Domain Adaptation–Can Quantity Compensate for Quality?

Shai Ben-David
David R. Cheriton
School of Computer Science
University of Waterloo
Waterloo, ON N2L 3G1
CANADA
shai@cs.uwaterloo.ca

Shai Shalev-Shwartz
School of Computer Science & Engineering
The Hebrew University of Jerusalem
Givat Ram, Jerusalem 91904
ISRAEL
shais@cs.huji.ac.il

Ruth Urner
David R. Cheriton
School of Computer Science
University of Waterloo
Waterloo, ON N2L 3G1
CANADA
rurner@cs.uwaterloo.ca

Abstract

The Domain Adaptation problem in machine learning occurs when the distribution generating the test data differs from the one that generates the training data. A common approach to this issue is to train a standard learner for the learning task with the available training sample (generated by a distribution that is different from the test distribution). In this work we address this approach, investigating whether there exist successful learning methods for which learning of a target task can be achieved by substituting the standard target-distribution generated sample by a (possibly larger) sample generated by a different distribution without worsening the error guarantee on the learned classifier. We give a positive answer, showing that this is possible when using a Nearest Neighbor algorithm. We show this under the assumptions of covariate shift as well as a bound on the ratio of the probability weights between the source (training) and target (test) distribution. We further show that these assumptions are not always sufficient to allow such a replacement of the training sample: For proper learning, where the output classifier has to come from a predefined class, we prove that any learner needs access to data generated from the target distribution.

Introduction

Much of the theoretical analysis of machine learning focuses on a model where 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. The data-generating distribution might change over time or there might simply not be any labeled data available from the relevant target domain to train a classifier on. The task of learning when the training and test data-generating distributions differ is referred to as *Domain Adaption* (DA) learning.

Domain Adaptation tasks occur in many practical situations and are frequently addressed in experimental research (e.g. recently in (Daumé III and Jagarlamudi 2011)). For example, in natural language processing one might be able to access labeled documents of a certain type, say from legal documents, but needs to build a classifier to label the content of documents of a different type, say medical documents.

Two general approaches have been employed to deal with Domain Adaptation: First, using a standard learner for the learning task on the target distribution and just feed this learner with labeled data generated by a different, the so-called *source* distribution (we call these type of learners *conservative* Domain Adaptation learners); Second, designing special methods that aim to make use of the knowledge that the labeled data comes from a possibly different distribution trying to utilize unlabeled data from the target distribution to overcome this flaw.

In this work, we are mainly concerned with conservative DA learners. We are interested in the question, for which successful learning methods we can replace the labeled training data from the target distribution by a (possibly larger) labeled sample from the source distribution without forfeiting the learning success? Or, put differently, when can a large quantity of low quality data (as generated by a different distribution) replace a high quality training sample, without worsening the error guarantee on the learned classifier? The main contribution of this paper is showing that there exist learning methods that enjoy this property.

Obviously, Domain Adaptation is not possible when the training data generating distribution is not related to the test distribution. We consider two, rather basic, assumptions about the relationship between the source and target distributions. The first one is that the conditional label distributions are the same for both distribution, which is often assumed and commonly referred to as the *covariate-shift assumption* (see e.g. (Sugiyama and Mueller 2005)). Besides the covariate shift assumption we assume a bound on the ratio of the probability weights between the two marginal (unlabeled) distributions for certain collections of subsets of the domain.

A simple (but also unrealistically restrictive) assumption is to consider such a bound on a point-wise weight ratio between source and target. We start by showing that under this condition whenever the error of some standard learner goes to 0 with increasing sample sizes, we can replace the input to this standard learner with a sample from a different distribution by increasing the size of the sample by a factor of this weight ratio (see Observation 5). However, this fails as soon as the distribution does not admit a zero-error classifier or the algorithm is only guaranteed to converge to the approximation error of a certain class.

In the main part of the paper, we extend this result to the case of non-zero Bayes error. Assuming a bound on the weight ratio of boxes in $\mathbb{R}^d$ (which is weaker than assuming the point-wise ratio bound), we show that the Nearest

Neighbor algorithm has the desired property: We can feed it with a sample from a source distribution (whose size depends on the usually required size and the box-wise weight ratio) without worsening the error guarantee.

This result gives rise to the question whether there exist other algorithms that allow a replacement of the target generated input sample by one generated by a different distribution. Is a bound on the point-wise weight ratio sufficient for every algorithm to allow such a replacement? Are there other algorithms for which a bound on a different collection of subsets is suitable to have this property?

In the last part of this paper we give a negative answer to the second question in the setting of *proper* DA learning, where the learner is required to output a predictor from some pre-determined class. Such learners are relevant when additional requirements are imposed on the learned predictor, such as being fast at prediction time. We show that there are cases where no standard learner can enjoy the same success when fed with labeled data from a different distribution even under the assumption of a bound on the point-wise weight ratio. As an aside, we present a non-conservative learning paradigm that is guaranteed to succeed in this setting.

