f-Domain-Adversarial Learning: Theory and Algorithms

David Acuna [1 2 3] **Guojun Zhang** [4 3] **Marc T. Law** [1] **Sanja Fidler** [1 2 3]

Abstract

Unsupervised domain adaptation is used in many machine learning applications where, during training, a model has access to unlabeled data in the target domain, and a related labeled dataset. In this paper, we introduce a novel and general domain-adversarial framework. Specifically, we derive a novel generalization bound for domain adaptation that exploits a new measure of discrepancy between distributions based on a variational characterization of f-divergences. It recovers the theoretical results from Ben-David et al. (2010a) as a special case, and supports divergences used in practice. Based on this bound, we derive a new algorithmic framework that introduces a key correction in the original adversarial training method of Ganin et al. (2016). We show that many regularizers and ad-hoc objectives introduced over the last years in this framework are then not required to achieve performance comparable to (if not better than) state-of-the-art domain-adversarial methods. Experimental analysis conducted on real world natural language and computer vision datasets show that our framework outperforms existing baselines, and obtains the best results for f-divergences that were not considered previously in domain-adversarial learning.

1. Introduction

The ability to learn new concepts from general-purpose data and transfer them to related but different contexts is critical in many modern applications. One such prominent scenario is called *unsupervised domain adaptation*. In domain adaptation, the learner has access to both a small (unlabeled) dataset on its domain of interest, and to a larger labeled dataset on a domain related to the target domain but with different distribution. The model is trained with both the labeled and unlabeled datasets, and it is expected to generalize well to the target dataset if the gap between both domains is not very significant.

The paramount importance of domain adaptation (DA) has led to remarkable advances in the field. From a theoretical point of view, (Ben-David et al., 2007; 2010a;b; Mansour et al., 2009) provided generalization bounds for unsupervised DA based on discrepancy measures that are a reduction of the Total Variation (TV). Zhang et al. (2019) recently proposed the Margin Disparity Discrepancy (MDD) with the aim of closing the gap between theory and algorithms. Their notion of discrepancy is tailored to margin losses and builds on the observation of only taking a single supremum over the class set to make optimization easier. Theories based on weighted combination of hypotheses for multiple source DA have also been developed (Hoffman et al., 2018a).

From an algorithmic perspective in the context of neural networks, Ganin & Lempitsky (2015); Ganin et al. (2016) proposed the idea of learning domain-invariant representations as an adversarial game. This approach led to a plethora of methods including state-of-the-art approaches such as Shu et al. (2018); Long et al. (2018); Hoffman et al. (2018b); Zhang et al. (2019). Although these methods were explained with insights from the theory of Ben-David et al. (2010a), and more recently through MDD (Zhang et al., 2019), both the $\mathcal{H}\Delta\mathcal{H}$ divergence (Ben-David et al., 2010a) and MDD are hard to optimize with deep neural networks. *Ad-hoc* objectives have thus been introduced to minimize the divergence between the source and target distributions in a common representation space. This has led to a disconnect between theory and the current SoTA practical methods. Specifically, the domain-classifier from Ganin et al. (2016) that gives rise to domain-adversarial training methods is inspired by the proxy $\mathcal{A}$-distance from Ben-David et al. (2007) which itself is an approximation of the empirical estimation of the $\mathcal{H}\Delta\mathcal{H}$-divergence. It has been shown however that the discrepancy being minimized in practice in this framework corresponds to the JS-divergence (Ganin & Lempitsky, 2015). Nonetheless, to the best of our knowledge, no clear connection between the DA theory and the algorithms that are typically employed has been made, i.e. generalization bounds for DA with f-divergences have not been derived.

Contributions. In this paper, we derive a more general do-

[1]NVIDIA [2]University of Toronto [3]Vector Institute [4]University of Waterloo. Correspondence to: David Acuna <davidj@cs.toronto.edu,dacunamarrer@nvidia.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

main adaptation generalization bound based on a variational characterization of f-divergences. These allow us to clearly connect domain-adversarial training methods with the domain adaptation theory from an f-divergence minimization perspective. The theoretical results from Ben-David et al. (2010a) can be seen as a special case of our work for a specific choice of divergence. For the Jensen-Shannon (JS) divergence, we show how to rectify the original domain-adversarial training method from Ganin et al. (2016). Our analysis shows that after a key correction, many regularizers and ad-hoc objectives introduced in the DANN framework are not required to achieve performance comparable to (if not better than) state-of-the-art unsupervised domain adaptation methods that rely on adversarial learning. We also study how learning invariant representations for different choices of divergence affects the transfer performance on real-world datasets. In particular, the choice of the Pearson χ^2 divergence is sufficient to outperform previous methods without additional techniques and/or additional hyperparameters.

2. Preliminaries

In this paper, we focus on the unsupervised domain adaptation task. During training, we assume that the learner has access to a source dataset of n_s *labeled* examples $\mathrm{S} = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$, and a target dataset of n_t *unlabeled* examples $\mathrm{T} = \{(x_i^t)\}_{i=1}^{n_t}$, where the source datapoints x_i^s are sampled i.i.d. from a distribution P_{s} (source distribution) over the input space $\mathcal{X}$ and the target inputs x_i^t are sampled i.i.d. from a distribution P_{t} (target distribution) over $\mathcal{X}$. Usually, in the case of binary classification, we have $\mathcal{Y} = \{0, 1\}$ and in the multiclass classification scenario, $\mathcal{Y} = \{1, ..., k\}$. When the definition of $\mathcal{X}$ or $\mathcal{Y}$ cannot be inferred from the context, we will mention it explicitly.

We denote a labeling function as $f : \mathcal{X} \to \mathcal{Y}$, and use indices f_s and f_t to refer to the source and target labeling functions, respectively. The task of unsupervised domain adaptation is to find a hypothesis function $h : \mathcal{X} \to \mathcal{Y}$ that generalizes to the target dataset T (i.e., to make as few errors as possible by comparing with the ground truth label $f_t(x_i^t)$). The risk of a hypothesis h w.r.t. the labeling function f, using a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ under distribution $\mathcal{D}$ is defined as: $R_{\mathcal{D}}^{\ell}(h, f) := \mathbb{E}_{x\sim\mathcal{D}}[\ell(h(x), f(x))]$. We also assume that ℓ satisfies the triangle inequality. For simplicity of notation, we define $R_S^{\ell}(h) := R_{P_s}^{\ell}(h, f_s)$ and $R_T^{\ell}(h) := R_{P_t}^{\ell}(h, f_t)$ where the indices S and T refer to the source and target domains, respectively. In the stochastic scenario, we let the labeling function be the optimal Bayes classifier i.e $f(x) = \operatorname{argmax}_{\hat{y}\in\mathcal{Y}} P(y = \hat{y}|x)$ (Mohri et al., 2018). $P(y|x)$ denotes the class conditional distribution for either the source ($P_s(y|x)$) or the target domain ($P_t(y|x)$), respectively. The empirical risks over the source dataset S and the target dataset T are denoted by $\hat{R}_S$ and $\hat{R}_T$.

