On Learning Invariant Representations for Domain Adaptation

Han Zhao [1] **Remi Tachet des Combes** [2] **Kun Zhang** [1] **Geoffrey J. Gordon** [1 2]

Abstract

Due to the ability of deep neural nets to learn rich representations, recent advances in unsupervised domain adaptation have focused on learning domain-invariant features that achieve a small error on the source domain. The hope is that the learnt representation, together with the hypothesis learnt from the source domain, can generalize to the target domain. In this paper, we first construct a simple counterexample showing that, contrary to common belief, the above conditions are not sufficient to guarantee successful domain adaptation. In particular, the counterexample exhibits *conditional shift*: the class-conditional distributions of input features change between source and target domains. To give a sufficient condition for domain adaptation, we propose a natural and interpretable generalization upper bound that explicitly takes into account the aforementioned shift. Moreover, we shed new light on the problem by proving an information-theoretic lower bound on the joint error of *any* domain adaptation method that attempts to learn invariant representations. Our result characterizes a fundamental tradeoff between learning invariant representations and achieving small joint error on both domains when the marginal label distributions differ from source to target. Finally, we conduct experiments on real-world datasets that corroborate our theoretical findings. We believe these insights are helpful in guiding the future design of domain adaptation and representation learning algorithms.

1. Introduction

The recent successes of supervised deep learning methods have been partially attributed to rich datasets and increasing computational power. However, in many critical applications, e.g., self-driving cars or personal healthcare, it is often prohibitively expensive and time-consuming to collect large-scale supervised training data. Unsupervised domain adaptation (DA) focuses on such limitations by trying to transfer knowledge from a labeled source domain to an unlabeled target domain, and a large body of work tries to achieve this by exploring domain-invariant structures and representations to bridge the gap. Theoretical results (Ben-David et al., 2010; Mansour et al., 2009a; Mansour & Schain, 2012) and algorithms (Glorot et al., 2011; Becker et al., 2013; Ajakan et al., 2014; Adel et al., 2017; Pei et al., 2018) under this setting are abundant.

Due to the ability of deep neural nets to learn rich feature representations, recent advances in domain adaptation have focused on using these networks to learn *invariant representations*, i.e., intermediate features whose distribution is the same in source and target domains, while at the same time achieving small error on the source domain. The hope is that the learnt intermediate representation, together with the hypothesis learnt using labeled data from the source domain, can generalize to the target domain. Nevertheless, from a theoretical standpoint, it is not at all clear whether aligned representations and small source error are sufficient to guarantee good generalization on the target domain. In fact, despite being successfully applied in various applications (Zhang et al., 2017; Hoffman et al., 2017), it has also been reported that such methods fail to generalize in certain closely related source/target pairs, e.g., digit classification from MNIST to SVHN (Ganin et al., 2016).

Given the wide application of domain adaptation methods based on learning invariant representations, we attempt in this paper to answer the following important and intriguing question:

> *Is finding invariant representations while at the same time achieving a small source error sufficient to guarantee a small target error? If not, under what conditions is it?*

Contrary to common belief, we give a negative answer to the above question by constructing a simple example showing that these two conditions are not sufficient to guarantee target generalization, even in the case of perfectly aligned representations between the source and target domains. In

[1]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA [2]Microsoft Research, Montreal, Canada. Correspondence to: Han Zhao <han.zhao@cs.cmu.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

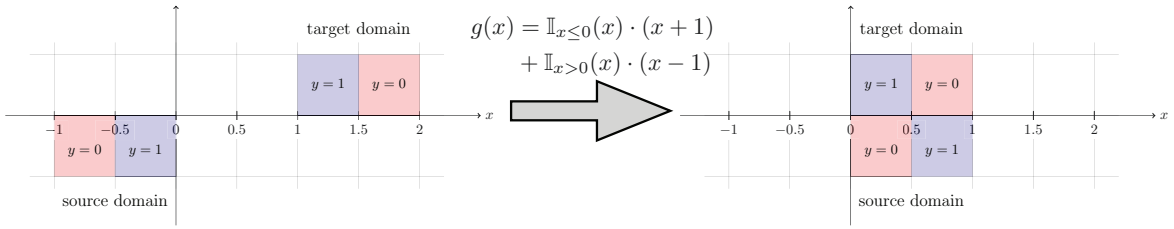


Figure 1. A counterexample where invariant representations lead to large joint error on source and target domains. Before transformation of $g(\cdot)$, $h^*(x) = 1$ iff $x \in (-1/2, 3/2)$ achieves perfect classification on both domains. After transformation, source and target distributions are perfectly aligned, but no hypothesis can achieve a small joint error.

fact, our example shows that the objective of learning invariant representations while minimizing the source error can actually be hurtful, in the sense that the better the objective, the larger the target error. At a colloquial level, this happens because learning invariant representations can break the originally favorable underlying problem structure, i.e., close labeling functions and conditional distributions. To understand when such methods work, we propose a generalization upper bound as a sufficient condition that explicitly takes into account the conditional shift between source and target domains. The proposed upper bound admits a natural interpretation and decomposition in domain adaptation; we show that it is tighter than existing results in certain cases.

Simultaneously, to understand what the necessary conditions for representation based approaches to work are, we prove an information-theoretic lower bound on the joint error of both domains for *any* algorithm based on learning invariant representations. Our result complements the above upper bound and also extends the constructed example to more general settings. The lower bound sheds new light on this problem by characterizing a fundamental tradeoff between learning invariant representations and achieving small joint error on both domains when the marginal label distributions differ from source to target. Our lower bound directly implies that minimizing source error while achieving invariant representation will only increase the target error. We conduct experiments on real-world datasets that corroborate this theoretical implication. Together with the generalization upper bound, our results suggest that adaptation should be designed to align the label distribution as well when learning an invariant representation (c.f. Sec. 4.3). We believe these insights will be helpful to guide the future design of domain adaptation and representation learning algorithms.

2. Preliminary

We first introduce the notations used throughout this paper and review a theoretical model for domain adaptation (DA) (Kifer et al., 2004; Ben-David et al., 2007; Blitzer et al., 2008; Ben-David et al., 2010).

