Information-Theoretic Analysis of Unsupervised Domain Adaptation

Ziqiao Wang[1] Yongyi Mao[1]

[1]*School of Electrical Engineering and Computer Science, University of Ottawa*

Overview

Unsupervised Domain Adaptation (UDA)

- Train a model on a labeled source sample and an unlabeled target sample.
- Goal: Find a model that performs well on the target domain.

Two Notions of Generalization Errors

- Population-to-population (PP) generalization error: KL based bounds.
- Expected empirical-to-population (EP) generalization error: algorithm-dependent bounds $\Longrightarrow$ two regularization strategies.

Problem Formulation

Setup

- Source data $Z=(X,Y)\sim\mu$; Target data $Z'=(X',Y')\sim\mu'$; Predictor space $\mathcal{F}=\{f_w:\mathcal{X}\to\mathcal{Y}|w\in\mathcal{W}\}$
- Source sample: $S=\{Z_i\}_{i=1}^n$; Target sample $S'_{X'}=\{X'_j\}_{j=1}^m$
- Learning algorithm: $\mathcal{A}:\mathcal{Z}^n\times\mathcal{X}^m\to\mathcal{W}$

Generalization Error

- Population risk of target domain: $R_{\mu'}(w)\triangleq\mathbb{E}_{Z'}[\ell(f_w(X'),Y')]$
- Empirical risk of source domain: $R_S(w)\triangleq\frac{1}{n}\sum_{i=1}^n\ell(f_w(X_i),Y_i)$
- Expected *EP* error:

$$\mathrm{Err}\triangleq\mathbb{E}_{W,S}\left[R_{\mu'}(W)-R_S(W)\right]=\mathbb{E}_{W,S,S'_{X'}}\left[R_{\mu'}(W)-R_S(W)\right]$$

- *PP* error for w: $\widetilde{\mathrm{Err}}(w)\triangleq R_{\mu'}(w)-R_\mu(w)$

Assumptions of the Loss Function

Assumption 1 (Boundedness). *$\ell(\cdot,\cdot)$ is bounded in $[0,M]$.*

Assumption 2 (Subgaussianity). *$\ell(f_w(X),Y)$ is R-subgaussian under μ.*

Assumption 3 (Lipschitzness). *$\ell(f_w(X),Y)$ is β-Lipschitz continuous i.e., $|\ell(f_w(x_1),y_1)-\ell(f_w(x_2),y_2)|\leq\beta d(z_1,z_2)$ for some metric d on $\mathcal{Z}$.*

Assumption 4 (Triangle and Symmetric). *$\ell(\cdot,\cdot)$ satisfies: $\ell(y_1,y_2)=\ell(y_2,y_1)$ and $\ell(y_1,y_2)\leq\ell(y_1,y_3)+\ell(y_3,y_2)$ for any $y_1,y_2,y_3\in\mathcal{Y}$.*

Key Ingredients

Lemma 1 (*DV variational formula of KL*). $\mathrm{D_{KL}}(Q||P)=\sup_f\mathbb{E}_{\theta\sim Q}[f(\theta)]-\log\mathbb{E}_{\theta\sim P}[\exp f(\theta)]$.

Lemma 2 (*KR duality*). $\mathbb{W}(P,Q)=\sup_{f\in 1-\mathrm{Lip}(\rho)}\int_\mathcal{X}f\,dP-\int_\mathcal{X}f\,dQ$.

Lemma 3. *If $g(\theta)$ is R-subgaussian, then*

$$\left|\mathbb{E}_{\theta'\sim Q}[g(\theta')]-\mathbb{E}_{\theta\sim P}[g(\theta)]\right|\leq\sqrt{2R^2\mathrm{D_{KL}}(Q||P)}.$$

Bounding PP Error by KL Divergence

Theorem 1. *If Assumption 2 holds, then* $\left|\widetilde{\mathrm{Err}}(w)\right|\leq\sqrt{2R^2\mathrm{D_{KL}}(\mu'||\mu)}$.

