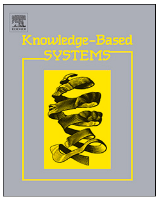
Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



# A novel domain adaptation theory with Jensen–Shannon divergence

Changjian Shui [a], Qi Chen [b], Jun Wen [c], Fan Zhou [b], Christian Gagné [a], Boyu Wang [d,*]

[a] Department of Electrical and Computer Engineering, Université Laval, Quebec, G1V 0A6, Canada
[b] Department of Computer Science and Software Engineering, Université Laval, Quebec, G1V 0A6, Canada
[c] Department of Biomedical Informatics (DBMI), Harvard Medical School, Boston, MA, USA
[d] Department of Computer Science, University of Western Ontario, London, N6A 5B7, Canada

ARTICLE INFO

*Article history:*
Received 25 March 2022
Received in revised form 19 July 2022
Accepted 27 August 2022
Available online 24 September 2022

*Keywords:*
Learning theory
Transfer learning
Domain adaptation
Representation learning
Deep learning

ABSTRACT

Domain adaptation aims to alleviate the shift between training and test distribution, where the DA theory is crucial in understanding the success of domain adaptation algorithms. In this paper, we reveal the incoherence between the empirical domain adversarial training and its generally assumed theoretical counterpart based on $\mathcal{H}$-divergence. Concretely, we find that $\mathcal{H}$-divergence is not equivalent to Jensen–Shannon divergence, the optimization objective in domain adversarial training. To this end, we establish a new theoretical framework by directly proving the upper and lower target risk bounds based on the joint distributional Jensen–Shannon divergence. We further derive bidirectional upper bounds for marginal and conditional shifts. Our framework exhibits inherent flexibility for different transfer learning problems, which is usable for various scenarios. From an algorithmic perspective, our theory enables a generic guideline of the unified principles of semantic conditional matching, feature marginal matching, and label marginal shift correction. We employ algorithms for each principle and empirically validate the benefits of our framework.

## 1. Introduction

In many practical machine learning scenarios, the model is trained on a fixed source but used for a different and related target. To alleviate the performance degradation caused by such a distribution, Domain adaptation (DA) [1] has been developed in various fields such as computer vision [2], natural language processing [3], and biomedical engineering [4].

Specifically, DA theory is crucial to the fundamental understanding and practical development of practical algorithms. Conventionally, such theoretical guarantees were typically established on the notion of $\mathcal{H}$-divergence [5,6] and its subsequent variants such as [7], where it requires a small $\mathcal{H}$-divergence between source–target and small joint risk. In the context of representation learning, this quantity ($\mathcal{H}$-divergence) is minimized via the well-known *domain adversarial training* such as [2,8,9], which is still a stimulating topic in current research.

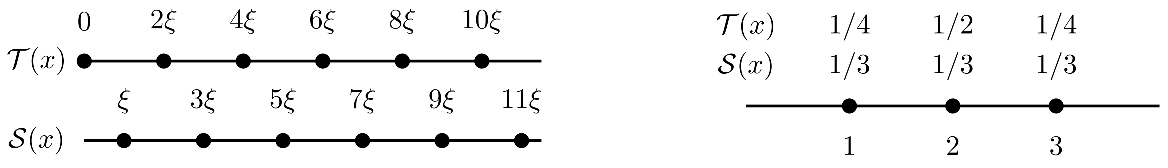
Domain adversarial training is widely successful in various DA problems such as open set DA [10–12] or conditional shift [4], however, the general assumed theoretical counterpart $\mathcal{H}$-divergence itself is rather limited to explain these working principles, which hampers the further practical advancement. It has been noted that the inherent principle of domain adversarial training is analogous to GANs [13], which is equivalent to minimize Jensen–Shannon divergence [14] between two distributions. Therefore, a DA theory established directly on the Jensen–Shannon divergence would provide a thorough understanding of adversarial training and help overcome the limitations imposed by the use of $\mathcal{H}$-divergence.

In this work, we reveal that $\mathcal{H}$-divergence is **not** consistent with the Jensen–Shannon divergence, indicating the improper adoption of $\mathcal{H}$-divergence theory to explain the domain adversarial training practice. We further build a DA theoretical framework *directly* based on Jensen–Shannon divergence. We establish that the upper bound of the target risk is determined by the source error and the Jensen–Shannon divergence of the two joint distribution (Section 3.1). Moreover, we derive the upper bounds of bidirectional shifts (Section 3.2), including (a) Feature Marginal Shift ($\mathcal{T}(x) \neq \mathcal{S}(x)$) and Label Conditional Shift ($\mathcal{T}(y|x) \neq \mathcal{S}(y|x)$); and (b) Label Marginal Shift ($\mathcal{T}(y) \neq \mathcal{S}(y)$) and Semantic (Feature) Conditional Shift ($\mathcal{T}(x|y) \neq \mathcal{S}(x|y)$). The theory provides a unified understanding of domain shifts, with covariate shift and label shift being its special cases, which provides theoretic insights and effective practice guidelines:

**Theoretical Insights** Jensen–Shannon divergence enables us to analyze the factors of label space that influence the transfer procedure, which remains elusive in the $\mathcal{H}$-divergence. Specifically, (I) we reveal that the intrinsic error of learning in the target-domain is controlled by the label-space size, the source domain intrinsic error and the similarity of the two domains (Section 3.3.1). (II) it also reveals why transfer learning is challenging

[*] Corresponding author.
*E-mail address:* bwang@csd.uwo.ca (B. Wang).

https://doi.org/10.1016/j.knosys.2022.109808
0950-7051/© 2022 Elsevier B.V. All rights reserved.



(a) (b)

**Fig. 1.** $D_{\mathrm{JS}}(\mathcal{T}(x) \parallel \mathcal{S}(x))$ cannot be viewed as the approximation of $d_{\mathcal{H}}(\mathcal{S}(x), \mathcal{T}(x))$: (a) for two uniform distributions with different supports, there exists $d_{\mathcal{H}}(\mathcal{T}(x), \mathcal{S}(x)) \ll D_{\mathrm{JS}}(\mathcal{T}(x) \parallel \mathcal{S}(x))$ if $0 < \xi \ll 1$; (b) while for two distributions with different probability mass, there exists $D_{\mathrm{JS}}(\mathcal{T}(x) \parallel \mathcal{S}(x)) < d_{\mathcal{H}}(\mathcal{T}(x), \mathcal{S}(x))$

if the label space of source and target are not identical (a.k.a. open set DA). We formally show that a smaller overlap over the label space leads to a difficult transfer (Section 3.3.2).

**Practical Implications** Our theory motivates new DA practice for representation learning, which is missing in $\mathcal{H}$-divergence. More concretely, we propose unified principles to control the target risk (Section 4.2): (I) re-weighted semantic conditional matching, to control the feature conditional shift $D_{\mathrm{JS}}(\mathcal{T}(x|y) \parallel \mathcal{S}(x|y))$; (II) label marginal shift correction, as the way to eliminate the label marginal shift $D_{\mathrm{JS}}(\mathcal{T}(y) \parallel \mathcal{S}(y))$; (III) constraining the feature marginal shift, an approach to prevent poor target pseudo label predictions (i.e. predicted labels), a common phenomena that can lead to negative transfer in semantic conditional matching. The proposed guideline enables us to select existing algorithms for each principle. The empirical results on real datasets verify the benefits of unified principles (Section 5).

## 2. $\mathcal{H}$-divergence based DA theory

In this paper, suppose we have the source distribution $\mathcal{S}$ and target distribution $\mathcal{T}$ over the *joint* input and output space $\mathcal{X} \times \mathcal{Y}$. According to [5,6], if the data is generated by a marginal distribution and underlying labeling function pair $(\mathcal{D}, h^\star)$, then the upper bound of the target risk error w.r.t. $\forall h \in \mathcal{H}$ is:

$$R_{\mathcal{T}}(h) \leq R_{\mathcal{S}}(h) + d_{\mathcal{H}}(\mathcal{T}(x), \mathcal{S}(x)) + \beta, \tag{1}$$

where $R_{\mathcal{D}}(h) = \mathbb{E}_{x\sim\mathcal{D}}|h(x) - h^\star(x)|$, $d_{\mathcal{H}}$ denotes the $\mathcal{H}$-divergence for measuring the marginal distribution similarities between $\mathcal{S}(x)$ and $\mathcal{T}(x)$ w.r.t. $x$, $\beta$ is the optimal joint risk over the two domains.

As pointed out by [5], it is generally impossible to exactly estimate the $\mathcal{H}$-divergence. Hence, this measure is approximated as a binary classification task where we are discriminating the source and the target samples. More specifically, the $\mathcal{H}$-divergence is approximated by distance $d_{\mathcal{A}} = 2(1 - 2\epsilon)$, with $\epsilon$ corresponding to the discrimination generalization error. Inspired by this intuition, [2] and subsequent approaches empirically adopted adversarial loss [13] between the domain classifier $d$ and feature extractor function $g$ in the context of representation learning:

$$\begin{aligned}&\min_g \max_d \mathbb{E}_{x\sim\mathcal{S}(x)} \log(d \circ g(x)) + \mathbb{E}_{x\sim\mathcal{T}(x)} \log(1 - d \circ g(x)),\\ &= \min_g D_{\mathrm{JS}}(\mathcal{S}(g(x)) \parallel \mathcal{T}(g(x))),\end{aligned} \tag{2}$$

where Eq. (2) is the dual term of Jensen–Shannon divergence [14].

### 2.1. JS divergence is not consistent with $\mathcal{H}$-divergence

From Eq. (2), domain adversarial training is essentially in learning representation to minimize the Jensen–Shannon divergence. However a $D_{\mathrm{JS}}$ is not equivalent to $d_{\mathcal{H}}$ in Eq. (1). We find these two metrics can be very different and present two counterexamples to illustrate it, shown in Fig. 1.

For the sake of simplicity, we design all examples over one dimensional space and use the threshold functions $\mathcal{H} = \{h_t : t \in \mathbb{R}\}$ as the hypothesis class. That is, for any $t \in \mathbb{R}$, the threshold function is defined by $h_t(x) = 1$ for $x < t$ and $h_t(x) = 0$ otherwise.

**Counterexample 1** We adopt the example of [15], showed in Fig. 1(a), with a small fixed $\xi \in (0, 1)$. Let the target $\mathcal{T}(x)$ be the uniform distribution over $\{2k\xi : k \in \mathbb{N}, 2k\xi \leq 1\}$ and the source $\mathcal{S}(x)$ be the uniform distribution over $\{(2k + 1)\xi : k \in \mathbb{N}, (2k + 1)\xi \leq 1\}$. We can compute $d_{\mathcal{H}}(\mathcal{T}(x), \mathcal{S}(x)) \approx d_{\mathcal{A}}(\mathcal{T}(x), \mathcal{S}(x)) = \xi$ while $D_{\mathrm{JS}}(\mathcal{T}(x) \parallel \mathcal{S}(x)) = 1$ since the two distributions have *disjoint* supports. Then $d_{\mathcal{H}}(\mathcal{T}(x), \mathcal{S}(x)) \ll D_{\mathrm{JS}}(\mathcal{T}(x) \parallel \mathcal{S}(x))$ when $\xi \ll 1$, indicating a *small $\mathcal{H}$-divergence can correspond to a very large Jensen–Shannon divergence.*

**Counterexample 2** Fig. 1(b) further illustrates that Jensen–Shannon divergence is *not* the upper bound of $\mathcal{H}$-divergence. We assume the source $\mathcal{S}(x)$ be the uniform distribution over $\{1, 2, 3\}$ and let the target $\mathcal{T}(x)$ be the distribution on the same support with different probability mass $\{\mathcal{T}(x = 1) = 1/4, \mathcal{T}(x = 2) = 1/2, \mathcal{T}(x = 3) = 1/4\}$. Then Jensen–Shannon divergence can be even smaller than $\mathcal{H}$-divergence: $D_{\mathrm{JS}}(\mathcal{T}(x) \parallel \mathcal{S}(x)) < d_{\mathcal{H}}(\mathcal{T}(x), \mathcal{S}(x))$.

Due to these differences, $\mathcal{H}$-divergence is *not a proper* theoretical tool for analyzing the practice that minimizes the Jensen–Shannon divergence (e.g. domain adversarial training and its variants such as [2]).

## 3. DA theory with JS divergence and theoretical insights

### 3.1. Upper and lower bound

We assume the data $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is generated from a *joint* distribution $\mathcal{D}$ and denote the hypothesis and loss function as $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and $L : \mathbb{R} \to \mathbb{R}$, where the hypothesis $h \in \mathcal{H}$ actually outputs a score of an observation $(x, y)$. We also denote $R_{\mathcal{D}}(h)$ the expected risk w.r.t. distribution $\mathcal{D}$: $R_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\, L(h(x, y))$. The complete proofs are demonstrated in Appendix A.

**Theorem 1** (*Upper Bound*). *Supposing the prediction loss $L$ is bounded within an interval $G$: $G = \max(L) - \min(L)$, then for all the hypothesis $h$ the expected risk w.r.t. the target domain can be upper bounded by:*

$$R_{\mathcal{T}}(h) \leq R_{\mathcal{S}}(h) + \frac{G}{\sqrt{2}}\sqrt{D_{\mathrm{JS}}(\mathcal{T} \parallel \mathcal{S})},$$

*where $D_{\mathrm{JS}}(\mathcal{T} \parallel \mathcal{S}) = \frac{1}{2}[D_{KL}(\mathcal{S} \parallel \mathcal{M}) + D_{KL}(\mathcal{T} \parallel \mathcal{M})]$ with $\mathcal{M} = \frac{1}{2}(\mathcal{T} + \mathcal{S})$ is the Jensen–Shannon divergence between the joint distribution $\mathcal{S}(x, y)$ and $\mathcal{T}(x, y)$.*

**Discussions** The theoretical result seamlessly connects the well-known assumptions in DA. When the *covariate shift* assumption holds ($\mathcal{T}(y|x) = \mathcal{S}(y|x)$), the upper bound can be expressed as $R_{\mathcal{T}}(h) \leq R_{\mathcal{S}}(h) + \frac{G}{\sqrt{2}}\sqrt{D_{\mathrm{JS}}(\mathcal{T}(x) \parallel \mathcal{S}(x))}$. Besides, when the *label shift* assumption holds ($\mathcal{T}(x|y) = \mathcal{S}(x|y)$), the upper bound can be alternatively expressed as $R_{\mathcal{T}}(h) \leq R_{\mathcal{S}}(h) + \frac{G}{\sqrt{2}}\sqrt{D_{\mathrm{JS}}(\mathcal{T}(y) \parallel \mathcal{S}(y))}$.