Comparing domains with f-divergences. A key component of domain adaptation is to study the discrepancy between the source and target distributions. In our work, we define new discrepancies between source and target distributions based on the variational characterization of popular choices of f-divergences. Thus, we start by providing the definition of f-divergences.

Definition 1 **(f-divergence, Csiszár (1967); Ali & Silvey (1966)).** *Let P_s and P_t be two distribution functions with densities p_s and p_t, respectively. Let p_s be absolutely continuous w.r.t p_t and both be absolutely continuous with respect to a base measure dx. Let $\phi : \mathbb{R}_+ \to \mathbb{R}$ be a convex, lower semi-continuous function that satisfies $\phi(1) = 0$. The f-divergence D_ϕ is defined as:*

$$D_\phi(P_s||P_t) = \int p_t(x)\, \phi\left(\frac{p_s(x)}{p_t(x)}\right) dx. \tag{2.1}$$

Variational characterization of f-divergences. Nguyen et al. (2010) derive a general variational method that estimates f-divergences from samples by turning the estimation problem into variational optimization. They show that any f-divergence can be written as (see details in Appendix A):

$$D_\phi(P_{\mathrm{s}}||P_{\mathrm{t}}) \geq \sup_{T\in\mathcal{T}} \mathbb{E}_{x\sim P_{\mathrm{s}}}[T(x)] - \mathbb{E}_{x\sim P_{\mathrm{t}}}[\phi^*(T(x))] \tag{2.2}$$

where ϕ^* is the (Fenchel) conjugate function of $\phi : \mathbb{R}_+ \to \mathbb{R}$ defined as $\phi^*(y) := \sup_{x\in\mathbb{R}_+}\{xy - \phi(x)\}$, and $T : \mathcal{X} \to \operatorname{dom}\phi^*$. The equality holds if $\mathcal{T}$ is the set of all measurable functions. Many popular divergences that are heavily used in machine learning and information theory are special cases of f-divergences. We summarize them and their conjugate function in Table 1. For simplicity, we assume in the following that $\mathcal{X} \subseteq \mathbb{R}^n$ and each density (i.e p_{s} and p_{t}) is absolutely continuous.

3. Discrepancies and Generalization Bounds

Domain adaptation bounds generally build upon the idea of bounding the gap between the source and target domains' error functions in terms of the discrepancy between their probability distributions. We first remind the reader of the seminal work of Ben-David et al. (2010a) that bounds the risk of any binary classifier in the hypothesis class $\mathcal{H}$ with the following theorem:

Theorem 1. *If $\ell(x, y) = |h(x) - y|$ and $\mathcal{H}$ is a class of functions, then for any $h \in \mathcal{H}$ we have:*

$$\begin{aligned} R_T^{\ell}(h) &\leq R_S^{\ell}(h) + D_{\mathrm{TV}}(P_s\|P_t) \\ &\quad + \min\{\mathbb{E}_{x\sim P_s}[|f_t(x) - f_s(x)|], \mathbb{E}_{x\sim P_t}[|f_t(x) - f_s(x)|]\}. \end{aligned} \tag{3.1}$$

Here,

$$D_{\mathrm{TV}}(P_{\mathrm{s}}\|P_{\mathrm{t}}) := \sup_{T\in\mathcal{T}} |\mathbb{E}_{x\sim P_{\mathrm{s}}}[T(x)] - \mathbb{E}_{x\sim P_{\mathrm{t}}}[T(x)]|$$

Table 1. Popular f-divergences, their conjugate functions and choices of a.

Divergence	$\phi(x)$	Conjugate $\phi^*(t)$	$\phi'(1)$	Activation func. $a(x)$
Kullback-Leibler (KL)	$x \log x$	$\exp(t-1)$	1	x
Reverse KL (KL-rev)	$-\log x$	$-1-\log(-t)$	-1	$-\exp x$
Jensen-Shannon (JS)	$-(x+1)\log\frac{1+x}{2} + x\log x$	$-\log(2-e^t)$	0	$\log\frac{2}{1+\exp(-x)}$
Pearson χ^2	$(x-1)^2$	$t^2/4+t$	0	x
Total Variation (TV)	$\frac{1}{2}\lvert x-1\rvert$	$\mathbf{1}_{-1/2\leq t\leq 1/2}$	$[-1/2, 1/2]$	$\frac{1}{2}\tanh x$

is the TV and $\mathcal{T}$ is the set of measurable functions. TV is an f-divergence such that $\phi(x) = |x-1|$ in Definition 1. For any function $\phi(x) \geq |x-1|$, one can replace $D_{\mathrm{TV}}(P_{\mathrm{s}}\|P_{\mathrm{t}})$ in Eq. (3.1) with $D_\phi(P_{\mathrm{s}}\|P_{\mathrm{t}})$. Theorem 1 thus bounds a classifiers target error in terms of the source error, the divergence between the two domains, and the dissimilarity of the labeling functions. Unfortunately, $D_{\mathrm{TV}}(P_{\mathrm{s}}\|P_{\mathrm{t}})$ cannot be estimated from finite samples of arbitrary distributions (Kifer et al., 2004). It is also a very loose upper bound as it involves the supremum over all measurable functions and does not account for the hypothesis class.

3.1. Measuring discrepancy with f-divergences

In the previous section, we have shown that measuring the similarity between P_{s} and P_{t} is critical in the derivation of generalization bounds and/or the design of algorithms. We now introduce a new discrepancy called $\mathrm{D}^\phi_{\mathcal{H}}$ that aims to generalize previous results to the family of f-divergences while solving the two aforementioned problems, namely **(1)** estimation of the divergence from finite samples of arbitrary distributions (Lemma 2) and **(2)** restriction of the discrepancy to the set including the hypothesis class $\mathcal{H}$. (Defs. 2 and 3). In Section 3.2 we show how this allows us to extend the bounds studied in Ben-David et al. (2010a).

