Support and Invertibility in Domain-Invariant Representations

Fredrik D. Johansson
MIT

David Sontag
MIT

Rajesh Ranganath
NYU

Abstract

Learning domain-invariant representations has become a popular approach to unsupervised domain adaptation and is often justified by invoking a particular suite of theoretical results. We argue that there are two significant flaws in such arguments. First, the results in question hold only for a fixed representation and do not account for information lost in non-invertible transformations. Second, domain invariance is often a far too strict requirement and does not always lead to consistent estimation, even under strong and favorable assumptions. In this work, we give generalization bounds for unsupervised domain adaptation that hold for any representation function by acknowledging the cost of non-invertibility. In addition, we show that penalizing distance between densities is often wasteful and propose a bound based on measuring the extent to which the support of the source domain covers the target domain. We perform experiments on well-known benchmarks that illustrate the short-comings of current standard practice.

1 Introduction

Domain transfer is a critical component of many machine learning problems: Self-driving cars must be robust to changes in weather conditions and landscape; Estimates of the efficacy of drugs that pass clinical trials should be valid for the population to which the drugs are prescribed; Policies for robotic control learned in simulated environments should be useful in the real world. In so-called *unsupervised domain adaptation*, labeled data are available only in a limited setting (e.g. driving only in San Francisco; patients restricted to a clinical trial cohort; simulated environments) called the *source domain*. The context in which models are ultimately applied is called the *target domain*.

When the label function is assumed stationary and source and target domains share statistical support, the classical solution to domain adaptation problems is *importance sampling* (IS) (Shimodaira, 2000). In IS methods, the influence of an observation on the learning algorithm is determined by its likelihood ratio between target and source domains. While asymptotically unbiased, IS estimators suffer from large variance (Cortes *et al.*, 2010) and are inapplicable when the target domain is not covered by the source. The latter is typical for many of the high-dimensional problems addressed in modern machine learning.

Domain-invariant representations have emerged as new, widely-used tools for domain transfer (Ben-David *et al.*, 2007; Ganin *et al.*, 2016; Long *et al.*, 2015) in problems where the label function is assumed fixed, but the covariate distribution changes between domains—so-called *covariate shift*. These methods work by uncovering predictive components of data that are distributed similarly across domains—an idea that has been justified by a string of theoretical work (Ben-David *et al.*, 2007; Mansour *et al.*, 2009; Ben-David *et al.*, 2010a; Cortes & Mohri, 2011). Crucially, these bounds do not rely on common support. Related ideas have been applied also under target (label) shift (Gong *et al.*, 2016).

Given the prevalence of algorithms learning domain-invariant representations, we ask: Under what conditions do these algorithms recover an optimal hypothesis? What are potential failure modes? We argue that there is discord between existing theoretical guarantees, how they are used to justify learning algorithms, and how these algorithms perform empirically. In particular, we give small example in which a) the objective of many algorithms is minimal but the target error is arbitrarily bad, and b) empirical performance is good but generalization bounds are surprisingly large.

First, we argue that regularizing representations to be domain invariant is too strict, in particular when domains (partially) overlap. We support this claim by

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

giving examples where empirical risk minimization on source data only outperforms domain-invariant representation learning algorithms. As an alternative, we give a generalization bound that measures the lack of *overlapping support* between domains. Our bound applies directly to learned representations and is tight when source and target domains are equal.

Second, for domain-invariant representation learning to succeed, the label must be predictable from the learned representation. When representations are regularized to reduce domain discrepancy, the class of admissible hypotheses shrinks and predictions worsen. This phenomenon may be asymmetric: a representation may be more suitable for the source domain than the target domain. We use this insight to characterize the unobservable adaptation error from losing information in non-invertible representations.

Finally, we study the performance of domain-invariant representation learning on a well-known benchmark task through the lens of our theoretical findings.

2 Background

We study *unsupervised domain adaptation*, defined as follows. Samples $\mathcal{D}_s = \{(x_i, y_i)\}_{i=1}^n$ of features $X \in \mathcal{X}$ and labels $Y \in \mathcal{Y}$ are observed from a *source domain*, distributed according to a density $p_s(X, Y)$. In addition, we observe *unlabeled* samples $\mathcal{D}_t = \{x'_i\}_{i=1}^m$ from a *target domain*, distributed according to a density $p_t(X)$. Unobserved labels in the target domain are distributed according to $p_t(Y \mid X)$. Based on $\mathcal{D}_s$ and $\mathcal{D}_t$, the unsupervised domain adaptation problem is to obtain an hypothesis $h \in \mathcal{H}$ that minimizes the *target risk* R_t as measured by a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$,

$$R_t^\ell(h) \coloneqq \mathop{\mathbb{E}}_{x,y\sim p_t}[\ell(h(x), y)] \ . \tag{1}$$

Analogously to (1), we define the *source risk* as $R_s^\ell(h) = \mathbb{E}_{x,y\sim p_s}[\ell(h(x), y)]$. When clear from context, we leave out the superscript ℓ indicating the loss function. In the sequel, unless otherwise stated, we let $\mathcal{Y} = \{0, 1\}$ and ℓ be the zero-one loss, $\ell(y, y') = \mathbb{1}[y \neq y']$. We call $R_t(h) - R_s(h)$ the *adaptation error*.

In this work, we make the *covariate shift* assumption which states that the conditional density of labels given features is stationary across domains. This is justifiable in some problems but not in all (Gong *et al.*, 2016).

Assumption 1. *Domains $p_s(X, Y)$ and $p_t(X, Y)$ satisfy the covariate shift assumption if*

$$p_s(Y \mid X) = p_t(Y \mid X) = p(Y \mid X) \ .$$

We say that Y is *realizable* in $\mathcal{H}$ if $p(Y \mid X) \in \mathcal{H}$ and that Y is *identifiable* over p_t if, under a set of assumptions on p_s, p_t, a function h may be obtained based on knowledge of $p_s(Y, X)$ and $p_t(X)$ such that $\forall x \in \operatorname{supp}(p_t) : h(x) = p(Y \mid X = x)$. In the case of deterministic hypotheses and labels, or when only label expectations are of interest, we may substitute conditional densities with appropriate mappings.

