Bridging Theory and Algorithm for Domain Adaptation

Yuchen Zhang Tianle Liu Mingsheng Long Michael I. Jordan

School of Software, Tsinghua University
National Engineering Lab for Big Data Software
University of California, Berkeley

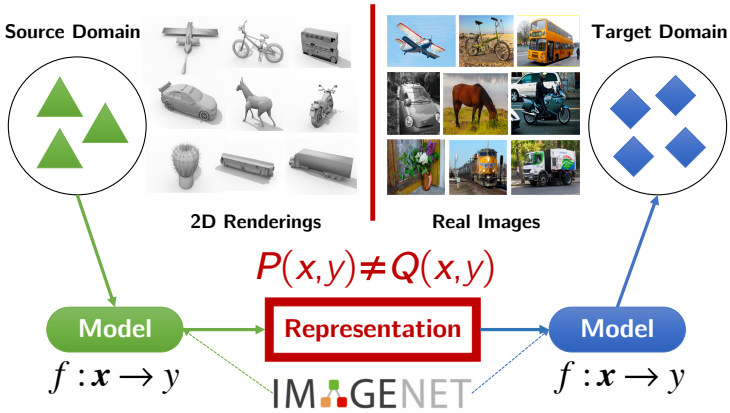
36th International Conference on Machine Learning

Outline

1. Transfer Learning
2. Previous Theory and Algorithm
3. MDD: Margin Disparity Discrepancy
   - Definition
   - Generalization Bounds
4. MDD: Theoretically Justified Algorithm
5. Experiments

Transfer Learning

- Machine learning across domains of Non-IID distributions $P \neq Q$
- How to design models that effectively bound the generalization error?



Notations and Assumptions

Notations:

- 0-1 risk: $\mathrm{err}_D(h) = \mathbb{E}_{(x,y)\sim D}\mathbb{1}[h(x) \neq y]$
- Empirical 0-1 risk: $\mathrm{err}_{\widehat{D}}(h) \triangleq \mathbb{E}_{(x,y)\sim \widehat{D}}\mathbb{1}[h(x) \neq y] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[h(x_i) \neq y_i]$
- Disparity: $\mathrm{disp}_D(h', h) \triangleq \mathbb{E}_D\mathbb{1}[h' \neq h]$,

Assumptions:

In unsupervised domain adaptation, there are two distinct domains, the source P and the target Q. The learner is trained on:

- A labeled sample $\widehat{P} = \{(x_i^s, y_i^s)\}_{i=1}^n$ drawn from source distribution P.
- An unlabeled sample $\widehat{Q} = \{x_i^t\}_{i=1}^m$ drawn from target distribution Q.

Key Problem:

How to control target domain expected risk $\mathrm{err}_Q(h)$?

Outline

1. Transfer Learning
2. Previous Theory and Algorithm
3. MDD: Margin Disparity Discrepancy
   - Definition
   - Generalization Bounds
4. MDD: Theoretically Justified Algorithm
5. Experiments

Previous Theory

In the seminal work [1], the $\mathcal{H}\Delta\mathcal{H}$-**divergence** was proposed to measure domain discrepancy and control the target risk,:

$$d_{\mathcal{H}\Delta\mathcal{H}}(P,Q) = \sup_{h,h'\in\mathcal{H}} \left|\mathrm{disp}_Q(h',h) - \mathrm{disp}_P(h',h)\right|. \tag{1}$$

[3] extended the $\mathcal{H}\Delta\mathcal{H}$-divergence to general loss functions, leading to the **discrepancy distance**:

$$\mathrm{disc}_L(P,Q) = \sup_{h,h'\in\mathcal{H}} |\mathbb{E}_Q L(h',h) - \mathbb{E}_P L(h',h)|, \tag{2}$$

where L should be a bounded function satisfying symmetry and triangle inequality. Note that many widely-used losses, e.g. **margin loss**, do not satisfy these requirements.

