ON LABEL SHIFT IN DOMAIN ADAPTATION VIA WASSERSTEIN DISTANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the label shift problem between the source and target domains in general domain adaptation (DA) settings. We consider transformations transporting the target to source domains, which enable us to align the source and target examples. Through those transformations, we define the label shift between two domains via optimal transport and develop theory to investigate the properties of DA under various DA settings (e.g., closed-set, partial-set, open-set, and universal settings). Inspired from the developed theory, we propose ***L**abel and **D**ata Shift **R**eduction via **O**ptimal **T**ransport* (LDROT) which can mitigate the data and label shifts simultaneously. Finally, we conduct comprehensive experiments to verify our theoretical findings and compare LDROT with state-of-the-art baselines.

1 INTRODUCTION

The remarkable success of deep learning can be largely attributed to computational power advancement and large-scale annotated datasets. However, in many real-world applications such as medicine and autonomous driving, labeling a sufficient amount of high-quality data to train accurate deep models is often prohibitively labor-expensive, error-prone, and time-consuming. Domain adaptation (DA) or transfer learning has emerged as a vital solution for this issue by transferring knowledge from a label-rich domain (a.k.a. source domain) to a label-scarce domain (a.k.a. target domain). Along with practical DA methods (Ganin & Lempitsky, 2015; Tzeng et al., 2015; Long et al., 2015; Shu et al., 2018; French et al., 2018) which have achieved impressive performance on real-world datasets, the theoretical results (Mansour et al., 2009; Ben-David et al., 2010; Redko et al., 2017; Zhang et al., 2019a; Cortes et al., 2019) are abundant to provide rigorous and insightful understanding of various aspects of transfer learning.

For domain adaptation, the source domain consists of the data distribution $\mathbb{P}^S$ with the density p^S, and the unknown ground-truth labeling function f^S assigning label y to source data x, whilst these are $\mathbb{P}^T$, p^T, and f^T for the target domain, respectively. Moreover, while the data shift can be characterized as a divergence between $\mathbb{P}^S$ and $\mathbb{P}^T$ (Mansour et al., 2009; Ben-David et al., 2010; Redko et al., 2017; Zhang et al., 2019a; Cortes et al., 2019), the label shift in these works is commonly characterized as $\mathbb{E}_{\mathbb{P}^T}\left[\left|f^S(x) - f^T(x)\right|\right]$ or $\mathbb{E}_{\mathbb{P}^S}\left[\left|f^S(x) - f^T(x)\right|\right]$ in which the binary classification with deterministic labeling functions $f^S(\cdot), f^T(\cdot) \in \{0,1\}$ was examined. Additionally, although this label shift term has occurred in the theoretical analysis of Mansour et al. (2009); Ben-David et al. (2010); Redko et al. (2017); Zhang et al. (2019a); Cortes et al. (2019), it is restricted in considering the shift between $f^S(x)$ and $f^T(x)$ at the same data x, which ignores the data shift between $\mathbb{P}^S$ and $\mathbb{P}^T$. This limitation is illustrated in Figure 1. In particular, for a white/square point x drawn from the target domain as in $\mathbb{E}_{\mathbb{P}^T}\left[\left|f^S(x) - f^T(x)\right|\right]$, the source labeling function f^S cannot give reasonable prediction probabilities for x, hence leading to inaccurate $\left|f^S(x) - f^T(x)\right|$.

Label shift has also been examined in an anti-causal setting (Lipton et al., 2018; Garg et al., 2020a), wherein an intervention on $p(y)$ induces the shift, but the process generating x given y is fixed, i.e., $p^S(x \mid y) = p^T(x \mid y)$. Although this setting is useful in some specific scenarios (e.g., a diagnostic problem in which diseases cause symptoms), it is not sufficiently powerful to cope with a general DA setting. Particularly, in an anti-causal setting, the source and target data distributions (i.e., $p^S(x)$ and $p^T(x)$) are just simply two different mixtures of the class conditional distributions $p^S(x \mid y) = p^T(x \mid y)$, hence sharing the same support set. This is certainly far from a general DA setting in

which both *data shift*: $p^S(x) \neq p^T(x)$ with arbitrarily separated support sets and *non-covariate shift*: $p^S(y \mid x) \neq p^T(y \mid x)$ appear.

Contribution. In this paper, we study the label shift for a general domain adaptation setting in which we have both *data shift*: $p^S(x) \neq p^T(x)$ with arbitrarily separated support sets and *non-covariate shift*: $p^S(y \mid x) \neq p^T(y \mid x)$. More specifically, our developed label shift is applicable to a general DA setting with a data shift between two domains and two totally different labeling functions (i.e., we cannot use f^S to predict accurately target examples and vice versa). To define the label shift between two given domains, we utilize transformation L to transport the target to the source data distributions (i.e., $L\#\mathbb{P}^T = \mathbb{P}^S$). This transformation allows us to align the data of two domains. Subsequently, the label shift between two domains is defined as the infimum of the label shift induced by such a transformation with respect to all feasible transformations. This viewpoint of label shift has a connection to optimal transport (Santambrogio, 2015; Villani, 2008; Peyré & Cuturi, 2019), which enables us to develop theory to quantify the label shift for various DA settings, e.g., anti-causal, closed-set, partial-set, open-set, and universal settings. Overall, our contributions can be summarized as follows:

1. We characterize the label shift for a general DA setting via optimal transport. From that, we develop a theory to estimate the label shift for various DA settings and study the trade-off of learning domain-invariant representations and WS label shift.

2. Inspired from the theoretical development, we propose ***L**abel and **D**ata Shift **R**eductions via **O**ptimal **T**ransport* (LDROT) which aims to mitigate both data and label shifts. We conduct comprehensive experiments to verify our theoretical findings and compare the proposed LDROT with the baselines to demonstrate the favorable performance of our method.

Related works. Several attempts have been proposed to characterize the gap between general losses of source and target domains in DA, notably (Mansour et al., 2009; Ben-David et al., 2010; Redko et al., 2017; Zhang et al., 2019a; Cortes et al., 2019). Ben-David & Urner (2014; 2012); Zhang et al. (2019a) study the impossibility theorems for DA, attempting to characterize the conditions under which it is nearly impossible to perform transferability between domains. PAC-Bayesian view on DA using weighted majority vote learning has been rigorously studied in Germain et al. (2013; 2016). Meanwhile, Zhao et al. (2019); Johansson et al. (2019) interestingly indicate the insufficiency of learning domain-invariant representation for successful adaptation. Specifically, Zhao et al. (2019) points out the degradation in target predictive performance if forcing domain invariant representations to be learned while two marginal label distributions of the source and target domains are overly divergent. Johansson et al. (2019) analyzes the information loss of non-invertible transformations and proposes a generalization upper bound that directly takes it into account. Le et al. (2021) employed a transformation to align two domains and developed theories based on this assumption. Moreover, label shift has been examined for the anti-causal setting Lipton et al. (2018); Garg et al. (2020a), which seems not sufficiently realistic for a general DA setting. Optimal transport theory has been theoretically leveraged with domain adaptation (Courty et al., 2017). We compare our proposed LDROT to DeepJDOT (Damodaran et al., 2018) (a deep DA approach based on the theory of (Courty et al., 2017)), and other OT-based DDA approaches, including SWD (Lee et al., 2019), DASPOT (Xie et al., 2019), ETD (Li et al., 2020), RWOT (Xu et al., 2020). Finally, in (Tachet des Combes et al., 2020) , a generator g is said to produce generalized label shift (GLS) representations if it transports source class conditional distributions to corresponding target ones. Further theories were developed to indicate that GLR representations are satisfied if we enforce clustering structure assumption assisting us in training a perfect classifier. Evidently, our work which focuses on how to quantify the label shift between two different domains taking into account the inherent data shift via optimal transport theory is totally different form that work in terms of motivation and developed theory.

2 Label Shift with Wasserstein Distance

2.1 Preliminaries

Notation. For a positive integer n and a real number $p \in [1,\infty)$, $[n]$ indicates the set $\{1,2,\ldots,n\}$ while $\|x\|_p$ denotes the l_p-norm of a vector $x \in \mathbb{R}^n$. Let $\mathscr{Y}^S$ and $\mathscr{Y}^T$ be the label sets of the source and target domains that have $M^S := \left|\mathscr{Y}^S\right|$ and $M^T := \left|\mathscr{Y}^T\right|$ elements, respectively. Meanwhile,

$\mathscr{Y} = \mathscr{Y}^S \cup \mathscr{Y}^T$ stands for the label set of both domains which has the cardinality of $M := |\mathscr{Y}|$. Subsequently, we denote $\mathscr{Y}_\Delta$, $\mathscr{Y}_\Delta^S$, and $\mathscr{Y}_\Delta^T$ as the simplices corresponding to $\mathscr{Y}, \mathscr{Y}^S$, and $\mathscr{Y}^T$ respectively. Finally, let $f^S(\cdot) \in \mathscr{Y}_\Delta$ and $f^T(\cdot) \in \mathscr{Y}_\Delta$ be the labeling functions of the source and target domains, respectively, by filling zeros for the missing labels.

We now examine a general supervised learning setting. Consider a hypothesis $h(\cdot) \in \mathscr{Y}_\triangle$ in a hypothesis class $\mathscr{H}$ and a labeling function $f(\cdot) \in \mathscr{Y}_\triangle$ where $\mathscr{Y}_\triangle := \{\boldsymbol{\pi} \in \mathbb{R}^M : \|\boldsymbol{\pi}\|_1 = 1 \text{ and } \boldsymbol{\pi} \geq \mathbf{0}\}$. Let d_Y be a metric over $\mathscr{Y}_\triangle$, we further define a general loss of the hypothesis h with respect to the labeling function f and the data distribution $\mathbb{P}$ as: $\mathscr{L}(h, f, \mathbb{P}) := \int d_Y(h(x), f(x)) \, \mathrm{d}\mathbb{P}(x)$.

Next, we consider a domain adaptation setting in which we have source space $\mathscr{X}^S$ endowed with a distribution $\mathbb{P}^S$ and the density function $p^S(x)$, and a target space $\mathscr{X}^T$ endowed with a distribution $\mathbb{P}^T$ and the density function $p^T(x)$. We examine various DA settings based on the labels of source and target domains including (1) *closed-set DA*: $\mathscr{Y}^S = \mathscr{Y}^T$, (2) *open-set DA*: $\mathscr{Y}^S \subset \mathscr{Y}^T$, (3) *partial-set DA*: $\mathscr{Y}^T \subset \mathscr{Y}^S$, and (4) *universal DA*: $\mathscr{Y}^S \subsetneq \mathscr{Y}^T$ and $\mathscr{Y}^T \subsetneq \mathscr{Y}^S$.

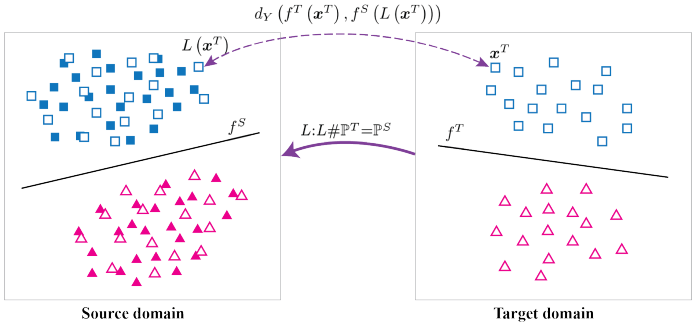


Figure 1: An illustration of our label shift definition. We employ a transformation $L : L\#\mathbb{P}^T = \mathbb{P}^S$ to align two domains. The white/square points $L(x^T)$ on the source domain correspond to the white/square points x^T on the target domain. We measure $d_Y\left(d_Y\left(f^T(x^T), f^S(L(x^T))\right)\right)$ and define the label shift w.r.t. L as $LS(S, T; L) := \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(f^T(x^T), f^S(L(x^T))\right)\right]$. Finally, we take infimum over all valid L to define the label shift between two domains.

2.2 Background on Label Shift

Together with data shift, the study of label shift is important for a general DA problem. However, due to the occurrence of data shift, it is challenging to formulate label shift in a general DA setting. Recent works (Lipton et al., 2018; Garg et al., 2020a) have studied label shift for the anti-causal setting in which an intervention on $p(y)$ induces the shift, but the process generating x given y is fixed, i.e., $p^S(x \mid y) = p^T(x \mid y)$. In spite of being useful in some specific cases, the anti-causal setting is restricted and cannot represent data shift broadly because the source data distribution $p^S(x)$ and the target data distribution $p^T(x)$ are simply just two different mixtures of identical class conditional distributions. Furthermore, the label shift framework from these works is non-trivial to generalize to all settings of DA.

In this paper, we address the issues of the previous works by defining a novel label shift framework via Wasserstein (WS) distance for a general DA setting that takes into account the data shift between two domains. We then develop a theory for our proposed label shift based on useful properties of WS distance, such as its horizontal view, numeric stability, and continuity (Arjovsky et al., 2017). We refer readers to Santambrogio (2015); Villani (2008); Peyré & Cuturi (2019) for a comprehensive knowledge body of optimal transport theory and WS distance, and Appendix A for necessary backgrounds of WS for this work.

2.3 Label Shift via Wasserstein Distance

To facilitate our ensuing discussion, we assume that the source (S) and target (T) distributions $\mathbb{P}^S$ and $\mathbb{P}^T$ are atomless distributions on Polish spaces. Therefore, there exists a transformation $L : \mathscr{X}^T \to \mathscr{X}^S$ such that $L\#\mathbb{P}^T = \mathbb{P}^S$ (Villani, 2008). Given that mapping L, a data example $x^T \sim \mathbb{P}^T$ with the ground-truth prediction probability $f^T(x^T)$ corresponds to another data example $x^S = L(x^T) \sim \mathbb{P}^S$

with the ground-truth prediction probability $f^S\left(x^S\right)$. Hence, it induces a label mismatch loss

$$d_Y\left(f^T\left(x^T\right), f^S\left(x^S\right)\right) = d_Y\left(f^T\left(x^T\right), f^S\left(L\left(x^T\right)\right)\right),$$

where d_Y is a given metric over $\mathscr{Y}_\Delta$. Based on that concept, the label shift between the source and target domains induced by the transformation L can be defined as

$$LS(S,T;L) := \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(f^T\left(x^T\right), f^S\left(L\left(x^T\right)\right)\right)\right]. \tag{1}$$

By finding the optimal mapping L, the label shift between two domains is defined as follows.

Definition 1. Let d_Y be a metric over the simplex $\mathscr{Y}_\Delta$. The label shift between the source and the target domains is defined as the infimum of the label shift induced by all valid transformations L:

$$LS(S,T) := \inf_{L:L\#\mathbb{P}^T=\mathbb{P}^S} LS(S,T;L) = \inf_{L:L\#\mathbb{P}^T=\mathbb{P}^S} \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(f^T\left(x^T\right), f^S\left(L\left(x^T\right)\right)\right)\right]. \tag{2}$$

We give an illustration for Definition 1 in Figure 1. The label shift in Eq. (2) suggests finding the optimal transformation L^* to optimally align the source and target domains with a minimal label mismatch.

Properties of label shift via Wasserstein distance: To show the connection between the aforementioned label shift and optimal transport, we introduce two ways of calculating the label shift via Wasserstein distance.

Proposition 2. *(i) Denote by $\mathbb{P}^S_{f^S}$ the joint distribution of $\left(x, f^S(x)\right)$, where $x \sim \mathbb{P}^S$, and $\mathbb{P}^T_{f^T}$ the joint distribution of $\left(x, f^T(x)\right)$, where $x \sim \mathbb{P}^T$. Then, we have: $LS(S,T) = \mathscr{W}_{d_Y}\left(\mathbb{P}^S_{f^S}, \mathbb{P}^T_{f^T}\right)$.*
(ii) Let $\mathbb{P}_{f^S}$ and $\mathbb{P}_{f^T}$ be the push-forward measures of $\mathbb{P}^S$ and $\mathbb{P}^T$ via f^S and f^T respectively, i.e., $\mathbb{P}_{f^S} = f^S\#\mathbb{P}^S$ and $\mathbb{P}_{f^T} = f^T\#\mathbb{P}^T$. Then, we have: $LS(S,T) = \mathscr{W}_{d_Y}\left(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}\right)$.

