On f-Divergence Principled Domain Adaptation: An Improved Framework

Ziqiao Wang[1] Yongyi Mao[1]

[1]University of Ottawa

June 4, 2024

Outline

1. Problem Setup
2. Previous Divergence-based Domain Learning Theory
3. Improved f-divergence Guided UDA Theory

Outline

1. Problem Setup
2. Previous Divergence-based Domain Learning Theory
3. Improved f-divergence Guided UDA Theory

Domain Adaptation

- Given data from a source domain, i.e. $\{X_i, Y_i\} \overset{i.i.d.}{\sim} \mu$
- Obtain a model for a target domain, i.e. $\{X, Y\} \sim \nu$
- **Practical Goal**: Efficiently transfer ML models between related populations at low cost.

Formal Notations

- Data space: $\mathcal{X} \times \mathcal{Y}$; Hypothesis space: $\mathcal{H} \triangleq \{h : \mathcal{X} \to \mathcal{Y}\}$;

Formal Notations

- Data space: $\mathcal{X} \times \mathcal{Y}$; Hypothesis space: $\mathcal{H} \triangleq \{h : \mathcal{X} \to \mathcal{Y}\}$;
- **Unsupervised Domain Adaptation (UDA)**:
  - Unknown distributions μ and ν
  - Labeled source-domain sample $\mathcal{S} = \{X_i, Y_i\}_{i=1}^{n} \sim \mu^{\otimes n}$
  - Unlabelled target-domain sample $\mathcal{T} = \{X_j\}_{j=1}^{m} \sim \nu^{\otimes m}$
  - **Target**: find a hypothesis $h \in \mathcal{H}$ "works well" on ν.

Formal Notations

- Data space: $\mathcal{X} \times \mathcal{Y}$; Hypothesis space: $\mathcal{H} \triangleq \{h : \mathcal{X} \to \mathcal{Y}\}$;
- **Unsupervised Domain Adaptation (UDA)**:
  - Unknown distributions μ and ν
  - Labeled source-domain sample $S = \{X_i, Y_i\}_{i=1}^{n} \sim \mu^{\otimes n}$
  - Unlabelled target-domain sample $T = \{X_j\}_{j=1}^{m} \sim \nu^{\otimes m}$
  - **Target**: find a hypothesis $h \in \mathcal{H}$ "works well" on ν.
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_0^+$.
- Target error: $R_\nu(h) \triangleq \mathbb{E}_{(X,Y)\sim\nu}\left[\ell(h(X), Y)\right]$, same way for the source error, $R_\mu(h)$.

Formal Notations

- Data space: $\mathcal{X} \times \mathcal{Y}$; Hypothesis space: $\mathcal{H} \triangleq \{h : \mathcal{X} \to \mathcal{Y}\}$;
- **Unsupervised Domain Adaptation (UDA)**:
  - Unknown distributions μ and ν
  - Labeled source-domain sample $S = \{X_i, Y_i\}_{i=1}^{n} \sim \mu^{\otimes n}$
  - Unlabelled target-domain sample $T = \{X_j\}_{j=1}^{m} \sim \nu^{\otimes m}$
  - **Target**: find a hypothesis $h \in \mathcal{H}$ "works well" on ν.
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_0^+$.
- Target error: $R_\nu(h) \triangleq \mathbb{E}_{(X,Y)\sim\nu}\left[\ell(h(X), Y)\right]$, same way for the source error, $R_\mu(h)$.
- We use $\ell(h, h')$ to denote $\ell(h(x), h'(x))$, i.e. the disagreement of h and h' on x.

Outline

$\mathcal{H}$-specified Discrepancy

By Ben-David et al. [2006, 2010], Mansour et al. [2009]:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mu,\nu) \triangleq \sup_{h,h'\in\mathcal{H}} \left|\mathbb{E}_\mu\left[\ell(h,h')\right] - \mathbb{E}_\nu\left[\ell(h,h')\right]\right|.$$

$\mathcal{H}$-specified Discrepancy

By Ben-David et al. [2006, 2010], Mansour et al. [2009]:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mu,\nu) \triangleq \sup_{h,h'\in\mathcal{H}} \left|\mathbb{E}_{\mu}\left[\ell(h,h')\right] - \mathbb{E}_{\nu}\left[\ell(h,h')\right]\right|.$$