As no labels are observed from the target domain, models that minimize risk (only) on the source domain are often biased. There are two common alternative strategies to minimize target risk: *importance-weighting* and *minimization of upper bounds on the target risk.*

2.1 Importance weighting

Under Assumption 1 (covariate shift), the target risk R_t may be approximated using importance-weighted samples $\mathcal{D}_s$ from the source density (Shimodaira, 2000),

$$\hat{R}_s^w(h) \coloneqq \frac{1}{n} \sum_{i=1}^n w(x_i)\ell(h(x_i), y_i) \ . \tag{2}$$

If the weighting function w is chosen to be $w(x) = p_t(x)/p_s(x)$, $\hat{R}_s^w(h)$ is a consistent estimator of $R_t(h)$ under the following assumption.

Assumption 2 (Sufficient support)**.** *We say that p_s has ϵ-sufficient support for p_t if $\forall x \in \operatorname{supp}(p_t) : p_s(x) \geq \epsilon$, with $\epsilon > 0$. This is also called ϵ-overlap.*

Cortes *et al.* (2010) give generalization bounds for importance-weighted estimates such as (2). These estimates have high variance when the largest ϵ in Assumption 2 is small; if there is no such $\epsilon > 0$, importance weighting is inapplicable without modification.

2.2 Upper bounds on target risk

When the source domain does not provide sufficient support, $\operatorname{supp}(p_t) \not\subseteq \operatorname{supp}(p_s)$, the target risk of a learned hypothesis may not be consistently estimated without further assumptions. However, we may bound the target risk from above and minimize this bound.

Ben-David *et al.* (2007) introduced the $\mathcal{H}\Delta\mathcal{H}$-distance to measure the worst-case loss from extrapolating between domains using binary hypotheses in a class $\mathcal{H}$. Let $R_p^\ell(h, h') \coloneqq \mathbb{E}_{x\sim p}[\ell(h(x), h'(x))]$ denote the expected disagreement between two hypotheses h, h'. Then, the $\mathcal{H}\Delta\mathcal{H}$-distance between p_s and p_t is[1]

$$d_{\mathcal{H}\Delta\mathcal{H}}(p_s, p_t) \coloneqq \sup_{h,h'\in\mathcal{H}} |R_s(h, h') - R_t(h, h')| \ . \tag{3}$$

By reducing the model class $\mathcal{H}$, the potential disagreement $d_{\mathcal{H}\Delta\mathcal{H}}$ between member functions may be

[1] The definition is sometimes given with a factor 2.

reduced—as well as the capacity of $\mathcal{H}$ to predict the label Y. The best-in-class joint hypothesis risk is

$$\lambda_{\mathcal{H}} \coloneqq \inf_{h \in \mathcal{H}} [R_s(h) + R_t(h)] \ . \tag{4}$$

These quantities lead to the following bound by applying the triangle-inequality of classification error.

Theorem 1 (Adaptation bound by Ben-David *et al.* (2010a))**.** *Under Assumption 1, for all $h \in \mathcal{H}$,*

$$R_t(h) \leq R_s(h) + d_{\mathcal{H}\Delta\mathcal{H}}(p_s, p_t) + \lambda_{\mathcal{H}} \ . \tag{5}$$

The result (5) may be bounded further based on the risk on a sample $(x_1, y_1), \dots, (x_n, y_n) \sim p_s(x, y)$ from the source domain, and an empirical estimate of $d_{\mathcal{H}\Delta\mathcal{H}}(p_s, p_t)$ (Ben-David *et al.*, 2010a). Similar results have also been obtained for continuous labels (Mansour *et al.*, 2009; Cortes & Mohri, 2011).

2.3 Domain-invariant representation learning

Theorem 1, and a suite of follow-up work, have been used to justify algorithms based on learning *domain-invariant representations*—transformations of features such that the source and target domains are approximately indistinguishable in the transformed space (Ben-David *et al.*, 2007). We describe these below.

Let a random variable $Z \in \mathcal{Z}$ be a representation of the input features X, parameterized by a *deterministic* function $\phi(X) =: Z$ with $\phi \in \mathcal{G} \subset \{\mathcal{X} \to \mathcal{Z}\}$. Hypotheses $h \in \mathcal{H}$ for Y are formed by compositions $h = f \circ \phi$ with prediction functions $f \in \mathcal{F} \subset \{\mathcal{Z} \to \mathcal{Y}\}$ operating in the representation space $\mathcal{Z}$, and $\mathcal{H} \coloneqq \{f \circ \phi : f \in \mathcal{F}, \phi \in \mathcal{G}\}$. The probability of a set $\mathbf{z} \subseteq \mathcal{Z}$ induced by ϕ is then $p(Z \in \mathbf{z}) = \int_{x \in X} p(X = x) \mathbb{1}[\phi(x) \in \mathbf{z}] dx$. If ϕ does not induce atoms, $p(Z)$ is a density; we consider only this case in the sequel.

We say that a representation $Z_\phi \coloneqq \phi(X)$ of X is *domain-invariant* if $p_s(Z_\phi) = p_t(Z_\phi)$. A common approach to learning approximately domain-invariant representations is to solve the following problem[2].

$$\underset{\phi \in \mathcal{G}, f \in \mathcal{F}}{\text{minimize}} \ \underbrace{\hat{R}_s(f \circ \phi)}_{\text{Source risk}} + \alpha \cdot \underbrace{d(\hat{p}_s(Z_\phi), \hat{p}_t(Z_\phi))}_{\text{Domain variance in } Z} \tag{6}$$

Here, $\hat{p}_s, \hat{p}_t$ denote empirical distributions of p_s and p_t, d is a distance function on densities, and α is a hyperparameter. In the next section, we describe several instantiations of (6).