The results of Proposition 2 indicate that we can compute the label shift via the Wasserstein distance on the simplex. For example, when $d_Y(y,y') = \|y-y'\|_p^p$, the label shift can be computed via the familiar $\mathscr{W}_p$ distance between $\mathbb{P}_{f^S}$ and $\mathbb{P}_{f^T}$, i.e., $LS(S,T) = \mathscr{W}_p^p(\mathbb{P}_{f^S}, \mathbb{P}_{f^T})$. Note that $\mathscr{W}_d\left(\mathbb{P}^S_{f^S}, \mathbb{P}^T_{f^T}\right)$ with $d = \lambda d_X + d_Y$ was studied in Courty et al. (2017) for proposing a DA method that can mitigate both label and data shifts. However, the concept label shift was not characterized and defined explicitly in that work. Moreover, our motivation and theory development in this work are different and independent from (Courty et al., 2017). To give a better understanding of our label shift definition, we now present some bounds for it in general and specific cases.

Proposition 3. *Denote by $p_Y^S = (p_Y^S(y))_{y=1}^M$ and $p_Y^T = (p_Y^T(y))_{y=1}^M$ the marginal distributions of the source and target domain labels, i.e., $p_Y^S(y) = \int_{\mathscr{X}^S} p^S(x,y)dx$ and $p_Y^T(y) = \int_{\mathscr{X}^T} p^T(x,y)dx$. For $d_Y(y,y') = \|y-y'\|_p^p$ when $p \geq 1$, the following holds:*

(i) $\mathscr{L}\left(h^T, f^T, \mathbb{P}^T\right) \leq LS(S,T) + \mathscr{L}\left(h^S, f^S, \mathbb{P}^S\right) + \mathscr{W}_{d_Y}\left(\mathbb{P}^S_{h^S}, \mathbb{P}^T_{h^T}\right) + const$ where the constant can be viewed as a reconstruction term: $\sup_{L,K:L\#\mathbb{P}^T=\mathbb{P}^S, K\#\mathbb{P}^S=\mathbb{P}^T} \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(f^T\left(K\left(L(x)\right)\right), f^T(x)\right)\right]$;

(ii) $LS(S,T) \geq \|p_Y^S - p_Y^T\|_p^p$;

(iii) In the setting that $\mathbb{P}^S$ and $\mathbb{P}^T$ are mixtures of well-separated Gaussian distributions, i.e., $p^a(x) = \sum_{y=1}^M p^a(y)\mathscr{N}(x|\mu_y^a, \Sigma_y^a)$ with $\|\mu_y^a - \mu_{y'}^a\|_2 \geq D \times \max\{\|\Sigma_y^a\|_{op}^{1/2}, \|\Sigma_{y'}^a\|_{op}^{1/2}\} \forall a \in \{S,T\}, y \neq y'$, in which $\|\cdot\|_{op}$ denotes the operator norm and D is sufficiently large, we have

$$\left|LS(S,T) - \mathscr{W}_p^p(\mathbb{P}_Y^S, \mathbb{P}_Y^T)\right| \leq \varepsilon(D), \tag{3}$$

where $\varepsilon(D)$ is a small constant depending on $D, p_Y^S, p_Y^T, (\Sigma_y^S, \Sigma_y^T)_{y=1}^M$, and it goes to 0 as $D \to \infty$.

A few comments on Proposition 3 are in order. The *inequality in (i)* bounds the target loss by the source loss and the label shift. Though this inequality has the same form as those in (Mansour et al., 2009; Ben-David et al., 2010; Redko et al., 2017; Zhang et al., 2019a; Cortes et al., 2019), the label shift in our inequality is more reasonably expressed. The *inequality in (ii)* reads that the *marginal label shift* $\|p_Y^S - p_Y^T\|_p^p$ is a lower bound of our label shift. Therefore, the label shift induced by the best transformation L^* can not be less than this quantity. A direct consequence is that $LS(S,T) = 0$

implies $p^S(y) = p^T(y)$, for all $y \in [M]$ (no label shift). Finally, the *inequality in (iii)* shows that when the classes are well-separated, the label shift will almost achieve the lower bound in the first inequality, which implies its tightness. The key step in the proof is proving that $\mathbb{P}_{f^S}$ and $\mathbb{P}_{f^T}$ will be concentrated around the vertices and it is also provable for sub-Gaussian distributions with some extra work. The bound gives a simple way to estimate the label shift in this scenario: instead of measuring the Wasserstein distance $\mathscr{W}(\mathbb{P}_{f^S}, \mathbb{P}_{f^T})$ on the simplex $\mathscr{Y}_\Delta$, we only need to measure the Wasserstein distance between the vertices equipped with the masses p_Y^S and p_Y^T. The first experiment in Section 4.1 also supports this finding.

Our label shift formulation can also serve as a tool to elaborate other aspects of DA, as we will see below.

Minimizing data shift while ignoring label shift can hurt the prediction on test set: Consider two classifiers on the source and target domains $h^S = h \circ g^S$ and $h^T = h \circ g^T$ where $g^S : \mathscr{X}^S \to \mathscr{Z}$, $g^T : \mathscr{X}^T \to \mathscr{Z}$ are the source and target feature extractors, and $h : \mathscr{Z} \to \mathscr{Y}_\Delta$ with $h \in \mathscr{H}$. We define a new metric d_Z with respect to the family $\mathscr{H}$ as follows:

$$d_Z(z_1, z_2) = \sup_{h \in \mathscr{H}} d_Y(h(z_1), h(z_2)),$$

where z_1 and z_2 lie on the latent space $\mathscr{Z}$. The necessary (also sufficient) condition under which d_Z is a proper metric on the latent space (see the proof in Appendix B) is realistic and not hard to be satisfied (e.g., the family $\mathscr{H}$ contains any bijection). We now can define a Wasserstein distance W_{d_Z} that will be used in the development of Theorem 4.

Theorem 4. *With regard to the latent space $\mathscr{Z}$, we can upper-bound the label shift as*

$$LS(S,T) \leq \mathscr{L}\left(h^S, f^S, \mathbb{P}^S\right) + \mathscr{L}\left(h^T, f^T, \mathbb{P}^T\right) + \mathscr{W}_{d_Z}\left(g^S \# \mathbb{P}^S, g^T \# \mathbb{P}^T\right).$$

Theorem 4 indicates a trade-off of learning domain-invariant representation by forcing $g^S \# \mathbb{P}^S = g^T \# \mathbb{P}^T$ (e.g., $\min \mathscr{W}_{d_Z}\left(g^S \# \mathbb{P}^S, g^T \# \mathbb{P}^T\right)$). It is evident that if the label shift between domains is significant, because $\mathscr{L}\left(h^S, f^S, \mathbb{P}^S\right)$ can be trained to be sufficiently small, learning domain-invariant representation by minimizing $\mathscr{W}_{d_Z}\left(g^S \# \mathbb{P}^S, g^T \# \mathbb{P}^T\right)$ leads to a hurt in the performance of the target classifier h^T on the target domain. Similar theoretical result was discovered in Zhao et al. (2019) for the binary classification (see Theorem 4.9 in that paper). However, our theory is developed based on our label shift formulation in a more general multi-class classification setting and uses WS distance rather than Jensen-Shannon (JS) distance (Endres & Schindelin, 2006) as in Zhao et al. (2019) for which the advantages of WS distance over JS distance have been thoughtfully discussed in Arjovsky et al. (2017).

Label shift under different settings of DA: An advantage of our method is that it can measure the label shift under various DA settings (i.e., open-set, partial-set, and universal DA). Doing this task is not straight-forward using other label shift methods. For example, if there is a label that appears in a domain but not in the other, it is not meaningful to measure the ratio between the marginal distribution of this label as in Garg et al. (2020b).

In what follows, we elaborate our label shift in those settings. Particularly, we provide some lower bounds for it, implying that the label shift is higher if there is label mismatch between two domains. Recall that when the source and target domains do not have the same number of labels, we can extend f^S and f^T to be functions taking values on $\mathscr{Y}_\Delta$ by filling 0 for the missing labels. For the sake of presenting the results, let $\mathscr{Y}_S \cap \mathscr{Y}_T = \{1, \ldots, C\}$ be the common labels of two domains, $\mathbb{Q}_S$ the marginal of $\mathbb{P}_{f_S}$ and $\mathbb{Q}_T$ the marginal of $\mathbb{P}_{f_T}$ on the first $(C-1)$ dimensions. $\mathbb{Q}_{S\setminus T}$ denotes the marginal of $\mathbb{P}_{f_S}$ in the space of variables having labels $\mathscr{Y}_S \setminus \mathscr{Y}_T$ and $\mathbb{Q}_{T\setminus S}$ denotes the marginal of $\mathbb{P}_{f_T}$ in the space of variables having labels $\mathscr{Y}_S \setminus \mathscr{Y}_T$.

Theorem 5. *Assume that $d_Y(y, y') = \|y - y'\|_p^p$. Then, the following holds:*

i) For the partial-set setting (i.e., $\mathscr{Y}^T \subset \mathscr{Y}^S$), we obtain

$$LS(S,T) \geq W_p^p(\mathbb{Q}_S, \mathbb{Q}_T) + \mathbb{E}_{X \sim \mathbb{Q}_{S\setminus T}}\left[\|X\|_p^p\right]. \tag{4}$$

ii) For the open-set setting (i.e., $\mathscr{Y}^S \subset \mathscr{Y}^T$), we obtain

$$LS(S,T) \geq W_p^p(\mathbb{Q}_S, \mathbb{Q}_T) + \mathbb{E}_{X \sim \mathbb{Q}_{T\setminus S}}\left[\|X\|_p^p\right]. \tag{5}$$

iii) For the universal setting (i.e., $\mathscr{Y}^T \subsetneq \mathscr{Y}^S$ and $\mathscr{Y}^S \subsetneq \mathscr{Y}^T$), we have

$$LS(S,T) \geq W_p^p(\mathbb{Q}_S, \mathbb{Q}_T) + \mathbb{E}_{X\sim\mathbb{Q}_{T\setminus S}}\left[\|X\|_p^p\right] + \mathbb{E}_{Y\sim\mathbb{Q}_{S\setminus T}}\left[\|Y\|_p^p\right]. \tag{6}$$

Theorem 5 reveals that the label shifts for the partial-set, open-set, or universal DA settings are higher than the vanilla closed-set setting due to the missing and unmatching labels of two domains (see Section 4.1). Our theory also implicitly indicates that by setting appropriate weights for source and target examples (i.e., low weights for the examples with missing and unmatching labels), we can deduct label shift by reducing the second term of lower-bounds in Eqs. (4), (5), and (6) to mitigate the negative transfer (Cao et al., 2019). We leave this interesting investigation to our future work. Moreover, further analysis can be found in Appendix C.

3 Label and Data Shift Reductions via Optimal Transport

Inspired by the developed theory in Section 2.3, in this section we propose a novel DA approach, named ***L**abel and **D**ata Shift **R**eductions via **O**ptimal **T**ransport* (LDROT), that aims to reduce both data and label shifts simultaneously.

3.1 Objective function of LDROT

We consider the source classifier $h^S = \bar{h}^S \circ g$ and the target classifier $h^T = \bar{h}^T \circ g$. The pathway of our method consists of three losses, which are as follows:

(i) *Standard loss $\mathscr{L}^S$*: We train the source classifier h^S on the labeled source data by minimizing the loss $\mathscr{L}^S := \mathscr{L}\left(h^S, f^S, \mathbb{P}^S\right)$;

(ii) *Shifting loss $\mathscr{L}^{shift}$*: Furthermore, to mitigate both label and data shifts, we propose to further regularize the loss $\mathscr{L}^S$ by $\mathscr{L}^{shift} := \mathscr{W}_d\left(\mathbb{P}^S_{h^S}, \mathbb{P}^T_{h^T}\right)$, where the ground metric d is defined as

$$d\left(z^S, z^T\right) = \lambda \cdot d_X\left(g\left(x^S\right), g\left(x^T\right)\right) + d_Y\left(h^S\left(x^S\right), h^T\left(x^T\right)\right), \tag{7}$$

where $z^S = \left(x^S, h^S\left(x^S\right)\right)$ with $x^S \sim \mathbb{P}^S$ and $z^T = \left(x^T, h^T\left(x^T\right)\right)$ with $x^T \sim \mathbb{P}^T$;

(iii) *Clustering loss $\mathscr{L}^{clus}$*: Finally, to boost the generalization ability of h^T, we enforce the clustering assumption (Chapelle & Zien, 2005) to enable h^T giving the same prediction for source and target examples on the same cluster. To employ the clustering assumption (Shu et al., 2018), we use Virtual Adversarial Training (VAT) (Miyato et al., 2019) in conjunction with minimizing the entropy of prediction (Grandvalet & Bengio, 2004): $\mathscr{L}^{clus} := \mathscr{L}^{ent} + \mathscr{L}^{vat}$ with

$$\mathscr{L}^{ent} = \mathbb{E}_{\mathbb{P}^T}\left[\mathbb{H}\left(h^T\left(g\left(x\right)\right)\right)\right], \mathscr{L}^{vat} = \quad \mathbb{E}_{0.5\mathbb{P}^S + 0.5\mathbb{P}^T}\left[\max_{x'\in B_\theta(x)} D_{KL}\left(h^T\left(g\left(x\right)\right), h^T\left(g\left(x'\right)\right)\right)\right],$$

where D_{KL} represents a Kullback-Leibler divergence, θ is a very small positive number, $B_\theta\left(x\right) := \{x' : \|x' - x\|_2 < \theta\}$, and $\mathbb{H}$ specifies the entropy.

Combining the above losses, we arrive at the following objective function of LDROT.

$$\inf_{g, \bar{h}^S, \bar{h}^T} \{\mathscr{L}^S + \alpha \mathscr{L}^{shift} + \beta \mathscr{L}^{clus}\}, \tag{8}$$

where $\alpha, \beta > 0$. In addition, we use the target classifier h^T to predict target examples.

Remark on the shifting term: We now explain why including the shifting term $\mathscr{W}_d\left(\mathbb{P}^S_{h^S}, \mathbb{P}^T_{h^T}\right)$ supports to reduce label and data shifts. First, we have the following inequality whose proof can be found in Appendix C:

$$\mathscr{W}_{d_X}\left(g\#\mathbb{P}^S, g\#\mathbb{P}^T\right) = \mathscr{W}_{d_X}\left(\mathbb{P}^S_{h^S}, \mathbb{P}^T_{h^T}\right) \leq \mathscr{W}_d\left(\mathbb{P}^S_{h^S}, \mathbb{P}^T_{h^T}\right) \tag{9}$$

since $d_X \leq d$. Therefore, by including the shifting term $\mathscr{W}_d\left(\mathbb{P}^S_{h^S}, \mathbb{P}^T_{h^T}\right)$, we aim to reduce $\mathscr{W}_{d_X}\left(g\#\mathbb{P}^S, g\#\mathbb{P}^T\right)$, which is useful for reducing the data shift on the latent space.

Second, we find that $\mathscr{W}_{d_Y}\left(\mathbb{P}^S_{h^S}, \mathbb{P}^T_{h^T}\right) \leq \mathscr{W}_d\left(\mathbb{P}^S_{h^S}, \mathbb{P}^T_{h^T}\right)$ since $d_Y \leq d$. The inequality (i) in Proposition 3 suggests that reducing the label shift $\mathscr{W}_{d_Y}\left(\mathbb{P}^S_{h^S}, \mathbb{P}^T_{h^T}\right)$ helps to reduces the loss on the target

domain, therefore increasing the quality of our DA method. Finally, by including $\mathscr{W}_d\left(\mathbb{P}^S_{h^S},\mathbb{P}^T_{h^T}\right)$, we aim to simultaneously reduce both terms $\mathscr{W}_{d_Y}\left(\mathbb{P}^S_{h^S},\mathbb{P}^T_{h^T}\right)$ and $\mathscr{W}_{d_X}\left(\mathbb{P}^S_{h^S},\mathbb{P}^T_{h^T}\right)$, which is equal to $\mathscr{W}_{d_X}\left(g\#\mathbb{P}^S, g\#\mathbb{P}^T\right)$. That step helps reduce the data shift between $g\#\mathbb{P}^S$and $g\#\mathbb{P}^T$, while forcing h^T to mimic h^S for predicting well on the target domain via reducing the label shift $\mathscr{W}_{d_Y}\left(\mathbb{P}^S_{h^S},\mathbb{P}^T_{h^T}\right)$.

3.2 Training procedure of LDROT

We now discuss a few important aspects of the training procedure of LDROT.

