is 1 everywhere. Thus we have the same $\mathcal{H}\Delta\mathcal{H}$ as in the previous example. Furthermore, we see that $\mathcal{D}_S$ and $\mathcal{D}_T$ are constructed in the same way as in the previous counterexample. We therefore have the same $\mathcal{H}\Delta\mathcal{H}$-divergence $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) = 0$ and the same weight ratio $C(\mathcal{D}_S, \mathcal{D}_T) = \frac{1}{2}$.

Now let us examine whether or not we have independence of mechanism according to our previously defined criterion: Let us first look at the case where f agrees with h_d on the target domain:

$$\mathrm{Cov}[f(X), p_T(X)] = \sum_{x \in \mathcal{X}} \frac{f(x) \cdot p_T(x)}{|\mathcal{X}|} - \left(\sum_{x \in \mathcal{X}} \frac{f(x)}{|\mathcal{X}|} \right) \left(\sum_{x \in \mathcal{X}} \frac{p_T(x)}{|\mathcal{X}|} \right) =$$

$$= \sum_{x \in \mathcal{X}_{a,t}} \frac{f(x) \cdot p_T(x)}{|\mathcal{X}|} + \sum_{x \in \mathcal{X}_{b,t}} \frac{f(x) \cdot p_T(x)}{|\mathcal{X}|} - \frac{|B|}{|\mathcal{X}|} \cdot \frac{1}{|\mathcal{X}|} = 0 + \frac{1}{2|\mathcal{X}|} - \frac{1}{2|\mathcal{X}|} = 0$$

Now let us look at the case where f agrees with h_c in the target domain:

$$Cov[f(X), p_T(X)] = \sum_{x \in \mathcal{X}} \frac{f(x) \cdot p_T(x)}{|\mathcal{X}|} - \left(\sum_{x \in \mathcal{X}} \frac{f(x)}{|\mathcal{X}|} \right) \left(\sum_{x \in \mathcal{X}} \frac{p_T(x)}{|\mathcal{X}|} \right) =$$

$$= \sum_{x \in \mathcal{X}_{a,t}} \frac{f(x) \cdot p_T(x)}{|\mathcal{X}|} + \sum_{x \in \mathcal{X}_{b,t}} \frac{f(x) \cdot p_T(x)}{|\mathcal{X}|} - \frac{1}{2|\mathcal{X}g|} = \frac{1}{2|\mathcal{X}|} + 0 - \frac{1}{2|\mathcal{X}|} = 0$$

Thus our criterion holds in this example.
We now only need to show that this example gives the same lower bound as in Ben-David and Urner (2012). For a given class $\mathcal{H}_C$ and its corresponding problem class

$$\mathcal{W}_{C,n} := \{(P_S, P_T, f) | P_S \text{ is uniform over } \mathcal{X},\ P_T \text{ is uniform over } T,$$

$$f = \mathbb{1}_A, \text{ where } |A| = |B| = |T| = |U|, U = \mathcal{X} \setminus T, B = \mathcal{X} \setminus A,$$

$$C = (A \cap T) \cup (B \cap U) \text{ , or } C = (A \cap U) \cup (B \cap T)\}$$

we can make a similar argument as in Ben-David and Urner (2012), that is we can show that the Left/Right-problem can be reduced to the DA problem $\mathcal{W}_{C,n}$ and use the lower bound for the Left/Right-problem $\mathcal{W}^{uni}_{\frac{n}{2}}$ over a set $\mathcal{X}'$ of half the size of $\mathcal{X}$ to derive the lower bound for the DA-problem. Note that the target support T for all pairs $(P_S, P_T) \in \mathcal{W}_{C,n}$ cannot be any subset of size $\frac{n}{2}$, but has to fulfill $|T \cap C| = |T \cap (\mathcal{X} \setminus C)| = \frac{n}{4}$. Therefore there are fewer pairs in $\mathcal{W}_{C,n}$ than in $\mathcal{W}_n$ and $\mathcal{W}^{\text{uni}}_n$. We will therefore not be able to achieve the same lower bound for $\mathcal{W}_{C,n}$ as for $\mathcal{W}_n$, but only the same bound as for $\mathcal{W}_{\frac{n}{2}}$. In order to construct samples over $\mathcal{X}$ from samples over $\mathcal{X}'$, we will need two copies of the domain $\mathcal{X}'$ of size $\frac{n}{2}$, such that one copy corresponds to the set C in $\mathcal{X}$ and the other copy corresponds to $\mathcal{X} \setminus C$. Thus each element of $\mathcal{X}'$ is assigned to two elements of $\mathcal{X}$ – one in C and one in $\mathcal{X} \setminus C$. We will then take the samples over $\mathcal{X}'$ and assign each sample element randomly to its corresponding element in either C or $\mathcal{X} \setminus C$. We will now show how this construction works.

Analogous to the proof of the original theorem, we suppose the samples $L' = \{l_1, \ldots, l_s\}$ and $R' = \{r_1, \ldots, r_s\}$ (each of size s) and a sample $M'$of size $t+1$ are samples from the the Left/Right-problem, coming from a triple $(U_{\tilde{A}}, U_{\tilde{B}}, U_{\tilde{X}})$ of distributions in $\mathcal{W}^{uni}_{\frac{n}{2}}$. We will denote the subset of C corresponding to the subset $\tilde{A} \subset X'$ as $\tilde{A}_1$ and the corresponding subset in $\mathcal{X} \setminus C$ as $\tilde{A}_2$. Similarly, we denote the corresponding subsets of $\tilde{B} \subset \mathcal{X}'$ in C and $\mathcal{X} \setminus C$ as $\tilde{B}_1$ and $\tilde{B}_2$ respectively.
For a given hypothesis class $\mathcal{H}$ and its corresponding set C we will construct the labeled source sample S' and the unlabeled target sample T' over $\mathcal{X}$ as follows.

For the construction of T' let $M'' = M' \setminus \{p\}$ for some uniformly chosen p from M'. For every element of $m_j \in M''$ we then randomly choose the corresponding element of m_j in C or in $\mathcal{X} \setminus C$ with probability $\frac{1}{2}$ each. The chosen elements then form the unlabeled

target sample T' over $\mathcal{X}$. We see that T' is uniformly distributed over either $\tilde{A}_1 \cup \tilde{A}_2$ or uniformly distributed over $\tilde{B}_1 \cup \tilde{B}_2$. T' can therefore be seen as distributed by the marginal $\mathcal{D}_T$ from a distribution P_T from $\mathcal{W}_{C,n}$.

Now we will construct S' in a similar way. With probability $\frac{1}{2}$ we will choose the next element of out L' or our of R' respectively. This element of $\mathcal{X}'$ has two corresponding elements in $\mathcal{X}$. We will now choose with probability $\frac{1}{2}$ its corresponding element in C or its corresponding element in $\mathcal{X} \setminus C$. If we choose l_i from L' first, we will give it the label 1 if we then chose the corresponding element in C and the label 0 if we then chose the corresponding element in $\mathcal{X} \setminus C$. Correspondingly, if we choose r_i from R, we will give it the label 0 if we then chose the corresponding element in C and the label 1 otherwise.

The resulting sample S' can be viewed as coming from a uniform source distribution $\mathcal{D}_S$ over $\mathcal{X} = \tilde{A}_1 \cup \tilde{B}_1 \cup \tilde{A}_2 \cup \tilde{B}_2$ with a labeling function f, mapping points from $\tilde{A}_1 \cup \tilde{B}_2$ to 1 and points from $\tilde{A}_2 \cup \tilde{B}_1$ to 0. Furthermore, T' can be considered to come from a target distribution $\mathcal{D}_T$ that is either equal to $U_{\tilde{A}}$ or to $U_{\tilde{B}}$. We can see that $(P_S, P_T) \in \mathcal{W}_{C,n}$.

Now suppose there exists a DA-learner $\mathcal{A}$ that $(\varepsilon, \delta, s, t)$-solves the DA-problem $\mathcal{W}_{C,n}$. Furthermore suppose $\mathcal{A}$ outputs the hypothesis h. We can now construct a learner for the Left/Right-problem that has error ε with probability $1-\delta$ in the following way. If $h(p) = 1$ and $p \in C$ or $h(p) = 0$ and $p \notin C$ the Left/Right-solver outputs $U_{\tilde{A}}$ and otherwise the Left/Right-solver outputs $U_{\tilde{B}}$. With this reduction of the Left/Right-problem to the DA-problem (which works analogous to the proof of Lemma 44 in Urner (2013)), we can derive almost the same lower bound as in Theorem 41 in Urner (2013). □

Since this lower bound also depends on the size of $\mathcal{X}$, this lower bound implies an impossibility result for DA-learnability in the case where $\mathcal{X}$ is infinite. We therefore see that the criterion formulated in Assumption 1 is not sufficient to change the lower bound in a meaningful way.
A potential weak point in this argument could be that we did not capture the whole idea of the original IGCI-criterion, since we only considered f but not its derivative f'. We will now examine whether the bound changes if we require $\mathrm{Cov}_{Z\sim \mathrm{Uni}(\mathcal{X})}[f'(Z), p_T(Z)]$ to be 0 instead of $\mathrm{Cov}_{Z\sim \mathrm{Uni}(\mathcal{X})}[f(Z), p_T(Z)]$. But first we need to provide $\mathcal{X}$ with the necessary structure to define f'. Since $\mathcal{X}$ is not continuous, we need an embedding into a metric space $\mathcal{X}'$ where we can define a derivative f'. We now can consider the evaluations of f' in $\mathcal{X}$.

