Impossibility Theorems for Domain Adaptation

Shai Ben-David and **Teresa Luu**
School of Computer Science
University of Waterloo
Waterloo, ON, CAN
{shai,t2luu}@cs.uwaterloo.ca

Tyler Lu
Dept. of Computer Science
University of Toronto
Toronto, ON, CAN
tl@cs.toronto.edu

Dávid Pál
Department of Computing Science
University of Alberta
Edmonton, AB, CAN
dpal@cs.ualberta.ca

Abstract

The domain adaptation problem in machine learning occurs when the test data generating distribution differs from the one that generates the training data. It is clear that the success of learning under such circumstances depends on similarities between the two data distributions. We study assumptions about the relationship between the two distributions that one needed for domain adaptation learning to succeed. We analyze the assumptions in an agnostic PAC-style learning model for a the setting in which the learner can access a labeled training data sample and an unlabeled sample generated by the test data distribution. We focus on three assumptions: (i) similarity between the unlabeled distributions, (ii) existence of a classifier in the hypothesis class with low error on both training and testing distributions, and (iii) the covariate shift assumption. I.e., the assumption that the conditioned label distribution (for each data point) is the same for both the training and test distributions. We show that without either assumption (i) or (ii), the combination of the remaining assumptions is not sufficient to guarantee successful learning. Our negative results hold with respect to *any* domain adaptation learning algorithm, as long as it does not have access to target labeled examples. In particular, we provide formal proofs that the popular covariate shift assumption is rather weak and does not relieve the necessity of the other assumptions.

We also discuss the intuitively appealing paradigm of re-weighting the labeled training sample according to the target unlabeled distribution and show that, somewhat counter intuitively, we show that paradigm cannot be trusted in the following sense. There are DA tasks that are indistinguishable as far as the training data goes but in which re-weighting leads to significant improvement in one task while causing dramatic deterioration of the learning success in the other.

1 Introduction

Much of the theoretical analysis of machine learning has focused on the case when the training and test data are generated by the *same* underlying distribution. While this may sometimes be a good approximation of reality, in many practical tasks this assumption cannot be justified. For example, when learning an automated part-of-speech tagger (see (Ben-David et al., 2006)), one's training data is limited to particular genres of documents (due to the cost of labeling), but of course the true goal is to have a good tagger for the types of documents occurring in future data sets. Here, one cannot expect the training data to be perfectly representative of future test data. The same issue occurs in building spam detectors when the training data, emails received from some addresses, is generated by a different distribution from the one generating the emails for the target user.

Nevertheless, this is not an "all-or-nothing" situation. While the training and test distributions may not be completely identical, they are often quite similar. Furthermore, in some such tasks, unlabeled examples, generated by the distribution governing the target domain, may be also available to the learner.

The hope of domain adaptation (DA) is to use training data from one source to help construct a predictor for data generated by a different but related target source. Clearly, the range of application to which such issues

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2010 by the authors.

apply is huge.

The development of algorithmic heuristics for specific DA applications brings about a growing need for a theory that could analyze, guide and support such tasks. Some high-level questions of interest are:

- What conditions, mainly on the relationship between the source training and target test distributions, allow DA to succeed? Which assumptions suffice to provide performance guarantees on the success of DA algorithms?
- Which algorithmic paradigms are likely to perform well under such given relatedness assumptions?

Clearly, DA works in practice, both in natural learning of animals and humans and in machine learning applications. On the other hand, it is obvious that some kind of relationship (or similarity) between the training and test domains is at the root of such successes. In our view, the big challenge that DA research faces is coming up with "appropriate" assumptions under which DA can be guaranteed to succeed. Such assumptions should balance the following competing requirements:

1. DA assumptions should be user friendly, in the sense that it would be reasonable to expect a domain expert to have an understanding (or some reliable intuition) of whether, and to what extent, the assumptions hold for a concrete learning task at hand.
2. The assumptions should be amenable to precise mathematical formalization.
3. They should suffice to allow the derivation of performance guarantees for DA algorithms.

Our impression is that very few of the currently applied DA assumptions meet these necessary requirements. In this paper we discuss some formal candidate DA assumptions from the perspective of the last requirement—to what extent they suffice to allow the existence of a DA algorithm with solid success guarantees.

This work focuses on the learning model in which there are two data generating distributions: a source distribution and a target distribution. Both generate labeled examples. The DA learner has access to an i.i.d. labeled sample from the source distribution, and to an i.i.d. unlabeled sample from the target distribution. The learner is expected to output a predictor, whose success is evaluated with respect to the target distribution.