Similarity-aware version of the ground metric d: The weight λ in the ground metric d in Eq. (7) represents the matching extent of $g\left(x^T\right)$ and $g\left(x^S\right)$. Ideally, we would like to replace this fixed constant λ by varied weights $\mathbf{w}\left(x^S,x^T\right)$ in such a way that $\mathbf{w}\left(x^S,x^T\right)$ is high if x^T and x^S share the same label and low otherwise. However, it is not possible because the label of x^T is unknown. As an alternative, it appears that if we can have a good way to estimate the pairwise similarity $s\left(x^S,x^T\right)$ of x^T and x^S, $s\left(x^S,x^T\right)$ seems to be high if x^T and x^S share the same label and low if otherwise. Based on this observation, we instead propose using a similarity-aware version of metric d as follows:

$$\bar{d}\left(z^S,z^T\right)=\mathbf{w}\left(x^S,x^T\right)d_X\left(g\left(x^S\right),g\left(x^T\right)\right)+d_Y\left(h^S\left(x^S\right),h^T\left(x^T\right)\right),$$

where the weight $\mathbf{w}\left(x^S,x^T\right)$ is estimated based on $s\left(x^S,x^T\right)$.

Entropic regularized version of shifting term: Since computing directly the shifting term $\mathscr{W}_d\left(\mathbb{P}^S_{h^S},\mathbb{P}^T_{h^T}\right)$ is expensive, we use the entropic regularized version of $\mathscr{W}_d\left(\mathbb{P}^S_{h^S},\mathbb{P}^T_{h^T}\right)$ instead, which we denote $\mathscr{W}^{\varepsilon}_d\left(\mathbb{P}^S_{h^S},\mathbb{P}^T_{h^T}\right)$ where ε is a positive regularized term (Detailed definition and discussion of the entropic regularized Wasserstein metric is in Appendix A). The dual-form (Genevay et al., 2016) of that entropic regularized term with respect to the ground metric $\bar{d}$ admits the following form:

$$\max_{\phi}\left\{\frac{1}{N_S}\sum_{i=1}^{N_S}\phi\left(g\left(x_i^S\right)\right)-\frac{\varepsilon}{N_T}\sum_{j=1}^{N_T}\left[\log\left(\frac{1}{N_S}\sum_{i=1}^{N_S}\exp\left\{\frac{\phi\left(g\left(x_i^S\right)\right)-\bar{d}\left(z_i^S,z_j^T\right)}{\varepsilon}\right\}\right)\right]\right\}, \tag{10}$$

where ϕ is a neural net named the Kantorovich potential network, $\left\{\left(x_i^S,y_i^S\right)\right\}_{i=1}^{N_S}$and $\left\{x_i^T\right\}_{i=1}^{N_T}$ are source and target data, $z_i^S=(x_i^S,h^S(x_i^S))$, $z_j^T=(x_j^T,h^T(x_j^T))$, and $\mathbf{w}_{ij}=\mathbf{w}\left(x_i^S,x_j^T\right)$.

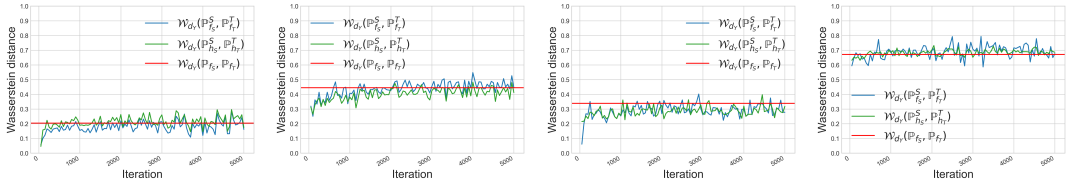
Evaluating the weights $\mathbf{w}_{ij}$: The weights $\mathbf{w}_{ij}$ are evaluated based on the similarity scores $s_{ij}:=s\left(x_i^S,x_j^T\right)$. Basically, we train from scratch or fine-tune a pre-trained deep net (e.g., ResNet (He et al., 2016)) using source dataset with labels and compute cosine similarity of latent representations r_j^T and r_i^S of x_j^T and x_i^S as $s_{ij}=s\left(x_i^S,x_j^T\right)=\text{cosine-sim}\left(r_i^S,r_j^T\right)$.

To ease the computation, we estimate the weights $\mathbf{w}_{ij}$ according to source and target batches. Specifically, we consider a balanced source batch of Mb source examples (i.e., M is the number of classes and b is source batch size for each class). For a target example x_j^T in the target batch, we sort the similarity array $[s_{ij}]_{i=1}^{Mb}$ in an ascending order. Ideally, we expect that the similarity scores of the target example x_j^T and the source examples x_i^S with the same class as x_j^T (i.e., totally we have b of theirs) are higher than other similarity scores in the current source batch. Therefore, we find μ_i as the $\frac{M-1}{M}$-percentile of the ascending similarity array $[s_{ij}]_{i=1}^{Mb}$ and compute the weights as $\mathbf{w}_{ij}=\exp\left\{\frac{s_{ij}-\mu_i}{\tau}\right\}$ with a temperature variable τ. It is worth noting that this weight evaluation strategy assists us in *sharpening* and *contrasting* the weights for the pairs in the similar and different classes. More specifically, for the pairs in the same classes, s_{ij} tend to be bigger than μ_i, whilst for the pairs in different classes, s_{ij} tend to be smaller than μ_i. Hence, with the support of exponential form and temperature variable τ, $\mathbf{w}_{ij}$ for the pairs in the same classes tend to be higher than those for the pairs in different classes.

Comparing to DeepJDOT (Damodaran et al., 2018): The $\mathscr{W}_d\left(\mathbb{P}^S_{h^S},\mathbb{P}^T_{h^T}\right)$ with $d=\lambda d_X+d_Y$ was investigated in DeepJDOT (Damodaran et al., 2018). However, ours is different from that work

Table 1: Classification accuracy (%) on Office-31 dataset for unsupervised DA (ResNet-50).

Method	A→W	A→D	D→W	W→D	D→A	W→A	Avg
ResNet-50 (He et al., 2016)	70.0	65.5	96.1	99.3	62.8	60.5	75.7
DeepCORAL (Sun & Saenko, 2016)	83.0	71.5	97.9	98.0	63.7	64.5	79.8
DANN (Ganin et al., 2016)	81.5	74.3	97.1	99.6	65.5	63.2	80.2
ADDA (Tzeng et al., 2017)	86.2	78.8	96.8	99.1	69.5	68.5	83.2
CDAN (Long et al., 2018)	94.1	92.9	98.6	**100.0**	71.0	69.3	87.7
TPN (Pan et al., 2019)	91.2	89.9	97.7	99.5	70.5	73.5	87.1
SAFN (Xu et al., 2019)	90.1	90.7	98.6	99.8	73.0	70.2	87.1
rRevGrad+CAT (Deng et al., 2019)	94.4	90.8	98.0	**100.0**	72.2	70.2	87.6
DeepJDOT (Damodaran et al., 2018)	88.9	88.2	98.5	99.6	72.1	70.1	86.2
ETD (Li et al., 2020)	92.1	88.0	**100.0**	**100.0**	71.0	67.8	86.2
RWOT (Xu et al., 2020)	95.1	94.5	99.5	**100.0**	77.5	77.9	90.8
LDROT	**95.6**	**98.0**	98.1	**100.0**	**85.6**	**84.9**	**93.7**



(a) Closed-set setting with $\mathscr{Y}^S = \mathscr{Y}^T = [9]$.

(b) Open-set setting with $\mathscr{Y}^S = [6]$, $\mathscr{Y}^T = [9]$.

(c) Partial-set setting with $\mathscr{Y}^S = [9]$, $\mathscr{Y}^T = [6]$.

(d) Universal setting with $\mathscr{Y}^S = [3] \cup \{4,5,6\}$, $\mathscr{Y}^T = [3] \cup \{7,8,9\}$.

Figure 2: Label shift estimation for various settings of DA when the source data set SVHN and the target data set is MNIST. Here, we denote $[C] := \{0,1,...,C\}$ for a positive integer number C.

in some aspects: (i) *similarity based dynamic weighting*, (ii) *clustering loss* for enforcing clustering assumption for target classifier, and (iii) *entropic dual form* for training rather than Sinkhorn as in Damodaran et al. (2018). More analysis of LDROT can be found in Section D.1.

4 Experiments

4.1 Experiments of theoretical part

Label shift estimation: In this experiment, we show how to evaluate the label shift if we know the labeling mechanisms of the source and target domains. We consider SVHN as the source domain and MNIST as the target domain. These two datasets have ten categorical labels which stand for the digits in $\mathscr{Y} = \{0,1,...,9\}$. Additionally, for each source or target example x, the ground-truth label of x is a categorical label in $\mathscr{Y} = \{0,1,...,9\}$. Since these labels have good separation, from part (iii) of Proposition 3, we can choose $f^S(x)$ and $f^T(x)$ as one-hot vectors on the simplex $\mathscr{Y}_\Delta$. Therefore, $\mathbb{P}_{f^S}$ and $\mathbb{P}_{f^T}$ are two discrete distributions over one-hot vectors representing the categorical labels, wherein each categorical label $y \in \{0,...,9\}$ corresponds to the one-hot vector $\mathbf{1}_{y+1} = [0,..,0,1_{y+1},0,...,0]$. The label shift between SVHN and MNIST is estimated by either $\mathscr{W}_{d_Y}\left(\mathbb{P}^S_{f^S}, \mathbb{P}^T_{f^T}\right)$ or $\mathscr{W}_{d_Y}\left(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}\right)$, where d_Y is chosen as L^1 distance. More specifically, $\mathscr{W}_{d_Y}\left(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}\right)$ is evaluated accurately via linear programming[1], while estimating $\mathscr{W}_{d_Y}\left(\mathbb{P}^S_{f^S}, \mathbb{P}^T_{f^T}\right)$ using the entropic regularized dual form $\mathscr{W}^\varepsilon_{d_Y}\left(\mathbb{P}^S_{f^S}, \mathbb{P}^T_{f^T}\right)$ with $\varepsilon = 0.1$ (Genevay et al., 2016). Moreover, to visualize the precision when using a probabilistic labeling mechanism to estimate the label shift, we train two probabilistic labeling functions h^S and h^T by minimizing the cross-entropy loss with respect to f^S and f^T respectively and subsequently estimate $\mathscr{W}_{d_Y}\left(\mathbb{P}_{h^S}, \mathbb{P}_{h^T}\right)$ using the entropic regularized dual form $\mathscr{W}^\varepsilon_{d_Y}\left(\mathbb{P}^S_{h^S}, \mathbb{P}^T_{h^T}\right)$ with $\varepsilon = 0.1$. Note that the prediction probabilities of h^S and h^T are now the points on the simplex $\mathscr{Y}_\Delta$.

We compute the label shift for four DA settings including the closed-set, partial-set, open-set, and universal settings. As shown in Figure 2, for all DA settings, the blue lines estimating $\mathscr{W}_{d_Y}\left(\mathbb{P}^S_{f^S}, \mathbb{P}^T_{f^T}\right)$ and the green lines estimating $\mathscr{W}_{d_Y}\left(\mathbb{P}_{h^S}, \mathbb{P}_{h^T}\right)$ along with batches tend to approach the red lines evaluating $\mathscr{W}_{d_Y}\left(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}\right)$ accurately, which illustrates the result of part (iii) in Propo-

[1] https://pythonot.github.io/all.html

Table 2: Classification accuracy (%) on Office-Home dataset for unsupervised DA (ResNet-50).

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
ResNet-50 (He et al., 2016)	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DANN (Ganin et al., 2016)	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
CDAN (Long et al., 2018)	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	**70.9**	56.7	81.6	65.8
TPN (Pan et al., 2019)	51.2	71.2	76.0	65.1	72.9	72.8	55.4	48.9	76.5	**70.9**	53.4	80.4	66.2
SAFN Xu et al. (2019)	52.0	71.7	76.3	64.2	69.9	71.9	63.7	51.4	77.1	**70.9**	57.1	81.5	67.3
DeepJDOT (Damodaran et al., 2018)	48.2	69.2	74.5	58.5	69.1	71.1	56.3	46.0	76.5	68.0	52.7	80.9	64.3
ETD (Li et al., 2020)	51.3	71.9	**85.7**	57.6	69.2	73.7	57.8	51.2	79.3	70.2	57.5	82.1	67.3
RWOT (Xu et al., 2020)	55.2	72.5	78.0	63.5	72.5	75.1	60.2	48.5	78.9	69.8	54.8	82.5	67.6
LDROT	**57.4**	**79.6**	82.5	**67.2**	**79.8**	**80.7**	**66.5**	**53.3**	**82.5**	**70.9**	**57.4**	**84.8**	**71.9**

sition 3. We also observe that the label shifts of partial-set, open-set, and universal settings are higher than the closed-set setting as discussed in Theorem 5.

Implication on target performance: In this experiment, we demonstrate that the theoretical finding of Theorem 4 indicating that forcing learning domain-invariant representations hurts the target performance. We train a classifier $h^{ST} = g \circ h$, where g is a feature extractor and h is a classifier on top of latent representations by solving $\min_{g,h}\{\mathscr{L}\left(h^{ST}, f^S, \mathbb{P}^S\right) + 0.1 \times \mathscr{W}_{L^1}^{\varepsilon}\left(g\#\mathbb{P}^S, g\#\mathbb{P}^T\right)\}$ where $\mathscr{W}_{L^1}^{\varepsilon}\left(g\#\mathbb{P}^S, g\#\mathbb{P}^T\right)$ is used to estimate $\mathscr{W}_{L^1}\left(g\#\mathbb{P}^S, g\#\mathbb{P}^T\right)$ for learning domain-invariant representations on the latent space. We conduct the experiments on the pairs A→W (Office-31) and P→I (ImageCLEF-DA) in which we measure the WS data shift on the latent space $\mathscr{W}_{L^1}^{\varepsilon}\left(g\#\mathbb{P}^S, g\#\mathbb{P}^T\right)$, and the source and target accuracies. As shown in Figure 3, along with the training process, while the WS data shift on the latent space consistently decreases (i.e., the latent representations become more domain-invariant), the source accuracies get saturated, but the target accuracies get hurt gradually.

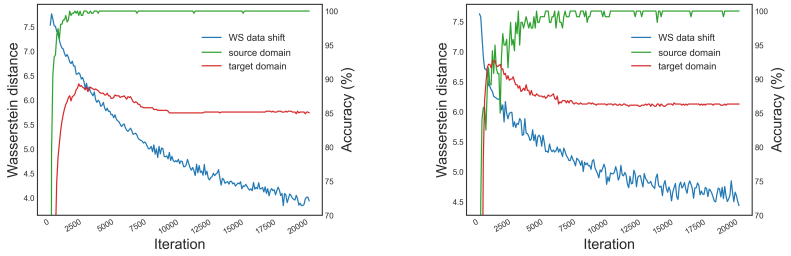


Figure 3: Illustration on target performance to show that forcing learning domain-invariant representations can hurt the target performance. Left: A→W (Office-31). Right: P→I (ImageCLEF-DA).

4.2 Experiments of LDROT on real-world datasets

We conduct the experiments on the real-world datasets: Digits, Office-31, Office-Home, and ImageCLEF-DA to compare our LDROT to the state-of-the-art baselines, especially OT-based ones DeepJDOT (Damodaran et al., 2018), SWD (Lee et al., 2019), DASPOT (Xie et al., 2019), ETD Li et al. (2020), and RWOT (Xu et al., 2020). Due to the space limit, we show the results for Office-31 and Office-Home in Tables 1 and 2, while *other results, parameter settings, and network architectures* can be found in Appendix D. The experimental results indicate that our proposed method outperforms the baselines.

4.3 Ablation Study for LDROT

We conduct comprehensive ablation studies to investigate the behavior of our LDROT. Due to the space limit, we leave to the ablation and analytic studies to Appendix D.7.

5 Conclusion

In this paper, we study label shift between the source and target domains in a general DA setting. Our main workaround is to consider valid transformations transporting the target to source domains allowing us to align the source and target examples and rigorously define the label shift between two domains. We then connect the proposed label shift to optimal transport theory and develop further theory to inspect the properties of DA under various DA settings (e.g., closed-set, partial-set, open-set, or universal setting). Furthermore inspired from theory development, we propose ***L**abel and **D**ata Shift **R**eduction via **O**ptimal **T**ransport* (LDROT) which can mitigate data and label shifts simultaneously. We conduct comprehensive experiments to verify our theoretical findings and compare LDROT against state-of-the-art baselines to demonstrate its merits.