Assumption 2 (Second IGCI criterion for binary classification)**.** *Let $\mathcal{X}$ be a finite domain. If we have a causal labeling $X \to Y$, that is, the features X cause the labels Y, then we obtain the following relation between the labeling function f and the cause (target) distribution p_T:*

$$\mathrm{Cov}_{Z\sim Uni(\mathcal{X})}[f'(Z), p_T(Z)] = 0$$

where Z is a random variable that is uniformly distributed over the source domain $\mathcal{X}$.

We can again give the same lower bound as in the previous DA-problem with this additional criterion:

Theorem 12 (Lower bound with second IGCI causal assumption)**.** *For every finite domain $\mathcal{X}$, for every ε and δ with $\varepsilon+\delta < \frac{1}{2}$, no algorithm can $(\varepsilon, \delta, s, t)$-solve the DA problem for the class $\mathcal{W}_{C,n}$ of pairs (P_S, P_T) satisfying the covariate shift with $C(\mathcal{D}_S, \mathcal{D}_T) \geq \frac{1}{2}$, $d_{\mathcal{H}_C \Delta \mathcal{H}_C}(\mathcal{D}_S, \mathcal{D}_T) = 0$, $opt_{P_T}(\mathcal{H}_C) = 0$ and the second IGCI criterion for binary classification, if*

$$s + t \leq \min \left\{ \sqrt{\ln(2)|\mathcal{X}|}, \sqrt{\ln(\frac{1}{2(\varepsilon+\delta)})|\mathcal{X}|} \right\} - 1.$$

Proof. We can again take $\mathcal{W}_n$ and $\mathcal{H}_{(1,0)}$ from the counterexample from Urner (2013). For this hypothesis class and the corresponding DA-problem it has already been shown that $C(\mathcal{D}_S, \mathcal{D}_T) \geq \frac{1}{2}$, $d_{\mathcal{H}_C \Delta \mathcal{H}_C}(\mathcal{D}_S, \mathcal{D}_T) = 0$ and $opt_{P_T}(\mathcal{H}_C) = 0$. Furthermore the bound from Theorem 12 has already been shown for this problem class in Urner (2013) (see Theorem 10).

The only thing that is left to show is that the second IGCI criterion for causality in binary classification holds in this counterexample.
We can find an embedding $\mathcal{X} \to \mathcal{X}'$, such that for every $(P_S, P_T) \in \mathcal{W}_{C,n}$ we can expand the corresponding labeling function f in such a way that it is differentiable in $\mathcal{X}'$. This embedding into a continuous domain $\mathcal{X}'$ is only restricted by the n values of f in $\mathcal{X}$. We can therefore find an embedding such that f is differentiable and we have $f'(x) = 0$ for all $x \in \mathcal{X}$. Intuitively, we can take any smooth[4] function $f : \mathcal{X}' \to [0, 1]$ that "connects" the values of $f(x)$ with $x \in \mathcal{X}$.[5] Thus, we get

$$\begin{aligned}\mathrm{Cov}_{Z \sim \mathrm{Uni}(\mathcal{X})}[f'(Z), p_T(Z)] =& \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} f'(x) p_T(x) - \left(\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} f'(x) \right) \left(\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} p_T(x) \right) \\ =& 0 + \frac{1}{|\mathcal{X}|} \cdot 0 = 0.\end{aligned}$$

Therefore the second IGCI criterion for binary classification is fulfilled. □

Before ending this subsection dealing with the usefulness of the IGCI model for DA in binary classification, let us have a brief look at the other requirements of the original IGCI-model and discuss their usefulness and applicability for DA.

[4] continuous differentiable
[5] $f'(x) = 0$ for $x \in \mathcal{X}$ is then implied by the fact that there are local extrema in x for all $x \in \mathcal{X}$.

Looking back at the original IGCI model, one of its requirements was that the density of the cause is non-zero on the whole domain. If this holds for source and target distributions, this would imply that they have the same support. One of the major assumptions we made in the setting we were considering was target realizability. If target- and source support are the same, this assumption would also imply source realizability. Combined with covariate shift, this gives us a joint hypothesis error of 0. With the assumption that $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) = 0$, we get DA-learnability in this case according to Theorem 7.

But in this case the shared support was the only criterion used from the IGCI-model. It is questionable to what extend this is a criterion for causality, since one can imagine an intervention that excludes a certain range of values and that produces the target domain. However the IGCI does not deal with interventions and it is easy to imagine an intervention violating any of the IGCI conditions. In particular its core criterion – the covariance between $f'(X)$ and $p(X)$ being 0 – can easily be violated by intervention in a causal setting. But even if we think about shifts in distribution not happening by an intervention of the observer, but by some different mechanism, it still seems possible that this shift may also imply a shift in the support of the distribution. Therefore positive density does not seem to be a relevant aspect of the IGCI model to our problem.

The last criterion we have not examined yet is the monotonicity of the diffeomorphism f in the original IGCI-model. This translates to a monotonously increasing labeling function f in our binary classification setting. For this definition to make sense, we first need to have an ordering of $\mathcal{X}$. Let us assume in the following that $\mathcal{X}$ is an ordered set. If we are to keep the deterministic nature of our DA setting, f can only have values 0 and 1 on $\mathcal{X}$. This implies that the labeling function f has the form of a threshold function, i.e. there is a $x \in \mathcal{X}$ such that $f(x') = 0$ for all $x' < x$ and $f(x') = 1$ for all $x' \geq x$ (or $f(x') = 0$ for all $x' \in \mathcal{X}$). With this quite restricted structure, we can simply learn the threshold function (since the corresponding threshold hypothesis class has VC-dimension 1) in the source-domain and select the hypothesis from $\mathcal{H}$ fitting the learned threshold function best in target domain. This will lead to the following theorem.

Theorem 13. *Let $\mathcal{X}$ be an ordered finite domain. Let furthermore $\mathcal{H}$ be a hypothesis class with finite VC-dimension d. There is a DA-learner outputting elements of $\mathcal{H}$ that is able to solve the DA-problem for any pair (P_S, P_T) of source- and target distributions P_S and P_T satisfying covariate shift with $C(\mathcal{D}_S, \mathcal{D}_T) \geq \frac{1}{2}$, $opt_{P_T}(\mathcal{H}_C) = 0$ and the corresponding labeling function f being deterministic and monotonously increasing. In particular there exists a constant c (independent of $|\mathcal{X}|$), such that for every $\varepsilon, \delta > 0$ for sample sizes $s \geq c\frac{1+\log(\frac{2}{\delta})}{(\frac{\varepsilon}{4})^2}$ and $t \geq c\frac{d\log(\frac{2}{\varepsilon})+\log(\frac{1}{\delta})}{(\frac{\varepsilon}{2})^2}$, with probability $1-\delta$ we have*

$$\mathbb{E}_{x\sim\mathcal{D}_T}[\mathcal{L}_{P_T}(\mathcal{A}(S,T))] \leq \varepsilon$$

Proof. Let $\varepsilon' = \frac{\varepsilon''}{2} = \frac{\varepsilon}{4}$ and $\delta' = \delta'' = \frac{\delta}{2}$. First we note that a monotonous deterministic

labeling function leads to realizability within the hypothesis class $\mathcal{H}_{thres}$ of threshold functions. We know that this hypothesis class has VC-dimension 1. According to Theorem 3, we can therefore learn the problem in the source domain up to error ε with probability $1-\delta'$ with a sample-complexity of $c\frac{1+\log\frac{1}{\delta}}{\varepsilon'^2}$ in the source domain. Since we have a weight ratio $C(\mathcal{D}_S, \mathcal{D}_T) \geq \frac{1}{2}$, we can now use Observation 1 to derive that we get error $2\varepsilon'$ with probability $1-\delta'$ for the same h' we learned with $c\frac{1+\log\frac{1}{\delta'}}{\varepsilon'^2}$ source samples. Now we only need to find an $h \in \mathcal{H}$ that agrees with the learned $h' \in \mathcal{H}_{thres}$ on the target domain. We will therefore label the unlabeled target sample U with the learned hypothesis $h' \in \mathcal{H}_{thres}$ and take the resulting labeled data U' as input for an empirical risk minimization algorithm $\mathcal{A}'$ for hypothesis class $\mathcal{H}$. According to Theorem 3 the labeling h' on $\mathcal{D}_T$ can be learned up to error ε'' with probability $1-\delta''$ with a sample complexity of $c\frac{d\log(\frac{1}{\varepsilon''})+\log\frac{1}{\delta''}}{\varepsilon''^2}$. Therefore $\mathcal{A}'(U')$ has error ε with probability $1-\delta$ if we the source sample has at least $c\frac{1+\log\frac{2}{\delta}}{(\frac{\varepsilon}{4})^2}$ elements and if the target sample has at least $c\frac{d\log(\frac{2}{\varepsilon})+\log\frac{2}{\delta}}{(\frac{\varepsilon}{2})^2}$ elements.