Notations We use $\mathcal{X}$ and $\mathcal{Y}$ to denote the input and output space, respectively. Similarly, $\mathcal{Z}$ stands for the representation space induced from $\mathcal{X}$ by a feature transformation $g : \mathcal{X} \mapsto \mathcal{Z}$. Accordingly, we use X, Y, Z to denote the random variables which take values in $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, respectively. In this work, *domain* corresponds to a distribution $\mathcal{D}$ on the input space $\mathcal{X}$ and a labeling function $f : \mathcal{X} \to [0, 1]$. In the domain adaptation setting, we use $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$ to denote the source and target domains, respectively. A *hypothesis* is a function $h : \mathcal{X} \to \{0, 1\}$. The *error* of a hypothesis h w.r.t. the labeling function f under distribution $\mathcal{D}_S$ is defined as: $\varepsilon_S(h, f) := \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_S}[|h(\mathbf{x}) - f(\mathbf{x})|]$. When f and h are binary classification functions, this definition reduces to the probability that h disagrees with f under $\mathcal{D}_S$: $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_S}[|h(\mathbf{x}) - f(\mathbf{x})|] = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_S}[\mathbb{I}(f(\mathbf{x}) \neq h(\mathbf{x}))] = \Pr_{\mathbf{x}\sim\mathcal{D}_S}(f(\mathbf{x}) \neq h(\mathbf{x}))$. In this work, we focus on the deterministic setting where the output $Y = f(X)$ is given by a deterministic labeling function f defined on the corresponding domain. For two functions g and h with compatible domains and ranges, we use $h \circ g$ to denote the function composition $h(g(\cdot))$. Other notations will be introduced in the context when necessary.

2.1. Problem Setup

We consider the unsupervised domain adaptation problem where the learning algorithm has access to a set of n labeled points $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ sampled i.i.d. from the source domain and a set of unlabeled points $\{\mathbf{x}_j\}_{j=1}^m \in \mathcal{X}^m$ sampled i.i.d. from the target domain. At a colloquial level, the goal of an unsupervised domain adaptation algorithm is to generalize well on the target domain by learning from labeled samples from the source domain as well as unlabeled samples from the target domain. Formally, let the *risk* of hypothesis h be the error of h w.r.t. the true labeling function under domain $\mathcal{D}_S$, i.e., $\varepsilon_S(h) := \varepsilon_S(h, f_S)$. As commonly used in computational learning theory, we denote by $\widehat{\varepsilon}_S(h)$ the empirical risk of h on the source domain. Similarly, we use $\varepsilon_T(h)$ and $\widehat{\varepsilon}_T(h)$ to mean the true risk and the empirical risk on the target domain. The problem of domain adaptation considered in this work can be stated as: under what conditions and by what algorithms can we

guarantee that a small training error $\widehat{\varepsilon}_S(h)$ implies a small test error $\varepsilon_T(h)$? Clearly, this goal is not always possible if the source and target domains are far away from each other.

2.2. A Theoretical Model for Domain Adaptation

To measure the similarity between two domains, it is crucial to define a discrepancy measure between them. To this end, Ben-David et al. (2010) proposed the $\mathcal{H}$-divergence to measure the distance between two distributions:

Definition 2.1 ($\mathcal{H}$-divergence)**.** Let $\mathcal{H}$ be a hypothesis class on input space $\mathcal{X}$, and $\mathcal{A}_\mathcal{H}$ be the collection of subsets of $\mathcal{X}$ that are the support of some hypothesis in $\mathcal{H}$, i.e., $\mathcal{A}_\mathcal{H} := \{h^{-1}(1) \mid h \in \mathcal{H}\}$. The distance between two distributions $\mathcal{D}$ and $\mathcal{D}'$ based on $\mathcal{H}$ is: $d_\mathcal{H}(\mathcal{D}, \mathcal{D}') := \sup_{A \in \mathcal{A}_\mathcal{H}} |\Pr_\mathcal{D}(A) - \Pr_{\mathcal{D}'}(A)|$. [1]

$\mathcal{H}$-divergence is particularly favorable in the analysis of domain adaptation with binary classification problems, and it had also been generalized to the *discrepancy distance* (Cortes et al., 2008; Mansour et al., 2009a;b; Cortes & Mohri, 2014) for general loss functions, including the one for regression problems. Both $\mathcal{H}$-divergence and the discrepancy distance can be estimated using finite unlabeled samples from both domains when $\mathcal{H}$ has a finite VC-dimension.

One flexibility of the $\mathcal{H}$-divergence is that its power on measuring the distance between two distributions can be controlled by the richness of the hypothesis class $\mathcal{H}$. To see this, first consider the situation where $\mathcal{H}$ is very restrictive so that it only contains the constant functions $h \equiv 0$ and $h \equiv 1$. In this case, it can be readily verified by the definition that $d_\mathcal{H}(\mathcal{D}, \mathcal{D}') = 0, \ \forall \ \mathcal{D}, \mathcal{D}'$. On the other extreme, if $\mathcal{H}$ contains all the measurable binary functions, then $d_\mathcal{H}(\mathcal{D}, \mathcal{D}') = 0$ iff $\mathcal{D}(\cdot) = \mathcal{D}'(\cdot)$ almost surely. In this case the $\mathcal{H}$-divergence reduces to the total variation, or equivalently the L_1 distance, between the two distributions.

Given a hypothesis class $\mathcal{H}$, we define its symmetric difference w.r.t. itself as: $\mathcal{H}\Delta\mathcal{H} = \{h(\mathbf{x}) \oplus h'(\mathbf{x}) \mid h, h' \in \mathcal{H}\}$, where $\oplus$ is the `xor` operation. Let h^* be the optimal hypothesis that achieves the minimum joint risk on both the source and target domains: $h^* := \arg\min_{h \in \mathcal{H}} \varepsilon_S(h) + \varepsilon_T(h)$, and let λ^* denote the joint risk of the optimal hypothesis h^*: $\lambda^* := \varepsilon_S(h^*) + \varepsilon_T(h^*)$. Ben-David et al. (2007) proved the following generalization bound on the target risk in terms of the empirical source risk and the discrepancy between the source and target domains:

Theorem 2.1 (Ben-David et al. (2007))**.** Let $\mathcal{H}$ be a hypothesis space of VC-dimension d and $\widehat{\mathcal{D}}_S$ (resp. $\widehat{\mathcal{D}}_T$) be the empirical distribution induced by a sample of size n drawn from $\mathcal{D}_S$ (resp. $\mathcal{D}_T$). Then w.p. at least $1 - \delta$, $\forall h \in \mathcal{H}$,

$$\varepsilon_T(h) \leq \widehat{\varepsilon}_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}_S, \widehat{\mathcal{D}}_T) + \lambda^* + O\left(\sqrt{\frac{d \log n + \log(1/\delta)}{n}}\right). \quad (1)$$

The bound depends on λ^*, the optimal joint risk that can be achieved by the hypotheses in $\mathcal{H}$. The intuition is the following: if λ^* is large, we cannot hope for a successful domain adaptation. Later in Sec. 4.3, we shall get back to this term to show an information-theoretic lower bound on it for any approach based on learning invariant representations.