Reproducibility Statement: Source codes for our experiments are provided in the supplementary of the paper. The details of experimental settings, computational infrastructure, and other used public libraries are given in the supplementary material. All datasets that we used in the paper are published and they are easy to find in the Internet.

Ethics Statement: Given the nature of the work, we do not foresee any negative societal and ethical impacts of the work.

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. D.4

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/arjovsky17a.html. 2.2, 2.3

S. Ben-David and R. Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, pp. 139–153, 2012. ISBN 9783642341052. 1

S. Ben-David and R. Urner. Domain adaptation—can quantity compensate for quality? *Annals of Mathematics and Artificial Intelligence*, 70(3):185–202, March 2014. ISSN 1012-2443. 1

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, May 2010. ISSN 0885-6125. 1, 2.3

Z. Cao, K. You, M. Long, J. Wang, and Q. Yang. Learning to transfer examples for partial domain adaptation. *CoRR*, abs/1903.12230, 2019. URL http://arxiv.org/abs/1903.12230. 2.3

O. Chapelle and A. Zien. Semi-supervised classification by low density separation. In *AISTATS*, volume 2005, pp. 57–64. Citeseer, 2005. 3.1

C. Cortes, M. Mohri, and Andrés M. Medina. Adaptation based on generalized discrepancy. *The Journal of Machine Learning Research*, 20(1):1–30, 2019. 1, 2.3

N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 3730–3739, 2017. 1, 2.3

B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11208 of *Lecture Notes in Computer Science*, pp. 467–483, 2018. 1, 3.2, 1, 2, 4.2, 4, 5

Z. Deng, Y. Luo, and J. Zhu. Cluster alignment with a teacher for unsupervised domain adaptation, 2019. 1, 4, 5

D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Trans. Inf. Theor.*, 49(7):1858–1860, 2006. ISSN 0018-9448. 2.3

G. French, M. Mackiewicz, and M. Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, 2018. 1

Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1180–1189, 2015. 1, 4

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1): 2096–2030, jan 2016. ISSN 1532-4435. 1, 2, 5

S. Garg, Y. Wu, S. Balakrishnan, and Z. Lipton. A unified view of label shift estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3290–3300. Curran Associates, Inc., 2020a. 1, 2.2

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3290–3300. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper/2020/file/219e052492f4008818b8adb6366c7ed6-Paper.pdf. 2.3

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/2a27b8144ac02f67687f76782a3b5d8f-Paper.pdf. 3.2, 4.1, A.2

P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, ICML'13, 2013. 1

P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A new PAC-Bayesian perspective on domain adaptation. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 859–868, 2016. 1

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper/2004/file/96f2b50b5d3613adf9c27049b2a888c7-Paper.pdf. 3.1

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. 3.2, 1, 2, D.2, D.5, 5

F. D. Johansson, D. Sontag, and R. Ranganath. Support and invertibility in domain-invariant representations. In *Proceedings of Machine Learning Research*, volume 89, pp. 527–536, 2019. 1

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. D.5

Trung Le, Tuan Nguyen, Nhat Ho, Hung Bui, and Dinh Phung. Lamda: Label matching deep domain adaptation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6043–6054. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/le21a.html. 1

Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced Wasserstein discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 10285–10295. Computer Vision Foundation / IEEE, 2019. 1, 4.2, 4, D.7

M. Li, Y. Zhai, Y. Luo, P. Ge, and C. Ren. Enhanced transport distance for unsupervised domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 1, 2, 4.2, 4, 5

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130. PMLR, 10–15 Jul 2018. 1, 2.2

M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In F. Bach and D. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 97–105, Lille, France, 2015. 1

M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 1640–1650. Curran Associates, Inc., 2018. 1, 2, 4, 5

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (eds.), *Advances in Neural Information Processing Systems 21*, pp. 1041–1048. 2009. 1, 2.3

T. Miyato, S. Maeda, M. Koyama, and S. Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, Aug 2019. ISSN 1939-3539. 3.1

Y. Pan, T. Yao, Y. Li, Y. Wang, C. Ngo, and T. Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *CVPR*, pp. 2234–2242, 2019. 1, 2, 4, 5

G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 1, 2.2

I. Redko, A. Habrard, and M. Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 737–753, 2017. 1, 2.3

F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, pp. 99–102, 2015. 1, 2.2, A.1, A.1, B.1

R. Shu, H. Bui, H. Narui, and S. Ermon. A DIRT-t approach to unsupervised domain adaptation. In *International Conference on Learning Representations*, 2018. 1, 3.1

B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In Gang Hua and Hervé Jégou (eds.), *Computer Vision – ECCV 2016 Workshops*, pp. 443–450, Cham, 2016. Springer International Publishing. 1, 4, 5

Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33, 2020. 1

E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. *CoRR*, 2015. 1

E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2962–2971, 2017. 1, 4, 5

L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. D.7

C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509. 1, 2.2, 2.3, A.1

Y. Xie, M. Chen, H. Jiang, T. Zhao, and H. Zha. On scalable and efficient computation of large scale optimal transport. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6882–6892, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 1, 4.2, 4

R. Xu, G. Li, J. Yang, and L. Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1426–1435, 2019. doi: 10.1109/ICCV.2019.00151. 1, 2, 5

R. Xu, P. Liu, L. Wang, C. Chen, and J. Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *CVPR 2020*, June 2020. 1, 1, 2, 4.2, 4, 5

Y. Zhang, Y. Liu, M. Long, and M. I. Jordan. Bridging theory and algorithm for domain adaptation. *CoRR*, abs/1904.05801, 2019a. 1, 2.3

Y. Zhang, H. Tang, K. Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5026–5035, 2019b. 5

H. Zhao, R. T. Des Combes, K. Zhang, and G. Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pp. 7523–7532, 2019. 1, 2.3

Supplement to "On Label Shift in Domain Adaptation via Wasserstein Distance"

In this appendix, we collect several proofs and remaining materials that are deferred from the main paper.

- In Appendix A, we present notations and definitions that are deferred from the main text including optimal transport and entropic regularized Wasserstein distance.
- In Appendix B, we present proofs of all the key results.
- In Appendix C, we present proofs of the remaining results, including the derivation of the dual form of entropic regularized optimal transport.
- In Appendix D, we provide training specification and additional experimental results.

A Notations and definitions

In this appendix, we provide notations, notions, and definitions that are used in the main text.

A.1 Optimal Transport

Given two probability measures $(\mathscr{X},\mathbb{P})$ and $(\mathscr{Y},Q)$ and a cost function or ground metric $d(x,y)$, under the conditions stated in the below theorem (cf. Theorems 1.32 and 1.33 (Santambrogio, 2015)), the *primal form* of Wasserstein (WS) distance (Santambrogio, 2015) is defined as:

$$\mathscr{W}_d(\mathbb{P},\mathbb{Q}) = \inf_{T:T\#\mathbb{P}=\mathbb{Q}} \mathbb{E}_{x\sim\mathbb{P}}\left[d(x,T(x))\right], \tag{11}$$

$$\mathscr{W}_d(\mathbb{P},\mathbb{Q}) = \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(x,y)\sim\gamma}\left[d(x,y)\right], \tag{12}$$

where $\Gamma(\mathbb{P},\mathbb{Q})$ specifies the set of joint distributions over $\mathscr{X}\times\mathscr{Y}$ which admits $\mathbb{P}$ and $\mathbb{Q}$ as marginals. The first definition is known as Monge problem (MP), while the second one is known as Kantorovich problem (KP). We now restate the sufficient conditions for which (MP) and (KP) are equivalent (cf. Theorems 1.32 and 1.33 (Santambrogio, 2015)).

Theorem. *If $\mathscr{X}$ and $\mathscr{Y}$ are compact, Polish metric spaces, $\mathbb{P}$ and $\mathbb{Q}$ are atomless, and d is a lower semi-continuous function, then (KP) is equivalent to (MP) in the sense that two infima are equal.*

In addition, under some mild conditions as stated in Theorem 5.10 in Villani (2008), we can replace the primal form by its corresponding dual form

$$\mathscr{W}_d(\mathbb{P},\mathbb{Q}) = \max_{\phi\in\mathscr{L}_1(\Omega,\mathbb{P})} \left\{\mathbb{E}_{\mathbb{Q}}\left[\phi^c(x)\right] + \mathbb{E}_{\mathbb{P}}\left[\phi(y)\right]\right\}, \tag{13}$$

where $\mathscr{L}_1(\Omega,\mathbb{P}) := \left\{\psi : \int_\Omega |\psi(y)|\,\mathrm{d}\mathbb{P}(y) < \infty\right\}$ and ϕ^c is the c-transform of function ϕ defined as $\phi^c(x) := \min_y \left\{d(x,y) - \phi(y)\right\}$.

A.2 Entropic Regularized Duality

To enable the application of optimal transport in machine learning and deep learning, Genevay et al. developed an entropic regularized dual form in Genevay et al. (2016). First, they proposed to add an entropic regularization term to the primal form in (12)

$$\mathscr{W}_d^\varepsilon(\mathbb{P},\mathbb{Q}) := \min_{\gamma\in\Gamma(\mathbb{Q},\mathbb{P})} \left\{\mathbb{E}_{(x,y)\sim\gamma}\left[d(x,y)\right] + \varepsilon D_{KL}\left(\gamma\|\mathbb{P}\otimes\mathbb{Q}\right)\right\}, \tag{14}$$

where ε is the regularization rate, $D_{KL}(\cdot\|\cdot)$ is the Kullback-Leibler (KL) divergence, and $\mathbb{P}\otimes\mathbb{Q}$ represents the specific coupling in which $\mathbb{Q}$ and $\mathbb{P}$ are independent. Note that when $\varepsilon\to 0$, $\mathscr{W}_d^\varepsilon(\mathbb{P},\mathbb{Q})$ approaches $\mathscr{W}_d(\mathbb{P},\mathbb{Q})$ and the optimal transport plan γ_ε^* of (14) also weakly converges to the optimal transport plan γ^* of (12). In practice, we set ε to be a small positive number, hence γ_ε^* is very close to γ^*. Second, using the Fenchel-Rockafellar theorem, they obtained the following dual form w.r.t. the potential ϕ

$$\mathscr{W}_d^{\varepsilon}(\mathbb{P},\mathbb{Q}) = \max_{\phi}\left\{\int \phi_{\varepsilon}^{c}(x)\,\mathrm{d}\mathbb{Q}(x) + \int \phi(y)\,\mathrm{d}\mathbb{P}(y)\right\} = \max_{\phi}\left\{\mathbb{E}_{\mathbb{Q}}\left[\phi_{\varepsilon}^{c}(x)\right] + \mathbb{E}_{\mathbb{P}}\left[\phi(y)\right]\right\}, \quad (15)$$

where $\phi_{\varepsilon}^{c}(x) := -\varepsilon \log\left(\mathbb{E}_{\mathbb{P}}\left[\exp\left\{\frac{-d(x,y)+\phi(y)}{\varepsilon}\right\}\right]\right)$.

A.3 Preliminaries

Notions. For a positive integer n and a real number $p \in [1,\infty)$, $[n]$ indicates the set $\{1,2,\ldots,n\}$ while $\|x\|_p$ denotes the l_p-norm of a vector $x \in \mathbb{R}^n$. Let $\mathscr{Y}^S$ and $\mathscr{Y}^T$ be the label sets of the source and target domains that have $M^S := \left|\mathscr{Y}^S\right|$ and $M^T := \left|\mathscr{Y}^T\right|$ elements, respectively. Meanwhile, $\mathscr{Y} = \mathscr{Y}^S \cup \mathscr{Y}^T$ stands for the label set of both domains which has the cardinality of $M := |\mathscr{Y}|$. Subsequently, we denote $\mathscr{Y}_{\Delta}$, $\mathscr{Y}_{\Delta}^S$, and $\mathscr{Y}_{\Delta}^T$ as the simplices corresponding to $\mathscr{Y}, \mathscr{Y}^S$, and $\mathscr{Y}^T$ respectively. Finally, let $f^S(\cdot) \in \mathscr{Y}_{\Delta}$ and $f^T(\cdot) \in \mathscr{Y}_{\Delta}$ be the labeling functions of the source and target domains, respectively, by filling zeros for the missing labels.

We first examine a general supervised learning setting. Consider a hypothesis h in a hypothesis class $\mathscr{H}$ and a labeling function f (i.e., $f(\cdot) \in \mathscr{Y}_{\triangle}$ and $h(\cdot) \in \mathscr{Y}_{\triangle}$ where $\mathscr{Y}_{\triangle} := \left\{\boldsymbol{\pi} \in \mathbb{R}^M : \|\boldsymbol{\pi}\|_1 = 1 \text{ and } \boldsymbol{\pi} \geq \mathbf{0}\right\}$ with the number of classes M). Let $d_{\mathscr{Y}}$ be a metric over $\mathscr{Y}_{\triangle}$. We further define the general loss of the hypothesis h w.r.t. the data distribution $\mathbb{P}$ and the labeling function f as: $\mathscr{L}(h,f,\mathbb{P}) := \int d_{\mathscr{Y}}(h(x),f(x))\,\mathrm{d}\mathbb{P}(x)$.

Next we consider a domain adaptation setting in which we have a source space $\mathscr{X}^S$ endowed with a distribution $\mathbb{P}^S$ and the density function $p^s(x)$ and a target space $\mathscr{X}^T$ endowed with a distribution $\mathbb{P}^T$ and the density function $p^t(x)$. Let $f^S(\cdot) \in \mathscr{Y}_{\Delta}$ and $f^T(\cdot) \in \mathscr{Y}_{\Delta}$ be the labeling functions of the source and target domains respectively. It appears that $p^S(x,y) = p^S(x) f^S(x,y)$ and $p^T(x,y) = p^T(x) f^T(x,y)$ are the source and target joint distributions of pairs (x,y) respectively. Note that for a categorical label $y \in \{1,\ldots,M\}$, $f^S(x,y)$ and $f^T(x,y)$ represent the $y-$th element of the prediction probabilities $f^S(x)$ and $f^T(x)$.

B Proofs of all the key results

In this appendix, we provide useful lemmas and proofs for main results in the paper.

B.1 Useful Lemmas

Lemma 6. *If $\gamma \in \Gamma\left(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}\right)$, there exists $\gamma' \in \Gamma\left(\mathbb{P}^S, \mathbb{P}^T\right)$ such that $\left(f^S, f^T\right)_{\#}\gamma' = \gamma$.*

Proof. Let denote γ^S as the joint distribution of the samples $\left(x^S, f^S\left(x^S\right)\right)$ where $x^S \sim \mathbb{P}^S$ and γ^T as the joint distribution of the samples $\left(x^T, f^T\left(x^T\right)\right)$ where $x^T \sim \mathbb{P}^T$. It is obvious that γ^S is a joint distribution of $\mathbb{P}^S$ and $\mathbb{P}^{f^S}$ and γ^T is a joint distribution of $\mathbb{P}^T$ and $\mathbb{P}^{f^T}$. According to the gluing lemma (see Lemma 5.5 in Santambrogio (2015)), there exists a joint distribution μ such that for any draw $\left(x^S, \tau^S, \tau^T, x^T\right) \sim \mu$ then $\left(x^S, \tau^S\right) \sim \gamma^S$, $\left(\tau^S, \tau^T\right) \sim \gamma$, and $\left(x^T, \tau^T\right) \sim \gamma^T$.

Let γ' be the distribution of samples $\left(x^S, x^T\right)$ (i.e., the projection of μ onto the first and fourth dimensions). This follows that γ' is a joint distribution of $\mathbb{P}^S$ and $\mathbb{P}^T$ (i.e., $\gamma' \in \Gamma\left(\mathbb{P}^S, \mathbb{P}^T\right)$). In addition, since $\left(x^S, \tau^S\right) \sim \gamma^S$, $\tau^S = f^S\left(x^S\right)$, since $\left(x^T, \tau^T\right) \sim \gamma^T$, $\tau^T = f^T\left(x^T\right)$, and $\left(\tau^S, \tau^T\right) \sim \gamma$. Therefore, we reach $\left(f^S, f^T\right)_{\#}\gamma' = \gamma$.