An interesting question in that learning scenario is whether a learner should just find the best hypothesis with respect to the labeled training source-domain sample and use the target-domain, unlabeled, sample just to evaluate the performance of that hypothesis on the target task (we call such paradigms "conservative DA"), or there maybe a way to utilize the information contained in the target-domain sample in the process of choosing the learner's classifier ("non-conservative DA"). While there have been several suggestions for non-conservative DA learning, there are no formal guarantees of the advantage of such methods. We address one such approach, re-weighting the labeled sample to match the marginal distribution of the target-domain, and show that in some cases it may lead to a major *increase* of the prediction error.

1.1 Related work

(Ben-David et al., 2006) define a formal model of DA and provide an upper bound on the error of the simplest algorithm—the empirical risk minimization (ERM) over the training data. The bound depends on the distance between distributions, as measured by the so-called $d_{\mathcal{A}}$ distance as introduced in (Kifer et al., 2004). The distinguishing feature of this distance is that it is estimable from an unlabeled sample alone, the size of which is distribution-free (it is determined by the VC dimension of $\mathcal{A}$).

A follow-up paper by (Mansour et al., 2009b) extends $d_{\mathcal{A}}$ distance to real-valued function classes and provides Rademacher-based bounds for more general loss functions. But the bounds are incomparable with those in (Ben-David et al., 2006). In addition, they propose re-weighting the examples of the source training data so that the re-weighted (unlabeled) empirical training distribution is closer to the (unlabeled) empirical target distribution. The idea bears some resemblance to importance sampling in Monte Carlo methods. See also (Sugiyama and Mueller, 2005; Cortes et al., 2008; Huang et al., 2007). Our work partly addresses the basic question regarding this method: does re-weighting training examples always do better than not re-weighting?

There are some other related work stemming from (Ben-David et al., 2006). In (Blitzer et al., 2007), the authors prove adaptation bounds when some target labeled data is available. Inspiring an algorithm that optimally re-weights the training and target data errors. (Crammer et al., 2008) consider the setting where there are multiple sources of training data, but the source unlabeled distributions must be the same. This direction is further explored in (Mansour et al., 2009a) where the target distribution is assumed to be a mixture of source distributions.

The covariate shift assumption, stating that the conditional distribution of the target and source data are the same, is a central element of most works on domain adaptation (e.g. (Huang et al., 2007; Sugiyama and Mueller, 2005)). However, the proposed methods such as instance weighting require very large sample sizes to reduce variance, which is very unrealistic. As discussed in (Quionero-Candela et al., 2009, Author Comments), under more realistic scenarios where the domain is much larger than the training and test sample sizes, it is unlikely that the same domain point will occur in both samples. That makes any label behavior over the training and test sample to be consistent with such a "bias-free" covariate shift assumption. Thus covariate shift cannot guarantee us the success of DA unless the points in the domain are visited several times. We give a concrete demonstration of this in Section 4.

1.2 Our Results

Note that most, if not all, of the theoretical guarantees on the performance of DA algorithms (e.g. (Ben-David et al., 2006; Blitzer et al., 2007), and Theorem 8 in (Mansour et al., 2009b)) concern the performance of "conservative" algorithms, that ignore available target-generated unlabeled data (e.g., ERM over the training data). In these papers, target-generated unlabeled data is used only to help analyze the success of the learner. The error bounds that are currently known contain the following two components: a component reflecting the discrepancy between the unlabeled distributions of the training and test data, as measured by $d_{\mathcal{A}}$, and a component that reflects the label discrepancy between these two distributions. Two questions naturally arise: first, how tight are these bounds? Are these components inevitable? A second, related question concerns the DA algorithm; while these components may be unavoidable for the naive ERM algorithm over the training data (or for conservative algorithms in general), can they be overcome by some smarter algorithms that utilize not just the labeled training sample but also the unlabeled sample of the target data distribution? (For example, the proposed re-weighting method of (Mansour et al., 2009b)).