Theorem 2.1 is the foundation of many recent works on unsupervised domain adaptation via learning invariant representations (Ajakan et al., 2014; Ganin et al., 2016; Zhao et al., 2018b; Pei et al., 2018; Zhao et al., 2018a). It has also inspired various applications of domain adaptation with adversarial learning, e.g., video analysis (Hoffman et al., 2016; Shrivastava et al., 2016; Hoffman et al., 2017; Tzeng et al., 2017), natural language understanding (Zhang et al., 2017; Fu et al., 2017), speech recognition (Zhao et al., 2019; Hosseini-Asl et al., 2018), to name a few.

At a high level, the key idea is to learn a rich and parametrized feature transformation $g : \mathcal{X} \mapsto \mathcal{Z}$ such that the induced source and target distributions (on $\mathcal{Z}$) are close, as measured by the $\mathcal{H}$-divergence. We call g an *invariant representation* w.r.t. $\mathcal{H}$ if $d_\mathcal{H}(\mathcal{D}_S^g, \mathcal{D}_T^g) = 0$, where $\mathcal{D}_S^g / \mathcal{D}_T^g$ is the induced source/target distribution. At the same time, these algorithms also try to find new hypothesis (on the representation space $\mathcal{Z}$) to achieve a small empirical error on the source domain. As a whole algorithm, these two procedures corresponds to simultaneously finding invariant representations and hypothesis to minimize the first two terms in the generalization upper bound of Theorem 2.1.

3. Related Work

A number of adaptation approaches based on learning invariant representations have been proposed in recent years. Although in this paper we mainly focus on using the $\mathcal{H}$-divergence to characterize the discrepancy between two distributions, other distance measures can be used as well, e.g., the maximum mean discrepancy (MMD) (Long et al., 2014; 2015; 2016), the Wasserstein distance (Courty et al., 2017b;a; Shen et al., 2018; Lee & Raginsky, 2018), etc.

Under the theoretical framework of the $\mathcal{H}$-divergence, Ganin et al. (2016) propose a domain adversarial neural network (DANN) to learn the domain invariant features. Adversarial training techniques that aim to build feature representations that are indistinguishable between source and target domains have been proposed in the last few years (Ajakan

[1] To be precise, Ben-David et al. (2007)'s original definition of $\mathcal{H}$-divergence has a factor of 2, we choose the current definition as the constant factor is inessential.

et al., 2014; Ganin et al., 2016). Specifically, one of the central ideas is to use neural networks, which are powerful function approximators, to approximate the $\mathcal{H}$-divergence between two domains (Kifer et al., 2004; Ben-David et al., 2007; 2010). The overall algorithm can be viewed as a zero-sum two-player game: one network tries to learn feature representations that can fool the other network, whose goal is to distinguish the representations generated on the source domain from those generated on the target domain. In a concurrent work, Johansson et al. (2019) also identified the insufficiency of learning domain-invariant representation for successful adaptation. They further analyzed the information loss of non-invertible transformations, and proposed a generalization upper bound that directly takes it into account. In our work, by showing an information-theoretic lower bound on the joint error of these methods, we show that although invariant representations can be achieved, it does not necessarily translate to good generalization on the target domain, in particular when the label distributions of the two domains differ significantly.

Causal approaches based on conditional and label shifts for domain adaptation also exist (Zhang et al., 2013; Gong et al., 2016; Lipton et al., 2018; Azizzadenesheli et al., 2018). One typical assumption made to simplify the analysis in this line of work is that the source and target domains share the same generative distribution and only differ at the marginal label distributions. It is worth noting that Zhang et al. (2013) and Gong et al. (2016) showed that both label and conditional shift can be successfully corrected when the changes in the generative distribution follow some parametric families. In this work we focus on representation learning and do not make such explicit assumptions.

4. Theoretical Analysis

Is finding invariant representations alone a sufficient condition for the success of domain adaptation? Clearly it is not. Consider the following simple counterexample: let $g_{\mathbf{c}} : \mathcal{X} \mapsto \mathcal{Z}$ be a constant function, where $\forall \mathbf{x} \in \mathcal{X}$, $g_{\mathbf{c}}(\mathbf{x}) = \mathbf{c} \in \mathcal{Z}$. Then for any discrepancy distance $d(\cdot, \cdot)$ over two distributions, including the $\mathcal{H}$-divergence, MMD, and the Wasserstein distance, and for any distributions $\mathcal{D}_S, \mathcal{D}_T$ over the input space $\mathcal{X}$, we have $d(\mathcal{D}_S^{g_{\mathbf{c}}}, \mathcal{D}_T^{g_{\mathbf{c}}}) = 0$, where we use $\mathcal{D}_S^{g_{\mathbf{c}}}$ (resp. $\mathcal{D}_T^{g_{\mathbf{c}}}$) to mean the induced source (resp. target) distribution by the transformation $g_{\mathbf{c}}$ over the representation space $\mathcal{Z}$. Furthermore, it is fairly easy to construct source and target domains $\langle \mathcal{D}_S, f_S \rangle$, $\langle \mathcal{D}_T, f_T \rangle$, such that for any hypothesis $h : \mathcal{Z} \mapsto \mathcal{Y}$, $\varepsilon_T(h \circ g_{\mathbf{c}}) \geq 1/2$, while there exists a classification function $f : \mathcal{X} \to \mathcal{Y}$ that achieves small error, e.g., the labeling function.

One may argue, with good reason, that in the counterexample above, the empirical source error $\widehat{\varepsilon}_S(h \circ g_{\mathbf{c}})$ is also large with high probability. Intuitively, this is because the simple constant transformation function $g_{\mathbf{c}}$ fails to retain the discriminative information about the classification task at hand, despite the fact that it can construct invariant representations.

Is finding invariant representations and achieving a small source error sufficient to guarantee small target error? In this section we first give a negative answer to this question by constructing a counterexample where there exists a non-trivial transformation function $g : \mathcal{X} \mapsto \mathcal{Z}$ and hypothesis $h : \mathcal{Z} \mapsto \mathcal{Y}$ such that both $\varepsilon_S(h \circ g)$ and $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S^g, \mathcal{D}_T^g)$ are small, while at the same time the target error $\varepsilon_T(h \circ g)$ is large. Motivated by this negative result, we proceed to prove a generalization upper bound that explicitly characterizes a sufficient condition for the success of domain adaptation. We then complement the upper bound by showing an information-theoretic lower bound on the joint error of *any* domain adaptation approach based on learning invariant representations.