□

This theorem shows that our previous lower bound that depended on the domain size $|\mathcal{X}|$ no longer holds if we assume the labeling function to be monotonous. Note that we did not make any assumption about causal direction or the dependence or independence of cause and mechanism.

In conclusion, we did not find a criterion inspired by the IGCI model for causality that made the DA problem easy for the binary classification case.

Since our formulated criterion is only a statement about covariance, it still does not necessarily imply that there is no dependence between the labeling function and the probability distribution of the target domain. Indeed, both target distribution and labeling function together with the function class $\mathcal{H}_C$ were constructed in a very dependent way, in order for the argument from Ben-David and Urner (2012) to hold. We will explore other formalizations later.

Before we discuss the regression case, we will briefly discuss whether our two criteria are good models of causality. As mentioned after introducing the IGCI model in Chapter 3 the model is best justified by a particular kind of generating process for f', where f' is generated as a piecewise constant function with $f'(x) = r_j$ for $x \in [\frac{j}{n}, \frac{j+1}{n})$ with r_j being independently and identically distributed. Under the assumption that one has chosen a good reference distribution U for P, this then implies that $\mathrm{Cov}_{Z\sim U}[f'(Z), p(X)]$ is small with high probability as Theorem 1 tells us. For our data-generating process, one could argue that in our case we have $f(j) = r_j$ with r_j being distributed according to Bernoulli distributions with probability $\frac{1}{2}$. However in our examples one could not argue, that

these r_j are independently distributed, since in our construction of $\mathcal{W}_{C,n}$, for a given p_T, $f(j) = 1$ implies $f(i) = 1$ for certain $i, j \in \mathcal{X}$. This would lead to the conclusion, that our construction, while fulfilling IGCI inspired correlation criteria, does not follow a data-generating process in which the generation of the labeling function f is independent from source and target distributions. Which in turn, by the Principle of Independence of Cause and Mechanism, would suggest that either Y is a cause of X or that X and Y have a common cause. This excludes the possibility that the only causal relation between X and Y is X being a direct cause for Y, which is the causal relation that our IGCI inspired assumptions should have modeled. However, if we do assume f to be generated by $f(j) = r_j$ with r_j being i.i.d from Bern($\frac{1}{2}$) and consider the problem class $\mathcal{W}'$ of all possible DA-problems that could result from that process, we get $\mathcal{W}_{C,n} \subset \mathcal{W}'$. Therefore, a counterexample for the DA-learnability for $\mathcal{W}_{C,n}$ would serve as a counterexample for the DA-learnability of $\mathcal{W}'$.

We note that $\mathcal{D}_S = \text{Uni}(\mathcal{X})$ does not seem to be a good reference distribution for $\mathcal{D}_T$, in the sense that $\int |p_S(x) - p_T(x)|dx$ is small. This is in contrast to an intuition that might justify the IGCI model in a causal setting. The assumption that $\int |p_S(x) - p_T(x)|dx$ is small, however, would again be a rather strong. It would for example imply that $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ is also small. Furthermore it would also imply that the risk under P_S and under P_T are close for any hypothesis $h \in \mathcal{H}$ if we have covariate shift. This would furthermore imply low joint hypothesis error, if we assume target realizability. Therefore $\int |p_S(x) - p_T(x)|dx$ being small makes the DA problem easy on its own.

IGCI model in regression

Before looking at other possible models for Independence of Cause and Mechanism, we will move away from binary classification and consider the IGCI assumptions in the regression case. We will again try to realize as many of the assumptions made by the original IGCI model as possible. Therefore we will only look at regression problems from domain $[0, 1]$ to $[0, 1]$. First, we will consider a non-monotonous and non-differentiable example which will work similar to our previous counterexamples.

We will again introduce two possible criteria for causality in the regression setting.

Assumption 3 (First IGCI criterion for regression)**.** *If we have a causal labeling $X \to Y$, that is, the features X cause the labels Y, then we obtain the following relation between the regression function f and the cause (target) distribution p_T:*

$$\text{Cov}_{Z \sim Uni([0,1])}[f(Z), p_T(Z)] = 0$$

where Z is a random variable that is uniformly distributed over the source domain $\mathcal{X}$.

Assumption 4 (Second IGCI criterion for regression)**.** *If we have a causal labeling $X \to Y$, that is, the features X cause the labels Y, then there is the following relation between the regression function[6] f and the cause (target) distribution p_T,*

$$\mathrm{Cov}_{Z\sim Uni([0,1])}[f'(Z), p_T(Z)] = 0$$

where Z is a random variable that is uniformly distributed over the source domain $\mathcal{X}$.

We will again construct an example where DA-learning is hard and where both IGCI criteria for regression as well as all the assumptions we had in Theorem 10 (like covariate shift, high weight ratio and low $\mathcal{H}$-divergence) hold.

Since we are now in the regression case, we first need to clarify what kind of loss function we will use. In the following section we will consider both 0-1-loss and ℓ^2-loss. In case of the ℓ^2-loss we will have to replace the assumption about $\mathcal{H}$-divergence with an assumption about its generalization the discrepancy distance as described in Definition 22. Let $\mathcal{L}^2_Q$ denote risk for the ℓ^2-loss under distribution Q. For the ℓ^2-loss the discrepancy distance of two distribution Q_1, Q_2 is

$$\begin{aligned}\mathrm{disc}_{\ell^2}(Q_1, Q_2) &= \max_{h,h'\in\mathcal{H}} |\mathcal{L}^2_{Q_1}(h', h) - \mathcal{L}^2_{Q_2}(h', h)| \\ &= \max_{h,h'\in\mathcal{H}} |\mathbb{E}_{x\sim Q_1}[|h(x) - h'(x)|^2]| - \mathbb{E}_{x\sim Q_2}[|h(x) - h'(x)|^2]|.\end{aligned}$$

We will now go on to construct a problem class $\mathcal{W}_{reg,n}$ of pairs of distributions (P_S, P_T) and its corresponding hypothesis class $\mathcal{H}_{C,reg}$.

In the following we will again assume n to be divisible by 4. Consider a partition of the domain $[0,1] = [0, \frac{1}{n}) \cup [\frac{1}{n}, \frac{2}{n}) \cup \cdots \cup [\frac{n-1}{n}, 1]$ into n intervals of the same size. Now consider only functions from $[0,1]$ to $[0,1]$ that map each of these intervals identically to either constant 0 or constant 1. Now let the (marginal) source distribution $\mathcal{D}_S$ be uniform over $[0,1]$ and the (marginal) target distribution $\mathcal{D}_T$ uniform over T a union of $\frac{n}{4}$ intervals $[\frac{i}{n}, \frac{i+1}{n})$ with $0 \le i \le \frac{n}{2} - 1$ and $\frac{n}{4}$ intervals $[\frac{j}{n}, \frac{j+1}{n})$ with $\frac{n}{2} \le j \le n - 1$. Now let $\mathcal{H}_{C,reg} = \{h_1, h_2\}$ with $h_1(x) = 0$ and $h_2(x) = 1$ for all $x \in [0, \frac{1}{2})$ and $h_1(x') = 1$ and $h_2(x') = 0$ for all $x \in [\frac{1}{2}, 1]$. Furthermore we will again assume P_S and P_T to have the same regression function, i.e., covariate shift holds. Now let this regression function $f : [0,1] \to [0,1]$ be defined by either

$$f(x) = \begin{cases} 0 & \text{, if } x \in ([0, \frac{1}{2}) \cap T) \cup ([\frac{1}{2}, 1] \setminus T) \\ 1 & \text{, if } x \in ([0, \frac{1}{2}) \setminus T) \cup ([\frac{1}{2}, 1] \cap T) \end{cases}$$

or

$$f(x) = \begin{cases} 1 & \text{, if } x \in ([0, \frac{1}{2}) \cap T) \cup ([\frac{1}{2}, 1] \setminus T) \\ 0 & \text{, if } x \in ([0, \frac{1}{2}) \setminus T) \cup ([\frac{1}{2}, 1] \cap T) \end{cases}$$

This construction follows the same idea as the previous counterexamples. We will now consider the learning problem $\mathcal{W}_{reg,n}$ consisting of all pairs (P_S, P_T) that result from a construction as described above. This DA learning problem is similar to $\mathcal{W}_{C,n}$.

