Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)
Sister Conferences Best Papers Track

Robust Domain Adaptation: Representations, Weights and Inductive Bias (Extended Abstract)

Victor Bouvier[1,2*], **Philippe Very**[3], **Clément Chastagnol**[4], **Myriam Tami**[1] and **Céline Hudelot**[1]
[1]Université Paris-Saclay (CentraleSupélec)
[2]Sidetrade
[3]Lend-Rx
[4]Alan
vbouvier@sidetrade.com, philippe.very@lend-rxtech.com, clement.chastagnol@alan.eu,
myriam.tami@centralesupelec.com, celine.hudelot@centralesupelec.com

Abstract

Domain Invariant Representations (IR) has improved drastically the transferability of representations from a labelled source domain to a new and unlabelled target domain. Unsupervised Domain Adaptation (UDA) in presence of *label shift* remains an open problem. To this purpose, we present a bound of the target risk which incorporates both weights and invariant representations. Our theoretical analysis highlights the role of inductive bias in aligning distributions across domains. We illustrate it on standard benchmarks by proposing a new learning procedure for UDA. We observed empirically that weak inductive bias makes adaptation robust to label shift. The elaboration of stronger inductive bias is a promising direction for new UDA algorithms.

1 Introduction

Deploying machine learning models in the real world often requires the ability to generalize to *unseen samples i.e.* samples significantly different from those seen during learning. Domain Adaptation (DA) [Quionero-Candela *et al.*, 2009; Pan and Yang, 2009] is a well-studied approach to bridge the gap between a *source* and a *target* distributions, respectively noted $p_S(x, y)$ and $p_T(x, y)$ where x are inputs and y are labels. Unsupervised Domain Adaptation (UDA) assumes that only unlabelled data from the target domain is available during training. In this context, a natural assumption, named *Covariate shift* [Shimodaira, 2000; Huang *et al.*, 2007], consists in assuming that the mapping from the inputs to the labels is conserved across domains, *i.e.* $p_T(y|x) = p_S(y|x)$. In this context, *Importance Sampling* (IS) performs adaptation by weighting the contribution of sample x in the loss by $w(x) = p_T(x)/p_S(x)$ [Quionero-Candela *et al.*, 2009]. Although IS seems natural when unlabelled data from the target domain is available, the covariate shift assumption is not sufficient to guarantee successful adaptation [Ben-David *et al.*, 2007]. Moreover, for high dimensional data [D'Amour *et al.*, 2017] such as texts or images, the shift between $p_S(x)$ and $p_T(x)$ results from non-overlapping supports leading to unbounded weights [Johansson *et al.*, 2019].

In this particular context, representations can help to reconcile non-overlapping supports [Ben-David *et al.*, 2007] by learning a so-called *Domain Invariant Representation* [Ganin and Lempitsky, 2015];

$$p_S(z) \approx p_T(z) \tag{1}$$

where $z := \varphi(x)$ for a given non-linear representation φ. These assume that the *transferability* of representations, defined as the combined error of an ideal classifier, remains low during learning. Unfortunately, this quantity involves target labels and is thus intractable. More importantly, looking for strict invariant representations ($p_S(z) = p_T(z)$) hurts the transferability of representations [Johansson *et al.*, 2019; Liu *et al.*, 2019; Wu *et al.*, 2019; Zhao *et al.*, 2019]. In particular, there is a fundamental trade-off between learning invariant representations and preserving transferability in presence of label shift ($p_T(y) \neq p_S(y)$) [Zhao *et al.*, 2019]. To mitigate this trade-off, some recent works suggest to relax domain invariance by weighting samples [Cao *et al.*, 2018a; Wu *et al.*, 2019; You *et al.*, 2019; Cao *et al.*, 2018b]. This strategy aligns a *weighted source* distribution with the target distribution

$$w(z)p_S(z) \approx p_T(z), \tag{2}$$

for some weights $w(z)$. We now have two tools, w and φ, which need to be calibrated to obtain distribution alignment. Which one should be promoted? How weights preserve good transferability of representations?