Related work The basic formal model of DA we follow in this work is defined in (Ben-David et al. 2006). It assumes that the learner has access to a labeled sample generated by the source distribution, but that the only information it has about the target distribution is conveyed by an unlabeled sample of that target distribution. Below, we discuss some of the assumptions (or parameters of the relatedness between the source and target tasks) that have been proposed to facilitate successful DA. We focus on assumptions that are related to those employed in this paper.

(Ben-David et al. 2006) examine the Domain Adaptation problem with respect to a given hypotheses class H. They propose to measure the relatedness of the two distributions by two parameters; the so-called $d_{\mathcal{A}}$ distance as introduced by (Kifer, Ben-David, and Gehrke 2004) (which is related to the weight ratio measures we introduce), and the minimum, over all hypotheses $h \in H$, of the sum of the errors of the hypothesis over the two tasks. That paper provides an upper bound, in terms of these parameters, on the error of the simplest conservative Domain Adaptation algorithm—the empirical risk minimization (ERM) over the training data. However, for the analysis provided in (Ben-David et al. 2006), feeding the ERM with examples from the source does not give the same guarantee as feeding the ERM with examples from the target. In the latter, the error will converge to the approximation error of the class H. In the former, the error will converge to the approximation error plus an additive error term that comes from the measure of discrepancy between the distributions. Thus, the error guarantee deteriorates.

A follow-up paper, (Mansour, Mohri, and Rostamizadeh 2009), extends the $d_{\mathcal{A}}$ distance to real-valued function classes and provides Rademacher-based bounds for more general loss functions. For the $0-1$ loss their bounds are incomparable with those in (Ben-David et al. 2006). In addition, they propose a non-conservative learning paradigm—re-weighting the examples of the source training data so that the re-weighted (labeled) empirical training distribution is closer to the (unlabeled) empirical target distribution.

The covariate shift assumption, stating that the conditional (label) distributions 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)). These papers utilize the covariate shift assumption by applying methods such as instance re-weighting. (Cortes, Mansour, and Mohri 2010) propose a non-conservative Domain Adaptation paradigm with provable success rates, assuming different relaxations of the point-wise weight ratio assumption.

Notation and basic definitions Let $(\mathcal{X}, \mu)$ be some domain set where $\mu : \mathcal{X}^2 \to \mathbb{R}^+$ is a metric over $\mathcal{X}$. We aim to learn a function $f : \mathcal{X} \to \{0,1\}$ that assigns labels to points in $\mathcal{X}$ with low error probability with respect to some *target distribution* P over $\mathcal{X} \times \{0,1\}$. For such a target distribution, P, and $h : \mathcal{X} \to \{0,1\}$, we define the *error* of h with respect to P as $\mathrm{Err}_P(h) = \Pr_{(x,y)\sim P}(y \neq h(x))$. We denote the Bayes optimal error for P by $\mathrm{opt}(P) := \min_{h\in\{0,1\}^{\mathcal{X}}} \mathrm{Err}_P(h)$. For a class H of hypotheses on $\mathcal{X}$, we denote the approximation error of H with respect to P by $\mathrm{opt}_H(P) := \min_{h\in H} \mathrm{Err}_P(h)$.

In the Domain Adaptation setup, where the training and test data generating distributions differ, we use the following notation: Let P_S and P_T be two distributions over $\mathcal{X} \times \{0,1\}$. We call P_S the *source distribution* and P_T the *target distribution*. We denote the marginal distribution of P_S over $\mathcal{X}$ by D_S and the marginal of P_T by D_T, and their labeling rules by $l_S : \mathcal{X} \to [0,1]$ and $l_T : \mathcal{X} \to [0,1]$, respectively (where, for a probability distribution P over $X \times \{0,1\}$, the associated labeling rule is the conditional probability of label 1 at any given point: $l(x) = \Pr_{(X,Y)\sim P}(Y = 1|X = x)$).

A Domain Adaptation learner takes as input a labeled i.i.d. sample S drawn according to P_S and an unlabeled i.i.d. sample T drawn according to D_T and aims to generate a good label predictor $h : \mathcal{X} \to \{0,1\}$ for P_T. Formally, a *Domain Adaptation (DA) learner* is a function

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

A Domain Adaptation learner A is *conservative* if it ignores the unlabeled sample it receives from the target distribution. Namely, $A(U,V) = A(U,W)$ for all $U \in \bigcup_{m=1}^{\infty}(\mathcal{X} \times \{0,1\})^m$ and $V, W \in \bigcup_{n=1}^{\infty} \mathcal{X}^n$.

Definition 1 (DA-learnablity)**.** Let $\mathcal{X}$ be some domain, $\mathcal{W}$ be a class of pairs (P_S, P_T) of distributions over $\mathcal{X} \times \{0,1\}$ and $\mathcal{A}$ be a DA learner.

- **General DA:** We say that $\mathcal{A}$ $(c, \epsilon, \delta, m, n)$*-solves DA for the class* $\mathcal{W}$, if, for all pairs $(P_S, P_T) \in \mathcal{W}$, when given access to a labeled sample S of size m, generated i.i.d. by P_S, and an unlabeled sample T of size n, generated i.i.d by D_T, with probability at least $1-\delta$ (over the choice of the samples S and T) $\mathcal{A}$ outputs a function h with $\mathrm{Err}_{P_T}(h) \leq c \cdot \mathrm{opt}(P_T) + \epsilon$.

