Supplementary Material
f-Domain-Adversarial Learning: Theory and Algorithms

A. Divergences between probability measures

As explained above, the difference term between source and target domains is important in bounding the target loss. We now provide more details about the $\mathcal{H}\Delta\mathcal{H}$-divergence and f-divergences that are used to compare both domains.

$\mathcal{H}\Delta\mathcal{H}$-divergence The $\mathcal{H}$-divergence is a restriction of total variation. For binary classification, define $I(h) := \{\mathbf{x} \in \mathcal{X} : h(\mathbf{x}) = 1\}$, then the $\mathcal{H}$-divergence between two measures μ and ν given the hypothesis class $\mathcal{H}$ is (Ben-David et al., 2010a):

$$d_{\mathcal{H}}(\mu, \nu) = 2 \sup_{h \in \mathcal{H}} |\mu(I(h)) - \nu(I(h))|. \tag{A.1}$$

Define $\mathcal{H}\Delta\mathcal{H} := \{h \oplus h' : h, h' \in \mathcal{H}\}$ ($\oplus$: XOR), then $d_{\mathcal{H}\Delta\mathcal{H}}(\mu, \nu)$ can be used to bound the difference between the source and target errors. $\mathcal{H}\Delta\mathcal{H}$ divergence has been extended to general loss functions (Mansour et al., 2009) and marginal disparity discrepancy (Zhang et al., 2019).

f-divergence Given two measures μ and ν with $\mu \ll \nu$ (μ absolute continuous w.r.t. ν), the f-divergence $D_\phi(\mu||\nu)$ is defined as (Csiszár, 1967; Ali & Silvey, 1966):

$$D_\phi(\mu \parallel \nu) = \int \phi\left(\frac{d\mu}{d\nu}\right) d\nu, \tag{A.2}$$

where $d\mu/d\nu$ is known as the Radon–Nikodym derivative (e.g. Billingsley, 2008). Assume ϕ is convex and lower semi-continuous, then from the Fenchel–Moreau theorem, $\phi^{**} = \phi$, with ϕ^* known as the Fenchel conjugate of ϕ:

$$\phi^*(\mathbf{y}) = \sup_{\mathbf{x} \in \operatorname{dom} \phi} \langle \mathbf{x}, \mathbf{y} \rangle - \phi(\mathbf{x}), \tag{A.3}$$

which is convex since it is a supremum of an affine function. In order for $\mathbf{x}$ to take the supremum, it is necessary and sufficient that $\mathbf{y} \in \partial\phi(\mathbf{x})$ using the stationarity condition. Therefore, with (A.2) and (A.3), $D_\phi(\mu \parallel \nu)$ can be written as:

$$D_\phi(\mu \parallel \nu) = \sup_{T \in \mathcal{T}} \mathbb{E}_{X \sim \mu}[T(X)] - E_{Z \sim \nu}[\phi^*(T(Z))], \tag{A.4}$$

where $\mathcal{T} = \{T : T \text{ is a measurable function and } T : \mathcal{X} \to \operatorname{dom} \phi^*\}$. In practice we restrict $\mathcal{T}$ to a subset as in Definition 2. For different choices of ϕ see Table 8.

(Nguyen et al., 2010) derive a general variational method to estimate f-divergences given only samples. (Nowozin et al., 2016) extend their method from merely estimating a divergence for a fixed model to estimating model parameters. While our method builds on this variational formulation, we use it in the context of domain adaptation.

B. Proofs

In this section, we provide the proofs for the different theorems and lemmas:

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:*

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

Divergence	$\phi(x)$	$\phi^*(t)$	$\phi'(1)$	$g(x)$
MDD	$x\log\frac{\gamma x}{1+\gamma x}+\frac{1}{\gamma}\log\frac{1}{1+\gamma x}$	$-\log(1-e^t)/\gamma$	$\log\frac{\gamma}{1+\gamma}$	$\log x$
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
Squared Hellinger (SH)	$(\sqrt{x}-1)^2$	$\frac{t}{1-t}$	0	$1-\exp x$
γ-weighted Pearson χ^2	$(\gamma x-1)^2/\gamma$	$(t^2/4+t)/\gamma$	0	x
Neynman χ^2	$\frac{(1-x)^2}{x}$	$2-2\sqrt{1-t}$	0	$1-\exp x$
γ-weighted total variation	$\frac{1}{2\gamma}\lvert\gamma x-1\rvert$	$(t/\gamma)\mathbf{1}_{-1/2\le t\le 1/2}$	$[-1/2,1/2]$	$\frac{1}{2}\tanh x$
Total Variation (TV)	$\frac{1}{2}\lvert x-1\rvert$	$\mathbf{1}_{-1/2\le t\le 1/2}$	$[-1/2,1/2]$	$\frac{1}{2}\tanh x$

Table 8. Popular f-divergences, their conjugate functions and choices of g. We take $\hat{l}(a,b)=g(b_{\text{argmax}\,a})$.