In this paper, we show that weights allow to design an interpretable generalization bound where transferability and invariance errors are uncoupled. In addition, we discuss the role of inductive design for both the classifier and the weights in addressing the lack of labelled data in the target domain. From these theoretical insights, we derive a new learning procedure for UDA that minimizes the transferability error while controlling representation invariance with weights. We provide an empirical illustration of our framework on two DA benchmarks (**Digits** and **Office31** datasets). We stress-test our learning scheme by modifying strongly the label distribution in the source domain. While methods based on invariant representations deteriorate considerably in this context, our procedure remains robust.

*Contact Author

2 Preliminaries

For two random variables (X, Y) on a given space $\mathcal{X} \times \mathcal{Y}$, we introduce two distributions: the source distribution $p_S(x, y)$ and the target distribution $p_T(x, y)$. Here, labels are one-hot encoded *i.e.* $y \in [0, 1]^C$ such that $\sum_c y_c = 1$ where C is the number of classes. We use the index notation S and T to differentiate source and target terms. We define the hypothesis class $\mathcal{H}$ as a subset of functions from $\mathcal{X}$ to $\mathcal{Y}$ which is the composition of a representation class Φ and a classifier class $\mathcal{G}$, *i.e.* $\mathcal{H} = \mathcal{G} \circ \Phi$. For the ease of reading, given a classifier $g \in \mathcal{G}$ and a representation $\varphi \in \Phi$, we note $g\varphi := g \circ \varphi$. Furthermore, in the definition $z := \varphi(x)$, we refer indifferently to z, φ, $Z := \varphi(X)$ as the *representation*. For two given h and $h' \in \mathcal{H}$ and ℓ the L^2 loss $\ell(y, y') = ||y - y'||^2$, the risk in domain $D \in \{S, T\}$ is noted:

$$\varepsilon_D(h) := \mathbb{E}_D[\ell(h(X), Y)] \tag{3}$$

and $\varepsilon_D(h, h') := \mathbb{E}_D[\ell(h(X), h'(X))]$. In the seminal works [Ben-David *et al.*, 2007; Mansour *et al.*, 2009], a theoretical limit of the target risk when using a representation φ has been derived:

Bound 1 (Ben David et al.) *For $\varphi \in \Phi$ and $g \in \mathcal{G}$*

$$\varepsilon_T(g\varphi) \leq \varepsilon_S(g\varphi) + d_{\mathcal{G}}(\varphi) + \lambda_{\mathcal{G}}(\varphi) \tag{4}$$

where $d_{\mathcal{G}}(\varphi) = \sup_{g,g' \in \mathcal{G}} |\varepsilon_S(g\varphi, g'\varphi) - \varepsilon_T(g\varphi, g'\varphi)|$ and $\lambda_{\mathcal{G}}(\varphi) = \inf_{g \in \mathcal{G}} \{\varepsilon_S(g\varphi) + \varepsilon_T(g\varphi)\}$.

This generalization bound ensures that the target risk $\varepsilon_T(g\varphi)$ is bounded by the sum of the source risk $\varepsilon_S(g\varphi)$, the disagreement risk between two classifiers from representations $d_{\mathcal{G}}(\varphi)$, and a third term, $\lambda_{\mathcal{G}}(\varphi)$, which quantifies the ability to perform well in both domains from representations. The latter is referred to as the *adaptability* error of representations. It is intractable in practice since it involves labels from the target distribution. Promoting distribution invariance of representations, *i.e.* $p_S(z)$ close to $p_T(z)$, results on a low $d_{\mathcal{G}}(\varphi)$. However, it induces an unexpected trade-off when learning domain invariant representations [Johansson *et al.*, 2019; Zhao *et al.*, 2019]:

Proposition 1 (Invariance hurts adaptability) *Let ψ be a representation which is a richer feature extractor than φ: $\mathcal{G} \circ \varphi \subset \mathcal{G} \circ \psi$. Then,*

$$d_{\mathcal{G}}(\varphi) \leq d_{\mathcal{G}}(\psi) \text{ while } \lambda_{\mathcal{G}}(\psi) \leq \lambda_{\mathcal{G}}(\varphi) \tag{5}$$

As a result of proposition 1, the benefit of representation invariance must be higher than the loss of adaptability, which is impossible to guarantee in practice.

