Domain Adaptation and Entanglement: an Optimal Transport Perspective

Okan Koç
RIKEN AIP

Alexander Soen
RIKEN AIP & Australian National University

Chao-Kai Chiang
The University of Tokyo

Masashi Sugiyama
The University of Tokyo & RIKEN AIP

Abstract

Current machine learning systems are brittle in the face of distribution shifts (DS), where the target distribution that the system is tested on differs from the source distribution used to train the system. This problem of robustness to DS has been studied extensively in the field of domain adaptation. For deep neural networks, a popular framework for unsupervised domain adaptation (UDA) is domain matching, in which algorithms try to align the marginal distributions in the feature or output space. The current theoretical understanding of these methods, however, is limited and existing theoretical results are not precise enough to characterize their performance in practice. In this paper, we derive new bounds based on optimal transport that analyze the UDA problem. Our new bounds include a term which we dub as *entanglement*, consisting of an expectation of Wasserstein distance between conditionals with respect to changing data distributions. Analysis of the entanglement term provides a novel perspective on the unoptimizable aspects of UDA. In various experiments with multiple models across several DS scenarios, we show that this term can be used to explain the varying performance of UDA algorithms.

1 INTRODUCTION

Deep learning has revolutionized machine learning and AI in the past decade and provided impressive solutions in computer vision, speech recognition and natural language processing, among others (Bengio et al., 2017).

However one of the reasons preventing their wider deployment is the trustworthiness and robustness issues surrounding deep neural networks (DNN). DNNs are often not robust to distribution shifts that occur in many critical environments, such as those encountered in robotics, healthcare and computer vision, to name a few.

The generalization guarantees of standard supervised learning, typically used to train DNNs, require the training and test data to come from the same distribution (Shalev-Shwartz and Ben-David, 2014). The performance of DNNs is often brittle whenever this assumption is violated, however, and degrades quickly in the face of even mild distribution shifts at times. Domain adaptation studies this robustness to distribution shift, in cases where the target distribution, from which unlabeled samples are available, is different from the source distribution that the model is trained on.

In this paper, we are interested in target risk upper bounds for *unsupervised* domain adaptation (UDA), in which there is no labeled data available from the target domain. In particular, we introduce novel upper bounds based on optimal transport (Villani, 2009; Peyré et al., 2019), and study them from the optimization perspective. Based on the unoptimizable part of the upper bound (which requires target labels), we introduce the notion of *entanglement*[1] which quantifies the central difficulty of UDA from the optimization perspective. Unlike previous bounds (Ben-David et al., 2010; Courty et al., 2017), the entanglement term makes explicit and quantifies the *hardness* of UDA: algorithms that minimize marginal distance (or divergence) can entangle different class-conditionals while

---

[1]Not to be confused with the notion of disentangled representations (Bengio et al., 2013) in representation learning research.
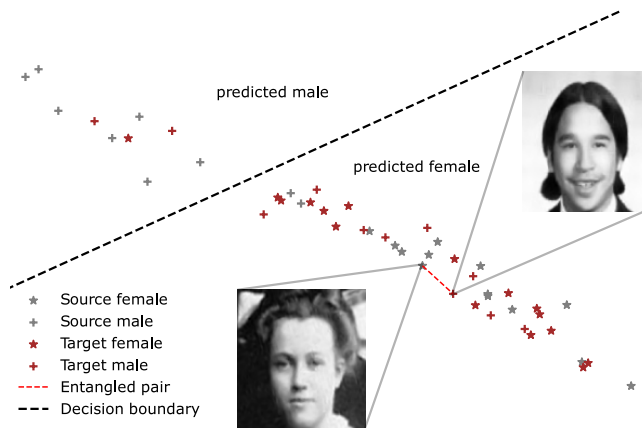
Figure 1: We derive new bounds for unsupervised domain adaptation based on optimal transport and introduce a term called entanglement, which quantifies the loss in accuracy that can happen when aligning marginals. During the marginal alignment step, optimal transport associates source inputs (gray) to target inputs (brown) and stochastic gradient descent tries to find neural network parameters that pull these coupled points closer together. Entanglement measures the average loss of associating pairs with *different* labels. In this figure we visualize such an entangled (source, target) output pair using the Portraits dataset (Ginosar et al., 2015). The labeled images of girls and boys until the year 1974 (first half of the dataset) are used as the source dataset and the remaining half of the original dataset (from 1974 till 2013) is used as the target. The dotted line indicates the decision boundary in the output space of a convolutional neural network separating male predictions from females.

pulling unlabeled samples closer together, making it hard (and often impossible) to find a good classifier that has high target accuracy. Previous work using the accuracy of the best model in source and target domains to bound the target accuracy hides this difficulty because this quantity can grow as the model is optimized (Johansson et al., 2019). Experiments confirm that the entanglement term is a suitable tool to analyze the UDA performance of many algorithms: depending on the (model, dataset, optimizer) triplet this term can grow during the optimization and cause target accuracy to degrade. Throughout the paper, by making various assumptions, such as on the loss function or the model capacity, we study the entanglement term and the resulting UDA problem in detail. Extensive experiments with different models, datasets and optimizers corroborate our conclusions.

2 RELATED WORK

Domain adaptation (DA) studies ways to transfer knowledge between similar *domains*, which consist of a joint input-label space and a probability distribution over this space. To categorize very briefly the algorithms proposed in the literature (see, e.g., Wilson and Cook (2020) for a survey): the earliest algorithms considered using *importance weighting* (IW) to adapt misspecified source models to the target distribution (Shimodaira, 2000; Huang et al., 2006). Under

the covariate shift and the equal support assumptions, one can show that importance weighting can be used to learn consistent models (Sugiyama and Kawanabe, 2012). *Domain matching* methods (Ganin et al., 2016; Damodaran et al., 2018; Shen et al., 2018; Tachet des Combes et al., 2020), try to find representations that minimize the difference between the source and the target distributions, quantified usually by a divergence term between the two. *Pseudolabeling*, widely used in the semi-supervised learning literature (Lee et al., 2013), can also be used to perform domain adaptation (Shu et al., 2018), although without any guarantees, unless other assumptions are made (Kumar et al., 2020).

On the theoretical side, the seminal paper of Ben-David et al. (2010) studies conditions under which a source-classifier does well on a related, but different target distribution. Under the condition that there exists a classifier that does well (i.e., has high accuracy) on both source and target domains, the knowledge transfer from source to target is then quantified by an $\mathcal{H}\Delta\mathcal{H}$ divergence that can be thought of informally as the projection of the total variation distance to the hypothesis space $\mathcal{H}$. Inspired by the theory, adversarial domain matching approaches (Ganin et al., 2016) try to find representations that minimize this divergence together with the source risk minimization. Despite the popularity of adversarial losses, such approaches lack rigorous guarantees, due to the lack of invertibil-

ity (Johansson et al., 2019) of the models or more generally the lack of invariant features (Zhao et al., 2019; Arjovsky et al., 2019).

A research direction in DA that is closely aligned with our work consists of changing the $\mathcal{H}\Delta\mathcal{H}$ divergence to various other dissimilarity functions. Some common dissimilarity functions considered are the KL-divergence (Nguyen et al., 2022) and – most relevant to this work – Wasserstein distances (Redko et al., 2017; Courty et al., 2017). The central property that these dissimilarity functions provide is the change of measure inequality, which allows the target risk to be upper bounded by a source risk plus the corresponding dissimilarity measure (Donsker and Varadhan, 1983). These bounds inspired several algorithms to optimize deep neural networks (Damodaran et al., 2018; Shen et al., 2018). We discuss the proposed algorithms and related work in more detail in Appendix C.

Notation. To set the notation for what is to follow: we are given input-label samples (x, y) from measurable metric spaces $(\mathcal{X}, d_x)$ and $(\mathcal{Y}, d_y)$. For a sample space $\mathcal{Z}$, we denote the set of probability measures as $\triangle(\mathcal{Z})$. We typically denote joint distributions as $p, q \in \triangle(\mathcal{X} \times \mathcal{Y})$. Their corresponding marginal distributions are denoted via a subscript, i.e., $p_x \in \triangle(\mathcal{X})$ or $q_y \in \triangle(\mathcal{Y})$. With slight abuse of notation, we will overload $\mathcal{Y}$ for the corresponding simplex $\triangle(\mathcal{Y})$ — which will become useful due to classifiers commonly outputting softmax probabilities. We denote the pushforward measure of a measurable function $f \colon \mathcal{X} \to \mathcal{Y}$ and a probability measure $\mu \in \triangle(\mathcal{X})$ by $f\sharp\mu$. With some abuse of notation, for a joint distribution $p \in \triangle(\mathcal{X}\times\mathcal{Y})$ we denote the pushforward $f\sharp p \doteq T_{(f,\mathrm{id})}\sharp p$, where $T_{(f,\mathrm{id})}$ is the map $(x, y) \mapsto (f(x), y)$.

3 ENTANGLEMENT IN DOMAIN MATCHING

In DA, we are interested in finding a model $f \colon \mathcal{X} \to \mathcal{Y}$ from a hypothesis class $\mathcal{H}$ which transfers well to a target domain: the model should not only achieve high accuracy on the source $p \in \triangle(\mathcal{X}\times\mathcal{Y})$ but *also* maintain similar levels of accuracy on a target distribution $q \in \triangle(\mathcal{X}\times\mathcal{Y})$. We consider the setting where we only have access to full samples from the source $(x, y) \sim p$, but only have unlabeled data from the target, i.e., y' is not observed in samples $(x', y') \sim q$. The performance on each of these distributions can be quantified via their corresponding risk. Given a loss function $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we define the source and target risks, respectively, as

$$\epsilon_p(f) \doteq \mathbb{E}_{(x,y)\sim p}[\ell(f(x), y)]; \; \epsilon_q(f) \doteq \mathbb{E}_{(x',y')\sim q}[\ell(f(x'), y')].$$

Motivated with the goal of keeping the target risk $\epsilon_q(f)$ small, in this section we provide novel bounds that connect target risk minimization to source risk minimization through optimal transport.

Optimal Transport. Amongst the set of distortion measures utilized in machine learning, optimal transport distances (Peyré et al., 2019; Villani, 2009) (using an appropriately chosen cost function) have been an appealing choice for comparing probability measures since (i) they satisfy the requirements of a metric between probability measures and (ii) they can be used between any two distributions μ, ν regardless of their support (unlike f-divergences (Ali and Silvey, 1966; Csiszár, 1967)).

Definition 3.1. Let $c \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a (non-negative) cost function. The α-Wasserstein distance w.r.t. c between measures $\mu \in \triangle(\mathcal{X})$ and $\nu \in \triangle(\mathcal{Y})$ is given by,

$$W_{\alpha,c}(\mu, \nu) \doteq \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)^\alpha \mathrm{d}\gamma(x, y) \right)^{1/\alpha},$$

where $\Gamma(\mu, \nu) \doteq \{\gamma \in \triangle(\mathcal{X} \times \mathcal{Y}) : \pi_1\sharp\gamma = \mu, \pi_2\sharp\gamma = \nu\}$, and $\pi_1(x, y) = x$, $\pi_2(x, y) = y$ are projection functions for the corresponding marginal constraints.

Typically, the cost function of a Wasserstein distance c is implicit in the sample space of $\mathcal{X}$ and $\mathcal{Y}$, i.e., when $\mathcal{X} = \mathcal{Y}$ is a metric space, one uses the underlying metric d_x. In our work, we will make the choice of the cost function explicit. In particular, we are interested in Wasserstein distances over probability measures with *joint sample spaces* when the cost function *decomposes* over the sample space's Cartesian product $\mathcal{X} \times \mathcal{Y}$.

Definition 3.2. We say that a cost function $c \colon (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ is decomposable if there exist cost functions $c_1 \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $c_2 \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ such that $c((x, y), (x', y')) = c_1(x, x') + c_2(y, y')$. We further denote the corresponding α-Wasserstein distance for decomposable costs by $W_{\alpha,c_1,c_2} = W_{\alpha,c}$.

It can be immediately seen that if a cost function decomposes into two metric cost functions c_1, c_2, then the original cost function must also be a metric cost function. This provides a simple method for defining a decomposable cost by the addition of metric functions $c_1 \colon (\mathcal{X} \times \mathcal{X}) \to \mathbb{R}$ and $c_2 \colon (\mathcal{Y} \times \mathcal{Y}) \to \mathbb{R}$.

3.1 Loss Function and Model Assumptions

In this section, we will establish bounds on the target risk $\epsilon_q(f)$ via Wasserstein distances of decomposable cost functions. In particular, the decomposable cost we will use will be a function of the loss function $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Before establishing our bounds, we first state the assumptions needed on the loss function ℓ and classifier f.

Assumption 3.1 (Metric Loss). The loss function $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is symmetric, non-negative (hence $\ell(x, y) \geq 0$ for all x, y), and respects the triangle inequality. In addition, the identity of indiscernibles, i.e., $\ell(y, y') = 0 \iff y = y'$ holds.

Assumption 3.1 is commonly made in the DA literature, in particular, the symmetry and triangle inequality properties of the loss function are often utilized to factor out terms involving the best model $f^\star$ (minimizing the source and target risks jointly) (Ben-David et al., 2010; Courty et al., 2017). It should be noted that although symmetry can be accounted for via symmetrizing the loss function – e.g., $\tilde{\ell}(y, y') = \frac{1}{2}(\ell(y, y') + \ell(y', y))$ – not all symmetric loss functions will satisfy the triangle inequality. Despite this, Assumption 3.1 can be replaced with an *approximate triangle inequality* (Crammer et al., 2008) with slight alteration to our results. We discuss the extension of our results to non-metric loss functions in Appendix D in detail.

Assumption 3.2 (Bounded Loss). The loss function $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is bounded such that for all $y, y' \in \mathcal{Y}$, $\ell(y, y') \leq L$ for $L > 0$.

When $\mathcal{Y}$ consists of a finite set $|\mathcal{Y}| < \infty$ and ℓ satisfies the identity of indiscernibles, then the loss function satisfies also a positive *lower* bound. Assumption 3.2 and the lower bound correspond to Lipschitz conditions w.r.t. a discrete topology of $\mathcal{Y}$. We require the Lipschitz conditions only in Section 4 to analyze our bounds in the presence of label shift.

In addition to loss function assumptions, we also consider the following assumption for the models $f \in \mathcal{H}$.

Assumption 3.3 (Surjective Models). The model $f\colon \mathcal{X} \to \mathcal{Y}$ is a surjective function.

The surjectivity simply states that model f will map an input to all possible outputs. It can be relaxed to hold only for some non-empty subset of $\mathcal{Y}$, however the requirement should hold uniformly over the hypothesis class of f whenever the bounds are minimized.

3.2 Bounds on Target Risk

We start with a simple, yet effective bound using suitable Wasserstein distances that use a decomposable cost function based on the loss function ℓ. To simplify notation, we will denote $(\ell \circ f)(x, x') = \ell(f(x), f(x'))$.

Lemma 3.1. *Suppose the loss function* $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ *satisfies Assumption 3.1. Then the* target *risk of a classifier* $f\colon \mathcal{X} \to \mathcal{Y}$ *is bounded by*

$$\epsilon_q(f) \leq \epsilon_p(f) + W_{1, \ell \circ f, \ell}(p, q) \tag{1}$$

$$\overset{(s)}{\leq} \epsilon_p(f) + W_{1, \ell, \ell}(f \sharp p, f \sharp q), \tag{2}$$

where the inequality (s) *holds for surjective* f. *Additionally,* (s) *is an* equality *whenever* f *is invertible. We remind the readers that* $(f \sharp p)(\hat{y}, y) = p(f^{-1}(\hat{y}), y)$.

Proof Sketch. The risk difference $\epsilon_q(f) - \epsilon_p(f)$ can be related to any coupling that integrates over the corresponding loss differences. The optimal transport cost is the lowest cost among such couplings and applying triangle inequality to its integrand yields the Wasserstein distance above. Surjectivity allows us to relate the optimal transport between p and q to that between $f \sharp p$ and $f \sharp q$. $\square$

It should be noted that the 1-Wasserstein distance in Lemma 3.1 can be easily replaced with a corresponding α-Wasserstein distance for $1 \leq \alpha < \infty$ due to the ordering of α-norms, i.e., $W_{\alpha,c} \leq W_{\beta,c}$ for any $\alpha < \beta$. This also indicates that out of the typical ranges of α-Wasserstein distances considered, the presentation of Lemma 3.1 taking $\alpha = 1$ is the most tight.

The second term in (1) is the Wasserstein distance between the *joint* distributions p and q when using the aggregate loss $\ell(f(x), f(x')) + \ell(y, y')$ between prediction-label pairs $(f(x), y)$ and $(f(x'), y')$. Since in UDA the labels y' from the target distribution are unavailable, the bound above is not immediately useful in our case. Loosening the bound is inevitable in order to derive practical optimizers, hence we next relate the bound to the marginal Wasserstein distance $W_{1, \ell \circ f}(p_\mathrm{x}, q_\mathrm{x})$ which was used in several DA works (Courty et al., 2017; Shen et al., 2018; Redko et al., 2017).

In order to relate the *joint* Wasserstein distance to Wasserstein distances over marginal distributions, we consider the following lemma that exploits the structure of decomposable joint costs.