- **Proper DA:** If H is a class of hypotheses, we say that $\mathcal{A}$ $(c, \epsilon, \delta, m, n)$*-solves proper DA for the class $\mathcal{W}$ relative to* H, if, for all pairs $(P_S, P_T) \in \mathcal{W}$, when given access to a labeled sample S of size m, generated i.i.d. by P_S, and an unlabeled sample T of size n, generated i.i.d by D_T, with probability at least $1 - \delta$ (over the choice of the samples S and T), $\mathcal{A}$ outputs an element h of H with $\mathrm{Err}_{P_T}(h) \leq c \cdot \mathrm{opt}_H(P_T) + \epsilon$.

Note the difference between these two learnability notions: For proper learning, we require that the output of the learner is a member of the class H, and the error of the algorithm is measured relative to the approximation error of H and not the Bayes error.

Properties that may help Domain Adaptation

Clearly, the success of Domain Adaptation (DA) learning cannot be achieved for every source-target pair of learning tasks. A major challenge for DA research is to discover conditions, or properties of learning tasks, that enable successful DA learning. Such properties express either some relationship between the source and target distributions or some "niceness" conditions of these distributions that facilitate learning. To be relevant to practical learning challenges, such properties should be conceivable from the point of view of realistic learning problems. In this chapter we define some such properties. The remainder of the paper is devoted to investigating the extent by which these properties indeed ease DA learning.

Covariate shift

The first property we mention is often assumed in Domain Adaptation analysis (e.g. (Sugiyama and Mueller 2005)). In this work, we assume this property throughout.

Definition 2 (Covariate shift)**.** We say that source and target distribution satisfy the *covariate shift* property if they have the same labeling function, *i.e.* if we have $l_S(x) = l_T(x)$ for all $x \in \mathcal{X}$.

In the sequel, we denote this common labeling function of P_S and P_T by l. The covariate shift assumption makes sense for many realistic DA tasks. For example, in many natural language processing (NLP) learning problems, such as parts of speech tagging, where a learner that trains on documents from one domain (say, news articles) is applied to a different domain (say, legal documents). For such tasks, it is reasonable to assume that the difference between the two tasks is only in their marginal distributions over English words rather than in the tagging of each word. While, on first thought, it may seem like under this assumption DA becomes easy, a closer look reveals that it is a rather weak assumption–a DA learner has no clue as to how the common labeling function may behave outside the scope of the source-generated labeled sample, as long as there are no restrictions on the labeling function.

Probabilistic Lipschitzness

We first recall that a function $f : \mathcal{X} \to \mathbb{R}$ satisfies the (standard) λ-Lipschitz property (with respect to the underlying metric μ), if $|f(x) - f(y)| \leq \lambda\, \mu(x, y)$ holds for all $x, y \in \mathcal{X}$. This condition can be readily applied to probabilistic labeling rules $l : X \to [0, 1]$. However, if the labeling function is deterministic, namely if $l(x) \in \{0, 1\}$ for all $x \in \mathcal{X}$, this requirement forces a $1/\lambda$ gap between differently labeled points, and will thus fail whenever l is non-constant on some connected subregion of its support. A natural relaxation is to require that the inequality will hold only with some high probability. Namely,

Definition 3 (Probabilistic Lipschitzness)**.** Let $\phi : \mathbb{R} \to [0, 1]$. We say that $f : \mathcal{X} \to \mathbb{R}$ is ϕ-Lipschitz w.r.t. a distribution D over $\mathcal{X}$ if, for all $\lambda > 0$:

$$\Pr_{x \sim D}[\exists y : |f(x) - f(y)| > \lambda\, \mu(x, y)] \leq \phi(\lambda)$$

This definition generalizes the standard definition since, given any $\lambda > 0$, setting $\phi(a) = 1$ for $a < \lambda$ and $\phi(a) = 0$ for $a \geq \lambda$ results in the standard λ–Lipschitzness condition.

It is worthwhile to note that this probabilistic Lipschitzness condition may be viewed as a way of formalizing the *cluster assumption* that is commonly made to account for the success of semi-supervised learning. It implies that the data can be divided into clusters that are almost label-homogeneous and are separated by low-density regions. See (Urner, Ben-David, and Shalev-Shwartz 2011) for such an application of a similar notion.

Weight ratio assumptions

One basic observation about DA learning is that it may become impossible when the source and target distributions are supported on disjoint domain regions. To guard against such scenarios, it is common to assume that there is some non-zero lower bound to the pointwise density ratio between the two distributions. However, this is often an unrealistic assumption. Going back to the NLP example, it is likely that there are some technical legal terms that may occur in legal documents but will never show up in any Reuters news article. Furthermore, such a pointwise assumption cannot be verified from finite samples of the domain and target distributions. To overcome these drawbacks, we propose the following relaxation of that assumption.