Definition 2 ($\mathrm{D}^\phi_{\mathcal{H}}$ discrepancy). *Let ϕ^* be the Fenchel conjugate of a convex, lower semi-continuous function ϕ that satisfies $\phi(1)=0$, and let $\hat{\mathcal{T}}$ be a set of measurable functions such that $\hat{\mathcal{T}} = \{\ell(h(x), h'(x)) : h, h' \in \mathcal{H}\}$. We define the discrepancy between P_s and P_t as:*

$$D^\phi_{\mathcal{H}}(P_s||P_t) := \sup_{h,h'\in\mathcal{H}} |\mathbb{E}_{x\sim P_s}[\ell(h(x),h'(x))] - \mathbb{E}_{x\sim P_t}[\phi^*(\ell(h(x),h'(x)))]|. \tag{3.2}$$

The $\mathrm{D}^\phi_{\mathcal{H}}$ discrepancy can be interpreted as a lower bound estimator of a general class of f-divergences (Lemma 1). Therefore, for any hypothesis class $\mathcal{H}$ and choice of ϕ, $\mathrm{D}^\phi_{\mathcal{H}}$ is never larger than its corresponding f-divergence. In Lemma 2 we show that its computation can be bounded in terms of finite examples. Finally, we recover the $\mathcal{H}\Delta\mathcal{H}$-divergence (Ben-David et al., 2010a) if we consider $\phi^*(t) = t$ and $\ell(h(x), h'(x)) = \mathbf{1}[h(x) \neq h'(x)]$, which is the TV.

Definition 3 ($\mathrm{D}^\phi_{h,\mathcal{H}}$ discrepancy). *Under the same conditions as above, the discrepancy between two distributions P_s and P_t is defined by:*

$$D^\phi_{h,\mathcal{H}}(P_s||P_t) := \sup_{h'\in\mathcal{H}} |\mathbb{E}_{x\sim P_s}[\ell(h(x),h'(x))] - \mathbb{E}_{x\sim P_t}[\phi^*(\ell(h(x),h'(x)))]|. \tag{3.3}$$

Taking the supremum of $\mathrm{D}^\phi_{h,\mathcal{H}}$ over $h \in \mathcal{H}$, we obtain $\mathrm{D}^\phi_{\mathcal{H}}$, and thus $\mathrm{D}^\phi_{h,\mathcal{H}}(P_{\mathrm{s}}\|P_{\mathrm{t}}) \leq \mathrm{D}^\phi_{\mathcal{H}}(P_{\mathrm{s}}\|P_{\mathrm{t}})$. This bound will be useful when deriving practical algorithms.

Lemma 1 (lower bound). *For any two functions h, h' in $\mathcal{H}$, we have:*

$$|R^\ell_S(h,h') - R^{\phi^*\circ\ell}_T(h,h')| \leq D^\phi_{h,\mathcal{H}}(P_s||P_t) \leq D^\phi_{\mathcal{H}}(P_s||P_t) \leq D_\phi(P_s||P_t). \tag{3.4}$$

Lemma 1 is fundamental in the derivation of divergence-based generalization bounds for DA. Specifically, it bounds the gap between the source and target domains' error functions in terms of the discrepancy between their distributions using f-divergences. We now show that the $\mathrm{D}^\phi_{h,\mathcal{H}}$ can be estimated from finite samples.

Lemma 2. *Suppose $\ell : \mathcal{Y}\times\mathcal{Y} \to [0,1]$, ϕ^* L-Lipschitz continuous, and $[0,1] \subset \operatorname{dom}\phi^*$. Let S and T be two empirical distributions corresponding to datasets containing n data points sampled i.i.d. from P_s and P_t, respectively. Let us note $\mathfrak{R}$ the Rademacher complexity of a given class of functions, and $\ell\circ\mathcal{H} := \{x \mapsto \ell(h(x), h'(x)) : h, h' \in \mathcal{H}\}$. $\forall\delta \in (0,1)$, we have with probability of at least $1-\delta$:*

$$|D^\phi_{h,\mathcal{H}}(P_s||P_t) - D^\phi_{h,\mathcal{H}}(S||T)| \leq 2\mathfrak{R}_{P_s}(\ell\circ\mathcal{H}) + 2L\mathfrak{R}_{P_t}(\ell\circ\mathcal{H}) + 2\sqrt{(-\log\delta)/(2n)}. \tag{3.5}$$

In Lemma 2, we have shown that the empirical $\mathrm{D}^\phi_{h,\mathcal{H}}$ converges to the true $\mathrm{D}^\phi_{h,\mathcal{H}}$ discrepancy. It can then be estimated using a set of finite samples from the two distributions. The gap is bounded by the complexity of the hypothesis class and the number of examples (n). This result will also be important in the derivation of Theorem 3.

3.2. Domain Adaptation: Generalization Bounds

We now provide a novel generalization bound to estimate the error of a classifier in the target domain using the proposed

$\mathrm{D}^{\phi}_{h,\mathcal{H}}$ divergence and results from the previous section. We also provide a generalization Rademacher complexity bound for a binary classifier[1] based on the estimation of the $\mathrm{D}^{\phi}_{h,\mathcal{H}}$ from finite samples. We show that our bound generalizes previous results in Appendix C.1.

Theorem 2 (generalization bound). *Suppose $\ell : \mathcal{Y}\times\mathcal{Y} \to [0,1] \subset \mathrm{dom}\,\phi^*$. Denote $\lambda^* := R^{\ell}_S(h^*) + R^{\ell}_T(h^*)$, and let h^* be the ideal joint hypothesis. We have:*

$$R^{\ell}_T(h) \le R^{\ell}_S(h) + D^{\phi}_{h,\mathcal{H}}(P_s||P_t) + \lambda^*. \tag{3.6}$$

The three terms in this upper bound share similarity with the bounds in Ben-David et al. (2010a) and Zhang et al. (2019). The main difference lies in the discrepancy being used to compare the two marginal distributions. Ben-David et al. (2010a) use the $\mathcal{H}\Delta\mathcal{H}$ divergence (a reduction of the TV), and Zhang et al. (2019) use the MDD. In our case, we use a reduction of a lower bound estimator of a variational characterization of the general f-divergences. This generalizes the TV (and thus (Ben-David et al., 2010a)) and also includes popular divergences typically used in practice (see Appendix C). Intuitively, the first term in the bound accounts for the source error, the second term corresponds to the discrepancy between the marginal distributions, and the third term measures the ideal joint hypothesis (λ^*). If $\mathcal{H}$ is expressive enough and the labeling functions are similar, this last term could be reduced to a small value. The ideal joint hypothesis incorporates the notion of adaptability: when the optimal hypothesis performs poorly in either domain, we cannot expect successful adaptation.