Previous Theory

Theorem

For every hypothesis $h \in \mathcal{H}$,

$$\mathrm{err}_Q(h) \leq \mathrm{err}_P(h) + d_{\mathcal{H}\Delta\mathcal{H}}(P, Q) + \lambda, \tag{3}$$

where $\lambda = \lambda(\mathcal{H}, P, Q)$ is the ideal combined loss:

$$\lambda = \min_{h^* \in \mathcal{H}} \{\mathrm{err}_P(h^*) + \mathrm{err}_Q(h^*)\}. \tag{4}$$

- $\mathrm{err}_P(h)$ depicts the performance of h on source domain.
- $d_{\mathcal{H}\Delta\mathcal{H}}$ bounds the performance gap caused by domain shift.
- λ quantifies the inverse of "*adaptability*" between domains.
- The order of complexity term is $O(\sqrt{d/m} + \sqrt{d/n})$, when d is the VC-dimension of $\mathcal{H}$.

Previous Algorithm

[2] sets a class of domain discriminator G to approximate function class $\mathcal{H}\Delta\mathcal{H} = \{\mathbb{1}[h' \neq h] | h, h' \in \mathcal{H}\}$ for computing $d_{\mathcal{H}\Delta\mathcal{H}}$:

$$d_{\mathcal{H}\Delta\mathcal{H}} \approx \sup_{g \in \mathcal{G}} (\mathbb{E}_Q \mathbb{1}[g(x) = 0] + \mathbb{E}_P \mathbb{1}[g(x) = 1])$$

[4] assumes that h and h' should agree on source domain. Then they use L1-loss of two classifiers' probabilistic outputs on target domain to approximate $d_{\mathcal{H}\Delta\mathcal{H}}$:

$$d_{\mathcal{H}\Delta\mathcal{H}} \approx \sup_{f, f'} \mathbb{E}_Q |f(x) - f'(x)|$$

There are two crucial directions for improvement:

- Generalization bound for classification with scoring functions and margin loss has not been formally studied in the DA setting.
- Computing the supremum requires an ergodicity over $\mathcal{H}\Delta\mathcal{H}$ increases the difficulty of optimization.

Outline

1. Transfer Learning
2. Previous Theory and Algorithm
3. MDD: Margin Disparity Discrepancy
   - Definition
   - Generalization Bounds
4. MDD: Theoretically Justified Algorithm
5. Experiments

DD: Hypothesis-induced Discrepancy

Definition (Disparity Discrepancy)

Given a hypothesis space $\mathcal{H}$ and a *specific classifier* $h \in \mathcal{H}$, the Disparity Discrepancy (DD) induced by $h' \in \mathcal{H}$ is defined by

$$d_{h,\mathcal{H}}(P, Q) = \sup_{h' \in \mathcal{H}} \left| \mathbb{E}_Q \mathbb{1}[h' \neq h] - \mathbb{E}_P \mathbb{1}[h' \neq h] \right|. \tag{5}$$

The supremum in the disparity discrepancy is taken **only over the hypothesis space** $\mathcal{H}$ and thus can be optimized more easily.

Theorem

For every hypothesis $h \in \mathcal{H}$,

$$\mathrm{err}_Q(h) \leq \mathrm{err}_P(h) + d_{h,\mathcal{H}}(P, Q) + \lambda, \tag{6}$$

where $\lambda = \lambda(\mathcal{H}, P, Q)$ *is the ideal combined loss.*

MDD: Towards an Informative Margin Theory

Notations for Multi-class Classification

- Scoring Function:

$$f \in \mathcal{F} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

- Labeling Function induced by f:

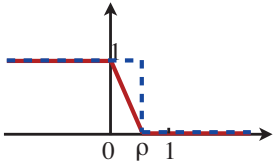
$$h_f : x \mapsto \arg\max_{y \in \mathcal{Y}} f(x, y). \tag{7}$$

- Margin of a Scoring Function:

$$\rho_f(x, y) = \frac{1}{2}(f(x, y) - \max_{y' \neq y} f(x, y'))$$

- Margin Loss:

$$\Phi_\rho(x) = \begin{cases} 0 & \rho \leqslant x \\ 1 - x/\rho & 0 \leqslant x \leqslant \rho \\ 1 & x \leqslant 0 \end{cases}$$