- Assumptions:
  - Triangle property: $\ell(y_1,y_2) \leq \ell(y_1,y_3) + \ell(y_3,y_2)$ for any $y_1,y_2,y_3 \in \mathcal{Y}$.
  - Bounded loss: e.g., $\ell \in [0,1]$

$\mathcal{H}$-specified Discrepancy

By Ben-David et al. [2006, 2010], Mansour et al. [2009]:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mu,\nu) \triangleq \sup_{h,h'\in\mathcal{H}} \left|\mathbb{E}_\mu\left[\ell(h,h')\right] - \mathbb{E}_\nu\left[\ell(h,h')\right]\right|.$$

- Assumptions:
  - Triangle property: $\ell(y_1,y_2) \leq \ell(y_1,y_3) + \ell(y_3,y_2)$ for any $y_1,y_2,y_3 \in \mathcal{Y}$.
  - Bounded loss: e.g., $\ell \in [0,1]$

Then, for any $h \in \mathcal{H}$,

$$R_\nu(h) \leq R_\mu(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mu,\nu) + \lambda^*,$$

where $\lambda^* = \min_{h^*\in\mathcal{H}} R_\nu(h^*) + R_\mu(h^*)$.

$\mathcal{H}$-specified Discrepancy

By Ben-David et al. [2006, 2010], Mansour et al. [2009]:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mu,\nu) \triangleq \sup_{h,h'\in\mathcal{H}} \left|\mathbb{E}_\mu\left[\ell(h,h')\right] - \mathbb{E}_\nu\left[\ell(h,h')\right]\right|.$$

- Assumptions:
  - Triangle property: $\ell(y_1,y_2) \leq \ell(y_1,y_3) + \ell(y_3,y_2)$ for any $y_1,y_2,y_3 \in \mathcal{Y}$.
  - Bounded loss: e.g., $\ell \in [0,1]$

Then, for any $h \in \mathcal{H}$,

$$R_\nu(h) \leq R_\mu(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mu,\nu) + \lambda^*,$$

where $\lambda^* = \min_{h^*\in\mathcal{H}} R_\nu(h^*) + R_\mu(h^*)$.

Can we extend $\mathcal{H}\Delta\mathcal{H}$-divergence to $\mathcal{H}$-specified f-divergence?

From $\mathcal{H}\Delta\mathcal{H}$-divergence to $\mathcal{H}$-specified f-divergence

- f-divergence: $\mathrm{D}_\phi(P||Q) \triangleq \mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\right]$, where ϕ is convex and $\phi(1)=0$.

From $\mathcal{H}\Delta\mathcal{H}$-divergence to $\mathcal{H}$-specified f-divergence

- f-divergence: $\mathrm{D}_\phi(P||Q) \triangleq \mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\right]$, where ϕ is convex and $\phi(1) = 0$.
- Its variational formula:

$$\mathrm{D}_\phi(P||Q) = \sup_{g\in\mathcal{G}} \mathbb{E}_{\theta\sim P}\left[g(\theta)\right] - \mathbb{E}_{\theta\sim Q}\left[\phi^*(g(\theta))\right].$$

From $\mathcal{H}\Delta\mathcal{H}$-divergence to $\mathcal{H}$-specified f-divergence

- f-divergence: $\mathrm{D}_\phi(P||Q) \triangleq \mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\right]$, where ϕ is convex and $\phi(1)=0$.
- Its variational formula:

$$\mathrm{D}_\phi(P||Q) = \sup_{g\in\mathcal{G}} \mathbb{E}_{\theta\sim P}\left[g(\theta)\right] - \mathbb{E}_{\theta\sim Q}\left[\phi^*(g(\theta))\right].$$

- By Acuna et al. [2021]:

$$\widetilde{\mathrm{D}}_\phi^{h,\mathcal{H}}(\mu||\nu) \triangleq \sup_{h'\in\mathcal{H}} \left|\mathbb{E}_\mu\left[\ell(h,h')\right] - \mathbb{E}_\nu\left[\phi^*(\ell(h,h'))\right]\right|.$$

$\Longrightarrow$ Additional absolute value function added.