**Theorem 2** (*Lower Bound*). *If we assume the loss $L$ as zero–one binary loss, then for any $h$, we can prove the target risk is lower bounded by:*

$$R_{\mathcal{T}}(h) \geq R_{\mathcal{S}}(h) - \sqrt{D_{\mathrm{JS}}(\mathcal{T} \parallel \mathcal{S})}.$$

The lower bound provides the insights of the *easy transfer* [16] scenario: learning the target domain can be easier than the source domain, and the gap is controlled (smaller than) by their distribution distance. For example, if we assume $R_{\mathcal{S}}(h) = 0.2$, $D_{\text{JS}}(\mathcal{T} \parallel \mathcal{S}) = 2 \times 10^{-4}$, then the target risk is also bounded: $R_{\mathcal{T}}(h) \in [0.186, 0.21]$. This indicates $R_{\mathcal{T}}(h)$ can be smaller than $R_{\mathcal{S}}(h)$ but not an arbitrary large gap.

### 3.2. Bi-directional marginal/conditional shifts

We can decompose the joint Jensen–Shannon divergence into bi-directional marginal and conditional shift upper bounds, according to the information theoretical chain rule [17].

**Corollary 1.** *The upper bound in Theorem 1 can be further decomposed as:*

$$\begin{aligned} R_{\mathcal{T}}(h) \le \quad & R_{\mathcal{S}}(h) + \frac{G}{\sqrt{2}} \underbrace{\sqrt{D_{\text{JS}}(\mathcal{T}(x) \parallel \mathcal{S}(x))}}_{\textit{Feature Marginal Shift}} \\ & + \frac{G}{\sqrt{2}} \underbrace{\sqrt{\mathbb{E}_{x\sim\mathcal{S}(x)} D_{\text{JS}}(\mathcal{T}(y|x) \parallel \mathcal{S}(y|x))}}_{\textit{Label Conditional Shift}} \\ & + \frac{G}{\sqrt{2}} \underbrace{\sqrt{\mathbb{E}_{x\sim\mathcal{T}(x)} D_{\text{JS}}(\mathcal{T}(y|x) \parallel \mathcal{S}(y|x))}}_{\textit{Label Conditional Shift}} \end{aligned} \tag{3}$$

$$\begin{aligned} R_{\mathcal{T}}(h) \le \quad & R_{\mathcal{S}}(h) + \frac{G}{\sqrt{2}} \underbrace{\sqrt{D_{\text{JS}}(\mathcal{T}(y) \parallel \mathcal{S}(y))}}_{\textit{Label Marginal Shift}} \\ & + \frac{G}{\sqrt{2}} \underbrace{\sqrt{\mathbb{E}_{y\sim\mathcal{S}(y)} D_{\text{JS}}(\mathcal{T}(x|y) \parallel \mathcal{S}(x|y))}}_{\textit{Semantic (Feature) Conditional Shift}} \\ & + \frac{G}{\sqrt{2}} \underbrace{\sqrt{\mathbb{E}_{y\sim\mathcal{T}(y)} D_{\text{JS}}(\mathcal{T}(x|y) \parallel \mathcal{S}(x|y))}}_{\textit{Semantic (Feature) Conditional Shift}} \end{aligned} \tag{4}$$

In particular, Eq. (4) provides an alternative direction for understanding DA. The target risk bound is alternatively controlled by the label marginal shift and the semantic (feature) conditional distribution shift. Generally, the source and target label marginal distribution, as well as the semantic (feature) conditional distributions are *both different*. For example, in the classification of different digit datasets (e.g., MNIST, USPS), when conditioning on the certain digit $Y = y$, it is clear that $\mathcal{S}(x|Y = y) \neq \mathcal{T}(x|Y = y)$, indicating the necessity of considering semantic information in DA.

### 3.3. Theoretical applications

One fundamental challenge in DA is to discover the relations and inherent properties of learning tasks, that ensure a successful transfer [15]. Jensen–Shannon divergence enables us to analyze the factor of label space that influences the transfer procedure, illustrated in two concrete scenarios.

#### 3.3.1. Application I: Target intrinsic error in DA

To characterize the inherent difficulty in the learning a task, we adopt the conditional entropy $H(Y_{\mathcal{D}}|X_{\mathcal{D}}) = \mathbb{E}_{x\sim\mathcal{D}(x)} H(Y|X = x)$ as the intrinsic error, an error in predicting the labels given that the underlying data distribution $\mathcal{D}$ is known [18,19]. For example, if $X$ does not provide any information for the label $Y$ such that $Y \perp\!\!\!\perp X$, then the conditional entropy arrives its maximum: $H(Y|X) = H(Y)$, indicating the impossibility to guarantee a small prediction error. However, in the context of $\mathcal{H}$-divergence [6], this property *cannot be analyzed* since the label is determined by a *fixed* labeling function, such that $H(Y|X) = \mathbb{E}_{x\sim\mathcal{D}(x)} H(h^\star(x)|X = x) \equiv 0$.

**Target Intrinsic Error: Upper Bound** In the context of DA, our goal is to ensure a small target risk, i.e., a small target intrinsic error is necessary. However, we never have the full target distribution $\mathcal{T}(x, y)$, indicating the impossibility to directly estimate target intrinsic error $H(Y_t|X_t)$. In contrast, we can have the information of source distribution, as well as the relations of source and target distribution. Then we can derive the target intrinsic error is controlled by the label space size, as well as the source intrinsic error and Jensen–Shannon divergence of two distributions. This result is also consistent with our intuition and the lower bound derived by Fano's inequality [17]: a smaller label space $|\mathcal{Y}|$ is generally easier to learn, if the other conditions are identical.

**Theorem 3.** *If we have: (1) Small source intrinsic error:* $H(Y_s|X_s) \le \epsilon$; *(2) Marginal distributions defined in Eq. (3) are close:* $D_{\text{JS}}(\mathcal{S}(x) \parallel \mathcal{T}(x)) \le \delta_1$; *(3) Conditional distributions defined in Eq. (3) are close:* $D_{\text{JS}}(\mathcal{S}(y|X = x) \parallel \mathcal{T}(y|X = x)) \le \delta_2$, $\forall x$, *Then the target intrinsic error can be upper bounded by:*

$$H(Y_t|X_t) \le \epsilon + \sqrt{\frac{\delta_2}{2}} + \frac{\sqrt{\delta_1}}{2} \log |\mathcal{Y}|.$$

#### 3.3.2. Application II: Inherent difficulty in learning open set DA

Our theory also proposes the analysis to understand when and what is difficult to transfer in Open Set DA, i.e., the source and target domain share only a portion of label space [11,12].

The key observation is that $\text{supp}\{\mathcal{T}(y)\} \cap \text{supp}\{\mathcal{S}(y)\} \neq \emptyset$. We suppose a small semantic conditional shift ($\forall y$, $D_{\text{JS}}(\mathcal{S}(x|y) \parallel \mathcal{T}(x|y)) \le \delta$ for a small $\delta > 0$), and a uniform label distributions over two different label spaces $\mathcal{Y}_1$ and $\mathcal{Y}_2$ such that $\mathcal{S}(y) \sim \text{Unif}(\mathcal{Y}_1)$, $\mathcal{T}(y) \sim \text{Unif}(\mathcal{Y}_2)$, $|\mathcal{Y}_1| = |\mathcal{Y}_2| = N$. We further assume the number of shared classes is $|\mathcal{Y}_1 \cap \mathcal{Y}_2| = \alpha N$, $0 < \alpha < 1$. Then if the loss is binary and based on Theorem 2 and Eq. (4), the target risk can be bounded:

$$R_{\mathcal{S}}(h) - \left(\sqrt{1-\alpha} + 2\sqrt{\delta}\right) \le R_{\mathcal{T}}(h) \le R_{\mathcal{S}}(h) + \frac{1}{\sqrt{2}} \left(\sqrt{1-\alpha} + 2\sqrt{\delta}\right).$$

When $\alpha \to 1$, $D_{\text{JS}}(\mathcal{T}(y) \parallel \mathcal{S}(y)) \to 0$, the source risk is closed the target risk, then simply minimizing the source risk and further semantic conditional matching (see Section 4) can effectively control the target risk. On the contrary, if $\alpha \to 0$, the gap between target and source risk is large, indicating that a small source risk and semantic conditional shift no more guarantee a small target risk. From the practical perspective, less label overlapping means that it is harder to transfer the exact corresponding semantic conditional information from the source to the target.

## 4. Practical principles for unsupervised DA

In this section, we instantiate our theoretical framework with practical principles for designing unsupervised DA algorithms in deep learning. We would like to point out that our theory is initially based on the labeled data information, but the practical principle can be applied in the unsupervised scenario.

We introduce a *feature embedding function* $g : \mathcal{X} \to \mathcal{Z}$ and denote latent variable (feature) $z = g(x)$. Our objective is to find a representation function $g$ and classifier $h$, following the principles in Table 1. We also denote $\hat{\mathcal{S}}(x, y) = \{(x_s^i, y_s^i)\}_{i=1}^{N_s}$, $\hat{\mathcal{T}}(x) = \{x_t^i\}_{i=1}^{N_t}$ as the observed (empirical) distribution.

### 4.1. Difficulty in controlling label conditional shift

In Eq. (3) in Corollary 1, it recovers the principles induced by $\mathcal{H}$-divergence. Specifically, the domain adversarial training is

**Table 1**
Empirical methods for bi-directional marginal/conditional shifts.

| Corollary 1 | | Source | Marginal Shift | Conditional Shift |
|---|---|---|---|---|
| Eq. (3) | Term | $R_{\mathcal{S}}(h)$ | $D_{\text{JS}}(\mathcal{T}(z)\\|\mathcal{S}(z))$ | $D_{\text{JS}}(\mathcal{T}(y\|z)\\|\mathcal{S}(y\|z))$ |
| | Method | ERM | Feature Marginal Matching | N/A |
| Eq. (4) | Term | $R_{\mathcal{S}}(h)$ | $D_{\text{JS}}(\mathcal{T}(y)\\|\mathcal{S}(y))$ | $D_{\text{JS}}(\mathcal{T}(z\|y)\\|\mathcal{S}(z\|y))$ |
| | Method | Label Marginal Shift Correction | | Semantic Distribution Matching |

equivalent to minimize the dual form of Jensen–Shannon divergence [14].

However, domain adversarial training *cannot* guarantee a small upper bound in Corollary 1. To this end, we can prove that merely minimizing $D_{\text{JS}}(\hat{\mathcal{T}}(z) \parallel \hat{\mathcal{S}}(z))$ can lead to an increase in the label conditional shift $D_{\text{JS}}(\hat{\mathcal{T}}(y|z) \parallel \hat{\mathcal{S}}(y|z))$, which is illustrated as the following:

$$\mathbb{E}_{z\sim\hat{\mathcal{T}}(z)}D_{\text{JS}}(\hat{\mathcal{S}}(y|z) \parallel \hat{\mathcal{T}}(y|z)) + \mathbb{E}_{z\sim\hat{\mathcal{S}}(z)}D_{\text{JS}}(\hat{\mathcal{S}}(y|z) \parallel \hat{\mathcal{T}}(y|z))$$
$$\geq 2\left(\sqrt{D_{\text{JS}}(\hat{\mathcal{T}}(y) \parallel \hat{\mathcal{S}}(y))} - \sqrt{D_{\text{JS}}(\hat{\mathcal{S}}(z) \parallel \hat{\mathcal{T}}(z))}\right)^2$$

The aforementioned inequality indicates that the third term in Eq. (3) is lower bounded by the gap between $D_{\text{JS}}(\hat{\mathcal{T}}(y) \parallel \hat{\mathcal{S}}(y))$ and $D_{\text{JS}}(\hat{\mathcal{S}}(z) \parallel \hat{\mathcal{T}}(z))$, then merely minimizing $D_{\text{JS}}(\hat{\mathcal{S}}(z) \parallel \hat{\mathcal{T}}(z))$ will be problematic if their label distributions are significantly different.

Moreover, controlling the label condition shift is practically difficult. Because it requires two identical *continuous* and high dimensional features such that $z_s = z_t$ with $z_s \in \hat{\mathcal{S}}(z)$, $z_t \in \hat{\mathcal{T}}(z)$, then minimizing $D_{\text{JS}}(\hat{\mathcal{T}}(y|Z = z_s) \parallel \hat{\mathcal{S}}(y|Z = z_t))$. Generally, it is not trivial to find such feature pairs $z_s = z_t$ from finite observational samples.

### 4.2. Proposed practice

According to Eq. (4) in Corollary 1, the target risk can be alternatively bounded by $R_{\mathcal{S}}(h)$, label marginal shift $D_{\text{JS}}(\mathcal{T}(y) \parallel \mathcal{S}(y))$, and semantic conditional shift i.e, $D_{\text{JS}}(\mathcal{T}(z|y) \parallel \mathcal{S}(z|y))$, which enable us to consider new principles in DA.

