On f-Divergence Principled Domain Adaptation: An Improved Framework

Ziqiao Wang [1] Yongyi Mao [2]

[1]Tongji University [2]University of Ottawa

Unsupervised Domain Adaptation

- 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 $\mathcal{T}=\{X_j\}_{j=1}^m \sim \nu_{\mathcal{X}}^{\otimes m}$
- Goal: Efficiently transfer ML models between related domains at low cost
  $\Longrightarrow$ Find a hypothesis $h\in\mathcal{H}$ "works well" on ν

Limitations of Previous f-Divergence-based DA Works

- Relying on a Weak Variational Representation (i.e. Eq. 1) 🤔
  $\Longrightarrow$ Cannot recover Donsker and Varadhans representation of KL divergence.
- Slow Rate of Sample Complexitiy Bound ☹️ $\Longrightarrow$ e.g., $\mathcal{O}(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}})$.
- Gap Between Theory and Algorithm 😟
  $\Longrightarrow$ e.g., Overestimation of the f-divergence in [1].

Main Contributions

- We design a novel f-divergence-based domain discrepancy measure, termed f-DD, and derive an upper bound for the target error.
- To improve the convergence rate of our f-DD-based bound, we refine it using a localization technique.
- Our f-DD outperforms previous f-divergence-based algorithms on three popular UDA benchmarks.

Background on f-Divergence

- (f-Divergence) Let P and Q be two distributions on Θ. Let $\phi:\mathbb{R}_+\to\mathbb{R}$ be a convex function with $\phi(1)=0$. If $P\ll Q$, then

$$\mathrm{D}_\phi(P||Q)\triangleq\mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\right],$$

  e.g., Total variation, KL, χ^2, squared Hellinger, Jeffreys, Jensen-Shannon, etc.
- Variational Representation of f-divergence.
  - Original Legendre Transformation:

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

  - Reparameterization of $g\to g+\alpha$ ("Shift Transformation") [2]

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

  Eq. (2) is point-wise "tighter" than Eq. (1)

  $\Downarrow$

- Example: Donsker and Varadhans (DV) representation of KL divergence: $\phi(x)=x\log x-x+1$, then $\phi^*(y)=e^y-1$
  - By Eq. (1)

$$\mathrm{D}_{\mathrm{KL}}(P||Q)=\sup_{g\in\mathcal{G}}\mathbb{E}_P[g(\theta)]-\mathbb{E}_Q\left[e^{g(\theta)}-1\right]. \quad (3)$$

  - By Eq. (2)

$$\mathrm{D}_{\mathrm{KL}}(P||Q)=\sup_{g\in\mathcal{G}}\mathbb{E}_P[g(\theta)]-\log\mathbb{E}_Q\left[e^{g(\theta)}\right]. \quad (4)$$

  Eq. (4) recovers the DV representation of KL, Eq. (4) is pointwise tighter than Eq. (3) by $\log(x)\le x-1$ for $x>0$.

Discrepancy-based DA Theory: Target Error Bound

- Additional Notations
  - Loss function $\ell:\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}_0^+$. Assumptions : $\begin{cases}\text{Triangle property}\\ \text{Bounded loss}\end{cases}$
  - Target error: $R_\nu(h)\triangleq\mathbb{E}_{(X,Y)\sim\nu}[\ell(h(X),Y)]$, Source error: $R_\mu(h)\triangleq\mathbb{E}_{(X,Y)\sim\mu}[\ell(h(X),Y)]$.
  - We use $\ell(h,h')$ to denote $\ell(h(x),h'(x))$, i.e. the disagreement of h and h' on x.
- [1] defines:

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

$\Longrightarrow$ Additional absolute value function added. 🧐

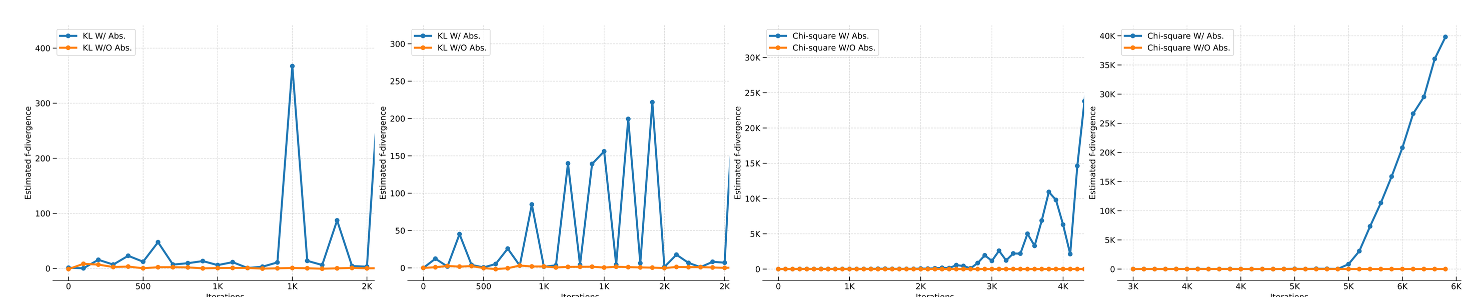
- Theory (Target Error Bound):

$$\text{Target Error}\le\text{Source Error}+\widetilde{\mathrm{D}}_\phi^{h,\mathcal{H}}(\mu||\nu)+\text{Ideal Joint Error}.$$

  $\Longrightarrow$ Absolute value function is necessary for establishing this bound
- f-Domain Adversarial Learning (f-DAL) Algorithm:

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

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



(a) KL (Office31) (b) KL (OfficeHome) (c) χ^2 (Office31) (d) χ^2 (OfficeHome)

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

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

- Our f-DD:

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

$\Longrightarrow$ Introducing the scaling parameter t (i.e. "Affine Transformation") ☺️.