3 Related work

Domain adaptation has been studied primarily under Assumption 2 (covariate shift) (Pan *et al.*, 2010), which is the setting also of this work. However, prediction under shift in the target $p(Y)$ and conditional $p(X \mid Y)$ has also been considered (Zhang *et al.*, 2013; Gong *et al.*, 2016; Lipton *et al.*, 2018). A common approach in both settings is to learn representations or projection of observed data that is invariant to the shift in question by minimizing adversarial losses (Ganin & Lempitsky, 2015; Bousmalis *et al.*, 2016; Tzeng *et al.*, 2017), integral probability metrics such as the maximum mean discrepancy (MMD) (Pan *et al.*, 2011; Long *et al.*, 2015, 2016; Baktashmotlagh *et al.*, 2013) and the Wasserstein distance (Shalit *et al.*, 2016; Courty *et al.*, 2017), or other divergences (Berisha *et al.*, 2016; Si *et al.*, 2010; Muandet *et al.*, 2013).

Many recent methods attempt to solve domain adaptation under covariate shift by optimizing objectives similar to (6) (Ganin & Lempitsky, 2015; Long *et al.*, 2015; Bousmalis *et al.*, 2016), with the distance d chosen to be a metric such that $d(p, q) = 0$ iff $p = q$. However, (Gong *et al.*, 2016) point out that it is not clear under what conditions $p_s(\phi(X)) \approx p_t(\phi(X))$ would imply $p_s(Y \mid \phi(X)) \approx p_t(Y \mid \phi(X))$. Ben-David *et al.* (2010b) showed that Assumption 1 (covariate shift) and small $d_{\mathcal{H}\Delta\mathcal{H}}$ are not sufficient on their own to identify Y. On the other hand, Ben-David & Urner (2012) showed that Assumption 2 (sufficient support) is sufficient by counterexample through a reduction of the Left-Right problem (Kelly *et al.*, 2010). Ben-David & Urner (2014) subsequently gave both upper and lower learning bounds for nearest-neighbor learners under Assumptions 2 and so-called probabilistic Lipschitzness.

Next, we argue that that searching for a representation Z such that $p_s(Z) \approx p_t(Z)$ is often undesirable, and that learning objectives in the style of (6) are insensitive to information lost in domain-invariant representations.

4 Limitations of domain-invariant representation learning

In this section, we give concrete examples of the failure modes of domain-invariant representation learning, and propose a shift in focus for future research.

4.1 Representation-induced adaptation error

When features X are high dimensional, they often contain information that is redundant or irrelevant for predicting the label Y but distinguishes the source and target domains. The adaptation bounds reviewed in the previous section suggest that removing such information may reduce the difference between source and target risk by making domains closer in density. However, doing so may also introduce an unobservable error, as we see this in the following example.

[2]We leave out additional regularization of ϕ and f.

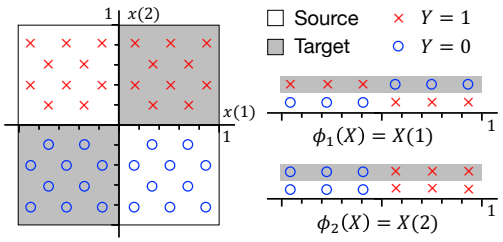


Figure 1: Illustration of Example 1 in which there are two optimal solutions to (6) with objective value 0 but with radically different target risk.

Example 1 (Variable selection)**.** Let $\mathcal{X} = [-1, 1]^2$ with $p_t(x) = .5$ if x is in the lower left or upper right quadrants, $\{[0, 1]\times[0, 1]\cup[-1, 0]\times[-1, 0]\}$, and $p_s(x) = .5$ if x is in the upper left or lower right quadrants, $\{[-1, 0] \times [0, 1] \cup [0, 1] \times [-1, 0]\}$. Further, let $Y = 1$ if $x(2) > 0$, and 0 otherwise (see Figure 1). Now, let $\mathcal{G}$ be the set of variable selections from $\mathbb{R}^2$ to $\mathbb{R}$ and let $\mathcal{F}$ be the set of threshold functions in $\mathbb{R}$. Then, for either selection of a single variable, $\phi_1(x) = x(1)$ or $\phi_2(x) = x(2)$, with $Z = \phi_i(X)$ we have that $p_s(Z) = p_t(Z)$ and $d_{\mathcal{F}\Delta\mathcal{F}}(p_s(Z), p_t(Z)) = 0$, and the function $f(z) = \mathbb{1}[z > 0]$ has $R_s(f \circ \phi_1) = R_s(f \circ \phi_2) = 0$. However, $R_t(f \circ \phi_2) = 0$, but $\forall f \in \mathcal{F} : R_t(f \circ \phi_2) \geq 1$. Hence, objective (6) is uninformative of the target risk.

Example 1 illustrates the impossibility of domain adaptation without overlap or other additional assumptions. Based on the observed data, there is nothing that distinguishes a failure case with maximum target risk from a successful case with minimal target risk. This is true even despite the fact that the problem satisfies the following strong condition.

Assumption 3 (Optimal domain-invariant representation)**.** *There exist a representation $\phi \in \mathcal{G}$ and $f \in \mathcal{F}$ such that $\forall x \in \operatorname{supp}_X(p_s) \cup \operatorname{supp}_X(p_t) : f(\phi(x)) = p(Y \mid X = x)$ and $p_s(\phi(X)) = p_t(\phi(X))$.*

Assumption 3 is by no means guaranteed to hold in practice. Often, variables that are distributed differently across domains are critical for prediction. Regardless, Assumption 3 is *necessary* for domain-invariant representation learning to be consistent. However, as strong as this assumption is, it is not *sufficient* for consistent domain adaptation—not with domain-invariant representations nor with any other method.