We note that in the above proof, we employ a general form of the gluing lemma for 4 distributions and spaces. The proof is mainly based on the gluing lemma for 3 distributions and spaces and trivial. □

Lemma 7. *Let d_Z be defined with respect to the family $\mathscr{H}$ as follows:*

$$d_Z(z_1,z_2) = \sup_{h \in \mathscr{H}} d_Y(h(z_1), h(z_2)),$$

where z_1 and z_2 lie on the latent space $\mathscr{Z}$. For any z_1 and z_2, if $h(z_1) = h(z_2), \forall h \in \mathscr{H}$ leads to $z_1 = z_2$, then d_Z is a proper metric.

Proof. First, $d_Z(z_1, z_2) \geq 0$ and $d_Z(z_1, z_2) = 0$ means $h(z_1) = h(z_2), \forall h \in \mathscr{H}$, which leads to $z_1 = z_2$. Second, it is obvious that $d_Z(z_1, z_2) = d_Z(z_2, z_1), \forall z_1, z_2$.

Given any z_1, z_2, z_3, we have

$$\begin{aligned} d_Z(z_1, z_3) &= \sup_{h \in \mathscr{H}} d_Y(h(z_1), h(z_3)) \leq \sup_{h \in \mathscr{H}} \left(d_Y(h(z_1), h(z_2)) + d_Y(h(z_2), h(z_3))\right) \\ &\leq \sup_{h \in \mathscr{H}} d_Y(h(z_1), h(z_2)) + \sup_{h \in \mathscr{H}} d_Y(h(z_2), h(z_3)) \\ &= d_Z(z_1, z_2) + d_Z(z_2, z_3). \end{aligned}$$

Therefore, d_Z is a proper metric. □

B.2 Proof and Corollary of Proposition 2

Proof. **(i)** First, we will prove that $\mathscr{W}_{d_Y}\left(\mathbb{P}^S_{f^S}, \mathbb{P}^T_{f^T}\right) \geq \mathscr{L}S(S,T)$. Let $H : \operatorname{supp}\left(\mathbb{P}^T_{f^T}\right) \to \operatorname{supp}\left(\mathbb{P}^S_{f^S}\right)$ be such that $H_\# \mathbb{P}^T_{f^T} = \mathbb{P}^S_{f^S}$ where supp indicates the support of a distribution. We can express H as

$$H\left(x, f^T(x)\right) := \left(H_1\left(x, f^T(x)\right), H_2\left(x, f^T(x)\right)\right),$$

with $H_1\left(x, f^T(x)\right) \in \mathscr{X}^S$ and $H_2\left(x, f^T(x)\right) \in \mathscr{Y}_\Delta$. Define $K(x) := H_1\left(x, f^T(x)\right)$. We claim that $K_\# \mathbb{P}^T = \mathbb{P}^S$. Observe first that for any $U_S \subset \mathscr{X}^S \times \mathscr{Y}_\triangle$, we have $\mathbb{P}^S_{f^S}(U_S) = \mathbb{P}^S(V_S)$ where $V_S := \left\{x \in \mathscr{X}^S \mid \left(x, f^S(x)\right) \in U_S\right\}$. Next, let $V_S \subset \mathscr{X}^S$ be any measurable set and denote $U_S := V_S \times \mathscr{Y}_\triangle$. Then by using the observation above and the fact $H_\# \mathbb{P}^T_{f^T} = \mathbb{P}^S_{f^S}$, we obtain

$$\mathbb{P}^S(V_S) = \mathbb{P}^S_{f^S}(U_S) = \mathbb{P}^T_{f^T}\left(H^{-1}(U_S)\right) = \mathbb{P}^T_{f^T}\left(K^{-1}(V_S) \times \mathscr{Y}_\triangle\right) = \mathbb{P}^T\left(K^{-1}(V_S)\right),$$

or equivalently, $K_\# \mathbb{P}^T = \mathbb{P}^S$. It follows from the fact $H_\# \mathbb{P}^T_{f^T} = \mathbb{P}^S_{f^S}$ that $H_2\left(x, f^T(x)\right) = f^S(K(x))$, which gives

$$\begin{aligned} \mathscr{W}_{d_Y}\left(\mathbb{P}^S_{f^S}, \mathbb{P}^T_{f^T}\right) &= \inf_{H : H_\# \mathbb{P}^T_{f^T} = \mathbb{P}^S_{f^S}} \mathbb{E}_{\left(x, f^T(x)\right) \sim \mathbb{P}^T_{f^T}}\left[d_Y\left(f^T(x), H_2\left(x, f^T(x)\right)\right)\right] \\ &\geq \inf_{K : K_\# \mathbb{P}^T = \mathbb{P}^S} \mathbb{E}_{x \sim \mathbb{P}^T}\left[d_Y\left(f^T(x), f^S(K(x))\right)\right] \\ &= \mathscr{L}S(S,T). \end{aligned}$$

In order to prove the reverse inequality, let us consider any maps K satisfying $K_\# \mathbb{P}^T = \mathbb{P}^S$. Define a map $H : \operatorname{supp}\left(\mathbb{P}^T_{f^T}\right) \to \operatorname{supp}\left(\mathbb{P}^S_{f^S}\right)$ as $H\left(x, f^T(x)\right) := \left(K(x), f^S(K(x))\right)$, we will show that $H_\# \mathbb{P}^T_{f^T} = \mathbb{P}^S_{f^S}$. Indeed, let $U_S \subset \mathscr{X}^S \times \mathscr{Y}_\triangle$ be any measurable sets and take $V_S := \left\{x \in \mathscr{X}^S \mid \left(x, f^S(x)\right) \in U_S\right\}$. Then, as $H^{-1}(U_S) = \left\{\left(x, f^S(x)\right) \mid K(x) \in V_S\right\} = \left\{\left(x, f^S(x)\right) \mid x \in K^{-1}(V_S)\right\}$, we have

$$\mathbb{P}^T_{f^T}\left(H^{-1}(U_S)\right) = \mathbb{P}^T\left(K^{-1}(V_S)\right) = \mathbb{P}^S(V_S) = \mathbb{P}^S_{f^S}(U_S),$$

which means $H_\# \mathbb{P}^T_{f^T} = \mathbb{P}^S_{f^S}$. As a result,

$$\mathscr{W}_{d_Y}\left(\mathbb{P}^S_{f^S}, \mathbb{P}^T_{f^T}\right) \leq \inf_{K : K_\# \mathbb{P}^T = \mathbb{P}^S} \mathbb{E}_{x \sim \mathbb{P}_T}\left[d_Y\left(f^T(x), f^S(K(x))\right)\right] = \mathscr{L}S(S,T).$$

By combining the above two inequalities, we obtain the desired equality

$$\mathscr{W}_{d_Y}\left(\mathbb{P}^S_{f^S}, \mathbb{P}^T_{f^T}\right) = \mathscr{L}S(S,T).$$

(ii) First, let $\gamma \in \Gamma\left(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}\right)$. According to Lemma 6, there exists $\gamma' \in \Gamma\left(\mathbb{P}^S, \mathbb{P}^T\right)$ such that $\left(f^S, f^T\right)_{\#}\gamma' = \gamma$. Then,

$$\begin{aligned}\mathbb{E}_{\left(y^S,y^T\right)\sim\gamma}\left[d_Y\left(y^S,y^T\right)\right] &= \mathbb{E}_{\left(x^S,x^T\right)\sim\gamma'}\left[d_Y\left(f^S\left(x^S\right),f^T\left(x^T\right)\right)\right] \\ &\geq \inf_{\gamma'\in\Gamma\left(\mathbb{P}^S,\mathbb{P}^T\right)} \mathbb{E}_{\left(x^S,x^T\right)\sim\gamma'}\left[d_Y\left(f^S\left(x^S\right),f^T\left(x^T\right)\right)\right] \\ &= \mathscr{W}_{d_Y}\left(\mathbb{P}^S_{f^S}, \mathbb{P}^T_{f^T}\right) = \mathscr{L}S(S,T).\end{aligned}$$

Therefore, we arrive at

$$\mathscr{W}_{d_Y}\left(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}\right) = \inf_{\gamma\in\Gamma\left(\mathbb{P}_{f^S},\mathbb{P}_{f^T}\right)} \mathbb{E}_{\left(y^S,y^T\right)\sim\gamma}\left[d_Y\left(y^S,y^T\right)\right] \geq \mathscr{L}S(S,T).$$

Second, let $\gamma' \in \Gamma\left(\mathbb{P}^S, \mathbb{P}^T\right)$, we denote $\gamma = \left(f^S, f^T\right)_{\#}\gamma'$. We then have

$$\begin{aligned}\mathbb{E}_{\left(x^S,x^T\right)\sim\gamma'}\left[d_Y\left(f^S\left(x^S\right),f^T\left(x^T\right)\right)\right] &= \mathbb{E}_{\left(y^S,y^T\right)\sim\gamma}\left[d_Y\left(y^S,y^T\right)\right] \\ &\geq \inf_{\gamma\in\Gamma\left(\mathbb{P}_{f^S},\mathbb{P}_{f^T}\right)} \mathbb{E}_{\left(y^S,y^T\right)\sim\gamma}\left[d_Y\left(y^S,y^T\right)\right] = \mathscr{W}_{d_Y}\left(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}\right).\end{aligned}$$

This follows that

$$\begin{aligned}\mathscr{L}S(S,T) = \mathscr{W}_{d_Y}\left(\mathbb{P}^S_{f^S}, \mathbb{P}^T_{f^T}\right) &= \inf_{\gamma'\in\Gamma\left(\mathbb{P}^S,\mathbb{P}^T\right)} \mathbb{E}_{\left(x^S,x^T\right)\sim\gamma'}\left[d_Y\left(f^S\left(x^S\right),f^T\left(x^T\right)\right)\right] \\ &\geq \mathscr{W}_{d_Y}\left(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}\right).\end{aligned}$$

Hence, the proof is completely done. □

Corollary 8. *The following inequality holds* $\mathscr{W}_{d_Y}\left(\mathbb{P}^S_{h^S}, \mathbb{P}^S_{f^S}\right) \leq \mathscr{L}\left(h^S, f^S, \mathbb{P}^S\right)$.

Proof. According to Proposition 1, we have:

$$\mathscr{W}_{d_Y}\left(\mathbb{P}^S_{h^S}, \mathbb{P}^S_{f^S}\right) = \inf_{L:L_{\#}\mathbb{P}^S=\mathbb{P}^S} \mathbb{E}_{x\sim\mathbb{P}^S}\left[d_Y\left(h^S(x), f^S(L(x))\right)\right].$$

Then by choosing L as the identity map (i.e., $L(x) = x$ for all x), we obtain:

$$\mathscr{W}_{d_Y}\left(\mathbb{P}^S_{h^S}, \mathbb{P}^S_{f^S}\right) \leq \mathbb{E}_{x\sim\mathbb{P}^S}\left[d_Y\left(h^S(x), f^S(L(x))\right)\right] = \mathscr{L}\left(h^S, f^S, \mathbb{P}^S\right).$$

□

B.3 Proof of Theorem 4

First, we will show that

$$\mathscr{L}S(S,T) \leq \mathscr{L}\left(h^S, f^S, \mathbb{P}^S\right) + \mathscr{L}\left(h^T, f^T, \mathbb{P}^T\right) + \mathscr{W}_{d_Y}\left(\mathbb{P}_{h^S}, \mathbb{P}_{h^T}\right).$$

By using the triangle inequality for the Wasserstein distance with respect to the metric d_Y, we have

$$\begin{aligned}\mathscr{L}S(S,T) &= \mathscr{W}_{d_Y}\left(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}\right) \\ &\overset{(1)}{\leq} \mathscr{W}_{d_Y}\left(\mathbb{P}_{f^S}, \mathbb{P}_{h^S}\right) + \mathscr{W}_{d_Y}\left(\mathbb{P}_{h^S}, \mathbb{P}_{h^T}\right) + \mathscr{W}_{d_Y}\left(\mathbb{P}_{h^T}, \mathbb{P}_{f^T}\right) \\ &= \mathscr{W}_{d_Y}\left(\mathbb{P}^S_{f^S}, \mathbb{P}^S_{h^S}\right) + \mathscr{W}_{d_Y}\left(\mathbb{P}_{h^S}, \mathbb{P}_{h^T}\right) + \mathscr{W}_{d_Y}\left(\mathbb{P}^T_{h^T}, \mathbb{P}^T_{f^T}\right) \\ &\overset{(2)}{\leq} \mathscr{L}\left(h^S, f^S, \mathbb{P}^S\right) + \mathscr{L}\left(h^T, f^T, \mathbb{P}^T\right) + \mathscr{W}_{d_Y}\left(\mathbb{P}_{h^S}, \mathbb{P}_{h^T}\right).\end{aligned}$$

Here we note that for $\overset{(1)}{\leq}$, we use the triangle inequality and for $\overset{(2)}{\leq}$, we invoke Corollary 8.

It is sufficient to prove that

$$\mathscr{W}_{d_Y}\left(\mathbb{P}_{h^S},\mathbb{P}_{h^T}\right)\leq \mathscr{W}_{d_Z}\left(g^S\#\mathbb{P}^S, g^T\#\mathbb{P}^T\right).$$

Indeed, let $\gamma\in\Gamma\left(g^S\#\mathbb{P}^S, g^T\#\mathbb{P}^T\right)$ and denote $\gamma'=h\#\gamma$. Then, we have $\gamma'\in\Gamma\left(\mathbb{P}_{h^S},\mathbb{P}_{h^T}\right)$, and

$$\mathbb{E}_{(y_1,y_2)\sim\gamma'}\left[d_Y\left(y_1,y_2\right)\right]=\mathbb{E}_{(z_1,z_2)\sim\gamma}\left[d_Y\left(h\left(z_1\right),h\left(z_2\right)\right)\right]\leq\mathbb{E}_{(z_1,z_2)\sim\gamma}\left[d_Z\left(z_1,z_2\right)\right].$$

Therefore, we obtain

$$\begin{aligned}\mathscr{W}_{d_Y}\left(\mathbb{P}_{h^S},\mathbb{P}_{h^T}\right)&=\inf_{\gamma'\in\Gamma\left(\mathbb{P}_{h^S},\mathbb{P}_{h^T}\right)}\mathbb{E}_{(y_1,y_2)\sim\gamma'}\left[d_Y\left(y_1,y_2\right)\right]\\&\leq\mathbb{E}_{(y_1,y_2)\sim\gamma'}\left[d_Y\left(y_1,y_2\right)\right]\leq\mathbb{E}_{(z_1,z_2)\sim\gamma}\left[d_Z\left(z_1,z_2\right)\right].\end{aligned}$$

Finally, we reach

$$\mathscr{W}_{d_Y}\left(\mathbb{P}_{h^S},\mathbb{P}_{h^T}\right)\leq\inf_{\gamma\in\Gamma\left(g^S\#\mathbb{P}^S,g^T\#\mathbb{P}^T\right)}\mathbb{E}_{(z_1,z_2)\sim\gamma}\left[d_Z\left(z_1,z_2\right)\right]=\mathscr{W}_{d_Z}\left(g^S\#\mathbb{P}^S,g^T\#\mathbb{P}^T\right).$$

Hence, we have proved our claim.

C Proof of the remaining results

C.1 Proof of Proposition 3

Denote by $p_Y^S=(p_Y^S(y))_{y=1}^M$ and $p_Y^T=(p_Y^T(y))_{y=1}^M$ the marginal distributions of the source and target domain labels, i.e., $p_Y^S(y)=\int_{\mathscr{X}^S}p^S(x,y)dx$ and $p_Y^T(y)=\int_{\mathscr{X}^T}p^T(x,y)dx$. Let $\mathbb{P}_Y^a$ be the discrete measure on the vertices of Δ^{M-1} putting mass $P_Y^a(y)$ on the one-hot representation of y, $\forall y\in[M], a\in\{S,T\}$. For $d_Y(y,y')=\|y-y'\|_p^p$ when $p\geq 1$, the following holds:

(i) $\mathscr{L}\left(h^T,f^T,\mathbb{P}^T\right)\leq\mathscr{L}S(S,T)+\mathscr{L}\left(h^S,f^S,\mathbb{P}^S\right)+\mathscr{W}_{d_Y}\left(\mathbb{P}^S_{h^S},\mathbb{P}^T_{h^T}\right)+const$, where the constant can be viewed as a reconstruction term: $\sup_{L,K:L\#\mathbb{P}^T=\mathbb{P}^S,K\#\mathbb{P}^S=\mathbb{P}^T}\mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(f^T\left(K\left(L\left(x\right)\right)\right),f^T\left(x\right)\right)\right]$;

(ii) $\mathscr{LS}(S,T)\geq\|p_Y^S-p_Y^T\|_p^p$;

(iii) In the setting that $\mathbb{P}^S$ and $\mathbb{P}^T$ are mixtures of well-separated Gaussian distributions, i.e.,

$$p^a(x)=\sum_{y=1}^M p^a(y)\mathscr{N}(x|\mu_y^a,\Sigma_y^a),\quad\forall a\in\{S,T\}$$

with $\|\mu_y^a-\mu_{y'}^a\|_2\geq D\times\max\{\|\Sigma_y^a\|_{op}^{1/2},\|\Sigma_{y'}^a\|_{op}^{1/2}\}\forall a\in\{S,T\},y\neq y'$, in which $\|\cdot\|_{op}$ denotes the operator norm and D is sufficiently large, we have

$$0\leq|\mathscr{LS}(S,T)-\mathscr{W}_p^p(\mathbb{P}_Y^S,\mathbb{P}_Y^T)|\leq\varepsilon(D),\tag{16}$$

where $\varepsilon(D)$ is a small constant depending on $D,p_Y^S,p_Y^T,(\Sigma_y^S,\Sigma_y^T)_{y=1}^M$, and it goes to 0 as $D\to\infty$.