We provide full answers to these questions. We prove that unless the learner has access to *labeled* examples from the target distribution, neither the pair of assumptions "covariate shift + small $d_{\mathcal{A}}$ between the unlabeled distributions" nor the pair "covariate shift + existence of a predictor (from a small VC class, known to the learner) that has low error on both the training and target domains" suffice to guarantee successful DA. These results hold with respect to ***any*** algorithm. It follows that the terms in the upper bounds on the error of the naive algorithm, that just minimizes the training error over a small VC class (derived in (Ben-David et al., 2006) and (Mansour et al., 2009b)) are both necessary as long as one does not make any further assumptions. Moreover, covariate shift does not help reduce the worst-case error of DA algorithms, regardless of the algorithm.

We also analyze the re-weighting paradigm, showing that a DA framework where all the target information is coming from an unlabeled target-domain sample, reweighting may be highly counter productive (even in cases where 'conservative' learning, that ignores the target sample, does well).

Our paper is structured as follows. Our definitions and learning setup are given in Section 2, while Section 3 briefly summarizes the common DA assumptions and describes a recently proposed, potentially promising re-weighting method. In Section 4 we demonstrate examples that give insight on what assumptions can cause popular DA approaches to succeed or fail, and finally we present our impossibility results in Section 5.

2 Preliminary Definitions and Notation

In this section we present the formal setup for DA. For simplicity, we consider binary classification tasks only. Let X be some domain set. We represent the learning tasks as probability distributions Q, P over $X \times \{0,1\}$. We call Q the *source distribution* and we call P the *target distribution* or the *test distribution*. We denote by Q_X and P_X the marginals of Q and P on X (informally, *unlabeled distributions*). For a distribution D over X, and a function $f : X \to [0,1]$ we define D_f over $X \times \{0,1\}$ by $D_f(1 \mid x \in X) = f(x)$.

In our DA scenario a learning algorithm receives an i.i.d. sample $L = ((x_1, y_1), (x_2, y_2), (x_m, y_m))$ from Q and an i.i.d. sample $U = (x'_1, x'_2, \ldots, x'_n)$ from P_X. The algorithm then outputs some $h : X \to \{0,1\}$. This is formalized below.

Definition 1 (Learner)**.** A *domain adaptation (DA) learner* is a function

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

The performance of a learner is measured by the error of the learned classifier with respect to the target distribution. If R is any distribution over $X \times \{0,1\}$, we let $\mathrm{Err}^R(h) = \Pr_{(x,y)\sim R}[h(x) \neq y]$ which we call the *R-error of h*. As usual in statistical learning theory, we measure the performance of a learner relative to the error of the best classifier in a hypothesis class. A *hypothesis class H* is a class of functions from X to $\{0,1\}$. We introduce the shorthand notation for the error of the best classifier in the hypothesis class. For a distribution R over $X \times \{0,1\}$, the error of the best classifier in the class is defined as $\mathrm{Err}^R(H) = \inf_{h \in H} \mathrm{Err}^R(h)$. Sometimes, we refer to $\mathrm{Err}^R(H)$ as the *approximation error* of H. We now define learnability in our model.

Definition 2 (Learnability)**.** Let P, Q be distributions over $X \times \{0,1\}$, H a hypothesis class, A a DA learner, $\epsilon, \delta > 0$, and m, n positive integers. We say that A *(ϵ, δ, m, n)-learns P from Q relative to H*, if when given access to a labeled sample L of size m, generated i.i.d. by Q, and an unlabeled sample U of size n, generated i.i.d by P_X, with probability at least $1-\delta$ (over the choice of the samples L and U), the learned classifier does not exceed the P-error of the best classifier in H by more than ϵ. In other words,

$$\Pr_{\substack{L \sim Q^m \\ U \sim (P_X)^n}} \left[\mathrm{Err}^P(A(L,U)) \le \mathrm{Err}^P(H) + \epsilon\right] \ge 1 - \delta \ .$$

Clearly, if the tasks P and Q are unrelated, or if H is not a suitable hypothesis class, then DA as formalized in the above definition, might not be possible. Thus, the question is what conditions on (P, Q, H) allow DA.

Definition 3 (Sufficient Condition for DA)**.** Let H be a hypothesis class and P and Q probability distributions over labeled data as before.

(a) We say that a *relation* (or a *condition*) R over triples (P, Q, H) *suffices for domain adaptation* if there exists a DA learner A such that for any $\epsilon, \delta > 0$ there exist integers $m := m(\epsilon, \delta)$ and $n := n(\epsilon, \delta)$ such that for every P, Q satisfying $R(P, Q, H)$, we have that A (ϵ, δ, m, n)-learns P from Q relative to H. (Can be extended to multiple relations.)