MDD: Margin Disparity Discrepancy

- Margin error: $\mathrm{err}_D^{(\rho)}(f) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim D}\left[\Phi_\rho \circ \rho_f(x, y)\right]$
- Margin disparity: $\mathrm{disp}_D^{(\rho)}(f', f) \triangleq \mathbb{E}_{z\sim D_x}[\Phi_\rho \circ \rho_{f'}(x, h_f(x))]$

Definition (**Margin Disparity Discrepancy**)

With the definition of margin disparity, we define Margin Disparity Discrepancy (MDD) and its empirical version by

$$
\begin{aligned}
d_{f,\mathcal{F}}^{(\rho)}(P, Q) &\triangleq \sup_{f'\in\mathcal{F}} \left(\mathrm{disp}_Q^{(\rho)}(f', f) - \mathrm{disp}_P^{(\rho)}(f', f)\right), \\
d_{f,\mathcal{F}}^{(\rho)}(\widehat{P}, \widehat{Q}) &\triangleq \sup_{f'\in\mathcal{F}} \left(\mathrm{disp}_{\widehat{Q}}^{(\rho)}(f', f) - \mathrm{disp}_{\widehat{P}}^{(\rho)}(f', f)\right).
\end{aligned}
\tag{8}
$$

The margin disparity discrepancy is well-defined since $d_{f,\mathcal{F}}^{(\rho)}(P, P) = 0$ and it satisfies the nonnegativity and subadditivity.

MDD: Bounding the Target Expected Error

Theorem

Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be a hypothesis set with $\mathcal{Y} = \{1, \cdots, k\}$ and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be the corresponding $\mathcal{Y}$-valued classifier class. For every scoring function $f \in \mathcal{F}$,

$$\mathrm{err}_Q(h_f) \leq \mathrm{err}_P^{(\rho)}(f) + d_{f,\mathcal{F}}^{(\rho)}(P, Q) + \lambda, \tag{9}$$

where $\lambda = \lambda(\rho, \mathcal{F}, P, Q)$ is the ideal combined margin loss:

$$\lambda = \min_{f^* \in \mathcal{H}} \{\mathrm{err}_P^{(\rho)}(f^*) + \mathrm{err}_Q^{(\rho)}(f^*)\}. \tag{10}$$

- This upper bound has a similar form with previous bound.
  - $\mathrm{err}_P^{(\rho)}(f)$ depicts the performance of f on source domain.
  - MDD bounds the performance gap caused by domain shift.
  - λ quantifies the inverse of "*adaptability*".
- **A new perspective for analyzing DA with respect to margin loss.**

MDD: Notations for Generalization Bounds

For deriving generalization bounds for MDD, we first introduce two function class:

Definition

Given a class of scoring functions $\mathcal{F}$, $\Pi_1(\mathcal{F})$ is defined as

$$\Pi_1\mathcal{F} = \{x \mapsto f(x, y) \big| y \in \mathcal{Y}, f \in \mathcal{F}\}, \tag{11}$$

We introduce a new function class $\Pi_{\mathcal{H}}\mathcal{F}$ that serves as **a "scoring" version** of the symmetric difference hypothesis space $\mathcal{H}\Delta\mathcal{H}$:

Definition

Given a class of scoring functions $\mathcal{F}$ and a class of the induced classifiers $\mathcal{H}$, we define $\Pi_{\mathcal{H}}\mathcal{F}$ as

$$\Pi_{\mathcal{H}}\mathcal{F} \triangleq \{x \mapsto f(x, h(x)) | h \in \mathcal{H}, f \in \mathcal{F}\}. \tag{12}$$

MDD: Notations for Generalization Bounds

Definition (Rademacher complexity)