Even in problems that are possible to solve consistently, a learned representation may be more predictive on the source domain than the target domain. To reason about this case, we must apply Theorem 1 to the hypothesis space $\mathcal{H}_\phi = \{f \circ \phi : f \in \mathcal{F}\}$ induced by the representation ϕ. Then, for all $f \in \mathcal{F}$,

$$R_t(f\circ\phi) \leq R_s(f\circ\phi)+d_{\mathcal{F}\Delta\mathcal{F}}(p_s(Z), p_t(Z))+\lambda_{\mathcal{H}_\phi}. \quad (7)$$

Here, $R_s(f \circ \phi)$ and $d_{\mathcal{F}\Delta\mathcal{F}}(p_s(Z), p_t(Z))$ may be bounded and minimized but, in contrast, $\lambda_{\mathcal{H}_\phi}$ *is unobserved and may increase when solving* (6).

Proposition 1. *For all $\phi \in \mathcal{G}, f \in \mathcal{F}$ as defined above, we have with $Z = \phi(X)$ and $\mathcal{H}_\phi = \{f \circ \phi : f \in \mathcal{F}\}$*

$$d_{\mathcal{F}\Delta\mathcal{F}}(p_s(Z), p_t(Z)) \leq d_{\mathcal{H}\Delta\mathcal{H}}(p_s(X), p_t(X)) \quad (8)$$

$$\lambda_{\mathcal{H}_\phi} \geq \lambda_{\mathcal{H}} \ . \quad (9)$$

Proof. The results follow immediately from the definitions of $d_{\mathcal{H}\Delta\mathcal{H}}$ and $\lambda_{\mathcal{H}}$, that $d_{\mathcal{F}\Delta\mathcal{F}}(p_s(Z), p_t(Z)) = d_{\mathcal{H}_\phi\Delta\mathcal{H}_\phi}(p_s(X), p_t(X))$, and that $\mathcal{H}_\phi \subseteq \mathcal{H}$. □

As a result of Proposition 1, solving (6) implies neither minimization of the RHS of (5) or (7). One interpretation of this result, and of Example 1, is that covariate shift (Assumption 1) need not hold with respect to the representation $Z = \phi(X)$, even if it does with respect to X. With $\phi^{-1}(z) = \{x : \phi(x) = z\}$,

$$p_t(Y \mid z) = \frac{\int_{x\in\phi^{-1}(z)} p(Y \mid x)p_t(x)dx}{\int_{x\in\phi^{-1}(z)} p_t(x)dx} \neq p_s(Y \mid z) \ .$$

Equality holds for general p_s, p_t only if ϕ is invertible. In Section 5, we define a quantity that measures the effect of this discrepancy and how it relates to invertibility, and use it to bound the target risk.

We summarize this section in a statement inspired by Lemma 2 in Bareinboim & Pearl (2013).

> If there are two distinct hypotheses h, h' for the label Y that are both consistent with $p_s(X, Y)$ and $p_t(X)$ and a set of assumptions $\mathcal{A}$, but result in different predictions on $p_t(X)$, Y is not identifiable over $p_t(X)$.

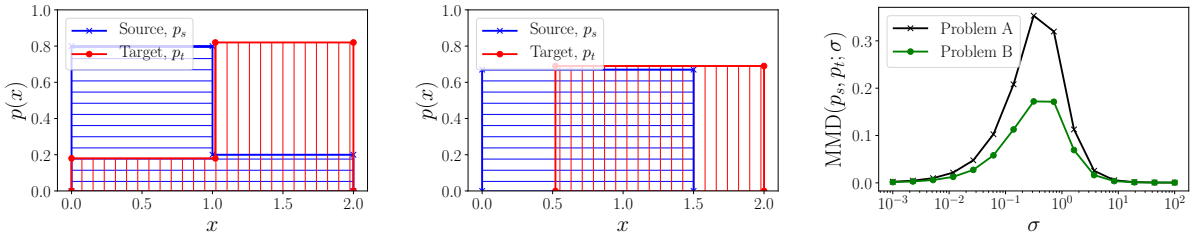
Like causal inference (Pearl, 2009), successful domain adaptation is often entirely reliant on making appropriate assumptions about unobservable quantities.

4.2 The cost of domain invariance

A desired property of adaptation bounds is that they are as tight as possible when Assumption 2 (sufficient support) holds, since consistent estimation is possible in this setting (Ben-David & Urner, 2012)[3]. However, bounds based on Theorem 1 do not always have this property, and their looseness is often independent of the observed risk on the source domain. We give an example of how unintuitive this can be below.

Example 2. We illustrate two examples of source and target densities in Figure 2 along with the estimated maximum mean discrepancy (MMD) (Gretton *et al.*, 2012) between domains for a Gaussian RBF-kernel

[3]By "consistency", we refer to the convergence of an estimate to a quantity of interest given enough samples.



(a) Problem A (b) Problem B (c) MMD for varying bandwidth, σ

Figure 2: Examples illustrating the counter-intuitive effects of using density distance metrics for regularizing domain adaptation methods. Despite the fact that sufficient support is satisfied in Problem A, typical adaptation bounds (using e.g. the RBF-kernel MMD, see (c)) are smaller for Problem B than for Problem A. In contrast, our proposed support sufficiency divergence with $\epsilon = 0.2$ (see Section 5) is 0 in Problem A and 0.33 for Problem B.

with varying bandwidth σ. The MMD has been used to bound $d_{\mathcal{H}\Delta\mathcal{H}}$ and the target risk in Gretton *et al.* (2009); Long *et al.* (2015); Pan *et al.* (2011); Gong *et al.* (2016), among others. Despite there being a significant *lack of overlap* between the support of source and target domains in Problem B, the MMD is smaller than in Problem A, in which the support of the target domain is completely covered by the source density. However, Problem A satisfies sufficient assumptions for identifiability, whereas Problem B does not. This illustrates a drawback of representation learning methods that penalize distributional distance between domains.

