Learning Theory for Domain Adaptation

Roshan Jarupla
A53301504

Tarun Kalluri
A53297589

Abstract

In this report, we consider the problem of domain adaptation, where the task is to transfer a model trained on one domain to another domain. Classical bounds for learning theory, which use PAC agnostic learning algorithms implicitly assume that the train and test distributions are the same. Firstly, we show results that relax this assumption in an unsupervised target setting along with some experimental results on MNIST-SVHN datasets. Next, we present results for cases involving semi-supervised target and multi source adaptation. In the second half, we show the hardness or impossibility of domain adaptation and specifically cover the importance of unlabeled target samples in designing Domain Adaptation (DA) algorithm.

1 Introduction

Traditional learning methods perform well when the train and test data arrive from the same distribution. However, this need not be the case. There might be cases where the training is done on a source domain, but the testing or deployment needs to be done on a completely different domain. For example, spam filters might be trained on data corresponding to one user, but it needs to be deployed on data pertaining to many other users. Similarly, a sentiment classification system trained on amazon reviews might not capture well the sentiments on 911 distress calls. In computer vision, which is a field which has seen massive growth due to machine learning, image recognition models trained on images captured in daytime do not transfer well to images captured well to night time. Therefore, it is very essential to model distribution shifts in machine learning.

This is important for two reasons. Firstly, it helps us better understand the limitations of existing methods from a theoretical standpoint, that is when do the existing methods trained on one dataset fail? Secondly, they help us to design appropriate algorithm which can decrease the obtained bounds to improve the performance. In fact most modern deep domain adaptation approaches [4, 6, 8] make use of the theory proposed in [1], and we show in the experiment section the connection between the theory and the empirical results.

In current report, we majorly follow two papers, namely [1] and [3]. In the first part, we address the following questions. First, under what conditions can a classifier learned from a source data can be expected to perform well on a different target data? Secondly, given a small amount of labeled target data, what should be an optimum combination strategy with the large labeled training examples from source data to achieve optimum performance? We also consider the case of multi source domain adaptation and provide bounds in terms of a combination of source domains.

In the latter half, we relax the assumption of existence of an optimum hypothesis that performs well on both source and target. Specifically, we discuss the case when the hypothesis is realizable on target and the divergence between source and target is bounded and also the density ratio is bounded, but still the sample complexity required to bound the adaptation error grows with the domain size. Furthermore, we also discuss the role of unlabeled samples in target, and examine if increasing the amount of unlabeled target samples can compensate for the explosion of labeled samples required from the source.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

The report is organized as follows. In section 2, we present the first part of the report and discuss the theory of learning from different domains. In section 3, we discuss the second part of the report and discuss the hardness results for Domain Adaptation and the utility of unlabeled target samples. Finally in Section 3, we present the experimental results from the first paper of the report and include our own experimental results.

2 A theory of learning from different domains

This part of the report is prepared by reading the paper *A theory of learning from different domains* [1]. The broad idea in this section is as follows. We first bound the classifier's target error in terms of its source error and the divergence between the domains. We use a classifier induced divergence metric that can be reliably estimated from finite unlabeled samples from the domains. Under the assumption of existence of an optimum joint hypothesis that performs well in both domains, the paper characterizes the target error of a source-trained classifier.

Secondly, the case of few shot target samples is addressed by bounding the target error of a model which minimizes a convex combination of the empirical source and target errors. Previous theoretical work only considered minimizing just the source error, just the target error, or weighting instances from the two domains equally - all of which were shown in this paper to be suboptimal in few situations compared to non trivial weighting. he resulting bound is always at least as tight as a bound which considers minimizing only the target error or an equal weighting of source and target errors.

2.1 Notation

In domain adaptation, we have a source domain $\mathcal{D}_s$ and target domain $\mathcal{D}_t$. A domain pair consists of samples $\mathcal{X}$ drawn from either $\mathcal{D}_s$ or $\mathcal{D}_t$ and a corresponding labeling function f_s, f_t for the source and target domains. In this report, we assume that the labeling functions of the source and target domains are the same or close. This is the so called coviariate shift assumption, where $P_s(x) \neq P_t(x)$ but $P_s(y|x) \sim P_t(y|x)$. This is a reasonable assumption since although the data coming from the different domains, the prediction rules will be the same. That is, given a cat image in day time or night time, it is still a cat. In general, it is a very difficult problem to solve in presence of both distribution and label shift. Also, this labeling function can be a real values between 0,1 since we are working in an agnostic PAC setting.

Furthermore, we denote using $h : \mathcal{X} \to \{0, 1\}$. The probability according to the distribution $\mathcal{D}_s$ that a hypothesis h disagrees with a labeling function f is defined as

$$\epsilon_s(h, f) = \mathrm{E}_{x \sim \mathcal{D}_s}[|h(x) - f(x)|] \tag{1}$$

which is also the error (or risk) for the hypothesis. Also, note the shorthand ϵ_s to refer to $\epsilon_s(h, f_s)$.