(b) We say that a real valued *parameter $p(P, Q, H)$ suffices for domain adaptation*, if there exists a DA learner A such that for any $\epsilon, \delta > 0$ there exist some value $p > 0$ and integers $m := m(\epsilon, \delta)$ and $n := n(\epsilon, \delta)$ such that for all P, Q satisfying $p(P, Q, H) < p$, we have that A (ϵ, δ, m, n)-learns P from Q relative to H. (Can be extended to many parameters.)

3 Common Domain Adaptation Assumptions and Methods

Several assumptions and parameters have been considered sufficient for DA. Below, we list some of the most common and/or successful ones. The focus of this work is providing some necessary conditions, without which no DA algorithm can be expected to enjoy guaranteed success. Surprisingly, it turns out that some of the most common assumptions do not suffice for any guarantees. In particular, we discuss two quantities that have been shown to be important in providing sufficient conditions on good domain adaptation. They are a measure of distance between distributions, called the $\mathcal{A}$-distance as introduced in (Kifer et al., 2004), and a measure of how well the distributions agree in its labels as seen in (Ben-David et al., 2006).

Covariate Shift. The first assumption that is often invoked to justify DA is that of "covariate shift" (see e.g. (Sugiyama and Mueller, 2005)). We say that P and Q satisfy the covariate shift assumption if the conditional label distribution does not change between the training and target distributions. That is, for all $x \in X$ and any $y \in \{0,1\}$, $P(y \mid x) = Q(y \mid x)$. This is a central element of much research on DA (e.g. see the book by (Quionero-Candela et al., 2009)). We show in Section 4 that it is an insufficient condition to guarantee DA success.

Similarity of the Unlabeled (Marginal) Distributions. Starting with (Ben-David et al., 2006), it becomes common to measure the distance between the marginal distributions of the source and test distributions by the so-called $\mathcal{A}$-distance.

Definition 4. Let Q_X and P_X be distributions over X. Let $\mathcal{A}$ be a collection of subsets of X such that each set is measurable with respect to Q_X and P_X. The *$\mathcal{A}$-distance* between Q_X and P_X is

$$d_{\mathcal{A}}(Q_X, P_X) = 2 \sup_{A \in \mathcal{A}} |Q_X(A) - P_X(A)|.$$

The particular choice of $\mathcal{A}$ that is used is $\mathcal{A} = H \Delta H$ where $H \Delta H$ is the set of all "symmetric differences" between elements of H. That is, $H \Delta H = \{\{x \in X \ : \ h(x) \neq h'(x)\} \ : \ h, h' \in H\}$.

Low-error Joint Prediction. This is a measure of the agreement between the labels of the two distributions from the perspective of the class H. (Ben-David et al., 2006) were first to define $\lambda_H(P, Q)$.

Definition 5. For two distributions P, Q over $X \times \{0,1\}$ and a hypothesis class H we define

$$\lambda_H(P, Q) = \inf_{h \in H} \left[\mathrm{Err}^Q(h) + \mathrm{Err}^P(h)\right] \ .$$

As an alternative (Mansour et al., 2009b) consider the optimal classifiers in H for the source and target distributions $h^*_Q = \mathrm{argmin}_{h \in H} \mathrm{Err}^P(h)$ and $h^*_P = \mathrm{argmin}_{h \in H} \mathrm{Err}^Q(h)$, and they measure the disagreement between the distributions of labels by $Q(h^*_P, h^*_Q) = Q_X(\{x \ : \ h^*_P(x) \neq h^*_Q(x)\})$.

Approximation Error. Finally, the approximation error of the class H with respect to the data-generating distributions plays a major role in determining how well one can predict labels using that class. Clearly, $\mathrm{Err}^P(H)$ (where P is the target distribution) should be small if one expects to achieve good prediction with members of H. Since we only have labels from the source distribution, Q, and must produce a label predictor based on that, it may probably be impossible to succeed if $\mathrm{Err}^Q(H)$ is not small.