Lemma 3.2. *Let* $c\colon (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ *be a cost function that decomposes into functions* $c_1\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *and* $c_2\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. *Then for* $1 \leq \alpha < \infty$,

$$W_{\alpha, c_1}(\mu_\mathrm{x}, \nu_\mathrm{x}) \leq W_{\alpha, c}(\mu, \nu) \leq W_{\alpha, c_1}(\mu_\mathrm{x}, \nu_\mathrm{x}) +$$
$$\underset{(x, x') \sim \gamma_x^\star}{\mathbb{E}} \left[W_{\alpha, c_2}^\alpha(\mu_{\mathrm{y}|\mathrm{x}}(\cdot \mid x), \nu_{\mathrm{y}|\mathrm{x}}(\cdot \mid x')) \right]^{1/\alpha}, \tag{3}$$

$$W_{\alpha, c_2}(\mu_\mathrm{y}, \nu_\mathrm{y}) \leq W_{\alpha, c}(\mu, \nu) \leq W_{\alpha, c_2}(\mu_\mathrm{y}, \nu_\mathrm{y}) +$$
$$\underset{(y, y') \sim \gamma_y^\star}{\mathbb{E}} \left[W_{\alpha, c_1}^\alpha(\mu_{\mathrm{x}|\mathrm{y}}(\cdot \mid y), \nu_{\mathrm{x}|\mathrm{y}}(\cdot \mid y')) \right]^{1/\alpha}, \tag{4}$$

where $\gamma_x^\star$ *and* $\gamma_y^\star$ *are optimal couplings corresponding to* $W_{\alpha, c_1}(\mu_x, \nu_x)$ *and* $W_{\alpha, c_2}(\mu_y, \nu_y)$, *respectively.*

Proof Sketch. To prove the lower bounds, we relate the optimal coupling between the joint distributions to the optimal coupling between the marginals by a transformation that disappears once we lower bound

the joint cost c with the marginal costs. The proof for the upper bounds hinges on the fact that transports that decompose the optimal coupling between $(x, y) \sim \mu$, $(x', y') \sim \nu$ pairs are sub-optimal: the couplings $\gamma(x, x')\gamma(y, y'|x, x')$ and $\gamma(y, y')\gamma(x, x'|y, y')$ incur a cost that upper bounds the optimal transport cost. □

For the rest of the paper, we will need both directions of the bound[2] for further analysis. In particular, we will use the first bound (3) to analyze DA *algorithmically*, while the second bound (4) will be a useful theoretical tool for that purpose.

3.3 Entanglement

Next, we establish a corollary that links the two previous lemmas. The corollary introduces an *entanglement* term between the source and target distributions, which fundamentally results from aligning marginal distributions during UDA without having access to target label information. First, we define two closely related entanglement terms.

Definition 3.3. For a model $f \colon \mathcal{X} \to \mathcal{Y}$, the label entanglement $\mathcal{E}_{\mathrm{y}}(f)$ and the prediction entanglement $\mathcal{E}_{\hat{\mathrm{y}}}(f)$ are defined as

$$\mathcal{E}_{\mathrm{y}}(f) \doteq \int_{\mathcal{Y} \times \mathcal{Y}} W_{1,\ell}(p_{\mathrm{y}|\mathrm{f}}(\cdot \mid \hat{y}), q_{\mathrm{y}|\mathrm{f}}(\cdot \mid \hat{y}'))\mathrm{d}\gamma_{\mathrm{f}}^{\star}(\hat{y}, \hat{y}'),$$

$$\mathcal{E}_{\hat{\mathrm{y}}}(f) \doteq \int_{\mathcal{Y} \times \mathcal{Y}} W_{1,\ell}(p_{\mathrm{f}|\mathrm{y}}(\cdot \mid y), q_{\mathrm{f}|\mathrm{y}}(\cdot \mid y'))\mathrm{d}\gamma_{\mathrm{y}}^{\star}(y, y'),$$

where $p_{\mathrm{y}|\mathrm{f}}(y \mid \hat{y})$ and $q_{\mathrm{y}|\mathrm{f}}(y \mid \hat{y}')$ correspond to the joint distributions p, q conditioned on $\hat{y} = f(x)$ and $\hat{y}' = f(x')$, while $p_{\mathrm{f}|\mathrm{y}}$ and $q_{\mathrm{f}|\mathrm{y}}$ denote the pushforwards of $p_{\mathrm{x}|\mathrm{y}}$ and $q_{\mathrm{x}|\mathrm{y}}$ conditioned on y and y', respectively: $f\sharp p_{\mathrm{x}|\mathrm{y}}(x \mid y)$, $f\sharp q_{\mathrm{x}|\mathrm{y}}(x \mid y')$. The couplings $\gamma_{\mathrm{f}}^{\star}$ and $\gamma_{y}^{\star}$ are optimal transport plans between pairs $f\sharp p_{\mathrm{x}}$, $f\sharp q_{\mathrm{x}}$ and p_y, q_y.

By combining Lemmas 3.1 and 3.2, letting $\alpha = 1$ and defining a decomposable cost function with $c_1 = c_2 = \ell$, we get the following corollary.

Corollary 3.1. *Suppose the loss function* $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ *satisfies Assumption 3.1 and a classifier* $f \colon \mathcal{X} \to \mathcal{Y}$ *satisfies Assumption 3.3 then*

$$\epsilon_q(f) \leq \epsilon_p(f) + W_{1,\ell}(f\sharp p_{\mathrm{x}}, f\sharp q_{\mathrm{x}}) + \mathcal{E}_{\mathrm{y}}(f), \quad (5)$$

$$\epsilon_q(f) \leq \epsilon_p(f) + W_{1,\ell}(p_{\mathrm{y}}, q_{\mathrm{y}}) + \mathcal{E}_{\hat{\mathrm{y}}}(f). \quad (6)$$

From the corollary we can see clearly the implications of aligning marginals during UDA. For instance, in the case of an optimal transport based alignment, learning a hypothesis f that brings the set of pairs $f(x)$ and $f(x')$ closer (thus minimizing the Wasserstein distance between $f\sharp p_{\mathrm{x}}$ and $f\sharp q_{\mathrm{x}}$) does not ensure that the *average* distance (with respect to the optimal coupling) between the conditionals $f\sharp p_{\mathrm{y}|\mathrm{x}}$ and $f\sharp q_{\mathrm{y}|\mathrm{x}}$ will also decrease. We refer to this potential increase in average conditional distance as *entanglement*[3].

3.4 Oracle Bound

As discussed above, during the optimal transport association process, the source inputs x can be associated to target inputs x' that have different matching labels. Minimizing the marginal Wasserstein distance between such *entangled* pairs (without including any entanglement terms from Definition 3.3) can cause the entanglement terms $\mathcal{E}_{\mathrm{y}}$ and $\mathcal{E}_{\hat{\mathrm{y}}}$ to increase, potentially increasing the target risk as well. In order to better understand the roles that these two terms play, we will next relate them to each other. For this purpose, we first introduce an *oracle* that we assume has access to the label entanglement $\mathcal{E}_{\mathrm{y}}$.

Definition 3.4. We denote the *oracle upper bound* (OUB) as

$$\mathcal{U}(f) \doteq \epsilon_p(f) + W_{1,\ell}(f\sharp p_{\mathrm{x}}, f\sharp q_{\mathrm{x}}) + \mathcal{E}_{\mathrm{y}}(f). \quad (7)$$

Although OUB includes only the label entanglement $\mathcal{E}_{\mathrm{y}}(f)$, we note that both of the entanglement terms are intrinsically connected.

Theorem 3.1. *Suppose Assumptions 3.1 and 3.3 hold. Then we have for all* $f \in \mathcal{H}$,

$$\mathcal{E}_{\mathrm{y}}(f) \leq \mathcal{E}_{\hat{\mathrm{y}}}(f) + \epsilon_p(f) + \epsilon_q(f) + W_{1,\ell}(p_{\mathrm{y}}, q_{\mathrm{y}}),$$

$$\mathcal{E}_{\hat{\mathrm{y}}}(f) \leq \mathcal{E}_{\mathrm{y}}(f) + \epsilon_p(f) + \epsilon_q(f) + W_{1,\ell}(f\sharp p_{\mathrm{x}}, f\sharp q_{\mathrm{x}}).$$

Proof Sketch. The proof is a result of applying triangle inequality to the loss function used to compute the entanglement terms, and then noting that each term corresponds to the source risk, target risk and the marginal Wasserstein distance respectively. The marginal Wasserstein distances can further be upper bounded by switching between the lower and the upper bounds in Lemma 3.2. □

As a consequence, minimization of the bounds (5) and (6) is also connected together.

---

[2]See Appendix E where we explore the bounds in the case of Gaussians.

[3]Note the relation to the notion of spurious features in the out-of-domain generalization (OOD) literature (Arjovsky et al., 2019): entanglement is the corresponding notion for optimal transport.

Corollary 3.2. *Suppose Assumptions 3.1 and 3.3 hold. Then we have for all $f \in \mathcal{H}$,*

$$\epsilon_p(f) + W_{1,\ell}(p_{\mathrm{y}}, q_{\mathrm{y}}) + \mathcal{E}_{\hat{\mathrm{y}}}(f) \in \left[\frac{1}{3} \cdot \mathcal{U}(f), 3 \cdot \mathcal{U}(f)\right].$$

A consequence of Corollary 3.2 is that minimizing either of the bounds (5) or (6) ($\mathcal{U}(f)$ corresponding to (5)) results in the minimization of the other.

4   ANALYSIS OF ENTANGLEMENT

We now study the bounds in Corollary 3.1 carefully and discuss how we can use them to analyze UDA and the unoptimizable entanglement term. In particular, we will focus on (a) assumptions necessary for further analysis, (b) effect of label shift, and (c) bounding the effect of the unoptimizable (without target labels) entanglement term $\mathcal{E}_{\mathrm{y}}(f)$ appearing in the oracle bound $\mathcal{U}(f)$.

4.1   Assumptions

In domain adaptation, a key quantity used to analyse the limit performance of a set of hypotheses $\mathcal{H}$ is the *ideal joint error* (Ben-David et al., 2010). In this paper, we will refer to it as *Low Joint Error*.

Assumption 4.1. The *Low Joint Error* assumption LJE(λ) holds for $\lambda > 0$ and a hypothesis class $\mathcal{H}$ if $\min_{f \in \mathcal{H}} \{\epsilon_p(f) + \epsilon_q(f)\} < \lambda$.

The LJE(λ) assumption states the existence of a classifier $f^{\star}_{\mathrm{LJE}} \in \mathcal{H}$ which has low risk on both the source and the target distributions. Without this assumption, we cannot hope to find a hypothesis with low source risk that transfers well to the target distribution (which is the goal of UDA as stated above).

We consider an alternative assumption which helps us control the label entanglement of a hypothesis set.

Assumption 4.2. The *Close Conditionals Assumption* CC(κ) holds for $\kappa > 0$ and a hypothesis class $\mathcal{H}$ if there exists a classifier $f \in \mathcal{H}$ such that $\epsilon_p(f) + \max_{y \in \mathcal{Y}} W_{1,\ell}(f \sharp p_{\mathrm{x}|\mathrm{y}}(\cdot \mid y), f \sharp q_{\mathrm{x}|\mathrm{y}}(\cdot \mid y)) < \kappa$.

Instead of requiring low risk on both the source and target distributions, CC(κ) requires a low source risk hypothesis $f^{\star}_{\mathrm{CC}} \in \mathcal{H}$ which *pushes* the conditionals close to each other: i.e, the assumption ensures that the pushforwards of conditionals $f^{\star}_{\mathrm{CC}} \sharp p_{\mathrm{x}|\mathrm{y}}(\cdot \mid y)$ and $f^{\star}_{\mathrm{CC}} \sharp q_{\mathrm{x}|\mathrm{y}}(\cdot \mid y)$ do not differ too much when conditioned on the *same* label y. CC is generally a valid assumption for DNNs with large enough capacity. We discuss this property of DNNs in Appendix E in detail.

One property of the CC(κ) assumption is that its violation (for a large enough $\kappa > 0$) precludes certain hypotheses from achieving good target risk.

Lemma 4.1. *Suppose Assumption 3.1 holds. If CC(κ) does not hold, then for all $f \in \mathcal{H}$,*

$$\kappa \cdot q_{\mathrm{y}}(y_{\min}) \leq \epsilon_p(f) \cdot q_{\mathrm{y}}(y_{\max}) + \epsilon_q(f) + \epsilon_r(f),$$

where $r(x, y) = p_{\mathrm{x}|\mathrm{y}}(x \mid y) \cdot q_{\mathrm{y}}(y) \in \triangle(\mathcal{X} \times \mathcal{Y})$ and $y_{\min} = \arg\min q_{\mathrm{y}}(y)$, $y_{\max} = \arg\max q_{\mathrm{y}}(y)$.

From this lemma we can infer that whenever CC(κ) is violated (i.e., κ is large) while LJE(λ) holds for some small λ, then either there is a large imbalance in the target label probabilities q_{y} or there must be a large discrepancy in the components of source risk for different labels. The latter case highlights the fact that in addition to low LJE(λ), we may also need a *balanced* (among minority classes) source risk for CC(κ) to hold for low κ. Consider the following example to illustrate this point further.

Example.   Consider the setting of binary classification where $\mathcal{Y} = \{0, 1\}$ and $q_{\mathrm{y}}(y) = 1/2$ for all $y \in \mathcal{Y}$. Now suppose that LJE(λ) holds and CC(κ) is violated for λ, κ satisfying $C \leq \kappa/2 - \lambda$ for some large $C > 0$. In this case, for any model achieving the LJE assumption $f^{\star}_{\mathrm{LJE}} \in \mathcal{H}$, Lemma 4.1 can be weakened to $C \leq \epsilon_r(f^{\star}_{\mathrm{LJE}})$. From LJE($\lambda$) we have $\epsilon_p(f^{\star}_{\mathrm{LJE}}) < \lambda$. Thus, for this to be true, we require $\mathbb{E}_p[\ell(f^{\star}_{\mathrm{LJE}}(x), y) \mid y = 0]$ to differ significantly from $\mathbb{E}_p[\ell(f^{\star}_{\mathrm{LJE}}(x), y) \mid y = 1]$. In particular, the performance on the minority label $y_{\min} = \arg\min_y p_{\mathrm{y}}(y)$ on the source will be poor, i.e., $\mathbb{E}_p[\ell(f^{\star}_{\mathrm{LJE}}(x), y) \mid y = y_{\min}]$ large.

4.2   Effect of Label Shift

In the following, we explore the effect of label shift between the source and target distributions p_{y} and q_{y}.

Lemma 4.2. *Suppose that Assumptions 3.1 to 3.3 hold and there exists an $l > 0$ such that $\min_{y \neq y'} \ell(y, y') > l$. Then given $W_{1,\ell}(p_y, q_y) < \delta$ holds for some $\delta > 0$ and CC(κ) for some $\kappa > 0$ then LJE(λ) holds with $\lambda = 2\kappa + \frac{L+l}{l} \cdot \delta$.*

Proof Sketch. The proof is a consequence of the fact that if the source and target label distributions are close together in Wasserstein distance then the corresponding coupling is concentrated on the diagonals: off diagonal terms are small and by using upper and lower bounds of the loss, we can cap their effect on the target risk using equation (6). □

The above Lemma 4.2 shows the heavy price we might have to pay (in terms of loss) for small deviations of the label distribution $p(y)$: small deviations δ, measured in terms of Wasserstein distance, can be amplified up

to L/l times. If the model f outputs probability distributions (i.e., values in the simplex), then we can use the diameter of the simplex as measured by the metric $\ell(f, f')$ as the bound L. For the lower bound l, we note that for discrete y, y' in the simplex $\min_{y \neq y'} \ell(y, y')$ can be seen as a scaling factor in the bounds.

Further note that if the marginal label distributions in the source and target domains are the same, then the LJE and CC assumptions provide a guarantee for each other.

Corollary 4.1. *Suppose Assumptions 3.1 to 3.3 hold and $p_y = q_y$. Then $\mathrm{CC}(\kappa)$ implies $\mathrm{LJE}(2\kappa)$ and $\mathrm{LJE}(\lambda)$ implies $\mathrm{CC}(\lambda \cdot (1 + q_y(y_{\max}))/q_y(y_{\min}))$.*

4.3 Optimization Criteria

As the entanglement computation requires target labels, we also define the objective function without the entanglement term below, as a regularized source risk, which is the quantity that is usually minimized in practice (Courty et al., 2017; Shen et al., 2018).

Definition 4.1. We denote the Wasserstein Regularized Risk (WRR) as

$$\mathcal{R}(f) \doteq \epsilon_p(f) + W_{1,\ell}(f \sharp p_x, f \sharp q_x). \qquad (8)$$

Based on the bound in (5) we analyze below the properties of the oracle entanglement upper bound $\mathcal{U}(f)$ (3.4) and the corresponding $\mathcal{R}(f)$ (4.1). We first start with a lower bound on $\mathcal{U}(f)$.

Lemma 4.3. *Suppose Assumptions 3.1 to 3.3 hold. Then $W_{1,\ell}(p_y, q_y) \leq \mathcal{U}(f)$.*

Proof. The proof follows from using the second lower bound and then the first upper bound in Lemma 3.2. $\square$

Lemma 4.3 implies that although there may be a hypothesis f^*_{CC} satisfying $\mathrm{CC}(\kappa)$ for a small $\kappa > 0$, if there is a shift in the label distributions, an algorithm minimizing $\mathcal{U}(f)$ is not able to recover such a hypothesis[4]. We now have what we need to analyze the tightness of the bounds in (5). We will show that $\mathrm{CC}(\kappa)$ and low label shift implies that the bound in (5) is tight (up to small multiples of κ and potentially large multiples of δ) at the best hypothesis $f^\star_{\mathrm{CC}}$.

Lemma 4.4. *Suppose that Assumptions 3.1 to 3.3 hold, $W_{1,\ell}(p_y, q_y) < \delta$, and there exists an $l > 0$ such that $\min_{y \neq y'} \ell(y, y') > l$. If $f^\star_{\mathrm{CC}} \in \mathcal{H}$ witnesses $\mathrm{CC}(\kappa)$, then*

$$\mathcal{U}(f^\star_{\mathrm{CC}}) \leq 3 \cdot \left(\kappa + \frac{L+l}{l} \cdot \delta \right).$$

---

[4]Note that similar lower bounds were presented in Zhao et al. (2019) using Jensen-Shannon divergence.