2.2 A bound relating source and target error

The H-divergence $d_{\mathcal{H}}$ The H-divergence is a classifier induced metric, and one of the key contributions of this paper is the proposal of such a classifier induced metric for cross domain learning. This can be defined as follows. Given a domain $\mathcal{X}$ with $\mathcal{D}$ and $\mathcal{D}'$ distributions over $\mathcal{X}$, the H-divergence, $d_{\mathcal{H}}$, is given by

$$d_{\mathcal{H}} = 2 \sup_{h \in \mathcal{H}} |Pr_D(I(h)) - Pr_{D'}(I(h))| \tag{2}$$

where $I(h)$ is the set of all points where $h(x) = 1$. This intuitively measures the dissimilarity of behaviour of hypothesis class over the distributions, and if there is a hypothesis in class that differentiates the domains, then they have high $d_{\mathcal{H}}$. This has two advantages. Firstly, for hypothesis classes of finite VC-dim, the $d_{\mathcal{H}}$can be computed accurately from finite sample. Secondly, the $d_{\mathcal{H}}$is always less than or equal to the TV (total variation) distance, so it does not inflate the bounds.

The relation between the empirical $d_{\mathcal{H}}$and real $d_{\mathcal{H}}$is given by the following lemma.

Lemma 1. *Let $\mathcal{H}$ be a hypothesis space on $\mathcal{X}$ with VC-dim d. Let U and U' are samples of size m from D and D' respectively, and let $\hat{d}_{\mathcal{H}}(U, U')$ be the empirical $d_{\mathcal{H}}$between them. Then, w.h.p,*

$$d_{\mathcal{H}}(D, D') \leq \hat{d}_{\mathcal{H}}(U, U') + 4\sqrt{\frac{d\log(2m) + \log(\frac{2}{\delta})}{m}} \tag{3}$$

Using this, we can only compute the empirical $d_{\mathcal{H}}$and use it in a bound as the real $d_{\mathcal{H}}$will always be upper bounded by it. This is kind of uniform convergence for $d_{\mathcal{H}}$.

Also, define the ideal joint hypothesis h^* as follows.

$$h^* = argmin_{h\in\mathcal{H}}\epsilon_s(h) + \epsilon_t(h). \tag{4}$$

and the combined error is given by

$$\lambda = \epsilon_s(h^*) + \epsilon_t(h^*) \tag{5}$$

The ideal joint hypothesis explicitly embodies a notion of adaptability. When this hypothesis performs poorly, adaptation using source data is difficult. On the other hand, we will show that if the ideal joint hypothesis performs well, we can measure adaptability of a source-trained classifier by using a divergence metric between the marginal source and target distributions $\mathcal{D}_s$ and $\mathcal{D}_t$.

$d_{\mathcal{H}\Delta\mathcal{H}}$divergence The $d_{\mathcal{H}\Delta\mathcal{H}}$ divergence metric is defined as follows.

$$d_{\mathcal{H}\Delta\mathcal{H}} = 2\sup_{h,h'\in\mathcal{H}} |Pr_{x\in\mathcal{D}_s}[h(x) \neq h'(x)] - Pr_{x\in\mathcal{D}_t}[h(x) \neq h'(x)]| \tag{6}$$

Intuitively, this measures the discrepancy between disagreement of two classifiers between source and target domains. If two classifiers disagree in the same manner in source and target, they are close and vice versa. Now, we provide a bound of the target error in terms of the source error of the hypothesis, the divergence metric and the optimum error.

Theorem 1. *Let $\mathcal{H}$ be a finite-dim hypothesis class. If U_s and U_t are unlabeled samples of size m' each drawn from $\mathcal{D}_s$ and $\mathcal{D}_t$ respectively, then for any $\delta \in (0, 1)$, w.h.p for every $h \in \mathcal{H}$*

$$\epsilon_t(h) \leq \epsilon_s(h) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(U_s, U_t) + 4\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m,}} + \lambda \tag{7}$$

Proof. Proof Sketch.

1. Step 1: We use the triangle inequality of classifiers to arrive at the relation

$$\epsilon_t(h) \leq \epsilon_t(h^*) + \epsilon_s(h, h^*) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t). \tag{8}$$

2. Step 2: We then apply Lemma 1, together with the fact that the VC-dim of delta class is twice as that of original class to arrive at the final result.

□

2.3 Learning bound combining source and target domains

Following from previous sections, we also report the case when βm samples are present from target $\mathcal{D}_t$ and $(1-\beta)m$ samples are present from source $\mathcal{D}_s$. Then, let

$$\epsilon_\alpha(h) = \alpha\epsilon_t(h) + (1-\alpha)\epsilon_s(h) \tag{9}$$

as the α-weighted combined source and target empirical error. Clearly, different values of α would weight the source and target differently. The bound on the target error for a classifier that minimizes the empirical α-weighted error is given by the following theorem.