Gap between Theory and Algorithm in Acuna et al. [2021]

$$\widetilde{\mathrm{D}}_{\phi}^{h,\mathcal{H}}(\mu||\nu) \triangleq \sup_{h'\in\mathcal{H}} \left|\mathbb{E}_{\mu}\left[\ell(h,h')\right] - \mathbb{E}_{\nu}\left[\phi^*(\ell(h,h'))\right]\right|.$$

▷ Theory (Target Error Bound):

$$R_{\nu}(h) \leq R_{\mu}(h) + \widetilde{\mathrm{D}}_{\phi}^{h,\mathcal{H}}(\mu||\nu) + \lambda^*,$$

Gap between Theory and Algorithm in Acuna et al. [2021]

$$\widetilde{\mathrm{D}}_\phi^{h,\mathcal{H}}(\mu||\nu) \triangleq \sup_{h'\in\mathcal{H}} \left|\mathbb{E}_\mu\left[\ell(h,h')\right] - \mathbb{E}_\nu\left[\phi^*(\ell(h,h'))\right]\right|.$$

- ▷ Theory (Target Error Bound):

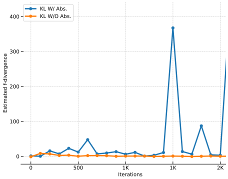
$$R_\nu(h) \le R_\mu(h) + \widetilde{\mathrm{D}}_\phi^{h,\mathcal{H}}(\mu||\nu) + \lambda^*,$$

- ▷ f-Domain Adversarial Learning (f-DAL) Algorithm:

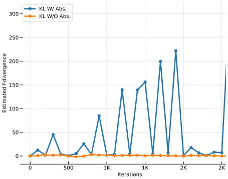
$$\min_h R_{\hat\mu}(h) + \underbrace{\max_{h'} \mathbb{E}_{\hat\mu}\left[\ell(h,h')\right] - \mathbb{E}_{\hat\nu}\left[\phi^*(\ell(h,h'))\right]}_{d(\hat\mu,\hat\nu;h)}.$$

$\Longrightarrow$ $d(\hat\mu,\hat\nu;h)$ drops the absolute value function compared with $\widetilde{\mathrm{D}}_\phi^{h,\mathcal{H}}(\mu||\nu)$

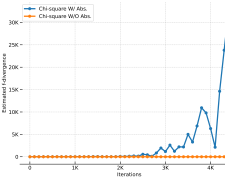
Overestimation by Absolute Value Function



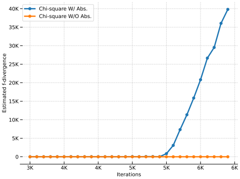
(a) KL (Office-31)



(b) KL (Office-Home)



(c) χ^2 (Office-31)



(d) χ^2 (Office-Home)

Figure 1: The y-axis is the estimated corresponding f-divergence and the x-axis is the number of iterations.

- f-DAL algorithm fails if the absolute value function is added.

Outline

Our work: New f-Domain Discrepancy (f-DD)

- Using "linear transformation" instead of the absolute value function:
  - Original variational formula:

$$\mathrm{D}_\phi(P||Q) = \sup_{g\in\mathcal{G}} \mathbb{E}_{\theta\sim P}\left[g(\theta)\right] - \mathbb{E}_{\theta\sim Q}\left[\phi^*(g(\theta))\right]. \tag{1}$$

  - Reparameterization of $g \to tg + \alpha$ (i.e. linear transformation):

$$\mathrm{D}_\phi(P||Q) = \sup_{g,t,\alpha} \mathbb{E}_{\theta\sim P}\left[tg(\theta)+\alpha\right] - \mathbb{E}_{\theta\sim Q}\left[\phi^*(tg(\theta)+\alpha)\right]. \tag{2}$$

Eq. (2) is tighter than Eq. (1)

Our work: New f-Domain Discrepancy (f-DD)

- Using "linear transformation" instead of the absolute value function:
  - Original variational formula:

$$\mathrm{D}_\phi(P||Q) = \sup_{g\in\mathcal{G}} \mathbb{E}_{\theta\sim P}\left[g(\theta)\right] - \mathbb{E}_{\theta\sim Q}\left[\phi^*(g(\theta))\right]. \quad (1)$$

  - Reparameterization of $g \to tg + \alpha$ (i.e. linear transformation):

$$\mathrm{D}_\phi(P||Q) = \sup_{g,t,\alpha} \mathbb{E}_{\theta\sim P}\left[tg(\theta)+\alpha\right] - \mathbb{E}_{\theta\sim Q}\left[\phi^*(tg(\theta)+\alpha)\right]. \quad (2)$$

  Eq. (2) is tighter than Eq. (1)