Corollary 1. *Let $f_w=g\circ h$ (where $h:\mathcal{X}\to\mathcal{T}$ and $g:\mathcal{T}\to\mathcal{Y}$), then*

$$R_\mu(w)-\sqrt{2R^2\mathrm{D_{KL}}(\mu'||\mu)}\leq R_{\mu'}(w)\leq R_\mu(w)+\sqrt{2R^2\mathrm{D_{KL}}(\mu'_h||\mu_h)}.$$

Corollary 2. *Assumption 1* $\Longrightarrow\left|\widetilde{\mathrm{Err}}(w)\right|\leq\frac{M}{2}\sqrt{\mathrm{D_{KL}}(\mu||\mu')+\mathrm{D_{KL}}(\mu'||\mu)}$.

Theorem 2. *Assumption 4 + $\ell(f_{w'}(X),f_w(X))$ is R-subgaussian* $\Longrightarrow\widetilde{\mathrm{Err}}(w)\leq\sqrt{2R^2\mathrm{D_{KL}}(P_{X'}||P_X)}+\lambda^*$, *where $\lambda^*=\min_{w\in\mathcal{W}}R_{\mu'}(w)+R_\mu(w)$.*

Bounding PP Error by Wasserstein Distance

Theorem 3. *If Assumption 3 holds, then* $\left|\widetilde{\mathrm{Err}}(w)\right|\leq\beta\mathbb{W}(\mu',\mu)$.

Corollary 3. *If Assumption 1 holds and let d be the discrete metric, then*

$$\left|\widetilde{\mathrm{Err}}(w)\right|\leq M\mathrm{TV}(\mu',\mu)\leq M\sqrt{\min\left\{\frac{1}{2}\mathrm{D_{KL}}(\mu'||\mu),1-e^{-\mathrm{D_{KL}}(\mu'||\mu)}\right\}}.$$

Theorem 4. *Assumption 4 + $\ell(f_w(X),f_{w'}(X))$ is β-Lipschitz* $\Longrightarrow\widetilde{\mathrm{Err}}(w)\leq\beta\mathbb{W}(P_{X'},P_X)+\lambda^*$, *where $\lambda^*=\min_{w\in\mathcal{W}}R_{\mu'}(w)+R_\mu(w)$.*

Mutual Information (MI) Bound for EP

Theorem 5. *Assume $\ell(f_w(X'),Y')$ is R-subgaussian then*

$$|\mathrm{Err}|\leq\underbrace{\frac{1}{nm}\sum_{j=1}^m\sum_{i=1}^n\mathbb{E}_{X'_j}\sqrt{2R^2I^{X'_j}(W;Z_i)}}_{\text{Generalization error on }\mu}+\underbrace{\sqrt{2R^2\mathrm{D_{KL}}(\mu||\mu')}}_{\text{PP error (Theorem 1)}},$$

where $I^{X'_j}(\cdot,\cdot)$ is the disintegrated version of mutual information.

Corollary 4. *Let Assumption 1 hold. Then*

$$|\mathrm{Err}|\leq\frac{M}{\sqrt{2}nm}\sum_{j=1}^m\sum_{i=1}^n\mathbb{E}_{X'_j}\sqrt{\min\left\{I^{X'_j}(W;Z_i),L^{X'_j}(W;Z_i)\right\}}+\frac{M}{\sqrt{2}}\sqrt{\min\left\{\mathrm{D_{KL}}(\mu||\mu'),\mathrm{D_{KL}}(\mu'||\mu)\right\}},$$

where $L^{X'_j}(\cdot;\cdot)$ is the disintegrated version of Lautum information.

Stronger Bounds for EP

Theorem 6. *Assume ℓ is Lipschitz for both $w\in\mathcal{W}$ and $z\in\mathcal{Z}$, then*

$$|\mathrm{Err}|\leq\frac{\beta'}{nm}\sum_{j=1}^m\sum_{i=1}^n\mathbb{E}_{X'_j,Z_i}\mathbb{W}(P_{W|Z_i,X'_j},P_{W|X'_j})+\beta\mathbb{W}(\mu,\mu').$$

Further, if Assumption 1 hold. Then

$$\left|\widetilde{\mathrm{Err}}\right|\leq\frac{M}{nm}\sum_{j=1}^m\sum_{i=1}^n\mathbb{E}_{X'_j,Z_i}\left[\mathrm{TV}(P_{W|Z_i,X'_j},P_{W|X'_j})\right]+M\mathrm{TV}(\mu,\mu')$$
$$\leq\frac{1}{nm}\sum_{j=1}^m\sum_{i=1}^n\mathbb{E}_{X'_j,Z_i}\sqrt{\frac{M^2}{2}\mathrm{D_{KL}}(P_{W|Z_i,X'_j}||P_{W|X'_j})}+\sqrt{\frac{M^2}{2}\mathrm{D_{KL}}(\mu||\mu')}.$$