[6] which we assume to be differentiable almost everywhere in [0,1]

Theorem 14 (DA-Hardness for regression under causal assumptions)**.** *Consider the regression DA-problem $\mathcal{W}_{reg,n}$ with the function class $\mathcal{H}_{C,reg}$ under either 0-1-loss or ℓ^2-loss. For every finite number[7] n there is no algorithm that can $(\varepsilon, \delta, s, t)$-solve the DA problem for the class $\mathcal{W}_{reg,n}$ of pairs (P_S, P_T) satisfying the covariate shift, $C(\mathcal{D}_S, \mathcal{D}_T) \geq \frac{1}{2}$, $d_{\mathcal{H}_{C,reg}\Delta\mathcal{H}_{C,reg}}(\mathcal{D}_S, \mathcal{D}_T) = 0$, $disc_{\ell^2}(\mathcal{D}_S, \mathcal{D}_T) = 0$, $opt_{P_T}(\mathcal{H}_{C,reg}) = 0$ and the first and second IGCI criteria for regression, if*

$$s + t \leq \min\left\{\sqrt{\ln(2)\frac{n}{2}}, \sqrt{\ln(\frac{1}{2(\varepsilon+\delta)})\frac{n}{2}}\right\} - 1$$

Proof. First we will show that for problems in $\mathcal{W}_{reg,n}$ with $\mathcal{H}_{C,reg}$ all of our assumptions hold. The pairs (P_S, P_T) of $\mathcal{W}_{reg,n}$ were constructed in such a way that covariate shift, $C(\mathcal{D}_S, \mathcal{D}_T) \geq \frac{1}{2}$ and $opt_{P_T}(\mathcal{H}_{C,reg}) = 0$ (for both 0-1-loss and ℓ^2-loss) hold. For the calculation of the $\mathcal{H}$-divergence we first note that $\mathcal{H}_{C,reg}\Delta\mathcal{H}_{C,reg} = \mathcal{H}_{0,1}$, since $\mathcal{H}_{C,reg}$ only contains two functions, which are each other's opposites. It is easy to see that $d_{\mathcal{H}_{0,1}}(Q_1, Q_2) = 0$ for all distributions Q_1, Q_2. In particular, $d_{\mathcal{H}_{C,reg}\Delta\mathcal{H}_{C,reg}}(\mathcal{D}_S, \mathcal{D}_T) = 0$. For the discrepancy distance we note that all elements of $\mathcal{H}_{C,reg}$ have only values in $\{0, 1\}$, therefore, we have $|h(x) - h'(x)|^2 = \mathbb{1}[h(x) \neq h'(x)]$ for all $h, h' \in \mathcal{H}_{C,reg}$. Thus, we have $\text{disc}_{\ell^2}(\mathcal{D}_S, \mathcal{D}_T) = d_{\mathcal{H}_{C,reg}\Delta\mathcal{D}_{C,reg}}(\mathcal{D}_S, \mathcal{D}_T) = 0$. For the covariance from the first IGCI criterion we get

$$\begin{aligned}
&\text{Cov}_{Z\sim\text{Uni}([0,1])}[f(Z), p_T(Z)] \\
=&\int_0^1 f(z)p_T(z)dz - \left(\int_0^1 f(z)dz\right)\underbrace{\left(\int_0^1 p_T(z)dz\right)}_{=1} \\
=&\int_0^1 f(z)p_T(z)dz - \int_0^1 f(z)dz \\
=&\sum_{i=0}^{n-1} \underbrace{\left(\int_{\frac{i}{n}}^{\frac{i+1}{n}} f(z)p_T(z)dz\right)}_{=0 \text{ for } i \text{ with } [\frac{i}{n}, \frac{i+1}{n}) \not\subset T \cap f^{-1}(1)} - \frac{1}{2} \\
=&\sum_{i\in\{1\ldots n-1\}:[\frac{i}{n}, \frac{i-1}{n})\subset T\cap f^{-1}(1)} \underbrace{\left(\int_{\frac{i}{n}}^{\frac{i+1}{n}} f(z)p_T(z)dz\right)}_{=\frac{2}{n}} - \frac{1}{2} \\
=&\frac{n}{4}\cdot\frac{2}{n} - \frac{1}{2} = 0
\end{aligned}$$

[7] that is divisible by 4

Thus the first ICGI criterion for regression is fulfilled. For the second ICGI criterion the calculation is quite similar:

$$
\begin{aligned}
&\mathrm{Cov}_{Z\sim\mathrm{Uni}([0,1])}[f'(Z), p_T(Z)] = \\
=&\int_0^1 f'(z)p_T(z)dz - \underbrace{\left(\int_0^1 f'(z)dz\right)}_{=1}\underbrace{\left(\int_0^1 p_T(z)dz\right)}_{=1} \\
=&\int_0^1 f'(z)p_T(z)dz - 1 \\
=&\sum_{i=0}^{n-1} \underbrace{\left(\int_{\frac{i}{n}}^{\frac{i+1}{n}} f(z)p_T(z)dz\right)}_{=0 \text{ for } i \text{ with } [\frac{i}{n},\frac{i+1}{n})\not\subset T} -1 \\
=&\sum_{i\in\{1\ldots n-1\}:[\frac{i}{n},\frac{i-1}{n})\subset T} \underbrace{\left(\int_{\frac{i}{n}}^{\frac{i+1}{n}} f'(z)p_T(z)dz\right)}_{=\frac{2}{n}} \quad - \quad 1 \\
=&\frac{n}{2}\cdot\frac{2}{n} - 1 = 0
\end{aligned}
$$

Therefore also the second ICGI criterion for regression holds.

Finally, we show that we can provide the same lower bound for the sample sizes s and t of the labeled source sample L and of the unlabeled target sample U respectively, where we substitute the domain size $|\mathcal{X}|$ by the number of intervals n. First, we again note that since both f and all elements of $\mathcal{H}_{C,reg}$ as well as only take values in $\{0,1\}$, the ℓ^2-loss becomes equivalent to the 0-1-loss: $|f(x)-h(x)|^2 = \mathbb{E}_{y\sim\mathrm{Bern}(f(x))}\mathbb{1}[h(x)\neq f(x)]$. It therefore suffices to prove the bound only for the 0-1-loss. Now consider a random mapping $g : \mathcal{X}\times\{0,1\} \to [0,1]\times\{0,1\}$ that maps every element $(x_i, y_i) \in \mathcal{X}\times\{0,1\}$ to a random variable that is uniformly distributed on $[\frac{i}{n}, \frac{i+1}{n}) \times \{y_i\}$ with $x_i \neq x_j$ for $i \neq j$. Under this mapping every learning problem $(P_S, P_T) \in \mathcal{W}_{C,n}$ with hypothesis class $\mathcal{H}_C$ has a corresponding learning problem $(P'_S, P'_T) \in \mathcal{W}_{reg,n}$ with hypothesis class $\mathcal{H}_{C,reg}$, in the sense that if one could solve the problem $(P'_S, P'_T) \in \mathcal{W}_{reg,n}$ it provides an equally good solution for $(P_S, P_T) \in \mathcal{W}_{C,n}$. Therefore a lower bound for the DA-learnability for $\mathcal{W}_{C,n}$ with $\mathcal{H}_C$ must then also hold for $\mathcal{W}_{reg,n}$ with $\mathcal{H}_{C,reg}$. Thus, by using Theorem 11 we achieve the lower bound as stated in the theorem. □

If the number of intervals n goes to infinity this theorem implies that the corresponding DA-problem cannot be learned with a finite number of samples. However, we still have not shown that we do not get DA-learnability in the original IGCI model, since we have not looked at a monotonous and differentiable example. We will finish

this section by providing a problem class $\mathcal{W}_{mon,n}$ of pairs (P_S, P_T) for which covariate shift and $C(\mathcal{D}_T, \mathcal{D}_S) \geq \frac{1}{2}$ hold and such that the corresponding regression function $f : [0,1] \to [0,1]$ is strictly monotonously increasing, continuous and almost everywhere differentiable. Furthermore we will provide a corresponding function class $\mathcal{H}_{C,mon}$ of strictly monotonous, continuous and almost everywhere differentiable function for which we have target realizability and $d_{\mathcal{H}_{C,mon}\Delta\mathcal{D}_{C,mon}}(\mathcal{D}_S, \mathcal{D}_T) = 0$ and $\text{disc}_{\ell^2}(\mathcal{D}_S, \mathcal{D}_T) = 0$. We will go on to prove a lower bound for the sizes of source and target samples necessary to solve the DA problems of $\mathcal{W}_{mon,n}$.

The marginal source and target distributions $\mathcal{D}_S$ and $\mathcal{D}_T$ of pairs $(P_S, P_T) \in \mathcal{W}_{mon,n}$ will be the same as for the pairs $(P'_S, P'_T) \in \mathcal{W}_{reg,n}$, with the same construction of the set T. For a given set T the regression function for (P_S, P_T) will now be defined by either

$$f(x) = \begin{cases} \frac{i}{n} + n(x - \frac{i}{n})^2 & \text{for } x \in [\frac{i}{n}, \frac{i+1}{n}) \subset ([0, \frac{1}{2}) \cap T) \cup ([\frac{1}{2}, 1] \setminus T) \\ \frac{i+1}{n} - n(\frac{i+1}{n} - x)^2 & \text{for } x \in [\frac{i}{n}, \frac{i+1}{n}) \subset ([0, \frac{1}{2}) \setminus T) \cup ([\frac{1}{2}, 1] \cap T) \end{cases}$$

or

$$f(x) = \begin{cases} \frac{i+1}{n} - n(\frac{i+1}{n} - x)^2 & \text{for } x \in [\frac{i}{n}, \frac{i+1}{n}) \subset ([0, \frac{1}{2}) \cap T) \cup ([\frac{1}{2}, 1] \setminus T) \\ \frac{i}{n} + n(x - \frac{i}{n})^2 & \text{for } x \in [\frac{i}{n}, \frac{i+1}{n}) \subset ([0, \frac{1}{2}) \setminus T) \cup ([\frac{1}{2}, 1] \cap T) \end{cases}$$

The hypothesis class $\mathcal{H}_{C,mon} = \{h_1, h_2\}$ will again consist of two functions that are opposed to each other. In this case they will be defined in the following way:

$$h_1(x) = \begin{cases} \frac{i}{n} + n(x - \frac{i}{n})^2 & \text{for } x \in [\frac{i}{n}, \frac{i+1}{n}) \subset ([0, \frac{1}{2}) \\ \frac{i+1}{n} - n(\frac{i+1}{n} - x)^2 & \text{for } x \in [\frac{i}{n}, \frac{i+1}{n}) \subset [\frac{1}{2}, 1] \end{cases}$$

and

$$h_2(x) = \begin{cases} \frac{i+1}{n} - n(\frac{i+1}{n} - x)^2 & \text{for } x \in [\frac{i}{n}, \frac{i+1}{n}) \subset ([0, \frac{1}{2}) \\ \frac{i}{n} + n(x + \frac{i}{n})^2 & \text{for } x \in [\frac{i}{n}, \frac{i+1}{n}) \subset [\frac{1}{2}, 1] \end{cases}$$

To get a better intuition for this problem class and hypothesis class we would refer the reader to Figure 5.2 which provides some illustrations. For this problem class the following theorem holds.