Proof. Rewriting the target loss we have:

$$\begin{aligned}R_T^\ell(h) &= R_T^\ell(h)-R_S^\ell(h,f_t)+R_S^\ell(h,f_t)-R_S^\ell(h)+R_S^\ell(h),\\ &\le R_S^\ell(h)+|R_S^\ell(h)-R_S^\ell(h,f_t)|+|R_T^\ell(h)-R_S^\ell(h,f_t)|\end{aligned}$$

where:

$$\begin{aligned}|R_S^\ell(h)-R_S^\ell(h,f_t)| &= |R_S^\ell(h,f_s)-R_S^\ell(h,f_t)|\\ &= |\mathbb{E}_{x\sim P_s}[|h(x)-f_t(x)|-|h(x)-f_s(x)|]|\\ &\le \mathbb{E}_{x\sim P_s}[|f_t(x)-f_s(x)|]\end{aligned}$$

and:

$$\begin{aligned}|R_T^\ell(h)-R_S^\ell(h,f_t)| &= |R_T^\ell(h,f_t)-R_S^\ell(h,f_t)|\\ &\le \int|p_t(x)-p_s(x)|\cdot|h(x)-f_t(x)|dx\\ &\le \int|(\frac{p_t(x)}{p_s(x)}-1)p_s(x)|dx = D_\phi(P_s||P_t)\end{aligned}$$

with $\phi(x)=|x-1|$ which represents the total divergence. □

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

$$\begin{aligned}|R_S^\ell(h,h')-R_T^{\phi^*\circ\ell}(h,h')| &\le D_{h,\mathcal{H}}^\phi(P_s||P_t)\le D_{\mathcal{H}}^\phi(P_s||P_t)\\ &\le D_\phi(P_s||P_t).\end{aligned}\tag{3.4}$$

Proof.

$$\mathrm{D}_{\mathcal{H}}^\phi(P_s||P_t)=\sup_{h\in\mathcal{H}}\mathrm{D}_{h,\mathcal{H}}^\phi(P_s||P_t)\ge\mathrm{D}_{h,\mathcal{H}}^\phi(P_s||P_t)\tag{B.1}$$

$$=\sup_{h'\in\mathcal{H}}|E_{x\sim P_s}[\ell(h(x),h'(x))]-\mathbb{E}_{x\sim P_t}[\phi^*(\ell(h(x),h'(x)))]|\tag{B.2}$$

$$\ge|E_{x\sim P_s}[\ell(h(x),h'(x))]-\mathbb{E}_{x\sim P_t}[\phi^*(\ell(h(x),h'(x)))]|\tag{B.3}$$

$$=|R_S^\ell(h,h')-R_T^{\phi^*\circ\ell}(h,h')|.\tag{B.4}$$

For the rightmost inequality in (3.4), it is well-known that f-divergence D_ϕ is nonnegative (e.g. Sason & Verdú, 2016), and thus

$$D_\phi(P_s\|P_t)=\sup_{T\in\mathcal{T}}|\mathbb{E}_{x\sim P_s}T(x)-\mathbb{E}_{x\sim P_t}\phi^*(T(x))|.\tag{B.5}$$

Restricting $\mathcal{T}$ to $\hat{\mathcal{T}}$ as in Definition 2 we obtain $D_\phi(P_\mathrm{s}\|P_\mathrm{t}) \geq \mathrm{D}^\phi_{\mathcal{H}}(P_\mathrm{s}||P_\mathrm{t})$. □

Lemma 2. *Suppose $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$, ϕ^* L-Lipschitz continuous, and $[0,1] \subset \mathrm{dom}\,\phi^*$. 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 $1-\delta$:*