**(I) Semantic Conditional Distribution Matching** Different from the controlling the label conditional shift $D_{\text{JS}}(\hat{\mathcal{T}}(y|Z = z) \parallel \hat{\mathcal{S}}(y|Z = z))$, controlling the semantic (feature) conditional shift $D_{\text{JS}}(\hat{\mathcal{T}}(z|Y = y) \parallel \hat{\mathcal{S}}(z|Y = y))$ is practically more efficient, since *labels are usually categorical variables with the finite classes*, comparing with continuous latent variable $Z$. However, there are no ground truth labels on the target domain, inducing the main issue in semantic conditional matching in DA. For addressing this, target *pseudo labels* $Y_p$, estimated from the classifier, are introduced as the approximation of the real target label. Then following insights of the third term in Eq. (4), the semantic conditional loss can be expressed as:

$$\sum_y (\hat{\mathcal{S}}(y) + \hat{\mathcal{T}}_p(y)) D_{\text{JS}}\left(\hat{\mathcal{T}}(z|Y_p = y) \parallel \hat{\mathcal{S}}(z|Y = y)\right), \tag{5}$$

where $\hat{\mathcal{T}}_p(y)$ is the target pseudo distribution predicted by the neural network. We notice that [20] encoded the label prediction information $h \circ g(x)$ as the conditional domain adversarial training, to implicitly minimize the conditional distribution divergence. However, semantic conditional matching requires *relative good pseudo-label prediction*. Otherwise the incorrect semantic (feature) feature alignment will lead to a negative transfer procedure for the target domain during the learning phase.

**(II) Label Marginal Shift Correction** *Is the semantic conditional matching sufficient to control the target risk?* From Eq. (4), the target risk is also controlled by label marginal shift. We can further extend this conclusion in the representation learning: if the semantic conditional distribution is matched, then the target risk is still controlled by the label marginal shift.

**Theorem 4.** *If any classifier h, feature learner g, and label $y \in \mathcal{Y} = \{-1, +1\}$ such that semantic conditional distribution is matched, $D_{\text{JS}}(\mathcal{S}(z|y), \mathcal{T}(z|y)) = 0$, then the target risk can be bounded:*

$$|R_{\mathcal{S}}(h \circ g) - R_{\mathcal{T}}(h \circ g)| \leq \sqrt{2D_{\text{JS}}(\mathcal{S}(y), \mathcal{T}(y))},$$

*where $R_{\mathcal{S}}(h \circ g) = R_{\mathcal{S}}(h(g(x), y))$ is the expected risk over the classifier h and feature learner g.*

As Theorem 4 suggests, we need to control label marginal shift $D_{\text{JS}}(\mathcal{T}(y) \parallel \mathcal{S}(y))$. Therefore we adopt the popular label re-weighted loss [21]:

$$\hat{R}^{\alpha}_{\mathcal{S}}(h \circ g) = \sum_{(x_s, y_s)\sim\hat{\mathcal{S}}(x,y)} \alpha(y_s) L(h(g(x_s), y_s))$$

with $\alpha(y) = \frac{\mathcal{T}(y)}{\mathcal{S}(y)}$. In addition, we can further prove the empirical re-weighted loss converges to $R_{\mathcal{T}}(h \circ g)$, if $D_{\text{JS}}(\mathcal{S}(z|y), \mathcal{T}(z|y)) = 0$ (see Appendix for details). As for estimating label weight $\hat{\alpha}$ from the data, several approaches have been proposed, e.g. Black Box Shift Learning (BBSL) [22].

**(III) Feature Marginal Matching as a Constraint** Although the aforementioned principles are theoretically appealing, we practically use the pseudo label $Y_p$ for the semantic conditional matching $D_{\text{JS}}(\hat{\mathcal{T}}(z|Y_p = y) \parallel \hat{\mathcal{S}}(z|Y = y))$, which can lead to negative transfer in the training loop if we face poor pseudo label predictions.

Can we derive the principle to recognize poor pseudo label prediction during learning? Theorem 5 reveals one consequence of poor target pseudo label prediction: it can lead to a large empirical feature marginal divergence $D_{\text{JS}}(\hat{\mathcal{S}}(z) \parallel \hat{\mathcal{T}}(z))$ (in Eq. (3)), under mild conditions.

**Theorem 5.** *We denote $\hat{\mathcal{S}}_p(y), \hat{\mathcal{T}}_p(y)$ as the prediction output (pseudo-label) distributions. If we have such a "bad" pseudo label prediction such that $D_{\text{JS}}(\hat{\mathcal{T}}(y) \parallel \hat{\mathcal{T}}_p(y)) = P$, small source prediction error $D_{\text{JS}}(\hat{\mathcal{S}}(y) \parallel \hat{\mathcal{S}}_p(y)) \leq \epsilon_1$ and small label ground truth empirical distribution divergence $D_{\text{JS}}(\hat{\mathcal{S}}(y) \parallel \hat{\mathcal{T}}(y)) \leq \epsilon_2$, then the feature marginal divergence on the latent space Z can be lower bounded by:*

$$D_{\text{JS}}(\hat{\mathcal{S}}(z) \parallel \hat{\mathcal{T}}(z)) \geq (\sqrt{P} - \sqrt{\epsilon_1} - \sqrt{\epsilon_2})^2.$$

Theorem 5 suggests that if $P \to 1$ and $\epsilon_1, \epsilon_2$ are small, $D_{\text{JS}}(\hat{\mathcal{S}}(z) \parallel \hat{\mathcal{T}}(z))$ can be very large. Therefore we add the constraint $\mathcal{D}_{\text{JS}}(\hat{\mathcal{S}}(z) \parallel \hat{\mathcal{T}}(z)) \leq \kappa$ as a broad adaptation step, to prevent the poor pseudo-label prediction (a.k.a. large $P$). In practice, we adopt Lagrangian relaxation as treating the constraint as a small regularization term, where $\kappa$ is the hyper-parameter.

**Practical Guideline** Based on these three principles, we propose a generic and iterative practical framework, where parameter optimization and pseudo-label prediction steps are conducted iteratively.

Moreover, we would like to emphasize that the realization of each principle is flexible. For example, the distribution matching can be done through either adversarial training by introducing the auxiliary domain discriminator $d$ or parametric distribution matching (e.g. statistical moment matching approach).

### *4.3. Training losses*

Based on the theoretical analysis, we proposed the following loss and algorithm in domain adaptation.

**I. Semantic Conditional Distribution Matching** As it is demonstrated, the first component is to match the semantic conditional distribution divergence. According to the upper bound of Jensen–Shannon divergence, the second term is upper bounded by $\sum_y(\hat{\mathcal{S}}(y) + \hat{\mathcal{T}}_p(y)) \parallel \hat{\mathcal{T}}(\cdot|y) - \hat{\mathcal{S}}(\cdot|y) \parallel_2$. We simply approximate the center of empirical distribution as the surrogate of the conditional distribution. Then we have:

$$\hat{\mathcal{S}}(g(x_s)|y) \approx \frac{1}{|\#y_s = y|} \sum_{(x_s,y_s)} \delta_{\{y_s=y\}} g(x_s)$$

$$\hat{\mathcal{T}}(g(x_t)|y) \approx \frac{1}{|\#y_t^p = y|} \sum_{(x_t,y_t^p)} \delta_{\{y_t^p=y\}} g(x_t)$$

Therefore the conditional matching term can be approximated as$\hat{R}_{\text{cond}}(g) = \sum_y(\hat{\mathcal{S}}(Y = y) + \hat{\mathcal{T}}_p(Y = y)) \parallel \hat{\mathcal{S}}(g(x_s)|Y = y) - \hat{\mathcal{T}}(g(x_t)|Y_p = y) \parallel_2^2$

**II. Labeling Marginal Shift Correction** The theoretical results also suggest the correcting the source data distribution, thus in the classification, we adopted the re-weighted cross-entropy:

$$\hat{R}_{\mathcal{S}}^{\hat{\alpha}}(f, g) = -\frac{1}{N_S} \sum_{(x_s,y_s)\sim\hat{\mathcal{S}}} \hat{\alpha}(y_s) \log(h \circ (g(x_s), y_s))$$

As for estimating $\hat{\alpha}$, we follow the popular BBSL estimator [22]. We first construct a source prediction confusion matrix $\hat{C} \in |\mathcal{Y}| \times |\mathcal{Y}|$ with $\hat{C}[i, j] = \mathbb{P}(\text{argmax}_y\ h(g(x_s), y) = i, y_s = j)$. The target pseudo-label $y^p$ and target pseudo-label distribution $\hat{\mathcal{T}}_p$ can be directly estimated from the neural network. Then the label re-weighting coefficient can be estimated as:

$$\hat{\alpha} = \hat{C}^{-1}\hat{\mathcal{T}}_p$$

**III. Marginal Feature Distribution Matching as the Constraint** Since the relative accurate pseudo-label estimation is important in the iterative algorithm, thus we introduce the marginal feature distribution matching as the training constraint. The main goal is to keep a good pseudo-label initialization. We just adopt the most popular Jensen–Shannon domain adversarial training.

$$\hat{R}_{\text{adv}}(d, g) = \mathbb{E}_{x_s\sim\hat{\mathcal{S}}(x)} \log(d \circ g(x_s)) + \mathbb{E}_{x_t\sim\hat{\mathcal{T}}(x)} \log(1 - d \circ g(x_t))$$

*Proposed algorithm* Based on this three losses, we proposed Algorithm 1.

*Complexity analysis* Thanks for your thoughtful comment. In terms of computational complexity, we should compute $\mathcal{O}(1)$ time adversarial loss and $\mathcal{O}(1)$ time re-weighted prediction loss. Besides, we should estimate the label ratio coefficient $\alpha$ to solve a quadratic optimization $\mathcal{O}(|\mathcal{Y}|)$ ($|\mathcal{Y}|$ is the number of classes) through gradient descent. As a result, the total computational complexity is $\mathcal{O}(|\mathcal{Y}|)$.

In terms of memory complexity, it requires $\mathcal{O}(1)$ domain discriminator and $\mathcal{O}(|\mathcal{Y}|)$ class feature centroid. Because the feature centroid is estimated in the *embedding space z*, the actual memory complexity can be much lower than that of the domain discriminator.

## 5. Experiments

We validate the proposed guideline by realizing each principle. We aim to show whether applying the unified principles is better than merely considering only one or two of them.

**Algorithm 1** Jensen–Shannon Principles in Unsupervised DA

```
Require: Labeled source $\hat{\mathcal{S}}$, Unlabelled Target $\hat{\mathcal{T}}$
Ensure: Label distribution ratio $\hat{\alpha}$. Feature embedding $g$, Classifier $h$, Domain discriminator $d$, class centroid for source $\mathbf{C}_s^y$ and target $\mathbf{C}^y$ ($\forall y \in \mathcal{Y}$).
▷ ▷ ▷ DNN Parameter Training Stage (fixed $\hat{\alpha}$) ◁ ◁ ◁
for mini-batch of samples $(\mathbf{x}_{\mathcal{S}}, \mathbf{y}_{\mathcal{S}}) \sim \hat{\mathcal{S}}$, $(\mathbf{x}_{\mathcal{T}}) \sim \hat{\mathcal{T}}$ do
    Predict target pseudo-label
        $\bar{\mathbf{y}}_{\mathcal{T}} = \text{argmax}_y h(g(\mathbf{x}_{\mathcal{T}}), y)$
    Compute unnormalized source confusion matrix for each batch.
        $C_{\hat{\mathcal{S}}} = \#[\text{argmax}_{y'} h(z, y') = y, Y = k]$
    Compute the *batched* class centroid for source $C_s^y$ and target $C^y$.
    Update source/target class centroid:
    Source class centroid update
    $\mathbf{C}_s^y = \epsilon_1 \times \mathbf{C}_s^y + (1 - \epsilon_1) \times C_s^y$
    Target class centroid update
    $\mathbf{C}^y = \epsilon_1 \times \mathbf{C}^y + (1 - \epsilon_1) \times C^y$
    Updating $g, h, d$ to minimize $\hat{R}_{\mathcal{S}}^{\hat{\alpha}}(f, g) + \hat{R}_{\text{cond}}(g) + \hat{R}_{\text{adv}}(d, g)$
end for
▷ ▷ ▷ Estimation $\hat{\alpha}$ ◁ ◁ ◁
Compute the global or normalized source confusion matrix
  $C_{\hat{\mathcal{S}}} = \hat{\mathcal{S}}[\text{argmax}_{y'} h(z, y') = y, Y = k]$ $(t = 1, \dots, T)$
Solve $\alpha' = C_{\hat{\mathcal{S}}}^{-1}\hat{\mathcal{T}}_p$.
Update $\alpha$ by moving average: $\alpha = \epsilon_1 \times \alpha + (1 - \epsilon_1) \times \alpha'$
```

### *5.1. Experimental settings*

We evaluate the proposed framework on two benchmarks.

**Digits Recognition**. It includes 3 domains: MNIST, SVHN [23] and USPS [24] dataset. MNIST is composed of gray images of size 28 × 28, USPS contains 16 × 16 gray digits; and SVHN consists of 32 × 32 color digits images, which are more challenging and can contain more than one digit in each image. We randomly sample 7K samples for each task. We evaluate our method by using the three typical adaptation tasks: USPS↔MNIST (two tasks) and SVHN→MNIST (one task).

**Office-31 dataset** [25]. It consists of 4,652 images and 31 categories collected from three different domains: Amazon (A) from amazon.com, Webcam (W) and DSLR (D), taken by web camera and digital SLR camera in different environmental settings, respectively.

**ImageCLEF** [26] This data is originally used for the ImageCLEF 2014 domain adaptation challenge consists of twelve common classes from three domains: ImageNet ILSVRC 2012 (I), Pascal VOC 2012 (P), and Caltech-256 (C). Each domain has 600 images in total and contains 50 images per class. We test 6 tasks by using all domain combinations.

**Amazon Review** [27] contains four domains with positive and negative product reviews. We follow the strategies of [28] to form a 5000-dimensional bag-of-words feature. Note that the label distribution in the original dataset is uniform. To further show the benefits of the proposed approach, we followed [29] by selecting specific domains (DVD, Electronics, and Kitchen) and creating a label distribution drifted task by randomly dropping 50% negative reviews of all the sources while keeping the target unchanged.