The above lemma implies that we can minimize $\mathcal{U}(f)$ to obtain small target risk if the label shift is small, assuming that we can find its minimizer. However, although the bound can be *tight* at the best hypothesis $f^\star_{\mathrm{CC}}$ satisfying $\mathrm{CC}(\kappa)$, the entanglement term may not be small at the minimizer of $\mathcal{R}(f)$. We next look at further assumptions we can make to force this term to be smaller.

Assumption 4.3 (Gradual Shift Assumption). The Gradual Shift assumption $\mathrm{GS}(a, b, \varepsilon, s)$ holds for a classifier $f \in \mathcal{H}$ if $\epsilon_p(f) < b$ implies that there exists $q^{(i)}_{x|y}(\cdot \mid y) \in \triangle(\mathcal{X})$ such that

$$q_{x|y}(x \mid y) = \sum_{i=1}^{s} r_i q^{(i)}_{x|y}(x \mid y)$$

for $0 \leq r_i \leq a$ and all $y \in \mathcal{Y}$; and for all $i \in \{1, \ldots, s\}$ we have $W_{1,\ell}(f \sharp q^{(i-1)}_{x|y}(\cdot \mid y), f \sharp q^{(i)}_{x|y}(\cdot \mid y)) < \varepsilon$, where $q^{(0)}_{x|y}(x \mid y) \doteq p_{x|y}(x \mid y)$.

Note that this definition is different from the usual gradual shift defined in the literature (Kumar et al., 2020), in that the indices of the gradually shifting distributions may not be available, and the samples from the mixture components can be scrambled. Moreover note the dependency on b: besides the smoothness assumption on the shifting conditionals, we are also implicitly assuming regularity properties on $\mathcal{H}$[5].

We first connect the GS assumption to that of CC.

Lemma 4.5. *Suppose that Assumptions 3.1 and $\mathrm{GS}(a, b, \varepsilon, s)$ hold. For a classifier $f \in \mathcal{H}$, if $\epsilon_p(f) < b$, then $\mathrm{CC}(\kappa)$ holds with $\kappa = b + \varepsilon \cdot \frac{a}{2} s(s + 1)$.*

Proof Sketch. Note that the GS assumption holds uniformly for each $y \in \mathcal{Y}$. The proof follows directly from the convexity of Wasserstein distances (with respect to their arguments) and the triangle inequality applied to the Wasserstein distances between the conditional source distributions $p_{x|y}$ and conditional target distributions $q_{x|y^{(i)}}$, for each $i = 1, \ldots, s$. $\square$

Given Lemma 4.5, we can guarantee the CC assumption whenever the GS assumption holds. Although the quadratic dependence on s seems strong, one can examine certain scenarios where it is weakened. First note that $\varepsilon' = s\varepsilon$ intuitively can be seen as the full aggregated distance of the source to target distributions. Thus fixing ε', increasing s only provides a linear increase. Secondly, the mixture constraint is bounded

---

[5]For example, if all hypotheses in $\mathcal{H}$ require the use of certain features to achieve low risk b, then we can relate the assumption further to the utilization of such *stable* features (Arjovsky et al., 2019).

as $a \in [1/s, 1]$. In the best case scenario, we have a balanced mixture $a = 1/s$. Then the dependence can be simplified as $\mathcal{O}(b + a\varepsilon' s) = \mathcal{O}(b + \varepsilon')$.

Finally we prove that the GS assumption entails slowly growing (depending on s) label entanglement for small label shift.

Theorem 4.1. *Suppose that Assumptions 3.1 to 3.3 and* $GS(a, b, \varepsilon, s)$ *hold;* $W_{1,l}(p_y, q_y) < \delta$, *and there exists an* $l > 0$ *such that* $\min_{y \neq y'} \ell(y, y') > l$. *For a classifier* $f \in \mathcal{H}$, *if* $\epsilon_p(f) < b$, *then*

$$\mathcal{E}_{\mathrm{y}}(f) \leq 2b + \varepsilon \cdot as(s+1) + 2\delta \cdot \frac{L+l}{l}.$$

Proof Sketch. We first bound the prediction entanglement $\mathcal{E}_{\hat{\mathrm{y}}}(f)$ using Lemma 4.5 and δ, following the same steps used to prove Lemma 4.2. The label entanglement $\mathcal{E}_{\mathrm{y}}(f)$ can then be bounded using Theorem 3.1. □

Theorem 4.1 states that under the previous assumptions, the label entanglement of a good source hypothesis ($\epsilon_p(f) < b$ for some small b) can be controlled. An immediate observation is that when $\delta = 0$, the label entanglement $\mathcal{E}_{\mathrm{y}}(f)$ is upper bounded with the same dependencies as CCA, as per Lemma 4.5.

5 EXPERIMENTS

In this section we experiment with several different models in different domain shift *scenarios*, and show the wide variety of behaviours we can observe from different (model, optimizer, dataset) triplets: domain matching algorithms optimizing only for marginal alignment are shown to be susceptible to entanglement, and this vulnerability cannot be prevented purely through algorithmic means (such as hyperparameter tuning). Moreover we evaluate the oracles LJE and CC to show the limit of what we can expect from the algorithms: if LJE holds for a small λ but $CC(\kappa)$ does not, then we show that they do not perform well in general[6].

We report in Table 1 mean source and target accuracies of various oracle and non-oracle UDA algorithms after 10 epochs, evaluated over three different initializations. All methods in Table 1 use crossentropy as the loss function, the Adam optimizer with a learning rate of 0.001 and a batchsize of 64. We test with five different models of various depths, using fully-connected and convolutional networks. See Appendix F for detailed explanations of the models,

---

scenarios and the algorithms used. Different ablations of various parameters and settings are conducted in Appendix G, including batchsize variations and other optimizer configurations. We also consider in that section a more difficult dataset together with a bigger model, and add two more UDA algorithms.

5.1 Results and Discussion

Experiments in Table 1 show overall that the vanilla Wasserstein distance regularized source risk minimization (WRR) in (4.1) that we analyze in this paper is not worse off than other methods in general. We see that it has positive transfer (i.e., the target accuracy is higher than that of ERM) using the best model (top table) in the first four scenarios and for the other models tested (bottom table) in the first three scenarios. More importantly, when minimizing $\mathcal{R}(f)$, experiments with different models show very different target accuracies resulting from various degrees of entanglement[7]. Depending on the model, even when the LJE assumption is satisfied, target risk varies significantly between cases of highly positive transfer to highly negative transfer. In Table 1 we underline all the positive transfer cases and among them box the best performing method.

Oracle Minimization and Entanglement. Next we select WRR for various extended experiments in Table 2 to get a more quantitative understanding into our assumptions underlying marginal distance minimization and the resulting entanglement estimates. We use the oracles to understand better what is happening during marginal alignment: we optimize not only the LJE oracle but also the CC oracle objective functions, as well as reporting estimates of the label entanglement $\hat{\mathcal{E}}_{\mathrm{y}}(f_{\mathcal{R}})$. Unlike Table 1, we use the Euclidean distance as the loss function to be able to compare the values of each component (source risk, the marginal Wasserstein distance and the entanglement estimate). Incidentally, we see that using the Euclidean distance as the loss function outperforms using cross-entropy (this is another ablation setting which we explore further in Appendix F).

We indicate some interesting findings from Table 2. For the CIFAR-10c corruptions scenario, the CC-oracle minimization is unable to find a good hypothesis that will generalize to both environments, even though the LJE assumption is satisfied for a relatively small λ (and in spite of the fact that label shift is zero for this scenario). This potentially explains why the target accuracy of WRR is so low for this dataset. Likewise, the estimates of label shift shown in the second col-

---

[6]Scripts used to generate the experiment results are available in `https://github.com/okankoc/Domain-Adaptation-and-Entanglement`.

[7]See Appendix F for detailed accuracy plots supporting this conclusion.

Okan Koç, Alexander Soen, Chao-Kai Chiang, Masashi Sugiyama

Table 1: Performance of different algorithms including the LJE oracle objective minimization tested over various distribution shift scenarios. Each cell reports "source accuracy / target accuracy" and approaches with positive transfer (i.e., better than ERM) are notated by an underline x̲ and the best overall method is notated by a box y̲. (Top table) presents the best model *conv1* and (Bottom table) presents a selection of other model results.

Scenario + Best Model	LJE Oracle	ERM	WRR	JDOT	DANN
MNIST → USPS / conv1	97.6 / 94.7	98.0 / 83.8	96.9 / 86.0̲	98.2 / 87.1̲	96.2 / 89.3
USPS → MNIST / conv1	95.2 / 98.0	94.8 / 71.4	94.8 / 78.8̲	95.6 / 89.3	93.3 / 77.9̲
MNIST → MNIST-M / conv1	99.2 / 96.8	98.9 / 47.5	98.8 / 51.2̲	99.1 / 69.3̲	98.8 / 79.1
SVHN → MNIST / conv1	92.3 / 99.1	91.2 / 64.9	90.4 / 65.3	90.4 / 48.0	86.6 / 10.3
Portraits 1905-1974 → 1974-2013 / conv1	96.9 / 95.5	96.8 / 86.3	94.8 / 83.8	96.5 / 85.2	95.8 / 87.3
CIFAR10 → CIFAR10c / conv1	72.7 / 83.8	75.5 / 37.7	71.2 / 10.0	10.0 / 10.0	73.7 / 60.9

Scenario + Other Models	LJE Oracle	ERM	WRR	JDOT	DANN
MNIST → USPS / MLP	92.4 / 91.4	92.6 / 16.0	90.8 / 59.3	92.4 / 37.8̲	91.4 / 28.8̲
USPS → MNIST / MLP	92.7 / 94.7	91.8 / 29.0	90.3 / 52.1	92.3 / 50.5̲	91.6 / 43.1̲
MNIST → MNIST-M / LeNet	98.7 / 94.5	98.9 / 48.4	93.6 / 56.3̲	99.0 / 58.5	89.6 / 19.9
SVHN → MNIST / conv2	90.9 / 98.9	90.1 / 64.4	19.6 / 11.3	90.6 / 68.7	80.6 / 18.7
Portraits 1905-1974 → 1974-2013 / MLP	94.8 / 90.5	94.5 / 83.5	90.9 / 77.9	94.6 / 84.1	94.6 / 83.0
CIFAR10 → CIFAR10c / small_cnn	67.4 / 88.0	70.3 / 39.6	58.2 / 10.0	10.0 / 10.1	68.3 / 52.7

Table 2: Comparison of the accuracy of minimizing LJE and CC with quantities associated with the WRR objective. Cells formatted as X / Y correspond to the metric evaluated on the source / target data. $W_1(\hat{p}_y, \hat{q}_y)$ corresponds to the empirical Wasserstein distance between the marginal label distributions. The values under the "Oracle Accuracy" header correspond to the accuracy values. The values under "WRR" correspond to empirical estimates of the accuracy, risk, WRR objective and label-entanglement, respectively.

Scenario + Model	$W_1(\hat{p}_y, \hat{q}_y)$	Oracle Accuracy		WRR			
		LJE	CC	Acc.	$\hat{\epsilon}_p$ / $\hat{\epsilon}_q$	$\hat{\mathcal{R}}$	$\hat{\mathcal{E}}_y$
MNIST → USPS / conv1	0.13	97.6 / 94.7	95.3 / 93.0	98.5 / 91.7	0.06 / 0.15	0.23	0.08
MNIST → USPS / MLP	0.13	92.4 / 91.4	93.1 / 87.9	95.6 / 74.3	0.14 / 0.42	0.37	0.19
MNIST → MNIST-M / conv1	0.00	99.2 / 96.8	98.6 / 95.3	98.8 / 70.7	0.02 / 0.42	0.19	0.30
SVHN → MNIST / conv2	0.20	90.9 / 98.9	86.8 / 93.9	89.1 / 64.8	0.19 / 0.50	0.38	0.34
Portraits 1905-1974 → 1974-2013 MLP	0.05	94.8 / 90.5	91.6 / 87.3	94.0 / 79.6	0.09 / 0.29	0.11	0.33
CIFAR10 → CIFAR10c / conv1	0.00	75.2 / 87.2	63.7 / 9.9	73.9 / 9.7	0.48 / 1.13	0.77	0.58

umn correlate with the relatively poorer performances of WRR, e.g., in the SVHN-MNIST scenario where we computed a label shift estimate of $W_1(\hat{p}_y, \hat{q}_y) = 0.2$, the highest shift among all tested scenarios. In both cases (high κ estimate and high label shift estimates) the entanglement estimates $\hat{\mathcal{E}}_y(f_{\text{WRR}})$ as well as the value of $\hat{\mathcal{R}}(f_{\text{WRR}})$ are quite high, indicating that besides entanglement, the WRR minimization is also unable to find minima with suitably low marginal distances.

6 CONCLUSION

In unsupervised domain adaptation (UDA), one common approach involves minimizing the source risk with a dissimilarity based regularizer that aligns marginal distributions. In the case of Wasserstein distances, we showed that regularizing the source risk by the marginal distance is not sufficient to get good target performance in general. Indeed, our bounds in Corollary 3.1 elucidate unoptimizable terms – such as the *label entanglement*, which may not decrease even when the source risk and marginal distance both decrease. This entanglement term intuitively corresponds to the average effect of not using *invariant features* (Arjovsky et al., 2019; Zhao et al., 2019). Through our analysis, we showed that the assumptions such as the Close Conditionals (CC) or Gradual Shift (GS) assumptions can lead to stronger theoretical guarantees, limiting the effect of the entanglement. In addition, experimentally we observed that the entanglement term is a good tool for analyzing the optimization landscape of UDA algorithms: depending on the (model, optimizer,

dataset) triplet we showed that the performance of UDA algorithms can vary significantly from cases of positive transfer to cases of highly negative transfer and that target accuracies correlate strongly with the degree of entanglement. Future work could include analyzing the optimization process more closely or exploring other reasonable assumptions to make UDA more tractable.

Acknowledgments

References

Acuna, D., Zhang, G., Law, M. T., and Fidler, S. (2021). f-domain adversarial learning: Theory and algorithms. In *International Conference on Machine Learning*, pages 66–75. PMLR.

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

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

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Bengio, Y., Goodfellow, I., and Courville, A. (2017). *Deep learning*, volume 1. MIT press Cambridge, MA, USA.

Bretagnolle, J. and Huber, C. (1979). Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47:119–137.

Canonne, C. L. (2022). A short note on an inequality between kl and tv. *arXiv preprint arXiv:2202.07198*.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30.

Crammer, K., Kearns, M., and Wortman, J. (2008). Learning from multiple sources. *Journal of machine learning research*, 9(8).

Csiszár, I. (1967). On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318.

Damodaran, B. B., Kellenberger, B., Flamary, R., Tuia, D., and Courty, N. (2018). Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 447–463.

Doberkat, E.-E. (2007). *Stochastic relations: foundations for Markov transition systems*. Chapman and Hall/CRC.

Donsker, M. D. and Varadhan, S. S. (1983). Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36(2):183–212.

Federici, M., Tomioka, R., and Forré, P. (2021). An information-theoretic approach to distribution shifts. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouve, A., and Peyré, G. (2019). Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021). Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

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

Ginosar, S., Rakelly, K., Sachs, S., Yin, B., and Efros, A. A. (2015). A century of portraits: A visual historical record of american high school yearbooks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–7.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks.

Gulrajani, I. and Lopez-Paz, D. (2020). In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hendrycks, D. and Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. (2006). Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19.

Johansson, F. D., Sontag, D., and Ranganath, R. (2019). Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536. PMLR.

Kirchmeyer, M., Rakotomamonjy, A., de Bezenac, E., and Gallinari, P. (2022). Mapping conditional distributions for domain adaptation under generalized target shift. In *International Conference on Learning Representations*.

Knuth, D. E. (1992). Two notes on notation. *The American Mathematical Monthly*, 99(5):403–422.

Kumar, A., Ma, T., and Liang, P. (2020). Understanding self-training for gradual domain adaptation. In *International conference on machine learning*, pages 5468–5479. PMLR.

Lee, D.-H. et al. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta.

Li, Y., Gong, M., Tian, X., Liu, T., and Tao, D. (2018). Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Nado, Z., Padhy, S., Sculley, D., D'Amour, A., Lakshminarayanan, B., and Snoek, J. (2020). Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*.

Nguyen, A. T., Tran, T., Gal, Y., Torr, P., and Baydin, A. G. (2022). KL guided domain adaptation. In *International Conference on Learning Representations*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E.,

DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library.

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

Pinsker, M. S. (1964). Information and information stability of random variables and processes. *Holden-Day*.

Redko, I., Habrard, A., and Sebban, M. (2017). Theoretical analysis of domain adaptation with optimal transport.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shen, J., Qu, Y., Zhang, W., and Yu, Y. (2018). Wasserstein distance guided representation learning for domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

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

Shu, R., Bui, H. H., Narui, H., and Ermon, S. (2018). A dirt-t approach to unsupervised domain adaptation. *ArXiv*, abs/1802.08735.

Sugiyama, M. and Kawanabe, M. (2012). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press.

Tachet des Combes, R., Zhao, H., Wang, Y.-X., and Gordon, G. J. (2020). Domain adaptation with conditional distribution matching and generalized label shift. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19276–19289. Curran Associates, Inc.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Wang, Z. and Mao, Y. (2023). Information-theoretic analysis of unsupervised domain adaptation. In *The Eleventh International Conference on Learning Representations*.

Wang, Z. and Mao, Y. (2024). On f-divergence principled domain adaptation: An improved framework. *arXiv preprint arXiv:2402.01887*.

Wilson, G. and Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46.

Xu, X., Zhou, X., Venkatesan, R., Swaminathan, G., and Majumder, O. (2019). d-sne: Domain adaptation using stochastic neighborhood embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2497–2506.

You, K., Long, M., Cao, Z., Wang, J., and Jordan, M. I. (2019). Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2720–2729.