3.1 The Re-weighting Method

(Mansour et al., 2009b) have recently proposed potentially promising method for utilizing the target-distribution unlabeled sample in the search for a label predictor. Consider the sample $L = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ generated by the source distribution, Q, and an unlabeled sample $U = (x'_1, \ldots, x'_n)$ generated by the target distribution, P, that are the input to a DA learner. Let $L_X = (x_1, x_2, \ldots, x_m)$ be the sample points in L (without their labels). A vector of non-negative weights, $\bar{w} = (w_1, \ldots, w_m)$, that sum up to 1, can be viewed as a probability distribution over L_X. We denote the new distribution $P^L_{\bar{w}}$ and call it a *re-weighted sample*. The idea of the re-weighting paradigm is to find such a distribution such that $\mathcal{A}$-distance between $(P^L_{\bar{w}})_X$ and U is as small as possible, where $\mathcal{A} = H\Delta H$.

(Mansour et al., 2009b) propose to use a linear programming method for computing such a sample re-weighting. The hope is that the new distribution over L will convey useful information about P. The re-weighted sample, denoted P^S, can then be fed into any standard supervised learning algorithm. As a canonical example, we can apply ERM to P^S and output the hypothesis $h^*_{P_S} = \mathrm{argmin}_{h \in H} \mathrm{Err}^{P_S}(h)$ with lowest error in H.

4 Demonstrating Pitfalls of Domain Adaptation

In this section we consider several concrete examples of P, Q, H. Our focus is on examining which of the assumptions from previous section suffices for DA learnability. Towards that, we shall compute, for each of the examples, the parameters $\lambda_H(P, Q)$, $d_{\mathcal{A}}(P, Q)$, $\mathrm{Err}^P(H)$, $\mathrm{Err}^Q(H)$, and the extent to which the covariate shift assumption holds. We also have a look at whether the re-weighting method achieves a low P-error or not.

A common strengthening of the covariate shift assumption, is that, on top of having a common labeling rule to the source and target distributions, it is also the case that the supports of the two distributions coincide (or at least that any measurable set that has non-zero weight under the target distribution also has non-zero weight under the source distribution). For the sake of clarity of presentation, we do not insist on this assumption in the examples we design below. However, this omission can be easily overcome by assigning some very small weight to any point that is assigned zero weight in the examples. Such a modifications will not change the phenomena demonstrated by these constructions.

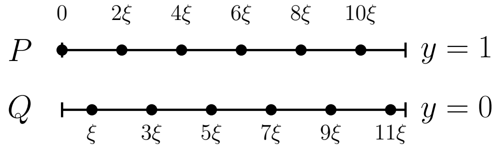

Figure 1: Picture shows the source and target distributions P, Q from Example 6 with parameter $\xi = 2/23$. P uniform over $\{(0\xi, 1), (2\xi, 1), (4\xi, 1), \ldots, (10\xi, 1)\}$ and the Q is uniform over $\{(\xi, 0), (3\xi, 0), (5\xi, 0), \ldots, (11\xi, 0)\}$.

For the sake of simplicity, we design all our examples over the unit interval $[0, 1]$ and use, throughout, the hypothesis class H of threshold functions. That is, for any $t \in [0, 1]$, we define a *threshold* function h_t by $h_t(x) = 1$ for $x < t$, and $h_t(x) = 0$ otherwise. The class of thresholds is $H = \{h_t \ : \ t \in [0, 1]\}$. Note that $H\Delta H$ becomes the class of half-open intervals, i.e. $\mathcal{A} = \{(a, b] \ : \ 0 \le a \le b \le 1\}$.

All the examples, and therefore the corollaries we derive from the analysis, can be extended to the Euclidean domain $\mathbb{R}^d$. In the examples, we will use h^*_Q to denote the classifier in H with the lowest Q-error (where Q is the source-distribution). Similarly, we denote by $h^*_{P_S}$ the classifier in H with the lowest P_S-error where P_S is the re-weighted sample. Our first example shows that the covariate shift assumption is not sufficient for good DA.

Example 6 (Inadequacy of Covariate Shift)**.** Fix some small $\xi \in (0, 1)$. Let the target distribution P be the uniform distribution over $\{2k\xi \ : \ k \in \mathbb{N},\ 2k\xi \le 1\} \times \{1\}$ and let the source distribution Q be the uniform distribution over $\{(2k+1)\xi \ : \ k \in \mathbb{N},\ (2k+1)\xi \le 1\} \times \{0\}$. See Figure 1.

Notice the following:

1. Covariate shift assumption holds for P and Q.
2. The distance $d_{\mathcal{A}}(P, Q) = \xi$ and hence it is arbitrarily small.
3. Both approximation errors $\mathrm{Err}^P(H), \mathrm{Err}^Q(H)$ are zero.
4. $\lambda_H(P, Q) = 1 - \xi$ and $\mathrm{Err}^P(h^*_Q) \ge 1 - \xi$ are large.