Theorem 15 (Hardness of domain adaptation for the IGCI model)**.** *For every finite number*[8] *n there is no algorithm with outputs in $\mathcal{H}_{C,mon}$ that can $(\varepsilon, \delta, s, t)$-solve the DA problem under the 0-1-loss for the class $\mathcal{W}_{mon,n}$ of pairs (P_S, P_T) satisfying the covariate shift, $C(\mathcal{D}_S, \mathcal{D}_T) \geq \frac{1}{2}$, $d_{\mathcal{H}_{C,mon}\Delta\mathcal{H}_{C,mon}}(\mathcal{D}_S, \mathcal{D}_T) = 0$, $disc_{\ell^2}(\mathcal{D}_S, \mathcal{D}_T) = 0$, $opt_{P_T}(\mathcal{H}_{C,mon}) = 0$ and the second IGCI criterion for regression, if*

$$s + t \leq \min\left\{\sqrt{\ln(2)\frac{n}{2}}, \sqrt{\ln\left(\frac{1}{2(\varepsilon + \delta)}\right)\frac{n}{2}}\right\} - 1$$

For the ℓ^2-loss we have $\mathcal{L}_{P_T}(h) \leq \frac{2}{15n^2}$ for all $h \in \mathcal{H}_{C,mon}$. But for $\varepsilon < \frac{2}{15n^2}$ the same bound holds as lower bound for the sample complexity needed to solve the DA-problem

[8] that is divisible by 4

under the ℓ^2-loss, i.e., no algorithm with outputs in $\mathcal{H}_{C,n}$ can $(\varepsilon, \delta, s, t)$-solve the DA problem under the ℓ^2-loss, if

$$s + t \leq \min\left\{\sqrt{\ln(2)\frac{n}{2}}, \sqrt{\ln\left(\frac{1}{2\,(\varepsilon+\delta)}\right)\frac{n}{2}}\right\} - 1.$$

Proof. The proof of this theorem works analogous to the proof of Theorem 14 for the 0-1-loss: Let $\mathcal{X} = \{x_0, x_2, \ldots, x_{n-1}\}$ and $C = \{x_0, \ldots, x_{\frac{n}{2}-1}\}$. We can define a random mapping $g : \mathcal{X} \times \{0,1\} \to [0,1] \times 0, 1$, that maps every element $(x_i, y_i) \in \mathcal{X} \times \{0,1\}$ to a random vector $U_i = (X_i, Y_i)$ such that X_i is uniformly distributed over $[\frac{i}{n}, \frac{i+1}{n})$ with $x_i \neq x_j$ for $i \neq j$ and

$$Y_i = \begin{cases} \frac{i+1}{n} - n(\frac{i+1}{n} - X_i)^2 & \text{if } y_i = 0 \\ \frac{i}{n} + n(X_i + \frac{i}{n})^2 & \text{if } y_i = 1. \end{cases}$$

Under this mapping every learning problem $(P_S, P_T) \in \mathcal{W}_{C,n}$ with hypothesis class $\mathcal{H}_C$ has a corresponding learning problem $(P'_S, P'_T) \in \mathcal{W}_{\text{mon},n}$ with hypothesis class $\mathcal{H}_{C,\text{mon}}$ in the sense that a solution for the problem $(P'_S, P'_T) \in \mathcal{W}_{\text{mon},n}$ provides an equally good solution for $(P_S, P_T) \in \mathcal{W}_{C,n}$. Therefore a lower bound for the DA-learnability of $\mathcal{W}_{C,n}$ with $\mathcal{H}$ must then also hold for $\mathcal{W}_{\text{mon},n}$ with $\mathcal{H}_{\text{mon},C}$.

For the ℓ^2-loss basically the same arguments hold, since it is equally hard to distinguish between the two hypotheses in $\mathcal{H}_{C,mon}$ in both the 0-1-loss and for the ℓ^2-loss. However the loss term gets smaller since for every interval $[\frac{i}{n}, \frac{i+1}{n})$ we have

$$\begin{aligned}
\int_{\frac{i}{n}}^{\frac{i+1}{n}} |h(x) - f(x)|^2\, dx &\leq \int_{\frac{i}{n}}^{\frac{i+1}{n}} \left| \frac{i}{n} + n\left(x - \frac{i}{n}\right)^2 - \left(\frac{i+1}{n} - n\left(\frac{i+1}{n} - x\right)^2\right)\right|^2 dx \\
&= \int_0^{\frac{1}{n}} \left| nx^2 - \left(\frac{1}{n} - n\left(\frac{1}{n} - x\right)^2\right)\right|^2 dx = \int_0^{\frac{1}{n}} \left|2nx^2 - 2x\right|^2 dx \\
&= \int_0^{\frac{1}{n}} 4|n^2x^4 - 2nx^3 + x^2| dx = 4\left[\frac{1}{5}n^2x^5 - \frac{1}{2}nx^4 + \frac{1}{3}x^3\right]_0^{\frac{1}{n}} \\
&= 4\left|\frac{n^2}{5} \cdot (\frac{1}{n})^5 - \frac{n}{2}(\frac{1}{n})^4 + \frac{1}{3}(\frac{1}{n})^3\right| = \frac{4(6-15+10)}{30n^3} = \frac{2}{15n^3}
\end{aligned}$$

for every $h \in \mathcal{H}_{C,mon}$ and a regression function from $\mathcal{W}_{mon,n}$. Therefore we get

$$\mathcal{L}_{P_T}(h) = \int_0^1 p_T(x)|f(x) - h(x)|^2 dx = \sum_{i=0}^{n-1} \int_{\frac{i}{n}}^{\frac{i+1}{n}} p_T(x)|f(x) - h(x)|^2 dx \leq \frac{n}{2} \cdot 2 \cdot \frac{2}{15n^3} = \frac{2}{15n^2}.$$

To be more precise, we either have $\mathcal{L}_{P_T}(h_1) = \frac{2}{15n^2}$ and $\mathcal{L}_{P_T}(h_2) = 0$ or $\mathcal{L}_{P_T}(h_1) = 0$ and $\mathcal{L}_{P_T}(h_2) = \frac{2}{15n^2}$. Thus $\mathcal{L}_{P_T}(h) < \frac{2}{15n^2}$ implies $mathcal L_{P_T}(h) = 0$ for $h \in \mathcal{H}_{C,\text{mon}}$.

In this case the risk of h under the 0-1-loss is also 0. Therfore, if a DA-learner $\mathcal{A}$ was able to $(\varepsilon, \delta, s, t)$-solve $\mathcal{W}_{\text{mon},n}$ for $\varepsilon < \frac{2}{15n^2}$ with respect to the ℓ^2-loss, then $\mathcal{A}$ would also $(\varepsilon, \delta, s, t)$-solve $\mathcal{W}_{\text{mon},n}$ with respect to the 0-1-loss. Therefore we get the same lower bound for $\mathcal{W}_{\text{mon},n}$ under the ℓ^2-loss for $\varepsilon < \frac{2}{15n^2}$.

Lastly, we have to check, that all the conditions hold. Obviously, we have covariate shift, target realizability, and a weight ratio of $\frac{1}{2}$.
For $d_{\mathcal{H}_{C,\text{mon}}\Delta\mathcal{H}_{C,\text{mon}}}(\mathcal{D}_S, \mathcal{D}_T)$, we again note that $\mathcal{H}_{C,\text{mon}}\Delta\mathcal{H}_{C,\text{mon}} = \mathcal{H}_{(1,0)}$, which makes $d_{\mathcal{H}_{C,\text{mon}}\Delta\mathcal{H}_{C,\text{mon}}}(\mathcal{D}_S, \mathcal{D}_T) = 0$ trivial.
The discrepancy distance is also easy to calculate, since there are only two hypotheses in $\mathcal{H}_{C,\text{mon}}$ $\text{disc}_{\ell^2}(\mathcal{D}_S, \mathcal{D}_T) = |\mathcal{L}_{\mathcal{D}_S}(h_1, h_2) - \mathcal{L}_{\mathcal{D}_T}(h_1, h_2)| = |\frac{2}{15n^2} - \frac{2}{15n^2}| = 0$.