Theorem 3 (generalization bound with Rademacher complexity). *Let $\ell : \mathcal{Y}\times\mathcal{Y} \to [0,1]$ and ϕ^* be L-Lipschitz continuous. Let S and T be two empirical distributions (i.e. datasets containing n data points sampled i.i.d. from P_s and P_t, respectively). Denote $\hat\lambda^* := \hat R^{\ell}_S(h^*) + \hat R^{\ell}_T(h^*)$. $\forall\delta \in (0,1)$, we have with probability of at least $1-\delta$:*

$$\begin{aligned} R^{\ell}_T(h) \le\ & \hat R^{\ell}_S(h) + D^{\phi}_{h,\mathcal{H}}(S||T) + \hat\lambda^* \\ & + 6\mathfrak{R}_S(\ell\circ\mathcal{H}) + 2(1+L)\mathfrak{R}_T(\ell\circ\mathcal{H}) \\ & + 5\sqrt{(-\log\delta)/(2n)}. \end{aligned} \tag{3.7}$$

Theorem 3 provides the computation of our generalization bound for a binary classifier in terms of the Rademacher complexity of the class $\mathcal{H}$. Under the assumption of an ideal joint hypothesis $\hat\lambda^*$, the generalization error can be reduced by jointly minimizing the risk in the source domain, the discrepancy between the two distributions, and regularizing the model to limit the complexity of the hypothesis class. We take all these into account when deriving practical algorithms in the next sections.

[1] Similar bounds can be derived for the multi-class scenario if we let $h : \mathcal{X}\times\mathcal{Y}$ being a score function and $\ell(x,y) = 1[\mathrm{argmax}_{\hat y} h(x,\hat y) \ne y]$ (i.e see (Mohri et al., 2018) Chapter 9).

4. Training Algorithm

We now exploit the results introduced above to derive a novel and practical domain-adversarial algorithm. We show how our framework for a particular divergence allows us to reinterpret and rectify the original domain-adversarial training method from Ganin et al. (2016). Our analysis highlights the differences between our adversarial training algorithm and that from Ganin et al. (2016). Finally, we analyze the use of γ weighted f-divergences. This sheds lights on why the practical objective from Zhang et al. (2019) outperforms DANN (Ganin et al. (2016)) and shows how, after a key correction of the latter, the performance gap vanishes.

4.1. f-Domain Adversarial Learning (f-DAL)

We now use the theory presented in the previous sections to derive f-DAL, a novel generalized domain adversarial learning framework.

Notation. Let the hypothesis h be the composition of $h = \hat h \circ g$ (i.e. let $\mathcal{H} := \{\hat h \circ g : \hat h \in \hat{\mathcal{H}}, g \in \mathcal{G}\}$ with $\hat{\mathcal{H}}$ another function class) where $g : \mathcal{X} \to \mathcal{Z}$. This can be interpreted as a mapping that pushes forward the two densities p_{s} and p_{t} to a representation space $\mathcal{Z}$ where a classifier $\hat h \in \hat{\mathcal{H}}$ operates. Consequently, we denote by $p^{\mathrm{z}}_{\mathrm{s}} := g\#p_{\mathrm{s}}$ and $p^{\mathrm{z}}_{\mathrm{t}} := g\#p_{\mathrm{t}}$ the push-forwards of the source and target domain densities, respectively. Figure 1 illustrates the f-DAL framework.

From Theorem 2, for adaptation to be possible in the representation space $\mathcal{Z}$, we assume the existence of some $\hat h \in \hat{\mathcal{H}}$ such that the ideal joint risk λ^* is negligible. This condition is necessary even if $p^{\mathrm{z}}_{\mathrm{s}} = p^{\mathrm{z}}_{\mathrm{t}}$. In other words, we need both, the difference between $p^{\mathrm{z}}_{\mathrm{s}}$ and $p^{\mathrm{z}}_{\mathrm{t}}$, and the ideal joint risk λ^* to be small. These are both sufficient and necessary conditions. We refer the reader to Ben-David et al. (2010b) for details on the impossibility theorems for DA. Thus, we assume that there exist some $g \in \mathcal{G}$ and $\hat h^* \in \hat{\mathcal{H}}$, such that the ideal joint risk (λ^*) is negligible. These assumptions are ubiquitous in modern DA methods, including SoTA methods (Ganin et al., 2016; Long et al., 2018; Hoffman et al., 2018b; Zhang et al., 2019) (sometimes not explicitly mentioned). It was recently shown in Zhao et al. (2019) that for this to be true in the present context, the label distributions between source and target must be close. In Appendix D.2, we provide further analysis and experimental results on the robustness of f-DAL to label shift. Moreover, we show that f-DAL can be simply combined with methods that deal with this setting, further boosting their performance. We emphasize however that dealing with label shift is outside of the scope of this work.

From Theorem 2, the target risk $R^{\ell}_T(h)$ can be minimized by jointly minimizing the error in the source domain and the discrepancy between the two distributions. Let y be the label of a source data point z, an optimization objective can be clearly written as:

$$\min_{\hat{h}\in\hat{\mathcal{H}}} \mathbb{E}_{z\sim p_s^z}[\ell(\hat{h}(z),y)] + \mathrm{D}^{\phi}_{\hat{h},\hat{\mathcal{H}}}(p_s^z||p_t^z). \tag{4.1}$$

Here, ℓ is a surrogate loss function used to minimize the empirical risk in the source domain. Under mild assumptions (see Proposition 1) and the use of Lemma 1, the minimization problem in (4.1) can be upper bounded (hence replaced) by the following min-max objective[2]:

$$\min_{\hat{h}\in\hat{\mathcal{H}}}\max_{\hat{h}'\in\hat{\mathcal{H}}} \mathbb{E}_{z\sim p_s^z}[\ell(\hat{h}(z),y)] + d_{s,t} \quad \text{where} \tag{4.2}$$

$$d_{s,t} := \mathbb{E}_{z\sim p_s^z}[\hat{\ell}(\hat{h}'(z),\hat{h}(z))] - \mathbb{E}_{z\sim p_t^z}[(\phi^*\circ\hat{\ell})(\hat{h}'(z),\hat{h}(z))].$$

We now formalize this result.

Proposition 1. *Suppose $d_{s,t}$ takes the form shown in (4.2) with $\hat{\ell}(\hat{h}'(z),\hat{h}(z)) \to \operatorname{dom}\phi^*$ and that for any $\hat{h} \in \hat{\mathcal{H}}$ (unconstrained), there exists $\hat{h}' \in \hat{\mathcal{H}}$ s.t. $\hat{\ell}(\hat{h}'(z),\hat{h}(z)) = \phi'(\frac{p_s^z(z)}{p_t^z(z)})$ for any $z \in \operatorname{supp}(p_t^z(z))$, with ϕ' the derivative of ϕ. The optimal $d_{s,t}$ is $D_\phi(P_s^z||P_t^z)$, i.e. $\max_{\hat{h}'\in\hat{\mathcal{H}}} d_{s,t} = D_\phi(P_s^z||P_t^z)$.*