Zhao, H., Des Combes, R. T., Zhang, K., and Gordon, G. (2019). On learning invariant representations for domain adaptation. In *International conference on machine learning*, pages 7523–7532. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Not Applicable

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. No. We will release our code if our submission is accepted.

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. Yes

   (b) Complete proofs of all theoretical results. Yes, we provide proofs in the appendix. We also have Proof sketches which we decided to omit due to space considerations.

   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). No, but we will release all code and necessary instructions if our paper is accepted.

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes, in the appendix we give a detailed description.

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes, in the appendix we provide these details.

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). No, but all our results were computed using our work laptops.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. Yes. In the appendix we cite all the datasets that we used.

   (b) The license information of the assets, if applicable. Not Applicable

   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable

   (d) Information about consent from data providers/curators. Not Applicable. All data that we used have permissible licenses and were obtained either by Pytorch or by downloading them from the data provider's website.

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. Not Applicable

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Appendix

A COUPLING PROPERTIES

Notation. In the proofs shown in the Appendix, for convenience we group the variables integrated over, e.g., $\mathrm{d}(y, y')$ indicates joint integration over the variables y and y'. We further utilize the Iverson brackets $[\![\texttt{predicate}]\!]$ which evaluate to 1 if `predicate` is true and 0 if it is false (Knuth, 1992).

Lemma A.1. *Suppose $\gamma_{\mathrm{x}} \in \Gamma(\mu_{\mathrm{x}}, \nu_{\mathrm{x}})$ and $\gamma_{\mathrm{y|x}}(\cdot \mid x, x') \in \Gamma(\mu_{\mathrm{y|x}}(\cdot \mid x), \nu_{\mathrm{y|x}}(\cdot \mid x'))$. Then for $\gamma = \gamma_{\mathrm{x}} \cdot \gamma_{\mathrm{y|x}}$, $\gamma \in \Gamma(\mu, \nu)$.*

Proof. Follows immediately from the calculation of marginals. $\qquad\square$

Lemma A.2. *Let $\mu, \nu \in \triangle(\mathcal{X} \times \mathcal{Y})$. Then for any coupling $\gamma \in \Gamma(\mu, \nu)$, there exists a factorization such that $\gamma = \gamma_{\mathrm{x}} \cdot \alpha$, where $\gamma_{\mathrm{x}} \in \Gamma(\mu_{\mathrm{x}}, \nu_{\mathrm{x}})$ and $\alpha(\cdot \mid x, x') \in \triangle(\mathcal{Y} \times \mathcal{Y})$ is some probability measure for each x, x'.*

Specifically,

$$\gamma_{\mathrm{x}}(x, x') = \int \gamma((x, y), (x', y'))\mathrm{d}(y, y').$$

Proof. Define γ_{x} as per the lemma. One can then verify that marginalization yields μ_{x} and ν_{x}. For example,

$$\begin{aligned}
\int \gamma_{\mathrm{x}}(x, x')\mathrm{d}x' &= \int \gamma((x, y), (x', y'))\mathrm{d}(y, y')\mathrm{d}x' \\
&= \int \gamma((x, y), (x', y'))\mathrm{d}(x', y')\mathrm{d}y \\
&= \int \mu(x, y)\mathrm{d}y \\
&= \mu_{\mathrm{x}}(x).
\end{aligned}$$

The opposite marginalization follows identically. α is a (conditional) probability measure as a result of the above marginalization of γ. Note that α is not necessarily a coupling of conditionals, i.e., $\alpha(\cdot \mid x, x') \in \Gamma(\mu_{\mathrm{y|x}}(\cdot \mid x), \nu_{\mathrm{y|x}}(\cdot \mid x')$ may not hold. $\qquad\square$

Lemma A.3. *Let $f \colon \mathcal{X} \to \mathcal{Y}$ be a measurable function and $\mu, \nu \in \triangle(\mathcal{X} \times \mathcal{Y})$. Suppose $\gamma \in \Gamma(\mu, \nu)$. Then,*

$$\bar{f}\sharp\gamma \in \Gamma(f\sharp\mu, f\sharp\nu). \tag{9}$$

where $\bar{f} \colon (x, y, x', y') \to (f(x), y, f(x'), y')$.

Proof. The result follows from simply utilizing the change of measure property of pushforward measures:

$$\begin{aligned}
\int (\bar{f}\sharp\gamma)((\hat{y}, y), (\hat{y}', y'))\mathrm{d}(\hat{y}', y') &= \int \gamma(f^{-1}(\hat{y}), y, f^{-1}(\hat{y}'), y')\mathrm{d}(\hat{y}', y') \\
&= \int \gamma(f^{-1}(\hat{y}), y, x', y')\mathrm{d}(x', y') \\
&= \mu(f^{-1}(\hat{y}), y) \\
&= f\sharp\mu(\hat{y}, y).
\end{aligned}$$

As the calculation holds symmetrically for marginalizing over the other argument, the result holds. $\qquad\square$

B   PROOFS

B.1   Proof of Lemma 3.1

Proof. To prove the result, for an arbitrary coupling $\gamma \in \Gamma(p, q)$ we consider the following:

$$
\begin{aligned}
\epsilon_q(f) &= \underset{q}{\mathbb{E}}[\ell(f(x), y)] \\
&= \underset{p}{\mathbb{E}}[\ell(f(x), y)] + \underset{q}{\mathbb{E}}[\ell(f(x), y)] - \underset{p}{\mathbb{E}}[\ell(f(x), y)] \\
&= \underset{p}{\mathbb{E}}[\ell(f(x), y)] + \underset{(x,y),(x',y')\sim\gamma}{\mathbb{E}} [\ell(f(x'), y') - \ell(f(x), y)] \\
&\leq \underset{p}{\mathbb{E}}[\ell(f(x), y)] + \underset{(x,y),(x',y')\sim\gamma}{\mathbb{E}} [\ell(f(x), f(x')) + \ell(f(x), y) + \ell(y, y') - \ell(f(x), y)] \\
&= \underset{p}{\mathbb{E}}[\ell(f(x), y)] + \underset{(x,y),(x',y')\sim\gamma}{\mathbb{E}} [\ell(f(x), f(x')) + \ell(y, y')] ,
\end{aligned}
$$

where the inequality holds due to our assumptions of symmetry and triangle inequality for ℓ. This proves the first result after taking an infimum over couplings. For the next result, we need to further examine the expectation

$$
\underset{(x,y),(x',y')\sim\gamma}{\mathbb{E}} [\ell(f(x), f(x')) + \ell(y, y')] = \int c_\ell(f(x), y, f(x'), y') \mathrm{d}\gamma(x, y, x', y')
$$
$$
= \int c_\ell(\hat{y}, y, \hat{y}', y') \mathrm{d}\bar{f}\sharp\gamma(\hat{y}, y, \hat{y}', y').
$$

From Lemma A.3, we have that $\bar{f}\sharp\gamma \in \Gamma(f\sharp p, f\sharp q)$ for $\bar{f} : (x, y, x', y') \to (f(x), y, f(x'), y')$. Now following Doberkat (2007, Proposition 1.101), as f is surjective, for any $\gamma' \in \Gamma(f\sharp p, f\sharp q)$ there exists a $\gamma \in \Gamma(p, q)$ such that the pushforward by $\bar{f}$ is equivalent to γ'. This implies that

$$
\Gamma(f\sharp p, f\sharp q) \subseteq \left\{ \bar{f}\sharp\gamma : \gamma \in \Gamma(p, q) \right\}.
$$

As such, the result follows after applying an infimum over couplings on the inequality. □

B.2   Proof of Lemma 3.2

To prove the result, we breakdown the proof into the cases of lower and upper bound.

Proof of Lower Bound. Let us denote $\gamma^\star \in \Gamma(\mu, \nu)$ as an optimal transport plan for $W_{1,c}(\mu, \nu)$. That is,

$$
W_{1,c}(\mu, \nu) = \iint c((x, y), (x', y'))\gamma^\star((x, y), (x', y'))\mathrm{d}(y, y')\mathrm{d}(x, x').
$$

Let $\gamma_\mathrm{x} \in \Gamma(\mu_\mathrm{x}, \nu_\mathrm{x})$ such that $\gamma^\star = \gamma_\mathrm{x} \cdot \alpha$, where $\alpha(\cdot \mid x, x') \in \triangle(\mathcal{Y} \times \mathcal{Y})$ for each x, x', as per Lemma A.2,

Thus we have

$$
\begin{aligned}
W_{p,c}^p(\mu, \nu) &= \iint c^p((x, y), (x', y'))\gamma^\star((x, y), (x', y'))\mathrm{d}(y, y')\mathrm{d}(x, x') \\
&\geq \iint c_1^p(x, x')\gamma^\star((x, y), (x', y'))\mathrm{d}(y, y')\mathrm{d}(x, x') \\
&= \int c_1^p(x, x') \int \gamma^\star((x, y), (x', y'))\mathrm{d}(y, y')\mathrm{d}(x, x') \\
&= \int c_1^p(x, x')\gamma_\mathrm{x}(x, x')\mathrm{d}(x, x') \\
&\geq \inf_{\gamma_\mathrm{x}\in\Gamma(\mu_\mathrm{x},\nu_\mathrm{x})} \int c_1^p(x, x')\gamma_\mathrm{x}(x, x')\mathrm{d}(x, x') \\
&= W_{p,c_1}^p(\mu_\mathrm{x}, \nu_\mathrm{x}),
\end{aligned}
$$

where the first inequality comes from the non-negativity of cost functions: $c^p = (c_1 + c_2)^p \geq c_1^p$.

The case for $W_{p,c_2}^p(\mu_\mathrm{y}, \nu_\mathrm{y})$ follows indentically. □

Proof of Upper Bound. We prove the first statement. The second inequality is proven identically, with x's and c_1's role exchanged with y's and c_2's. We start from the definition of the Wasserstein p-distance

$$
W_{p,c}(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \left\{ \int c((x,y),(x',y'))^p \gamma((x,y),(x',y')) \mathrm{d}(y,y') \mathrm{d}(x,x') \right\} \right)^{1/p}
$$

$$
= \inf_{\gamma \in \Gamma(\mu,\nu)} \left\{ \int (c_1(x,x') + c_2(y,y'))^p \, \gamma((x,y),(x',y')) \mathrm{d}(y,y') \mathrm{d}(x,x') \right\}^{1/p}
$$

$$
\leq \inf_{\gamma \in \Gamma(\mu,\nu)} \left\{ \left(\int c_1^p(x,x') \gamma((x,y),(x',y')) \mathrm{d}(y,y') \mathrm{d}(x,x') \right)^{1/p} \right.
$$

$$
\left. + \left(\int c_2^p(y,y') \gamma((x,y),(x',y')) \mathrm{d}(y,y') \mathrm{d}(x,x') \right)^{1/p} \right\}
$$

where we applied Minkowski's inequality. We now define the sets,

$$
\Gamma_{\mathrm{x}}(\mu,\nu) = \Gamma(\mu_x, \nu_x)
$$

$$
\Gamma_{\mathrm{y|x}}(\mu,\nu \mid x,x') = \Gamma(\mu_{\mathrm{y|x}}(\cdot \mid x), \nu_{\mathrm{y|x}}(\cdot \mid x'))
$$

$$
\Gamma^f(\mu,\nu) = \{ \gamma \mid \gamma(x,y,x',y') = \gamma_{\mathrm{x}}(x,x') \cdot \gamma_{\mathrm{y|x}}(y,y' \mid x,x'),
$$

$$
\gamma_{\mathrm{x}} \in \Gamma(\mu_x, \nu_x), \gamma_{\mathrm{y|x}}(\cdot \mid x,x') \in \Gamma_{\mathrm{y|x}}(\mu,\nu \mid x,x') \}.
$$

One can easily verify that $\Gamma^f(\mu,\nu) \subseteq \Gamma(\mu,\nu)$. Furthermore, let $\gamma_{\mathrm{x}}^\star = \arg\inf_{\gamma_{\mathrm{x}} \in \Gamma_{\mathrm{x}}(\mu,\nu)} \int c_1^p(x,x') \mathrm{d}\gamma_{\mathrm{x}}(x,x')$. As a result, we have

$$
W_{p,c}(\mu,\nu) \leq \inf_{\gamma \in \Gamma^f(\mu,\nu)} \left\{ \left(\int c_1^p(x,x') \gamma((x,y),(x',y')) \mathrm{d}(y,y') \mathrm{d}(x,x') \right)^{1/p} \right.
$$

$$
\left. + \left(\int c_2^p(y,y') \gamma((x,y),(x',y')) \mathrm{d}(y,y') \mathrm{d}(x,x') \right)^{1/p} \right\}
$$

$$
= \inf_{\gamma_{\mathrm{x}} \in \Gamma_{\mathrm{x}}(\mu,\nu)} \inf_{\gamma_{\mathrm{y|x}}(\cdot|x,x') \in \Gamma_{\mathrm{y|x}}(\mu,\nu|x,x')} \left\{ \left(\int c_1^p(x,x') \gamma_{\mathrm{x}}(x,x') \mathrm{d}(x,x') \right)^{1/p} \right.
$$

$$
\left. + \left(\int c_2^p(y,y') \gamma_{\mathrm{x}}(x,x') \gamma_{\mathrm{y|x}}(y,y' \mid x,x') \mathrm{d}(y,y') \mathrm{d}(x,x') \right)^{1/p} \right\}
$$

$$
\leq \inf_{\gamma_{\mathrm{y|x}}(\cdot|x,x') \in \Gamma_{\mathrm{y|x}}(\mu,\nu|x,x')} \left\{ \left(\int c_1^p(x,x') \gamma_{\mathrm{x}}^\star(x,x') \mathrm{d}(x,x') \right)^{1/p} \right.
$$

$$
\left. + \left(\int c_2^p(y,y') \gamma_{\mathrm{x}}^\star(x,x') \gamma_{\mathrm{y|x}}(y,y' \mid x,x') \mathrm{d}(y,y') \mathrm{d}(x,x') \right)^{1/p} \right\}
$$

$$
= \left(\int c_1^p(x,x') \gamma_{\mathrm{x}}^\star(x,x') \mathrm{d}(x,x') \right)^{1/p}
$$

$$
+ \inf_{\gamma_{\mathrm{y|x}}(\cdot|x,x') \in \Gamma_{\mathrm{y|x}}(\mu,\nu|x,x')} \left\{ \left(\int c_2^p(y,y') \gamma_{\mathrm{x}}^\star(x,x') \gamma_{\mathrm{y|x}}(y,y' \mid x,x') \mathrm{d}(y,y') \mathrm{d}(x,x') \right)^{1/p} \right\}
$$

$$
= W_{p,c_1}(\mu_{\mathrm{x}}, \nu_{\mathrm{x}})
$$

$$
+ \inf_{\gamma_{\mathrm{y|x}}(\cdot|x,x') \in \Gamma_{\mathrm{y|x}}(\mu,\nu|x,x')} \left\{ \left(\int c_2^p(y,y') \gamma_{\mathrm{x}}^\star(x,x') \gamma_{\mathrm{y|x}}(y,y' \mid x,x') \mathrm{d}(y,y') \mathrm{d}(x,x') \right)^{1/p} \right\}
$$

$$
= W_{p,c_1}(\mu_{\mathrm{x}}, \nu_{\mathrm{x}}) + \left(\int \gamma_{\mathrm{x}}^\star(x,x') \cdot W_{p,c_2}^p(\mu_{\mathrm{y|x}}(\cdot \mid x), \nu_{\mathrm{y|x}}(\cdot \mid x')) \mathrm{d}(x,x') \right)^{1/p}
$$

$$
= W_{p,c_1}(\mu_{\mathrm{x}}, \nu_{\mathrm{x}}) + \mathbb{E}_{\gamma_x^\star} \left[W_{p,c_2}^p(\mu_{\mathrm{y|x}}(\cdot \mid x), \nu_{\mathrm{y|x}}(\cdot \mid x')) \right]^{1/p}.
$$

(We note that the infimum $\inf_{\gamma_{\mathrm{y|x}}(\cdot|x,x') \in \Gamma_{\mathrm{y|x}}(\mu,\nu|x,x')}$ uses a slight abuse of notation where we are taking individual infimums over each x, x' pairs.) □

B.3   Proof of Corrollary 3.1

Proof. Follows immediately from Lemma 3.1 and 3.2. □

B.4   Proof of Theorem 3.1

We first prove the following general Lemma.

Lemma B.1. *Suppose that Assumption 3.1 holds. Then for all $f \in \mathcal{H}$*

$$\mathcal{E}_{\mathrm{y}}(f) \le \epsilon_p(f) + \epsilon_q(f) + W_{1,\ell}(f \sharp p_{\mathrm{x}}, f \sharp q_{\mathrm{x}}).$$

Proof. Let $f \in \mathcal{H}$ be arbitrary. We begin by first upper bounding the label entanglement

$$\mathcal{E}_{\mathrm{y}}(f) = \int W_{1,\ell}(p_{\mathrm{y|f}}(\cdot \mid \hat{y}), q_{\mathrm{y|f}}(\cdot \mid \hat{y}')) \mathrm{d}\gamma_{\mathrm{f}}^{\star}(\hat{y}, \hat{y}')$$

$$= \iint \ell(y, y') \mathrm{d}\gamma_{\mathrm{y|f}}(y, y' \mid \hat{y}, \hat{y}') \mathrm{d}\gamma_{\mathrm{f}}^{\star}(\hat{y}, \hat{y}'),$$

where $\gamma_{\mathrm{f}}^{\star}$ is the optimal transport plan between $f \sharp p_{\mathrm{x}}$ and $f \sharp q_{\mathrm{x}}$ and $\gamma_{\mathrm{y|f}}(\cdot \mid \hat{y}, \hat{y}')$ is the optimal transport plan between $p_{\mathrm{y|f}}$ and $q_{\mathrm{y|f}}$, with respect to their corresponding Wasserstein distances.