The problem illustrated in Example 2 has practical consequences, as we see in Section 6.2. When label marginal distributions differ in a classification task, but domains partially overlap, requiring domain invariance is often too strict. In fact, in our examples, training using only source labels often does better than domain-invariant representation learning.

5 A new support-based bound

We proceed to bound the target risk of an hypothesis in terms of its error on the source domain and the expected lack of sufficient support. This bound is aimed at overcoming limitations of existing bounds by a) explicitly characterizing the risk induced by non-invertible representations and b) avoiding unnecessary side effects of domain invariance.

We say that there is lack of *sufficient support* at a point x if the target density is larger than the source density and the source density is small, as defined by $\delta_{p,q}(x)$,

$$\delta_{p,q}(x) = \mathbb{1}[q(x) \geq p(x) \;\text{ and }\; p(x) < \epsilon] \,. \tag{10}$$

We let $\delta_{s,t}(x)$ serve as short-hand for $\delta_{p_s,p_t}(x)$. Below, we define the *support sufficiency divergence*.

Definition 1. For distributions, p, q, the *support sufficiency divergence* from p to q is defined by

$$d^{\epsilon}_{\text{supp}}(p \parallel q) := \mathop{\mathbb{E}}_{q}[\delta_{p,q}(x)] - \mathop{\mathbb{E}}_{p}[\delta_{p,q}(x)]$$

Note that $d^{\epsilon}_{\text{supp}}$ is not symmetric, but is 0 for $p = q$. Crucially however, it is 0 also when $p \neq q$ for some choices of ϵ, if $\text{supp}(p) = \text{supp}(q)$. Further, it holds that $0 \leq d^{\epsilon}_{\text{supp}} \leq 1$ and the bounds are tight (see Appendix A.1 for a proof).

Our main result builds on the idea that we can expect an hypothesis to be accurate on the target domain in regions where the source density is sufficiently high. First, let $w^{\epsilon}_{p,q}(x)$ be a weighting function such that

$$w^{\epsilon}_{p,q}(x) = \begin{cases} q(x)/p(x) & \text{if } p(x) \geq \epsilon \\ 1 & \text{othewrise} \end{cases} \tag{11}$$

We may state the following result.

Lemma 1. *Let $p_s(x), p_t(x)$ be densities over $\mathcal{X}$. Further, let $\ell : \mathcal{X} \to \mathbb{R}_+$ be a function such that $\exists M > 0 : \forall x \in \mathcal{X} : \ell(x) \in [0, M)$. Then, with $\epsilon > 0$,*

$$\mathbb{E}_{p_t}[\ell(x)] \leq \underbrace{\mathbb{E}_{p_s}\left[w^{\epsilon}_{p_s,p_t}(x)\ell(x)\right]}_{\text{Weighted expectation}} + \underbrace{M \cdot d^{\epsilon}_{\text{supp}}(p_t \parallel p_s)}_{\text{Support discrepancy}} \,.$$

Equality holds if $p_t = p_s$ or if $\forall x \in \text{supp}(p_t) : p_s(x) = \epsilon$. The second term is 0 if and only if Assumption 2 holds with $\epsilon \leq \inf_{x:p_t(x)\geq p_s(x)} p_s(x)$, by definition. The proof can be found in Appendix A.2.

Before we state our main result, we define a measure of the impact of non-invertibility in representations.

Definition 2. Given are domains p_s and p_t, a prediction function $f \in \mathcal{F}$, a label Y, a loss ℓ and a representation $Z = \phi(X)$. Let

$$\Delta_{q,p}(x) := \mathop{\mathbb{E}}_{q(y|\phi(x))}[\ell(f(\phi(x)), y)] - \mathop{\mathbb{E}}_{p(y|x)}[\ell(f(\phi(x)), y)]$$

Then, the *excess target information loss* is

$$\eta_\phi^\ell(f, Y) = \mathop{\mathbb{E}}_{p_t(x)} \left[\Delta_{p_t,p}(x) - \Delta_{p_s,p}(x)\right]$$

We say that the information loss induced by the representation $\phi(x)$ is *symmetric* if $\eta_\phi^\ell(f, y) = 0$. Both Δ and η are always 0 for invertible ϕ. Note also that η may be negative, although we don't expect this in practice as we explain later.

By Lemma 1 and Definition 2, we have the following.

Theorem 2. *Consider any feature representation $z = \phi(x)$ with $\phi \in \mathcal{G}$ and prediction function $f \in \mathcal{F}$, and define $h = f \circ \phi$. Further, let $p_s(Z)$ and $p_t(Z)$ be the two distributions induced by the representation ϕ applied to X distributed according to $p_s(X), p_t(X)$. Further, assume that for any hypothesis $h \in \mathcal{H}$ and a loss function ℓ, $\sup_{x\in\mathcal{X},y\in\mathcal{Y},h\in\mathcal{H}}[\ell(h(x), y)] \leq M$. For any $\epsilon > 0$,*

$$R_t(f \circ \phi) \leq \underbrace{\mathbb{E}_{p_s}\left[w^\epsilon_{p_s,p_t}(z)\ell(f(z), y)\right]}_{\text{Observable}} + M \underbrace{d^\epsilon_{\text{supp}}(p_s(z) \parallel p_t(z))}_{\text{Observable}} + \underbrace{\eta_\phi^\ell(f, y)}_{\text{Unobservable}} . \tag{12}$$

Proof sketch. For any $h = f \circ \phi$, we have that

$$E_{p_t(x,y)}[\ell(h(x), y)] = E_{p_t(z,y)}[\ell(f(z), y)] .$$

By adding and subtracting $E_{p_t(z)p_s(y|z)}[\ell(f(z), y)]$,

$$E_{p_t(z,y)}[\ell(f(z), y)] = E_{p_t(z)p_s(y|z)}[\ell(f(z), y)] + E_{p_t(z)p_t(y|z)}[\ell(f(z), y)] - E_{p_t(z)p_s(y|z)}[\ell(f(z), y)]$$