If we let the feature extractor $g \in \mathcal{G}$ be the one that minimizes both the source error and the discrepancy term, Eq. (4.2) can be rewritten as:

$$\begin{aligned}\min_{\hat{h}\in\hat{\mathcal{H}},g\in\mathcal{G}}\max_{\hat{h}'\in\hat{\mathcal{H}}} &\mathbb{E}_{x\sim p_s}[\ell(\hat{h}\circ g,y)] + \mathbb{E}_{x\sim p_s}[\hat{\ell}(\hat{h}'\circ g,\hat{h}\circ g)] \\ &- \mathbb{E}_{x\sim p_t}[(\phi^*\circ\hat{\ell})(\hat{h}'\circ g,\hat{h}\circ g)].\end{aligned} \tag{4.3}$$

We let $\hat{\ell}(c,b) = a(b_{\text{argmax } c})$, where argmax a is the index of the largest element of vector a. For the choice of $a(.)$, we follow Nowozin et al. (2016) and choose it to be a monotonically increasing function when possible. This implies that we choose the domain of $\hat{\ell}$ to be $\mathbb{R}^k \times \mathbb{R}^k$ with k categories. Intuitively, $\hat{h}'$ is an auxiliary per-category domain classifier. This makes our framework different from DANN.

4.2. Revisiting Domain-Adversarial Training (DANN)

The original idea of domain-adversarial training was introduced in Ganin et al. (2016) and motivated with the theoretical results of Ben-David et al. (2010a). Specifically, the *domain-classifier/regularizer* is inspired by the proxy $\mathcal{A}$-distance (Ben-David et al., 2007) which is an approximation of the empirical estimation of the $\mathcal{H}\Delta\mathcal{H}$ divergence. While it has been shown that under mild assumptions the discrepancy being minimized in DANN corresponds to the JS divergence (see Appendix C), the connection between this and the DA theory has not been made clear since, to the best of our knowledge, generalization bounds for DA with f-divergences has not been derived.

[2] $d_{s,t}$ can be seen as an upper bound of the $\mathrm{D}^{\phi}_{h,\mathcal{H}}$ discrepancy.

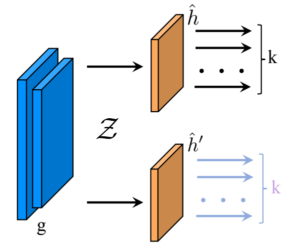

Figure 1. **f-DAL framework**. We interpret $h : \mathcal{X} \to \mathcal{Y}$ as the composition of two networks $h = \hat{h} \circ g$, where $g : \mathcal{X} \to \mathcal{Z}$ and $\hat{h}$ is a classifier operating in a representation space $\mathcal{Z}$. Inspired by our bounds, we let $\hat{h}'$ be a network of the same topology as $\hat{h}$. This is interpreted as a per-category domain classifier. Unlike us, Ganin et al. (2016) use a global domain-classifier or discriminator.

In this section, we use our bounds and algorithmic framework to revisit the domain-adversarial training method from Ganin et al. (2016). The analysis shows that while both can be interpreted as minimizing the JS divergence and thus are in line with our theoretical results (Theorem 2, Lemma 1 and Appendix C), DANN ignores the contribution of the source classifier which is not desirable or intuitive. Experimental results confirm that this apparently subtle difference leads to significant gains (using the same JS divergence, see tables 2 and 13). To explicitly see this, let us first rewrite the $d_{s,t}$ term in f-DAL (Equation (4.3)) using the JS divergence (shifted up to a constant that does not alter optimization). We then have $\hat{\ell}(h',h) = \log\sigma(h'_{\text{argmax}h})$ and $\phi^*(t) = -\log(1-e^t)$, where $\sigma(x) := \frac{1}{1+\exp(-x)}$ is the sigmoid function.

Plugging all together and rewriting conveniently, we obtain:

$$\begin{aligned} d_{s,t} &= \mathbb{E}_{x_s\sim p_s}\log\sigma\circ\left[\hat{h}'\circ g(x_s)\right]_{\text{argmax}h} \\ &+ \mathbb{E}_{x_t\sim p_t}\log\left(1-\sigma\circ\left[\hat{h}'\circ g(x_t)\right]_{\text{argmax}h}\right)\end{aligned} \tag{4.4}$$

which is the resulting $d_{s,t}$ term of f-DAL for the JS divergence. Assuming the output of the source classifier $\hat{h}$ is constant in terms of the argmax operator (e.g. $\hat{h} = e_i$, with e_i any standard basis vector), we obtain after manipulation the second part of the expression shown in Equation (9) in Ganin et al. (2016). Effectively, this shows that DANN ignores the contribution of the source classifier $\hat{h}$. In fact, *it assumes that the output of the source classifier is always constant* (e.g. $\hat{h} = e_i$), which is problematic. Moreover, the motivation of DANN through the proxy $\mathcal{A}$-distance ignores the topology/architecture of the discriminator network. This is in contrast with our formulation which suggests that the topology of the per-category domain classifier $\hat{h}'$ should be identical to that of $\hat{h}$ since both $\hat{h}', \hat{h} \in \hat{\mathcal{H}}$ (Figure 1) .

We additionally notice that f-DAL can explain DANN and connect it with the DA theory directly from a JS minimization perspective (i.e. without relying on an approximation of the empirical $\mathcal{H}\Delta\mathcal{H}$ divergence as in Ganin et al. (2016)). This result follows from Lemma 1 and details can be found in Appendix C. This allows us to compare head-to-head f-DAL JS vs DANN, in which scenario f-DAL can be understood as the *corrected/revisited* version of DANN.

4.3. On γ-weighted f-divergences

If we relax the need for $\phi(1) = 0$ in Proposition 1, the new objective only shifts by a constant, e.g., $\max_{\hat{h}' \in \hat{\mathcal{H}}} d_{s,t} = D_{\hat{\phi}}(P_s^z || P_t^z) + \phi(1)$ with $\hat{\phi}(x) := \phi(x) - \phi(1)$. By Lemma 4 (Appendix C), we can rescale ϕ^*, and ϕ will change accordingly. These can be done for the general family of divergences, accommodating a larger family of distributions.

γ-weighted JS Divergence. We recall that the objective from MDD (Zhang et al., 2019) (i.e. the one introduced to deal with the practical issues of the MDD discrepancy) corresponds to the γ-JS divergence (up to a constant that does not alter optimization). This result gives insight into the big performance gap observed when comparing MDD vs DANN (see Appendix C). That gap is due to the fact that DANN considers the output of the source classifier as a constant (see section 4.2). After revisiting DANN (Equation (4.3) and Section 4.2), experimental results (Table 3) show that the γ-weighted-JS divergence only performs comparably to the JS divergence with per-dataset extra-tuning of the γ parameter. A statistical analysis shows that this difference in performance (if any) does not justify the expensive introduction of the new hyperparameter γ.

5. Experimental Results

We now experimentally analyze and compare the proposed framework vs previous adversarial methods. We perform experiments on both toy datasets (digits) and real-world problems (natural language and visual tasks).