3 Theory

To overcome the limitation raised in proposition 1, we expose a new bound of the target risk which embeds a new trade-off between invariance and transferability. We show this new bound remains inconsistent with the presence of label shift and we expose the role of weights to address this problem.

3.1 A New Trade-Off Between Invariance and Transferability

We introduce here two important tools that will guide our analysis. They are built upon $\mathcal{F}$ and $\mathcal{F}_C$; two *suitable*[1] classes of critics functions *i.e.*, subset of applications from $\mathcal{Z} \to [-1, 1]$ and $\mathcal{Z} \to [-1, 1]^C$.

- $\mathrm{INV}(\varphi)$, named *invariance error*, that aims at capturing the difference between source and target distribution of representations, corresponding to:

$$\mathrm{INV}(\varphi) := \sup_{f \in \mathcal{F}} \{\mathbb{E}_T[f(Z)] - \mathbb{E}_S[f(Z)]\} \tag{6}$$

- $\mathrm{TSF}(\varphi)$, named *transferability error*, that is dedicated to control if aligned representations have the same labels across domains. For that, we use our class of functions $\mathcal{F}_C$ and we compute the IPM of $Y \cdot \mathbf{f}(Z)$ ($\mathbf{f} \in \mathcal{F}_C$ and $Y \cdot \mathbf{f}(Z)$ is the scalar product) between the source and the target domains:

$$\mathrm{TSF}(\varphi) := \sup_{\mathbf{f} \in \mathcal{F}_C} \{\mathbb{E}_T[Y \cdot \mathbf{f}(Z)] - \mathbb{E}_S[Y \cdot \mathbf{f}(Z)]\} \tag{7}$$

Using $\mathrm{INV}(\varphi)$ and $\mathrm{TSF}(\varphi)$, we can provide a new bound of the target risk:

Bound 2 *For $\varphi \in \Phi$ and $g \in \mathcal{G}$*

$$\varepsilon_T(g\varphi) \leq \varepsilon_S(g\varphi) + 6 \cdot \mathrm{INV}(\varphi) + 2 \cdot \mathrm{TSF}(\varphi) + \varepsilon_T(\mathbf{f}_T\varphi) \tag{8}$$

In contrast with bound 1 (Eq. 5), here two IPMs are involved to compare representations ($\mathrm{INV}(\varphi)$ and $\mathrm{TSF}(\varphi)$). A new term, $\varepsilon_T(\mathbf{f}_T\varphi)$, reflects the level of noise when fitting labels from representations. Bounding the target risk using IPMs has two advantages. First, it allows to better control the invariance / transferability trade-off since $\varepsilon_T(\mathbf{f}_T\varphi) \leq \lambda_{\mathcal{G}}(\varphi)$. This is paid at the cost of $4 \cdot \mathrm{INV}(\varphi) \geq d_{\mathcal{G}}(\varphi)$. Second, $\varepsilon_T(\mathbf{f}_T\varphi)$ is source free and indicates whether there is enough information in representations for learning the task in the target domain at first.

An interesting property of the bound, named tightness, is the case when $\mathrm{INV}(\varphi) = 0$ and $\mathrm{TSF}(\varphi) = 0$ simultaneously. The condition of tightness of the bound provides rich information on the properties of representations.