(iv) In the anti-causal setting, where $p^S(x|y)=p^T(x|y)$ for all x,y,

$$\mathscr{LS}(S,T)\leq M^p\|p_Y^S-p_Y^T\|_1+\min\{\mathbb{E}_{\mathbb{P}^S}\|f^S-f^T\|_p^p,\mathbb{E}_{\mathbb{P}^T}\|f^S-f^T\|_p^p\};\tag{17}$$

Proof. **(i)** Let L and K be two arbitrary maps such that $L\#\mathbb{P}^T=\mathbb{P}^S$ and $K\#\mathbb{P}^S=\mathbb{P}^T$. We have the following triangle inequality:

$$\begin{aligned}d_Y\left(h^T\left(x\right),f^T\left(x\right)\right)\leq\;&d_Y\left(h^T\left(x\right),h^S\left(L\left(x\right)\right)\right)+d_Y\left(h^S\left(L\left(x\right)\right),f^S\left(L\left(x\right)\right)\right)\\&+d_Y\left(f^S\left(L\left(x\right)\right),f^T\left(K\left(L\left(x\right)\right)\right)\right)+d_Y\left(f^T\left(K\left(L\left(x\right)\right)\right),f^T\left(x\right)\right).\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(h^T(x), f^T(x)\right)\right] &\leq \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(h^T(x), h^S(L(x))\right)\right] + \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(h^S(L(x)), f^S(L(x))\right)\right] \\
&+ \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(f^S(L(x)), f^T(K(L(x)))\right)\right] + \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(f^T(K(L(x))), f^T(x)\right)\right] \\
&\overset{(1)}{=} \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(h^T(x), h^S(L(x))\right)\right] + \mathbb{E}_{\mathbb{P}^S}\left[d_Y\left(h^S(x), f^S(x)\right)\right] \\
&+ \mathbb{E}_{\mathbb{P}^S}\left[d_Y\left(f^S(x), f^T(K(x))\right)\right] + \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(f^T(K(L(x))), f^T(x)\right)\right] \\
&= \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(h^T(x), h^S(L(x))\right)\right] + \mathscr{L}\left(h^S, f^S, \mathbb{P}^S\right) \\
&+ \mathbb{E}_{\mathbb{P}^S}\left[d_Y\left(f^S(x), f^T(K(x))\right)\right] + \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(f^T(K(L(x))), f^T(x)\right)\right].
\end{aligned}$$

Note that, the derivation in $\overset{(1)}{=}$ is due to $L\#\mathbb{P}^T = \mathbb{P}^S$, hence gaining

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(h^S(L(x)), f^S(L(x))\right)\right] &= \mathbb{E}_{\mathbb{P}^S}\left[d_Y\left(h^S(x), f^S(x)\right)\right], \\
\mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(f^S(L(x)), f^T(K(L(x)))\right)\right] &= \mathbb{E}_{\mathbb{P}^S}\left[d_Y\left(f^S(x), f^T(K(x))\right)\right].
\end{aligned}$$

As a consequence, we find that

$$\begin{aligned}
\mathscr{L}\left(h^T, f^T, \mathbb{P}^T\right) &\leq \inf_{L,K:L\#\mathbb{P}^T=\mathbb{P}^S, K\#\mathbb{P}^S=\mathbb{P}^T} \Big\{ \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(h^T(x), h^S(L(x))\right)\right] + \mathscr{L}\left(h^S, f^S, \mathbb{P}^S\right) \\
&\qquad + \mathbb{E}_{\mathbb{P}^S}\left[d_Y\left(f^S(x), f^T(K(x))\right)\right] + \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(f^T(K(L(x))), f^T(x)\right)\right] \Big\} \\
&\leq \inf_{L:L\#\mathbb{P}^T=\mathbb{P}^S} \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(h^T(x), h^S(L(x))\right)\right] + \inf_{K:K\#\mathbb{P}^S=\mathbb{P}^T} \mathbb{E}_{\mathbb{P}^S}\left[d_Y\left(f^S(x), f^T(K(x))\right)\right] \\
&+ \mathscr{L}\left(h^S, f^S, \mathbb{P}^S\right) + \sup_{L,K:L\#\mathbb{P}^T=\mathbb{P}^S, K\#\mathbb{P}^S=\mathbb{P}^T} \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(f^T(K(L(x))), f^T(x)\right)\right] \\
&= \mathscr{W}_{d_Y}\left(\mathbb{P}^S_{h^S}, \mathbb{P}^T_{h^T}\right) + \mathscr{L}\mathscr{S}(S,T) + \mathscr{L}\left(h^S, f^S, \mathbb{P}^S\right) \\
&+ \sup_{L,K:L\#\mathbb{P}^T=\mathbb{P}^S, K\#\mathbb{P}^S=\mathbb{P}^T} \mathbb{E}_{\mathbb{P}^T}\left[d_Y\left(f^T(K(L(x))), f^T(x)\right)\right].
\end{aligned}$$

(ii) We will show that for all transformation L satisfying $L\#\mathbb{P}^T = \mathbb{P}^S$,

$$\mathbb{E}_{\mathbb{P}^T}\left\|f^T(x) - f^S(L(x))\right\|_p^p \geq \left\|p_Y^S - p_Y^T\right\|_p^p, \tag{18}$$

and then take the infimum of the LHS, which directly leads to the conclusion. Indeed, by applying Jensen inequality, we find that

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^T}\left\|f^T(x) - f^S(L(x))\right\|_p^p &= \sum_{y=1}^M \mathbb{E}_{\mathbb{P}^T}\left|p^T(y|x) - p^S(y|L(x))\right|^p \\
&\geq \sum_{y=1}^M \left|\mathbb{E}_{\mathbb{P}^T}\left(p^T(y|x) - p^S(y|L(x))\right)\right|^p \\
&= \sum_{y=1}^M \left|\mathbb{E}_{\mathbb{P}^T} p^T(y|x) - \mathbb{E}_{\mathbb{P}^S} p^S(y|x)\right|^p \\
&= \sum_{y=1}^M \left|p_Y^T(y) - p_Y^S(y)\right|^p \\
&= \left\|p_Y^S - p_Y^T\right\|_p^p
\end{aligned}$$

We have thus proved our claim.

(iii) Consider y's as one-hot vectors, i.e., vertices of the simplex. By the fact that Wasserstein distances on simplex are no greater than M, we have

$$\begin{aligned}
|\mathscr{W}_p^p(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}) - \mathscr{W}_p^p(\mathbb{P}_Y^S, \mathbb{P}_Y^T)| &= \left|\sum_{i=0}^{p-1} \mathscr{W}_p^i(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}) \mathscr{W}_p^{p-1-i}(\mathbb{P}_Y^S, \mathbb{P}_Y^T)\right| |\mathscr{W}_p(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}) - \mathscr{W}_p(\mathbb{P}_Y^S, \mathbb{P}_Y^T)| \\
&\leq pM^{p-1} |\mathscr{W}_p(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}) - \mathscr{W}_p(\mathbb{P}_Y^S, \mathbb{P}_Y^T)|.
\end{aligned}$$

Besides, by triangle inequalities,

$$|\mathscr{W}_p(\mathbb{P}_{f^S},\mathbb{P}_{f^T}) - \mathscr{W}_p(\mathbb{P}_Y^S,\mathbb{P}_Y^T)| \le \mathscr{W}_p(\mathbb{P}_{f^S},\mathbb{P}_Y^S) + \mathscr{W}_p(\mathbb{P}_{f^T},\mathbb{P}_Y^T). \tag{19}$$

Thus, we only need to prove the claimed bounds for $W_p(\mathbb{P}_{f^S},\mathbb{P}_Y^S)$ and $W_p(\mathbb{P}_{f^T},\mathbb{P}_Y^T)$. Because the proofs are similar for the source and target, in the followings, we drop the superscript S,T for the ease of notations. We first show that the mass of $\mathbb{P}_f$ concentrates near the vertices, i.e., there exists $(\alpha_y,\varepsilon_y)_{y\in[M]}$ being small numbers depends on D such that, for $Z\sim f(X)$,

$$0 \le p_Y(y) - \mathbb{P}_f(\|Z-y\|_p^p < \alpha_y) \le \varepsilon_y \quad \forall y\in[M]. \tag{20}$$

Indeed, for all y, let $B_y = \{x : \|\Sigma_y^{-1/2}(x-\mu_y)\| \le \sqrt{D}\}$. Denote the dimension of X to be d and χ_d^2 the Chi-square distribution with d degree of freedom. We have the following tail bound:

$$\mathbb{P}(X\in B_y|y) = P(\chi_d^2 \le D) \ge 1 - e^{-(D-2d)/4}.$$

Hence, we obtain that

$$\mathbb{P}(X\in B_y) = \sum_{y=1}^M \mathbb{P}(X\in B_y|y)p_Y(y) \ge \mathbb{P}(\chi_d^2\le D)p_Y(y) \ge p_Y(y) - \varepsilon_y,$$

where $\varepsilon_y = e^{-(D-2d)/4}$. Besides, if $x\in B_y$ then for any $y'\neq y$, by triangle inequalities and the definition of the operator norm,

$$\begin{aligned}
\|\Sigma_{y'}^{-1/2}(x-\mu_{y'})\|_2 &\ge \|\Sigma_{y'}\|_{op}^{-1/2}\|(x-\mu_{y'})\|_2 \\
&\ge \|\Sigma_{y'}\|_{op}^{-1/2}(\|(\mu_y-\mu_{y'})\|_2 - \|(x-\mu_y)\|_2) \\
&\ge \|\Sigma_{y'}\|_{op}^{-1/2}(\|(\mu_y-\mu_{y'})\|_2 - \|(\mu_y-\mu_{y'})\|_2/\sqrt{D}) \\
&\ge D - \sqrt{D},
\end{aligned}$$

where the above inequalites are due to our assumption and the fact that

$$\|x-\mu_y\|_2 \le \|\Sigma_y\|_{op}^{1/2}\|\Sigma_y^{-1/2}(x-\mu_y)\|_2 \le \|\Sigma_y\|_{op}^{1/2}\sqrt{D} \le \|\mu_y-\mu_{y'}\|_2/\sqrt{D}.$$

Hence, for all $x\in B_y$ and $y'\neq y$, we have

$$\frac{p(x|y')}{p(x|y)} = \frac{|\Sigma_y|^{1/2}}{|\Sigma_{y'}|^{1/2}} e^{(\|\Sigma_y^{-1/2}(x-\mu_y)\|_2^2 - \|\Sigma_{y'}^{-1/2}(x-\mu_{y'})\|_2^2)/2} \lesssim e^{D-(D-\sqrt{D})^2}.$$

Combining the above inequality with Bayes' rule leads to

$$p(y|x) = \frac{1}{1+\sum_{y'\neq y}\frac{p(y')p(x|y')}{p(y)p(x|y)}} \ge 1 - \sum_{y'\neq y}\frac{p(y')p(x|y')}{p(y)p(x|y)} \ge 1-\gamma_y,$$

where $\gamma_y \lesssim e^{-D(D-2\sqrt{D})}$. This means the difference between labeling function at $x\in B_y$ and y is bounded as follows

$$\|f(x)-y\|_p^p = (1-p(y|x))^p + \sum_{y'\neq y} p(y'|x)^p \le 2(1-p(y|x))^p \le 2\gamma_y^p.$$

Choosing $\alpha_y = 2\gamma_y^p$, by the fact that $x\in B_y$ implies $\|f(x)-y\|_p^p \le \alpha_y$, we have

$$\mathbb{P}_f(\|Z-y\|_p^p \le \alpha_y) \ge \mathbb{P}(X\in B_y) \ge p_Y(y) - \varepsilon_y.$$

Putting the above results together, we find that

$$p_Y(y) - \mathbb{P}_f(\|Z-y\|_p^p \le \alpha_y) \le \varepsilon_y.$$

Due to the continuity of $\mathbb{P}_f$, we can also shrink α_y such that the inequality still holds and the left-hand side is positive and we get our claim (20).

Now let $E_y = \{z : \|z-y\|_p^p \le \alpha_y\}$ and D_y be a set containing E_y for all $y\in[M]$ satisfying $\mathbb{P}_f(D_y) = p_Y(y)$ and $\{D_y\}_{y=1}^M$ is a partition of Δ^{M-1}. Let p_f be the density of $\mathbb{P}_f$. It can be seen that

$$\pi(z,y) = p_f(z)1[z\in D_y]$$

is the density function of a coupling between $\mathbb{P}_f$ and $\mathbb{P}_Y$. Hence, we have the following inequalities:

$$\begin{aligned}
\mathscr{W}_p^p(\mathbb{P}_f,\mathbb{P}_Y) &\leq \sum_{y=1}^{M}\int_{z\in\Delta^{M-1}} \|y-z\|_p^p\,\pi(z,y)dz \\
&\leq \sum_{y=1}^{M}\int_{z\in D_y} \|y-z\|_p^p\,\pi(z,y)dz \\
&\leq \sum_{y=1}^{M}\left(\int_{z\in E_y} \|y-z\|_p^p\,\pi(z,y)+\int_{z\in D_y\setminus E_y} \|y-z\|_p^p\,\pi(z,y)\right)dz \\
&\leq \sum_{y=1}^{M}\left(\int_{z\in E_y} \alpha_y\pi(z,y)dz+\int_{z\in D_y\setminus E_y} M^p\pi(z,y)dz\right)dz \\
&\leq \sum_{y=1}^{M}\alpha_y+M^p\varepsilon_y \lesssim e^{-pD(D-2\sqrt{D})}+e^{-(D-2d)/4},
\end{aligned}$$

which goes to 0 exponentially fast when D grows to infinity. Plugging the above inequality into equation (19), we obtain the conclusion of part (iii) of the proposition.