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

Proof. For reference, we refer the reader to Chapter 3 of (Mohri et al., 2018). Using the notations of R and $\hat{R}$ that represent the true and empirical risks, we have:

$$\begin{aligned} \mathrm{D}^\phi_{h,\mathcal{H}}(P_\mathrm{s}||P_\mathrm{t}) - \mathrm{D}^\phi_{h,\mathcal{H}}(\mathrm{S}||\mathrm{T}) &= \sup_{h' \in \mathcal{H}} \{|R^\ell_S(h,h') - R^{\phi^* \circ \ell}_T(h,h')|\} \\ &- \sup_{h' \in \mathcal{H}} \{|\hat{R}^\ell_S(h,h') - \hat{R}^{\phi^* \circ \ell}_T(h,h')|\} \\ &\leq \sup_{h' \in \mathcal{H}} ||R^\ell_S(h,h') - R^{\phi^* \circ \ell}_T(h,h')| - |\hat{R}^\ell_S(h,h') - \hat{R}^{\phi^* \circ \ell}_T(h,h')|| \\ &\leq \sup_{h' \in \mathcal{H}} |R^\ell_S(h,h') - R^{\phi^* \circ \ell}_T(h,h') - \hat{R}^\ell_S(h,h') + \hat{R}^{\phi^* \circ \ell}_T(h,h')| \\ &= \sup_{h' \in \mathcal{H}} |R^\ell_S(h,h') - \hat{R}^\ell_S(h,h')| + |R^{\phi^* \circ \ell}_T(h,h') - \hat{R}^{\phi^* \circ \ell}_T(h,h')| \\ &\leq 2\mathfrak{R}_{P_\mathrm{s}}(\ell \circ \mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}} + 2\mathfrak{R}_{P_\mathrm{t}}(\phi^* \circ \ell \circ \mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}} \end{aligned} \tag{B.6}$$

where: $|R^\ell_S(h,h') - \hat{R}^\ell_S(h,h')| \leq 2\mathfrak{R}_{P_\mathrm{s}}(\ell \circ \mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}$ (Theorem 3.3 of (Mohri et al., 2018)). Similarly, by Talagrand's lemma (Lemma 5.7 and Definition 3.2 of (Mohri et al., 2018)) we have: $\mathfrak{R}_{P_\mathrm{t}}(\phi^* \circ \ell \circ \mathcal{H}) \leq \mathrm{L}\mathfrak{R}_{P_\mathrm{t}}(\ell \circ \mathcal{H})$, with $\phi^* \circ \ell \circ \mathcal{H} := \{x \mapsto \phi(\ell(h(x), h'(x))) : h, h' \in \mathcal{H}\}$. □

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:*

$$R^\ell_T(h) \leq R^\ell_S(h) + D^\phi_{h,\mathcal{H}}(P_s||P_t) + \lambda^*. \tag{3.6}$$

Proof. We first introduce the following lemma for our proof:

Lemma 3. *For any function ϕ that satisfies $\phi(1) = 0$ we have $\phi^*(t) \geq t$ where ϕ^* is the Fenchel conjugate of ϕ.*

Proof. From the definition of Fenchel conjugate, $\phi^*(t) = \sup_{x \in \mathrm{dom}\,\phi}(xt - \phi(x)) \geq t - \phi(1) = t$. □

$$\begin{aligned} R^\ell_T(h, f_t) &\leq R^\ell_T(h, h^*) + R^\ell_T(h^*, f_t) && \text{(triangle inequality } \ell\text{)} && \text{(B.7)} \\ &= R^\ell_T(h, h^*) + R^\ell_T(h^*, f_t) - R^\ell_S(h, h^*) + R^\ell_S(h, h^*) && && \text{(B.8)} \\ &\leq R^{\phi^* \circ \ell}_T(h, h^*) - R^\ell_S(h, h^*) + R^\ell_S(h, h^*) + R^\ell_T(h^*, f_t) && \text{(Lemma 3)} && \text{(B.9)} \\ &\leq |R^{\phi^* \circ \ell}_T(h, h^*) - R^\ell_S(h, h^*)| + R^\ell_S(h, h^*) + R^\ell_T(h^*, f_t) && && \text{(B.10)} \\ &\leq \mathrm{D}^\phi_{h,\mathcal{H}}(P_\mathrm{s}||P_\mathrm{t}) + R^\ell_S(h, h^*) + R^\ell_T(h^*, f_t) && \text{(Lemma 1)} && \text{(B.11)} \\ &\leq \mathrm{D}^\phi_{h,\mathcal{H}}(P_\mathrm{s}||P_\mathrm{t}) + R^\ell_S(h, f_s) + \underbrace{R^\ell_S(h^*, f_s) + R^\ell_T(h^*, f_t)}_{\lambda^*}. && && \text{(B.12)} \end{aligned}$$