The above example demonstrates that the covariate assumption can be satisfied without affecting the suc-

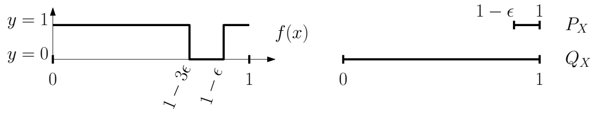

Figure 2: The figure shows the function f and the marginals P_X, Q_X from Example 7. The distributions P, Q satisfies covariate shift and f is their common conditional distribution. That is, for all $x \in X$, $Q(y = 1 \mid x) = P(y = 1 \mid x) = f(x)$. The marginal Q_X is the uniform distribution over $[0, 1]$ and the marginal P_X is the uniform distribution over $[1 - \epsilon, 1]$.

cess of learning—it does not help to guarantee DA. We will make a stronger statement along this line in Section 5 below.

Example 7 (Success of Re-weighting)**.** Fix some small $\epsilon \in (0, 1/4)$. We make sure that covariate assumption holds. That is, for any $x \in X$, $P(y = 1 \mid x) = Q(y = 1 \mid x) = f(x)$. We define $f : X \to [0, 1]$ as follows. For $x \in [1 - 3\epsilon, 1 - \epsilon]$ we set $f(x) = 0$ and otherwise we set $f(x) = 1$. To fully specify Q and P, it remains to specify their marginals Q_X, P_X. We let Q_X be the uniform distribution over $[0, 1]$ and we let P_X to be the uniform distribution over $[1 - \epsilon, 1]$. See Figure 2.

Note that,

$$\lambda_H(P, Q) = 2\epsilon \quad d_{\mathcal{A}}(P_X, Q_X) = 1 - \epsilon \quad \mathrm{Err}^P(h^*_Q) = 1$$
$$\mathrm{Err}^P(H) = 0 \qquad \mathrm{Err}^Q(H) = \epsilon \ .$$

Furthermore, it is not hard to see that $\mathrm{Err}^P(h^*_{P^S}) \to 0$ in probability as $n, m \to \infty$, which can be considered a tremendous success of the re-weighting method. On the other hand, as a sample from Q gets large, ERM on it gets close to h^*_Q which has P-error 1.

(Ben-David et al., 2006) prove that for any $h \in H$, $|\,\mathrm{Err}^P(h) - \mathrm{Err}^Q(h)| \leq \lambda_H(P, Q) + d_{\mathcal{A}}(P_X, Q_X)$. This example shows that the upper bound on the gain of re-weighting that we mentioned above, $\lambda_H(P, Q) + d_{H\Delta H}(P, Q)$, is really tight and the ratio between this gain and $\lambda_H(P, Q)$ can be arbitrarily large.

Example 8 (Total Failure of Re-weighting)**.** This example is the same as the previous, except that the labels of P are flipped. That is, Q_X is uniform over $[0, 1]$ and the conditional distribution $Q(y = 0 \mid x) = 1$ for any $x \in [1 - 3\epsilon, 1 - \epsilon]$ and for any $x \in X \setminus [1 - 3\epsilon, 1 - \epsilon]$ the conditional distribution is $Q(y = 1 \mid x) = 1$. The marginal distribution P_X is the uniform distribution over $[1 - \epsilon, 1]$. However, in contrast with the previous example $P(y = 0 \mid x) = 1$ for all $x \in X$.

Note that,

$$\lambda_H(P, Q) = \epsilon \quad d_{\mathcal{A}}(P_X, Q_X) = 1 - \epsilon \quad \mathrm{Err}^P(h^*_Q) = 0$$
$$\mathrm{Err}^P(H) = 0 \qquad \mathrm{Err}^Q(H) = \epsilon \ .$$

It is not hard to see that $\mathrm{Err}^P(h^*_{P^S}) \to 1$ in probability as $n, m \to \infty$, which is an embarrassing failure of the re-weighting method. However, ERM on a sample of Q will do very well.

When we compare Examples 7 and 8, we see that all the relevant parameters that we were considering, $\lambda_H(P, Q)$ and $d_{\mathcal{A}}(P, Q)$, have not changed by much. However, there is a dramatic shift in the performance (i.e. P-error) of the re-weighting method and equally dramatic, but opposite, shift is in the performance of ERM on a sample of Q (which is close to h^*_Q).