Proposition 2 *$p_S(y, z) = p_T(y, z)$ if and only if $\mathrm{INV}(\varphi) = \mathrm{TSF}(\varphi) = 0$.*

3.2 Reconciling Weights and Invariant Representations

We propose to adapt the bound by incorporating weights. More precisely, we study the effect of modifying the source distribution $p_S(z)$ to a *weighted source* distribution $w(z)p_S(z)$ where w is a positive function which verifies $\mathbb{E}_S[w(Z)] = 1$. By replacing $p_S(z)$ by $w(z)p_S(z)$ (distribution referred as $w \cdot S$) in bound 2, we obtain a new bound of the target risk:

Bound 3 *For $\varphi \in \Phi$, $g \in \mathcal{G}$ and $w : \mathcal{Z} \to \mathbb{R}^+$ such that $\mathbb{E}_S[w(z)] = 1$:*

$$\varepsilon_T(g\varphi) \leq \varepsilon_{w \cdot S}(g\varphi) + 6 \cdot \mathrm{INV}(w, \varphi) + 2 \cdot \mathrm{TSF}(w, \varphi) + \varepsilon_T(\mathbf{f}_T\varphi) \tag{9}$$

[1] See the paper for details about assumptions on critics.

where $\mathrm{INV}(w,\varphi) := \sup_{f\in\mathcal{F}} \{\mathbb{E}_T[f(Z)] - \mathbb{E}_S[w(Z)f(Z)]\}$
and $\mathrm{TSF}(w,\varphi) := \sup_{\mathbf{f}\in\mathcal{F}_C} \{\mathbb{E}_T[Y\cdot\mathbf{f}(Z)] - \mathbb{E}_S[w(Z)Y\cdot\mathbf{f}(Z)]\}$.

As for the previous bound 2, the property of tightness, *i.e.* when invariance and transferability are null simultaneously, leads to interesting observations:

Proposition 3 $\mathrm{INV}(w,\varphi) = \mathrm{TSF}(w,\varphi) = 0$ *if and only if* $w(z) = \frac{p_T(z)}{p_S(z)}$ *and* $\mathbb{E}_T[Y|Z=z] = \mathbb{E}_S[Y|Z=z]$.

This proposition means that the nullity of invariance error implies distribution alignment: $w(z)p_S(z) = p_T(z)$. This is of strong interest since both representations and weights are involved for achieving domain invariance. The nullity of the transferability error implies that labelling functions are conserved across domains. Furthermore, the equality $\mathbb{E}_T[Y|Z] = \mathbb{E}_S[Y|Z]$ interestingly resonates with a promising line of work [Arjovsky *et al.*, 2019]. Incorporating weights in the bound thus brings two benefits:

1. It raises the inconsistency issue of invariant representations in presence of label shift, as mentioned in section 2. Indeed, tightness is not conflicting with label shift.
2. $\mathrm{TSF}(w,\varphi)$ and $\mathrm{INV}(w,\varphi)$ have two disctinct roles: the former promotes domain invariance of representations while the latter controls whether aligned representations share the same labels across domains.

4 The Role of Inductive Bias

Inductive Bias refers to the set of assumptions which improves generalization, such as a specific neural network architecture or a well-suited regularization. First, we provide a theoretical analysis of the role of inductive bias for addressing the lack of labelling data in the target domain, which is the most challenging part of *Unsupervised* Domain Adaptation. Second, we describe the effect of weights to induce invariance property on representations.

4.1 Inductive Design of a Classifier

General Formulation.

Our strategy consists in approximating target labels error through a classifier $\tilde{g} \in \mathcal{G}$. We refer to the latter as the inductive design of the classifier. Our proposition follows the intuitive idea which states that the best source classifier, $g_S := \arg\min_{g\in\mathcal{G}} \varepsilon_S(g\varphi)$, is not necessarily the best target classifier *i.e.* $g_S \neq \arg\min_{g\in\mathcal{G}} \varepsilon_T(g\varphi)$. For instance, a well-suited regularization in the target domain, noted $\Omega_T(g)$ may improve performance, *i.e.* setting $\tilde{g} := \arg\min_{g\in\mathcal{G}} \varepsilon_S(g\varphi) + \lambda\cdot\Omega_T(g)$ may lead to $\varepsilon_T(\tilde{g}\varphi) \leq \varepsilon_T(g_S\varphi)$. We formalize this idea through the following definition:

Definition 1 (Inductive design of a classifier) *We say that there is an inductive design of a classifier at level* $0 < \beta \leq 1$ *if for any representations* φ*, noting* $g_S = \arg\min_{g\in\mathcal{G}} \varepsilon_S(g\varphi)$*, we can determine* $\tilde{g}$ *such that:*

$$\varepsilon_T(\tilde{g}\varphi) \leq \beta\varepsilon_T(g_S\varphi) \tag{10}$$

We say the inductive design is $\beta-$*strong when* $\beta < 1$ *and weak when* $\beta = 1$.

In this definition, β does not depend of φ, which is a strong assumption, and embodies the strength of the inductive design. The closer to 1 is β, the less improvement we can expect using the inductive classifier $\tilde{g}$. We now study the impact of the inductive design of a classifier in our previous bound 3. Thus, we introduce the approximated transferability error:

$$\widehat{\mathrm{TSF}}(w,\varphi,\tilde{g}) = \sup_{\mathbf{f}\in\mathcal{F}_C} \{\mathbb{E}_T[\tilde{g}(Z)\cdot\mathbf{f}(Z)] - \mathbb{E}_S[w(Z)Y\cdot\mathbf{f}(Z)]\} \tag{11}$$

leading to a bound of the target risk where transferability is target labels free:

Bound 4 (Inductive Bias and Guarantee) *Let* $\varphi \in \Phi$ *and* $w : \mathcal{Z} \to \mathbb{R}^+$ *such that* $\mathbb{E}_S[w(z)] = 1$ *and a* $\beta-$*strong inductive classifier* $\tilde{g}$ *and* $\rho := \frac{\beta}{1-\beta}$ *then:*

$$\varepsilon_T(\tilde{g}\varphi) \leq \rho\{\varepsilon_{w\cdot S}(g_{w\cdot S}\varphi) + 6\cdot\mathrm{INV}(w,\varphi) + 2\cdot\widehat{\mathrm{TSF}}(w,\varphi,\tilde{g}) + \varepsilon_T(\mathbf{f}_T\varphi)\} \tag{12}$$

Here, the target labels are only involved in $\varepsilon_T(\mathbf{f}_T\varphi)$ which reflects the level of noise when fitting labels from representations. Therefore, transferability is now free of target labels. This is an important result since the difficulty of UDA lies in the lack of labelled data in the target domain. It is also interesting to note that the weaker the inductive bias ($\beta \to 1$), the higher the bound and vice versa.

Predicted labels play an important role in UDA. In light of the inductive classifier, this means that $\tilde{g}$ is simply set as $g_{w\cdot S}$. This is a weak inductive design ($\beta = 1$), thus, theoretical guarantee from bound 4 is not applicable. However, there is empirical evidence that showed that predicted labels help in UDA [Grandvalet and Bengio, 2005; Long *et al.*, 2018]. A better understanding of this phenomenon is left for future work. See the paper connections between $\widehat{\mathrm{TSF}}(w,\varphi,\tilde{g})$ and popular approaches of the literature.

4.2 Inductive Design of Weights

While the bounds introduced in the present work involve weights in the representation space, there is an abundant literature that builds weights in order to relax the domain invariance of representations [Cao *et al.*, 2018a; Wu *et al.*, 2019; You *et al.*, 2019; Cao *et al.*, 2018b]. In the rest of the paper we focus on weights

$$w(z) = \frac{p_T(z)}{p_S(z)}. \tag{13}$$

It is worth noting that it controls the invariance error ($\mathrm{INV}(w,\varphi) = 0$). One can quantify the effect of a feature transformation ψ when designing weights;