The last remaining condition is $\text{Cov}_{Z\sim\text{Uni}([0,1])}[f'(Z), p_T(Z)] = 0$. For every of the n intervals we get

$$\int_{\frac{i}{n}}^{\frac{i+1}{n}} f'(x)dx = f\left(\frac{i+1}{n}\right) - f\left(\frac{i}{n}\right) = \frac{i+1}{n} - \frac{i}{n} = \frac{1}{n}.$$

Therefore

$$\begin{aligned}
&\text{Cov}_{Z\sim\text{Uni}([0,1])}[f'(Z), p_T(Z)] = \int_0^1 f'(x)p_T(x)dx - \left(\int_0^1 p_T(x)dx\right)\left(\int_0^1 f'(x)dx\right) \\
&= \sum_{i:[\frac{i}{n},\frac{i+1}{n})\subset T} \int_{\frac{i}{n}}^{\frac{i+1}{n}} f'(x)p_T(x)dx + \sum_{i:[\frac{i}{n},\frac{i+1}{n})\not\subset T} \int_{\frac{i}{n}}^{\frac{i+1}{n}} f'(x)p_T(x)dx - 1\cdot 1 \\
&= \frac{n}{2}\cdot(\frac{1}{n}\cdot 2) + \frac{n}{2}\cdot(\frac{1}{n}\cdot 0) - 1 = 1 + 0 - 1 = 0.
\end{aligned}$$

□

Note that for $n \to \infty$ the ℓ^2-loss becomes 0 for every choice of $h \in \mathcal{H}_{C,mon}$ and is therefore easy to learn. However, this fact does not change, if we allow similarly constructed regression functions as in $\mathcal{W}_{mon,n}$ that would violate the second IGCI criterion for regression. Therefore the reason the risk is low in this case does not seem to be causality.

There are only two conditions of the original IGCI model left that are not met in the previous theorem. The first of these conditions is that the mechanism $f : [0,1] \to [0,1]$ is supposed to be a diffeomorphism. However, in our bound it is only differentiable almost everywhere. But we could easily construct a similar function that is differentiable everywhere and that works in the same way, since f is continuous everywhere and differentiable in all but a finite number of points.

Lastly, the IGCI model assumed the density of the cause to be positive in everywhere in $[0,1]$, but our construction violated this for the target distribution. However, if we

did not allow the density of the target distribution to be 0 anywhere in $[0, 1]$, target realizability would imply source realizability. In this case we get DA-learnability, but again this is not due to any causal criterion.
In conclusion the IGCI model does not seem to help to get rid of the lower bounds as provided in Ben-David and Urner (2012), Urner (2013), unlike we initially hoped.

Independence of Cause and Mechanism as Statistical Independence

Another way of formulating the independence of cause and mechanism could be the definition of a meta-distribution over the marginal-distribution over pairs $(\mathcal{D}_T, f)$ and the postulation of statistical independence between (the random variables) $\mathcal{D}_T$ and f. To give some intuition for this, one can imagine that the mechanism and the distribution change between data sets. However, if we are in a causal scenario the changes of the marginal distribution $\mathcal{D}$ and the labeling function f would not affect one another. Thus, given a mechanism every distribution of the cause will be as likely as if the mechanism was not given. To check if this statistical independence condition is fulfilled, we would first need to define a meta-distribution. This is were we encounter a problem: Since only one target and one source distribution was given, this leaves us very little information to infer a meta-distribution. Therefore this meta-distribution will likely only have theoretical relevance and not result in a condition that can be checked empirically, if we only have one source domain.
Given some problem set, it might seem most natural, to define the meta-distribution as a uniform distribution over all problems considered, if possible. Indeed, in the setup of our counterexample this is possible, since the problem set $\mathcal{W}_{C,n}$ from Theorem 11 is a finite set.
But does statistical independence hold in this case? Unfortunately not, since for some pairs $(\mathcal{D}_T, f)$ the corresponding distribution P_T is not contained in $\mathcal{W}_{C,n}$, while there might be another pair (P'_S, P'_T) in $\mathcal{W}_{C,n}$ such that the corresponding labeling function is f. To be more precise, if $\mathcal{D}_T$ is determined, then only those labeling functions f will be allowed that are indicator functions $\mathbb{1}_A$ for a set A with $|A \cap T| = |A \cap U|$, where T is the support of $\mathcal{D}_T$ and U its complement. Therefore all other f are excluded and we will not get statistical independence. Maybe this finally yields a criterion for causality that helps with domain adaptation?
Unfortunately not – at least if we do not change our notion of DA-learnability. We can define a meta-distribution in a way that gives us statistical independence. For this we will define the problem set

$$\begin{aligned}\mathcal{W}' := \{(P_S, P_T) |\ &\text{the marginal } \mathcal{D}_S \text{ is uniform over } \mathcal{X},\ \text{the marginal } \mathcal{D}_T \\ &\text{is uniform over some set } T \subset \mathcal{X},\ \text{covariate shift holds and the corresponding} \\ &\text{labeling function } f \text{ is the indicator function } \mathbb{1}_A, \text{over some set } A \subset \mathcal{X}\}.\end{aligned}$$

Obviously, we have $\mathcal{W}_{C,n} \subset \mathcal{W}'$.
Now, if we do not change our notion of learnability in the presence of a meta-distribution,

it will again suffice to show that for every DA-learner $\mathcal{A}$ there exists a pair (P_S, P_T) in the support of our meta-distribution, i.e. $\mathcal{W}'$, such that $\mathcal{A}$ fails to learn (P_S, P_T). We have already seen that for any learner $\mathcal{A}$ there exists a pair $(P_S, P_T) \in \mathcal{W}_{C,n} \subset \mathcal{W}'$, such that $\mathcal{A}$ fails on (P_S, P_T).
One could argue that the meta-distribution should contain both target and source distributions, and that they should be considered as identically and independently distributed random variables. We can obtain a meta-distribution fulfilling this condition if we define the problem set

$$\mathcal{W}'' = \{(P_S, P_T) | \mathcal{D}_S \text{ is uniform over some set } D \subset \mathcal{X}, \mathcal{D}_T \text{ is uniform over some set } T \subset \mathcal{X}, f \text{ is the indicator function of some set } A \subset \mathcal{X}\}$$

and take the uniform distribution over $\mathcal{W}''$. In this case, we of course lose the weight-ratio assumption for most instances of (P_S, P_T). However, we can still require possible counterexamples to fulfill this condition. And indeed, since $\mathcal{W}' \subset \mathcal{W}''$ we can construct the same counterexamples for this meta-distribution as for the previous meta-distribution.
In conclusion, statistical independence between the distribution of the features and the labeling function does not seem to help for the DA-problem. If anything, the independence required seems to make the problem even more difficult.

Independence of Cause and Mechanism as Algorithmic Independence

A kind of "independence between cause and mechanism" we have not examined yet, is the algorithmic independence criterion, introduced in Chapter 3. Here we can make the argument that for our counterexample from Theorem 11 to work, knowing the set T (which serves as target domain) gives us information for the set A (which was constructed, such that we have either $f = \mathbb{1}_A$ or $f = \mathbb{1}_{\mathcal{X} \setminus A}$ for the labeling function f). Thus knowing $\mathcal{D}_T$ (and its support T) in our problem class $\mathcal{W}_{C,n}$ restricts the choice of A from $\binom{|\mathcal{X}|/2}{|\mathcal{X}|/4} \cdot \binom{|\mathcal{X}|/2}{|\mathcal{X}|/4}$ to only two possible choices, if we also know the corresponding hypothesis class $\mathcal{H}_C$. Without any formal proof, it now seems plausible that the conditional minimum description length for f given $\mathcal{D}_T$ is shorter than the minimum description length for f, without knowing $\mathcal{D}_T$. Therefore f and $\mathcal{D}_T$ would seem causally related.
However, this description length would also depend on the hypothesis class $\mathcal{H}_C$ and would be different for other hypothesis classes. While we would probably often get some reduction of description length in other hypothesis classes it will never be as drastic as in this example. But a good criterion for (objective) causality should probably not depend on our choice of hypothesis class. It is therefore unclear if this argument is an actual sign for causality. Furthermore it is unclear how this could result in a positive criterion making domain adaptation easy.