Now, applying triangle inequality to the loss function ℓ, assumed among the metric properties of ℓ, we have

$$\mathcal{E}_{\mathrm{y}}(f) = \iint \ell(y, y') \mathrm{d}\gamma_{\mathrm{y|f}}^{\star}(y, y' \mid \hat{y}, \hat{y}') \mathrm{d}\gamma_{\mathrm{f}}^{\star}(\hat{y}, \hat{y}')$$

$$\le \iint \left(\ell(\hat{y}, y) + \ell(\hat{y}', y') + \ell(\hat{y}, \hat{y}') \right) \mathrm{d}\gamma_{\mathrm{y|f}}^{\star}(y, y' \mid \hat{y}, \hat{y}') \mathrm{d}\gamma_{\mathrm{f}}^{\star}(\hat{y}, \hat{y}')$$

$$= \epsilon_p(f) + \epsilon_q(f) + \iint \ell(\hat{y}, \hat{y}') \mathrm{d}\gamma_{\mathrm{y|f}}^{\star}(y, y' \mid \hat{y}, \hat{y}') \mathrm{d}\gamma_{\mathrm{f}}^{\star}(\hat{y}, \hat{y}')$$

$$= \epsilon_p(f) + \epsilon_q(f) + \int \ell(\hat{y}, \hat{y}') \mathrm{d}\gamma_{\mathrm{f}}^{\star}(\hat{y}, \hat{y}')$$

$$= \epsilon_p(f) + \epsilon_q(f) + W_{1,\ell}(f \sharp p_{\mathrm{x}}, f \sharp q_{\mathrm{x}}),$$

where the last equality follows from the optimality of the transportation plan. □

Proof of Theorem 3.1. The proofs of the two inequalities are symmetric. As such we only provide a proof of the former label entanglement. From Lemma 3.2, we can switch from a lower bound to an upper bound to yield,

$$W_{1,\ell}(f \sharp p_{\mathrm{x}}, f \sharp q_{\mathrm{x}}) \le W_{1,\ell,\ell}(f \sharp p, f \sharp q) \le W_{1,\ell}(p_{\mathrm{y}}, q_{\mathrm{y}}) + \mathcal{E}_{\hat{\mathrm{y}}}(f).$$

Combining this fact together with Lemma B.1 completes the proof. □

B.5   Proof of Corollary 3.2

Proof. We first prove the lower bound. Using Theorem 3.1, we have

$$\mathcal{U}(f) \le 2\epsilon_p(f) + W_{1,\ell}(p_{\mathrm{y}}, q_{\mathrm{y}}) + \epsilon_q(f) + \mathcal{E}_{\hat{\mathrm{y}}}(f) + W_{1,\ell}(f \sharp p_{\mathrm{x}}, f \sharp q_{\mathrm{x}}).$$

Now, we can upper bound $W_{1,\ell}(f \sharp p_{\mathrm{x}}, f \sharp q_{\mathrm{x}})$ using the same reasoning that was used in the proof of Theorem 3.1, yielding

$$\mathcal{U}(f) \le 2\epsilon_p(f) + W_{1,\ell}(p_{\mathrm{y}}, q_{\mathrm{y}}) + \epsilon_q(f) + \mathcal{E}_{\hat{\mathrm{y}}}(f) + W_{1,\ell}(f \sharp p_{\mathrm{x}}, f \sharp q_{\mathrm{x}})$$

$$\le 2\epsilon_p(f) + 2W_{1,\ell}(p_{\mathrm{y}}, q_{\mathrm{y}}) + \epsilon_q(f) + 2\mathcal{E}_{\hat{\mathrm{y}}}(f)$$

$$\le 3\epsilon_p(f) + 3W_{1,\ell}(p_{\mathrm{y}}, q_{\mathrm{y}}) + 3\mathcal{E}_{\hat{\mathrm{y}}}(f).$$

As required.

The upper bound is proven identically, except by starting with $\epsilon_p(f) + W_{1,\ell}(p_{\mathrm{y}}, q_{\mathrm{y}}) + \mathcal{E}_{\hat{\mathrm{y}}}(f)$ and switching the roles of $\mathcal{E}_{\hat{\mathrm{y}}}(f)$ and $\mathcal{E}_{\mathrm{y}}(f)$ in the proof above. □

B.6 Proof of Lemma 4.1

Proof. As $CC(\kappa)$ does not hold, for all $f \in \mathcal{H}$, we have that

$$\kappa < \epsilon_p(f) + \max_y W_{1,\ell}(f \sharp p_{\mathrm{x}|\mathrm{y}}(\cdot \mid y), f \sharp q_{\mathrm{x}|\mathrm{y}}(\cdot \mid y)).$$

Let us fix f for now and denote $\bar{y}$ as the label of the conditional which satisfies the maximum.

Now we have,

$$q_{\mathrm{y}}(\bar{y}) \cdot \kappa < q_{\mathrm{y}}(\bar{y}) \cdot \epsilon_p(f) + q_{\mathrm{y}}(\bar{y}) \cdot W_{1,\ell}(f \sharp p_{\mathrm{x}|\mathrm{y}}(\cdot \mid \bar{y}), f \sharp q_{\mathrm{x}|\mathrm{y}}(\cdot \mid \bar{y}))$$

$$\leq q_{\mathrm{y}}(\bar{y}) \cdot \epsilon_p(f) + \int q_{\mathrm{y}}(y) \cdot W_{1,\ell}(f \sharp p_{\mathrm{x}|\mathrm{y}}(\cdot \mid y), f \sharp q_{\mathrm{x}|\mathrm{y}}(\cdot \mid y)) \mathrm{d}y.$$

Now denote $\gamma^\star(\hat{y}, \hat{y}' \mid y)$ as the optimal transport plan corresponding to each Wasserstein distance between the conditionals. Thus, we have

$$q_{\mathrm{y}}(\bar{y}) \cdot \kappa < q_{\mathrm{y}}(\bar{y}) \cdot \epsilon_p(f) + \int q_{\mathrm{y}}(y) \int \ell(\hat{y}, \hat{y}') \gamma^\star(\hat{y}, \hat{y}' \mid y) \mathrm{d}(\hat{y}, \hat{y}') \mathrm{d}y$$

$$\leq q_{\mathrm{y}}(\bar{y}) \cdot \epsilon_p(f) + \int q_{\mathrm{y}}(y) \int (\ell(\hat{y}, y) + \ell(\hat{y}', y)) \gamma^\star(\hat{y}, \hat{y}' \mid y) \mathrm{d}(\hat{y}, \hat{y}') \mathrm{d}y$$

$$= q_{\mathrm{y}}(\bar{y}) \cdot \epsilon_p(f) + \int q_{\mathrm{y}}(y) \left(\int \ell(\hat{y}, y) f \sharp p_{\mathrm{x}|\mathrm{y}}(\hat{y}) \mathrm{d}\hat{y} + \int \ell(\hat{y}', y) f \sharp q_{\mathrm{x}|\mathrm{y}}(\hat{y}') \mathrm{d}\hat{y}' \right) \mathrm{d}y$$

$$= q_{\mathrm{y}}(\bar{y}) \cdot \epsilon_p(f) + \int q_{\mathrm{y}}(y) \left(\int \ell(f(x), y) p_{\mathrm{x}|\mathrm{y}}(x) \mathrm{d}x + \int \ell(f(x'), y) q_{\mathrm{x}|\mathrm{y}}(x') \mathrm{d}x' \right) \mathrm{d}y$$

$$= q_{\mathrm{y}}(\bar{y}) \cdot \epsilon_p(f) + \epsilon_q(f) + \epsilon_{p_{\mathrm{x}|\mathrm{y}} q_{\mathrm{y}}}(f).$$

The second inequality follows from the triangle inequality (among the assumed metric properties of ℓ).

Taking maximum and minimums to make the bound hold uniformly over $f \in \mathcal{H}$ yields the result. $\qquad\square$

B.7 Proof of Lemma 4.2

Proof. Using the second upper bound in Lemma 3.2 for 1-Wasserstein distances

$$\epsilon_q(f^\star) \leq \kappa + \delta + \int [\![y \neq y']\!] W_{1,\ell}(f^\star \sharp p_{x|y}(\cdot \mid y), f^\star \sharp q_{x|y}(\cdot \mid y')) \mathrm{d}\gamma_{\mathrm{y}}^\star(y, y'),$$

where $f^\star$ witnesses our assumption of $CC(\kappa)$.

Now all we need to do is to bound the terms on the conditionals. We utilize the boundedness assumption on ℓ, yielding,

$$\epsilon_q(f^\star) \leq \kappa + \delta + L \cdot \int [\![y \neq y']\!] \mathrm{d}\gamma_{\mathrm{y}}^\star(y, y').$$

Now the summation over the coupling can be bounded by lower bounding $W_{1,\ell}(p_{\mathrm{y}}, q_{\mathrm{y}})$,

$$l \int [\![y \neq y']\!] \mathrm{d}\gamma_{\mathrm{y}}^\star(y, y') \leq \int \ell(y, y') \mathrm{d}\gamma_{\mathrm{y}}^\star(y, y') < \delta \implies \int [\![y \neq y']\!] \mathrm{d}\gamma_{\mathrm{y}}^\star(y, y') \leq \delta/l,$$

where we use the assumption that $\ell(y, y') = 0$ when $y = y'$ and that $W_{1,\ell}(p_{\mathrm{y}}, q_{\mathrm{y}}) < \delta$.

This yields,

$$\epsilon_q(f^\star) < \kappa + \delta + \frac{L}{l} \cdot \delta.$$

Thus together, we have

$$\epsilon_p(f^\star) + \epsilon_q(f^\star) < 2\kappa + \delta + \frac{L}{l} \cdot \delta.$$

As required. $\qquad\square$

B.8   Proof of Corollary 4.1

We split the proof of this corollary into two sub-results.

Corollary B.1. *Suppose Assumption 3.1 to 3.3 holds and $p_y = q_y$. If $\mathrm{CC}(\kappa)$ holds for $\kappa > 0$, then $\mathrm{LJE}(\lambda)$ holds for $\lambda = 2\kappa$.*

Proof. Hold via Lemma 4.2 with $\delta = 0$. □

Corollary B.2. *Suppose Assumption 3.1 holds and $p_y = q_y$. If $\mathrm{LJE}(\lambda)$ holds for $\lambda > 0$, then $\mathrm{CC}(\kappa)$ holds for $\kappa = \lambda \cdot \frac{1 + \max_y q_y(y)}{\min q_y(y)}$.*

Proof. Holds via contra-positive of Lemma 4.1. □

Proof of Corollary 4.1. Follows from Corollary B.1 and B.2. □

B.9   Proof of Lemma 4.4

Proof. Let $f \in \mathcal{H}$ satisfy $\mathrm{CC}(\kappa)$. Using Corollary 3.2, we have that

$$
\begin{aligned}
\mathcal{U}(f) &\leq 3\epsilon_p(f) + 3W_{1,\ell}(p_y, q_y) + 3\mathcal{E}_{\hat{y}}(f) \\
&< 3\delta + 3\epsilon_p(f) + 3\mathcal{E}_{\hat{y}}(f) \\
&= 3\delta + 3\epsilon_p(f) + 3\int_{\mathcal{Y}\times\mathcal{Y}} W_{1,\ell}(p_{f|y}(\cdot \mid y), q_{f|y}(\cdot \mid y'))\mathrm{d}\gamma_y^\star(y, y') \\
&= 3\delta + 3\epsilon_p(f) + 3\int_{\mathcal{Y}\times\mathcal{Y}} [\![y = y']\!] \cdot W_{1,\ell}(p_{f|y}(\cdot \mid y), q_{f|y}(\cdot \mid y'))\mathrm{d}\gamma_y^\star(y, y') \\
&\qquad + 3\int_{\mathcal{Y}\times\mathcal{Y}} [\![y \neq y']\!] \cdot W_{1,\ell}(p_{f|y}(\cdot \mid y), q_{f|y}(\cdot \mid y'))\mathrm{d}\gamma_y^\star(y, y') \\
&< 3\delta + 3\kappa + 3\int_{\mathcal{Y}\times\mathcal{Y}} [\![y \neq y']\!] \cdot W_{1,\ell}(p_{f|y}(\cdot \mid y), q_{f|y}(\cdot \mid y'))\mathrm{d}\gamma_y^\star(y, y') \\
&\leq 3\delta + 3\kappa + 3L\int_{\mathcal{Y}\times\mathcal{Y}} [\![y \neq y']\!]\mathrm{d}\gamma_y^\star(y, y').
\end{aligned}
$$

Using the same argument as Lemma 4.2, we have

$$
\int_{\mathcal{Y}\times\mathcal{Y}} [\![y \neq y']\!]\mathrm{d}\gamma_y^\star(y, y') \leq \delta/l.
$$

The result follows immediately after. □

A simpler version of the lemma can also be proven:

Corollary B.3. *Suppose that Assumption 3.1. Then $f \in \mathcal{H}$ satisfying $\mathrm{LJE}(\lambda)$, implies that*

$$
\mathcal{U}(f) \leq 2\lambda + 2W_{1,\ell}(f\sharp p_x, f\sharp q_x).
$$

Proof. The result yields immediately after upper bounding $\mathcal{U}(f)$ using the $\mathrm{LJE}(\lambda)$ assumption and Lemma B.1. □

B.10   Proof of Lemma 4.5

Proof. Let f satisfy the requirements for GS.

Following the convexity of Wasserstein distance, we have that

$$
W_{1,\ell}(f\sharp p_{x|y}(\cdot \mid y), f\sharp q_{x|y}(\cdot \mid y)) \leq \sum_{i=1}^{s} r_i \cdot W_{1,\ell}(f\sharp p_{x|y}(\cdot \mid y), f\sharp q_{x|y}^{(i)}(\cdot \mid y)).
$$

Note that for each of i, applying triangle inequality

$$W_{1,\ell}(f\sharp p_{\mathsf{x}|\mathsf{y}}(\cdot \mid y), f\sharp q_{\mathsf{x}|\mathsf{y}}^{(i)}(\cdot \mid y)) \le \sum_{j=1}^{i} W_{1,\ell}(f\sharp q_{\mathsf{x}|\mathsf{y}}^{(j-1)}(\cdot \mid y), f\sharp q_{\mathsf{x}|\mathsf{y}}^{(j)}(\cdot \mid y))$$

$$< i \cdot \varepsilon.$$

Thus together, we have for any $y \in \mathcal{Y}$

$$W_{1,\ell}(f\sharp p_{\mathsf{x}|\mathsf{y}}(\cdot \mid y), f\sharp q_{\mathsf{x}|\mathsf{y}}(\cdot \mid y)) < \varepsilon \cdot \sum_{i=1}^{s} i \cdot r_i$$

$$< \varepsilon \cdot a \cdot \sum_{i=1}^{s} i$$

$$< \varepsilon \cdot a \cdot \frac{s(s+1)}{2}. \tag{10}$$

where the last inequality holds from taking an arithmetic sequence. Using the definition of the $\mathrm{CC}(\kappa)$ assumption, the result follows immediately. $\qquad\square$

B.11 Proof of Theorem 4.1

Proof. To prove our upper bound on the label entanglement, we will utilize Theorem 3.1 and first upper bound the prediction entanglement. Consider the following decomposition of the prediction entanglement:

$$\mathcal{E}_{\hat{\mathsf{y}}}(f) = \int W_{1,\ell}(p_{\mathsf{f}|\mathsf{y}}(\cdot \mid y), q_{\mathsf{f}|\mathsf{y}}(\cdot \mid y'))\mathrm{d}\gamma_{\mathsf{y}}^{\star}(y, y')$$

$$= \int [\![y \ne y']\!] \cdot W_{1,\ell}(p_{\mathsf{f}|\mathsf{y}}(\cdot \mid y), q_{\mathsf{f}|\mathsf{y}}(\cdot \mid y'))\mathrm{d}\gamma_{\mathsf{y}}^{\star}(y, y') + \int [\![y = y']\!] \cdot W_{1,\ell}(p_{\mathsf{f}|\mathsf{y}}(\cdot \mid y), q_{\mathsf{f}|\mathsf{y}}(\cdot \mid y'))\mathrm{d}\gamma_{\mathsf{y}}^{\star}(y, y').$$

We upper bound the first term corresponding to the case of $y \ne y'$. For any $y, y' \in \mathcal{Y}$, we have

$$W_{1,\ell}(p_{\mathsf{f}|\mathsf{y}}(\cdot \mid y), q_{\mathsf{f}|\mathsf{y}}(\cdot \mid y')) = W_{1,\ell}(f\sharp p_{\mathsf{x}|\mathsf{y}}(\cdot \mid y), f\sharp q_{\mathsf{x}|\mathsf{y}}(\cdot \mid y'))$$

$$\le W_{1,\ell}(f\sharp p_{\mathsf{x}|\mathsf{y}}(\cdot \mid y), f\sharp p_{\mathsf{x}|\mathsf{y}}(\cdot \mid y')) + W_{1,\ell}(f\sharp p_{\mathsf{x}|\mathsf{y}}(\cdot \mid y'), f\sharp q_{\mathsf{x}|\mathsf{y}}(\cdot \mid y'))$$

$$= W_{1,\ell}(f\sharp p_{\mathsf{x}|\mathsf{y}}(\cdot \mid y'), f\sharp q_{\mathsf{x}|\mathsf{y}}(\cdot \mid y')) + \int \ell(\hat{y}, \hat{y}')\mathrm{d}\gamma_{\mathsf{f}}^{p}(\hat{y}, \hat{y}' \mid y, y'),$$

where $\gamma_{\mathsf{f}}^{p}(\hat{y}, \hat{y}' \mid y, y')$ denotes the optimal transport plan between $p_{\mathsf{f}|\mathsf{y}}(\cdot \mid y)$ and $p_{\mathsf{f}|\mathsf{y}}(\cdot \mid y')$. This latter term can be further bounded,

$$\int \ell(\hat{y}, \hat{y}')\mathrm{d}\gamma_{\mathsf{f}}^{p}(\hat{y}, \hat{y}' \mid y, y') \le \int \ell(\hat{y}, \hat{y}')\mathrm{d}f\sharp\tilde{\gamma}_{\mathsf{x}}^{p}(\hat{y}, \hat{y}' \mid y, y')$$

$$= \int \ell(f(x), f(x'))\mathrm{d}\tilde{\gamma}_{\mathsf{x}}^{p}(x, x' \mid y, y'),$$

where the inequality follows from the surjectivity of the pushforward (similar reasoning as per proof of Lemma 3.1). The coupling $\tilde{\gamma}_{\mathsf{x}}^{p}$ corresponds to the resulting coupling of said pushforward.