Theorem 2. *Let H be a hypothesis space of VC dimension d.Let $\mathcal{U}_s$ and $\mathcal{U}_t$ be unlabeled samples of size m' each, drawn from $\mathcal{D}_s$ and $\mathcal{D}_t$ respectively. Let S be a labeled sample of size m generated by drawing βm points from $\mathcal{D}_t$ and (1−β)m points from $\mathcal{D}_s$ and labeling them according to f_S and f_T,respectively. If $\hat{h} \in \mathcal{H}$ is the empirical minimizer of α error ϵ_α and h^* is the target error minimizer,then for any $\delta in(0,1)$, w.h.p we have*

$$\epsilon_t(h) \leq \epsilon_t(h^*) + 4\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{(1-\beta)}} + \sqrt{\frac{2d\log(2m) + 2\log(\frac{8}{\delta})}{m}} + 2(1-\alpha)\left(\frac{1}{2}\hat{d}_{\mathcal{H}}(U_s, U_t) + 4\sqrt{\frac{2d\log(2m') + 2\log(\frac{8}{\delta})}{m'}}\right)$$

Proof. The proof is not entirely clear to me. However, the broad steps involved are bounding the α error in terms of target empirical error, followed by uniform convergence of α error and applying the classical learning theory bound. □

3 Hardness of Domain Adaptation and the Utility of Unlabeled Target Samples

This part of the report is prepared by reading the paper *On the hardness of domain adaptation and the utility of unlabeled target samples* [3]. While there have been previous works for bounding the sample complexity required for Domain Adaptation (DA) in the restricted setting of proper DA learning, this paper assumes the general DA learning framework, when the learner is allowed to output arbitrary predictors.

3.1 Definitions and Assumptions

We formalize the notion of Domain Adaptation learnability. Formally, a Domain Adaptation learner is a function that given a set of labeled sample drawn from source distribution and a set of unlabeled sample drawn from target distribution, will produce a label predictor or hypothesis.

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

Here $\mathcal{X}$ is the domain over which two different probability distributions exist - the source distribution P_S and the target distribution P_T. We work with the covariant shift assumption which means the same labeling function $l : \mathcal{X} \rightarrow \{0,1\}$ is used to label examples coming from both the distributions.

Definition 1. ***(DA learnability)*** *Let $\mathcal{W}$ be a class triples (P_S, P_T, l) of source and target distributions over some domain $\mathcal{X}$ and a labeling function, and let $\mathcal{A}$ be a DA learner. We say that $\mathcal{A}$ ()-solves DA for the class $\mathcal{W}$, if, for all triples $(P_S, P_T, l) \in \mathcal{W}$, when given a sample S of size m, generated i.i.d. by P_S and labeled by l, and an unlabeled sample T of size n, generated i.i.d. by P_T, with probability at least $1 - \delta$ (over the choice of the samples S and T) $\mathcal{A}$ outputs a function h with $Err_T^l(h) \leq \epsilon$.*

Definition 2. ***(Weight Ratio)*** *Let $\mathcal{B} \subseteq 2^{\mathcal{X}}$ be a collection 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 target distribution w.r.t. $\mathcal{B}$ 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-measurable and P_T-measurable by $C(P_S, P_T)$.

In the case of discrete distributions this is the point-wise weight ratio $C(P_S, P_T) = C_{\{\{x\}:x\in\mathcal{X}\}}(P_S, P_T)$. Note that for every $\mathcal{B} \subseteq 2^{\mathcal{X}}, C(P_S, P_T) \leq C_{\mathcal{B}}(P_S, P_T)$.

Definition 3. ***(ϵ-net)*** *Let $\mathcal{X}$ be some domain, $\mathcal{W} \subseteq 2^{\mathcal{X}}$ a collection of subsets of $\mathcal{X}$ and P a distribution over $\mathcal{X}$. An ϵ-net for $\mathcal{W}$ with respect to P is a subset $N \subseteq \mathcal{X}$ that intersects every member of $\mathcal{W}$ that has P-weight at least ϵ.*

Lemma 2. *Let $\mathcal{X}$ be some domain, $\mathcal{W} \subseteq 2^{\mathcal{X}}$ a collection of subsets of $\mathcal{X}$, P_S and P_T a source and target distribution over $\mathcal{X}$ with $\mathcal{C} := \mathcal{C}_{\mathcal{W}}(P_S, P_T) \geq 0$. Then every $(\mathcal{C}\epsilon)$-net for $\mathcal{W}$ with respect to P_S is an ϵ-net for $\mathcal{W}$ with respect to P_T.*

Definition 4. ***(Margin Assumption)*** *Let $\mathcal{X} \subseteq \mathbb{R}^d$, P a distribution over $\mathcal{X}$ and $h : \mathcal{X} \to \{0, 1\}$ a classifier. We say that h is a γ-margin classifier with respect to P if for all $x \in \mathcal{X}$ whenever $P(B_\gamma(x)) > 0$, where $B_\gamma(x)$ is a ball of radius γ around x, then $h(y) = h(z)$ holds for all $y, z \in B_\gamma(x)$.*

We say that a class H realizes the distribution with margin γ if the optimal zero-error classifier is a γ-margin classifier.

3.2 Lower bounds for Realizable Domain Adaptation

The lower bound in this section shows that no small amount of labeled source and unlabeled target data suffices for DA under covariate shift. Even in the case where the learner knows that the target is realizable by the class $H_{1,0}$, that contains only the all-1 or the all-0 labeling functions, the size of source sample plus size of target sample needs to be as large as $\sqrt{|\mathcal{X}|}$ which can be infinite (i.e., the "hardness" result).