5.1. Setup

Digits. We evaluate our method on two digits datasets **MNIST** and **USPS** with two transfer tasks (M $\to$ U and U $\to$ M). We adopt the splits and evaluation protocol from (Long et al., 2018) which constitute of 60,000 and 7,291 training images and the standard test set of 10,000 and 2,007 test images for MNIST and USPS, respectively.

Visual Tasks. We use two visual benchmarks: **(1)** the *Office-31* dataset (Saenko et al., 2010) contains 4,652 images and 31 categories, collected from three distinct domains: Amazon **(A)**, Webcam **(W)** and DSLR **(D)**. **(2)** the *Office-Home* dataset (Venkateswara et al., 2017) contains 15,500 images from four different domains: Artistic images, Clip Art, Product images, and Real-world images.

NLP Tasks. For this task, we consider the Amazon product

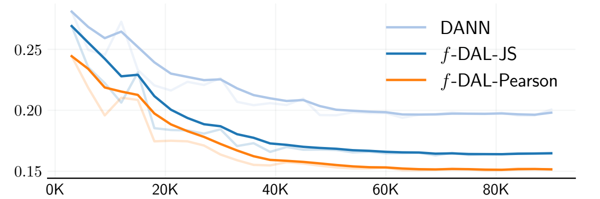

Figure 2. Target Domain Loss on the Digits Datasets M→ U.

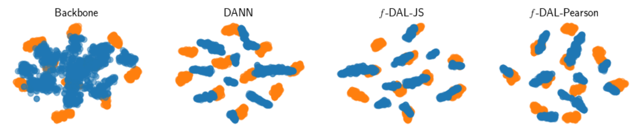

Figure 3. t-SNE Visualization of the last layer features on the Digits Dataset M→ U.

reviews dataset (Blitzer et al., 2006) which contains online reviews of different products collected on the Amazon website. We follow the splits and evaluation protocol from (Courty et al., 2017; Dhouib et al., 2020). We choose 4 of its subsets corresponding to different product categories, namely: books, dvd, electronics and kitchen (denoted by B, D, E, K, respectively) and leads to 12 domain adaptation tasks of varying difficulty. The problem is to predict positive (higher than 3 stars) or negative (3 stars or less) notation of reviews. For each task, we use predefined sets of 2000 instances of source and target data samples for training, and keep 4000 instances of the target domain for testing.

Baselines. Our main baseline is DANN (Ganin et al., 2016). For the JS divergence, our method can be seen as the revisited interpretation of DANN. We then study whether this interpretation based on our bounds correlates well with experimental results. We also compare with recent methods such as CDAN (Long et al., 2018) for Digits and JDOT and MADAOT (Courty et al., 2017; Dhouib et al., 2020) for the NLP benchmark. MDD (Zhang et al., 2019) is the γ-JS divergence in our framework, we also use it for comparison in visual tasks where results for the method are available.

Implementation Details: We implement our algorithm in PyTorch. For the Digits datasets, the implementation details follows (Long et al., 2018). Thus, the backbone network is LeNet (LeCun et al., 1998). The main classifier ($\hat{h}$) and auxiliary classifier ($\hat{h}'$) are both 2 linear layers with ReLU non-linearities and Dropout (0.5) in the last layer. For the NLP task, we follow the standard protocol from Courty et al. (2017); Ganin et al. (2016) and use a simple 2-layer model with sigmoid activation function. For the visual datasets, we use ResNet-50 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) as the backbone network. The main classifier ($\hat{h}$) and auxiliary classifier ($\hat{h}'$) are both 2 layers neural nets with Leaky-ReLU activation functions. We use spectral normalization (SN) as in (Miyato et al., 2018) only for these two (i.e $\hat{h}$ and $\hat{h}'$). We did not see any transfer improvement by using it. The reason for this was

Table 2. Comparison of the f-DAL framework vs DANN on different datasets.

Method	Datasets: Toy Digits	NLP Amazon Reviews	Vision: Office-31	Office-Home	Significance
DANN (Ganin et al., 2016)	93.3	76.3	82.2	57.6	-
f-DAL (JS)	**96.6**	80.0	88.8	66.8	×✓✓✓
f-DAL (Pearson χ^2)	96.3	**81.6**	**89.2**	**68.3**	×✓✓✓

Table 3. Comparison of γ weighted divergences

	γ	Avg Digits	Avg Office-31	Avg
f-DAL (JS)	-	**96.6**	88.8	92.7
f-DAL (Pearson χ^2)	-	96.3	**89.2**	**92.8**
f-DAL(γ-JS) MDD	2	96.0	88.1	92.0
	3	*96.3*	88.5	92.4
	4	96.2	*88.9*	92.5

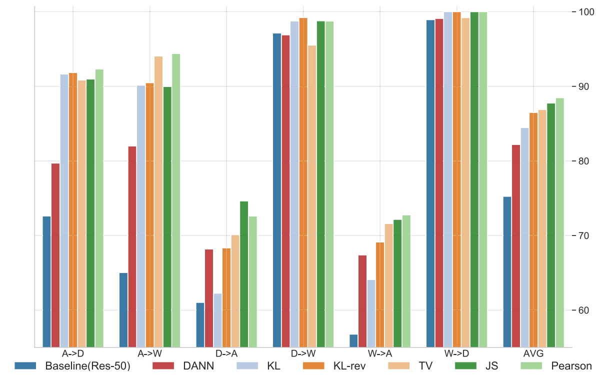

Figure 4. Transfer performance of a model trained using f-DAL for different f-divergences and transfer tasks on Office-31. Baseline is ResNet-50 source only. We show the performance of DANN (Table 4). When compared with f-DAL (JS), a performance boost is observed. This is in line with our bounds which suggest the use of a per-category domain classifier vs a discriminator.

to avoid gradient issues and instabilities during training for some divergences in the first epochs. For the first two tasks, hyperparameters are determined based on a subset (10%) of the training set for one task (e.g. M $\rightarrow$ U and B $\rightarrow$ D) and kept constant for the others. For the visual tasks, we use the hyperparameters and same training protocol from MDD (Zhang et al. (2019)). We report the average accuracies over 3 experiments. Full details are in Appendix E.

5.2. Experimental Analysis

Revisited DANN. We now compare the performance of f-DAL (JS) vs DANN on the four datasets. In this scenario, f-DAL (JS) is the corrected version of DANN as discussed in Section 4.2. We can see that f-DAL (JS) always outperforms DANN. To further corroborate the statistical significance of this, we conducted a two sided Wilcoxon signed rank test. With the exception of the Digits datasets (for which performance is beyond 90%), f-DAL (JS) is statistically significantly better than DANN (5% significance, 95% confidence, Table 13). For the digits dataset, we provide training losses in the target domain in Fig. 2 and t-SNE (Maaten & Hinton, 2008) visualizations of the last layer input (perplexity=30) in Fig. 3. f-DAL (JS) converges faster and the resulting features are also better aligned.