The last two terms equal $\eta_\phi^\ell(h, y)$ as $p_s(y \mid x) = p_t(y \mid x)$ by Assumption 1. Note that the marginal density over z is equal in both of the last terms. The first term may be decomposed by the support of p_s. With $L_s(z) = E_{p_s(y|z)}[\ell(f(z), y) \mid Z = z]$, we get

$$\begin{aligned}(*) &:= E_{p_t(z)p_s(y|z)}[\ell(f(z), y)] \\ &= \int_{z:p_s(z)\geq\epsilon} p_t(z)L_s(z)dz + \int_{z:p_s(z)<\epsilon} p_t(z)L_s(z)dz\end{aligned}$$

Adding and subtracting $\int_{z:p_s(z)<\epsilon} p_s(z)L_s(z)dz$, we get

$$\begin{aligned}(*) &= \mathbb{E}_{p_s}\left[w^\epsilon_{p_s,p_t}(z)\ell(f(z), y)\right] \\ &+ \int_{\substack{z:p_s(z)<\epsilon \\ p_s\leq p_t}} \underbrace{(p_t(z) - p_s(z))L_s(z)}_{\geq 0} dz \\ &+ \underbrace{\int_{\substack{z:p_s(z)<\epsilon \\ p_s> p_t}} (p_t(z) - p_s(z))L_s(z)dz}_{\leq 0} .\end{aligned}$$

Bounding the second term by $d^\epsilon_{\text{supp}}(p_s \parallel p_t)$ and removing the third non-positive term, we obtain the result. For a full proof, see Appendix A.3. □

Theorem 2 is consistent with our intuition that increasing the sufficiency of the support of p_s for p_t leads to better adaptation. If this overlap is increased without losing information, such as through collection of additional samples, this is usually preferable.

Unlike bounds based on the triangle inequality (Ben-David *et al.*, 2010a; Mansour *et al.*, 2009; Cortes & Mohri, 2011), the bound in Theorem 2 is tight when $p_s(X) = p_t(X)$. On the other hand, when the supports of $p_s(Z)$ and $p_t(Z)$ are completely disjoint, the bound is non-informative. In Section 5.1 we obtain a tighter bound for the disjoint case by incorporating additional assumptions. In many problems, however, there is partial overlap, such as under label marginal shift.

For domains with common and bounded support, ϵ may be chosen such that minimizing the bound of Theorem 2 reduces to importance sampling. In fact, we may view Theorem 2 as a middle-ground between importance sampling estimates and upper bounds on the target risk, using importance sampling where feasible. The choice of ϵ in Lemma 2 trades off the sizes of the two middle terms in (12)—small ϵ, larger first term and vice versa. Additionally, if $\ell(x)$ is 0 everywhere on $\text{supp}(p)$, the first term is 0. The choice of ϵ also affects the variance in Monte-Carlo estimates of these terms. If ϵ is close to 0, the weights $w^\epsilon_{p_s,p_t}(z)$ are potentially larger, and variance increases (Cortes *et al.*, 2010).

When ϕ is invertible, $\eta_\phi^\ell(f, y) = 0$ as $p_t(y \mid \phi(x)) = p_t(y \mid x)$. Shalit *et al.* (2016) gave a bound based on integral probability metrics in the style of Theorem 1, with the additional restriction that ϕ is invertible. However, this is a strong restriction as such ϕ cannot increase the sufficiency of support w.r.t. $p_s(z)$ and $p_t(z)$. We conjecture that under appropriate assumptions of smoothness, η is larger for less invertible ϕ. By encouraging ϕ to be near-invertible, this is mitigated. This would serve as justification for reconstruction losses used by for example Bousmalis *et al.* (2016). Alternatively, $\eta_\phi^\ell(f, y) = 0$ if any information lost in ϕ is equally important for predicting labels in the source domain as in the target. If $\eta = 0$ is always true, Assumption 3 is sufficient for identification of the label.

5.1 Incorporating assumptions on the loss

In Theorem 2, the loss at points outside of the overlap between domains is bounded from above by a constant, M. As a result, the bound is uninformative for disjoint domains. If prior knowledge about the label function is available, we may address this by making assumptions about how the label function extrapolates, akin to Theorem 1. Below, we give an alternative bound based on an assumption that the loss $\ell(f \circ \phi)$ of hypotheses using a representation ϕ belongs to a known

(a) "0" in MNIST (b) "0" in MNIST-M

Figure 3: A benchmark for domain adaptation: MNIST→MNIST-M (Ganin & Lempitsky, 2015).

family $\mathcal{L}$. Critically, this new bound remains qualitatively different from previous work as a) it penalizes extrapolation between domains *only in regions where the source density is low* and b) it explicitly characterizes the excess target risk due to information lost in the learned representation.

Definition 3. We define the *integral probability metric (IPM) support sufficiency divergence* between densities p, q on $\mathcal{X}$ with respect to a class of functions $\mathcal{L}$ by

$$d^{\mathcal{L},\epsilon}_{\text{supp}}(p \parallel q) \coloneqq \sup_{\ell \in \mathcal{L}} \left| \mathop{\mathbb{E}}_{q}[\delta_p(x)\ell(x)] - \mathop{\mathbb{E}}_{p}[\delta_p(x)\ell(x)] \right| \quad (13)$$

where $\delta_p(x) = \mathbb{1}[p(x) < \epsilon]$.