□

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 $1-\delta$:*

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

Proof. We show in the following that:

$$R^\ell_T(h) \leq \hat{R}^\ell_S(h) + \mathrm{D}^\phi_{h,\mathcal{H}}(\mathrm{S}||\mathrm{T}) + \hat{\lambda}^*_\phi \tag{B.13}$$

$$+ 6\mathfrak{R}_S(\ell \circ \mathcal{H}) + 2(1+L)\mathfrak{R}_T(\ell \circ \mathcal{H}) + 5\sqrt{(-\log\delta)/(2n)}. \tag{B.14}$$

This follows from Theorem 2 where: $R^\ell_T(h) \leq R^\ell_S(h) + \mathrm{D}^\phi_{h,\mathcal{H}}(P_\mathrm{s}||P_\mathrm{t}) + R^\ell_S(h^*) + R^\ell_T(h^*)$. We also have: $|R^\ell_D(h) - \hat{R}^\ell_D(h)| \leq 2\mathfrak{R}_D(\ell \circ \mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}$ (Theorem of 3.3 (Mohri et al., 2018)). From Lemma 2, $\mathrm{D}^\phi_{h,\mathcal{H}}(P_\mathrm{s}||P_\mathrm{t}) \leq 2\mathfrak{R}_{P_\mathrm{s}}(\ell \circ \mathcal{H}) + 2\mathrm{L}\mathfrak{R}_{P_\mathrm{t}}(\ell \circ \mathcal{H}) + 2\sqrt{\frac{\log\frac{1}{\delta}}{2n}}$. Plugging in and rearranging gives the desired results. □

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

Proof. We first rewrite from the definition of $d_{s,t}$ in (4.2):

$$d_{s,t} = \mathbb{E}_{z\sim p^z_\mathrm{s}}[\hat{\ell}(\hat{h}'(z), \hat{h}(z))] - \mathbb{E}_{z\sim p^z_\mathrm{t}}[(\phi^* \circ \hat{\ell})(\hat{h}'(z), \hat{h}(z))] \tag{B.15}$$

$$= \int [p^z_\mathrm{s}(z)\hat{\ell}(\hat{h}'(z), \hat{h}(z)) - p^z_\mathrm{t}(z)(\phi^* \circ \hat{\ell})(\hat{h}'(z), \hat{h}(z))]dz \tag{B.16}$$

$$= \int p^z_\mathrm{t}(z) \left[\frac{p^z_\mathrm{s}(z)}{p^z_\mathrm{t}(z)}\hat{\ell}(\hat{h}'(z), \hat{h}(z)) - (\phi^* \circ \hat{\ell})(\hat{h}'(z), \hat{h}(z))\right] dz. \tag{B.17}$$

Maximizing w.r.t h' and assuming $\hat{\mathcal{H}}$ is unconstrained we have: $\frac{p^z_\mathrm{s}(z)}{p^z_\mathrm{t}(z)} \in (\partial\phi^*)(\hat{\ell}(\hat{h}'(z), \hat{h}(z))$ for any $z \in \mathrm{supp}(p^z_\mathrm{t})$. From the definition of Fenchel conjugate we have:

$$x \in \partial\phi^*(t) \iff \phi(x) + \phi^*(t) = xt \iff \phi'(x) = t.$$

Plugging $x = p^z_\mathrm{s}(z)/p^z_\mathrm{t}(z)$ and $t = \ell(\hat{h}'(z), \hat{h}(z))$ we obtain $\ell(\hat{h}'(z), \hat{h}(z)) = \phi'(p^z_\mathrm{s}(z)/p^z_\mathrm{t}(z))$. Hence, from the definition of f-divergences (Definition 1) and its variational characterization (eq. 2.2), we write:

$$\max_{\hat{h}' \in \hat{\mathcal{H}}} d_{s,t} = D_\phi(P^z_\mathrm{s}||P^z_\mathrm{t}). \tag{B.18}$$

□

C. Connection to previous frameworks

In this appendix we show that f-DAL encompasses previous frameworks on domain adaptation, including $\mathcal{H}\Delta\mathcal{H}$-divergence, DANN (Ganin et al., 2016) and MDD (Zhang et al., 2019).