Theorem 3. *For every finite domain $\mathcal{X}$ for every ϵ and δ with $\epsilon + \delta < 1/2$, no algorithm can $(\epsilon, \delta, s, t)-$solve the DA problem for the class $\mathcal{W}$ of triples (P_S, P_T, l) with $C(P_S, P_T) \geq 1/2$, $d_{\mathcal{H}_{1,0}\Delta\mathcal{H}_{1,0}} = 0$ and $opt_T^l(H_{1,0}) = 0$ if $s + t < \sqrt{(1 - 2(\epsilon + \delta)|\mathcal{X}|} - 2$.*

Proof of Theorem.

We obtain the bound by reducing the following problem to DA:

The Left/Right Problem. We consider the problem of distinguishing two distributions from finite samples. This problem was introduced in [5].

Input: Three finite samples L, R, and M of points from some domain $\mathcal{X}$.

Output: Assuming that L is an i.i.d. sample from some distribution P over $\mathcal{X}$, R is an i.i.d. sample from some distribution Q over $\mathcal{X}$, and M is an i.i.d. sample generated by one of these two probability distributions, was M generated by P or by Q?

We first derive a lower bound on the sample size needed to solve Left/Right problem. Then we reduce this problem to DA, thereby obtaining a lower bound on the sample size needed to solve DA.

Lower bound for the Left/Right problem: We say that a (randomized) algorithm $(\delta, l, r, m)-$solves the left/right problem if, given samples L, R and M of sizes l, r, m respectively, it gives the correct answer with probability at least $1 - \delta$. We show that for any sample sizes l, r, and m and for any $\gamma < 1/2$, there exists a finite domain $\mathcal{X} = \{1, \ldots, n\}$ and a finite class $\mathcal{W}_n^{uni}$ of triples of distributions over $\mathcal{X}$ such that no algorithm can $(\gamma, l, r, m)-$solve the Left/Right problem for this class.

We construct our class $\mathcal{W}_n^{uni} = \{(U_A, U_B, U_C) : A \cup B = \{1, \ldots, n\}, A \cap B = \emptyset, |A| = |B|, C = A \text{ or } C = B\}$, where the distribution generating L (i.e., U_A) and the distribution generating R (i.e., U_B) are uniform over half of the points in $\mathcal{X}$, but their supports are disjoint. U_Y denotes uniform distribution over set Y.

Lemma 3. *For any given sample sizes l for L , r for R, and m for M and any $0 < \gamma < 1/2$, if $k = max l, r + m$, then for $n > (k + 1)^2/(1 - 2\gamma)$ no algorithm has probability of success greater than $1 - \gamma$ over the class $\mathcal{W}_n^{uni}$.*

We refer the reader to [5] for proof of the above lemma. Now that we have a bound on the size of sample needed to solve Left/Right problem, we turn our attention to the reduction from this problem to DA.

Reducing the Left/Right problem to Domain Adaptation learning: We define a class of DA problems that corresponds to the class of triples $\mathcal{W}_n^{uni}$, for which we have proven a lower bound on the sample sizes needed for solving the Left/Right problem.

For a number n, let $\mathcal{W}_n^{DA}$ be the class of triples (P_S, P_T, l), where P_S is uniform over some finite set $\mathcal{X}$ of size n, P_T is uniform over some subset U of $\mathcal{X}$ of size $n/2$ and l assigns point in U to 1 and point in $\mathcal{X}\backslash U$ to 0 or vice versa. Observe that $C(P_S, P_T) = 1/2$ and $d_{\mathcal{H}_{1,0}\Delta\mathcal{H}_{1,0}}(P_S, P_T) = 0$ for all (P_S, P_T, l) in $\mathcal{W}_n^{DA}$. Further, for the class $\mathcal{H}_{1,0}$ that contains only the constant 1 function and the constant 0 function, we have $opt_T^l(\mathcal{H}_{1,0})$ for all elements of $\mathcal{W}_n^{DA}$.

Lemma 4. *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 (ϵ, δ, s, t)-solve DA for a class $\mathcal{W}$ that includes $\mathcal{W}_n^{DA}$, we can construct an algorithm that $(\epsilon + \delta, s, s, t + 1)$-solves the Left/Right problem on $\mathcal{W}_n^{uni}$.*

Proof. We can assume we are given samples $L = \{l_1, \ldots, l_s\}$ and $R = \{r_1, \ldots, r_s\}$ of size s and a sample M of size $t+1$ for the Left/Right problem coming from a triple (U_A, U_B, U_C) of distributions in $\mathcal{W}_n^{uni}$. We construct an input to DA by setting the unlabeled target sample $T = M\backslash\{p\}$ where p is a point from M chosen uniformly at random and construct the labeled source sample S as follows: We select s elements from $L \times \{0\} \cup R \times \{1\}$ by successively flipping an unbiased coin and depending on the output choosing the next element from $L \times \{0\}$ or $R \times \{1\}$.