Definition 4 (Weight ratio)**.** Let $\mathcal{B} \subseteq 2^{\mathcal{X}}$ be a collection of subsets of the domain $\mathcal{X}$. For some $\eta > 0$ we define the η*-weight ratio* of the source distribution and the target distribution with respect to $\mathcal{B}$ as

$$C_{\mathcal{B},\eta}(D_S, D_T) = \inf_{\substack{b \in \mathcal{B} \\ D_T(b) \geq \eta}} \frac{D_S(b)}{D_T(b)},$$

Further, we define the *weight ratio* of the source distribution and the target distribution with respect to $\mathcal{B}$ as

$$C_{\mathcal{B}}(D_S, D_T) = \inf_{\substack{b \in \mathcal{B} \\ D_T(b) \neq 0}} \frac{D_S(b)}{D_T(b)},$$

These measures become relevant for Domain Adaptation when bounded away from zero.

Note that the pointwise weight ratio mentioned above can be obtained by setting $\mathcal{B} = \{\{x\} : x \in \mathcal{X}\}$. For every $\mathcal{B} \subseteq 2^{\mathcal{X}}$ we have $C_{\{\{x\}:x \in \mathcal{X}\}}(D_S, D_T) \leq C_{\mathcal{B}}(D_S, D_T)$, thus bounding the pointwise weight ratio away from 0 is the strongest restriction.

A basic DA error bound

Observation 5. *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 $C_{\{\{x\}:x\in\mathcal{X}\}}(D_S,D_T) > 0$. Then we have $\mathrm{Err}_{P_T}(h) \le \frac{1}{C}{}_{\{\{x\}:x\in\mathcal{X}\}}\mathrm{Err}_{P_S}(h)$ for all $h : \mathcal{X} \to \{0,1\}$.*

Note that this result implies that any learning algorithm that can achieve arbitrarily small target error when it can access target generated training samples, can also achieve this based on only source generated samples. However, if there is a positive lower bound on the error guarantee of the algorithm (e.g., due to a non-zero Bayes error, or to an approximation error of the algorithm) then Observation 5 becomes meaningless as soon as the weight-ratio, $C_{\{\{x\}:x\in\mathcal{X}\}}$, is smaller than that error lower bound.

General DA-learning

In this section we consider general DA learning bounds that are meaningful regardless of the algorithm being able to predict with arbitrarily small error. We show that for the Nearest Neighbor algorithm a target generated sample can be replaced by a source generated sample while maintaining the error guarantee. For this, we employ a Lipschitzness assumption on the labeling function and a weight-ratio assumption w.r.t. the class of axis-aligned rectangles. Note that if $\mathcal{B}$ is this set of axis aligned rectangles, we can estimate the η-weight ratio from finite samples (see Theorem 3.4 and the subsequent discussion of (Kifer, Ben-David, and Gehrke 2004)).

Let $\mathrm{NN}(P_S)$ be the Nearest Neighbor method w.r.t. the source labeled training sample. Given a labeled sample $S \subseteq \mathcal{X} \times \{0,1\}$, $\mathrm{NN}(P_S)$ outputs a function h_{NN} that assigns to each point the label of its nearest neighbor in the sample S. We will analyze the performance of $\mathrm{NN}(P_S)$ as a function of the Lipschitzness and the weight ratio.

Let $S_{\mathcal{X}}$ denote the sample points of S without labels (namely, $S_{\mathcal{X}} := \{x \in \mathcal{X} \mid \exists y \in \{0,1\} : (x,y) \in S\}$). For any $x \in S_{\mathcal{X}}$, let $l_S(x)$ denote the label of the point x in the sample S. Given some labeled sample set S and a point $x \in \mathcal{X}$, let $N_S(x)$ denote the nearest neighbor to x in S, $N_S(x) = \mathrm{argmin}_{z\in S_{\mathcal{X}}}\mu(x,z)$. We define h_{NN} for all points $x \in \mathcal{X}$ by $h_{\mathrm{NN}}(x) = l_S(N_S(x))$.

We will assume that our domain is the unit cube in $\mathbb{R}^d$. Furthermore, we will assume that the weight ratio for the class B of axis-aligned rectangles in $\mathbb{R}^d$ is bounded away from zero. We begin with a basic lemma.

Lemma 6. *Let $C_1, C_2, \ldots, C_r$ be subsets of some domain set $\mathcal{X}$ and let S be a set of points of size m sampled i.i.d. according to some distribution P over that domain. Then, $\mathbb{E}_{S\sim P^m}\left[\sum_{i:C_i\cap S=\emptyset} P[C_i]\right] \le \frac{r}{me}$.*

Proof. From the linearity of expectation, we get

$$\mathop{\mathbb{E}}_{S\sim P^m}\Big[\sum_{i:C_i\cap S=\emptyset} P[C_i]\Big] = \sum_{i=1}^{r} P[C_i] \mathop{\mathbb{E}}_{S\sim P^m}[\mathbf{1}_{C_i\cap S=\emptyset}] .$$

Next, for each i we have

$$\begin{aligned}\mathop{\mathbb{E}}_{S\sim P^m}[\mathbf{1}_{C_i\cap S=\emptyset}] &= \Pr_{S\sim P^m}[C_i \cap S = \emptyset] \\ &= (1-P[C_i])^m \le \mathrm{e}^{-P[C_i]\,m} .\end{aligned}$$

Combining the above two equations and using $\max_a a\mathrm{e}^{-ma} \le \frac{1}{me}$ conclude our proof. □

We now show that if we assume that the (possibly deterministic) labeling function satisfies the probabilistic Lipschitzness and we have a lower bound on the weight ratio w.r.t. the class $\mathcal{B}$ above, then the nearest neighbor algorithm solves the Domain Adaptation problem.