While the bound from (Ben-David et al., 2006) implies that $\mathrm{Err}^P(h^*_Q)$ is bounded by $\mathrm{Err}^P(H) + \lambda_H(P, Q) + d_{H\Delta H}(P, Q)$, one could have hoped that, by re-weighting the sample S to reflect the distribution P_X, the term $d_{H\Delta H}(P^S, Q)$ in that bound would be diminished. Example 8 shows that this may not be the case. The P-error of $ERM(P^S)$ maybe as bad as that bound allows.

5 Impossibility Results

We are interested in the question "under what conditions is DA possible?" Clearly, the success of DA is conditioned upon some type of "similarity" or "relatedness" between the data distribution that we get our labels from and the target data distribution that we use to evaluate our error by. In our formalism this translates to the question "what relations between probability distributions suffice for DA?"

The next theorem shows that some intuitive conditions, that have been proposed in the literature for that purpose, do not suffice to guarantee the success of DA learning. In particular, among the three assumptions that we have been discussing—covariate shift, small $d_{\mathcal{A}}$ distance between the unlabeled distributions and the existence of hypotheses that mutually succeed on both the training and test domains (small λ_H), the last two are both necessary (and, as we know from previous results, are also sufficient).

Theorem 1 (Necessity of small $d_{\mathcal{A}}(P_X, Q_X)$)**.** *Let X be some domain set, and H a class of functions over X. Assume that, for some $A \subseteq X$, $\{h^{-1}(1) \cap A : h \in H\}$ contains more than two sets and is linearly ordered by inclusion. Then, the conditions covariate shift plus small λ_H do not suffice for DA. In particular, for every $\epsilon > 0$ there exists probability distributions Q over $X \times \{0, 1\}$, P over X such that for **every** domain adaptation learner A, every integers $m, n > 0$, there exists a labeling function $f : X \to \{0, 1\}$ such that*

1. *$\lambda_H(P_f, Q) \leq \epsilon$.*
2. *P_f and Q satisfy the covariate shift assumption.*

3. $\Pr_{\substack{L\sim Q^m\\ U\sim P^n}}\left[\mathrm{Err}^{P_f}(A(L,U)) \geq 1/2\right] \geq 1/2.$

Proof Sketch. The idea is to construct a (labeled) distribution Q, (unlabeled) distribution P (over X), and two labelers g and g' so that, as far as the input to a DA learner is concerned, the triples (Q, P_g, H) and $(Q, P_{g'}, H)$ are indistinguishable, while, at the same time, any hypothesis that has small error on P_g fails badly on $P_{g'}$ and vice versa. It follows, that A cannot tell which of the two target distributions it is trying to learn, and any hypothesis it outputs will have an error of at least 0.5 w.r.t. one of these potential targets. So given any A, we'll pick the target on which it fails. We demonstrate the idea of the construction for the simple case of X being the real line and H the class of thresholds. It is easy to extend the examples underlying this proof to the case of a general H stated in the theorem.

Consider Examples 7 and 8 except we modify Q_X so that on the interval $[1-\epsilon, 1]$ we apply the "odd points" construction of Example 6 (also see Figure 1)[1]. Similarly, for P we apply the "even points" construction restricted to $[1-\epsilon, 1]$ and make it uniform. We label Q as in Example 7 except in $[1-\epsilon, 1]$ we only label the "odd points" as 1 (the rest of the points in that interval will be labeled by either g or g'). The support of Q and P are disjoint, so we can satisfy covariate shift by defining g, g' to be consistent with Q on support of Q while on the support of P (the "even points") g will label 1 and g' label 0 (λ_H will be small whether g or g' is used for f). Note that on any sample L, U whatever A outputs the sum of the errors (w.r.t. P_g and $P_{g'}$) is 1. For sake of argument suppose A is deterministic (idea holds if allowed randomization), then we look at the sets

$$\begin{aligned} G &= \{(L,U) : |L| = m, |U| = n, \mathrm{Err}^{P_g}(A(L,U)) \geq 1/2\} \\ G' &= \{(L,U) : |L| = m, |U| = n, \mathrm{Err}^{P_{g'}}(A(L,U)) > 1/2\} \end{aligned}$$

and of course G and G' have disjoint union the set of all (L, U) with $|L| = m, |U| = n$. We choose $f = g$ if $\Pr(G) \geq 1/2$ and g' otherwise. It is easy to see that (3) follows. □

This theorem also applies to broader hypothesis classes, such as linear, homogeneous halfspaces in $\mathbb{R}^d$. Note that we could have included the assumption of (Mansour et al., 2009b) in the theorem since both (Q, P_g) and $(Q, P_{g'})$ satisfy their assumption.