These sets can be considered as input to DA generated from a source distribution $P_S = U_{(A\cup B)}$ that is uniform over $A \cup B$. The target distribution P_T of this instance has marginal equal to U_A or to U_B (depending on whether M was a sample from U_A or from U_B). The labeling function of this DA instance is $l(x) = 0$ if $x \in A$ and $l(x) = 1$ if $x \in B$. Observe that $C(P_S, P_T) = 1/2$, $opt_T^l(\mathcal{H}_{1,0}) = 0$, and $(P_S, P_T, l) \in \mathcal{W}_n^{DA}$. Assume that h is the output of $\mathcal{A}$ on input S and T. The algorithm for the Left/Right problem then outputs U_A if $h(p) = 0$ and U_B if $h(p) = 1$ and the claim follows.

Since $n = |\mathcal{X}|$, we have shown from the above two lemmas that no algorithm can solve the DA problem for $\mathcal{W}_n^{DA}$, even under the assumption of realizability by $\mathcal{H}_{1,0}$, if the sample sizes of the source and target sample satisfy $|S| + |T| < \sqrt{(1 - 2(\epsilon + \delta))|\mathcal{X}|} - 2$.

3.3 Utility of Unlabeled Data

In many scenarios unlabeled data is abundantly available while the labeled data is hard to obtain. Here we present a Domain Adaptation algorithm that also utilizes the unlabeled target data, known as an Adaptive Domain Adaptation algorithm, and show separate bounds on the size of labeled data and the size of the unlabeled data to achieve DA learnability.

We start by presenting a Domain Adaptation algorithm for the case where the labeling function satisfies the Lipschitz property and the target is realizable with a margin. Later, we present a modified algorithm for the case of finite domain with arbitrary labeling functions and hypothesis classes of finite VC-dimension (under target-realizability assumption).

3.3.1 Realizability with a Margin

We propose the following adaptive Domain Adaptation Algorithm:

Algorithm $\mathcal{A}$

Input An i.i.d. sample S from P_S labeled by l, an unlabeled i.i.d. sample T from P_T and a margin parameter γ.

Step 1 Partition the domain $[0, 1]^d$ into a collection $\mathcal{B}$ of boxes (axis-aligned rectangles) with side length $(\gamma/\sqrt{d})$.

Step 2 Obtain sample S' by removing every point in S, which is sitting in a box that is not hit by T.

Step 3 Output an ERM classifier from H for the sample S'.

We now present the theorem that gives bounds on the sizes of the labeled sample and unlabeled sample separately that are required for this adaptive DA algorithm $\mathcal{A}$ to succeed.

Theorem 4. *Let $\mathcal{X} = [0, 1]^d, \gamma > 0$ a margin parameter, H be a hypothesis class of finite VC dimension and $\mathcal{W}$ be a class of triples of source distribution, target distribution and labeling function with*

– *$C_{\mathcal{I}}(P_S, P_T) > 0$ for the class $\mathcal{I} = (\mathcal{H}\Delta\mathcal{H}) \cap \mathcal{B}$ where $\mathcal{B}$ is a partition of $[0,1]^d$ into boxes of sidelength $\gamma/\sqrt{d}$*

– *P_T is realizable by H with margin γ*

– *the labeling function l is a γ-margin classifier with respect to P_T.*

Then there is a constant $c > 1$ such that, for all $\epsilon > 0, \delta > 0$ and all $P_S, P_T, l) \in \mathcal{W}$, when given i.i.d. sample S from P_S, labeled by l of size

$$|S| \geq c\left(\frac{VC(H) + log(1/\delta)}{C_{\mathcal{I}}(P_S, P_T)(1-\epsilon)\epsilon} log(\frac{VC(H)}{C_{\mathcal{I}}(P_S, P_T)(1-\epsilon)\epsilon})\right)$$

and an i.i.d. sample T from P_T of size $|T| \geq \frac{2(\sqrt{d}/\gamma)^d ln(3(\sqrt{d}/\gamma)^d/\delta)}{\epsilon}$ then algorithm $\mathcal{A}$ outputs a classifier h with $Err_T^l(h) \leq \epsilon$ with probability at least $1-\delta$.

Proof of Theorem. Let $\epsilon > 0$, $\delta > 0$ be given and set $C = C_{\mathcal{I}}(P_S, P_T)$. We set $\epsilon' = \epsilon/2$ and $\delta' = \delta/3$ divide the space $\mathcal{X}$ up into *heavy* and *light* boxes from $\mathcal{B}$, by defining a box to be light if $P_T(b) \leq \epsilon'/|\mathcal{B}|$ and heavy otherwise. Let $\mathcal{X}^l$ denote the union of light boxes, and $\mathcal{X}^h$ the union of heavy boxes. Let P_S^h and P_T^h denote restrictions of source and target distributions to $\mathcal{X}^h$. $P_S^h = P_S(U)/P_S(\mathcal{X}^h)$ for all $U \subseteq \mathcal{X}^h$ and $P_S^h = 0$ for all $U \nsubseteq \mathcal{X}^h$. Similarly, for P_T^h. Note that $|B| = (\sqrt{d}/\gamma)^d$, we have $P_T(\mathcal{X}^h) \geq 1 - \epsilon'$ and thus, $P_S(\mathcal{X}^h) \geq C(1-\epsilon')$.