Theorem 7. *Let our domain $\mathcal{X}$ be the unit cube, $[0,1]^d$, and for some $C > 0$, let $\mathcal{W}$ be a class of pairs (P_S, P_T) of source and target distributions over $\mathcal{X} \times \{0,1\}$ satisfying the covariate shift assumption, with $C_{\mathcal{B}}(D_S,D_T) \ge C$, and their common labeling function $l : \mathcal{X} \to [0,1]$ satisfying the ϕ-probabilistic-Lipschitz property with respect to the target distribution, for some function ϕ. Then, for all λ,*

$$\mathbb{E}_{S\sim P_S^m}[\mathrm{Err}_{P_T}(h_{\mathrm{NN}})] \le 2\mathrm{opt}(P_T)+\phi(\lambda)+4\lambda\frac{\sqrt{d}}{C}m^{-\frac{1}{d+1}} .$$

Proof. We start by proving that

$$\begin{aligned}&\mathop{\mathbb{E}}_{S\sim P_S^m}[\mathrm{Err}_{P_T}(h_{NN})] \qquad (1)\\ &\le\ 2\mathrm{opt}(P_T) + \phi(\lambda) + \lambda \mathop{\mathbb{E}}_{S\sim P_S^m, x\sim D_T}[\|x - N_S(x)\|_2]\end{aligned}$$

We first note that given two instances x, x' we have

$$\begin{aligned}\Pr_{y\sim l(x), y'\sim l(x')}[y \ne y'] &= l(x)(1-l(x')) + l(x')(1-l(x)) \\ &\le 2l(x)(1-l(x)) + |l(x') - l(x)| ,\end{aligned}$$

where the last inequality follows by standard algebraic manipulations. The error of the NN procedure can be therefore written as

$$\begin{aligned}&\mathbb{E}_{S\sim P_S^m}[\mathrm{Err}_{P_T}(h_{\mathrm{NN}})]\\ &= \mathbb{E}_{S\sim P_S^m}\mathbb{E}_{x\sim P_T} \Pr_{y\sim l(x), y'\sim l(N_S(x))}[y \ne y'] \\ &\le \mathbb{E}_{S\sim P_S^m}\mathbb{E}_{x\sim P_T}\left[2l(x)(1-l(x)) + |l(N_S(x)) - l(x)|\right] \\ &\le 2\mathrm{opt}(P_T) + \mathbb{E}_{S\sim P_S^m}\mathbb{E}_{x\sim P_T}[|l(N_S(x)) - l(x)|] .\end{aligned}$$

Using the definition of probabilistic Lipschitzness and the fact that the range of l is $[0,1]$, no matter what S is, we have

$$\mathbb{E}_{x\sim P_T}[|l(N_S(x)) - l(x)|] \le \phi(\lambda)+\lambda\mathbb{E}_{x\sim P_T}[\|N_S(x)-x\|_2] ,$$

which yields equation 1. Thus, in order to prove learnability, we need an upper bound on $\mathbb{E}_{S\sim P_S^m, x\sim D_T}[\|x - N_S(x)\|_2]$.

Now, fix some $\gamma > 0$ and let $C_1, \ldots, C_r$ be the cover of the set $[0,1]^d$ using boxes of side-length γ. We have $D_T(C_i) \le \frac{1}{C_{\mathcal{B}}(D_S,D_T)} D_S(C_i) \le \frac{1}{C} D_S(C_i)$ for all boxes C_i. Thus, Lemma 6 yields

$$\begin{aligned}&\mathbb{E}_{S\sim D_S^m}\left[\sum_{i:C_i\cap S=\emptyset} D_T[C_i]\right] \le \mathbb{E}_{S\sim D_S^m}\left[\sum_{i:C_i\cap S=\emptyset} \tfrac{1}{C} D_S[C_i]\right]\\ &\le \tfrac{1}{C}\,\mathbb{E}_{S\sim D_S^m}\left[\sum_{i:C_i\cap S=\emptyset} D_S[C_i]\right] \le \tfrac{r}{Cme}\end{aligned}$$

For each x, x' in the same box we have $\|x - x'\|_2 \leq \sqrt{d}\,\gamma$. Otherwise, $\|x - x'\|_2 \leq 2\sqrt{d}$. For $x \in \mathcal{X}$ we let $C_x \in \{C_1, \ldots, C_r\}$ denote the box that contains the point x. Therefore,

$$\begin{aligned} & \mathop{\mathbb{E}}_{S \sim P_S^m, x \sim D_T} [\|x - N_S(x)\|_2] \\ \leq \; & \mathop{\mathbb{E}}_{S \sim P_S^m} [\mathop{\Pr}_{x \sim D_T} [C_x \cap S = \emptyset]\, 2\sqrt{d} + \mathop{\Pr}_{x \sim D_T} [C_x \cap S \neq \emptyset]\, \sqrt{d}\,\gamma] \\ \leq \; & \sqrt{d}(\tfrac{2r}{m\mathrm{e}\,C} + \gamma)\,. \end{aligned}$$