Applications

Gradient Penalty as an Universal Regularizer

Theorem 7. *Consider a "noisy" iterative algorithm for updating W, e.g., SGLD,*

$$|\mathrm{Err}|\leq\sqrt{\frac{R^2}{n}\sum_{t=1}^T\frac{\eta_t^2}{\sigma_t^2}\mathbb{E}_{S'_{X'},W_{t-1},S}\left[\left|\left|G_t-\mathbb{E}_{Z_{B_t}}[G_t]\right|\right|^2\right]}+\sqrt{2R^2\mathrm{D_{KL}}(\mu||\mu')}.$$

Restrict the gradient norm $\Longrightarrow$ Reduce $|\mathrm{Err}|$!

Controlling Label Information for KL Guided Marginal Alignment

- Nguyen et al. (2022): $\mathrm{D_{KL}}(P_{Y'|T'}||P_{Y|T})\leq\mathrm{D_{KL}}(P_{Y'|X'}||P_{Y|X})$ if $I(X;Y)=I(T;Y)$.
- $I(X;Y)\neq I(T;Y)$ when ℓ is cross-entropy: $I(X;Y)\geq I(T;Y)=H(Y)-H(Y|T)$.

$$\mathbb{E}_{W,Z_i}\left[\ell(f_W(T_i),Y_i)\right]=H(Y_i|T_i)+\mathbb{E}_{T_i,W}\left[\mathrm{D_{KL}}(P_{Y_i|T_i,W}||Q_{Y_i|T_i,W})\right]-I(W;Y_i|T_i).$$

Minimizing cross-entropy $\not\Rightarrow$ Minimizing $H(Y|T)$

- $I^{T'_j}(W;Y_i|T_i)\leq\mathcal{O}\left(||W-\widetilde{W}||^2\right)$: Creating $f_{\widetilde{w}}$ that does not depend on Y.
  - Train $f_{\widetilde{w}}$ by pseudo labels of f_w
  - Adding $||W-\widetilde{W}||^2$ as a regularizer in the training of W.

Experimental Results

Table 1: RotatedMNIST and Digits. Results of baselines are reported from Nguyen et al. (2022).

	RotatedMNIST (0° as source domain)						Digits			
Method	**15°**	**30°**	**45°**	**60°**	**75°**	**Ave**	**M $\to$ U**	**U $\to$ M**	**S $\to$ M**	**Ave**
ERM	97.5±0.2	84.1±0.8	53.9±0.7	34.2±0.4	22.3±0.5	58.4	73.1±4.2	54.8±6.2	65.9±1.4	64.6
DANN	97.3±0.4	90.6±1.1	68.7±4.2	30.8±0.6	19.0±0.6	61.3	90.7±0.4	91.2±0.8	71.1±0.5	84.3
MMD	97.5±0.1	95.3±0.4	73.6±2.1	44.2±1.8	32.1±2.1	68.6	91.8±0.3	94.4±0.5	82.8±0.3	89.7
CORAL	97.1±0.3	82.3±0.3	56.0±2.4	30.8±0.2	27.1±1.7	58.7	88.0±1.9	83.3±0.1	69.3±0.6	80.2
WD	96.7±0.3	93.1±1.2	64.1±3.3	41.4±7.6	27.6±2.0	64.6	88.2±0.6	60.2±1.8	68.4±2.5	72.3
KL	97.8±0.1	97.1±0.2	93.4±0.8	75.5±2.4	68.1±1.8	86.4	98.2±0.2	97.3±0.5	92.5±0.9	96.0
ERM-GP	97.5±0.1	86.2±0.5	62.0±1.9	34.8±2.1	26.1±1.2	61.2	91.3±1.6	72.7±4.2	68.4±0.2	77.5
KL-GP	98.2±0.2	96.9±0.1	95.0±0.6	**88.0±8.1**	**78.1±2.5**	**91.2**	98.8±0.1	**97.8±0.1**	**93.8±1.1**	**96.8**
KL-CL	**98.4±0.2**	**97.3±0.2**	**95.6±0.1**	83.0±8.2	73.6±4.0	89.6	**98.9±0.1**	97.7±0.1	93.0±0.3	96.5