We claim the following and then use these claims to complete the proof.

Claim 1. With probability at least $1-\delta'$ an i.i.d. P_T-sample T of size as stated in the Theorem hits every heavy box.

Claim 2. With probability at least $1-2\delta'$ the intersection of S and $\mathcal{X}^h$, where S is an i.i.d. P_S-sample S of size as stated in the Theorem is an ϵ'-net for $\mathcal{H}\Delta\mathcal{H}$ with respect to P_T^h

Proof of Claim 1: Let b be a heavy box, thus $P_T(b) \geq \epsilon'/|\mathcal{B}|$. When drawing i.i.d. sample T from P_T the probability of not hitting b is at most $(1 - (\epsilon'/|\mathcal{B}|))^{|T|}$. Now, the union bound implies that the probability that there is a box in $\mathcal{B}^h$ that does not get hit by the sample T is at most

$$|\mathcal{B}^h|(1 - (\epsilon'/|\mathcal{B}|))^{|T|} \leq |\mathcal{B}|(1 - (\epsilon'/|\mathcal{B}|))^{|T|} \leq e^{-\epsilon'|T|/|\mathcal{B}|}.$$

Thus, if $|T| \geq \frac{|\mathcal{B}|ln(|\mathcal{B}|/\delta')}{\epsilon'} = \frac{2(\sqrt{d}/\gamma)^d ln(3(\sqrt{d}/\gamma)^d/\delta)}{\epsilon}$ the sample T hits every heavy box with probability at least $1-\delta'$.

Proof of Claim 2: Let $S^h := S \cap \mathcal{X}^h$. Note that, as S is an i.i.d. P_S sample, we can consider S^h to be an i.i.d. P_S^h sample. We have the following bound on the weight ratio between P_S^h and P_T^h:

$$C_{\mathcal{I}}(P_S^h, P_T^h) = inf_{p \in \mathcal{I}, P_T^h(p) > 0} \frac{P_S^h(p)}{P_T^h(p)} = inf_{p \in \mathcal{I}, P_T^h(p) > 0} \frac{P_S(p)P_T(\mathcal{X}^h)}{P_T(p)P_S(\mathcal{X}^h)} \geq C\frac{P_T(\mathcal{X}^h)}{P_S(\mathcal{X}^h)} \geq C(1-\epsilon')$$

Note that every element in $\mathcal{H}\Delta\mathcal{H}$ can be partitioned into elements from $\mathcal{I}$, therefore we obtain the same bound on the weight ratio for the symmetric differences of H: $C_{\mathcal{H}\Delta\mathcal{H}}(P_S^h, P_T^h) \geq C(1-\epsilon')$.

It is well known that there is a constant $c > 1$ such that, conditioned on S^h having size at least $M := c\left(\frac{VC(\mathcal{H}\Delta\mathcal{H}) + log(1/\delta')}{C(1-\epsilon')\epsilon'} log(\frac{VC(\mathcal{H}\Delta\mathcal{H})}{C(1-\epsilon')\epsilon'})\right)$, with probability $\geq 1-\delta'$ it is a $C(1-\epsilon')\epsilon'$-net with respect to P_S^h and thus ϵ'-net with respect to P_T^h. (Follows from Lemma 2 and property of ϵ-nets)

It remains to be shown that with probability $\geq 1-\delta'$ we have $|S^h| \geq M$. As we have $P_S(\mathcal{X}^h) \geq C(1-\epsilon')$, we can view the sampling of the points of S and checking whether they hit $\mathcal{X}^h$ as a Bernoulli variable with mean $\mu = P_S(\mathcal{X}^h) \geq C(1-\epsilon')$. Applying Hoeffding's inquality, we have for all $t > 0$, $Pr(\mu|S| - |S^h| \geq t|S|) \leq e^{-2t^2|S|}$. If we set $C' = C(1-\epsilon')$ and $t = C'/2$, and take $|S| \geq \frac{2M}{C'}$, we obtain $Pr(|S^h| < M) \leq Pr(\mu|S| - |S^h| \geq \frac{C'}{2}|S|) \leq e^{-\frac{C'^2|S|}{2}}$.

It can be shown that $VC(\mathcal{H}\Delta\mathcal{H}) \leq 2VC(H) + 1$ [2]. Now $|S| \geq \frac{2M}{C'} > \frac{VC(\mathcal{H}\Delta\mathcal{H}) + log(1/\delta')}{C^2(1-\epsilon')^2\epsilon'}$ implies that $e^{-\frac{C'^2|S|}{2}} \leq \delta'$ and this proves that S^h is an ϵ'-net of $\mathcal{H}\Delta\mathcal{H}$ with probability at least $(1-\delta')^2 \geq 1 - 2\delta'$.