Since the number of boxes is $(2/\gamma)^d$ we get that

$$\mathop{\mathbb{E}}_{S \sim P_S^m, x \sim D_T} [\|x - N_S(x)\|] \leq \sqrt{d} \left(\frac{2^{d+1}\,\gamma^{-d}}{m\,\mathrm{e}\,C} + \gamma \right)\,.$$

Combining this with equation 1, we get

$$\begin{aligned} & \mathbb{E}_{S \sim P_S^m}[\mathrm{Err}_{P_T}(h_{\mathrm{NN}})] \\ \leq \quad & 2\mathrm{opt}(P_T) + \phi(\lambda) + \lambda\sqrt{d} \left(\frac{2^{d+1}\,\gamma^{-d}}{m\,\mathrm{e}\,C} + \gamma \right) \\ \leq \quad & 2\mathrm{opt}(P_T) + \phi(\lambda) + \lambda\sqrt{d}\,\frac{1}{C} \left(\frac{2^{d+1}\,\gamma^{-d}}{m\,\mathrm{e}} + \gamma \right) \end{aligned}$$

as we have $0 < C \leq 1$. Setting $\gamma = 2m^{-\frac{1}{d+1}}$ and noting that

$$\begin{aligned} \frac{2^{d+1}\gamma^{-d}}{me} + \gamma \quad &= \quad \frac{2^{d+1}\gamma^{-d}m^{d/(d+1)}}{m\mathrm{e}} + 2m^{-1/(d+1)} \\ &= \quad 2m^{-1/(d+1)}(1/\mathrm{e} + 1) \leq 4m^{-1/(d+1)} \end{aligned}$$

we obtain $\mathbb{E}_{S \sim P_S^m}[\mathrm{Err}_{P_T}(h_{\mathrm{NN}})] \leq 2\mathrm{opt}(P_T) + \phi(\lambda) + 4\lambda\sqrt{d}\frac{1}{C}m^{-\frac{1}{d+1}}$. □

Note that, if source and target data are the same, then the same asnalysis leads to an error bound of $\mathbb{E}_{S \sim P^m}[\mathrm{Err}_P(h_{\mathrm{NN}})] \leq 2\mathrm{opt}(P) + \phi(\lambda) + 4\lambda\sqrt{d}m^{-\frac{1}{d+1}}$. Thus, replacing a target labeled sample of size m by a source labeled sample of size $C^{d+1} \cdot m$ leads to the same error guarantee. If the labeling function is λ-Lipschitz in the standard sense of Lipschitzness and the labels are deterministic, then we have $\mathrm{opt}(P_T) = 0$ and $\phi(a) = 0$ for all $a \geq \lambda$. Applying Markov's inequality then yields the following sample size bound:

Corollary 8. *Let our domain $\mathcal{X}$ be the unit cube in $\mathbb{R}^d$ and for some $C > 0$, let $\mathcal{W}$ be a class of pairs (P_S, P_T) of source and target distributions over $\mathcal{X} \times \{0, 1\}$ with $C_{\mathcal{B}}(D_S, D_T) \geq C$ satisfying the covariate shift assumption and their common labeling function $l : \mathcal{X} \to \{0, 1\}$ satisfying the λ-Lipschitz property. Then, for all $\epsilon > 0$, $\delta > 0$ and all $m \geq \left(\frac{4\,\lambda\,\sqrt{d}}{C\,\epsilon\,\delta}\right)^{d+1}$ the nearest neighbor algorithm applied to a sample of size m, has, with probability at least $1 - \delta$, error of at most ϵ w.r.t. the target distribution for any pair $(P_S, P_T) \in \mathcal{W}$.*

Remark 1: For the results of Theorem 7 and 8 we can actually settle for the η-weight ratio (see Definition 4). For boxes of very small target-weight, we do not need to require the source distribution to have any weight at all. More precisely, since the number of boxes we are using to cover the space in the proof of Theorem 7 is $(2/\gamma)^d$, aiming for some value of ϵ, we could waive the requirement for boxes that have target weight less than $\gamma^d\epsilon/2$. Thus by assuming a lower bound on the $\gamma^d\epsilon/2$-weight ratio, the potential misclassification of these boxes sum up to at most ϵ and thus we only produce an additional error of ϵ.

Remark 2: It is well known that the exponential dependence on the dimension of the space in the bound of Theorem 7 and Corollary 8 is inevitable. To see this, consider a domain of points arranged on a grid of side-length $1/\lambda$ for some $\lambda > 0$. Every labelling function on these points is λ-Lipschitz. But as there are λ^d such points in a grid in the unit cube, a no-free-lunch argument shows that no algorithm can be guaranteed to learn a low-error classifier for the class of all distributions with λ-Lipschitz labelings unless it sees a sample of size in the order of λ^d. Note that this classical learning setting can be viewed as a Domain Adaptation task where we have a pointwise weight ratio 1 between source and target. Thus, this lower bound also applies to any Domain Adaptation learner for classes that satisfy the Lipschitz and bounded weight ratio conditions.