Theorem (informal): f-DD-based Bound

For any $h\in\mathcal{H}$,

$$\text{Target Error}\le\text{Source Error}+\inf_{t\ge0}\frac{\mathrm{D}_\phi^{h,\mathcal{H}}(\nu||\mu)+K_\mu(t)}{t}+\text{Ideal Joint Error},$$

where $K_\mu(t)$ is the upper bound for the "general CGF" for μ.

- Ideal joint error can be $\min_{h^*\in\mathcal{H}}R_\nu(h^*)+R_\mu(h^*)$ [3] or $\min\{R_\nu(f_\mu),R_\mu(f_\nu)\}$ [5].
- If ϕ is twice differentiable and ϕ'' is monotone, then $\inf_{t\ge0}\frac{\mathrm{D}_\phi^{h,\mathcal{H}}(\nu||\mu)+K_\mu(t)}{t}=\sqrt{\frac{2}{\phi''(1)}\mathrm{D}_\phi^{h,\mathcal{H}}(\nu||\mu)}$, e.g., $\phi''(1)=1$ for KL recovers [4, Theorem 4.2].
- Sample complexity bound: w.h.p.,

$$\mathrm{D}_\phi^{h,\mathcal{H}}(\nu||\mu)\le\mathrm{D}_\phi^{h,\mathcal{H}}(\hat{\nu}||\hat{\mu})+\text{Complexitiy Terms}+\mathcal{O}\left(1/\sqrt{n}+1/\sqrt{m}\right).$$

Shaper Bound: Localization Technique

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

$$\boxed{\mathrm{D}_\phi^{h,\mathcal{H}_r}(\nu||\mu)\triangleq\sup_{h'\in\mathcal{H}_r,t\ge0}\mathbb{E}_\nu[t\ell(h,h')]-\inf_{\alpha\in\mathbb{R}}\mathbb{E}_\mu[\phi^*(t\ell(h,h')+\alpha)-\alpha].}$$

Theorem (informal): Localized KL-DD-based Bound

For any $h\in\mathcal{H}_{r_1}$, w.h.p.

$$\text{Target Error}\le\text{Source Train Error}+\mathcal{O}(\mathrm{D}_{\mathrm{KL}}^{h,\mathcal{H}_r}(\hat{\nu}||\hat{\mu}))+\mathcal{O}\left(1/n+1/m\right)$$
$$+\mathcal{O}\left(\sqrt{(r+r_1)/n}+\sqrt{r/m}\right)+(r,r_1)\text{-Related Terms}+\text{Complexity}.$$

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

Algorithms and Experiments

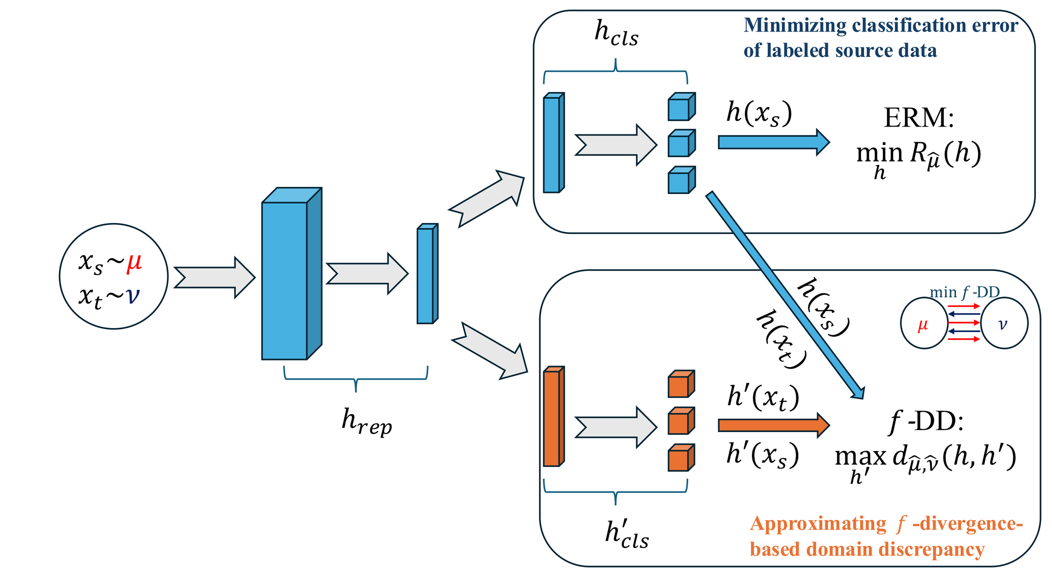


Figure 2. Overview of f-DD.

- 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
f-DAL [1]	89.5	68.5	96.3
Ours (KL-DD)	89.8	69.4	96.9
Ours (χ^2-DD)	89.7	69.2	96.4
Ours (Jefferys-DD)	90.1	70.2	97.1

Key Takeways

1. Utilizing stronger variational formulations of f-divergences can obtain improved results in both theoretical analysis and algorithmic performance.
2. The best performance is achieved by Jeffreys-DD, suggesting the value of manually adjusting the asymmetric properties in domain adaptation.

References

[1] 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.

[2] Rohit Agrawal and Thibaut Horel. Optimal bounds between f-divergences and integral probability metrics. In International Conference on Machine Learning, pages 115–124. PMLR, 2020.

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

[4] Ziqiao Wang and Yongyi Mao. Information-theoretic analysis of unsupervised domain adaptation. In International Conference on Learning Representations, 2023.

[5] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In International Conference on Machine Learning, pages 7523–7532. PMLR, 2019.