We further visualize the label distribution of digits and Office-31, showing in Fig. 2. We observe the non-uniform label distributions over these two tasks. We implement the digits dataset based on the LeNet5 [30]. All digit images are resized to 28 × 28 for fair comparisons. As for Office-31 and Image CLEF task, we implement it on the Pre-trained AlexNet [31]. As for Amazon

**Table 2**
Accuracy (%) on Office-31 Dataset.

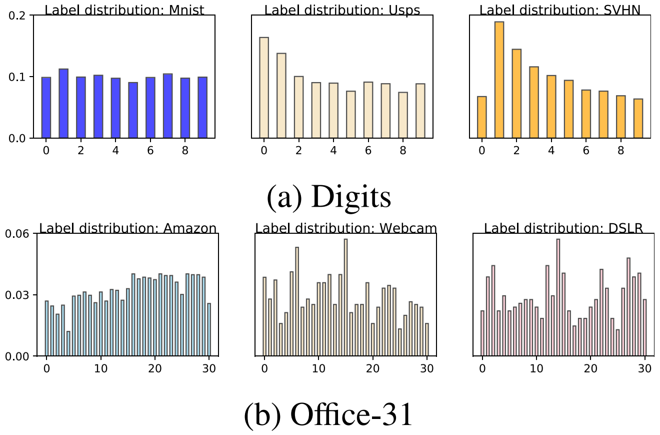
| Method | A → D | A → W | D → W | W → D | W → A | D → A | Ave |
|---|---|---|---|---|---|---|---|
| Without DA | 63.8 ± 0.5 | 61.6 ± 0.5 | 95.4 ± 0.3 | 99.0 ± 0.2 | 49.8 ± 0.4 | 51.1 ± 0.6 | 70.1 |
| DANN [2] | 72.3 ± 0.3 | 73.0 ± 0.5 | 96.4 ± 0.3 | 99.2 ± 0.3 | 51.2 ± 0.5 | 52.4 ± 0.4 | 74.1 |
| CDAN [20] | 76.3 ± 0.1 | 78.3 ± 0.2 | 97.2 ± 0.1 | **100.0**±0.0 | 57.5 ± 0.4 | 57.3 ± 0.2 | 77.7 |
| (I + III) | 72.6 ± 0.4 | 73.5 ± 0.4 | 96.2 ± 0.2 | 99.3 ± 0.5 | 51.4 ± 0.2 | 52.8 ± 0.5 | 74.3 |
| (I + II) | 75.3 ± 0.7 | 79.4 ± 1.1 | 97.1 ± 0.5 | 97.5 ± 0.5 | 58.2 ± 0.9 | 61.8 ± 0.8 | 78.2 |
| (II + III) | 75.7 ± 0.1 | 79.2 ± 0.7 | 96.8 ± 0.1 | 99.8 ± 0.1 | 59.5 ± 0.4 | 58.7 ± 0.3 | 78.3 |
| (I + II + III) | **76.7**±0.4 | **80.8**±0.4 | **97.5**±0.2 | 99.8 ± 0.1 | **59.8**±0.4 | **62.3**±0.2 | **79.5** |

**Table 3**
Accuracy (%) on digits dataset.

| Method | SVHN → MNIST | MNIST → USPS | USPS → MNIST |
|---|---|---|---|
| Without DA | 62.1 ± 1.2 | 87.1 ± 0.9 | 78.1 ± 0.6 |
| DANN [2] | 73.8 ± 1.8 | 89.1 ± 0.6 | 83.0 ± 0.8 |
| CDAN [20] | 86.7 ± 0.8 | 93.2 ± 0.6 | 93.0 ± 0.5 |
| (I + III) | 76.7 ± 0.8 | 89.4 ± 0.7 | 84.6 ± 1.4 |
| (I + II) | 87.3 ± 0.6 | 94.6 ± 0.7 | 94.7 ± 0.5 |
| (II + III) | 88.6 ± 0.9 | 95.5 ± 0.8 | 95.5 ± 0.7 |
| (I + II + III) | **89.6**±1.1 | **96.5**±0.6 | **97.0**±0.6 |

**Table 4**
Accuracy (%) on CLEF dataset.

| Method | I → C | I → P | C → I | P → I | C → P | P → C | Ave |
|---|---|---|---|---|---|---|---|
| Without DA | 84.3 ± 0.2 | 66.2 ± 0.2 | 71.3 ± 0.4 | 70.0 ± 0.2 | 59.3 ± 0.5 | 84.5 ± 0.3 | 73.9 |
| DANN [2] | 89.0 ± 0.4 | 66.5 ± 0.3 | 79.8 ± 0.4 | 81.8 ± 0.3 | 63.5 ± 0.5 | 88.7 ± 0.3 | 78.2 |
| CDAN [20] | 91.8 ± 0.2 | 67.7 ± 0.3 | 81.5 ± 0.2 | 83.3 ± 0.1 | 63.0 ± 0.2 | 91.5 ± 0.3 | 79.8 |
| (I + III) | 89.3 ± 0.2 | 67.0 ± 0.6 | 80.0 ± 0.7 | 81.9 ± 0.3 | 62.9 ± 0.4 | 89.2 ± 0.2 | 78.4 |
| (I + II) | 90.2 ± 0.5 | 66.7 ± 0.6 | 80.3 ± 0.5 | 82.7 ± 0.7 | 62.5 ± 0.7 | 90.7 ± 0.6 | 78.8 |
| (II + III) | 91.5 ± 0.1 | 67.3 ± 0.3 | 81.7 ± 0.3 | 82.8 ± 0.2 | 63.5 ± 0.4 | 91.2 ± 0.2 | 79.9 |
| (I + II + III) | **92.1**±0.2 | **68.2**±0.2 | **82.1**±0.2 | **84.0**±0.2 | **64.2**±0.2 | **91.9**±0.1 | **80.4** |



(a) Digits

(b) Office-31

**Fig. 2.** Label distribution on Digits and Office-31 dataset.

review, we followed the same network structure and adopted the code from [29]. We use the same hyper-parameter training strategy with DANN [2]. We update the neural network parameters and $\hat{\alpha}$. We compare the baselines of merely considering feature marginal [2], conditional matching [20], and our principles. We repeat the experiments five times and report the average and std (see Table 4).

### 5.2. Results and analysis

We report the empirical performances in Tables 2, 3 and 5. The empirical results indicate the improved performance on the unified principles, comparing with merely one or two principles. We observe the empirical benefit of semantic conditional matching (II) is relative more notable.

Fig. 3 further reveals the properties of the proposed principles. Specifically, Fig. 3(a) shows the evolution of each principle (loss) during the training, which is exact coherent with the goals in the guideline. The semantic conditional shift (Principle II) and the weighted source classification error (Principle I) gradually diminish and $D_{\text{JS}}(\hat{\mathcal{T}}(z) \parallel \hat{\mathcal{S}}(z))$ (Principle III) restricts within a small value. In addition, we trace the target domain prediction accuracy of different principles combinations in Fig. 3(b), for demonstrating the impact of each principle. The results indicate the importance of considering semantic (feature) conditional distribution matching (II), with a significant performance influence ($\sim$ 4.2%). On the other hand, the influences of principle (I) and (III) are relatively modest ($\sim$ 1.3%). Fig. 3(c) revealed estimated $\hat{\alpha}$ and its ground truth value, which verified the correctness of the proposed principle. In Amazon review dataset Table 5, the conditional matching and marginal matching have significant performance drop, because label distributions in the source are different from the target. This observation is also consistent with [29].

**Ablation Study: Label-drifted DA** To further elaborate the role of proposed principles, we simulate a significant label drifted in DA.

In Office-31 (A→ W dataset), we randomly drop out 25% samples in the first half of classes within source, and 25% samples in latter half of classes in target. We visualize result in Fig. 4(b), which verifies the strong practical benefits of semantic conditional matching (principle II, with improvement $\sim 5.8 - 9.4\%$). Besides, principle (III) empirically offers a coarse adaptation step to improve pseudo-label prediction. E.g., in Fig. 4(b), introducing principle (III) improves the prediction performance by $\sim$ 1.1% In Digits dataset (SVHN → MNIST), based on [4], we randomly drop

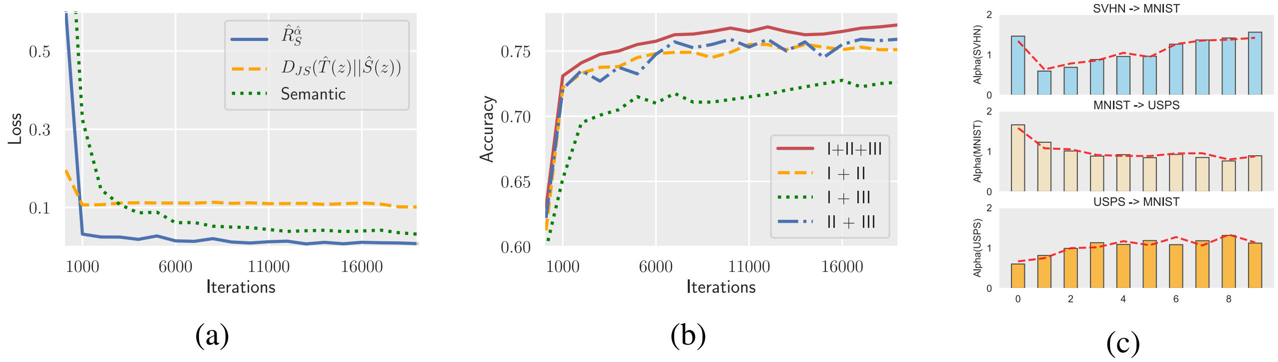


**Fig. 3.** Analysis of proposed principles. (a) Office-31, Domain A→D. Evolution of each loss during the training. (b) Office-31, Domain A→D. Evolution of accuracy during the training. (c) In digits dataset, we visualize the estimated $\hat{\alpha}$ (red dot curve) and ground truth value (bar plot).

**Table 5**
Accuracy (%) on label-shifted Amazon Review.

| Method | E → D | K → D | D → E | K → E | D → K | E → K | Ave |
|---|---|---|---|---|---|---|---|
| Without DA | 80.3 ± 0.3 | 73.4 ± 0.2 | 81.8 ± 0.3 | 71.8 ± 0.3 | 76.3 ± 0.1 | 76.5 ± 0.2 | 76.1 |
| DANN [2] | 70.3 ± 0.7 | 70.6 ± 0.9 | 69.1 ± 2.4 | 65.2 ± 1.7 | **77.9** ± 0.3 | 59.9 ± 3.4 | 68.8 |
| CDAN [20] | 71.7 ± 0.5 | 65.8 ± 1.4 | 82.3 ± 0.5 | 61.0 ± 2.1 | 73.9 ± 0.7 | 74.9 ± 0.5 | 71.6 |
| (I + III) | **80.5**± 0.4 | 70.7 ± 0.6 | 80.5 ± 0.3 | 66.1 ± 0.7 | 76.6 ± 0.4 | 78.4 ± 0.7 | 75.5 |
| (II + III) | 79.9 ± 0.4 | 68.6 ± 1.4 | 81.7 ± 0.5 | 70.8 ± 0.7 | 77.7 ± 0.8 | **79.5** ± 0.6 | 76.4 |
| (I + II + III) | 80.4 ± 0.6 | **74.1** ± 0.8 | **82.0** ± 0.8 | **72.3** ± 0.4 | 77.3 ± 0.5 | 79.4 ± 0.6 | **77.6** |

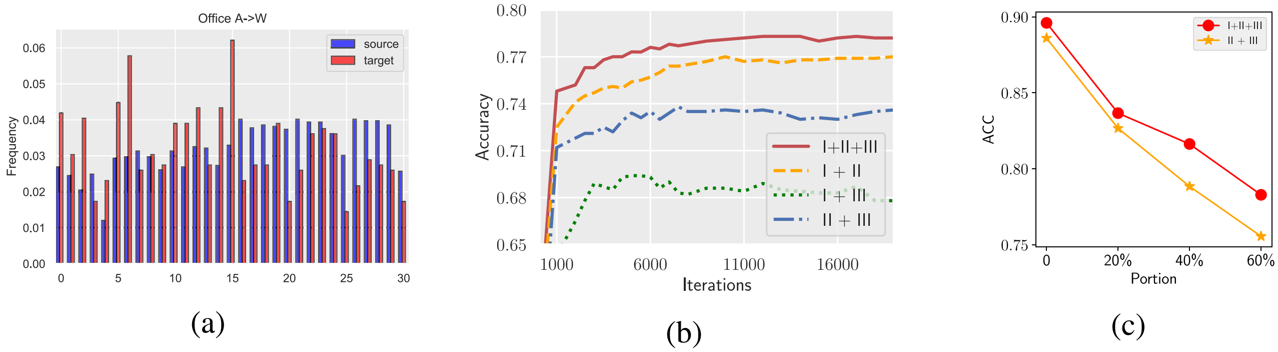


**Fig. 4.** Ablation Study: Label-drifted DA. (a) Office-31. Label distribution of drifted A→W. (b) Label drifted A→W. Evolution of accuracy of different principles during the training. (c) Label drifted Digits, SVHN → MNIST. Evolution of accuracy under different label drifts.

**Table 6**
Ablation Study: The performance (%) with and without label shift correction (component I) in CLEF dataset.

| Method | (II + III) | (I + II + III) |
|---|---|---|
| C → I | 77.1 ± 0.1 | 79.3 ± 0.4 |
| I → C | 87.7 ± 0.3 | 89.3 ± 0.2 |
| I → P | 63.9 ± 0.2 | 67.1 ± 0.3 |
| P → I | 78.2 ± 0.2 | 81.2 ± 0.3 |
| C → P | 58.8 ± 0.2 | 60.8 ± 0.2 |
| P → C | 86.6 ± 0.4 | 88.9 ± 0.2 |
| Average | 75.4 | 77.8 |

out different portion % of samples in latter half of classes (i.e digits $5-9$) in source domain. To show the role of label shift correction, we visualize the results in Fig. 4(c). We observe that in a relative large label drift, the re-weighted loss (principle I) improves $\sim 3\%$ performance.