Proper DA learning

Recall that a DA algorithm is called proper if its output is a member of a predetermined hypothesis class. This requirement is important in several applications. For example, in some situations runtime of the learned classifier is an important factor, and one would prefer a faster classifier even at the expense of somewhat poorer predictions. If the hypothesis class only contains fast computable functions, then the properness of the DA algorithm guarantees that the algorithm will output a fast predictor. Another example is a user being interested in the explanatory aspects of the predictor, requiring the output hypothesis to belong to a family of functions that are readily interpretable. Linear classifiers are an obvious example of such desirable predictors, under both of these scenarios.

In this section, we show that in the context of proper Domain Adaptation, the use of algorithms that utilize target-generated data, is necessary. We show that there are classes that can not be properly learned without access to data from the test distribution:

Theorem 9. *Let our domain set be the unit ball in $\mathbb{R}^d$, for some d. Consider the class H of half-spaces as our target class. Let x and z be a pair of antipodal points on the unit sphere and let $\mathcal{W}$ be a set that contains two pairs (P_S, P_T) and (P_S, P'_T) of distributions with:*

1. *both pairs satisfy the covariate shift assumption,*
2. *$l(x) = l(z) = 1$ and $l(\bar{0}) = 0$ for their common labeling function l,*
3. *$D_S(x) = D_S(z) = D_S(\bar{0}) = 1/3$,*
4. *$D_T(x) = D_T(\bar{0}) = 1/2$ or $D'_T(z) = D'_T(\bar{0}) = 1/2$.*

Then, for any number m, any constant c, no proper DA learning algorithm can $(c, \epsilon, \delta, m, 0)$ solve the Domain Adaptation learning task for $\mathcal{W}$ with respect to H, if $\epsilon < 1/2$ and $\delta < 1/2$. (In other words, every conservative DA learner fails to solve the Domain Adaptation learning problem w.r.t. $\mathcal{W}$.)

Proof. Clearly, no halfspace can correctly classify the three points, $x, \bar{0}$ and y. Note that for any halfspace h, we have $\mathrm{Err}_{P_T}(h) + \mathrm{Err}_{P'_T}(h) \geq 1$, which implies $\mathrm{Err}_{P_T}(h) \geq 1/2$ or $\mathrm{Err}_{P'_T}(h) \geq 1/2$. Thus for every learner, there exists a target distribution (either P_T or P'_T) such that, with probability at least $1/2$ over the sample, outputs a function of error at least $1/2$. Lastly, note that the approximation error of the class of halfspaces for the target distributions is 0, thus the result holds for any constant c. □

In the example of the above Theorem it becomes crucial for the learning algorithm to estimate whether the support of the target distribution is x and $\bar{0}$ or z and $\bar{0}$. This information cannot be obtained without access to a sample of the target distribution despite of a point-wise weight ratio as large as $1/2$. Thus, no amount "low quality" (as source generated) data can compensate for having a sample from the target distribution.

We now present a general method for proper DA learning. The basic idea of our construction is to apply a simple two step procedure, similar to the one suggested in (Urner, Ben-David, and Shalev-Shwartz 2011) in the context of semi-supervised learning. In the first step, we use the labeled examples from the source distribution to learn an arbitrary predictor, which should be rather accurate on the target distribution. For example, as we have shown in the previous section, this predictor can be the NN rule. In the second step, we will apply that predictor to the unlabeled examples from the target distribution and feed this constructed (now labeled) sample to a standard agnostic learner for the usual supervised learning setting. Recall the definition of an agnostic learner: For $\epsilon >, \delta > 0$, $m \in \mathbb{N}$ we say that an algorithm (ϵ, δ, m) (agnostically) learns a hypothesis class H, if for all distributions P, when given an *i.i.d.* sample of size at least m, it ouputs a classifier of error at most $\mathrm{opt}_H(P) + \epsilon$ with probability at least $1 - \delta$. If the ouput of the algorithm is always a member of H, we call it a agnostic proper learner for H.

To prove that this two step procedure works, we first prove that agnostic learners are robust with respect to small changes in the input distribution.

Lemma 10. *Let P be a distribution over $\mathcal{X} \times \{0,1\}$, let $f : \mathcal{X} \rightarrow \{0,1\}$ be a function with $\mathrm{Err}_P(f) \leq \epsilon_0$, let $\mathcal{A}$ be an agnostic learner for some hypothesis class H over $\mathcal{X}$ and let $m : (0,1)^2 \rightarrow \mathbb{N}$ be a function such that A, for all $\epsilon, \delta > 0$ is guaranteed to $(\epsilon, \delta, m(\epsilon, \delta))$-learn H. Then, with probability at least $(1 - \delta)$ over an i.i.d. sample of size $m(\epsilon, \delta)$ from P's marginal labeled by f, $\mathcal{A}$ outputs a hypothesis h with $\mathrm{Err}_P(h) \leq \mathrm{opt}_H(P) + 2\epsilon_0 + \epsilon$.*