(iv) Let $\pi_y = P(\cdot|y)$ be conditional measure of X given $Y=y$. By using the law of total probability, we have

$$\mathbb{P}^S = \sum_{y=1}^{M} p_Y^S(y)\pi_y, \quad \mathbb{P}^T = \sum_{y=1}^{M} p_Y^T(y)\pi_y. \tag{21}$$

Given the above equations, some simple algebraic transformations would lead to

$$\mathbb{P}_{f^S} = f^S\#\mathbb{P}^S = \sum_{y=1}^{M} p_Y^S(y) f^S\#\pi_y. \tag{22}$$

Now let $q_{yy} = \min\{p_Y^S(y), p_Y^T(y)\}$ and choose $(q_{yy'})_{y\neq y'}$ such that $(q_{yy'})_{y=\overline{1,M},y'=\overline{1,M}}$ is a valid coupling of p_Y^S and p_Y^T. By the convexity of Wasserstein distance, we have

$$W_p^p(\mathbb{P}_{f^S},\mathbb{P}_{f^T}) \leq \sum_{y,y'} q_{yy'} W_p^p(f^S\#\pi_y, f^T\#\pi_{y'}). \tag{23}$$

As the distance between two arbitrary points on the simplex Δ^{M-1} is not greater than M, neither is the Wasserstein distance between any two measures on Δ^{M-1}. Therefore,

$$\sum_{y\neq y'} q_{yy'} W_p^p(f^S\#\pi_y, f^T\#\pi_{y'}) \leq M^p \sum_{y\neq y'} q_{yy'} \leq M^p \|p_Y^S - p_Y^T\|_1. \tag{24}$$

Besides,

$$\begin{aligned}
\sum_{y=1}^{M} q_{yy} W_p^p(f^S\#\pi_y, f^T\#\pi_y) &\leq \sum_{y=1}^{M} p_Y^S(y) \inf_{\pi\in\Gamma(\pi_y,\pi_y)} \int_{\mathscr{X}\times\mathscr{X}} \left\|f^S(x)-f^T(x')\right\|_p^p d\pi(x,x') \\
&\leq \sum_{y=1}^{M} p_Y^S(y) \int_{\mathscr{X}} \left\|f^S(x)-f^T(x)\right\|_p^p d\pi_y(x) \\
&= \sum_{y=1}^{M} p_Y^S(y) \int_{\mathscr{X}} \left\|f^S(x)-f^T(x)\right\|_p^p p(x|y)dx \\
&= \int_{\mathscr{X}} \left\|f^S(x)-f^T(x)\right\|_p^p p^S(x)dx \\
&= \mathbb{E}_{\mathbb{P}^S}\left\|f^S-f^T\right\|_p^p,
\end{aligned}$$

Similarly, since $q_{yy} \leq p^T(y)$ for all $y\in[M]$, we could also obtain

$$\sum_{y=1}^{M} q_{yy} W_p^p(f^S\#\pi_y, f^T\#\pi_y) \leq \mathbb{E}_{\mathbb{P}^T}\left\|f^S-f^T\right\|_p^p.$$

Consequently,

$$\sum_{y=1}^{M} q_{yy} W_p^p(f^S \# \pi_y, f^T \# \pi_y) \leq \min\{\mathbb{E}_{\mathbb{P}^S} \left\| f^S - f^T \right\|_p^p, \mathbb{E}_{\mathbb{P}^T} \left\| f^S - f^T \right\|_p^p\}. \tag{25}$$

Combining equations (24) and (25), we have the conclusion of part (iv). □

C.2 Proof of Theorem 5

Before providing the proof of Theorem 5, we first introduce a lemma which facilitates our later arguments.

Lemma 9. *Let μ and ν be two probability measures on $\mathbb{R}^d$. Denote by μ_1 (ν_1) and μ_2 (ν_2) the marginal distributions of μ (ν) on the first k dimensions and the last $d-k$ dimensions, respectively, where $0 \leq k \leq d$. We have*

$$\mathscr{W}_p^p(\mu, \nu) \geq \mathscr{W}_p^p(\mu_1, \nu_1) + \mathscr{W}_p^p(\mu_2, \nu_2) \tag{26}$$

Proof. Let π be the optimal coupling of μ and ν, and $(X_1, \ldots, X_d, Y_1, \ldots, Y_d)$ is a random vector having law π, we have $(X_1, \ldots, X_d) \sim \mu, (Y_1, \ldots, Y_d) \sim \nu$. Denote by π_1 and π_2 the marginal distribution of $(X_1, \ldots, X_k, Y_1, \ldots, Y_k)$ and $(X_{k+1}, \ldots, X_d, Y_{k+1}, \ldots, Y_d)$, respectively. It can be seen that π_1 is a coupling of μ_1 and ν_1 while π_2 is a coupling of μ_2 and ν_2. Hence,

$$\begin{aligned}
\mathscr{W}_p^p(\mu, \nu) &= \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_p^p d\pi(x, y) \\
&= \int_{\mathbb{R}^k \times \mathbb{R}^k} \|x_1 - y_1\|_p^p d\pi_1(x, y) + \int_{\mathbb{R}^{d-k} \times \mathbb{R}^{d-k}} \|x_2 - y_2\|_p^p d\pi_2(x, y) \\
&\geq \mathscr{W}_p^p(\mu_1, \nu_1) + \mathscr{W}_p^p(\mu_2, \nu_2),
\end{aligned}$$

where $x_1 \in \mathbb{R}^k$ is a vector including the first k coordinates of x, whereas $x_2 \in \mathbb{R}^{d-k}$ contains the last $d-k$ elements of x, and similar definitions apply for y_1 and y_2. □

Now, we come back to the proof of Theorem 2.

Proof of Theorem 5:

(i) We consider two random vectors $(X_1, \ldots, X_M) \sim \mathbb{P}_{f^S}$ and $(Y_1, \ldots, Y_M) \sim \mathbb{P}_{f^T}$. Recall that $\mathbb{Q}_S$ and $\mathbb{Q}_T$ are the marginal distributions of $\mathbb{P}_{f^S}$ and $\mathbb{P}_{f^T}$ on their first $C-1$ dimensions, respectively, while $\mathbb{Q}_{S \setminus T}$ denotes the marginal of $\mathbb{P}_{f_S}$ on the space of variables having labels in the set $\mathscr{Y}_S \setminus \mathscr{Y}_T$. Since $\mathscr{Y}_T \subset \mathscr{Y}_S$, we have

1. $(X_1, \ldots, X_{C-1}) \sim \mathbb{Q}_S$, $(Y_1, \ldots, Y_{C-1}) \sim \mathbb{Q}_T$;
2. $(X_{C+1}, \ldots, X_M) \sim \mathbb{Q}_{S \setminus T}$, $(Y_{C+1}, \ldots, Y_M) \sim \delta_{\mathbf{0}_{M-C}}$,

where $\mathbf{0}_m$ denotes the zero vector in $\mathbb{R}^m$ for $m \in \mathbb{N}$. Let μ be the distribution of $(X_1, \ldots, X_{C-1}, X_{C+1}, \ldots, X_M)$, ν the distribution of $(Y_1, \ldots, Y_{C-1}, Y_{C+1}, \ldots, Y_M)$. As the simplex Δ^{M-1} is an $M-1$ dimensional manifold in $\mathbb{R}^M$, the Wasserstein distance between $\mathbb{P}_{f^S}$ and $\mathbb{P}_{f^T}$ can be written as

$$\mathscr{W}_p^p(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}) = \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathbb{R}^{C-1} \times \mathbb{R}^{C-1}} \sum_{i=1}^{M} |x_i - y_i|^p d\gamma(x_{\widehat{C}}, y_{\widehat{C}}),$$

where $x_{\widehat{C}} = (x_1, \ldots, x_{C-1}, x_{C+1}, \ldots, x_M), y_{\widehat{C}} = (y_1, \ldots, y_{C-1}, y_{C+1}, \ldots, y_M)$ and $x_C := 1 - \sum_{k \neq C} x_k, y_C := 1 - \sum_{k \neq C} y_k$. As $|x_C - y_C|^p \geq 0$, it can be deduced that

$$\mathscr{W}_p^p(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}) \geq \mathscr{W}_p^p(\mu, \nu).$$

Besides, according to Lemma 9, we get

$$\mathscr{W}_p^p(\mu, \nu) \geq \mathscr{W}_p^p(\mathbb{Q}_S, \mathbb{Q}_T) + \mathscr{W}_p^p(\mathbb{Q}_{S \setminus T}, \delta_{\mathbf{0}_{M-C}}) = \mathscr{W}_p^p(\mathbb{Q}_S, \mathbb{Q}_T) + \mathbb{E}_{X \sim \mathbb{Q}_{S \setminus T}} \left[\|X\|_p^p \right].$$

Putting the above two inequalities together, we obtain the conclusion that

$$\mathscr{W}_p^p(\mathbb{P}_{f^S}, \mathbb{P}_{f^T}) \geq \mathscr{W}_p^p(\mathbb{Q}_S, \mathbb{Q}_T) + \mathbb{E}_{X \sim \mathbb{Q}_{S \setminus T}} \left[\|X\|_p^p \right]. \tag{27}$$

(ii) Part (ii) is done similarly to part (i). Therefore, it is omitted.

(iii) Let $(X_1,\ldots,X_M)\sim\mathbb{P}_{f^S}$ and $(Y_1,\ldots,Y_M)\sim\mathbb{P}_{f^T}$. Assume that $\mathscr{Y}_S=\{1,\ldots,C,C+1,\ldots,D\}$ and $\mathscr{Y}_T=\{1,\ldots,C,D+1,\ldots,M\}$. It follows from the definitions of $\mathbb{Q}_S,\mathbb{Q}^T,\mathbb{Q}_{S\setminus T}$ and $\mathbb{Q}_{T\setminus S}$ that

1. $(X_1,\ldots,X_{C-1})\sim\mathbb{Q}_S$, $(Y_1,\ldots,Y_{C-1})\sim\mathbb{Q}_T$;
2. $(X_{C+1},\ldots,X_D)\sim\mathbb{Q}_{S\setminus T}$, $(Y_{C+1},\ldots,Y_D)\sim\delta_{\mathbf{0}_{D-C}}$;
3. $(X_{D+1},\ldots,X_M)\sim\delta_{\mathbf{0}_{M-D}}$, $(Y_{D+1},\ldots,Y_M)\sim\mathbb{Q}_{T\setminus S}$.

Let μ be the distribution of $(X_1,\ldots,X_{C-1},X_{C+1},\ldots,X_M)$, and ν be the distribution of $(Y_1,\ldots,Y_{C-1},Y_{C+1},\ldots,Y_M)$. By using the same arguments as in part (i), we get $\mathscr{W}_p^p(\mathbb{P}_{f^S},\mathbb{P}_{f^T})\geq\mathscr{W}_p^p(\mu,\nu)$. Next, applying Lemma 9 twice, we obtain

$$\begin{aligned}\mathscr{W}_p^p(\mu,\nu)&\geq\mathscr{W}_p^p(\mathbb{Q}_S,\mathbb{Q}_T)+\mathscr{W}_p^p(\mathbb{Q}_{S\setminus T},\delta_{\mathbf{0}_{D-C}})+\mathscr{W}_p^p(\delta_{\mathbf{0}_{M-D}},\mathbb{Q}_{\mathbb{T}\setminus\mathbb{S}})\\&=\mathscr{W}_p^p(\mathbb{Q}_S,\mathbb{Q}_T)+\mathbb{E}_{Y\sim\mathbb{Q}_{S\setminus T}}\left[\|Y\|_p^p\right]+\mathbb{E}_{X\sim\mathbb{Q}_{T\setminus S}}\left[\|X\|_p^p\right].\end{aligned}$$

As a consequence, we have proved our claim in part (iii).

C.3 Proofs of claims in paragraph "Remark on shifting term"

Lemma 10. *The followings are true:*

(i) $\mathscr{W}_{d_X}(g\#\mathbb{P}^S,g\#\mathbb{P}^T)=\mathscr{W}_{d_X}(\mathbb{P}^S_{h^S},\mathbb{P}^T_{h^T})$;

(ii) $\mathscr{W}_{d_Y}\left(\mathbb{P}^T_{h^T},\mathbb{P}^T_{f^T}\right)\leq\mathscr{LS}(S,T)+\mathscr{L}\left(h^S,f^S,\mathbb{P}^S\right)+\mathscr{W}_{d_Y}\left(\mathbb{P}^S_{h^S},\mathbb{P}^T_{h^T}\right)$.

Proof. **(i)** Applying the same arguments as in Proposition 1, we have

$$\mathscr{W}_{d_X}(\mathbb{P}^S_{h^S},\mathbb{P}^T_{h^T})=\inf_{H:H\#\mathbb{P}^T_{h^T}=\mathbb{P}^S_{h^S}}\mathbb{E}_{\mathbb{P}^T_{h^T}}\left[d_X[g(x),g(H_1(x))]\right]=\inf_{L:L\#\mathbb{P}^T=\mathbb{P}^S}\mathbb{E}_{\mathbb{P}^T}\left[d_X(g(x),g(L(x)))\right],\quad(28)$$

where $H((x,h^T(x)))=(H_1((x,h^T(x))),H_2((x,h^T(x))))$ such that $H_1((x,h^T(x)))\in\mathscr{X}^S$. So now we only need to prove that

$$\mathscr{W}_{d_X}(g\#\mathbb{P}^S,g\#\mathbb{P}^T)=\inf_{L:L\#\mathbb{P}^T=\mathbb{P}^S}\mathbb{E}_{\mathbb{P}^T}\left[d_X(g(x),g(L(x)))\right].\quad(29)$$

Due to the equivalence of Monge and Kantorovich problem, we can write the RHS as

$$\inf_{L:L\#\mathbb{P}^T=\mathbb{P}^S}\mathbb{E}_{\mathbb{P}^T}\left[d_X(g(x),g(L(x)))\right]=\inf_{\gamma\in\Gamma(\mathbb{P}^S,\mathbb{P}^T)}\mathbb{E}_{(x^S,x^T)\sim\gamma}\left[d_X(g(x^S),g(x^T))\right].\quad(30)$$

To prove Eq. (29), we will show that RHS is not less than LHS and inversely. Indeed, for any coupling $\gamma\in\Gamma(\mathbb{P}^S,\mathbb{P}^T)$, we have $\gamma'=(g,g)\#\gamma$ as a coupling of $(g\#\mathbb{P}^S,g\#\mathbb{P}^T)$, therefore

$$\mathbb{E}_{(x^S,x^T)\sim\gamma}\left[d_X(g(x^S),g(x^T))\right]=\mathbb{E}_{(\mathbf{y}^S,\mathbf{y}^T)\sim\gamma'}\left[d_X(\mathbf{y}^S,\mathbf{y}^T)\right]\geq\inf_{\gamma'\in\Gamma(g\#\mathbb{P}^S,g\#\mathbb{P}^T)}\mathbb{E}_{(\mathbf{y}^S,\mathbf{y}^T)\sim\gamma'}\left[d_X(\mathbf{y}^S,\mathbf{y}^T)\right].$$

Taking the infimum with respect to γ,

$$\inf_{\gamma\in\Gamma(\mathbb{P}^S,\mathbb{P}^T)}\mathbb{E}_{(x^S,x^T)\sim\gamma}\left[d_X(g(x^S),g(x^T))\geq\mathscr{W}_{d_X}(g\#\mathbb{P}^S,g\#\mathbb{P}^T)\right].\quad(31)$$

Conversely, thanks to Lemma 6, for each coupling γ' of $(g\#\mathbb{P}^S,g\#\mathbb{P}^S)$, there exists a coupling γ of $(\mathbb{P}^S,\mathbb{P}^S)$ such that $\gamma'=(g,g)\#\gamma$, which deduces that

$$\mathbb{E}_{(\mathbf{y}^S,\mathbf{y}^T)\sim\gamma'}\left[d_X(\mathbf{y}^S,\mathbf{y}^T)\right]=\mathbb{E}_{(x^S,x^T)\sim\gamma}\left[d_X(g(x^S),g(x^T))\right]\geq\inf_{\gamma\in\Gamma(\mathbb{P}^S,\mathbb{P}^T)}\mathbb{E}_{(x^S,x^T)\sim\gamma}\left[d_X(g(x^S),g(x^T))\right].$$

Taking the infimum with respect to γ',

$$\mathscr{W}_{d_X}(g\#\mathbb{P}^S,g\#\mathbb{P}^T)\geq\inf_{\gamma\in\Gamma(\mathbb{P}^S,\mathbb{P}^T)}\mathbb{E}_{(x^S,x^T)\sim\gamma}\left[d_X(g(x^S),g(x^T))\right].\quad(32)$$

Inequalities (31) and (32) together imply equation (29) and finish proof of part (i).