Proof of Theorem (Cont'd). We have from Claim 1, T hits every heavy box with probability at least $1-\delta'$. Thus, with high probability $S^h = S \cup \mathcal{X}^h \subseteq S'$ where S' is as defined in the algorithm $\mathcal{A}$ the intersection of S with boxes that are hit by T. From Claim 2, S^h is an ϵ'-net for $\mathcal{H}\Delta\mathcal{H}$ with respect to P_T^h with probability at least $1-2\delta'$. Therefore, with probability at least $1-3\delta' = 1-\delta$, the set S' is an ϵ'-net for $\mathcal{H}\Delta\mathcal{H}$ with respect to P_T^h. We also know that an ϵ'-net for $\mathcal{H}\Delta\mathcal{H}$ with respect to P_T^h is an ϵ-net with respect to P_T.

Finally, we need to show that S' being an ϵ-net for the set $\mathcal{H}\Delta\mathcal{H}$ with respect to P_T suffices for the ERM-classifier from the target class to have target error at most ϵ. Let $h_T^* \in H$ denote the γ-margin classifier of zero target error. Note that every box in $\mathcal{B}$ is labeled homogeneously with label 1 or label 0 by the labeling function l as l is a γ-margin classifier as well. Let $s \in S'$ be a sample point and $b_s \in \mathcal{B}$ be the box that contains s. As h_T^* is a γ-margin classifier and $P_T(b_s) > 0$ (b_s was hit by T by the definition of S'), b_s is labeled homogeneously by h_T^* as well and as h_T^* has zero target error this label has to correspond to the labeling by l. Thus, $h_T^* = l(s)$ for all $s \in S'$, which means that the empirical error with respect to S' of h_T^* is zero.

Now consider a classifier h_ϵ with $Err_T^l(h_\epsilon) \geq \epsilon$. Let $s \in S'$ be a sample point in $h_T^* \Delta h_\epsilon$ (which exists as S' is an ϵ-net). As $s \in h_T^* \Delta h_\epsilon$ we have $h_\epsilon(s) \neq h_T^*(s) = l(s)$ thus h_ϵ has an empirical error larger than zero, which implies that no classifier of error larger than ϵ can be chosen by ERM on input S'. This completes the proof.

3.3.2 Finite Domain

The procedure $\mathcal{A}$ from the previous section can be modified to work on any finite domain with arbitrary labeling functions and hypothesis classes of finite VC dimension (under the target-realizability assumption). Modifying the previous algorithm to apply to this case and using the modified algorithm $\mathcal{A}$, it is observed that the bound on size of the source sample S needed to guarantee success does not change, however the bound on size of the target sample T now depends on the size of the domain.

Theorem 5. *Let $\mathcal{X}$ be some domain, H be a hypothesis class of finite VC dimension and $\mathcal{W} = \{(P_S, P_T, l) | C(P_S, P_T) > 0, opt_T^l(H) = 0\}$ be a class of pairs of source and target distributions with bounded weight ratio where the target is realizable by H. Then there is a constant $c > 1$ such that, for all $\epsilon > 0$, $\delta > 0$,and all $(P_S, P_T, l) \in \mathcal{W}$, when given an i.i.d. sample S from P_S, labeled by l of size $|S| \geq c\left(\frac{VC(H)+log(1/\delta)}{C(P_S,P_T)^2(1-\epsilon)^2\epsilon} log(\frac{VC(H)}{C(P_S,P_T)^2(1-\epsilon)^2\epsilon})\right)$ and an i.i.d. sample T from P_T of size $|T| \geq \frac{2|\mathcal{X}|ln(3|\mathcal{X}|/\delta)}{\epsilon}$ then algorithm A outputs a classifier h with $Err_T^l(h) \leq \epsilon$ with probability at least $1-\delta$.*

4 Experiments

In this section, we first discuss the experimental result given in the paper, and next discuss our own implementation and experiments of computing $d_{\mathcal{H}\Delta\mathcal{H}}$ divergence.

Experiment Setting The authors use a sentiment classification task to prove the empirical significance of their work. They use four domains, Books, DVD, Electronics and Kitchen with varying measures of domain divergence. In Figure 1, in each graph, the y-axis represents the error and x-axis represents the $\alpha-$ value. Plots on on the top row show the value given by the approximation to the bound, and plots on the bottom row show the empirical test set error.

Firstly, the leftmost plot depicts the variation of source datasets for fixed value of m_s and m_t, the source and target samples respectively. It is clear that as α ratio of $\frac{m_t}{m_t+m_s}$ is ideal for all the datasets irrespective of the divergence value. The middle plot then varies the target samples m_T for a fixed dataset *pair*. Again, the empirical and theoretical errors also match and an optimal value of α is found out to be equal to β, rather than using all source, all target or equal weighting of empirical errors.