4.1. Invariant Representation and Small Source Risk are Not Sufficient

In this section, we shall construct a simple 1-dimensional example where there exists a function $h^* : \mathbb{R} \mapsto \{0, 1\}$ that achieves zero error on *both* source and target domains. Simultaneously, we show that there exists a transformation function $g : \mathbb{R} \mapsto \mathbb{R}$ under which the induced source and target distributions are perfectly aligned, but *every* hypothesis $h : \mathbb{R} \mapsto \{0, 1\}$ incurs a large joint error on the induced source and target domains. The latter further implies that if we find a hypothesis that achieves small error on the source domain, then it has to incur a large error on the target domain. We illustrate this example in Fig. 1.

Let $\mathcal{X} = \mathcal{Z} = \mathbb{R}$ and $\mathcal{Y} = \{0, 1\}$. For $a \leq b$, we use $U(a, b)$ to denote the uniform distribution over $[a, b]$. Consider the following source and target domains:

$$\mathcal{D}_S = U(-1, 0), \qquad f_S(x) = \begin{cases} 0, & x \leq -1/2 \\ 1, & x > -1/2 \end{cases}$$

$$\mathcal{D}_T = U(1, 2), \qquad f_T(x) = \begin{cases} 0, & x \geq 3/2 \\ 1, & x < 3/2 \end{cases}$$

In the above example, it is easy to verify that the interval hypothesis $h^*(x) = 1$ iff $x \in (-1/2, 3/2)$ achieves perfect classification on *both* domains.

Now consider the following transformation:

$$g(x) = \mathbb{I}_{x \leq 0}(x) \cdot (x + 1) + \mathbb{I}_{x > 0}(x) \cdot (x - 1).$$

Since $g(\cdot)$ is a piecewise linear function, it follows that $\mathcal{D}_S^Z = \mathcal{D}_T^Z = U(0, 1)$, and for any distance metric $d(\cdot, \cdot)$ over distributions, we have $d(\mathcal{D}_S^Z, \mathcal{D}_T^Z) = 0$. But now for any hypothesis $h : \mathbb{R} \mapsto \{0, 1\}$, and $\forall x \in [0, 1]$, $h(x)$ will

make an error in exactly one of the domains, hence

$$\forall h : \mathbb{R} \mapsto \{0, 1\}, \quad \varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) = 1.$$

In other words, under the above invariant transformation g, the smaller the source error, the larger the target error.

One may argue that this example seems to contradict the generalization upper bound from Theorem 2.1, where the first two terms correspond exactly to a small source error and an invariant representation. The key to explain this apparent contradiction lies in the third term of the upper bound, λ^*, i.e., the optimal joint error achievable on *both* domains. In our example, when there is no transformation applied to the input space, we show that h^* achieves 0 error on both domains, hence $\lambda^* = \min_{h \in \mathcal{H}} \varepsilon_S(h) + \varepsilon_T(h) = 0$. However, when the transformation g is applied to the original input space, we prove that every hypothesis has joint error 1 on the representation space, hence $\lambda_g^* = 1$. Since we usually do not have access to the optimal hypothesis on both domains, although the generalization bound still holds on the representation space, it becomes vacuous in our example.

An alternative way to interpret the failure of the constructed example is that the labeling functions (or conditional distributions in the stochastic setting) of source and target domains are far away from each other in the representation space. Specifically, in the induced representation space, the optimal labeling function on the source and target domains are:

$$f'_S(x) = \begin{cases} 0, & x \leq 1/2 \\ 1, & x > 1/2 \end{cases}, \quad f'_T(x) = \begin{cases} 0, & x > 1/2 \\ 1, & x \leq 1/2 \end{cases},$$

and we have $||f'_S - f'_T||_1 = \mathbb{E}_{x \sim U(0,1)}[|f'_S(x) - f'_T(x)|] = 1$.

4.2. A Generalization Upper Bound

For most of the practical hypothesis spaces $\mathcal{H}$, e.g., half spaces, it is usually intractable to compute the optimal joint error λ^* from Theorem 2.1. Furthermore, the fact that λ^* contains errors from both domains makes the bound very conservative and loose in many cases. In this section, inspired by the constructed example from Sec. 4.1, we aim to provide a general, intuitive, and interpretable generalization upper bound for domain adaptation that is free of the pessimistic λ^* term. Ideally, the bound should also explicitly characterize how the shift between labeling functions of both domains affects domain adaptation. Due to space constraints, we refer the interested reader to the Appendix for the proofs of our technical lemmas, and mainly focus in the following on interpretations and results.

Because of its flexibility in choosing the witness function class $\mathcal{H}$ and its natural interpretation as adversarial binary classification, we still adopt the $\mathcal{H}$-divergence to measure the discrepancy between two distributions. For any hypothesis space $\mathcal{H}$, it can be readily verified that $d_{\mathcal{H}}(\cdot, \cdot)$ satisfies the triangular inequality:

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') \leq d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}'') + d_{\mathcal{H}}(\mathcal{D}'', \mathcal{D}'),$$

where $\mathcal{D}, \mathcal{D}', \mathcal{D}''$ are any distributions over the same space. We now introduce a technical lemma that will be helpful in proving results related to the $\mathcal{H}$-divergence:

Lemma 4.1. Let $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$ and $\mathcal{D}, \mathcal{D}'$ be two distributions over $\mathcal{X}$. Then $\forall h, h' \in \mathcal{H}$, $|\varepsilon_{\mathcal{D}}(h, h') - \varepsilon_{\mathcal{D}'}(h, h')| \leq d_{\tilde{\mathcal{H}}}(\mathcal{D}, \mathcal{D}')$, where $\tilde{\mathcal{H}} := \{\text{sgn}(|h(\mathbf{x}) - h'(\mathbf{x})| - t) \mid h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$.

As a matter of fact, the above lemma also holds for any function class $\mathcal{H}$ (not necessarily a hypothesis space) where there exists a constant $M > 0$, such that $||h||_\infty \leq M$ for all $h \in \mathcal{H}$. Another useful lemma is the following triangular inequality:

Lemma 4.2. Let $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$ and $\mathcal{D}$ be any distribution over $\mathcal{X}$. For any $h, h', h'' \in \mathcal{H}$, we have $\varepsilon_{\mathcal{D}}(h, h') \leq \varepsilon_{\mathcal{D}}(h, h'') + \varepsilon_{\mathcal{D}}(h'', h')$.