C.1. $\mathcal{H}\Delta\mathcal{H}$-divergence

We now show that Theorem 2 generalizes the bound proposed in (Ben-David et al., 2010a). Let the pair $\{\phi(x), \phi^*(t)\} = \{\frac{1}{2}|x-1|, t\}$ for $t \in [0,1]$, such that $\mathrm{D}^\phi_{h,\mathcal{H}} = \mathrm{D}^{\mathrm{TV}}_{h,\mathcal{H}}$ and $\sup_{h\in\mathcal{H}} \mathrm{D}^{\mathrm{TV}}_{h,\mathcal{H}} = \mathrm{D}^{\mathrm{TV}}_{\mathcal{H}} = \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}$, with $d_{\mathcal{H}\Delta\mathcal{H}}$ defined in (Ben-David et al., 2010a) (see also (A.1)). Theorem 2 gives us that $R^\ell_T(h) \leq R^\ell_S(h) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}} + \lambda^*$, recovering Theorem 2 of (Ben-David et al., 2010a).

C.2. DANN formulation and JS divergence

The DANN formulation by Ganin & Lempitsky (2015) can also be incorporated in our framework if one takes $\hat{\ell}(\hat{h}' \circ g(x), e_1) = \log \sigma(e_1 \cdot \hat{h}' \circ g(x))$ and $\phi^*(t) = -\log(1 - e^t)$, where $\sigma(x) := \frac{1}{1+\exp(-x)}$ is the sigmoid function, and e_1 corresponds to the standard basis vector. Reinterpreting $\hat{h}' := e_1 \cdot \hat{h}'$, sustituting and computing $d_{s,t}$ we obtain:

$$d_{s,t} = \mathbb{E}_{x_s \sim p_s} \log \sigma \circ \hat{h}' \circ g(x_s) + \mathbb{E}_{x_t \sim p_t} \log \Big(1 - \sigma \circ \hat{h}' \circ g(x_t)\Big) \tag{C.1}$$

$$= -\left[\mathbb{E}_{x_s \sim p_s} \log \frac{1}{\sigma \circ \hat{h}' \circ g(x_s)} + \mathbb{E}_{x_t \sim p_t} \log \frac{1}{1 - \sigma \circ \hat{h}' \circ g(x_t)}\right], \tag{C.2}$$

which is equivalent with the second part of the expression show in equation 9 in (Ganin et al., 2016).

Effectively, this formulation ignores the contribution of the source classifier $\hat{h}'$. In fact, *it assumes the output of the source classifier is always constant* (e.g $\hat{h} = e_1$). Notice that this is corrected in f-DAL where $\hat{\ell}(a, b) = g(b_{\text{argmax } a})$. We experimentally also observed that this formulation leads to an inferior performance. Nonetheless, the following proposition shows that under the assumption of an optimal domain classifier $\hat{h}'$, $d_{s,t}$ achieves JS-divergence (up to a constant shift), which upper bounds the $\mathrm{D}^{\mathrm{JS}}_{\hat{h},\mathcal{H}}$.

Proposition 2. *Suppose $d_{s,t}$ follows the form of eq. C.1 and $\hat{h}$ is the optimal domain classifier which is unconstrained, then* $\max_{\hat{h}'} d_{s,t} = D_{JS}(S||T) - 2\log 2$.

Proof. For simplicity in the notation let $\hat{h}' := \sigma \circ (e_1 \cdot \hat{h}')$, rewritting eq. C.1 we have:

$$d_{s,t}(\hat{h}', g) = \int_{\mathcal{Z}} p^z_{\mathrm{s}}(z) \log \hat{h}'(z) + p^z_{\mathrm{t}}(z) \log(1 - \hat{h}'(z)) dz. \tag{C.3}$$

By taking derivatives and finding the optimal $\hat{h}^*(z)$, we get : $h^*(z) = \frac{p^z_s(z)}{p^z_{\mathrm{s}}(z) + p^z_{\mathrm{t}}(z)}$.

By plugging $\hat{h}^*(z)$ into (C.1), rearranging, and using the definition of the Jensen-Shanon (JS) divergence, we get the desired result. □

It is worth noting that the additional negative constant $-2\log 2$ does not affect the optimization.