Now putting the above upper-bounds together, we can bound the prediction entanglement as follows:

$$\mathcal{E}_{\hat{\mathsf{y}}}(f) \le \int [\![y \ne y']\!] \cdot \left(W_{1,\ell}(p_{\mathsf{f}|\mathsf{y}}(\cdot \mid y'), q_{\mathsf{f}|\mathsf{y}}(\cdot \mid y')) + L\right) \mathrm{d}\gamma_{\mathsf{y}}^{\star}(y, y')$$

$$+ \int [\![y = y']\!] \cdot W_{1,\ell}(p_{\mathsf{f}|\mathsf{y}}(\cdot \mid y'), q_{\mathsf{f}|\mathsf{y}}(\cdot \mid y'))\mathrm{d}\gamma_{\mathsf{y}}^{\star}(y, y')$$

$$= \int W_{1,\ell}(p_{\mathsf{f}|\mathsf{y}}(\cdot \mid y'), q_{\mathsf{f}|\mathsf{y}}(\cdot \mid y'))\mathrm{d}\gamma_{\mathsf{y}}^{\star}(y, y') + L \cdot \int [\![y \ne y']\!]\mathrm{d}\gamma_{\mathsf{y}}^{\star}(y, y'),$$

where we used the assumed upper bound L on the loss function. Now to finalize the upper bound on prediction entanglement, we make two additional observations. First, from (10) in Lemma 4.5's proof, for any $y' \in \mathcal{Y}$, we have

$$W_{1,\ell}(p_{\mathrm{f}|\mathrm{y}}(\cdot \mid y'), p_{\mathrm{f}|\mathrm{y}}(\cdot \mid y')) \leq \kappa' \doteq \varepsilon \cdot \frac{a}{2} s(s+1).$$

Second, from the lower bound on the loss ℓ for $y \neq y'$, we have,

$$\int [\![y \neq y']\!] \cdot \mathrm{d}\gamma_{\mathrm{y}}^{\star}(y, y') \leq \frac{W_{1,\ell}(p_{\mathrm{y}}, q_{\mathrm{y}})}{l}.$$

This finally gives us,

$$\mathcal{E}_{\hat{\mathrm{y}}}(f) \leq \int W_{1,\ell}(p_{\mathrm{f}|\mathrm{y}}(\cdot \mid y'), q_{\mathrm{f}|\mathrm{y}}(\cdot \mid y')) \mathrm{d}\gamma_{\mathrm{y}}^{\star}(y, y') + L \cdot \int [\![y \neq y']\!] \mathrm{d}\gamma_{\mathrm{y}}^{\star}(y, y')$$

$$\leq \kappa' + L \cdot \frac{\delta}{l}.$$

Note that via assumption, we have $W_{1,\ell}(p_{\mathrm{y}}, q_{\mathrm{y}}) < \delta$. Now switching to label entanglement, from Theorem 3.1, we have that

$$\mathcal{E}_{\mathrm{y}}(f) \leq \mathcal{E}_{\hat{\mathrm{y}}}(f) + W_{1,\ell}(p_{\mathrm{y}}, q_{\mathrm{y}}) + \epsilon_p(f) + \epsilon_q(f)$$

$$= \kappa' + \delta \cdot \frac{L + l}{l} + \epsilon_p(f) + \epsilon_q(f).$$

Using the second bound in Corollary 3.1, we can bound the target risk $\epsilon_q(f)$ as

$$\epsilon_q(f) \leq \epsilon_p(f) + W_{1,l}(p_y, q_y) + \mathcal{E}_{\hat{\mathrm{y}}}(f) \leq b + \delta + \kappa' + L \cdot \frac{\delta}{l},$$

which gives us,

$$\mathcal{E}_{\mathrm{y}}(f) \leq 2\kappa' + 2b + 2\delta \cdot \frac{L + l}{l}.$$

Substituting κ' yields the desired result. $\qquad \square$

C RELATED WORK

Besides the UDA approaches mentioned in Section 2, Domain Adaptation (DA) was also studied in a number of specialized contexts, ranging from semi-supervised DA (where we can expect a limited amount of target labels (Xu et al., 2019)), and gradual DA (where the distributions typically shift slowly over time (Kumar et al., 2020)), to test-time DA (where source domain inputs are not accessible at test-time (Nado et al., 2020), e.g., due to privacy concerns) and universal DA (where the domain can enlarge to contain new classes) (You et al., 2019).

C.1 Comparisons to Existing Theory

When compared to previous work, we avoid removing the role of the loss ℓ and model f from the statistical divergence term. Indeed, prior work involving the Wasserstein distance has involved either partially removing the presence of the model by assuming that a probabilistic Lipschitz property holds for f (Courty et al., 2017, Theorem 3.1); or assumes that the mapping $(x, y) \mapsto \ell(f(x), y)$ is Lipschitz itself (Wang and Mao, 2023, Theorem 4.4). For the latter case of Wang and Mao (2023), Assumption 3.1 is not needed to establish a similar bound to Lemma 3.1; however, Assumption 3.1 is eventually made for further theoretic analysis.

From a technical perspective, all these bounds involving the Wasserstein distance (ours included) are based on exploiting the Lipschitz constant of different metric spaces. Indeed, in our case, the cost function is already a metric and thus is 1-Lipschitz to itself. See Appendix D for how to relax Assumption 3.1 while maintaining the essential properties of the proved bounds.

C.2 Relationship to Other Change of Measure Inequalities

A key point of comparison between our work and other domain matching methods which utilize divergences (over marginals) is the change of measure inequality used to switch from an expectation in q to an expectation in p (i.e., switching from the target risk to the source risk via an upper bound). In general, the change of measure inequalities yield variants of Lemma 3.1 that use different divergences (dissimilarity functions).

Our work is based on Wasserstein distances that use a particular choice of (decomposable) cost function derived by the loss function ℓ of the prediction task, Lemma 3.1. Several other approaches also utilize the Wasserstein distance, but differ in the specific cost functions used (Courty et al., 2017; Damodaran et al., 2018; Wang and Mao, 2023). One can, as a corollary, derive many of these variants of Lemma 3.1 by making additional Lipschitz continuity assumptions for the loss function ℓ or the model f.

Alternatively, a common choice of dissimilarity function is the KL-divergence via the Donskar-Varadhan change of measure inequality (Donsker and Varadhan, 1983). With the addition of the sub-Gaussian loss assumption, square-root of the KL-divergence was used in domain adaptation (Nguyen et al., 2022; Wang and Mao, 2023). More recent work has also modified the variational form of f-divergences to achieve similar upper bounds (Acuna et al., 2021; Wang and Mao, 2024). We discuss the relation of our Wasserstein distance based bounds to the KL-divergence in more detail in Appendix E.

C.3 Relationship to Other Conditional Discrepancy Terms

The entanglement terms that we proposed in section 3.3 are examples of *conditional discrepancy* terms introduced in the literature. An *average* conditional discrepancy term was first introduced in Theorem 1 of Ben-David et al. (2010), however it was not used in the central theorem of that paper (Theorem 2) that utilized the $H\Delta H$ divergence to bound the target risk. The authors in Zhao et al. (2019) proposed (Theorem 4.1) using the conditional discrepancy term in the proof of Theorem 2 of Ben-David et al. (2010), thereby replacing the λ of LJE(λ). This quantity, the joint error of the best performing hypothesis in the hypothesis class, was shown by Johansson et al. (2019) to be not invariant over the (typically) non-invertible layers of deep neural networks.

The Error Decomposition Theorem 3.1 of Tachet des Combes et al. (2020) forms a counterpart to the Theorem 4.1 in Zhao et al. (2019), in that the latter corresponds to the first bound (5) of Corollary 3.1 in our paper, while the former corresponds to the second bound (6). While we used the bound (6) only for theoretical analysis in Section 4, the authors in Tachet des Combes et al. (2020) use Theorem 3.1 to come up with a weighted domain matching algorithm. This approach was justified by making a *Generalized Label Shift* (GLS) assumption (Tachet des Combes et al., 2020), which is a sufficient condition to enforce zero (prediction) entanglement, however, it is in general not possible to satisfy such an assumption without target labels (as can be seen by the dependency of the upper bound in Theorem 3.4 of (Tachet des Combes et al., 2020) on the target risk). The entanglement terms thus capture the hardness of UDA (from an information-theoretic perspective) in a concise but general way: during any domain matching procedure that does not have access to target labels, the optimized model may entangle the conditionals (in the output space), and we cannot expect any assumptions such as GLS to be a useful guide during this optimization.

The authors in Federici et al. (2021) use the KL-divergence chain rule and mutual information to introduce their *latent concept shift*, which can be viewed as another conditional discrepancy term. The introduced bounds utilize the KL-divergence and its symmetric counterpart the JS-divergence, both of which require the shared support assumption, unlike the optimal-transport based bounds. We highlight the connection between optimal transport and KL-divergence based bounds in more detail in Appendix C.

As for works that utilized Wasserstein distances, Courty et al. (2017); Redko et al. (2017); Shen et al. (2018); Damodaran et al. (2018) all used the λ of the low joint error assumption (LJE(λ) in place of the entanglement terms appearing in our bounds. Our analysis, derived through the simple bound of Lemma 3.1 and the decomposition rules in 3.2, was not considered before in its generality.

Similar to Tachet des Combes et al. (2020), Kirchmeyer et al. (2022) also proposes an IW-based framework to handle both label shift *and* conditional shift. Yet unlike most UDA methods, Kirchmeyer et al. (2022) does not constrain the model to learn only invariant representations but trains a target classifier directly on transported source inputs. See also Shu et al. (2018) for an approach which considers adapting a domain invariant model on unlabeled target inputs (using pseudolabeling) assuming potential violations of the LJE(λ) assumption.

D   APPROXIMATE TRIANGLE INEQUALITY

In this section we study how we can relax the Assumption 3.1 while maintaining the essential properties of the proved bounds. A key property of Assumption 3.1 is the triangle inequality axiom. This is crucial for proving Lemma 3.1 without removing the occurrence of ℓ. In the latter case, removing ℓ is typically done via the Lipschitz continuity of ℓ: but this simply switches the notion of distance (originally defined by a semi-metric ℓ) to the underlying metric space of $\mathcal{X} \times \mathcal{Y}$ (which does have a triangle inequality property).

Many loss function violate the simple triangle inequality, but can satisfy a weaker condition. To weaken the metric assumption, one can consider an approximate triangle inequality.

Definition D.1 ($\varkappa$-Approximate Triangle Inequality (Crammer et al., 2008)). Let $d\colon \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a function. It satisfies the $\varkappa$-Approximate Triangle Inequality for $\varkappa \geq 1$ if,

$$d(z, z') \leq \varkappa \cdot (d(z, z'') + d(z'', z')), \tag{11}$$

for all $z, z', z'' \in \mathcal{Z}$.

Note that $\varkappa = 1$ reduces Definition D.1 to the usual triangle inequality.

In prior work, Wang and Mao (2023) explored the approximate triangle inequality in the context of domain adaptation. However, this was specifically for dealing with the LJE(λ) assumption.

We note that many of our results hold when utilizing the following weaker assumption.

Assumption D.1 ($\varkappa$-Almost-Metric Loss). The loss function ℓ is a semi-metric (i.e., Assumption 3.1 without the triangle inequality) and also satisfies Definition D.1 with $\varkappa$.

This immediately helps us to obtain a variant of our key Lemma 3.1.

Lemma D.1. *Suppose the loss function $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfies Assumption D.1 with $\varkappa \geq 1$. Then the target risk of a classifier $f\colon \mathcal{X} \to \mathcal{Y}$ is bounded by:*

$$\epsilon_q(f) \leq \varkappa^2 \cdot \epsilon_p(f) + \varkappa \cdot W_{1,\ell \circ f, \varkappa \ell}(p, q) \tag{12}$$

$$\overset{\text{(s)}}{\leq} \varkappa^2 \cdot \epsilon_p(f) + \varkappa \cdot W_{1,\ell,\varkappa \ell}(f \sharp p, f \sharp q), \tag{13}$$

where the inequality (s) *holds for surjective f. Additionally,* (s) *is an* equality *whenever f is invertible. We remind that $(f \sharp p)(\hat{y}, y) = p(f^{-1}(\hat{y}), y)$.*

Proof of Lemma D.1. The proof follows identically to Lemma 3.1, except for the initial application of the triangle inequality. Examine the following for $\gamma \in \Gamma(q, p)$,

$$
\begin{aligned}
\epsilon_q(f) &= \mathbb{E}_q[\ell(f(x), y)] \\
&= \varkappa^2 \cdot \mathbb{E}_p[\ell(f(x), y)] + \mathbb{E}_{(x,y),(x',y')\sim\gamma}\left[\ell(f(x'), y') - \varkappa^2 \cdot \ell(f(x), y)\right] \\
&\leq \varkappa^2 \cdot \mathbb{E}_p[\ell(f(x), y)] + \mathbb{E}_{(x,y),(x',y')\sim\gamma}\left[\varkappa \cdot \ell(f(x), f(x')) + \varkappa \cdot \ell(f(x), y') - \varkappa^2 \cdot \ell(f(x), y)\right] \\
&\leq \varkappa^2 \cdot \mathbb{E}_p[\ell(f(x), y)] + \mathbb{E}_{(x,y),(x',y')\sim\gamma}\left[\varkappa \cdot \ell(f(x), f(x')) + \varkappa^2 \cdot \ell(f(x), y) + \varkappa^2 \cdot \ell(y, y') - \varkappa^2 \cdot \ell(f(x), y)\right] \\
&= \varkappa^2 \cdot \mathbb{E}_p[\ell(f(x), y)] + \mathbb{E}_{(x,y),(x',y')\sim\gamma}\left[\varkappa \cdot \ell(f(x), f(x')) + \varkappa^2 \cdot \ell(y, y')\right] \\
&= \varkappa^2 \cdot \mathbb{E}_p[\ell(f(x), y)] + \varkappa \cdot \mathbb{E}_{(x,y),(x',y')\sim\gamma}\left[\ell(f(x), f(x')) + \varkappa \cdot \ell(y, y')\right].
\end{aligned}
$$

Then, with this change, the proof of Lemma 3.1 follows without change to yield the result. □

To further see how switching from Assumption 3.1 to Assumption D.1 affects our results, we also derive an equivalence to Corollary 3.1.

Corollary D.1. *Suppose the loss function $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfies Assumption D.1 with $\varkappa \geq 1$ and a classifier $f\colon \mathcal{X} \to \mathcal{Y}$ satisfies Assumption 3.3 then*

$$\epsilon_q(f) \leq \varkappa^2 \cdot \epsilon_p(f) + \varkappa \cdot W_{1,\ell}(f\sharp p_{\mathrm{x}}, f\sharp q_{\mathrm{x}}) + \varkappa^2 \cdot \mathcal{E}_{\mathrm{y}}(f), \tag{14}$$

$$\epsilon_q(f) \leq \varkappa^2 \cdot \epsilon_p(f) + \varkappa^2 \cdot W_{1,\ell}(p_{\mathrm{y}}, q_{\mathrm{y}}) + \varkappa \cdot \mathcal{E}_{\hat{\mathrm{y}}}(f). \tag{15}$$

Proof of Corollary D.1. Result follows immediately from Lemma D.1 and Lemma 3.2 whilst noticing that $W_{1,\varkappa\ell} = \varkappa \cdot W_{1,\ell}$. $\qquad\square$

Our remaining results follow similarly. Note that a consequence of the difference between Corollary 3.1 and Corollary D.1 is that the Oracle Upper Bound (OUB) $\mathcal{U}$ changes to

$$\mathcal{U}_{\varkappa}(f) \doteq \varkappa^2 \cdot \epsilon_p(f) + \varkappa \cdot W_{1,\ell}(f\sharp p_{\mathrm{x}}, f\sharp q_{\mathrm{x}}) + \varkappa^2 \cdot \mathcal{E}_{\mathrm{y}}(f). \tag{16}$$

Notice that as $\varkappa$ increases, so does the influence of the label-entanglement term with respect to the marginal distance. Also from an optimization perspective, the Wasserstein Regularized Risk (WRR) equivalent

$$\mathcal{R}_{\varkappa}(f) \doteq \varkappa^2 \cdot \epsilon_p(f) + \varkappa \cdot W_{1,\ell}(f\sharp p_{\mathrm{x}}, f\sharp q_{\mathrm{x}}), \tag{17}$$

now requires a scaling by $\varkappa$ depending on the $\varkappa$-approximate triangle inequality. This scaling is equivalent to setting the correct regularization constant when optimizing WRR.

E   ADDITIONAL REMARKS

Remark.   An interesting case of the bounds in Lemma 3.2 occurs when the marginal source and target distributions are identical. For example, if $p_{\mathrm{y}} = q_{\mathrm{y}}$, then the first term of Eq. (4) vanishes and $\gamma_{\mathrm{y}}^{\star}(y, y')$ is diagonal. As a result, the inequality further reduces to $W_{\alpha,c}(p, q) \leq \mathbb{E}_{y \sim p_{\mathrm{y}}}\left[W_{\alpha,c_2}^{\alpha}(p_{\mathrm{x}|\mathrm{y}}(\cdot \mid y), q_{\mathrm{x}|\mathrm{y}}(\cdot \mid y))\right]^{1/\alpha}$. Moreover, the results can be upper bounded further using any feasible coupling, including, e.g., the Cartesian coupling $(x, x') \mapsto p(x)q(x')$ as opposed to $\gamma_x^{\star}$ in (3).