- Our f-DD:

$$\mathrm{D}_\phi^{h,\mathcal{H}}(\nu||\mu) \triangleq \sup_{t\in\mathbb{R},h'} \mathbb{E}_\nu\left[t\ell(h,h')\right] - \inf_{\alpha\in\mathbb{R}} \mathbb{E}_\mu\left[\phi^*(t\ell(h,h')+\alpha)-\alpha\right].$$

f-DD-based Theory

$$\mathrm{D}_{\phi}^{h,\mathcal{H}}(\nu||\mu) \triangleq \sup_{t\in\mathbb{R},h'} \mathbb{E}_{\nu}\left[t\ell(h,h')\right] - \inf_{\alpha\in\mathbb{R}} \mathbb{E}_{\mu}\left[\phi^*(t\ell(h,h')+\alpha)-\alpha\right].$$

- Target Error Bound: For any $h \in \mathcal{H}$,

$$R_{\nu}(h) \leq R_{\mu}(h) + \inf_{t\geq 0} \frac{\mathrm{D}_{\phi}^{h,\mathcal{H}}(\nu||\mu) + K_{\mu}(t)}{t} + \lambda^*, \tag{3}$$

  where $K_{\mu}(t)$ is the upper bound for the "cumulant generating function (CGF)" for μ.

- If ϕ is twice differentiable and ϕ'' is monotone, then

$$R_{\nu}(h) \leq R_{\mu}(h) + \sqrt{\frac{2}{\phi''(1)}\mathrm{D}_{\phi}^{h,\mathcal{H}}(\nu||\mu)} + \lambda^*. \tag{4}$$

Localization Technique

- Restricted Hypothesis Space (Rashomon set): $\mathcal{H}_r \triangleq \{h \in \mathcal{H} | R_\mu(h) \leq r\}$
- Localized f-DD: For a given $h \in \mathcal{H}_{r_1}$

$$\mathrm{D}_\phi^{h,\mathcal{H}_r}(\nu||\mu) \triangleq \sup_{h' \in \mathcal{H}_r, t \geq 0} \mathbb{E}_\nu \left[t\ell(h,h')\right] - \inf_{\alpha \in \mathbb{R}} \mathbb{E}_\mu \left[\phi^*(t\ell(h,h') + \alpha) - \alpha\right].$$

Localization Technique

- Restricted Hypothesis Space (Rashomon set): $\mathcal{H}_r \triangleq \{h \in \mathcal{H} | R_\mu(h) \leq r\}$
- Localized f-DD: For a given $h \in \mathcal{H}_{r_1}$

$$\mathrm{D}_\phi^{h,\mathcal{H}_r}(\nu||\mu) \triangleq \sup_{h' \in \mathcal{H}_r, t \geq 0} \mathbb{E}_\nu \left[t\ell(h,h')\right] - \inf_{\alpha \in \mathbb{R}} \mathbb{E}_\mu \left[\phi^*(t\ell(h,h') + \alpha) - \alpha\right].$$

- Target Error Bound:
For any h, h' and $C_1, C_2 > 0$ satisfying
$\inf_\alpha \mathbb{E}_\mu \left[\phi^*(C_1\ell(h,h') + \alpha) - \alpha\right] \leq C_1(1 + C_2)\mathbb{E}_\mu \left[\ell(h,h')\right]$, then:

$$R_\nu(h) \leq R_\mu(h) + \frac{1}{C_1}\mathrm{D}_\phi^{h,\mathcal{H}_r}(\nu||\mu) + C_2 R_\mu^r(h) + \lambda_r^*,$$

where $\lambda_r^* = \min_{h^* \in \mathcal{H}_r} R_\mu(h^*) + R_\nu(h^*)$ and $R_\mu^r(h) = \sup_{h' \in \mathcal{H}_r} \mathbb{E}_\mu \left[\ell(h,h')\right]$.

Localization Technique

- Target Error Bound:

$$R_\nu(h) \leq R_\mu(h) + \frac{1}{C_1}\mathrm{D}_\phi^{h,\mathcal{H}_r}(\nu||\mu) + C_2 R_\mu^r(h) + \lambda_r^*.$$

  - $R_\mu^r(h) \leq r + r_1 \Longrightarrow$ Small r, r_1
  - If $r < \lambda^*$, then it's possible that $\lambda_r^* > \lambda^* \Longrightarrow$ Large r

Localization Technique

- Target Error Bound:

$$R_\nu(h) \leq R_\mu(h) + \frac{1}{C_1}\mathrm{D}_\phi^{h,\mathcal{H}_r}(\nu||\mu) + C_2 R_\mu^r(h) + \lambda_r^*.$$