C.3. MDD formulation and γ-weighted JS divergence

Now let us demonstrate how our f-DAL framework incorporates MDD naturally. Suppose $\phi^*(t) = -\frac{1}{\gamma}\log(1 - e^t)$ and $\hat{\ell}(\hat{h}(z), \hat{h}'(z)) = \log \hat{h}'(z)_{\text{argmax } \hat{h}(z)}$. We retrieve the following result as in Zhang et al. (2019):

Proposition 3 (Zhang et al. (2019)). *Suppose $d_{s,t}$ takes the form of MDD, i.e,*

$$\gamma d_{s,t} = \gamma \mathbb{E}_{z \sim p^z_s} \log \hat{h}'(z)_{\mathit{argmax}\, \hat{h}(z)} + \mathbb{E}_{z \sim p^z_t} \hat{h}(z) \cdot \log(1 - \hat{h}'(z)_{\mathit{argmax}\, \hat{h}(z)}). \tag{C.4}$$

With unconstrained function class $\hat{\mathcal{H}}$, the optimal $d_{s,t}$ satisfies:

$$\max_{\hat{h}'} \gamma d_{s,t} = (\gamma + 1)\mathrm{JS}_\gamma(p^z_s \| p^z_t) + \gamma \log \gamma - (\gamma + 1)\log(\gamma + 1), \tag{C.5}$$

where $\mathrm{JS}_\gamma(p^z_s \| p^z_t)$ is γ-weighted Jensen–Shannon divergence (Huszár, 2015; Nowozin et al., 2016):

$$\mathrm{JS}_\gamma(p^z_s \| p^z_t) = \frac{\gamma}{\gamma + 1}\mathrm{KL}(p^z_s \| \frac{\gamma p^z_s + p^z_t}{\gamma + 1}) + \frac{1}{\gamma + 1}\mathrm{KL}(p^z_t \| \frac{\gamma p^z_s + p^z_t}{\gamma + 1}). \tag{C.6}$$

We remark that when $\gamma = 1$, $\mathrm{JS}_\gamma(p^z_{\mathrm{s}} \| p^z_{\mathrm{t}})$ is the original Jensen–Shannon divergence. One should also note the the additional negative constant $\gamma \log \gamma - (\gamma + 1)\log(\gamma + 1)$, which attributes to the negativity of MDD, does not affect the optimization.

$\phi^*(t) = -\frac{1}{\gamma}\log(1 - e^t)$ can be considered by rescaling the ϕ^* for the usual JS divergence (see Table 8). In general we can rescale ϕ^* for any f-divergence with the following lemma:

Lemma 4 (Boyd & Vandenberghe (2004)). *For any $\lambda > 0$, the Fenchel conjugate of $\lambda\phi$ is $(\lambda\phi)^*(t) = \lambda\phi^*(t/\lambda)$, with* $\mathrm{dom}(\lambda\phi)^* = \lambda \,\mathrm{dom}\, \phi^*$.

C.4. Revisiting MCD (Saito et al., 2018)

Let's now use f-DAL to revisit MCD. This will allow us to understand the cause of the performance gap. For example, MCD(86.5) vs Ours (89.5) on Office-31. Moreover, it will show us how to improve MCD. Let $\hat{\ell}(c,b) = |c-b|$ in Equation (4.3), and choose ϕ to be the TV (Table 1). We have:

$$\min_{\hat{h}\in\hat{\mathcal{H}},g\in\mathcal{G}} \max_{\hat{h}'\in\hat{\mathcal{H}}} R_s[\hat{h}\circ g] + \mathbb{E}_{p_s}[|\hat{h}'\circ g - \hat{h}\circ g|] - \mathbb{E}_{p_t}[|\hat{h}'\circ g - \hat{h}\circ g|] \tag{C.7}$$

where $\hat{\ell}$ should be in $[-0.5, 0.5]$ to satisfy requirements on ϕ^* (Table 1). Comparing this with MCD we can see **3** key differences. **1)** MCD ignores the second term based on assumptions, further requires careful initialization for $\hat{h}, \hat{h}'$. **2)** The max operator in their case goes over $\hat{h}$ and $\hat{h}'$. This makes optimization harder (see Zhang et al. (2019)). We do not need this because our bounds are based on $D^{\phi}_{h,\mathcal{H}} \leq D^{\phi}_{\mathcal{H}}$ (definitions 2 and 3, Lemma 1). **3)** The restriction on the $\hat{\ell}(c,b)$ is not taken into account (should be re-weighted or the act. function follow Tab 1). As mentioned in MCD (Eq. 9), $I[c \neq b]$ is similar, but in this context not the same as $|c-b|$. Thus, 1,2,3 could explain the difference in performance 86.5 vs Ours (89.5). We believe using these recommendations on MCD could lead to a powerful algorithm but we defer that to further work.