Besides, in CLEF dataset, we also randomly drop out 25% samples in the first half of classes within source, and 25% samples in latter half of classes in target. The empirical result is shown in Table 6, which suggested a significant improvement in the label-shift correction term.

## 6. Related work

*DA theory* An important aspect in DA is to establish the proper theory to understand how it influences the target risk. The most popular approach is based on $\mathcal{H}$-divergence [6], which is set on the deterministic labeling function and binary loss. Then, variants of hypothesis based discrepancy have been proposed such as distribution discrepancy [32], Margin disparity discrepancy [33], etc. However, these theoretical results mainly focus on the relation of feature marginal discrepancy $d(\mathcal{S}(x) \parallel \mathcal{T}(x))$ and the difficulty to analyze various scenarios such as target shift, open-set DA, etc.

An alternative is to adopt the statistical divergence. [34] proposed Rényi-$\alpha$ divergence to measure the feature marginal discrepancy. Then [35,36] analyzed Rényi divergence on the joint source target distribution, with binary and cross entropy loss, respectively. However, they generally focus on covariate shift settings by assuming $\mathcal{S}(y|x) = \mathcal{T}(y|x)$. Moreover, the aforementioned theories did not discuss the inspired practice under the representation learning, which restricts its utility in deep learning. Another popular choice is the Wasserstein distance such as [37], but still focus on the feature marginal distance $W_1(\mathcal{S}(x) \parallel \mathcal{T}(x))$, since the chain rule generally does not hold on Wasserstein distance.

*DA principles for the representation learning* Deriving principles for DA problems in the representation learning is crucial for the real-world applications. From the conventional DA theories such as [6,15,38–40], a small joint optimal risk $\beta$ is important to ensure a small target risk. Therefore, different empirical approaches have speculated various ideas to control a small $\beta$. From the theoretical prospective, [41] adopted Jensen–Shannon divergence to derive the lower bound of $\beta$, indicating necessarily of considering the target shift. However, it is still not clear how the algorithms explicitly guarantee a small $\beta$. Indeed, our work can further extend this by proving a new theoretical upper bound through target shift and feature conditional shift, which enable the possible practice to explicitly control the target risk.

We also notice that [42,43] analyzed feature conditional shift from the causal prospective in RKHS space, which is generally difficult to adapt in the large-scale dataset. From empirical aspects, [4,44–49] proposed various strategies for eliminating conditional shift, which speculated one or two principles to improve the prediction performance. We formally demonstrate the unified three principles, as a way to control the target risk. In addition, our $D_{\text{JS}}$ analysis provides justifications to explain these empirical success e.g., [49], which in fact are *not particularly focused on previous theories but already achieved meaningful results for current deep DA problems.*

## 7. Conclusion

We proposed a new theoretical framework based on Jensen–Shannon divergence for analyzing DA problems. Our theory established bi-directional marginal/conditional shifts for the target risk bound. We further demonstrated its flexibility in various theoretical and algorithmic applications. It is worth mentioning that our theoretical framework is not only suitable for DA, but also extendable to analyzing the real shift problems such as fair representation learning [50,51], individual treatment effect estimation [52]. We anticipate that our theory can open up a pathway towards new algorithm designs for DA, driven by the advantages of fundamental understanding.

## CRediT authorship contribution statement

**Changjian Shui:** Conceptualization, Methodology, Writing – original draft. **Qi Chen:** Methodology, Software. **Jun Wen:** Writing – original draft. **Fan Zhou:** Software, Validation. **Christian Gagné:** Supervision, Writing – review & editing. **Boyu Wang:** Supervision, Writing – review & editing.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Boyu Wang reports financial support was provided by Natural Sciences and Engineering Research Council of Canada. Changjian Shui and Christan Gagné report financial support were provided by Natural Sciences and Engineering Research Council of Canada and CIFAR. Qi Chen reports finanical support was provided by NSERC Discovery Grant and the China Scholarship Council. Fan Zhou reports finanical support was provided by China Scholarship Council.

## Appendix A. $\mathcal{H}$-divergence v.s. Jensen–Shannon divergence

### A.1. Counterexample one

We take the example proposed by [15] (Example 6), which has already computed the $d_{\mathcal{H}}(\mathcal{S}(x), \mathcal{T}(x)) = \xi$. However, since $\text{supp}(\mathcal{S}(x)) \cap \text{supp}(\mathcal{T}(x)) = \emptyset$, $D_{\text{JS}}(\mathcal{S}(x) \parallel \mathcal{T}(x)) = 1$.

### A.2. Counterexample two

We have $\mathcal{S} = \text{Unif}\{1, 2, 3\}$ and $\mathcal{T} = \{\mathbb{P}(X = 1) = \frac{1}{4}, \mathbb{P}(X = 2) = \frac{1}{2}, \mathbb{P}(X = 3) = \frac{1}{4}\}$.

*Computing $d_{\mathcal{H}}$* It is also related to the optimal classification error.

$$\text{err}(h) = \begin{cases} 1/2 & \text{if} \quad t < 1, t > 3 \\ 11/24 & \text{if} \quad 1 < t < 2 \\ 13/24 & \text{if} \quad 2 < t < 3 \end{cases}$$

Then the $\mathcal{H}$ divergence is $d_{\mathcal{H}}(\mathcal{T}(x), \mathcal{S}(x)) = 1 - 2\min_h[\text{err}(h)] = \frac{1}{12} \approx 0.0833$

*Computing $D_{\text{JS}}(\mathcal{T}(x) \parallel \mathcal{S}(x))$* Since the two distributions hold the same support, we can compute the mixture distribution $\mathcal{M} = \{\mathbb{P}(X = 1) = \frac{7}{24}, \mathbb{P}(X = 2) = \frac{5}{12}, \mathbb{P}(X = 3) = \frac{7}{24}\}$, We can compute the Jensen–Shannon divergence with $D(\mathcal{S} \parallel \mathcal{M}) \approx 0.02110$ and $D(\mathcal{T} \parallel \mathcal{M}) \approx 0.02032$.

Then $D_{\text{JS}}(\mathcal{T}(x) \parallel \mathcal{S}(x)) = \frac{1}{2}(0.0211+0.02032) = 0.0207$. In this scenario, the $D_{\text{JS}}(\mathcal{T}(x) \parallel \mathcal{S}(x)) < d_{\mathcal{H}}(\mathcal{T}(x), \mathcal{S}(x))$, therefore, the $D_{\text{JS}}$ cannot be viewed as an upper bound of $d_{\mathcal{H}}$.

## Appendix B. Domain adaptation: Upper bound

We first prove an intermediate lemma:

**Lemma 1.** *Let $Z \in \mathcal{Z}$ be the real valued integrable random variable, let $P$ and $Q$ are two distributions on a common space $\mathcal{Z}$ such that $Q$ is absolutely continuous w.r.t. $P$. If for any function $f$ and $\lambda \in \mathbb{R}$ such that $\mathbb{E}_P[e^{\lambda f(z) - \mathbb{E}_P(f(z))}] < \infty$, then we have:*

$$\lambda(\mathbb{E}_Q f(z) - \mathbb{E}_P f(z)) \leq D_{KL}(Q \parallel P) + \log \mathbb{E}_P[e^{\lambda f(z) - \mathbb{E}_P(f(z))}]$$

*Where $D_{KL}(Q \parallel P)$ is the Kullback–Leibler divergence between distribution $Q$ and $P$, and the equality arrives when $f(z) = \mathbb{E}_P f(z) + \frac{1}{\lambda}\log(\frac{dQ}{dP})$.*

**Proof.** We let $g$ be **any** function such that $\mathbb{E}_P[e^{g(z)}] < \infty$, then we define a random variable $Z_g(z) = \frac{e^{g(z)}}{\mathbb{E}_P[e^{g(z)}]}$, then we can verify that $\mathbb{E}_P(Z_g) = 1$. We assume another distribution $Q$ such that $Q$ (with distribution density $q(z)$) is absolutely continuous w.r.t. $P$ (with distribution density $p(z)$), then we have:

$$\begin{aligned} \mathbb{E}_Q[\log Z_g] &= \mathbb{E}_Q[\log \frac{q(z)}{p(z)} + \log(Z_g \frac{p(z)}{q(z)})] \\ &= D_{\text{KL}}(Q \parallel P) + \mathbb{E}_Q[\log(Z_g \frac{p(z)}{q(z)})] \\ &\leq D_{\text{KL}}(Q \parallel P) + \log \mathbb{E}_Q[\frac{p(z)}{q(z)} Z_g] \\ &= D_{\text{KL}}(Q \parallel P) + \log \mathbb{E}_P[Z_g] \end{aligned}$$

Since $\mathbb{E}_P[Z_g] = 1$ and according to the definition we have $\mathbb{E}_Q[\log Z_g] = \mathbb{E}_Q[g(z)] - \mathbb{E}_Q \log \mathbb{E}_P[e^{g(z)}] = \mathbb{E}_Q[g(z)] - \log \mathbb{E}_P[e^{g(z)}]$ (since $\mathbb{E}_P[e^{g(z)}]$ is a constant w.r.t. $Q$) and we therefore have:

$$\mathbb{E}_Q[g(z)] \leq \log \mathbb{E}_P[e^{g(z)}] + D_{\text{KL}}(Q \parallel P) \tag{6}$$

Since this inequality holds for any function $g$ with finite moment generation function, then we let $g(z) = \lambda(f(z) - \mathbb{E}_P f(z))$ such that $\mathbb{E}_P[e^{f(z) - \mathbb{E}_P f(z)}] < \infty$. Therefore we have $\forall \lambda$ and $f$ we have:

$$\mathbb{E}_Q \lambda(f(z) - \mathbb{E}_P f(z)) \leq D_{\text{KL}}(Q \parallel P) + \log \mathbb{E}_P[e^{\lambda f(z) - \mathbb{E}_P f(z)}]$$

Since we have $\mathbb{E}_Q \lambda(f(z) - \mathbb{E}_P f(z)) = \lambda \mathbb{E}_Q(f(z) - \mathbb{E}_P f(z)) = \lambda(\mathbb{E}_Q f(z) - \mathbb{E}_P f(z))$, therefore we have:

$$\lambda(\mathbb{E}_Q f(z) - \mathbb{E}_P f(z)) \leq D_{\text{KL}}(Q \parallel P) + \log \mathbb{E}_P[e^{\lambda(\mathbb{E}_Q f(z) - \mathbb{E}_P f(z))}]$$

As for the attainment in the equality of Eq. (6), we can simply set $g(z) = \log(\frac{q(z)}{p(z)})$, then we can compute $\mathbb{E}_P[e^{g(z)}] = 1$ and the equality arrives. Therefore in Lemma 1, the equality reaches when $\lambda(f(z)-\mathbb{E}_P f(z)) = \log(\frac{dQ}{dP})$. □

In the classification problem, we define the observation pair $z=(x,y)$. We also define the loss function $\ell(z) = L \circ h(z)$ with deterministic hypothesis $h$ and prediction loss function $L$. Then for abuse of notation, we simply denote the loss function $\ell(z)$ in this part.

Supposing the prediction loss $L$ is bounded with interval $G$ with $G=\max(L)-\min(L)$, then the expected risk in the target domain can be upper bounded by:

$$R_\mathcal{T}(h) \le R_\mathcal{S}(h) + \frac{G}{\sqrt{2}}\sqrt{D_{\mathrm{JS}}(\mathcal{T}\parallel\mathcal{S})}$$

Where $D_{\mathrm{JS}} = \frac{1}{2}\big(D(\mathcal{T}\parallel\frac{1}{2}(\mathcal{T}+\mathcal{S})) + D(\mathcal{S}\parallel\frac{1}{2}(\mathcal{T}+\mathcal{S}))\big)$ is the joint Jensen–Shannon divergence.