Comparing f-divergences. We compare the performance of f-divergences on *Office-31*. Specifically, we evaluate the model on the six combinations of transfer tasks with different divergences. All hyperparameters are kept constant for all divergences in this experiment. As shown in Figure 4, the JS and Pearson χ^2 divergences achieve the best results, with the *Pearson* χ^2 *achieving the best overall result among all the transfer tasks on this benchmark*. This is also the case for the Digits, NLP and Office-Home datasets. It is worth noting that this divergence was never used before to learn invariant representations in the context of DA. The excellent performance of χ^2 is also reminiscent of histogram-based (visual) bags of words representations that were shown to work better with χ^2 distances than with ℓ_2 and ℓ_1 distances for image and text classification tasks (Li et al., 2013).

Comparing γ-weighted divergences. We now investigate the significance of introducing the hyper-parameter γ to define the γ-weighted divergences. We compare in Table 3 the performance of using γ-JS vs JS and Pearson in two benchmarks: (1) Digits and (2) Office-31. The γ-JS divergence only outperforms the JS after tuning the hyperarameter γ. The difference is only of 0.1% in average in the Office-31 dataset giving a p-val=0.89 using the Wilcoxon signed rank test. This means that after correction with our framework DANN/f-DAL-JS is as good as γ-JS without additional hyperparameter tuning. In general, we found the use Pearson χ^2 divergence gives slightly better numerical results.

Training Dynamics. Fig. 2 and Fig. 5 illustrate the target loss curves and the values of $\hat{\ell}$ for JS and Pearson, respectively. In both cases our framework converges faster and achieves lower cost (see Figure 2). Figure 5 illustrates the value of $\hat{\ell}$ for both source and target where $\hat{\ell} \approx \phi'(1) = 0$, which implies $p_s^z \approx p_t^z$ (Proposition 1) as desired. It is worth noting that while this is true in both cases, domain invariance is achieved faster (almost after the first epoch) with the Pearson χ^2. This could also give intuition about the noticeable performance gap while using this divergence.

Results. We compare our method vs. recent state-of-the-art domain adversarial approaches in Tables 4 to 7. *Ours* in the tables correspond to f-DAL using the Pearson χ^2 divergence, with the exception of D $\rightarrow$ W and D $\rightarrow$ A in Table 4, and Ar $\rightarrow$ Pr in Table 5 where we use JS divergence. A detailed version of these with every divergence's performance can be found in Appendix D. In all cases, our approach outperforms previous methods, including MDD which is also

Table 4. Accuracy represented in (%) with average and standard deviation on the Office-31 benchmark.

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
ResNet-50 (He et al., 2016)	68.4±0.2	96.7±0.1	99.3±0.1	68.9±0.2	62.5±0.3	60.7±0.3	76.1
DANN (Ganin et al., 2016)	82.0±0.4	96.9±0.2	99.1±0.1	79.7±0.4	68.2±0.4	67.4±0.5	82.2
JAN (Long et al., 2017)	85.4±0.3	97.4±0.2	99.8±0.2	84.7±0.3	68.6±0.3	70.0±0.4	84.3
GTA (Sankaranarayanan et al., 2018)	89.5±0.5	97.9±0.3	99.8±0.4	87.7±0.5	72.8±0.3	71.4±0.4	86.5
MCD (Saito et al., 2018)	88.6±0.2	98.5±0.1	**100.0**±.0	92.2±0.2	69.5±0.1	69.7±0.3	86.5
CDAN (Long et al., 2018)	94.1±0.1	98.6±0.1	**100.0**±.0	92.9±0.2	71.0±0.3	69.3±0.3	87.7
f-DAL (γ-JS) / MDD (Zhang et al., 2019)	94.5±0.3	98.4±0.1	**100.0**±.0	93.5±0.2	74.6±0.3	72.2±0.1	88.9
Ours (f-DAL)	**95.4** ±0.7	98.8±0.1	**100.0**±.0	93.8 ±0.4	**74.9** ±1.5	74.2 ±0.5	**89.5**
Ours (f-DAL Pearson) + Alignment	93.4±0.4	**99.0**±0.1	**100.0**±.0	**94.8**±0.6	73.6±0.2	**74.6**±0.4	89.2

Table 5. Accuracy (%) on the Office-Home benchmark.

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
ResNet-50 (He et al., 2016)	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DANN (Ganin et al., 2016)	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN (Long et al., 2017)	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CDAN (Long et al., 2018)	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
f-DAL (γ-JS) / MDD (Zhang et al., 2019)	54.9	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	72.5	60.2	82.3	68.1
Ours (f-DAL)	54.7	71.7	77.8	61.0	**72.6**	72.2	60.8	53.4	80.0	**73.3**	**60.6**	**83.8**	68.5
Ours (f-DAL - Pearson) + Alignment	**56.7**	**77.0**	**81.1**	**63.1**	72.2	**75.9**	**64.5**	**54.4**	**81.0**	72.3	58.4	83.7	**70.0**

Table 6. Accuracy on the Amazon Reviews data sets

Method	B→D	B→E	B→K	D→B	D→E	D→K	E→B	E→D	E→K	K→B	K→D	K→E	Avg
JDOTNN (Courty et al., 2017)	79.5	78.1	79.4	76.3	78.8	82.1	74.9	73.7	87.2	72.8	76.5	84.5	78.7
MADAOT (Dhouib et al., 2020)	82.4	75.0	80.4	**80.9**	73.5	81.5	**77.2**	78.1	**88.1**	75.6	75.9	87.1	79.6
DANN (Dhouib et al., 2020; Ganin et al., 2016)	80.6	74.7	76.7	74.7	73.8	76.5	71.8	72.6	85.0	71.8	73.0	84.7	76.3
Ours (f-DAL)	**84.0**	**80.9**	**81.4**	80.6	**81.8**	**83.9**	76.7	**78.3**	87.9	**76.5**	**79.5**	**87.5**	**81.6**

Table 7. Accuracy on the Digits datasets

Method	M→U	U→M	Avg
DANN (Ganin et al., 2016)	91.8	94.7	93.3
CDAN (Long et al., 2018)	93.9	96.9	95.4
Ours (f-DAL)	**95.3**	**97.3**	**96.3**

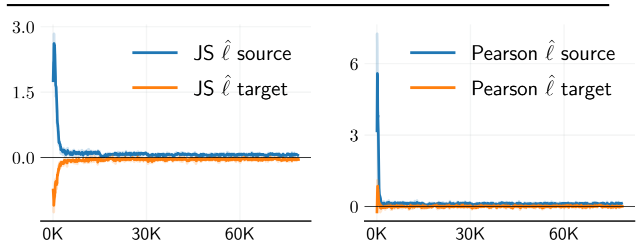

Figure 5. Values of $\hat{\ell}(\hat{h}', \hat{h})$ for source and target on Digits M→ U. $\hat{\ell} \approx \phi'(1) = 0$, which implies $p_s^z \approx p_t^z$ (see Proposition 1)

included in our framework (Section 4.3), and requires tuning of the hyperparameter γ. What is most impressive is that, unlike our approach, some methods listed in the tables can be interpreted as DANN + additional techniques to improve their performance (i.e. CDAN). It would be interesting to see if these techniques still introduce gains after correcting DANN (i.e. f-DAL JS) or if they were necessary because of the disconnect between theory and algorithms.