D. Additional Experimental Results

Table 9. 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
f-DAL (JS)	93.0±1.4	98.8±0.1	100.0±.0	92.8±0.4	74.9±1.5	73.3±0.1	88.8
f-DAL (Pearson χ^2)	95.4±0.7	98.4±0.2	100.0±.0	93.8±0.4	73.5±1.1	74.2±0.5	**89.2**
f-DAL(γ-JS) / MDD + Alignment (Jiang et al., 2020)	90.3±0.2	98.7±0.1	99.8±.0	92.1±0.5	75.3±0.2	74.9±0.3	88.8
f-DAL (Pearson χ^2) + 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 10. 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
f-DAL (JS)	53.7	71.7	76.3	60.2	68.4	69.0	60.2	52.6	76.9	71.4	59.0	81.8	66.8
f-DAL (Pearson χ^2)	54.7	69.4	77.8	61.0	72.6	72.2	60.8	53.4	80.0	73.3	60.6	83.8	68.3
f-DAL(γ-JS) / MDD + Alignment (Jiang et al., 2020)	56.2	77.9	79.2	64.4	73.1	74.4	64.2	54.2	79.9	71.2	58.1	83.1	69.5
f-DAL (Pearson χ^2) + 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 11. 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	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
f-DAL (JS)	83.2	78.8	80.4	80.2	79.4	82.9	72.3	76.3	87.8	74.7	78.5	87.0	80.1
f-DAL (Pearson χ^2)	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**

D.1. Experimental results with others γ-shifted divergences

In this section, we show experiments on the Digits Benchmark (Avg on 3 runs) for a shifted γ-Pearson χ^2. We follow Section 4.3 and let $\hat{\phi}(x) = \phi(x) - \gamma x$. Results shown in Table 14 are similar to those obtained for the γ-JS (Table 3), for

Table 12. 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
f-DAL (JS)	95.3	98.0	**96.6**
f-DAL (Pearson χ^2)	95.3	97.3	96.3

Table 13. p-values Significance Test (Wilcoxon signed rank test)

	Digits	NLP	Office-31	Office-Home
Avg DANN	93.3	76.3	82.2	57.6
Avg f-DAL JS	96.6	80.1	88.8	66.8
p-val	0.5	0.0025	0.031	0.0025

which our test showed no significance to have γ. We also conducted experiments for the other modality, e.g. NLP data, with γ-JS. Similarly, we observed results are not significant wrt JS(γ=3, Avg=80.4) and slightly worse than Pearson.

Table 14. γ-shifted Pearson χ^2 Digits Benchmark.

γ	Avg Digits
-	*96.3*
2	96.2
3	**96.4**
4	96.3

D.2. Robustness to Label Shift

In this section, we compare the robustness to label-shift of f-DAL-JS vs DANN on the digits benchmark. Specifically, we consider the task M→ U and artificially generate different version of the target dataset where data-points are re-sampled in terms of its classes. This way we can have control over the JS divergence between the label distribution (i.e $JS(P_s(y)||P_t(y))$) and compare at different levels. Figure 7 shows the results. Firstly, we can observe that both methods performance degrades as the distance between label distributions increases. This is an expected behavior in DA, and can also be explained with our theory. For example, as this distance increases, the term λ^* in Theorem 2 simply increases, and thus this cannot be assumed to be negligible. To explicitly see why, we refer the reader to Zhao et al. (2019) where the authors derived a lower bound for joint risk. It is important to also have in mind that λ^* incorporates the notion of adaptability. That is, if the optimal hypothesis performs poorly in either domain, adaptation is simply not possible and thus assumptions are need it. Secondly, from the figure, we can also see our method is more robust to label-shift than DANN. Indeed, we fit linear regression models to highlight the trend and show the value of the slope in each case. The performance comparison is noticeable. We emphasize the aim of this experiment is to showcase the robustness of f-DAL-JS vs DANN when label-shift exists. Our method does not propose any additional correction or term to deal with this and doing so (i.e dealing explicitly with label-shift) is out-of-the-scope of this work. Our algorithm follows the common assumption stated on adversarial DA methods and let λ^* to be negligible. We believe the better performance of f-DAL-JS vs DANN under label-shift is just a consequence of directly connecting theory and algorithm. We additionally show f-DAL can be perfectly combined with methods that deal with label shift such as Implicit Alignment (i.e Jiang et al. (2020)) (Tables 9 and 10). Indeed, doing so leads to SoTA results on the Office-Home dataset (Table 10). This again showcases the versatility of f-DAL.