Theorem 3. *Assume that for any representation $\phi \in \mathcal{G}$, and any $f \in \mathcal{F}$, $\mathbb{E}_{p_s(y|\phi(x))}[\ell f(\phi(x), y)] \in \mathcal{L}$. Under the conditions of Theorem 2, we have*

$$R_t(f \circ \phi) \leq \underbrace{\mathbb{E}_{p_s}\left[w^{\epsilon}_{p_s,p_t}(z)\ell(f(z),y)\right]}_{\text{Observable}} + \underbrace{d^{\mathcal{L},\epsilon}_{\text{supp}}(p_s(z) \parallel p_t(z))}_{\text{Observable}} + \underbrace{\eta^{\ell}_{\phi}(f,y)}_{\text{Unobservable}} .$$

Remark 1. Theorem 3 provides a tighter bound than Theorem 2 at the cost of stronger assumptions. With $M \geq \sup_{x\in\mathcal{X},y\in\mathcal{Y},\ell\in\mathcal{L}} \ell(h(x),y)$, and $\ell > 0$,

$$d^{\mathcal{L},\epsilon}_{\text{supp}}(p \parallel q) \leq M \max\{d^{\epsilon}_{\text{supp}}(p \parallel q), d^{\epsilon}_{\text{supp}}(q \parallel p)\} \leq M.$$

For the first inequality to be tight, the maximizer ℓ^* of (13) must be flexible enough to always be equal to M when $q > p$ and always equal to 0 when $q < p$. This is unlikely to be true of the actual loss when the supports of p and q overlap. Instead, it is common to assume that ℓ^* obeys some smoothness conditions. In Appendix A.4 we show how $d^{\mathcal{L},\epsilon}_{\text{supp}}(p \parallel q)$ may be estimated using kernel evaluations if $\mathcal{L}$ is a reproducing-kernel Hilbert space, following Gretton *et al.* (2012).

6 Empirical results

We revisit previous empirical results in light of our theoretical findings with emphasis on Domain-Adversarial Neural Networks (DANN) by (Ganin *et al.*, 2016)[4].

6.1 Plausibility of sufficient assumptions

The most common benchmarks for domain adaptation algorithms are computer vision and natural language processing tasks. One example is the MNIST → MNIST-M task (Ganin & Lempitsky, 2015), in which the goal is to learn to classify handwritten digits overlayed with random photographs (MNIST-M) based on labeled images of digits alone (MNIST) (LeCun *et al.*, 1998) (see Figure 3). For this task, we can immediately rule out Assumption 2 of sufficient support, as MNIST-M images are full-color images that have measure 0 in MNIST. Still, previous work have achieved target accuracy of $> 55\%$ when training on source data alone, and $> 80\%$ when using unlabeled target data, compared to $> 95\%$ when using labeled target data (Ganin & Lempitsky, 2015; Bousmalis *et al.*, 2016). These results support Assumption 3—that there exists a domain-invariant representation in which the labeling function is approximately realizable.

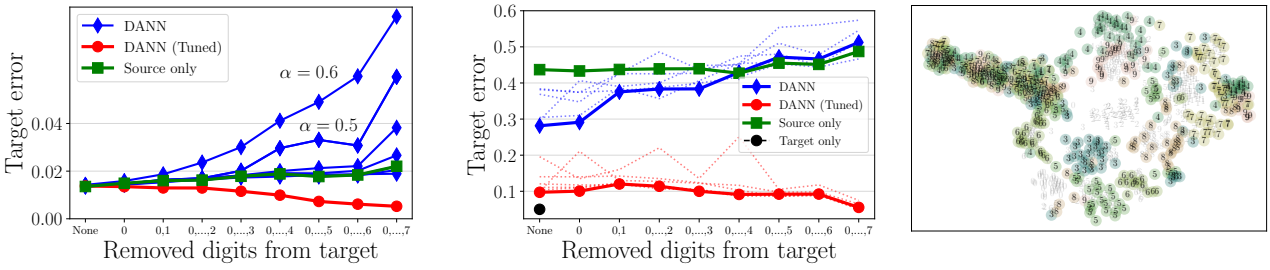
6.2 Contrasting support and domain variance

When label marginal distributions differ under covariate shift, $p_s(Y) \neq p_t(Y), p_s(Y \mid X) = p_t(Y \mid X)$, such as when objects of a certain class appear more often in one domain, a distance between feature marginals, $p_s(X), p_t(X)$, is induced. Authors have studied this restricted setting in detail (Zhang *et al.*, 2013; Lipton *et al.*, 2018). If additionally the target domain is made up of a subset of the source domain, encouraging domain invariance may cause more harm than good. We study a) the performance of DANN models under domain shift with sufficient support, and b) the realizability of the label in the learned representation.

We create a task in which the source domain is the standard MNIST dataset and the target domain is a version of MNIST for which domain shift is induced by successively removing digit classes from the support of the target domain, leaving the source domain fixed. In this setup, the support of the target domain is contained in the source domain, and empirical risk minimization based on source data alone should be a good baseline. We compare to the case where the target is replaced by MNIST-M, but perturbed in the same way.

The DANN model optimizes (6), with d an adversarial neural network classifying images by domain, and α a hyperparameter controlling the strength of this penalty in the objective $O = 2(1 - |\alpha - 0.5|)((1-\alpha)\hat{R}_s + \alpha d)$. In this way, we interpolate between empirical risk

[4]Our implementation is based on that of `https://github.com/pumpikano/tf-dann`



(a) MNIST→MNIST (b) MNIST→MNIST-M (c) MNIST→MNIST-M$\backslash\{0,1,2\}$

Figure 4: Left: Target error as a function of marginal label distribution. For each setting, a DANN model is trained on unlabeled target data and labeled source data. We compare the accuracy of this model to a model tuned on target labels but with a fixed representation given by the first model. Different lines of the same color indicate different values of penalty strength $\alpha \in \{0.0, ..., 0.6\}$. Right: Embeddings learned by DANN with equal (top) and unequal (bottom) label marginal distributions. In MNIST-M$\backslash\{0,1,2\}$, all images of digits 0,1,2 have been removed. Grey digits are from the source domain and black digits from the target domain.

minimization ($\alpha = 0$), the standard DANN formulation ($\alpha = 0.5$) and prioritizing domain invariance ($\alpha > 0.5$). We compare the error of two different models: 1) The standard DANN estimator $h_{dann} = f_{dann} \circ \phi_{dann}$, and 2) A model $h_{tuned} = f_{tuned} \circ \phi_{dann}$ in which the learned representation ϕ_{dann} from 1) is fixed and the prediction function f is fit to the *target labels* (Tuned). The latter serves to give an upper bound on best-case risk when predicting from the representations learned by DANN.