**Proof.** According to Lemma 1, $\forall\lambda>0$ we have:

$$\mathbb{E}_Q f(z) - \mathbb{E}_P f(z) \le \frac{1}{\lambda}(\log\mathbb{E}_P\ e^{[\lambda(f(z)-\mathbb{E}_P f(z))]} + D_{\mathrm{KL}}(Q\parallel P)) \quad (7)$$

And $\forall\lambda<0$ we have:

$$\mathbb{E}_Q f(z) - \mathbb{E}_P f(z) \ge \frac{1}{\lambda}(\log\mathbb{E}_P\ e^{[\lambda(f(z)-\mathbb{E}_P f(z))]} + D_{\mathrm{KL}}(Q\parallel P)) \quad (8)$$

Then we introduce an intermediate distribution $\mathcal{M}(z) = \frac{1}{2}(\mathcal{S}(z)+\mathcal{T}(z))$, then $\mathrm{supp}(\mathcal{S}) \subseteq \mathrm{supp}(\mathcal{M})$ and $\mathrm{supp}(\mathcal{T}) \subseteq \mathrm{supp}(\mathcal{M})$, and let $f=\ell$. Since the random variable $\ell$ is bounded through $G=\max(L)-\min(L)$, then according to [53](Chapter 2.1.2), $\ell-\mathbb{E}_P\ell$ is sub-Gaussian with parameter at most $\sigma = \frac{G}{2}$, then we can apply Sub-Gaussian property to bound the log moment generation function:

$$\log\mathbb{E}_P\ e^{[\lambda(\ell(z)-\mathbb{E}_P\ell(z))]} \le \log e^{\frac{\lambda^2\sigma^2}{2}} \le \frac{\lambda^2 G^2}{8}.$$

In Eq. (7), we let $Q=\mathcal{T}$ and $P=\mathcal{M}$, then $\forall\lambda>0$ we have:

$$\mathbb{E}_\mathcal{T}\ \ell(z) - \mathbb{E}_\mathcal{M}\ \ell(z) \le \frac{G^2\lambda}{8} + \frac{1}{\lambda}D_{\mathrm{KL}}(\mathcal{T}\parallel\mathcal{M}) \quad (9)$$

In Eq. (8), we let $Q=\mathcal{S}$ and $P=\mathcal{M}$, then $\forall\lambda<0$ we have:

$$\mathbb{E}_\mathcal{S}\ \ell(z) - \mathbb{E}_\mathcal{M}\ \ell(z) \ge \frac{G^2\lambda}{8} + \frac{1}{\lambda}D_{\mathrm{KL}}(\mathcal{S}\parallel\mathcal{M}) \quad (10)$$

In Eq. (9), we denote $\lambda=\lambda_0>0$ and $\lambda=-\lambda_0<0$ in Eq. (10). Then Eq. (9), Eq. (10) can be reformulated as:

$$\begin{aligned}\mathbb{E}_\mathcal{T}\ \ell(z) - \mathbb{E}_\mathcal{M}\ \ell(z) &\le \frac{G^2\lambda_0}{8} + \frac{1}{\lambda_0}D_{\mathrm{KL}}(\mathcal{T}\parallel\mathcal{M}) \\ \mathbb{E}_\mathcal{M}\ \ell(z) - \mathbb{E}_\mathcal{S}\ \ell(z) &\le \frac{G^2\lambda_0}{8} + \frac{1}{\lambda_0}D_{\mathrm{KL}}(\mathcal{S}\parallel\mathcal{M})\end{aligned} \quad (11)$$

Adding the two inequalities in Eq. (11), we therefore have:

$$\mathbb{E}_\mathcal{T}\ \ell(z) \le \mathbb{E}_\mathcal{S}\ \ell(z) + \frac{1}{\lambda_0}\big(D_{\mathrm{KL}}(\mathcal{S}\parallel\mathcal{M}) + D_{\mathrm{KL}}(\mathcal{T}\parallel\mathcal{M})\big) + \frac{\lambda_0}{4}G^2 \quad (12)$$

Since the inequality holds for $\forall\lambda_0$, then by taking $\lambda_0 = \frac{2}{G}\sqrt{D_{\mathrm{KL}}(\mathcal{S}\parallel\mathcal{M}) + D_{\mathrm{KL}}(\mathcal{T}\parallel\mathcal{M})}$ we finally have:

$$\mathbb{E}_\mathcal{T}\ \ell(z) \le \mathbb{E}_\mathcal{S}\ \ell(z) + \frac{G}{\sqrt{2}}\sqrt{D_{\mathrm{JS}}(\mathcal{T}\parallel\mathcal{S})} \quad □ \quad □ \quad (13)$$

### B.1. Extension to unbounded loss

The advantage of proposed theory can be naturally extended to the unbounded loss.

**Corollary 2** (*Sub-Gaussian Upper Bound*). *If the loss function satisfies $\sigma$-Sub Gaussian property: $\log\mathbb{E}_P\ e^{[\lambda(\ell(z)-\mathbb{E}_P\ell(z))]} \le \frac{\lambda^2\sigma^2}{2}$, then the expected risk in the target domain can be upper bounded by:*

$$R_\mathcal{T}(h) \le R_\mathcal{S}(h) + \sigma\sqrt{2D_{\mathrm{JS}}(\mathcal{T}\parallel\mathcal{S})}$$

**Proof.** The proof is trivial by simply plugging in the Sub-Gaussian condition in the moment generation function. □

**Corollary 3** (*Sub-Gamma Upper Bound*). *If the loss function satisfies $(\sigma,a)$-Sub Gamma property: $\log\mathbb{E}_P\ e^{[\lambda(\ell(z)-\mathbb{E}_P\ell(z))]} \le \frac{\lambda^2\sigma}{2(1-a|\lambda|)}$, for $0<|\lambda|<\frac{1}{a}$. Then the expected risk in the target domain can be upper bounded by:*

$$R_\mathcal{T}(h) \le R_\mathcal{S}(h) + (\sigma+1)\sqrt{2D_{\mathrm{JS}}(\mathcal{T}\parallel\mathcal{S})} + 2aD_{\mathrm{JS}}(\mathcal{T}\parallel\mathcal{S})$$

**Proof.** For the same step for the moment generation function, by taking $\lambda_0 \in (0,\frac{1}{a})$, then analogously we have:

$$\begin{aligned}\mathbb{E}_\mathcal{T}\ \ell(z) - \mathbb{E}_\mathcal{M}\ \ell(z) &\le \frac{\lambda_0\sigma}{2(1-a\lambda_0)} + \frac{1}{\lambda_0}D_{\mathrm{KL}}(\mathcal{T}\parallel\mathcal{M}) \\ \mathbb{E}_\mathcal{M}\ \ell(z) - \mathbb{E}_\mathcal{S}\ \ell(z) &\le \frac{\lambda_0\sigma}{2(1-a\lambda_0)} + \frac{1}{\lambda_0}D_{\mathrm{KL}}(\mathcal{S}\parallel\mathcal{M})\end{aligned}$$

Therefore we have

$$\begin{aligned}\mathbb{E}_\mathcal{T}\ \ell(z) - \mathbb{E}_\mathcal{S}\ \ell(z) &\le \frac{\lambda_0\sigma}{(1-a\lambda_0)} + \frac{1}{\lambda_0}\,(D_{\mathrm{KL}}(\mathcal{T}\parallel\mathcal{M}) + D_{\mathrm{KL}}(\mathcal{S}\parallel\mathcal{M})) \\ &= \frac{\lambda_0\sigma}{(1-a\lambda_0)} + \frac{1}{\lambda_0}\,\big(2D_{\mathrm{JS}}(\mathcal{T}\parallel\mathcal{S})\big)\end{aligned}$$

We let $\lambda_0 = \frac{\sqrt{2D_{\mathrm{JS}}(\mathcal{T}\parallel\mathcal{S})}}{\sigma+a\sqrt{2D_{\mathrm{JS}}(\mathcal{T}\parallel\mathcal{S})}} \in (0,\frac{1}{a})$ and we can simplify the upper bound as:

$$\mathbb{E}_\mathcal{T}\ \ell(z) - \mathbb{E}_\mathcal{S}\ \ell(z) \le (\sigma+1)\sqrt{2D_{\mathrm{JS}}(\mathcal{T}\parallel\mathcal{S})} + 2aD_{\mathrm{JS}}(\mathcal{T}\parallel\mathcal{S}) \quad □$$

The extended upper bounds can be much tighter than the conclusion in Theorem 1, particularly when the loss is in a large range with a small variance.

## Appendix C. Domain adaptation theory: Lower bound

We firstly introduce several information theoretical tools:

**Lemma 2** (*Pinsker's Inequality*).*If $P$ and $Q$ are two probability distribution on the measurable space $(\Omega,\mathcal{F})$, then*

$$\mathrm{TV}(P,Q) \le \sqrt{2D_{KL}(P\parallel Q)}$$

*Where $D(P\parallel Q)_{KL}$ is the Kullback–Leibler divergence between distribution $P$ and $Q$ and $TV(P\parallel Q) = \sum_z |P(z)-Q(z)|$*

**Lemma 3** ([17]). *[$f$-divergence data processing inequality] Consider a channel that produces $Y$ given $X$ on the deterministic function $g$. If $P_Y$ is the distribution of $Y$ when $X$ is generated by $P_X$ and $Q_Y$ is the distribution of $Y$ when $X$ is generated by $Q_X$, then for any $f$-divergence $D_f(\cdot\parallel\cdot)$:*

$$D_f(P_Y\parallel Q_Y) \le D_f(P_X\parallel Q_X)$$

If we restrict the zero–one loss $L\in\{0,1\}$, then we can prove the target risk be lower bounded by:

$$R_\mathcal{T}(h) \ge R_\mathcal{S}(h) - \sqrt{D_{\mathrm{JS}}(\mathcal{T}\parallel\mathcal{S})}$$

**Proof.** Again we denote the observation pair $z=(x,y)$. For abuse of notation, we simply denote the loss function $\ell = L\circ h$ with $\ell\in\{0,1\}$.

According to $f$-divergence data processing inequality, if we set the deterministic function $g$ as $g(Z) = \mathbf{1}_E(Z)$ for any event $E$, then

$Y$ is Bernoulli distribution with parameter $P(E)$ or $Q(E)$ and the data processing inequality becomes:

$$D_f(\text{Bern}(P(E)) \parallel \text{Bern}(Q(E))) \leq D_f(P_Z \parallel Q_Z)$$

If we define the event $E$ as we make an error in the prediction (a.k.a $l(z) = 1$), then $P(E) = P(\text{making an error}) = E_P\mathbf{1}\{\text{making an error}\} = \mathbb{E}_P[\ell(z)]$. Therefore we have:

$$D_f(\text{Bern}(\mathbb{E}_P[\ell(z)]) \parallel \text{Bern}(\mathbb{E}_Q[\ell(z)])) \leq D_f(P_Z \parallel Q_Z)$$

Again we introduce the intermediate distribution $\mathcal{M} = \frac{1}{2}(\mathcal{S}+\mathcal{T})$. According to the data processing inequality on the expectation of random variables, if we adopt KL divergence by letting $f(t) = t\log(t)$, then we have:

$$D_{\text{KL}}(\text{Bern}(\mathbb{E}_{\mathcal{T}}[\ell(z)]) \parallel \text{Bern}(\mathbb{E}_{\mathcal{M}}[\ell(z)])) \leq D_{\text{KL}}(\mathcal{T} \parallel \mathcal{M})$$
$$D_{\text{KL}}(\text{Bern}(\mathbb{E}_{\mathcal{S}}[\ell(z)]) \parallel \text{Bern}(\mathbb{E}_{\mathcal{M}}[\ell(z)])) \leq D_{\text{KL}}(\mathcal{S} \parallel \mathcal{M})$$

We notice $\mathbb{E}_{\mathcal{T}}(\ell(z)) \in [0,1]$, $\mathbb{E}_{\mathcal{S}}(\ell(z)) \in [0,1]$. Then we can adopt Pinsker's inequality by treating the expected value as the Bernoulli distribution parameters. Then we can compute their Total Variation (TV) distance.

$$\text{TV}(\text{Bern}(p), \text{Bern}(q)) = |p-q| + |1-p-1+q| = 2|p-q|$$

Then we have:

$$\begin{aligned}
&2|\mathbb{E}_{\mathcal{T}}[\ell(z)] - \mathbb{E}_{\mathcal{M}}[\ell(z)]| = \text{TV}(\text{Bern}(p), \text{Bern}(q))\\
&\leq \sqrt{2D_{\text{KL}}(\text{Bern}(\mathbb{E}_{\mathcal{T}}[\ell(z)]) \parallel \text{Bern}(\mathbb{E}_{\mathcal{M}}[\ell(z)]))}\\
&\leq \sqrt{2D_{\text{KL}}(\mathcal{T} \parallel \mathcal{M})}
\end{aligned}$$

Similarity we have $2|\mathbb{E}_{\mathcal{S}}[\ell(z)] - \mathbb{E}_{\mathcal{M}}[\ell(z)]| \leq \sqrt{2D_{\text{KL}}(\mathcal{S} \parallel \mathcal{M})}$. We add these two item together. Then We adopt the inequality $\sqrt{a}+\sqrt{b} \leq \sqrt{2(a+b)}$ with $a \geq 0$ and $b \geq 0$, then we have

$$\sqrt{D_{\text{KL}}(\mathcal{T} \parallel \mathcal{M})} + \sqrt{D_{\text{KL}}(\mathcal{S} \parallel \mathcal{M})} \leq 2\sqrt{D_{\text{JS}}(\mathcal{T} \parallel \mathcal{S})}.$$

We also have

$$\begin{aligned}
&2|\mathbb{E}_{\mathcal{S}}[\ell(z)] - \mathbb{E}_{\mathcal{M}}[\ell(z)]| + 2|\mathbb{E}_{\mathcal{T}}[\ell(z)] - \mathbb{E}_{\mathcal{M}}[\ell(z)]|\\
&\geq 2|\mathbb{E}_{\mathcal{S}}[\ell(z)] - \mathbb{E}_{\mathcal{M}}[\ell(z)] - \mathbb{E}_{\mathcal{T}}[\ell(z)] + \mathbb{E}_{\mathcal{M}}[\ell(z)]|\\
&= 2|\mathbb{E}_{\mathcal{S}}[\ell(z)] - \mathbb{E}_{\mathcal{T}}[\ell(z)]|
\end{aligned}$$

Given the aforementioned results, we have the following the two side inequality:

$$|\mathbb{E}_{\mathcal{S}}[\ell(z)] - \mathbb{E}_{\mathcal{T}}[\ell(z)]| \leq \sqrt{D_{\text{JS}}(\mathcal{T} \parallel \mathcal{S})}$$

We have $-\sqrt{D_{\text{JS}}(\mathcal{T} \parallel \mathcal{S})} \leq \mathbb{E}_{\mathcal{T}}[\ell(z)] - \mathbb{E}_{\mathcal{S}}[\ell(z)] \leq \sqrt{D_{\text{JS}}(\mathcal{T} \parallel \mathcal{S})}$ and finally we have the lower bound:

$$\mathbb{E}_{\mathcal{T}}[\ell(z)] \geq \mathbb{E}_{\mathcal{S}}[\ell(z)] - \sqrt{D_{\text{JS}}(\mathcal{T} \parallel \mathcal{S})}$$

*Remark* We should point out the derived upper bound is looser and restrictive than that we derived from Theorem 1, with a scale $\frac{1}{\sqrt{2}}$ when we restrict the loss in $\{0,1\}$ and Theorem 1 can be extended to any bounded loss while this proof **cannot**. □

## Appendix D. Joint JS divergence decomposition

In this section, we will provide an upper bound of the chain rule in Jensen–Shannon divergence. According to the definition of Jensen–Shannon divergence and the chain rule of KL divergence we have:

$$\begin{aligned}
&2D_{\text{JS}}(\mathcal{T}(x,y) \parallel \mathcal{S}(x,y))\\
&= D_{\text{KL}}(\mathcal{T}(x,y) \parallel \mathcal{M}(x,y)) + D_{\text{KL}}(\mathcal{S}(x,y) \parallel \mathcal{M}(x,y))\\
&= D_{\text{KL}}(\mathcal{T}(x) \parallel \mathcal{M}(x)) + \mathbb{E}_{x\sim\mathcal{T}(x)}D_{\text{KL}}(\mathcal{T}(y|x) \parallel \mathcal{M}(y|x))\\
&+ D_{\text{KL}}(\mathcal{S}(x) \parallel \mathcal{M}(x)) + \mathbb{E}_{x\sim\mathcal{S}(x)}D_{\text{KL}}(\mathcal{S}(y|x) \parallel \mathcal{M}(y|x))\\
&= 2D_{\text{JS}}(\mathcal{T}(x) \parallel \mathcal{S}(x)) + \mathbb{E}_{x\sim\mathcal{T}(x)}D_{\text{KL}}(\mathcal{T}(y|x) \parallel \mathcal{M}(y|x))\\
&+ \mathbb{E}_{x\sim\mathcal{S}(x)}D_{\text{KL}}(\mathcal{S}(y|x) \parallel \mathcal{M}(y|x))
\end{aligned}$$

In general, for continuous random variable, the $D_{\text{KL}}$ divergence does not exist an exact upper bound. While we can simple upper bound these by adding two complementary terms.