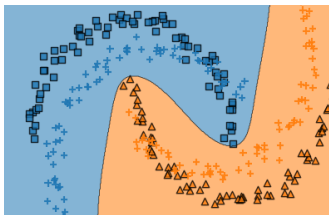

Figure 6. Domain Adaptation. A learner trained on abundant labeled data (marked as squares, colors are categories) is expected to perform well in the target domain (marked as +). Decision boundaries correspond to a 2-layers neural net trained using f-DAL.

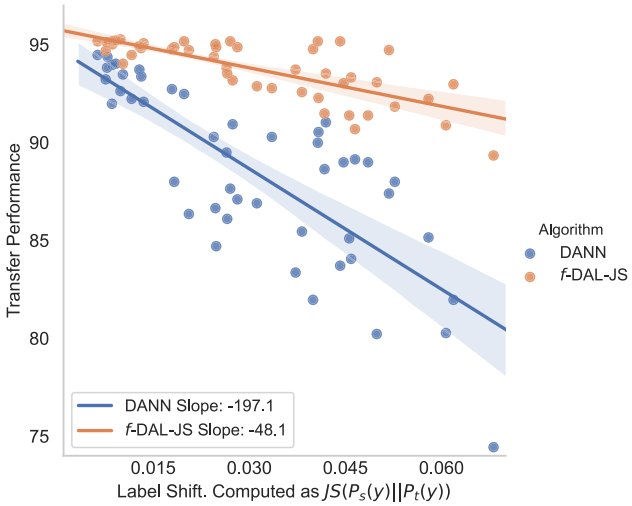

Figure 7. Robustness to Label Shift f-DAL-JS vs DANN. The x-axis represents the Jensen-Shanon distance between the label distributions. We can observe that f-DAL-JS is more robust to label shift than DANN. Linear regression models are fit to highlight the trend(slope is also shown). (Dataset M $\rightarrow$ U).

E. More Details on Experimental Setup

Our algorithm is implemented 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. We train for 30 epochs, the optimizer is SGD with Nesterov Momentum (momentum 0.9, batch size 128), the learning rate is 0.01. The regularization term for the discrepancy is set to 0.5 and the GRL coefficient set to 0.6. We use a weight decay coefficient of 0.002. Hyperparameters follow closely the ones used by Long et al. (2018), if some differ slightly, they were determined in a subset(10%) of the training set of the task M$\rightarrow$U and kept constant for the other task. We use three different seeds (i.e 1,2,3) and report the average over the runs.

For the NLP task, we follow the standard protocol from Courty et al. (2017); Ganin et al. (2016) and use simple 2-layer model with sigmoid activation function. Thus, the main classifier ($\hat{h}$) and auxiliary classifier ($\hat{h}'$) are a simple linear layer with BN. We train for 10 epochs, the optimizer is SGD with Nesterov Momentum (momentum 0.9, batch size 16), the learning rate is 0.001. We use three different seeds (i.e 1,2,3) and report the average over the runs. The regularization term for the discrepancy is set to 1 and the GRL coefficient set to 0.1. We use a weight decay coefficient of 0.002. Hyper-parameters are empirically determined in a subset(10%) of the training set of the task (B$\rightarrow D$) and kept constant for the others.

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 to avoid gradient issues and instabilities during training for some divergences in the first epochs. We use the hyperparams and same training protocol from MDD (Zhang et al. (2019) and CDAN (Long et al. (2018)). We report the average accuracies over 3 experiments.

Experiments are conducted on NVIDIA Titan V (Digits, NLP) and V100 (Visual Tasks) GPU cards.