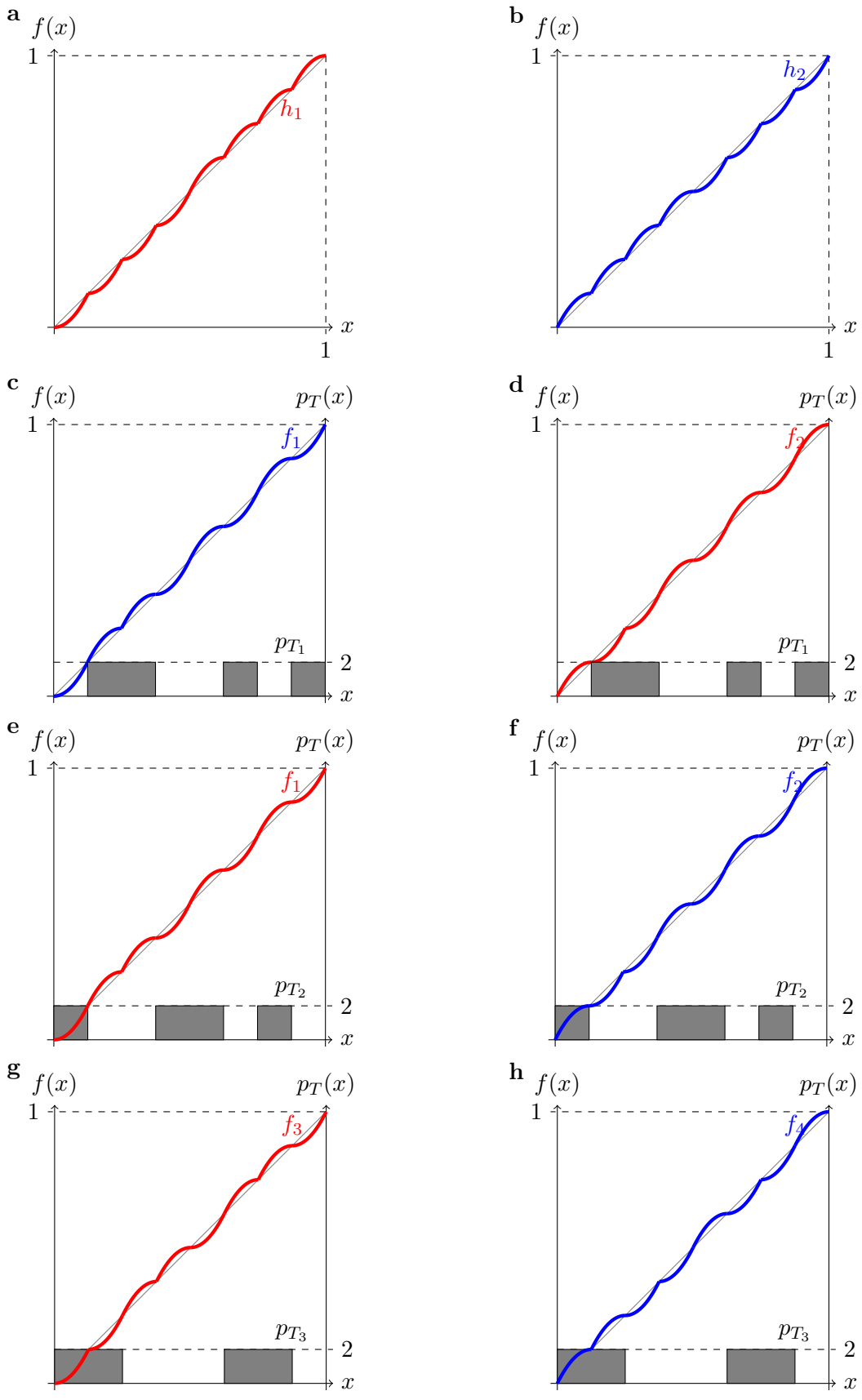


Figure 5.2.: This figure illustrates the problem class $\mathcal{W}_{\text{mon},8}$ used for Theorem 15. **a** and **b** show $h_1, h_2 \in \mathcal{H}_{C,\text{mon}}$ the elements of of the corresponding hypothesis class. **c** and **d** show two opposing labeling functions f_1 and f_2 that co-occur with some target distribution P_{T_1}. **e** and **f** show the same labeling functions f_1 and f_2 but with the opposite P_{T_2} (with $T_2 = [0,1] \setminus T_1$). **g** and **h** give another example of two opposing labeling functions f_3 and f_4 and a different target distribution P_{T_3}. Regression functions drawn as red/blue coincide with h_1/h_2 respectively on the target domain.

6. Discussion

In this thesis we introduced the problem of Domain Adaptation in the context of Learning Theory, as it was used in Ben-David et al. (2010a, 2006, 2010b), Ben-David and Urner (2012). Furthermore, we introduced some possible formalizations for causality. We then tried to formulate a criterion for causality, in the hope that it would be useful for Domain Adaptation learnability. In particular, we tried to find a criterion that would overcome the impossibility results from Ben-David et al. (2010b) or the lower bound from Ben-David and Urner (2012).

We focused primarily on the causal direction as opposed to the anti-causal direction. First we noted that causality criteria like Structural Equation Models that make use of the randomness within one domain allow deterministic labeling functions as were used in Ben-David and Urner (2012) and Ben-David et al. (2010b) to be considered as causal scenarios. Criteria resulting from the randomness within one domain, would therefore not help to eliminate the impossibility results of Ben-David et al. (2010b) or the lower bound of Ben-David and Urner (2012). The only criteria we could obtain for the distribution shift between domains resulting from SCMs were covariate shift and a generalization of covariate shift. Since covariate shift already holds for the results from Ben-David et al. (2010b) and Ben-David and Urner (2012), it obviously does not help to overcome these results.

We therefore looked at causality criteria, in particular the Principle of Independence of Cause and Mechanism, that we hoped might distinguish between the causal and the anti-causal direction for deterministic processes. Thus, we looked at the IGCI-model for regression, as it was used in Janzing et al. (2012), Janzing and Schölkopf (2015). Inspired by this IGCI-model, we gave several attempts for a criterion that indicates a causal relationship between the feature vectors and the labels (i.e. the feature vectors cause the label) in a binary setting. For each criterion we came up with, we could construct lower bounds similar to the one given in Ben-David and Urner (2012) for the sample complexity needed to solve the DA-problem. These lower bounds were dependent on the cardinality of the feature space $|\mathcal{X}|$ and thus lead to counter examples for DA learnability. We concluded that causality criteria inspired by the IGCI model do not seem to help for DA learnability in binary classification. We then looked at a regression model that matched most of the criteria for the IGCI-model[1] and were again able to construct a similar counterexample as the one given in Ben-David and Urner (2012).[2]

[1] in fact it matched all, but "$p_T(x) = 0$ for all $x \in [0, 1]$," which on its own would have lead to DA-learnability combined with the other criteria from Ben-David and Urner (2012).

[2] However, for the ℓ^2-loss the risk was low for all hypothesis in the corresponding hypothesis class.

We then discussed briefly whether a formulation of the Principle of Cause and Mechanism in terms of statistical independence for a meta-distribution of distributions of causes and mechanisms (i.e. labelings) would lead to DA-learnability. We came to the conclusion that even if we knew this meta-distribution – which in practice we likely would not, making a possible criterion like this not very useful – a statistical criterion would not exclude scenarios as in the counterexample from Ben-David and Urner (2012). Rather, a meta-distribution could tell us that these scenarios are unlikely (i.e., the meta-distribution assigns them low probability). For finite sample bounds this independence seemed to make our DA-problem even harder, since the problem classes can get larger.

Lastly, we briefly discussed whether a formulation of the Principle of Cause and Mechanism in terms of algorithmic independence would make the DA-problem easy. While in our particular DA problem $\mathcal{W}$ (which was given in Ben-David and Urner (2012)), knowing $\mathcal{H}$ and knowing the target distribution p_T significantly reduced the number of possible labeling functions f and therefore arguably its description length, we were not able to formulate any positive causality criterion that would make the DA-problem easy. Furthermore, as argued for the statistical independence criterion, we could embed $\mathcal{W}$ into a larger problem class $\mathcal{W}'$, where knowing $\mathcal{H}$ and knowing p_T would not reduce the number of possibilities for f. We are therefore not convinced that algorithmic information criteria would help to solve the DA-problem from Ben-David and Urner (2012).

In summary, we explored several formalizations for the causal direction as an independence between the marginal distribution of the cause and the labeling and showed that the DA problem could still be hard under these causal assumptions. We also do not see, how any *independence statement* would help to achieve finite sample bounds, since this absence of information seems to make the problem harder. In the best case this independence would imply that only low probabilities [3] are assigned to possible counterexamples. For these scenarios it would also be necessary to make further assumptions about the generating process of cause and mechanism, which might be difficult in practice. We therefore believe that, in order to provide meaningful results for domain adaptation learnability in binary classification in a causal scenario one would need to change the notion of learnability to something that is not distribution free or consider the anti-causal direction. While the anti-causal direction does not give us any justification for the covariate shift assumption, a *dependence* between the distribution of the cause and the mechanism, might be helpful to make use of the unlabeled target data. We also briefly discussed the use of causal assumptions for domain adaptation in regression. While we could construct similar lower bounds for our examples, we do not believe our discussion of the regression case to exhaustive. Since causality seems to be formalized more clearly in the continuous case, we therefore believe that exploring the use of causality in domain adaptation for regression scenarios could still be interesting. This concludes the summary of our results.

We will now give a discussion about the use of causal assumptions in learning theory in general - in particular their use for finite sample bounds in binary classification. We

[3] in terms of the generation process for cause and mechanism

will then discuss possible directions for future work. While many more complex causal models are possible, will only discuss two the two most simple causal scenarios here – the *causal* and the *anti-causal* case.

In the *causal* case, it is likely that if in practice we have prior knowledge of the data-generating process, it is knowledge of the mechanism determining the effect from the cause. This could be viewed as information about a restricted hypothesis class $\mathcal{H}$ that is realizable, or has small approximation error[4]. On the other hand, a causal mechanism, would likely give us no prior information about the distribution of the feature vectors. These intuitions are captured well in common assumptions in learning theory – i.e., we often assume a restricted hypothesis class $\mathcal{H}$ in order to get distribution free bounds on the estimation error[5], which arguably only yields meaningful results, if the approximation error of $\mathcal{H}$ is small.