Let $f_S : \mathcal{X} \to [0, 1]$ and $f_T : \mathcal{X} \to [0, 1]$ be the optimal labeling functions on the source and target domains, respectively. In the stochastic setting, $f_S(\mathbf{x}) = \Pr_S(y = 1 \mid \mathbf{x})$ corresponds to the optimal Bayes classifier. With these notations, the following theorem holds:

Theorem 4.1. Let $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$ be the source and target domains, respectively. For any function class $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$, and $\forall h \in \mathcal{H}$, the following inequality holds:

$$\begin{aligned} \varepsilon_T(h) \leq \varepsilon_S(h) + d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) \\ + \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\}. \end{aligned}$$

Remark The three terms in the upper bound have natural interpretations: the first term is the source error, the second one corresponds to the discrepancy between the marginal distributions, and the third measures the distance between the labeling functions from the source and target domains. Altogether, they form a sufficient condition for the success of domain adaptation: besides a small source error, not only do the marginal distributions need to be close, but so do the labeling functions.

Comparison with Theorem 2.1. It is instructive to compare the bound in Theorem 4.1 with the one in Theorem 2.1. The main difference lies in the λ^* in Theorem 2.1 and the $\min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\}$ in Theorem 4.1. λ^* depends on the choice of the hypothesis class $\mathcal{H}$, while our term does not. In fact, our quantity reflects the underlying structure of the problem, i.e., the conditional shift. Finally, consider the example given in the left panel of Fig. 1. It is easy to verify that

we have $\min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\} = 1/2$ in this case, while for a natural class of hypotheses, i.e., $\mathcal{H} := \{h(x) = 0 \Leftrightarrow a \le x \le b \mid a < b\}$, we have $\lambda^* = 1$. In that case, our bound is tighter than the one in Theorem 2.1.

In the covariate shift setting, where we assume the conditional distributions of $Y \mid X$ between the source and target domains are the same, the third term in the upper bound vanishes. In that case the above theorem says that to guarantee successful domain adaptation, it suffices to match the marginal distributions while achieving small error on the source domain. In general settings where the optimal labeling functions of the source and target domains differ, the above bound says that it is not sufficient to simply match the marginal distributions and achieve small error on the source domain. At the same time, we should also guarantee that the optimal labeling functions (or the conditional distributions of both domains) are not too far away from each other. As a side note, it is easy to see that $\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|] = \varepsilon_S(f_T)$ and $\mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|] = \varepsilon_T(f_S)$. In other words, they are essentially the cross-domain errors. When the cross-domain error is small, it implies that the optimal source (resp. target) labeling function generalizes well on the target (resp. source) domain.

Both the error term $\varepsilon_S(h)$ and the divergence $d_{\tilde{H}}(\mathcal{D}_S, \mathcal{D}_T)$ in Theorem 4.1 are with respect to the true underlying distributions $\mathcal{D}_S$ and $\mathcal{D}_T$, which are not available to us during training. In the following, we shall use the Rademacher complexity to provide for both terms a data-dependent bound from empirical samples from $\mathcal{D}_S$ and $\mathcal{D}_T$.

Definition 4.1 (Empirical Rademacher Complexity)**.** Let $\mathcal{H}$ be a family of functions mapping from $\mathcal{X}$ to $[a, b]$ and $\mathbf{S} = \{\mathbf{x}_i\}_{i=1}^n$ a fixed sample of size n with elements in $\mathcal{X}$. Then, the *empirical Rademacher complexity* of $\mathcal{H}$ with respect to the sample X is defined as

$$\mathrm{Rad}_{\mathbf{S}}(\mathcal{H}) := \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(\mathbf{x}_i)\right],$$

where $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1}^n$ and σ_i are i.i.d. uniform random variables taking values in $\{+1, -1\}$.

With the empirical Rademacher complexity, we can show that w.h.p., the empirical source error $\widehat{\varepsilon}_S(h)$ cannot be too far away from the population error $\varepsilon_S(h)$ for all $h \in \mathcal{H}$:

Lemma 4.3. Let $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$, then for all $\delta > 0$, w.p. at least $1 - \delta$, the following inequality holds for all $h \in \mathcal{H}$: $\varepsilon_S(h) \le \widehat{\varepsilon}_S(h) + 2\mathrm{Rad}_{\mathbf{S}}(\mathcal{H}) + 3\sqrt{\log(2/\delta)/2n}$.

Similarly, for any distribution $\mathcal{D}$ over $\mathcal{X}$, let $\widehat{\mathcal{D}}$ be its empirical distribution from sample $\mathbf{S} \sim \mathcal{D}^n$ of size n. Then for any two distributions $\mathcal{D}$ and $\mathcal{D}'$, we can also use the empirical Rademacher complexity to provide a data-dependent bound for the perturbation between $d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}')$ and $d_{\mathcal{H}}(\widehat{\mathcal{D}}, \widehat{\mathcal{D}}')$:

Lemma 4.4. Let $\tilde{\mathcal{H}}$, $\mathcal{D}$ and $\widehat{\mathcal{D}}$ be defined above, then for all $\delta > 0$, w.p. at least $1 - \delta$, the following inequality holds for all $h \in \tilde{\mathcal{H}}$: $\mathbb{E}_{\mathcal{D}}[\mathbb{I}_h] \le \mathbb{E}_{\widehat{\mathcal{D}}}[\mathbb{I}_h] + 2\mathrm{Rad}_{\mathbf{S}}(\tilde{\mathcal{H}}) + 3\sqrt{\log(2/\delta)/2n}$.