Theorem 2 (Necessity of small $\lambda_H(P,Q)$)**.** *Let X be some domain set, and H be a class of functions over X whose VC dimension is much smaller than $|X|$ (in particular, any H with a finite VC dimension over an infinite X will do). Then, the conditions covariate shift plus small $d_{H\Delta H}(P,Q)$ do not suffice for DA. In particular, for every $\epsilon > 0$ there exist probability distributions Q over $X \times \{0,1\}$, P over X such that for **every** DA learner A, every integers $m, n > 0$, there exists a labeling function $f : X \to \{0,1\}$ such that*

1. $d_{H\Delta H}(P, Q_X) \leq \epsilon.$

2. The covariate shift assumption holds.

3. $\Pr_{\substack{L\sim Q^m\\ U\sim P^n}}\left[\mathrm{Err}^{P_f}(A(L,U)) \geq 1/2\right] \geq 1/2.$

Proof Sketch. The proof follows the same path as the above proof. Consider Example 6, and use the Q distribution there (concentrated on "odd points" with labels being 0), let P be the unlabeled distribution in the example (on "even points"), so that (1) is satisfied. Now there are two candidate target functions g, g' constructed by making them both agree with Q on its support, but on support of P, g will label all 1's and g' labels 0. Note that covariate shift holds regardless if we pick f to be g or g'. Further, A cannot tell whether g or g' is the true labeler on P and makes total error of 1 on P_g and $P_{g'}$ combined. Using similar arguments in the above proof, we can establish (3). □

6 Conclusion

We have analyzed the problem of domain adaptation in the setting where the learner has access to labeled examples drawn from the training data distribution and to unlabeled examples drawn from the target distribution. We considered three types of assumptions concerning the relationship between the domain and target distributions. These are:

1. The training and target distributions are close w.r.t. the $d_{H\Delta H}$ distance
2. There exist a hypothesis in H that has low error on both distributions.
3. The covariate shift assumption—the labeling function does not change between the training and target data.

We conclude that neither of the assumption combinations 1+3 nor 2+3 suffices to guarantee successful domain adaptation. These results hold w.r.t. any possible learning algorithm, as long as no further relatedness assumption are imposed on the training and target distributions. Recalling that (Ben-David et al., 2006) have shown that assuming 1+2 does imply learnability (even with the most straightforward learning algorithm), our results fully clarify the implications of

[1] Q_X is not strictly uniform, but piecewise uniform on two regions

these assumptions. In particular, the results demonstrate that the popular covariate shift assumption (see e.g. (Sugiyama and Mueller, 2005)) is in a sense irrelevant in this setting—its addition cannot replace any of the other assumptions, and it becomes redundant when the other two assumptions hold.

The natural follow-up challenges for understanding domain adaptation is to come up with and formalize other types of relatedness assumptions that may reflect the intuition of domain experts when domain adaptation works for their setting, and find ways to prove their utility. Another follow-up question is to relax our model of domain adaptation by allowing the learner access to some labeled examples from the target domain. The questions in such a setting are mainly to find conditions under which the labeled training-domain data can be utilize to improve learnability based on just target examples.

References

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

J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *Advances in Neural Information Processing Systems*, 2007.

C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample selection bias correction theory. In *Proceedings of the 19th Annual Conference on Algorithmic Learning Theory*, 2008.

K. Crammer, M. Kearns, and J. Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9:1757–1774, 2008.

Jiayuan Huang, Arthur Gretton, Bernhard Schlkopf, Alexander J. Smola, and Karsten M. Borgwardt. Correcting sample selection bias by unlabeled data. In *In NIPS*. MIT Press, 2007.

D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. In *Very Large Databases*, 2004.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems*, 2009a.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In Sanjoy Dasgupta and Adam Klivans, editors, *Proceedings of the Conference on Learning Theory*, 2009b.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors. *Dataset Shift in Machine Learning*. The MIT Press, 2009. ISBN 0262170051, 9780262170055.

M. Sugiyama and K. Mueller. Generalization error estimation under covariate shift. In *Workshop on Information-Based Induction Sciences*, 2005.