Close Conditionals Assumption.   We give an explicit construction to show that CCA is a valid assumption for networks with large enough capacity. Assume that there's a network $f\colon \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}^n$ that satisfies the LJE assumption, i.e., $\lambda = \min_f \epsilon_p(f) + \epsilon_q(f)$ is low for such a network. We will compose the output of the network with a nonlinear mapping $g(\cdot)$ such that the resulting enlarged network $h = g \circ f$ satisfies CCA: $\kappa = \min_h \epsilon_p(h) + \max_y W_{1,\ell}(h\sharp p(x|y), h\sharp q(x|y))$ for some low κ.

Consider the function $g(\cdot)$ composed of mappings $g_i(\cdot)$ such that $g = g_m \circ g_{m-1} \ldots \circ g_1$, where for each $g_i = g_{in} \circ g_{i(n-1)} \ldots \circ g_{i1}$, $i \in \mathbb{Z}^m$, g_{ij} switches between affine operations

$$g_{ij}(x) = \begin{cases} A_{ij}x + b_{ij} & \text{if } x_j > x_k \text{ for all } k \neq j, \\ x & \text{else.} \end{cases}$$

Such a mapping can be implemented approximately using, e.g., a ReLU network that uses an *and* logic internally. The affine operations $A_{ij}x + b_{ij}$ can be defined such that each decreases the Wasserstein distance between the j-th conditionals sufficiently, while making sure that the conditionals are still linearly separable from the rest, e.g., a mapping that shrinks the two conditionals closer towards their mutual mean.

Relation to KL-divergence.   It is possible to relate Lemma 3.1 to the KL-divergence by: (1) upper-bounding $W_{1,\ell,\ell}$ via a Lipschitz continuity assumption on ℓ to get $W_{1,d}$, where d is the metric distance of $\mathcal{X} \times \mathcal{Y}$; (2) further bounding the Wasserstein distance with the total variation (TV) distance, taking into account the diameter of the space; and finally (3) using the Pinsker's inequality (Pinsker, 1964) or the Bretagnolle-Huber bound (Bretagnolle and Huber, 1979). For the last step, one can take the minimum of both inequalities as they do not strictly dominate each other (Canonne, 2022). A specific combination of these steps has been previously leveraged in Wang and Mao (2023).

In what follows, we present a version of Corollary 3.1 which switches from our Wasserstein distance to a KL-divergence. Instead of doing both (1) and (2) in the above, we consider the diameter of the space w.r.t. the specific metric used in the Wasserstein distance $W_{1,\ell}$.

Corollary E.1. *Suppose that the pre-conditions of Corollary 3.1 hold and that ℓ satisfies Assumption 3.2. Then*

$$\epsilon_q(f) \le \epsilon_p(f) + L \cdot \sqrt{\frac{1}{2} \cdot \min\{\mathrm{KL}(f\sharp p_{\mathrm{x}} \parallel f\sharp q_{\mathrm{x}}), \mathrm{KL}(f\sharp q_{\mathrm{x}} \parallel f\sharp p_{\mathrm{x}})\}} + \mathcal{E}_{\mathrm{y}}(f) \tag{18}$$

$$\epsilon_q(f) \le \epsilon_p(f) + L \cdot \sqrt{\frac{1}{2} \cdot \min\{\mathrm{KL}(p_{\mathrm{y}} \parallel q_{\mathrm{y}}), \mathrm{KL}(q_{\mathrm{y}} \parallel p_{\mathrm{y}})\}} + \mathcal{E}_{\hat{\mathrm{y}}}(f). \tag{19}$$

Proof. Follows immediately from bounding the diameter of $\mathcal{Y}$ w.r.t. the bounded loss ℓ and Pinsker's inequality. $\square$

One could generalize the above by replacing L with the diameter of $\mathcal{Y}$ under the metric ℓ. A practical consequence of the above corollary is that one could potentially switch the Wasserstein regularized risk (WRR) to use the above upper bound, i.e., modifying WRR to

$$\mathcal{R}_{\mathrm{KL}}(f) \doteq \epsilon_p(f) + L \cdot \sqrt{\frac{1}{2} \cdot \min\{\mathrm{KL}(f\sharp p_{\mathrm{x}} \parallel f\sharp q_{\mathrm{x}}), \mathrm{KL}(f\sharp q_{\mathrm{x}} \parallel f\sharp p_{\mathrm{x}})\}} \le \epsilon_p(f) + L \cdot \sqrt{\frac{1}{2} \cdot \mathrm{JS}(f\sharp p_{\mathrm{x}} \parallel f\sharp q_{\mathrm{x}})}, \tag{20}$$

presents an alternative divergence to minimize in practice. One could also extend the above corollary to define a KL-divergence version of the entanglement term. One should, however, note that KL-divergences evaluate to ∞ whenever the support of the distributions considered are disjoint.

Gaussians. We explore the bound (3) in Lemma 3.2 in the case of Gaussians with scaled covariances, using the Euclidean distance for ℓ and setting $\alpha = 2$. The arguments apply due to symmetry (by switching $y \leftrightarrow x$) to (4) as well.

Assume that $x, y \sim p = \mathcal{N}(\mu, \Sigma)$ and $x', y' \sim q = \mathcal{N}(\mu', \Sigma')$ where

$$\mu \doteq (\mu_x^{\mathrm{T}}, \mu_y^{\mathrm{T}})^{\mathrm{T}},$$

$$\Sigma \doteq \begin{pmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{pmatrix},$$

with corresponding notation for μ', Σ'. We first start by showing that whenever $\Sigma' = s^2\Sigma$ for some $s \in \mathbb{R}$, the Wasserstein-2 distances satisfy

$$W_{2,\ell}^2(p, q) = W_{2,\ell}^2(p_y, q_y) + \int \pi^\star(y, y') W_{2,\ell}^2(p_{x|y}(\cdot \mid y), q_{x|y}(\cdot \mid y')) \mathrm{d}(y, y'),$$

where $p(x, y)$ is decomposed as $p(x, y) = p_y(y) p_{x|y}(x|y)$ (sym. q) and $\pi^\star(y, y')$ is the optimal transport map between p_y, q_y. The Wasserstein-2 distance between the joint Gaussians p, q is given by (Peyré et al., 2019)

$$W_{2,\ell}^2(p, q) = \|\mu - \mu'\|^2 + \mathrm{tr}(\Sigma + \Sigma' - 2(\Sigma^{1/2}\Sigma'\Sigma^{1/2})^{1/2}).$$

Plugging in $\Sigma' = s^2\Sigma$ we get

$$W_{2,\ell}^2(p, q) = \|\mu - \mu'\|^2 + (s - 1)^2 \mathrm{tr}(\Sigma),$$

$$W_{2,\ell}^2(p_y, q_y) = \|\mu_y - \mu_y'\|^2 + (s - 1)^2 \mathrm{tr}(\Sigma_y),$$

where the distance between the marginals p_y, q_y has the optimal transport given by a deterministic mapping $T(y)$

$$\pi^\star(y, y') = \delta_{(y, T(y))},$$

$$T(y) \doteq \mu_y' + \Sigma_y^{-1/2} \left(\Sigma_y^{1/2}\Sigma_y'\Sigma_y^{1/2}\right)^{1/2} \Sigma_y^{-1/2}(y - \mu_y)$$

$$= \mu_y' + s(y - \mu_y).$$

Given that $\mu_{x|y} = \mu_x + \Sigma_{xy}\Sigma_y^{-1}(y - \mu_y)$ and $\Sigma_{x|y} = \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}$, the 2-Wasserstein distance between $p(x \mid y)$ and $q(x \mid y)$ is given by

$$W_{2,\ell}^2(p_{x|y}, q_{x|y}) = \| \left(\mu_x + \Sigma_{xy}\Sigma_y^{-1}(y - \mu_y)\right) - \left(\mu_x' + \Sigma_{xy}\Sigma_y^{-1}(y' - \mu_y')\right) \|^2 + (s - 1)^2 \mathrm{tr}(\Sigma_{x|y}).$$

Plugging into the integral above and using $T(y)$ we get

$$\int W_{2,\ell}^2(p_{x|y}(\cdot \mid y), q_{x|y}(\cdot \mid \mu_y' + s(y - \mu_y)))\mathrm{d}y$$

$$= \int \| \left(\mu_x + \Sigma_{xy}\Sigma_y^{-1}(y - \mu_y)\right) - \left(\mu_x' + \Sigma_{xy}\Sigma_y^{-1}(\mu_y' + s(y - \mu_y) - \mu_y')\right) \|^2 \mathrm{d}y + (s-1)^2 \mathrm{tr}(\Sigma_{x|y})$$

$$= \|\mu_x - \mu_x'\|^2 + 2(1-s)(\mu_x - \mu_x')^{\mathrm{T}}\Sigma_{xy}\Sigma_y^{-1} \int (y - \mu_y)\mathrm{d}y +$$

$$(s-1)^2 \mathrm{tr}\left((\Sigma_{xy}\Sigma_y^{-1})^{\mathrm{T}}(\Sigma_{xy}\Sigma_y^{-1}) \int (y - \mu_y)(y - \mu_y)^{\mathrm{T}}\mathrm{d}y\right) + (s-1)^2 \mathrm{tr}(\Sigma_{x|y})$$

$$= \|\mu_x - \mu_x'\|^2 + (s-1)^2 \mathrm{tr}(\Sigma_y^{-1}\Sigma_{yx}\Sigma_{xy} + \Sigma_{x|y}),$$

since $\int (y - \mu_y)\mathrm{d}y = 0$ and $\int (y - \mu_y)(y - \mu_y)^{\mathrm{T}}\mathrm{d}y = \Sigma_y$. Now since $\mathrm{tr}(ABC) = \mathrm{tr}(BCA)$ for any matrix triple A, B, C, $\mathrm{tr}(\Sigma_{x|y})$ can be written as $\mathrm{tr}(\Sigma_{x|y}) = \mathrm{tr}(\Sigma_x) - \mathrm{tr}(\Sigma_y^{-1}\Sigma_{yx}\Sigma_{xy})$. Plugging into the above and adding $W_2^2(p_y, q_y)$ we get

$$W_{2,\ell}^2(p_y, q_y) + \int \pi^*(y, y') W_{2,\ell}^2(p_{x|y}(\cdot \mid y), q_{x|y}(\cdot \mid y'))\mathrm{d}(y, y')$$

$$= \|\mu_y - \mu_y'\|^2 + (s-1)^2 \mathrm{tr}(\Sigma_y) + \|\mu_x - \mu_x'\|^2 + (s-1)^2 \mathrm{tr}(\Sigma_x)$$

$$= \|\mu - \mu'\|^2 + (s-1)^2 \mathrm{tr}(\Sigma)$$

$$= W_{2,\ell}^2(p, q),$$

which proves the above. Now we relate (by switching between the couplings) the joint Wasserstein distance in Lemma 3.2 that uses the composite cost $c(x, y, x', y') = \ell(x, x') + \ell(y, y')$ to $W_{2,\ell}^2(p, q)$

$$W_{2,\ell,\ell}^2(p, q) \le W_{2,\ell}^2(p, q) + 2 \int \ell(x, x')\ell(y, y')\pi^\star(x, y, x', y')\mathrm{d}((x, y), (x', y')),$$

where $\pi^\star(x, y, x', y')$ is the optimal map corresponding to $W_{2,\ell}(p, q)$ and is decomposed (due to $\Sigma' = s^2\Sigma$) as

$$\pi^\star(x, y, x', y') = \delta_{x, T_1(x)}\delta_{y, T_2(y)},$$
$$T_1(x) = \mu_x' + s(x - \mu_x),$$
$$T_2(y) = \mu_y' + s(y - \mu_y).$$

Hence the integral above can be written as

$$\int \ell(x, x')\ell(y, y')\pi^\star(x, y, x', y')\mathrm{d}((x, y), (x', y')) = \int \ell(x, T_1(x))\ell(y, T_2(y))\mathrm{d}(x, y).$$

Applying Cauchy-Schwartz inequality we see that finally

$$W_{2,\ell,\ell}^2(p, q) \le W_{2,\ell}^2(p_y, q_y) + \underset{\pi^\star(y,y')}{\mathbb{E}}[W_2^2(p_{x|y}(\cdot \mid y), q_{x|y}(\cdot \mid y'))] + 2W_{2,\ell}(p_y, q_y)\left(\underset{\pi^\star(y,y')}{\mathbb{E}}[W_2^2(p_{x|y}(\cdot \mid y), q_{x|y}(\cdot \mid y'))]\right)^{1/2}.$$

Taking square root of both sides we get the inequality (3) for $\alpha = 2$.

F EXPERIMENTAL DETAILS

In this section we discuss in more detail the experimental setup used to produce the results in Section 5. We generate accuracy plots over the epochs for the methods shown in Table 1, together with entanglement estimates, to support the conclusions that were drawn.

Scenarios. The scenarios considered are: MNIST to USPS, USPS to MNIST, MNIST to MNIST-M, SVHN to MNIST, CIFAR-10 corruptions and the Portraits dataset. MNIST and USPS are standard *digits* datasets, with handwritten digit styles showing significant variations between them. SVHN is a more challenging dataset

composed of colored street view images of house numbers. MNIST-M is a dataset that is causally related to MNIST: it is created by changing the background of MNIST inputs with colored images from another dataset, see Ganin et al. (2016) for details. Portraits (Ginosar et al., 2015) is a set of yearbook images from early 1900s to early 2000s, in particular, it is sometimes presented as a dataset suitable for *gradual* domain adaptation (Kumar et al., 2020)[8]. In our case the source dataset uses labeled images of girls and boys until the year 1974 (first half of the dataset) and the remaining half of the dataset (from 1974 till 2013) is used as the target. Hairstyles and lighting conditions, in particular, undergo significant change over the years. Since the Portraits dataset has only two classes, it is a convenient dataset to use for visualizing entanglement, see Figure 1. Finally CIFAR-10 corruptions (CIFAR-10c) are created using different noise models, such as changes in illumination or motion blur (Hendrycks and Dietterich, 2019). In our case we report the average result of the three most difficult (in terms of degradation of ERM accuracy) noise models, which are fog, frost and snow-based corruption models.

For each scenario, we utilize the default train-test splits for source and target datasets if provided; otherwise we split the data using %80/20 split ratio. When the input dimensions of the datasets do not match in a distribution shift scenario, we either downsample (e.g., in the case of MNIST $\rightarrow$ USPS, we resize the MNIST images to 16×16) or upsample (e.g., in the case of SVHN $\rightarrow$ MNIST, we resize the grayscale MNIST images to $32 \times 32 \times 3$). For the Portraits dataset, we resize the images down to 32×32[9].

Models. We run each UDA scenario on different models of varying complexity, in order to get a better picture of the dependency of entanglement on the model capacity and type. The first model that we test is a two hidden layer multilayer perceptron (MLP) with hidden layer neurons of size $(200, 100)$, referred to as *MLP*. The rest are convolutional neural networks (CNN) with various depth sizes and regularizations: *small_cnn* is a CNN with two convolutional layers (with ReLU nonlinearity and max pooling) followed by two fully connected layers, *LeNet* is a CNN with two convolutional layers using average pooling followed by three fully connected layers. The last two bigger networks used are referred to as *conv1* and *conv2*: *conv1* is a deeper CNN than the others (six convolutional layers followed by two fully-connected layers) and *conv2* (two convolutional layers followed by three fully-connected layers) has additional batch-norm and dropout regularization between the layers. The model *conv2* was used to visualize the effect of entanglement during UDA in Figure 1.

Due to space constraints, in Table 1, we present only the final source and target accuracies at the last epoch from the best model (*conv1*) and choose another randomly (different for each scenario) from the models mentioned above that satisfy LJE oracle accuracy above 90% for both the source and target domains (except for the CIFAR-10 corruptions scenario). Fixing the final epoch can be misleading as an indicator of performance, as the accuracy curves tend to show significant oscillatory behaviour, which gives useful indication about the quality of the marginal alignment and the effect of the entanglement. Hence, in addition to the results presented in Table 1, in this section we plot the resulting accuracy curves over the epochs as well.

Methods. There are five methods presented in Table 1. The LJE *Oracle* trains the model on both source *and* target labels and is provided here as a reference to check whether the LJE assumption holds to some degree. *ERM* trains the model only on source labels, and similarly is provided as a reference to check for *negative transfer* (i.e., the potential loss in accuracy after adapting to the target input distribution). *DANN* (Ganin et al., 2016) is an adversarial UDA approach that uses an additional *discriminator* network (a two-layer CNN with batch-norm) to minimize the divergence between the source and target. DANN and related adversarial methods can be shown to approximately minimize a Jensen-Shannon divergence (Goodfellow et al., 2014). *WRR* minimizes the regularized source risk presented in 4.1, without any adjustments such as scaling. *JDOT* (Courty et al., 2017; Damodaran et al., 2018) is a method that adds a feature space Wasserstein distance minimization (with multiplier $\alpha = 0.001$) in addition WRR, which potentially trades off entanglement for a higher marginal distance. As we see in the experiments, this strategy to constrain the network Lipschitz bound and hence the entanglement, is sometimes effective (when compared to the vanilla WRR) but not always. WRR and JDOT both minimize cross-entropy as the loss function in Table 1 but JDOT scales it by a factor of 0.001.

In Table 2 we present an additional oracle that trains the model by optimizing the CC objective function:

---

[8]However preliminary experiments conducted using different methods, including pseudolabeling typically used in gradual domain adaptation, see no benefit in adapting to images gradually according to year, compared to domain matching that ignores the gradual shift.