Since $\tilde{\mathcal{H}}$ is a hypothesis class, by definition we have:

$$\begin{aligned} d_{\tilde{\mathcal{H}}}(\mathcal{D}, \widehat{\mathcal{D}}) &= \sup_{A\in\mathcal{A}_{\tilde{\mathcal{H}}}} |\Pr_{\mathcal{D}}(A) - \Pr_{\widehat{\mathcal{D}}}(A)| \\ &= \sup_{h\in\tilde{\mathcal{H}}} |\mathbb{E}_{\mathcal{D}}[\mathbb{I}_h] - \mathbb{E}_{\widehat{\mathcal{D}}}[\mathbb{I}_h]|. \end{aligned}$$

Hence combining the above identity with Lemma 4.4, we immediately have w.p. at least $1 - \delta$:

$$d_{\tilde{\mathcal{H}}}(\mathcal{D}, \widehat{\mathcal{D}}) \le 2\mathrm{Rad}_{\mathbf{S}}(\tilde{\mathcal{H}}) + 3\sqrt{\log(2/\delta)/2n}. \quad (2)$$

Now use a union bound and the fact that $d_{\tilde{\mathcal{H}}}(\cdot, \cdot)$ satisfies the triangle inequality, we have:

Lemma 4.5. Let $\tilde{\mathcal{H}}$, $\mathcal{D}, \mathcal{D}'$ and $\widehat{\mathcal{D}}, \widehat{\mathcal{D}}'$ be defined above, then for $\forall\delta > 0$, w.p. at least $1 - \delta$, for $\forall h \in \tilde{\mathcal{H}}$:

$$d_{\tilde{\mathcal{H}}}(\mathcal{D}, \mathcal{D}') \le d_{\tilde{\mathcal{H}}}(\widehat{\mathcal{D}}, \widehat{\mathcal{D}}') + 4\mathrm{Rad}_{\mathbf{S}}(\tilde{\mathcal{H}}) + 6\sqrt{\log(4/\delta)/2n}.$$

Combine Lemma 4.3, Lemma 4.5 and Theorem 4.1 with a union bound argument, we get the following main theorem that characterizes an upper bound for domain adaptation:

Theorem 4.2. Let $\langle \mathcal{D}_S, f_S\rangle$ and $\langle \mathcal{D}_T, f_T\rangle$ be the source and target domains, and let $\widehat{\mathcal{D}}_S, \widehat{\mathcal{D}}_T$ be the empirical source and target distributions constructed from sample $\mathbf{S} = \{\mathbf{S}_S, \mathbf{S}_T\}$, each of size n. Then for any $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$ and $\forall h \in \mathcal{H}$:

$$\begin{aligned} \varepsilon_T(h) \le \widehat{\varepsilon}_S(h) + d_{\tilde{\mathcal{H}}}(\widehat{\mathcal{D}}_S, \widehat{\mathcal{D}}_T) + 2\mathrm{Rad}_{\mathbf{S}}(\mathcal{H}) + 4\mathrm{Rad}_{\mathbf{S}}(\tilde{\mathcal{H}}) \\ + \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\} \\ + O\left(\sqrt{\log(1/\delta)/n}\right), \end{aligned}$$

where $\tilde{\mathcal{H}} := \{\mathrm{sgn}(|h(\mathbf{x}) - h'(\mathbf{x})| - t) | h, h' \in \mathcal{H}, t \in [0, 1]\}$.

Essentially, the generalization upper bound can be decomposed into three parts: the first part comes from the domain adaptation setting, including the empirical source error, the empirical $\mathcal{H}$-divergence, and the shift between labeling functions. The second part corresponds to complexity measures of our hypothesis space $\mathcal{H}$ and $\tilde{\mathcal{H}}$, and the last part describes the error caused by finite samples.

4.3. An Information-Theoretic Lower Bound

In Sec. 4.1, we constructed an example to demonstrate that learning invariant representations could lead to a feature space where the joint error on both domains is large. In this section, we extend the example by showing that a similar result holds in more general settings. Specifically, we shall prove that for *any* approach based on learning invariant representations, there is an intrinsic lower bound on the joint

error of source and target domains, due to the discrepancy between their marginal label distributions. Our result hence highlights the need to take into account task related information when designing domain adaptation algorithms based on learning invariant representations.

Before we proceed to the lower bound, we first define several information-theoretic concepts that will be used in the analysis. For two distributions $\mathcal{D}$ and $\mathcal{D}'$, the Jensen-Shannon (JS) divergence $D_{\text{JS}}(\mathcal{D} \,||\, \mathcal{D}')$ is defined as:

$$D_{\text{JS}}(\mathcal{D} \,||\, \mathcal{D}') := \frac{1}{2} D_{\text{KL}}(\mathcal{D} \,||\, \mathcal{D}_M) + \frac{1}{2} D_{\text{KL}}(\mathcal{D}' \,||\, \mathcal{D}_M),$$

where $D_{\text{KL}}(\cdot \,||\, \cdot)$ is the Kullback–Leibler (KL) divergence and $\mathcal{D}_M := (\mathcal{D} + \mathcal{D}')/2$. The JS divergence can be viewed as a symmetrized and smoothed version of the KL divergence, and it is closely related to the L_1 distance between two distributions through Lin's lemma (Lin, 1991).

Unlike the KL divergence, the JS divergence is bounded: $0 \leq D_{\text{JS}}(\mathcal{D} \,||\, \mathcal{D}') \leq 1$. Additionally, from the JS divergence, we can define a distance metric between two distributions as well, known as the JS distance (Endres & Schindelin, 2003):

$$d_{\text{JS}}(\mathcal{D}, \mathcal{D}') := \sqrt{D_{\text{JS}}(\mathcal{D} \,||\, \mathcal{D}')}.$$

With respect to the JS distance and for any (stochastic) mapping $h : \mathcal{Z} \mapsto \mathcal{Y}$, we can prove the following lemma via the celebrated data processing inequality:

Lemma 4.6. Let $\mathcal{D}_S^Z$ and $\mathcal{D}_T^Z$ be two distributions over $\mathcal{Z}$ and let $\mathcal{D}_S^Y$ and $\mathcal{D}_T^Y$ be the induced distributions over $\mathcal{Y}$ by function $h : \mathcal{Z} \mapsto \mathcal{Y}$, then

$$d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \leq d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z). \tag{3}$$

For methods that aim to learn invariant representations for domain adaptation, an intermediate representation space $\mathcal{Z}$ is found through feature transformation g, based on which a common hypothesis $h : \mathcal{Z} \mapsto \mathcal{Y}$ is shared between both domains (Ganin et al., 2016; Tzeng et al., 2017; Zhao et al., 2018b). Through this process, the following Markov chain holds:

$$X \xrightarrow{g} Z \xrightarrow{h} \hat{Y}, \tag{4}$$

where $\hat{Y} = h(g(X))$ is the predicted random variable of interest. Hence for any distribution $\mathcal{D}$ over $\mathcal{X}$, this Markov chain also induces a distribution $\mathcal{D}^Z$ over $\mathcal{Z}$ and $\mathcal{D}^{\hat{Y}}$ over $\mathcal{Y}$. By Lemma 4.6, we know that $d_{\text{JS}}(\mathcal{D}_S^{\hat{Y}}, \mathcal{D}_T^{\hat{Y}}) \leq d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$. With these notations, noting that the JS distance is a metric, the following inequality holds:

$$d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \leq d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_S^{\hat{Y}}) + d_{\text{JS}}(\mathcal{D}_S^{\hat{Y}}, \mathcal{D}_T^{\hat{Y}}) + d_{\text{JS}}(\mathcal{D}_T^{\hat{Y}}, \mathcal{D}_T^Y).$$