For the *anti-causal* direction, it also seems plausible, that the prior knowledge we have is knowledge of the mechanism. However, in the anti-causal direction in classification, the mechanism would be a generative model, where the labels generate the feature vectors. The knowledge of a class of gernerative models could be thought of as knowledge of a class of conditional distributions $\mathcal{P}$ of feature vectors, given the labels. One could then use this class to find a restricted hypothesis class $\mathcal{H}$ that contains all possible Bayes classifiers resulting from combining conditionals in $\mathcal{P}$ and marginal distributions over the labels. But our knowledge of the data would not only be restricted to this hypothesis class $\mathcal{H}$, but also contain information about the possible marginal distributions of feature vectors x, that could result from a class of conditionals $\mathcal{P}$. We believe, it would be possible to use that knowledge to obtain better finite sample bounds. These bounds would not be distribution free, in the sense that they make *no assumption* about the marginal distribution of the feature vectors. However, it might be possible to find bounds that are independent of the marginal distribution of feature vectors x, given that the distribution arises from a generative process as described by $\mathcal{P}$. An example for upper and lower finite sample bounds can be seen for isotropic Gaussian conditionals in Li et al. (2017).

For these generative processes it seems furthermore likely that one could provide semi-supervised learning guarantees, as these generative processes might lead to clusterability, which in turn would imply the usefulness of unlabeled data in semi-supervised learning and domain adaptation scenarios. This, however, would not be the case for all possible generative models. For example if the conditional distributions for the two labels are uniform in $[0, 1]$ and in $[1, 2]$ respectively, then no clusterability arises.

However, for sufficiently smooth distributions (e.g., distributions that fulfill the Lipschitz-condition), we might get clusterablity due to the Lipschitzness that would be implied in the labeling function (compare to Urner et al. (2012)). It is possible that under some

[4]the approximation error of an hypothesis class is defined by $\inf h \in \mathcal{H}\mathcal{L}_P(h)$

[5]The estimation error of a hypothesis the difference between its risk and the best risk achievable in the hypothesis, i.e., $\mathcal{L}_P(h) - \inf_{h' \in \mathcal{H}} \mathcal{L}_P(h')$

conditions this might lead to *probabilistic Lipschitzness*, a property which was shown to be useful for semi-supervised learning and domain adaptation scenarios in Urner (2013).[6]

One example one might explore are Gaussian conditionals. These could be seen as as Gaussian additive noise models, as introduced in Chapter 3 for a binary classification setting. In Li et al. (2017), upper and lower bounds for classification and clustering in a generative model with isotropic Gaussians are introduced. Combined with the results of Castelli and Cover (1995) and Castelli and Cover (1996), these bounds could provide a comparison between supervised and semi-supervised learning techniques and show that unlabeled data indeed helps in this generative example.[7]

To conclude, we did not find any positive result for DA learnability that resulted from the Principle of Independence of Cause and Mechanism. For future work, we suggest to consider the anti-causal direction instead and to examine whether generative models help to make use of unlabeled data for semi-supervised as well as for domain adaptation scenarios.

[6] Note that Urner (2013) primarily focuses on probabilistic Lipschitzness in deterministic scenarios. However, generative models will likely give rise to non-deterministic labeling functions. The use of probabilistic Lipschitzness in these scenarios remains – to the best of our knowledge – unexplored.

[7] Note that for a fair comparison, it is indeed necessary to look at lower bounds for supervised learning methods that assume a generative model.

A. VC-dimension of $\mathcal{H}\Delta\mathcal{H}$

In the bound from Theorem 2, there still remains an unknown quantity: the VC-dimension of $\mathcal{H}\Delta\mathcal{H}$. In Ben-David et al. (2010a) it is stated that

$$VC(\mathcal{H}\Delta\mathcal{H}) \leq 2 \cdot VC(\mathcal{H}) \tag{A.1}$$

However, the sketch of proof of the paper does not actually lead to this result, but gives a slightly worse bound of

$$VC(\mathcal{H}\Delta\mathcal{H}) \leq 4VC(\mathcal{H}) \log(VC(\mathcal{H})). \tag{A.2}$$

In another paper of the same author the statement (A.1.) is claimed Ben-David and Urner (2012) again, but even though the given sketch of proof is different, it only serves to give the bound including the additional log-factor. In Mansour et al. (2009) (A.1) is claimed without a proof or reference as well. It is still unclear if the better bound holds. [1] For this master thesis, we will therefore only use the weaker result and provide a proof, even though it is very likely that the stronger version still holds.

Proposition 3. *Let $VC(\mathcal{H}) = d$. Then we get the following bound for the VC-dimension of $\mathcal{H}\Delta\mathcal{H}$:*

$$VC(\mathcal{H}\Delta\mathcal{H}) \leq 4d \log_2(4dl)$$

Proof. Let $C \subset \mathcal{X}$ with $|C| = n$. By Sauer's Lemma, we get the following bound on the number of hypotheses in $\mathcal{H}_C$:

$$|\mathcal{H}_C| \leq \sum_{i=0}^{d} \binom{n}{i} \leq (n+1)^d$$

Note that $h\Delta h' = h'\Delta h$ for all $h, h' \in \mathcal{H}$. Thus we only have to consider half of all possible combinations between different h, h', when counting the hypotheses in $\mathcal{H}_C\Delta\mathcal{H}_C$.

[1] If $\mathcal{H}$ has a compression scheme of size $VC(\mathcal{H})$, we have a compression scheme of size $2VC(\mathcal{H})$ for $\mathcal{H}\Delta\mathcal{H}$. Therefore the VC-dimension of $\mathcal{H}\Delta\mathcal{H}$ is actually bounded by $2VC(\mathcal{H})$ in this case. There are many classes $\mathcal{H}$ that have been shown to have a compression scheme of size $VC(\mathcal{H})$, but it still remains an open question, if there is always a compression scheme of that size. Therefore it would also be interesting to find an example, where $VC(\mathcal{H}\Delta\mathcal{H}) > 2VC(\mathcal{H})$, because it would also serve to disprove the claim about the existence of a compression scheme of size $VC(\mathcal{H})$.

Furthermore $h\Delta h = h'\Delta h'$ for all $h, h' \in \mathcal{H}$. So this function needs to be counted only once. Therefore:

$$|(\mathcal{H}\Delta\mathcal{H})_C| = |\mathcal{H}_C\Delta\mathcal{H}_C| \leq \frac{|\mathcal{H}_C|(|\mathcal{H}_C|-1)}{2} + 1 \leq \frac{(n+1)^{2d}}{2}$$

We would now like to find an n such that:

$$\begin{aligned} & \frac{(n+1)^{2d}}{2} < 2^n \\ \Leftrightarrow \quad & 2d < \frac{(n+1)}{\log_2(n+1)} \end{aligned}$$

Setting $n := 4d\log_2(4d)$, we see:

$$\begin{aligned} \frac{4d\log_2(4d)+1}{\log_2(4d\log_2(4d)+1)} &> 2d\frac{2\log_2(4d)+1}{\log_2((4d+1)\log_2(4d))} \\ &> 2d\frac{2\log_2(4d)+1}{\log_2(4d+1)+\log_2(\log_2(4d))} \\ &> 2d\frac{2\log_2(4d)+1}{\log_2(4d)+1+\log_2(\log_2(4d))} \\ &> 2d \end{aligned}$$

Thus $VC(\mathcal{H}\Delta\mathcal{H}) \leq 4d\log_2(4d)$. □

Bibliography

Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *Algorithmic Learning Theory, ALT 2012*, pages 139–153, 2012.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Neural Information Processing Systems,NIPS 2006*, pages 137–144, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010a.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2010*, pages 129–136, 2010b.

Shai Ben-David, Shai Shalev-Shwartz, and Ruth Urner. Domain adaptation–can quantity compensate for quality? In *International Symposium on Artificial Intelligence and Mathematics, ISAIM 2012*, 2012.

Vittorio Castelli and Thomas M. Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 16(1):105–111, 1995.

Vittorio Castelli and Thomas M. Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Trans. Information Theory*, 42(6):2102–2117, 1996.

Dominik Janzing and Bernhard Schölkopf. Semi-supervised interpolation in an anticausal learning scenario. *Journal of Machine Learning Research*, 16:1923–1948, 2015.

Dominik Janzing, Joris M. Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniusis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artif. Intell.*, 182-183:1–31, 2012.

Benjamin G. Kelly, Thitidej Tularak, Aaron B. Wagner, and Pramod Viswanath. Universal hypothesis testing in the learning-limited regime. In *IEEE International Symposium on Information Theory, ISIT 2010*, pages 1478–1482, 2010.

Tianyang Li, Xinyang Yi, Constantine Caramanis, and Pradeep Ravikumar. Minimax gaussian classification & clustering. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, pages 1–9, 2017.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Neural Information Processing Systems, NIPS 2009*, pages 1041–1048, 2009.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. The MIT Press, Cambridge, MA, USA, 2017.

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 2018.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. *Semi-supervised learning in causal and anticausal settings*, chapter 13, pages 129–141. Festschrift in Honor of Vladimir Vapnik. Springer, 2013.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Ruth Urner. *Learning with non-Standard Supervision*. PhD thesis, University of Waterloo, Ontario, Canada, 2013.

Ruth Urner, Shai Ben-David, and Ohad Shamir. Learning from weak teachers. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2012*, pages 1252–1260, 2012.