Proposition 4 (Inductive design of w and invariance) *Let* $\psi : \mathcal{Z} \to \mathcal{Z}'$ *such that* $\mathcal{F}\circ\psi \subset \mathcal{F}$ *and* $\mathcal{F}_C\circ\psi \subset \mathcal{F}_C$*. Let* $w : \mathcal{Z}' \to \mathbb{R}^+$ *such that* $\mathbb{E}_S[w(Z')] = 1$ *and we note* $Z' := \psi(Z)$*. Then,* $\mathrm{INV}(w,\varphi) = \mathrm{TSF}(w,\varphi) = 0$ *if and only if:*

$$w(z') = \frac{p_T(z')}{p_S(z')} \quad \text{and} \quad p_S(z|z') = p_T(z|z') \tag{14}$$

while both $\mathbf{f}_S^\varphi = \mathbf{f}_T^\varphi$ *and* $\mathbf{f}_S^\psi = \mathbf{f}_T^\psi$.

Since we do not leverage any transformation ψ, we frame it as a *weak inductive design*.

Method	Office 31	Digits
DANN	67.8	63.2
CDAN	81.6	73.2
IWAN	75.0	81.6
CDAN_w	81.8	83.2
RUDA (ours)	**83.8**	**86.5**

Table 1: Summary of results of adaptation in context of label shift. RUDA outperforms baselines. See the paper for more details.

5 Towards Robust Domain Adaptation

5.1 Algorithm

We expose a new learning procedure which relies on weak inductive design of both weights and the classifier. This procedure focuses on the transferability error since the inductive design of weights naturally controls the invariance error. Our learning procedure is then a bi-level optimization problem, named RUDA (Robust UDA):

$$\begin{cases} \varphi^\star & = \arg\min_{\varphi\in\Phi} \ \varepsilon_{w(\varphi)\cdot S}(g_{w\cdot S}\varphi) + \lambda\cdot\widehat{\text{TSF}}(w,\varphi,g_{w\cdot S}) \\ & \text{such that } w(\varphi) = \arg\min_{w} \text{INV}(w,\varphi) \end{cases}$$

where $\lambda > 0$ is a trade-off parameter. $\widehat{\text{TSF}}(w,\varphi,g_{w\cdot S})$ and $\text{INV}(w,\varphi)$ are computed in an adversarial manner involving two discriminators. See the paper for more details.

5.2 Experiments

Datasets. We investigate two digits datasets: **MNIST** and **USPS** transfer tasks MNIST to USPS (M→U) and USPS to MNIST (U→M). We used standard train / test split for training and evaluation. **Office-31** is a dataset of images containing objects spread among 31 classes captured from different domains: **Amazon**, **DSLR** camera and a **Webcam** camera. **DSLR** and **Webcam** are very similar domains but images differ by their exposition and their quality. We stress-test our approach by investigating more challenging settings where the label distribution shifts strongly across domains. For the **Digits** dataset, we explore a wide variety of shifts by keeping only 5%, 10%, 15% and 20% of digits between 0 and 5 of the original dataset (refered as $\% \times [0 \sim 5]$). For the **Office-31** dataset, we explore the shift where the object spread in classes 16 to 31 are duplicated 5 times (refered as $5 \times [16 \sim 31]$).

Comparison with the state-of-the-art. For all tasks, we report results from DANN [Ganin and Lempitsky, 2015] and CDAN [Long *et al.*, 2018]. We report IWAN [Zhang *et al.*, 2018], a weighted DANN where weights are learned from a second discriminator, and CDAN_w a weighted CDAN where weights are added in the same setting than RUDA_w. A summary of results is presented in Table 1.

6 Related Works

This paper makes several contributions, both in terms of theory and algorithm. Concerning theory, our bound provides a risk suitable for domain adversarial learning with weighting strategies. Existing theories for non-overlapping supports [Ben-David *et al.*, 2007; Mansour *et al.*, 2009] and importance sampling [Cortes *et al.*, 2010; Quionero-Candela *et al.*, 2009] do not explore the role of representations neither the aspect of adversarial learning. In [Ben-David *et al.*, 2007], analysis of representation is conducted and connections with our work is discussed in the paper. The work [Johansson *et al.*, 2019] is close to ours and introduces a distance which measures support overlap between source and target distributions under covariate shift. Our analysis does not rely on such assumption, its range of application is broader.