Then, the empirical Rademacher complexity of $\mathcal{F}$ with respect to the sample $\widehat{D}$ is defined as

$$\widehat{\mathfrak{R}}_{\widehat{D}}(\mathcal{F}) = \mathbb{E}_{\sigma} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i). \tag{13}$$

where σ_i's are independent uniform random variables taking values in $\{-1, +1\}$. The Rademacher complexity is

$$\mathfrak{R}_{n,D}(\mathcal{F}) = \mathbb{E}_{\widehat{D} \sim D^n} \widehat{\mathfrak{R}}_{\widehat{D}}(\mathcal{F}). \tag{14}$$

Definition (Covering Number)

(Informal) A *covering number* $\mathcal{N}_2(\tau, \mathcal{G})$ is the minimal number of $\mathcal{L}_2$ balls of radius $\tau > 0$ needed to cover a class $\mathcal{G}$ of bounded functions $g : \mathcal{X} \to \mathbb{R}$

MDD: Rademacher Generalization Bounds

With the Rademacher complexity, we proceed to show that MDD can be well estimated through finite samples.

Lemma

For any $\delta > 0$, with probability $1 - 2\delta$, the following holds **simultaneously** *for any scoring function $f \in \mathcal{F}$,*

$$\begin{aligned} &|d_{f,\mathcal{F}}^{(\rho)}(\widehat{P}, \widehat{Q}) - d_{f,\mathcal{F}}^{(\rho)}(P, Q)| \\ \leq &\frac{2k}{\rho}\mathfrak{R}_{n,P}(\Pi_{\mathcal{H}}\mathcal{F}) + \frac{2k}{\rho}\mathfrak{R}_{m,Q}(\Pi_{\mathcal{H}}\mathcal{F}) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}. \end{aligned} \tag{15}$$

This lemma justifies that the expected MDD with respect to f can be uniformly approximated by the empirical one computed on samples.

MDD: Margin Theory for Domain Adaptation

Combining previous theorems, we obtain a Rademacher complexity based generalization bound of the expected target error.

Theorem (**Generalization Bound**)

For any $\delta > 0$, with probability $1 - 3\delta$, we have the following uniform generalization bound for all scoring functions $f \in \mathcal{F}$,

$$\begin{aligned}\operatorname{err}_Q(h_f) \leq & \operatorname{err}_{\widehat{P}}^{(\rho)}(f) + d_{f,\mathcal{F}}^{(\rho)}(\widehat{P}, \widehat{Q}) + \lambda \\ & + \frac{2k^2}{\rho}\mathfrak{R}_{n,P}(\Pi_1 \mathcal{F}) + \frac{2k}{\rho}\mathfrak{R}_{n,P}(\Pi_{\mathcal{H}} \mathcal{F}) + 2\sqrt{\frac{\log \frac{2}{\delta}}{2n}} \\ & + \frac{2k}{\rho}\mathfrak{R}_{m,Q}(\Pi_{\mathcal{H}} \mathcal{F}) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}},\end{aligned} \tag{16}$$

where $\lambda = \lambda(\rho, \mathcal{F}, P, Q)$ is the ideal combined margin loss.

MDD: Rademacher Bound of Linear Classifier

We need to check the variation of $\mathfrak{R}_{n,D}(\Pi_{\mathcal{H}}\mathcal{F})$ with the growth of n. First, we include a simple example of binary linear classifiers.

Theorem

Let $S \subseteq \mathcal{X} = \{\mathbf{x} \in \mathbb{R}^s | \|\mathbf{x}\|_2 \leq r\}$ be a sample of size m and suppose

$$\mathcal{F} = \left\{ f : \mathcal{X} \times \{\pm 1\} \to \mathbb{R} \;\middle|\; f(\mathbf{x}, y) = \operatorname{sgn}(y)\ \mathbf{w} \cdot \mathbf{x},\ \|\mathbf{w}\|_2 \leq \Lambda \right\},$$
$$\mathcal{H} = \left\{ h \mid h(\mathbf{x}) = \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x}),\ \|\mathbf{w}\|_2 \leq \Lambda \right\}.$$

Then the empirical Rademacher complexity of $\Pi_{\mathcal{H}}\mathcal{F}$ can be bounded as follows:

$$\widehat{\mathfrak{R}}_S(\Pi_{\mathcal{H}}\mathcal{F}) \leq 2\Lambda r \sqrt{\frac{d \log \frac{em}{d}}{m}},$$

where d is the VC-dimension of $\mathcal{H}$.