In Figure 4, we observe that models trained without target supervision (DANN) perform steadily worse on MNIST→MNIST, the more the label marginal distribution is perturbed. This holds also for MNIST→MNIST-M, where sufficient support is not satisfied. There, DANN is beneficial for small label shift, but eventually does no better than a model trained using only source data. Learning with a domain-adversarial loss appears to have little impact on the *realizability* of the target label in the representation; the target-tuned models achieve almost as good performance as the fully target-trained lower bound. In Figure 4c, we see that the embeddings learned using DANN models under label marginal shift show worse separation between classes, than the embeddings learned under equal label marginal distributions (see Appendix C).

7 Discussion

We have studied algorithms for unsupervised domain adaptation based on domain-invariant representation learning and the theoretical arguments used to support them. We find that, despite empirical success, the theoretical justification of these algorithms is flawed in that oft-cited generalization bounds are not minimized by the learned representations. In particular, the literature has failed to characterize conditions under which domain-invariant representations lead to consistent estimation. We have found through examples and experiments on domain adaptation benchmarks that domain invariance is often too strong a requirement for learning, both when there is overlap between domains and when there is not. This stems from the fact that overlapping support is sufficient for domain transfer, and equality in densities is not necessary.

We have proposed alternative bounds that measure distance in support instead of density and that explicitly recognize loss incurred by non-invertible representations. Our bounds suggest several ways to design new algorithms. First, minimizing the second term in our bound, the support sufficiency divergence, may be achieved by replacing indicator functions by hinge losses (see Appendix B). This increases the looseness of the bound, but makes its derivative informative. In the same spirit, we may design new heuristics that regularize representations only in points at which the source density is much smaller than the target density. Second, while the excess adaptation error induced by learning non-invertible transformations is unobservable, it is associated with the information loss of the representation. To avoid this, we may attempt to maintain a small excess by imposing a reconstruction loss on the representation, similar to Bousmalis *et al.* (2016).

Acknowledgements

We thank Zach Lipton, Alexander D'Amour, Christina X Ji and Hunter Lang for insightful feedback. This work was supported in part by Office of Naval Research Award No. N00014-17-1-2791 and the MIT-IBM Watson AI Lab.

References

Baktashmotlagh, M., Harandi, M.T., Lovell, B.C. & Salzmann, M. (2013). Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision*, 769–776.

Bareinboim, E. & Pearl, J. (2013). A general algorithm for deciding transportability of experimental results. *Journal of causal Inference*, **1**, 107–134.

Ben-David, S. & Urner, R. (2012). On the hardness of domain adaptation and the utility of unlabeled target samples. In *International Conference on Algorithmic Learning Theory*, 139–153, Springer.

Ben-David, S. & Urner, R. (2014). Domain adaptation–can quantity compensate for quality? *Annals of Mathematics and Artificial Intelligence*, **70**, 185–202.

Ben-David, S., Blitzer, J., Crammer, K. & Pereira, F. (2007). Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, 137–144.

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

Ben-David, S., Lu, T., Luu, T. & Pál, D. (2010b). Impossibility theorems for domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, 129–136.

Berisha, V., Wisler, A., Hero, A.O. & Spanias, A. (2016). Empirically estimable classification bounds based on a nonparametric divergence measure. *IEEE Transactions on Signal Processing*, **64**, 580–591.

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

Cortes, C. & Mohri, M. (2011). Domain adaptation in regression. In *International Conference on Algorithmic Learning Theory*, 308–323, Springer.

Cortes, C., Mansour, Y. & Mohri, M. (2010). Learning bounds for importance weighting. In *Advances in neural information processing systems*, 442–450.

Courty, N., Flamary, R., Habrard, A. & Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, 3733–3742.

Ganin, Y. & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 1180–1189.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M. & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, **17**, 2096–2030.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C. & Schölkopf, B. (2016). Domain adaptation with conditional transferable components. In *International Conference on Machine Learning*, 2839–2848.

Gretton, A., Smola, A.J., Huang, J., Schmittfull, M., Borgwardt, K.M. & Schölkopf, B. (2009). Covariate shift by kernel mean matching.

Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B. & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, **13**, 723–773.

Kelly, B.G., Tularak, T., Wagner, A.B. & Viswanath, P. (2010). Universal hypothesis testing in the learning-limited regime. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, 1478–1482, IEEE.

LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**, 2278–2324.

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

Long, M., Cao, Y., Wang, J. & Jordan, M.I. (2015). Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*.

Long, M., Zhu, H., Wang, J. & Jordan, M.I. (2016). Deep transfer learning with joint adaptation networks. *arXiv preprint arXiv:1605.06636*.

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

Muandet, K., Balduzzi, D. & Schölkopf, B. (2013). Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, 10–18.

Pan, S.J., Yang, Q. *et al.* (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, **22**, 1345–1359.

Pan, S.J., Tsang, I.W., Kwok, J.T. & Yang, Q. (2011). Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, **22**, 199–210.

Pearl, J. (2009). *Causality*. Cambridge university press.

Shalit, U., Johansson, F. & Sontag, D. (2016). Estimating individual treatment effect: generalization bounds and algorithms. *arXiv preprint arXiv:1606.03976*.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood

function. *Journal of statistical planning and inference*, **90**, 227–244.

Si, S., Tao, D. & Geng, B. (2010). Bregman divergence-based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, **22**, 929.

Tzeng, E., Hoffman, J., Saenko, K. & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 4.

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