Concerning algorithms, the covariate shift adaptation has been well-studied in the literature [Huang *et al.*, 2007; Gretton *et al.*, 2009; Sugiyama *et al.*, 2007]. Importance sampling to address label shift has also been investigated [Storkey, 2009], notably with kernel mean matching [Zhang *et al.*, 2013]. Recently, a scheme for estimating labels distribution ratio with consistency guarantee has been proposed [Lipton *et al.*, 2018]. Learning domain invariant representations has also been investigated in the fold of [Ganin and Lempitsky, 2015; Long *et al.*, 2015] and mainly differs by the metric chosen for comparing distribution of representations.

Using both weights and representations is also an active topic [Cao *et al.*, 2018b; You *et al.*, 2019]. Our work shares strong connections with [Combes *et al.*, 2020] which uses consistent estimation of true labels distribution from [Lipton *et al.*, 2018]. We suggest a very similar empirical evaluation and we also investigate the effect of weights on CDAN loss [Long *et al.*, 2018] with a different weighting scheme since our approach computes weights in the representation space. All these works rely on an assumption at some level, *e.g. Generalized Label Shift* in [Combes *et al.*, 2020], when designing weighting strategies. Our discussion on the role of inductive design of weights may provide a new theoretical support for these approaches.

7 Conclusion

The present work introduces a new bound of the target risk which unifies weights and representations in UDA. We conduct a theoretical analysis of the role of inductive bias when designing both weights and the classifier. In light of this analysis, we propose a new learning procedure which leverages two weak inductive biases, respectively on weights and the classifier. To the best of our knowledge, this procedure is original while being close to straightforward hybridization of existing methods. We illustrate its effectiveness on two benchmarks. The empirical analysis shows that weak inductive bias can make adaptation more robust even when stressed by strong label shift between source and target domains. This work leaves room for in-depth study of stronger inductive bias by providing both theoretical and empirical foundations.

References

[Arjovsky *et al.*, 2019] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[Ben-David *et al.*, 2007] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, 2007.

[Cao *et al.*, 2018a] Yue Cao, Mingsheng Long, and Jianmin Wang. Unsupervised domain adaptation with distribution matching machines. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Cao *et al.*, 2018b] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150, 2018.

[Combes *et al.*, 2020] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoff Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *arXiv preprint arXiv:2003.04475*, 2020.

[Cortes *et al.*, 2010] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.

[D'Amour *et al.*, 2017] Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*, 2017.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015.

[Grandvalet and Bengio, 2005] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005.

[Gretton *et al.*, 2009] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

[Huang *et al.*, 2007] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2007.

[Johansson *et al.*, 2019] Fredrik Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536, 2019.

[Lipton *et al.*, 2018] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pages 3122–3130, 2018.

[Liu *et al.*, 2019] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In *International Conference on Machine Learning*, pages 4013–4022, 2019.

[Long *et al.*, 2015] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, 2015.

[Long *et al.*, 2018] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018.

[Mansour *et al.*, 2009] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *22nd Conference on Learning Theory, COLT 2009*, 2009.

[Pan and Yang, 2009] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[Quionero-Candela *et al.*, 2009] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

[Shimodaira, 2000] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[Storkey, 2009] Amos Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, pages 3–28, 2009.

[Sugiyama *et al.*, 2007] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÃžller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

[Wu *et al.*, 2019] Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. In *International Conference on Machine Learning*, pages 6872–6881, 2019.

[You *et al.*, 2019] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[Zhang *et al.*, 2013] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827, 2013.

[Zhang *et al.*, 2018] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8156–8164, 2018.

[Zhao *et al.*, 2019] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532, 2019.