  - $R_\mu^r(h) \leq r + r_1 \Longrightarrow$ Small r, r_1
  - If $r < \lambda^*$, then it's possible that $\lambda_r^* > \lambda^* \Longrightarrow$ Large r

- Localized KL-DD:
$\inf_\alpha \mathbb{E}_\mu \left[\phi^*(C_1\ell(h,h') + \alpha) - \alpha\right] \leq C_1(1 + C_2)\mathbb{E}_\mu \left[\ell(h,h')\right]$

Localization Technique

- Target Error Bound:

$$R_\nu(h) \leq R_\mu(h) + \frac{1}{C_1}\mathrm{D}_\phi^{h,\mathcal{H}_r}(\nu||\mu) + C_2 R_\mu^r(h) + \lambda_r^*.$$

  - $R_\mu^r(h) \leq r + r_1 \Longrightarrow$ Small r, r_1
  - If $r < \lambda^*$, then it's possible that $\lambda_r^* > \lambda^* \Longrightarrow$ Large r

- Localized KL-DD:
$\inf_\alpha \mathbb{E}_\mu \left[\phi^*(C_1\ell(h,h') + \alpha) - \alpha\right] \leq C_1(1+C_2)\mathbb{E}_\mu \left[\ell(h,h')\right]$
$$\Longleftarrow \begin{cases} C_1 > 0 \\ C_2 \in (0,1) \\ \left(e^{C_1} - C_1 - 1\right)\left(1 + C_2^2 \min\{r_1 + r, 1\}\right) \leq C_1 C_2 \end{cases}$$

Generalization Bound via Localized f-DD

Theorem (informal)

For any $h \in \mathcal{H}_{r_1}$, w.p. at least $1-\delta$, we have

$$R_\nu(h) \leq R_{\hat{\mu}}(h) + \frac{\mathrm{D}_{\mathrm{KL}}^{h,\mathcal{H}_r}(\hat{\nu}||\hat{\mu})}{C_1} + C_2 R^r_\mu(h) + \mathcal{O}\left(\frac{\log(1/\delta)}{n} + \frac{\log(1/\delta)}{m}\right) + \mathcal{O}\left(\sqrt{\frac{(r_1+r)\log(1/\delta)}{n}} + \sqrt{\frac{r\log(1/\delta)}{m}}\right) + \text{Complexity.} + \lambda^*_r.$$

Small $r, r_1 \implies$ fast decaying rate (i.e. $\mathcal{O}\left(\frac{1}{n}+\frac{1}{m}\right)$).

Experiments

- Three specific discrepancy measures:
  - KL-DD, χ^2-DD, the weighted Jeffereys-DD: $\gamma_1 \mathrm{D}_{\mathrm{KL}}(\hat{\mu}||\hat{\nu}) + \gamma_2 \mathrm{D}_{\mathrm{KL}}(\hat{\nu}||\hat{\mu})$
- Objective Function:
  Bounded $\ell \to$ Unbounded $\hat{\ell}$ (Optimizing over t may not be necessary)

$$\min_h R_{\hat{\mu}}(h) + \max_{h'} \left\{ \mathbb{E}_{\hat{\mu}} \left[\hat{\ell}(h,h') \right] - \inf_\alpha \mathbb{E}_{\hat{\nu}} \left[\phi^*(\hat{\ell}(h,h') + \alpha) - \alpha \right] \right\}.$$

Table 1: Accuracy (%) on UDA Classification Tasks

Method	Office-31	Office-Home	Digits
Acuna et al. [2021]	89.5	68.5	96.3
Our weighted Jeffereys-DD	**90.1**	**70.2**	**97.1**

Summary

- Significant gap between previous f-divergence-based domain learning theory and algorithm in Acuna et al. [2021]
- We propose new f-divergence-based domain learning theory
- We further improve the target error bound by the localization technique
- Our weighted Jeffereys-DD outperforms previous methods
- For further details, including optimization on t, t-SNE visualization, and more, please refer to our paper available at: https://arxiv.org/pdf/2402.01887

References I

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *The 22nd Conference on Learning Theory*, 2009.

David Acuna, Guojun Zhang, Marc T Law, and Sanja Fidler. f-domain adversarial learning: Theory and algorithms. In *International Conference on Machine Learning*, pages 66–75. PMLR, 2021.