(ii) Using triangle inequality, we have

$$\begin{aligned}\mathscr{W}_{d_Y}\left(\mathbb{P}^T_{h^T},\mathbb{P}^T_{f^T}\right) &= \mathscr{W}_{d_Y}\left(\mathbb{P}_{h^T},\mathbb{P}_{f^T}\right) \leq \mathscr{W}_{d_Y}\left(\mathbb{P}_{h^T},\mathbb{P}_{h^S}\right) + \mathscr{W}_{d_Y}\left(\mathbb{P}_{h^S},\mathbb{P}_{f^S}\right) + \mathscr{W}_{d_Y}\left(\mathbb{P}_{f^S},\mathbb{P}_{f^T}\right) \\ &= \mathscr{W}_{d_Y}\left(\mathbb{P}^T_{h^T},\mathbb{P}^S_{h^S}\right) + \mathscr{L}\left(h^S,f^S,\mathbb{P}^S\right) + \mathscr{LS}\left(S,T\right).\end{aligned}$$

As a consequence, we obtain the conclusion of part (ii). □

D Additional experiment results

D.1 More analysis about rationale of the terms used in the objective function of LDROT

This objective function consists of three losses: (i) standard loss $\mathscr{L}^S$, (ii) shifting loss $\mathscr{L}^{shift}$, and (iii) clustering loss $\mathscr{L}^{clus}$. The standard loss $\mathscr{L}^S$ is trained on the labeled source domain. The shifting loss aims to reduce both data and label shift simultaneously on the latent space by minimizing $\mathscr{W}_d\left(\mathbb{P}^S_{h^S},\mathbb{P}^T_{h^T}\right)$ where $d = \lambda d_X + d_Y$ for which d_X is data distance on the latent space and d_Y is distance on the label simplex. Based on the theory developed, we demonstrate that minimizing this term helps to reduce both data shift (i.e., $\mathscr{W}_{d_X}\left(g_\#\mathbb{P}^S, g_\#\mathbb{P}^T\right)$ and label shift (i.e., $\mathscr{W}_{d_Y}\left(\mathbb{P}^S_{h^S},\mathbb{P}^T_{h^T}\right)$). Finally, the $\mathscr{L}^{clus}$ assists us in enforcing the clustering assumption to boost the generalization of the target classifier $\bar{h}^T$. By enforcing the clustering assumption, classifiers are encouraged to preserve the cluster structure and give the same predictions for data representations in the same cluster. It appears that when pushing target latent toward source representations via minimizing $\mathscr{W}_d\left(\mathbb{P}^S_{h^S},\mathbb{P}^T_{h^T}\right)$, source and target representations tend to group in clusters, hence we can strengthen and boost generalization of target classifier $\bar{h}^T$ by enforcing it to preserve the predictions in the same clusters.

In addition, we propose a dynamic weighting for λ using similarities of pairs between source and target examples. For our similarity-based weighting distance, we base on pre-trained similarities to decide if we push more or fewer pairs of source and target latent representations together to reduce data and label shifts more efficiently. Definitely, if we can push groups of source and target representations with the same labels together more efficiently, we can certainly reduce both data and label shifts simultaneously.

D.2 Data preparation and pre-processing

Digits. We resize the resolution of each sample in the dataset to 32×32, and normalize the value of each pixel to the range of $[-1, 1]$.

Office-31, Office-Home, and **ImageCLEF-DA**. We use 2048-dimensional features extracted from ResNet-50 (He et al., 2016) pretrained on ImageNet.

D.3 Algorithm of LDROT

We present peusocode of LDROT in Algorithm 1.

D.4 Network architecture

There are 2 types of the architecture described in Table 3, which are small (**S**) and large (**L**) networks. We use **L** network for *Digits* and **S** network for the other datasets. Additionally, excluding dense layers in the ϕ network, we add the batch normalization layers on top of convolutional and dense layers to reduce the overfitting problem. Finally, we implement our LDROT in Python using TensorFlow (version 1.9.0) (Abadi et al., 2016), an open-source software library for Machine Intelligence developed by the Google Brain Team. All experiments are run on a computer with an NVIDIA Tesla V100 SXM2 with 16 GB memory.

Algorithm 1 Pseudocode for training our LDROT.

Input: A source batch $\mathscr{B}^S = \left\{\left(x_i^S, y_i^S\right)\right\}_{i=1}^{Mb}$, a target batch $\mathscr{B}^T = \left\{\left(x_j^T, y_j^T\right)\right\}_{j=1}^{b}$, b is the batch size.

Output: Classifier $\bar{h}^{*S}\bar{h}^{*T}$, generator g^*.

Evaluate $\left\{r_i^S\right\}_{i=1}^{Mb}$ and $\left\{r_j^T\right\}_{j=1}^{b}$ based on $\mathscr{B}^S$ and $\mathscr{B}^T$ respectively.

Compute the weights $\mathbf{w}_{ij}$ ibased on $\left\{r_i^S\right\}_{i=1}^{Mb}$ and $\left\{r_j^T\right\}_{j=1}^{b}$.

for number of training iterations **do**
  for k steps **do**
    Update ϕ according to Eq. (13).
  end for
  Update $\bar{h}^S, \bar{h}^T$ and g according to Eq. (10).
end for

Table 3: Small and large networks for LDROT. The Leaky ReLU (lReLU) parameter a is set to 0.1.

Architecture	S	L
Input size	2048	$32 \times 32 \times 3$
Generator g		instance normalization
	256 dense, ReLU	3×3 conv. 64 lReLU
	dropout, $p = 0.5$	3×3 conv. 64 lReLU
	Gaussian noise, σ = 1	3×3 conv. 64 lReLU
		2×2 max-pool, stride 2
		dropout, $p = 0.5$
		Gaussian noise, σ = 1
		3×3 conv. 64 lReLU
		3×3 conv. 64 lReLU
		3×3 conv. 64 lReLU
		2×2 max-pool, stride 2
		dropout, $p = 0.5$
		Gaussian noise, σ = 1
		3×3 conv. 8 lReLU
		2×2 max-pool, stride 2
Classifier $\bar{h}^S, \bar{h}^T$	*#classes* dense, softmax	3×3 conv. 8 lReLU
		3×3 conv. 8 lReLU
		3×3 conv. 8 lReLU
		global average pool
		#classes dense, softmax
ϕ	1 dense, linear	100 dense, ReLU
		1 dense, linear

D.5 Implementation details

We first present our procedure to compute the weights $\mathbf{w}_{ij}$, which derives from the feature extraction process. For *Digits*, we design a network to train from scratch on labeled source examples. Then source and target features are extracted via this pretrained model. For the other datasets, we use extracted ResNet-50 features (He et al., 2016) and design a small network to train LDROT. During training, the features are used for first computing pairwise similarity scores and the weights $\mathbf{w}_{ij}$ after that.

For LDROT, we find that some hyper-parameters contributes substantially to the model performance, namely τ and ε. The temperature parameter τ, which contributes to sharpening and contrasting the weights $\mathbf{w}_{ij}$, is fixed to 0.5. Tweaking the regularization rate ε is vital for scaling $\phi_\varepsilon^c(x)$ and we select $\varepsilon = 0.1$. For trade-off parameters α, β, we choose $\alpha = 0.1$ and $\beta = 0.5$ for all settings. We apply Adam optimizer Kingma & Ba (2014) ($\beta_1 = 0.5, \beta_2 = 0.999$) with Polyak averaging. The

Table 4: Classification accuracy (%) on Digits dataset for unsupervised domain adaptation.

Method	S→M	M→U	U→M	Avg
DANN (Ganin & Lempitsky, 2015)	85.5	84.9	86.3	85.6
ADDA (Tzeng et al., 2017)	89.2	85.4	96.5	90.4
DeepCORAL (Sun & Saenko, 2016)	88.3	84.1	93.6	88.7
CDAN (Long et al., 2018)	89.2	95.6	98.0	94.3
TPN (Pan et al., 2019)	93.0	92.1	94.1	93.1
rRevGrad+CAT (Deng et al., 2019)	98.8	94.0	96.0	96.3
SWD (Lee et al., 2019)	98.9	98.1	97.1	98.0
DeepJDOT (Damodaran et al., 2018)	96.7	95.7	96.4	96.3
DASPOT (Xie et al., 2019)	96.2	97.5	96.5	96.7
ETD (Li et al., 2020)	97.9	96.4	96.3	96.9
RWOT (Xu et al., 2020)	98.8	**98.5**	97.5	98.3
LDROT	**99.0**	98.2	**99.1**	**98.8**

Table 5: Classification accuracy (%) on ImageCLEF-DA dataset for unsupervised domain adaptation (ResNet-50).

Method	I→P	P→I	I→C	C→I	C→P	P→C	Avg
ResNet-50 (He et al., 2016)	74.8	83.9	91.5	78.0	65.5	91.2	80.7
DeepCORAL (Sun & Saenko, 2016)	75.1	85.5	92.0	85.5	69.0	91.7	83.1
DANN (Ganin et al., 2016)	75.0	86.0	96.2	87.0	74.3	91.5	85.0
ADDA (Tzeng et al., 2017)	75.5	88.2	96.5	89.1	75.1	92.0	86.0
CDAN (Long et al., 2018)	77.7	90.7	97.7	91.3	74.2	94.3	87.7
TPN (Pan et al., 2019)	78.2	92.1	96.1	90.8	76.2	95.1	88.1
SymNets (Zhang et al., 2019b)	80.2	93.6	97.0	93.4	78.7	96.4	89.9
SAFN (Xu et al., 2019)	79.3	93.3	96.3	91.7	77.6	95.3	88.9
rRevGrad+CAT (Deng et al., 2019)	77.2	91.0	95.5	91.3	75.3	93.6	87.3
DeepJDOT (Damodaran et al., 2018)	77.5	90.5	95.0	88.3	74.9	94.2	86.7
ETD (Li et al., 2020)	81.0	91.7	**97.9**	93.3	79.5	95.0	89.7
RWOT (Xu et al., 2020)	81.3	92.9	**97.9**	92.7	79.1	96.5	90.0
LDROT	**81.7**	**96.7**	97.5	**94.2**	**80.4**	**96.7**	**91.2**

learning rate is set to 0.001 and 0.0001 for Digits and the other datasets respectively. Additionally, in nature, our model solves the minimax optimization problem (see Eq. (13) in the main paper) in which ϕ and $\bar{h}^S, \bar{h}^T, g$ are updated sequentially in each iteration with five times for ϕ and one time for $\bar{h}^S, \bar{h}^T, g$. Finally, we use the cosine distance for d_X and Kullback-Leibler (KL) divergence for d_Y.

D.6 Additional results for Digits and ImageCLEF-DA

We additionally present the experimental results for Digits and ImageCLEF-DA datasets in Tables 4 and 5. It can be observed that LDROT also outperforms the baselines on these datasets.

D.7 Ablation studies

Effects of the label shift term d_Y and the weights $\mathbf{w}_{ij}$: To answer the question of *how will the model performance be affected when the weights $\mathbf{w}_{ij}$ or d_Y in Eq. (10) is removed?*, we evaluate our model on four different settings: LDROT without both the weights $\mathbf{w}_{ij}$ and the label shift $d_Y(\cdot,\cdot)$ (LDROT$-wd$), without the weights $\mathbf{w}_{ij}$ (LDROT$-w$), without minimizing the label shift $d_Y(\cdot,\cdot)$ (LDROT$-d$) and a complete model (LDROT$+wd$). The results on *Office-Home* in Table 6 dedicate the significance of $\mathbf{w}_{ij}$ and $d_Y(\cdot,\cdot)$, where these components remarkably contribute to reducing the data and label shifts with 4.1% improvements on average.

Rationale of weigh strategy. In Figure 4, we visualize the similarity scores of a randomly selected target example x_j^T in a batch w.r.t. source examples x_i^S using the source and target domains *Amazon*

Table 6: Accuracy (%) of ablation study on *Office-Home*.

Method	Ar→Cl	Cl→Ar	Cl→Rw	Pr→Ar	Rw→Ar	Avg
LDROT$-wd$	51.7	62.6	78.5	63.3	66.1	64.4
LDROT$-w$	56.7	62.3	79.3	64.0	66.8	65.8
LDROT$-d$	55.4	65.5	80.1	65.0	68.4	66.9
LDROT$+wd$	**57.4**	**67.2**	**80.7**	**66.5**	**70.9**	**68.5**

and *Dslr* of *Office-31* dataset, respectively. The orange points represent the similarity scores for the same class, while the blue points represent those for different classes. It is evident that the orange values tend to bigger than the blue ones except in some outlier cases, hence if we choose μ_i as indicated, we can separate well the orange and blue values.

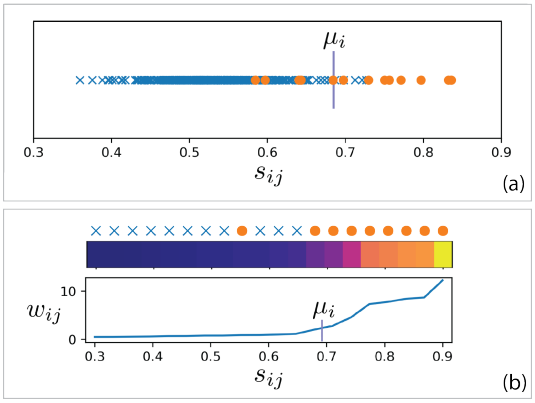
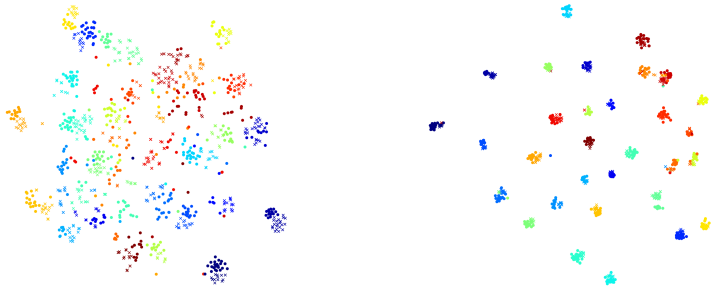


Figure 4: (a) 1D visualization of finding an appropriate μ_i. It is able to split same-label pairs (orange points) and different-label pairs (blue points) if μ_i is set to $\frac{M-1}{M}$-percentile of this array. (b) To observe the weight values w_{ij} of those pairs, we randomly picked a target example to compute similarity scores with 20 representative source points in a batch, and then sort them in ascending order. After computing the weights, the figures for the same-label pairs tend to be much higher than that for different-label pairs. A heat-map color is used to represent the weights magnitude (the brighter means higher value).



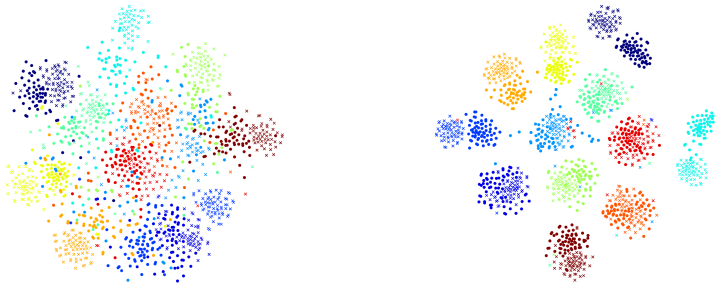
(a) ResNet.

(b) LDROT.

Figure 5: The t-SNE visualization of **A**→**D** (Figure a, b) tasks with label and domain information. Each color denotes a class while the circle and cross markers represent the source and target data respectively.

Feature visualization. We visualize the features of ResNet-50 and our methods on **A**→**D** (*Office-31*) and **P**→**C** (*ImageCLEF-DA*) tasks by t-SNE van der Maaten & Hinton (2008) in Figure 5. The visualizations in Figure 5a and 6a show that ResNet-50 classifies quite well on source domains (**A** and **P**) but poorly on target domains (**D** and **C**). While the representation in Figure 5b and 6b is

generated by our method with better alignment. LDROT achieves exactly 31 and 12 clusters corresponding to 31 and 12 classes of *Office-31* and *ImageCLEF-DA*, which represents generalization ability of our model in which the classifier generalizes well not only on the source domain but also on the target domain.



(a) ResNet.

(b) LDROT.

Figure 6: The t-SNE visualization of and **P**→**C** (Figure a, b) tasks with label and domain information. Each color denotes a class while the circle and cross markers represent the source and target data respectively.

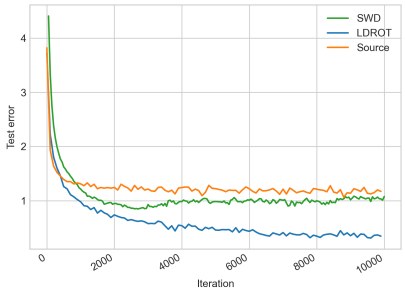


Figure 7: Comparision of convergence performance between LDROT and other approaches on the transfer task **A**→**D**.

Convergence. We testify the convergence of our LDROT with the test errors on **A**→**D** task, as shown in Figure 7. We conduct experiments on three methods including *Source* (test error is achieved with classifier trained on source data without adaptation), *SWD* Lee et al. (2019), and our *LDROT*. For fair comparison, the methods are applied the same optimizer (Adam with learning rate of 0.0001) and batch size. The results show that the error of *LDROT* on the target domain is remarkably lower, which illustrates better generalization capability of the source classifier. During training, our method encourages a target sample to actively moving to a suitable group or cluster of source examples in a similarity-aware manner. This phenomenon implies that *LDROT* enjoys faster and stable convergence than the other settings.