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{T}(x)}D_{\text{KL}}(\mathcal{T}(y|x) \parallel \mathcal{M}(y|x))\\
&\leq \mathbb{E}_{x\sim\mathcal{T}(x)}D_{\text{KL}}(\mathcal{T}(y|x) \parallel \mathcal{M}(y|x)) + \mathbb{E}_{x\sim\mathcal{T}(x)}D_{\text{KL}}(\mathcal{S}(y|x) \parallel \mathcal{M}(y|x))\\
&= 2\mathbb{E}_{x\sim\mathcal{T}(x)}D_{\text{JS}}(\mathcal{T}(y|x) \parallel \mathcal{S}(y|x))
\end{aligned}$$

## Appendix E. Target intrinsic error upper bound

If $H(Y_s|X_s) \leq \epsilon$, the source target marginal and conditional distribution are close $D_{\text{JS}}(\mathcal{S}(x) \parallel \mathcal{T}(x)) \leq \delta_1$, $\forall x$, we have $D_{\text{JS}}(\mathcal{S}(y|x) \parallel \mathcal{T}(y|x)) \leq \delta_2$. Then the target distribution conditional entropy can be upper bounded by:

$$H(Y_t|X_t) \leq \epsilon + \sqrt{\frac{\delta_2}{2}} + \frac{\sqrt{\delta_1}}{2}\log|\mathcal{Y}|$$

**Proof.** Since $\frac{1}{2}\text{TV}(P,Q)^2 \leq D_{\text{JS}}(P \parallel Q) \leq TV(P,Q)$ [54], then for $\forall x$ we have:

$$\parallel \mathcal{S}(y|x) - \mathcal{T}(y|x) \parallel_1 \leq \sqrt{2\delta_2}$$

Then for conditional entropy for the target distribution, we have:

$$\begin{aligned}
&H(Y_t|X_t) = \mathbb{E}_{x\sim\mathcal{T}(x)}H(Y_t|X_t = x)\\
&= \mathbb{E}_{x\sim\mathcal{T}(x)}(H(Y_t|X=x) - H(Y_s|X=x)) + \mathbb{E}_{x\sim\mathcal{T}(x)}H(Y_s|X=x)\\
&\leq \mathbb{E}_{x\sim\mathcal{T}(x)}|H(Y_t|X=x) - H(Y_s|X=x)| + \mathbb{E}_{x\sim\mathcal{T}(x)}H(Y_s|X=x)
\end{aligned}$$

Since the Entropy function is $\frac{1}{2}$ Lipschitz w.r.t. $L_1$ norm, then we have

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{T}(x)}|H(Y_t|X=x) - H(Y_s|X=x)|\\
&\leq \mathbb{E}_{x\sim\mathcal{T}(x)}\frac{1}{2} \parallel \mathcal{T}(y|x) - \mathcal{S}(y|x) \parallel_1 \leq \sqrt{\frac{\delta_2}{2}}
\end{aligned}$$

Then we need to bound $\mathbb{E}_{x\sim\mathcal{T}(x)}H(Y_s|X=x)$,

$$\begin{aligned}
&E_{x\sim\mathcal{T}(x)}H(Y_s|X=x) = E_{x\sim\mathcal{S}(x)}H(Y_s|X=x) +\\
&E_{x\sim\mathcal{T}(x)}H(Y_s|X=x) - E_{x\sim\mathcal{S}(x)}H(Y_s|X=x)\\
&\leq \epsilon + E_{x\sim\mathcal{T}(x)}H(Y_s|X=x) - E_{x\sim\mathcal{S}(x)}H(Y_s|X=x)
\end{aligned}$$

We still adopt the conclusion when we proof Theorem 1, i.e the transport inequality of the gaps of same function under different marginal distribution measures by assuming $z = x$. We can compute $G = H(Y_s|X=x) \leq H(Y_s) \leq \log|\mathcal{Y}|$, then we have:

$$\begin{aligned}
E_{x\sim\mathcal{T}(x)}H(Y_s|X=x) &\leq \epsilon + \frac{\log|\mathcal{Y}|}{\sqrt{2}}\sqrt{D_{\text{JS}}(\mathcal{T}(x) \parallel \mathcal{S}(x))}\\
&\leq \epsilon + \sqrt{\frac{\delta_1}{2}}\log|\mathcal{Y}|
\end{aligned}$$

Putting all them together we have the aforementioned conclusion. □

## Appendix F. Inherent difficulty for controlling label conditional shift

### F.1. Lower bound of label conditional shift

We can prove the label-conditional shift can be lower bounded by:

$$\begin{aligned}
&\mathbb{E}_{z\sim\hat{\mathcal{T}}(z)}D_{\text{JS}}(\hat{\mathcal{S}}(y|z) \parallel \hat{\mathcal{T}}(y|z)) + \mathbb{E}_{z\sim\hat{\mathcal{S}}(z)}D_{\text{JS}}(\hat{\mathcal{S}}(y|z) \parallel \hat{\mathcal{T}}(y|z))\\
&\geq 2\left(\sqrt{D_{\text{JS}}(\hat{\mathcal{T}}(y) \parallel \hat{\mathcal{S}}(y))} - \sqrt{D_{\text{JS}}(\hat{\mathcal{S}}(z) \parallel \hat{\mathcal{T}}(z))}\right)^2
\end{aligned}$$

We notice the square form of Jensen–Shannon divergence is the valid statistical distance. Then we have:

$$\sqrt{D_{\text{JS}}(\hat{\mathcal{T}}(y) \parallel \hat{\mathcal{S}}(y))} = \sqrt{D_{\text{JS}}(\sum_z \hat{\mathcal{T}}(y|z)\hat{\mathcal{T}}(z) \parallel \sum_z \hat{\mathcal{S}}(y|z)\hat{\mathcal{S}}(z))}$$

$$\leq \sqrt{D_{\text{JS}}(\sum_z \hat{\mathcal{T}}(y|z)\hat{\mathcal{S}}(z) \parallel \sum_z \hat{\mathcal{S}}(y|z)\hat{\mathcal{S}}(z))} +$$

$$\sqrt{D_{\text{JS}}(\sum_z \hat{\mathcal{T}}(y|z)\hat{\mathcal{S}}(z) \parallel \sum_z \hat{\mathcal{T}}(y|z)\hat{\mathcal{T}}(z))}$$

$$\leq \sqrt{\mathbb{E}_{z\sim\hat{\mathcal{S}}(z)} D_{\text{JS}}(\hat{\mathcal{S}}(y|z) \parallel \hat{\mathcal{T}}(y|z))} + \sqrt{D_{\text{JS}}(\hat{\mathcal{S}}(z) \parallel \hat{\mathcal{T}}(z))}$$

We derive the inequality according to (1) Jensen–Shannon distance is a valid statistical metric; (2) The convex property of the Jensen–Shannon divergence w.r.t. the empirical distribution; (3) The $f$-divergence data-processing inequality. Then we have the following the results:

$$\mathbb{E}_{z\sim\hat{\mathcal{S}}(z)} D_{\text{JS}}(\hat{\mathcal{S}}(y|z) \parallel \hat{\mathcal{T}}(y|z)) \geq$$
$$\left(\sqrt{D_{\text{JS}}(\hat{\mathcal{T}}(y) \parallel \hat{\mathcal{S}}(y))} - \sqrt{D_{\text{JS}}(\hat{\mathcal{S}}(z) \parallel \hat{\mathcal{T}}(z))}\right)^2$$

We can analogously derive:

$$\mathbb{E}_{z\sim\hat{\mathcal{T}}(z)} D_{\text{JS}}(\hat{\mathcal{S}}(y|z) \parallel \hat{\mathcal{T}}(y|z)) \geq$$
$$\left(\sqrt{D_{\text{JS}}(\hat{\mathcal{T}}(y) \parallel \hat{\mathcal{S}}(y))} - \sqrt{D_{\text{JS}}(\hat{\mathcal{S}}(z) \parallel \hat{\mathcal{T}}(z))}\right)^2$$

By combining these two terms we finally derive the lower bounded . Which exactly recovers the result of [41]: over-matching the marginal distribution divergence to zero can increase this lower bound of the third term.

## Appendix G. New practical principles

In this section, we firstly prove the lower bound in context of conditional distribution matching. We demonstrate that in the presence of conditional distribution matching, we still need to control the label shift term to control a small lower bound.

### G.1. Necessity of considering label shift

In this section, we suppose there exist a more general stochastic representation learning function $g$ with a conditional probability distribution $g(z|x)$. Then the marginal distribution and conditional distribution w.r.t. latent variable can be reformulated as $\mathcal{S}(z) = \int_x g(z|x)\mathcal{S}(x)dx$ and $\mathcal{S}(z|y) = \int_x g(z|x)\mathcal{S}(x|Y=y)dx$.

If $\forall$ classifier $h$, feature function $g$, and label $y \in \mathcal{Y} = \{-1,+1\}$ such that semantic conditional distribution is matched: $D_{\text{JS}}(\mathcal{S}(z|y), \mathcal{T}(z|y)) = 0$, then the target risk can be bounded:

$$R_{\mathcal{S}}(h\circ g) - \sqrt{2D_{\text{JS}}(\mathcal{S}(y),\mathcal{T}(y))} \leq R_{\mathcal{T}}(h\circ g)$$
$$\leq R_{\mathcal{S}}(h\circ g) + \sqrt{2D_{\text{JS}}(\mathcal{S}(y),\mathcal{T}(y))}$$

Where $R_{\mathcal{S}}(h\circ g) = R_{\mathcal{S}}(h(g(x),y))$ the expected risk over the classifier $h$ and feature learner $g$.

**Proof.** For simplifying the analysis, we only focus on the binary classification with margin style loss with $L(h(z),y) = L(yh(z))$, including $0-1$ loss, hinge loss, logistic loss, etc.). Throughout the whole analysis, we will simply adopt the $0-1$ loss. We additionally define the following distributions:

$$\mu^{\mathcal{S}}(z) = \mathcal{S}(Y=1,Z=z) = \mathcal{S}(Y=1)\mathcal{S}(Z=z|Y=1)$$
$$\pi^{\mathcal{S}}(z) = \mathcal{S}(Y=-1,Z=z) = \mathcal{S}(Y=-1)\mathcal{S}(Z=z|Y=-1)$$
$$\mu^{\mathcal{T}}(z) = \mathcal{T}(Y=1,Z=z) = \mathcal{T}(Y=1)\mathcal{T}(Z=z|Y=1)$$
$$\pi^{\mathcal{T}}(z) = \mathcal{T}(Y=-1,Z=z) = \mathcal{T}(Y=-1)\mathcal{T}(Z=z|Y=-1)$$

Then in the source distribution and target distribution for the common feature extractor $Q$ and hypothesis $h$, we have:

$$R_{\mathcal{S}}(h\circ g) = \mathbb{E}_{\mathcal{S}}\mathbf{1}\{yh(z)\leq 0\}$$

$$R_{\mathcal{T}}(h\circ g) = \mathbb{E}_{\mathcal{T}}\mathbf{1}\{yh(z)\leq 0\}$$

According to [55], the risk can be reformulated as

$$R_{\mathcal{S}}(h\circ g) = \sum_z \mathbf{1}\{h(z)\leq 0\}\mu^{\mathcal{S}}(z) + \mathbf{1}\{h(z)>0\}\pi^{\mathcal{S}}(z)$$
$$R_{\mathcal{T}}(h\circ g) = \sum_z \mathbf{1}\{h(z)\leq 0\}\mu^{\mathcal{T}}(z) + \mathbf{1}\{h(z)>0\}\pi^{\mathcal{T}}(z)$$

Then we have:

$$R_{\mathcal{T}}(h\circ g) - R_{\mathcal{S}}(h\circ g) \geq \sum_z \min\{\mu^{\mathcal{T}}(z)-\mu^{\mathcal{S}}(z), \pi^{\mathcal{T}}(z)-\pi^{\mathcal{S}}(z)\}$$