Improving f-DAL with Sampling-Based Alignment. In this experiment, we show that if the distance between the label marginals is not negligible f-DAL is still effective and can simply be combined with SoTA methods that deal with the label shift such as Jiang et al. (2020). We refer to this in Tables 4 and 5 as "+Alignment." For this experiment, we follow the setting from Jiang et al. (2020) but replace the adversarial method for f-DAL-Pearson. We also remove their masking scheme as we did not find it necessary with f-DAL. Clearly, in the Office-31 dataset (Table 4) the distance between the label marginals is not significantly different and we did not see any improvement by introducing implicit alignment. This is in contrast with Table 5 (Office-Home dataset) where our method notably benefits from the sampling-based alignment scheme. This again showcases the versatility of f-DAL. We refer to Appendix D.2 for more details and experiments on label-shift.

6. Related Work

Theory. The domain adaptation problem has been rigorously investigated in (Ben-David et al., 2007; 2010a; Mansour et al., 2009; Zhao et al., 2019; Zhang et al., 2019) where a classifiers target error is bounded in terms of its source error and the divergence between the two domains. We propose a measure of discrepancy between distributions based on a variational characterization of f-divergences. Our method includes the $\mathcal{H}\Delta\mathcal{H}$-divergence as a particular case but also other divergences used in practice. Moreover, our bounds based on f-divergences allow us to connect theory and practical algorithms without surrogate objectives.

Domain-Adversarial Algorithms. Ganin et al. (2016) introduced domain-adversarial training with insights from Ben-David et al. (2010a). This algorithm has been heavily adopted in the context of neural networks (Long et al., 2018; Hoffman et al., 2018b; Zhang et al., 2019). We propose a general adversarial framework for the family of f-divergences based on our bounds. We show how to correct the training algorithm from Ganin et al. (2016), and how to incorporate a large family of f-divergences. We explain why MDD (Zhang et al., 2019) outperforms Ganin et al. (2016) and show how the gap vanishes after correction.

Variational f-divergences. Nguyen et al. (2010) propose a derivation of the variational characterization of f-divergences that was later used for GANs (Nowozin et al., 2016). These were used in the context of DA in an example in Wu et al. (2019) to rewrite the domain-regularizer from Ganin et al. (2016). We derive f-divergence based generalization bounds from which we derive an algorithmic framework different from Ganin et al. (2016). Our analysis shows how to correct DANN. Morever, experimental results showing the performance of f-divergences in the context of domain-adversarial learning has not been provided.

7. Conclusions

We have provided a novel perspective on the domain-adversarial problem by deriving a general domain adaptation framework. Our bounds are based on a variational characterization of f-divergences and recover the theoretical results from seminal works as a special case, and also support divergences typically used in practice. We have derived a general algorithmic framework that is practical for neural networks. It allows us to reinterpret and correct the original domain-adversarial training method. We also show through large-scale experiments that several f-divergences can be used to minimize the discrepancy between source and target domains. We showed that some divergences that do not require additional techniques and/or hyperparameter tuning can help achieve state-of-the-art performance.

Acknowledgements. We would like to thank Rafid Mahmood, Mark Brophy and the anonymous reviewers for helpful discussions and feedback on earlier versions of this manuscript.

References

Ali, S. M. and Silvey, S. D. 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, 1966.

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

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010a.

Ben-David, S., Lu, T., Luu, T., and Pál, D. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 129–136, 2010b.

Billingsley, P. *Probability and measure*. John Wiley & Sons, 2008.

Blitzer, J., McDonald, R., and Pereira, F. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 120–128, 2006.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. Joint distribution optimal transportation for domain adaptation, 2017.

Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Dhouib, S., Redko, I., and Lartizien, C. Margin-aware adversarial domain adaptation with optimal transport. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2514–2524, Virtual, 13–18 Jul 2020. PMLR.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

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

Hoffman, J., Mohri, M., and Zhang, N. Algorithms and theory for multiple-source adaptation. In *Advances in Neural Information Processing Systems*, pp. 8246–8256, 2018a.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., and Darrell, T. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pp. 1989–1998. PMLR, 2018b.

Huszár, F. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*, 2015.

Jiang, X., Lao, Q., Matwin, S., and Havaei, M. Implicit class-conditioned domain alignment for unsupervised domain adaptation. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4816–4827. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/jiang20d.html.

Kifer, D., Ben-David, S., and Gehrke, J. Detecting change in data streams. In *VLDB*, volume 4, pp. 180–191. Toronto, Canada, 2004.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, P., Samorodnitsk, G., and Hopcroft, J. Sign cauchy projections and chi-square kernel. In *Advances in Neural Information Processing Systems*, pp. 2571–2579, 2013.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. Deep transfer learning with joint adaptation networks, 2017.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 1640–1650, 2018.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. In *Proceedings of The 22nd Annual Conference on Learning Theory (COLT 2009)*, Montreal, Canada, 2009.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56 (11):5847–5861, 2010.

Nowozin, S., Cseke, B., and Tomioka, R. f-GAN: Training generative neural samplers using variational divergence minimization. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 271–279. Curran Associates, Inc., 2016.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226. Springer, 2010.

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3723–3732, 2018.

Sankaranarayanan, S., Balaji, Y., Castillo, C. D., and Chellappa, R. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8503–8512, 2018.

Sason, I. and Verdú, S. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

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

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *(IEEE) Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Wu, Y., Winston, E., Kaushik, D., and Lipton, Z. Domain adaptation with asymmetrically-relaxed distribution alignment. *arXiv preprint arXiv:1903.01689*, 2019.

Zhang, Y., Liu, T., Long, M., and Jordan, M. Bridging theory and algorithm for domain adaptation. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7404–7413, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Zhao, H., Combes, R. T. d., Zhang, K., and Gordon, G. J. On learning invariant representation for domain adaptation. *arXiv preprint arXiv:1901.09453*, 2019.