[9]We also tried domain adaptation with the original sizes, however, the source and target accuracies were significantly lower in that case for the models that we use in this paper.

$\min_f \epsilon_p(f) + \max_{y \in \mathcal{Y}} W_{1,\ell}(f \sharp p_{x|y}(\cdot \mid y), f \sharp q_{x|y}(\cdot \mid y))$. Although the max operator is not differentiable, Pytorch keeps track of the index of the conditionals that obtain the maximum distance and differentiates through it, which corresponds to a subdifferential.

Libraries. We use the Pytorch library (Paszke et al., 2019) to optimize the Deep Neural Network models and for the implemention of the methods mentioned above. For the OT computations in WRR and JDOT, we use the GeomLoss library (Feydy et al., 2019) which computes the entropy-regularized Wasserstein distance (i.e., the Sinkhorn divergence (Peyré et al., 2019)), where we set the regularization to the default value of 0.05. We use the 2-Wasserstein distance rather than the 1-Wasserstein distance for computational reasons, although the difference is generally small. To estimate the entanglement term in Table 2 we use the POT library (Flamary et al., 2021), as we need the optimal transport map $\gamma^\star(x, x')$ available as an output (GeomLoss makes available only the value and not the mapping).

Plots. In Figures 2 to 4 we pick three scenarios and models to present target accuracy errorbar plots for each method shown in Table 1. We present mean +/- standard deviations over three different seeds, over 10 epochs and plot on the right hand side the entanglement estimates of each method. The entanglement estimates are averaged over the mini-batches. It can be seen that the entanglement estimates have a strong negative correlation with the method's target accuracies. The ordering of the accuracies is also generally reversed in the case of entanglement.

In Figures 5 to 7 we pick three scenarios and models to plot the WRR performances using the Euclidean distance as the loss function. These scenario/model cases correspond to the results shown in Table 2. The entanglement estimates together with the WRR values are plotted together with the target loss. The WRR values together with the entanglement estimates upper bound the target loss for all cases considered. Although we have not considered the optimization landscape of WRR in this article, results suggest that the WRR objective function is not minimized effectively.
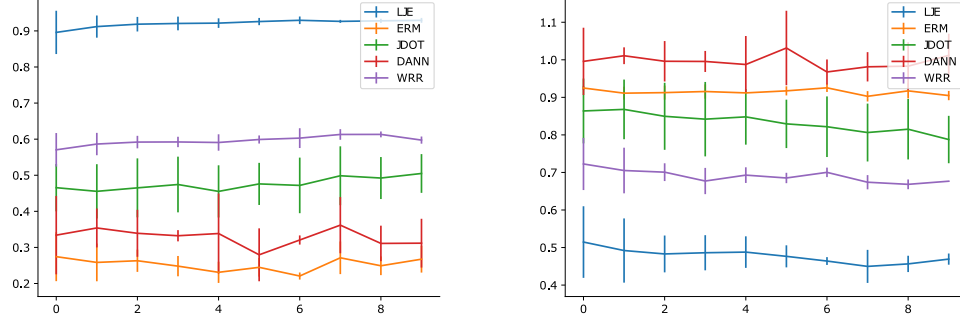
Figure 2: Target accuracy plots for the MNIST → USPS scenario shown in Table 1 using the model MLP. Corresponding entanglement estimates are shown on the right hand figure.
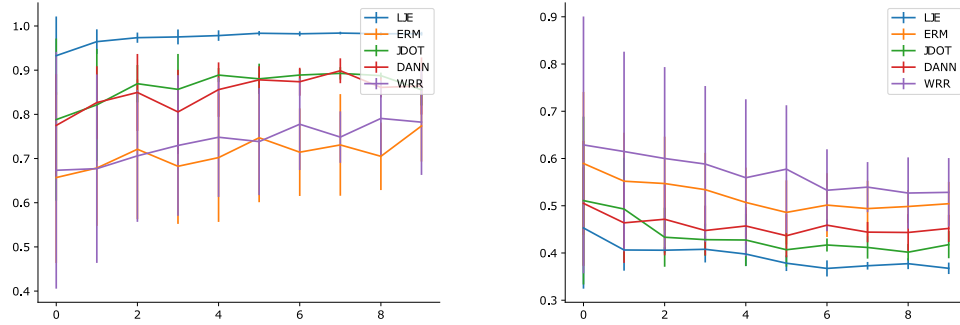

Figure 3: Target accuracy plots for the USPS → MNIST scenario shown in Table 1 using the model conv1. Corresponding entanglement estimates are shown on the right hand figure.
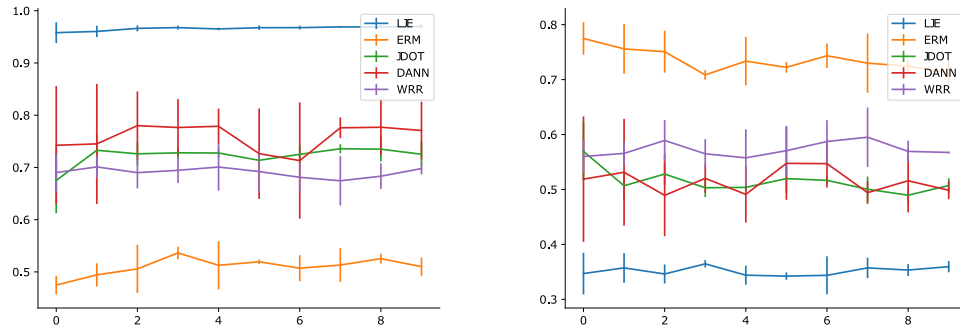

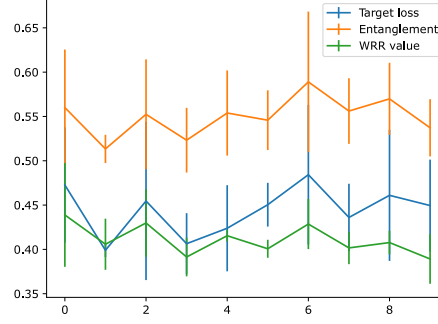Figure 4: Target accuracy plots for the MNIST → MNIST-M scenario shown in Table 1 using the model conv1. Corresponding entanglement estimates are shown on the right hand figure.

Figure 5: Target loss over the epochs for the MNIST → MNIST-M scenario shown in Table 2 using the model conv1. Corresponding entanglement estimates as well as the WRR values are included for comparison.
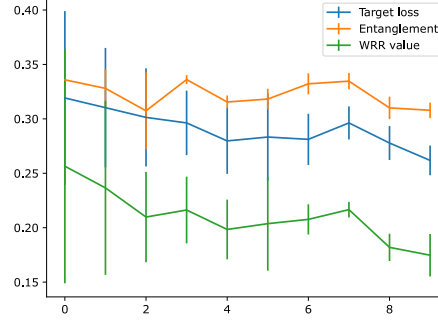
Figure 6: Target loss over the epochs for the Portraits dataset shown in Table 2 using the model MLP. Corresponding entanglement estimates as well as the WRR values are included for comparison.
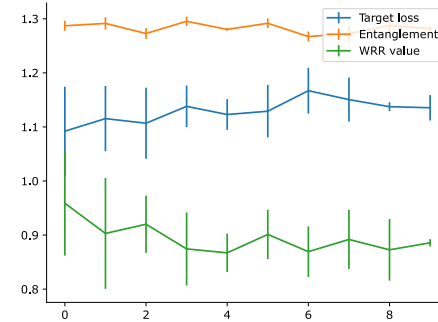
Figure 7: Target loss over the epochs for the CIFAR10 → CIFAR10c scenario shown in Table 2 using the model conv1. Corresponding entanglement estimates as well as the WRR values are included for comparison.

G ABLATIONS

In this article, we highlighted the (model, dataset, optimizer) dependency of entanglement. This triple dependency has severe consequences for UDA: changes in the target accuracy can be attributed to hyperparameters, model selection, other algorithmic choices or any preprocessing during the dataset creation (e.g., centering, label shift removal). In this section, we select several interesting ablations that highlight this manifold dependency.

Batch size. We explore the change in the performance of the UDA methods as the batch size increases from 64 to 512 in Figures 8 and 9. In the USPS to MNIST scenario, with a batch size of 64, WRR shows a high-variance but higher accuracy performance, whereas increasing the batch size to 512 makes the WRR performance decrease significantly. We attribute this decrease to the fact the scaling is set as 1. JDOT with the scaling set to 0.001 does significantly better.

Training for more epochs. We find that training for more epochs can sometimes change the ranking of the methods in terms of accuracy. In Figure 10 showing the USPS $\to$ MNIST scenario, we see that the performance of WRR worsens, potentially due to the high variance discussed above.

Effect of the loss function. As noted in the discussion for Table 2, to be able to relate the entanglement estimates and the WRR values during the UDA optimization to our upper bound, we switched the loss function from cross-entropy to the Euclidean distance. We explore here in more detail the change in target accuracy resulting from this switch. Table 3 indicates that in several cases there can be significant gains in accuracy. We present the results at the fifth epoch.

Effect of label shift. As noted in section 5, label shift deteriorates performance (potentially by increasing label-entanglement). We explore here the effect further by artificially removing label shift from two datasets. We implement class balanced sampling in the MNIST $\to$ USPS and SVHN $\to$ MNIST cases below in Table 4 using a weighted sampler in PyTorch. We present the results at the fifth epoch. It can be seen that the entanglement estimates decrease significantly and as a result the target accuracies are significantly higher. The effect is more pronounced for the SVHN $\to$ MNIST case, where there is a label shift (in Wasserstein-1 distance) of 0.2, as opposed to 0.13 in the MNIST $\to$ USPS case.

W_1 vs W_2 distance minimization. We used W_2 distance minimization in all the experiments, as we suspected that the WRR objective function may not be easy to differentiate for $\alpha \neq 2$, e.g., when $\alpha = 1$. We investigate this claim numerically by switching between the W_2 and W_1 minimization of the WRR objective function in Table 5. WRR uses the Euclidean distance as the loss function and the accuracies are evaluated at the fifth epoch of training. From the results it is not possible to claim that one is strictly better than the other: although there is some significant gain in accuracy in the Portraits dataset and in the MNIST $\to$ USPS scenario of using $\alpha = 2$, in the USPS $\to$ MNIST scenario, we fail to see a significant gain.

Additional methods. Besides the methods shown in Table 1, we also evaluated two more approaches: one based on *MMD* (see, e.g., Li et al. (2018)) and a more recent approach called *Reverse-KL* (Nguyen et al., 2022). We report in Table 6 the mean source/target accuracies obtained with the best model (conv1).

We included the ERM performance in Table 6 as well, in order to reflect on the cases of positive and negative transfer. Note that MMD does not show any cases of positive transfer. Our implementation of MMD was based on the DomainBed (Gulrajani and Lopez-Paz, 2020) repository. We attribute the uniformly poor performance to either (1) bad specification of kernel parameters or (2) our choice of the feature space.

As for the Reverse-KL algorithm (our implementation[10] is a modification of the code in the Reverse-KL repository), note the uneven performance: Reverse-KL replaces DANN and JDOT methods as the best method in the scenarios MNIST $\to$ USPS, USPS $\to$ MNIST and Portraits, yet drops down to chance levels in others. We attribute this to our choice of hyperparameters ($\beta = 0.1$ and $\beta_{\mathrm{aux}} = 0.1$ in Nguyen et al. (2022)): by varying them we can improve in the scenarios with poor results but then the accuracies in the above mentioned scenarios get significantly worse. Note that, similarly for JDOT (Damodaran et al., 2018), we observe much better results

---

[10]Note that Reverse-KL (Nguyen et al., 2022) implementation uses *probabilistic representation networks*, hence our models in Table 1 had to be adapted in order to run the algorithm.

by changing hyperparameters between experiments. This excessive hyperparameter-dependency is an instance of the (model, optimizer, dataset) triplet dependency of the entanglement that we tried to highlight in this article.

Additional dataset and model. We looked at an additional dataset called *Office-Home* (Venkateswara et al., 2017) which consists of four different subdatasets (Art, Clipart, Product and Real World) with corresponding domain adaptation scenarios (e.g., Art → Clipart). The models that we tried in Table 1 are unable to get high accuracy in these challenging datasets (e.g., LJE and CCA accuracies are very low) due to the high resolution of the images, small dataset size and wide distribution shifts observed between the subdatasets, among other potential difficulties. As a remedy, we used an ImageNet-pretrained *ResNet18* He et al. (2016) as our model (with the last layer suitably changed to output 65 classes). We used the same settings as reported in the experiments, except for lowering the Adam learning rate to 1×10^{-4}. We provide in Table 7 some preliminary results on the Product → Real World scenario (denoted as Pr → Rw).

From our preliminary results, it seems that JDOT is the best performing algorithm here (and the only one that shows positive transfer, comparing to ERM target accuracy). Interestingly WRR, which can be seen as a bare-bones version of JDOT (without the intermediate feature space alignment and without the scaling) performs poorly here, which highlights again the dependence on hyperparameter selection, as part of the optimizer component of entanglement's (model, optimizer, dataset) triplet dependency.

Table 3: The effect of switching from cross entropy to the Euclidean distance as the WRR loss function.

Scenario / Model	Cross-entropy Acc.	Euclidean distance Acc.
MNIST → USPS / small_cnn	86.8 ± 1.0	94.7 ± 1.1
USPS → MNIST / LeNet	72.6 ± 17.3	93.5 ± 2.0
SVHN → MNIST / small_cnn	41.2 ± 0.4	57.3 ± 10.7
MNIST → MNIST-M / MLP	12.2 ± 1.1	16.7 ± 1.4

Table 4: The effect of class-balanced sampling during the WRR optimization.

Scenario / Model	$W_1(\hat{p}_y, \hat{q}_y)$	Unbalanced Acc. / Entanglement	Balanced Acc. / Entanglement
MNIST → USPS / LeNet	0.13 / 0.0	84.7 ± 2.7 / 0.45 ± 0.02	91.4 ± 0.7 / 0.35 ± 0.01
SVHN → MNIST / conv1	0.2 / 0.0	83.3 ± 7.7 / 0.51 ± 0.06	95.3 ± 0.83 / 0.38 ± 0.01

Table 5: The effect of switching from W_2 to W_1.

Scenario / Model	W_2-minimization Acc.	W_1-minimization Acc.
MNIST → USPS / small_cnn	94.7 ± 1.1	92.3 ± 0.29
USPS → MNIST / LeNet	95.3 ± 0.4	93.5 ± 2.0
Portraits 1905–1974 → 1974–2013 / conv2	84.07 ± 1.05	80.0 ± 1.0

Table 6: The source/target accuracy of two other UDA methods (with ERM performance added as a baseline).

Scenario / Model	ERM	MMD	Reverse-KL
MNIST → USPS / conv1	98.0 / 83.8	98.5 / 80.7	98.9 / 90.2
USPS → MNIST / conv1	98.9 / 83.8	95.7 / 48.2	96.4 / 97.3
MNIST → MNIST-M / conv1	98.9 / 47.5	98.7 / 44.0	97.7 / 10.3
SVHN → MNIST / conv1	91.2 / 64.9	89.0 / 61.3	19.4 / 11.2
Portraits / conv1	96.8 / 86.3	95.5 / 79.1	96.7 / 87.7
CIFAR10 → CIFAR10c / conv1	75.5 / 37.7	62.6 / 30.9	21.1 / 10.7

Table 7: The source/target accuracy of different methods using an ImageNet-pretrained ResNet18 model on the Office-Home dataset, Product → Real World distribution shift scenario.

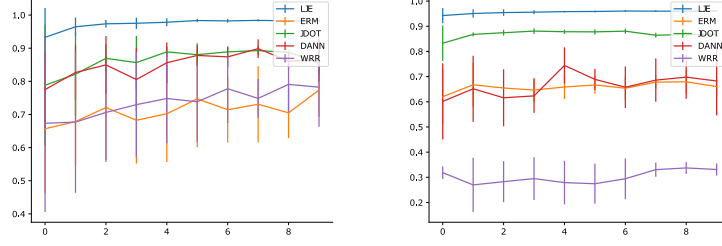| Scenario / Model	LJE	ERM	WRR	JDOT	DANN	MMD
Office-Home Pr → Rw	84.4 / 76.9	86.5 / 54.0	2.8 / 1.7	88.5 / 63.9	84.4 / 48.0	84.1 / 53.7

Figure 8: Target accuracy plots for the USPS → MNIST scenario shown in Table 1 using the model conv1. Left/right hand sides correspond to batch sizes 64 and 512 respectively.
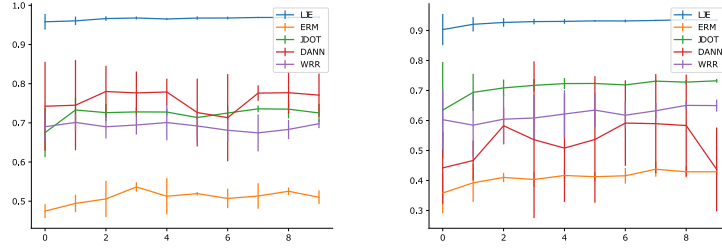

Figure 9: Target accuracy plots for the MNIST → MNIST-M scenario shown in Table 1 using the model conv1. Left/right hand sides correspond to batch sizes 64 and 512 respectively.
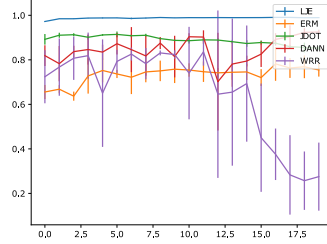

Figure 10: A case in the USPS → MNIST scenario for the conv1 model where increasing the number of epochs worsens the accuracy of the methods. All hyperparameter settings and model configurations are the same as in Table 1 except for the number of epochs.