If we define the conditional distribution matching as there exists a distribution $\exists g^{\star}$ such that $\mathcal{S}(z|y) = \mathcal{T}(z|y) = \mathcal{D}(z|y)$, then we can simplify as

$$\sum_z \min\{\mu^{\mathcal{T}}(z)-\mu^{\mathcal{S}}(z), \pi^{\mathcal{T}}(z)-\pi^{\mathcal{S}}(z)\}$$
$$\geq -|\mathcal{S}(y=1)-\mathcal{T}(y=1)|\sum_z \max\{\mathcal{D}(z|y=1), \mathcal{D}(z|y=-1)\}$$
$$= -\frac{1}{2}d_{\text{TV}}(\mathcal{S}(y),\mathcal{T}(y))\frac{1}{2}1 + d_{\text{TV}}(\mathcal{D}(z|y=1),\mathcal{D}(z|y=-1))$$
$$\geq -\frac{1}{2}d_{\text{TV}}(\mathcal{S}(y),\mathcal{T}(y))\frac{1}{2}(1+1) = -\frac{1}{2}d_{\text{TV}}(\mathcal{S}(y),\mathcal{T}(y))$$
$$\geq -\sqrt{2D_{\text{JS}}(\mathcal{S}(y),\mathcal{T}(y))}$$

As for the upper bound, since we have:

$$R_{\mathcal{T}}(h\circ g) - R_{\mathcal{S}}(h\circ g) \leq \sum_z \max\{\mu^{\mathcal{T}}(z)-\mu^{\mathcal{S}}(z), \pi^{\mathcal{T}}(z)-\pi^{\mathcal{S}}(z)\}$$

Given the conditional shift, we have:

$$\sum_z \max\{\mu^{\mathcal{T}}(z)-\mu^{\mathcal{S}}(z), \pi^{\mathcal{T}}(z)-\pi^{\mathcal{S}}(z)\}$$
$$\leq |\mathcal{S}(y=1)-\mathcal{T}(y=1)|\sum_z \max\{\mathcal{D}(z|y=1), \mathcal{D}(z|y=-1)\}$$
$$= \frac{1}{2}d_{\text{TV}}(\mathcal{S}(y),\mathcal{T}(y))\frac{1}{2}(1 + d_{\text{TV}}(\mathcal{D}(z|y=1),\mathcal{D}(z|y=-1)))$$
$$\leq \frac{1}{2}d_{\text{TV}}(\mathcal{S}(y),\mathcal{T}(y))\frac{1}{2}(1+1) = \frac{1}{2}d_{\text{TV}}(\mathcal{S}(y),\mathcal{T}(y))$$
$$\leq \sqrt{2D_{\text{JS}}(\mathcal{S}(y),\mathcal{T}(y))}$$

Finally we have the two side bound. □

### G.2. Detecting poor pseudo-label

We can prove if we have poor pseudo-label, the marginal divergence can be very large. If we assume $D_{\text{JS}}(\hat{\mathcal{T}}(y) \parallel \hat{\mathcal{T}}_p(y)) = P$, and small source prediction error $D_{\text{JS}}(\hat{\mathcal{S}}(y) \parallel \hat{\mathcal{S}}_p(y)) \leq \epsilon_1$ and small source target ground truth distribution $D_{\text{JS}}(\hat{\mathcal{S}}(y) \parallel \hat{\mathcal{T}}(y)) \leq \epsilon_2$, then we can prove

$$D_{\text{JS}}(\hat{\mathcal{S}}(z) \parallel \hat{\mathcal{T}}(z)) \geq (\sqrt{P} - \sqrt{\epsilon_1} - \sqrt{\epsilon_2})^2$$

**Proof.** Since in the DA, we adopt the same classifier $h$ to predict both domains, the empirical label prediction output distribution (pseudo-label distribution) is defined as:

$$\hat{\mathcal{S}}_p(y) = \sum_z h(y|z)\hat{\mathcal{S}}(z) \qquad \hat{\mathcal{T}}_p(y) = \sum_z h(y|z)\hat{\mathcal{T}}(z)$$

According to the $f$-divergence data-processing inequality, we have:

$$D_{\mathrm{JS}}(\hat{\mathcal{S}}(z) \parallel \hat{\mathcal{T}}(z)) \geq D_{\mathrm{JS}}(\hat{\mathcal{S}}_p(y) \parallel \hat{\mathcal{T}}_p(y))$$

Since Jensen–Shannon distance is a valid statistical distance, then we have:

$$\sqrt{D_{\mathrm{JS}}(\hat{\mathcal{S}}_p(y) \parallel \hat{\mathcal{T}}_p(y))} + \sqrt{D_{\mathrm{JS}}(\hat{\mathcal{S}}_p(y) \parallel \hat{\mathcal{S}}(y))}$$
$$+ \sqrt{D_{\mathrm{JS}}(\hat{\mathcal{S}}(y) \parallel \hat{\mathcal{T}}(y))} \geq \sqrt{D_{\mathrm{JS}}(\hat{\mathcal{T}}(y) \parallel \hat{\mathcal{T}}_p(y))} = \sqrt{P}$$

Since we have a small source prediction error, a small empirical label shift, then we have:

$$\sqrt{D_{\mathrm{JS}}(\hat{\mathcal{S}}_p(y) \parallel \hat{\mathcal{T}}_p(y))} \geq \sqrt{P} - \sqrt{\epsilon_1} - \sqrt{\epsilon_2}$$

Combining together we have $D_{\mathrm{JS}}(\hat{\mathcal{S}}(z) \parallel \hat{\mathcal{T}}(z)) \geq (\sqrt{P} - \sqrt{\epsilon_1} - \sqrt{\epsilon_2})^2$ □

## References

[1] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.

[2] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, J. Mach. Learn. Res. 17 (1) (2016) 2030–2096.

[3] H. Guo, R. Pasunuru, M. Bansal, Multi-source domain adaptation for text classification via DistanceNet-bandits, 2020, arXiv preprint arXiv:2001.04362.

[4] Y. Li, M. Murias, S. Major, G. Dawson, D. Carlson, On target shift in adversarial domain adaptation, in: The 22nd International Conference on Artificial Intelligence and Statistics, 2019, pp. 616–625.

[5] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, Analysis of representations for domain adaptation, in: Advances in Neural Information Processing Systems, 2007, pp. 137–144.

[6] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J.W. Vaughan, A theory of learning from different domains, Mach. Learn. 79 (1–2) (2010) 151–175.

[7] I. Redko, E. Morvant, A. Habrard, M. Sebban, Y. Bennani, A survey on domain adaptation theory, 2020, arXiv preprint arXiv:2004.11829.

[8] M. Long, Y. Cao, J. Wang, M.I. Jordan, Learning transferable features with deep adaptation networks, 2015, arXiv preprint arXiv:1502.02791.

[9] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7167–7176.

[10] P. Panareda Busto, J. Gall, Open set domain adaptation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 754–763.

[11] Z. Cao, L. Ma, M. Long, J. Wang, Partial adversarial domain adaptation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 135–150.

[12] K. You, M. Long, Z. Cao, J. Wang, M.I. Jordan, Universal domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2720–2729.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[14] S. Nowozin, B. Cseke, R. Tomioka, f-gan: Training generative neural samplers using variational divergence minimization, in: Advances in Neural Information Processing Systems, 2016, pp. 271–279.

[15] S. Ben-David, T. Lu, T. Luu, D. Pál, Impossibility theorems for domain adaptation, in: International Conference on Artificial Intelligence and Statistics, 2010, pp. 129–136.

[16] S. Hanneke, S. Kpotufe, On the value of target data in transfer learning, in: Advances in Neural Information Processing Systems, 2019, pp. 9867–9877.

[17] Y. Polyanskiy, Y. Wu, Lecture notes on information theory, 2019.

[18] A. Achille, S. Soatto, Emergence of invariance and disentanglement in deep representations, J. Mach. Learn. Res. 19 (1) (2018) 1947–1980.

[19] P. Zhang, H. Wang, N. Naik, C. Xiong, R. Socher, DIME: An information-theoretic difficulty measure for AI datasets, 2020, URL: https://openreview.net/forum?id=H1lNb0NtPH.

[20] M. Long, Z. Cao, J. Wang, M.I. Jordan, Conditional adversarial domain adaptation, in: Advances in Neural Information Processing Systems, 2018, pp. 1640–1650.

[21] C. Cortes, Y. Mansour, M. Mohri, Learning bounds for importance weighting, in: Advances in Neural Information Processing Systems, 2010, pp. 442–450.

[22] Z. Lipton, Y.X. Wang, A. Smola, Detecting and correcting for label shift with black box predictors, in: International Conference on Machine Learning, 2018, pp. 3122–3130.

[23] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A.Y. Ng, Reading digits in natural images with unsupervised feature learning, 2011.

[24] J.J. Hull, A database for handwritten text recognition research, IEEE Trans. Pattern Anal. Mach. Intell. 16 (5) (1994) 550–554.

[25] K. Saenko, B. Kulis, M. Fritz, T. Darrell, Adapting visual category models to new domains, in: European Conference on Computer Vision, Springer, 2010, pp. 213–226.

[26] M. Villegas, H. Müller, A. Gilbert, L. Piras, J. Wang, K. Mikolajczyk, A.G.S. De Herrera, S. Bromuri, M.A. Amin, M.K. Mohammed, et al., General overview of imageCLEF at the CLEF 2015 labs, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2015, pp. 444–461.

[27] J. Blitzer, M. Dredze, F. Pereira, Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 440–447.

[28] M. Chen, Z. Xu, K.Q. Weinberger, F. Sha, Marginalized denoising autoencoders for domain adaptation, in: Proceedings of the 29th International Coference on International Conference on Machine Learning, 2012, pp. 1627–1634.

[29] C. Shui, Z. Li, J. Li, C. Gagné, C.X. Ling, B. Wang, Aggregating from multiple target-shifted sources, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 9638–9648, URL: https://proceedings.mlr.press/v139/shui21a.html.

[30] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.

[31] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[32] C. Cortes, M. Mohri, A.M. Medina, Adaptation based on generalized discrepancy, J. Mach. Learn. Res. 20 (1) (2019) 1–30.

[33] Y. Zhang, T. Liu, M. Long, M. Jordan, Bridging theory and algorithm for domain adaptation, in: International Conference on Machine Learning, 2019, pp. 7404–7413.

[34] Y. Mansour, M. Mohri, A. Rostamizadeh, Multiple source adaptation and the Rényi divergence, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2009, pp. 367–374.

[35] P. Germain, A. Habrard, F. Laviolette, E. Morvant, A new PAC-Bayesian perspective on domain adaptation, in: International Conference on Machine Learning, 2016, pp. 859–868.

[36] J. Hoffman, M. Mohri, N. Zhang, Algorithms and theory for multiple-source adaptation, in: Advances in Neural Information Processing Systems, 2018, pp. 8246–8256.

[37] J. Shen, Y. Qu, W. Zhang, Y. Yu, Wasserstein distance guided representation learning for domain adaptation, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[38] S. Ben-David, R. Urner, Domain adaptation–can quantity compensate for quality? Ann. Math. Artif. Intell. 70 (3) (2014) 185–202.

[39] P. Germain, A. Habrard, F. Laviolette, E. Morvant, A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers, in: International Conference on Machine Learning, 2013, pp. 738–746.

[40] F. Johansson, D. Sontag, R. Ranganath, Support and invertibility in domain-invariant representations, in: The 22nd International Conference on Artificial Intelligence and Statistics, 2019, pp. 527–536.

[41] H. Zhao, R.T. Des Combes, K. Zhang, G. Gordon, On learning invariant representations for domain adaptation, in: International Conference on Machine Learning, 2019, pp. 7523–7532.

[42] K. Zhang, B. Schölkopf, K. Muandet, Z. Wang, Domain adaptation under target and conditional shift, in: International Conference on Machine Learning, 2013, pp. 819–827.

[43] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, B. Schölkopf, Domain adaptation with conditional transferable components, in: International Conference on Machine Learning, 2016, pp. 2839–2848.

[44] S. Tan, X. Peng, K. Saenko, Generalized domain adaptation with covariate and label shift CO-alignment, 2019, arXiv preprint arXiv:1910.10320.

[45] M. Long, J. Wang, G. Ding, J. Sun, P.S. Yu, Transfer feature learning with joint distribution adaptation, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2200–2207.

[46] K. Saito, Y. Ushiku, T. Harada, Asymmetric tri-training for unsupervised domain adaptation, in: Proceedings of the 34th International Conference on Machine Learning, Vol. 70, JMLR. org, 2017, pp. 2988–2997.

[47] C. Chen, W. Xie, W. Huang, Y. Rong, X. Ding, Y. Huang, T. Xu, J. Huang, Progressive feature alignment for unsupervised domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 627–636.

[48] S. Xie, Z. Zheng, L. Chen, C. Chen, Learning semantic representations for unsupervised domain adaptation, in: International Conference on Machine Learning, 2018, pp. 5423–5432.

[49] R. Cai, Z. Li, P. Wei, J. Qiao, K. Zhang, Z. Hao, Learning disentangled semantic representation for domain adaptation, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, 2019, pp. 2060–2066.

[50] C. Louizos, K. Swersky, Y. Li, M. Welling, R. Zemel, The variational fair autoencoder, 2015, arXiv preprint arXiv:1511.00830.

[51] H. Edwards, A. Storkey, Censoring representations with an adversary, 2015, arXiv preprint arXiv:1511.05897.

[52] U. Shalit, F.D. Johansson, D. Sontag, Estimating individual treatment effect: generalization bounds and algorithms, in: Proceedings of the 34th International Conference on Machine Learning, Vol. 70, JMLR. org, 2017, pp. 3076–3085.

[53] M.J. Wainwright, High-Dimensional Statistics: A Non-Asymptotic Viewpoint, vol. 48, Cambridge University Press, 2019.

[54] K.K. Thekumparampil, A. Khetan, Z. Lin, S. Oh, Robustness of conditional gans to noisy labels, in: Advances in Neural Information Processing Systems, 2018, pp. 10271–10282.

[55] X. Nguyen, M.J. Wainwright, M.I. Jordan, et al., On surrogate loss functions and f-divergences, Ann. Statist. 37 (2) (2009) 876–904.