Combining the above inequality with Lemma 4.6, we immediately have:

$$d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \leq d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z) + d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_S^{\hat{Y}}) + d_{\text{JS}}(\mathcal{D}_T^Y, \mathcal{D}_T^{\hat{Y}}). \tag{5}$$

Intuitively, $d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_S^{\hat{Y}})$ and $d_{\text{JS}}(\mathcal{D}_T^Y, \mathcal{D}_T^{\hat{Y}})$ measure the distance between the predicted label distribution and the ground truth label distribution on the source and target domain, respectively. With the help of Lemma **??**, the following result establishes a relationship between $d_{\text{JS}}(\mathcal{D}^Y, \mathcal{D}^{\hat{Y}})$ and the accuracy of the prediction function h:

Lemma 4.7. Let $Y = f(X) \in \{0, 1\}$ where $f(\cdot)$ is the labeling function and $\hat{Y} = h(g(X)) \in \{0, 1\}$ be the prediction function, then $d_{\text{JS}}(\mathcal{D}^Y, \mathcal{D}^{\hat{Y}}) \leq \sqrt{\varepsilon(h \circ g)}$.

We are now ready to present the key lemma of the section:

Lemma 4.8. Suppose the Markov chain $X \xrightarrow{g} Z \xrightarrow{h} \hat{Y}$ holds, then

$$d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \leq d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z) + \sqrt{\varepsilon_S(h \circ g)} + \sqrt{\varepsilon_T(h \circ g)}.$$

Remark This lemma shows that if the marginal label distributions are significantly different between the source and target domains, then in order to achieve a small joint error, the induced distributions over $\mathcal{Z}$ from source and target domains have to be significantly different as well. Put another way, if we are able to find an invariant representation such that $d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z) = 0$, then the joint error of the composition function $h \circ g$ has to be large:

Theorem 4.3. Suppose the condition in Lemma 4.8 holds and $d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then:

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2} \left(d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z) \right)^2.$$

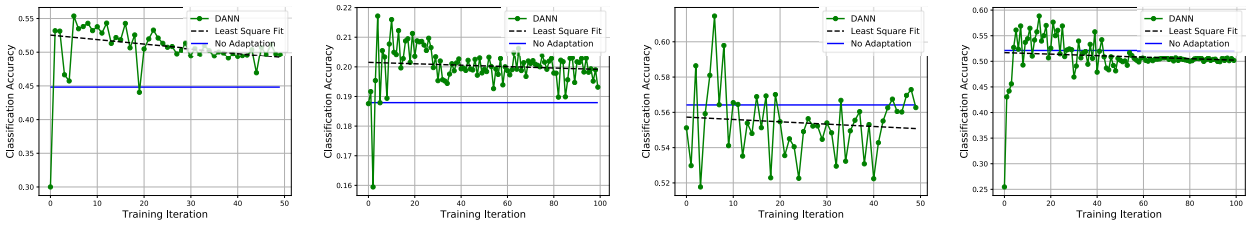
Remark The lower bound gives us a necessary condition on the success of any domain adaptation approach based on learning invariant representations: if the marginal label distributions are significantly different between source and target domains, then minimizing $d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$ and the source error $\varepsilon_S(h \circ g)$ will only increase the target error. In fact, Theorem 4.3 can be extended to hold in the setting where different transformation functions are applied in source and target domains:

Corollary 4.1. Let g_S, g_T be the source and target transformation functions from $\mathcal{X}$ to $\mathcal{Z}$. Suppose the condition in Lemma 4.8 holds and $d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then:

$$\varepsilon_S(h \circ g_S) + \varepsilon_T(h \circ g_T) \geq \frac{1}{2} \left(d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z) \right)^2.$$

Recent work has also explored using different transformation functions to achieve invariant representations (Bousmalis et al., 2016; Tzeng et al., 2017), but Corollary 4.1 shows that this is not going to help if the marginal label distributions differ between two domains.

We conclude this section by noting that our bound on the joint error of both domains is not necessarily the tightest one. This can be seen from the example in Sec. 4.1, where $d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z) = d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) = 0$, and we have $\varepsilon_S(h \circ g) +$



(a) USPS → MNIST (b) USPS → SVHN (c) SVHN → MNIST (d) SVHN → USPS

Figure 2. Digit classification on MNIST, USPS and SVHN. The horizontal solid line corresponds to the target domain test accuracy without adaptation. The green solid line is the target domain test accuracy under domain adaptation with DANN. We also plot the least square fit (dashed line) of the DANN adaptation results to emphasize the negative slope.

$\varepsilon_T(h \circ g) = 1$, but in this case our result gives a trivial lower bound of 0. Nevertheless, our result still sheds new light on the importance of matching marginal label distributions in learning invariant representation for domain adaptation, which we believe to be a promising direction for the design of better adaptation algorithms.

5. Experiments

Our theoretical results on the lower bound of the joint error imply that over-training the feature transformation function and the discriminator may hurt generalization on the target domain. In this section, we conduct experiments on real-world datasets to verify our theoretical findings. The task is digit classification on three datasets of 10 classes: MNIST, USPS and SVHN. MNIST contains 60,000/10,000 train/test instances; USPS contains 7,291/2,007 train/test instances, and SVHN contains 73,257/26,032 train/test instances. We show the label distribution of these three datasets in Fig. 3.

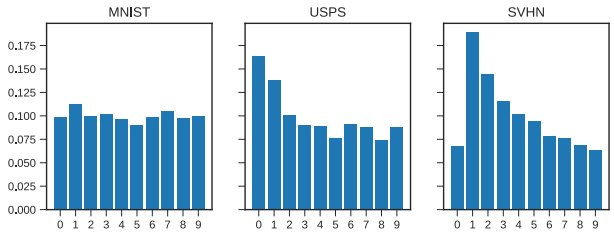


Figure 3. The label distributions of MNIST, USPS and SVHN.