We also setup our experiments to examine the influence of *adaptation algorithms* on the *domain divergence*. Specifically, we chose two adaptation algorithms. One is called MMD [7] for maximum mean discrepancy, which minimizes the means of feature spaces from two datasets in an auxiliary feature space. Secondly, we use DANN [4] which uses an adversarial GAN based objective to explicitly minimize the domain divergence. The code and models have been used from the respective

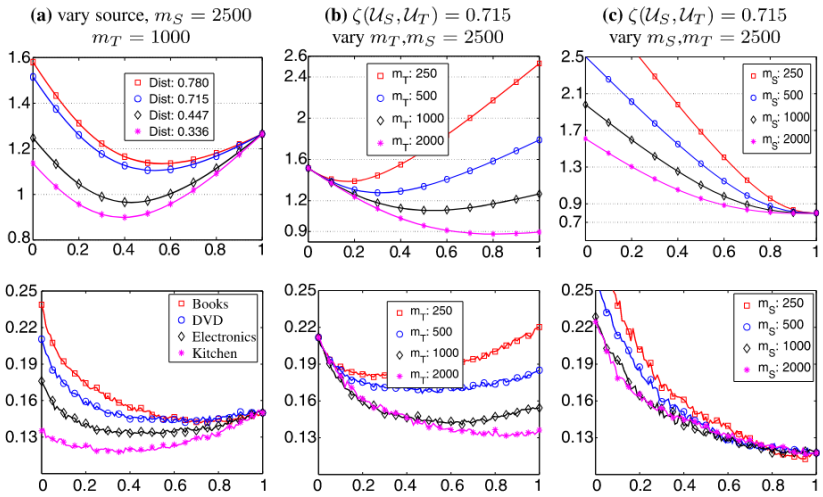

Figure 1: Comparing the bound with the actual empirical value. First row is the analytical plot, and second row is the empirical plots on various domains. The columns correspond to different domains, source and target data. The x-axis in each plot corresponds to the α value.

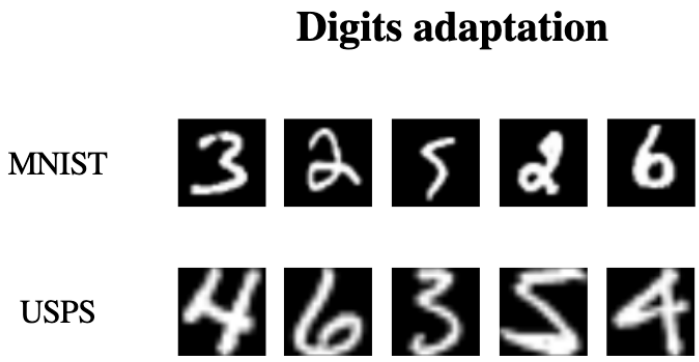

Figure 2: Examples of images from MNIST and USPS datasets. Each dataset consists of digits from 0-9.

papers. As for datasets, we use MNIST and USPS because they are small datasets. Few samples from these are given in Figure 2 We use a LeNet backbone following the original paper on DANN [4].

We plot our result in Figure 3. We plot the empirical approximation of $\mathcal{H}\Delta\mathcal{H}$ divergence, which is obtained by computing the error ϵ of a classifier trained to discriminate between the domains. Then, the $d_{\mathcal{H}\Delta\mathcal{H}}$divergence is given by $\mathcal{H}\Delta\mathcal{H} = 2(1 - 2\epsilon)$. As we can see, the adaptation algorithms have lesser values of $d_{\mathcal{H}\Delta\mathcal{H}}$compared to directly classifying on source domains. This indicates the effectiveness of the concept of $d_{\mathcal{H}\Delta\mathcal{H}}$divergence.

References

[1] 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):151–175, 2010.

[2] Shai Ben-David and Ami Litman. Combinatorial variability of vapnik-chervonenkis classes with applications to sample compression schemes. *Discrete Applied Mathematics*, 86(1):3–25, 1998.

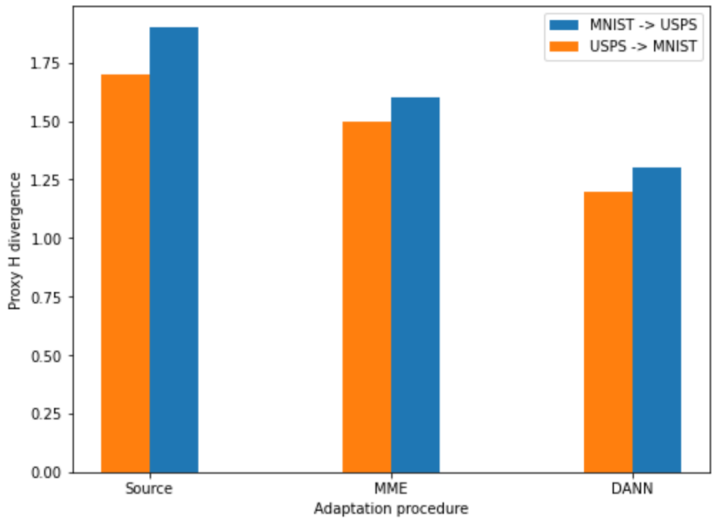

Figure 3: The plot comparing the proxy $\mathcal{H}\Delta\mathcal{H}$ divergence for various adaptation methods. The source classifier is without using any adaptation, while the second and third use two forms of adaptation on MNIST and USPS datasets.

[3] Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *International Conference on Algorithmic Learning Theory*, pages 139–153. Springer, 2012.

[4] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[5] B.G. Kelly, T. Tularak, A.B. Wagner, and P. Viswanath. Universal hypothesis testing in the learning-limited regime. *IEEE International Symposium on Information Theory (ISIT)*, 2010.

[6] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *arXiv preprint arXiv:1705.10667*, 2017.

[7] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 2200–2207, 2013.

[8] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.