Proof. Let P' be the distribution that has the same marginal as P and f as its deterministic labeling rule. Note that for the optimal hypothesis h^* in H with respect to P we have $\mathrm{Err}_{P'}(h^*) \leq \mathrm{opt}_H(P) + \epsilon_0$. This implies, that when we feed the P'-generated sample S to the agnostic learner, it outputs an $h \in H$ with $\mathrm{Err}_{P'}(h) \leq \mathrm{opt}_H(P) + \epsilon_0 + \epsilon$ with probability at least $(1 - \delta)$ and this yields $\mathrm{Err}_P(h) \leq \mathrm{opt}_H(P) + 2\epsilon_0 + \epsilon$. □

Applying this lemma we readily get:

Theorem 11. *Let $\mathcal{X}$ be some domain and $\mathcal{W}$ be a class of pairs (P_S, P_T) of distributions over $\mathcal{X} \times \{0,1\}$ with $\mathrm{opt}(P_T) = 0$ such that there is an algorithm $\mathcal{A}$ and functions $m : (0,1)^2 \rightarrow \mathbb{N}$, $n : (0,1)^2 \rightarrow \mathbb{N}$ such that $\mathcal{A}$ $(0, \epsilon, \delta, m(\epsilon, \delta), n(\epsilon, \delta))$-solves the Domain Adaptation learning task for $\mathcal{W}$ for all $\epsilon, \delta > 0$. Let H be some hypotheses class for which there exists an agnostic proper learner. Then, the H-proper Domain Adaptation problem w.r.t. the class $\mathcal{W}$ can be $(1, \epsilon, \delta, m(\epsilon/3, \delta/2), n(\epsilon/3, \delta/2) + m'(\epsilon/3, \delta/2))$-solved, where m' is the sample complexity function for agnostically learning H.*

Proof. Given the parameters ϵ and δ, let S be a P_S-sample of size at least $m(\epsilon/3, \delta/2)$ and T be an unlabeled D_T-sample of size $n(\epsilon/3, \delta/2) + m'(\epsilon/3, \delta/2)$. Divide the unlabeled sample into a sample T_1 of size $n(\epsilon/3, \delta/2)$ and T_2 of size $m'(\epsilon/3, \delta/2)$. Apply $\mathcal{A}(S, T_1)$, the predictor resulting form applying the learner $\mathcal{A}$ to the S and T_1, to label all members of T_2, and then feed the now-labeled T_2 as input to the agnostic proper learner for H. The claimed performance of the output hypothesis now follows from Lemma 10. □

The algorithm $\mathcal{A}$ used in this theorem could be the Nearest Neighbor algorithm, $\mathrm{NN}(P_S)$, if the class $\mathcal{W}$ satisfies the conditions for Theorem 7. Overall, we have shown that with a non-conservative DA algorithm, that employs unlabeled examples from the target distribution, we can agnostically learn a member of the hypotheses class for the target distribution, whereas whithout target-generated data we can not.

Conclusion and Open Questions

When analyzing the generalization error of learning algorithms, it is common to decompose the error into three terms: **(1)** the *Bayes error*, which measures the inherent non-determinism in the labeling mechanism. **(2)** the *approximation error*, which is the minimum generalization error achievable by a predictor in a reference class of hypotheses. **(3)** the *estimation error*, which is a result of the training error being only an estimate of the true error.

In this paper we study Domain Adaptation problems, in which the source and target distributions are different. This introduces a forth error term: **(4)** The *distribution discrepancy error*, which is a result of the training examples and test examples being sampled from different distributions.

The main question we address is: "Which assumptions on the discrepancy between the two distributions make it possible to decrease the distribution discrepancy error by requiring more examples." This poses an interesting tradeoff between quality (how much the training examples reflect the target distribution) and quantity (how many examples we have). We showed that for Nearest Neighbor, with the covariate shift and a bound on weight ratio of boxes assumptions, quantity compensates for quality. We also showed that for proper DA, even infinite number of source examples cannot compensate for the distribution discrepancy, but unlabeled examples from the target distribution (which is another form of low quality examples) can compensate for the distribution discrepancy error. A major open question is whether there are additional algorithms for which quantity can compensate for quality.

References

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2006. Analysis of representations for domain adaptation. In *NIPS*, 137–144.

Cortes, C.; Mansour, Y.; and Mohri, M. 2010. Learning bounds for importance weighting. In Lafferty, J.; Williams, C. K. I.; Shawe-Taylor, J.; Zemel, R.; and Culotta, A., eds., *Advances in Neural Information Processing Systems 23*. 442–450.

Daumé III, H., and Jagarlamudi, J. 2011. Domain adaptation for machine translation by mining unseen words. In *Association for Computational Linguistics*.

Huang, J.; Gretton, A.; Schölkopf, B.; Smola, A. J.; and Borgwardt, K. M. 2007. Correcting sample selection bias by unlabeled data. In *In NIPS*. MIT Press.

Kifer, D.; Ben-David, S.; and Gehrke, J. 2004. Detecting change in data streams. In *VLDB*, 180–191.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2009. Domain adaptation: Learning bounds and algorithms. *CoRR* abs/0902.3430.

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

Urner, R.; Ben-David, S.; and Shalev-Shwartz, S. 2011. Unlabeled data can speed up prediction time. In *ICML*.