Before training, we preprocess all the samples into gray scale single-channel images of size 16×16, so they can be used by the same network. In our experiments, to ensure a fair comparison, we use the same network structure for all the experiments: 2 convolutional layers, one fully connected hidden layer, followed by a softmax output layer with 10 units. The convolution kernels in both layers are of size 5×5, with 10 and 20 channels, respectively. The hidden layer has 1280 units connected to 100 units before classification. For domain adaptation, we use the original DANN (Ganin et al., 2016) with gradient reversal implementation. The discriminator in DANN takes the output of convolutional layers as its feature input, followed by a 500×100 fully connected layer, and a one-unit binary classification output.

We plot four adaptation trajectories in Fig. 2. Among the four adaptation tasks, we can observe two phases in the adaptation accuracy. In the first phase, the test set accuracy rapidly grows, in less than 10 iterations. In the second phase, it gradually decreases after reaching its peak, despite the fact that the source training accuracy keeps increasing smoothly. Those phase transitions can be verified from the negative slopes of the least squares fit of the adaptation curves (dashed lines in Fig. 2). We observe similar phenomenons on additional experiments using artificially unbalanced datasets trained on more powerful networks in Appendix **??**. The above experimental results imply that over-training the feature transformation and discriminator does not help generalization on the target domain, but can instead hurt it when the label distributions differ (as shown in Fig. 3). These experimental results are consistent with our theoretical findings.

6. Conclusion and Future Work

In this paper we theoretically and empirically study the important problem of learning invariant representations for domain adaptation. We show that learning an invariant representation and achieving a small source error is not enough to guarantee target generalization. We then prove both upper and lower bounds for the target and joint errors, which directly translate to sufficient and necessary conditions for the success of adaptation. We believe our results take an important step towards understanding deep domain adaptation, and also stimulate future work on the design of stronger deep domain adaptation algorithms that align conditional distributions. Another interesting direction for future work is to characterize what properties the feature transformation function should have in order to decrease the conditional shift. It is also worth investigating under which conditions the label distributions can be aligned without explicit labeled data from the target domain.

Acknowledgements

HZ thank Yifan Wu for inspiring discussion on possible failure cases of deep domain adaptation, and Jeremy Cohen for reviewing drafts of this work. HZ and GG would like to acknowledge the support from DARPA XAI project, contract #FA87501720152. KZ would like to acknowledge the support by National Institutes of Health (NIH) under Contract No. NIH-1R01EB022858-01, FAINR01EB022858, NIH-1R01LM012087, NIH-5U54HG008540-02, and FAIN-U54HG008540, by the United States Air Force under Contract No. FA8650-17-C-7715, and by National Science Foundation (NSF) EAGER Grant No. IIS-1829681. The NIH, the U.S. Air Force, and the NSF are not responsible for the views reported in this article.

References

Adel, T., Zhao, H., and Wong, A. Unsupervised domain adaptation with a relaxed covariate shift assumption. In *AAAI*, pp. 1691–1697, 2017.

Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., and Marchand, M. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

Azizzadenesheli, K., Liu, A., Yang, F., and Anandkumar, A. Regularized learning for domain adaptation under label shifts. 2018.

Becker, C. J., Christoudias, C. M., and Fua, P. Non-linear domain adaptation with boosting. In *Advances in Neural Information Processing Systems*, pp. 485–493, 2013.

Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Wortman, J. Learning bounds for domain adaptation. In *Advances in neural information processing systems*, pp. 129–136, 2008.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. Domain separation networks. In *Advances in Neural Information Processing Systems*, pp. 343–351, 2016.

Cortes, C. and Mohri, M. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

Cortes, C., Mohri, M., Riley, M., and Rostamizadeh, A. Sample selection bias correction theory. In *International Conference on Algorithmic Learning Theory*, pp. 38–53. Springer, 2008.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 3730–3739, 2017a.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9): 1853–1865, 2017b.

Endres, D. M. and Schindelin, J. E. A new metric for probability distributions. *IEEE Transactions on Information theory*, 2003.

Fu, L., Nguyen, T. H., Min, B., and Grishman, R. Domain adaptation for relation extraction with domain adversarial neural network. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pp. 425–429, 2017.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Glorot, X., Bordes, A., and Bengio, Y. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 513–520, 2011.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., and Schölkopf, B. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pp. 2839–2848, 2016.

Hoffman, J., Wang, D., Yu, F., and Darrell, T. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A. A., and Darrell, T. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

Hosseini-Asl, E., Zhou, Y., Xiong, C., and Socher, R. Augmented cyclic adversarial learning for domain adaptation. *arXiv preprint arXiv:1807.00374*, 2018.

Johansson, F. D., Ranganath, R., and Sontag, D. Support and invertibility in domain-invariant representations. *arXiv preprint arXiv:1903.03448*, 2019.

Kifer, D., Ben-David, S., and Gehrke, J. Detecting change in data streams. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pp. 180–191. VLDB Endowment, 2004.

Lee, J. and Raginsky, M. Minimax statistical learning with wasserstein distances. In *Advances in Neural Information Processing Systems*, pp. 2692–2701, 2018.

Lin, J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

Lipton, Z. C., Wang, Y.-X., and Smola, A. Detecting and correcting for label shift with black box predictors. *arXiv preprint arXiv:1802.03916*, 2018.

Long, M., Wang, J., Ding, G., Sun, J., and Yu, P. S. Transfer joint matching for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1410–1417, 2014.

Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pp. 97–105, 2015.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pp. 136–144, 2016.

Mansour, Y. and Schain, M. Robust domain adaptation. In *ISAIM*, 2012.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009a.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Multiple source adaptation and the rényi divergence. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 367–374. AUAI Press, 2009b.

Pei, Z., Cao, Z., Long, M., and Wang, J. Multi-adversarial domain adaptation. 2018.

Shen, J., Qu, Y., Zhang, W., and Yu, Y. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. Learning from simulated and unsupervised images through adversarial training. *arXiv preprint arXiv:1612.07828*, 2016.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. *arXiv preprint arXiv:1702.05464*, 2017.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pp. 819–827, 2013.

Zhang, Y., Barzilay, R., and Jaakkola, T. Aspect-augmented adversarial networks for domain adaptation. *arXiv preprint arXiv:1701.00188*, 2017.

Zhao, H., Zhang, S., Wu, G., Gordon, G. J., et al. Multiple source domain adaptation with adversarial learning. In *International Conference on Learning Representations*, 2018a.

Zhao, H., Zhang, S., Wu, G., Moura, J. M., Costeira, J. P., and Gordon, G. J. Adversarial multiple source domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 8568–8579, 2018b.

Zhao, H., Hu, J., Zhu, Z., Coates, A., and Gordon, G. J. Deep generative and discriminative domain adaptation. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2019.