MDD: Generalization Bound with Covering Numbers

For more general settings, we derive bound based on covering number:

Theorem

Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be a hypothesis set with $\mathcal{Y} = \{1, \cdots, k\}$ and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be the corresponding $\mathcal{Y}$-valued classifier class. Suppose $\Pi_1\mathcal{F}$ is bounded in $\mathcal{L}_2$ by L. Fix $\rho > 0$. For all $\delta > 0$, with probability $1 - 3\delta$ the following inequality holds for all hypothesis $f \in \mathcal{F}$:

$$\begin{aligned}\mathrm{err}_Q(h_f) \leq \mathrm{err}^{(\rho)}_{\widehat{P}}(f) + d^{(\rho)}_{f,\mathcal{F}}(\widehat{P}, \widehat{Q}) + \lambda + 2\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \\ + \sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \frac{16k^2\sqrt{k}}{\rho}\inf_{\epsilon\geq 0}\Big\{\epsilon + 3\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\big) \\ \big(\int_{\epsilon}^{L}\sqrt{\log \mathcal{N}_2(\tau, \Pi_1\mathcal{F})}d\tau + L\int_{\epsilon/L}^{1}\sqrt{\log \mathcal{N}_2(\tau, \Pi_1\mathcal{H})}\mathrm{d}\tau\big)\Big\}.\end{aligned} \tag{17}$$

Outline

1. Transfer Learning
2. Previous Theory and Algorithm
3. MDD: Margin Disparity Discrepancy
   - Definition
   - Generalization Bounds
4. MDD: Theoretically Justified Algorithm
5. Experiments

MDD: Theoretically Justified Algorithm

- MDD is defined as the supremum over hypothesis space $\mathcal{F}$.
- Minimizing MDD is a **minimax game**.
- Because the max-player is still too strong, we introduce a feature extractor ψ to make the min-player stronger.
- The overall optimization problem can be written as

$$\begin{aligned} \min_{f,\psi} \ & \mathrm{err}^{(\rho)}_{\psi(\widehat{P})}(f) + (\mathrm{disp}^{(\rho)}_{\psi(\widehat{Q})}(f^*, f) - \mathrm{disp}^{(\rho)}_{\psi(\widehat{P})}(f^*, f)), \\ & f^* = \max_{f'} (\mathrm{disp}^{(\rho)}_{\psi(\widehat{Q})}(f', f) - \mathrm{disp}^{(\rho)}_{\psi(\widehat{P})}(f', f)). \end{aligned} \tag{18}$$

- To enable representation-based domain adaptation, we need to learn new representation ψ such that MDD is minimized.

MDD: Theoretically Justified Algorithm

- We design an adversarial learning algorithm to solve this problem.
- We introduce an **auxiliary classifier** f' sharing the same hypothesis space with f.

$$\min_{f,\psi} \max_{f'} \ \mathrm{err}^{(\rho)}_{\psi(\widehat{P})}(f) + (\mathrm{disp}^{(\rho)}_{\psi(\widehat{Q})}(f', f) - \mathrm{disp}^{(\rho)}_{\psi(\widehat{P})}(f', f)), \tag{19}$$

- Multiclass margin loss causes the problem of gradient vanishing.
- Denote by σ the softmax function, $\sigma_j(\mathbf{z}) = \frac{e^{z_j}}{\sum_{i=1}^{k} e^{z_i}}$, for $j = 1, \ldots, k$.
- We choose **combined cross-entropy** loss to approximate MDD:

$$\begin{aligned} \mathcal{E}(\widehat{P}) &= -\mathbb{E}_{(x^s,y^s)\sim\widehat{P}} \log[\sigma_{y^s}(f(\psi(x^s)))], \\ \mathcal{D}(\widehat{P}, \widehat{Q}) &= \mathbb{E}_{x^t\sim\widehat{Q}} \log[1 - \sigma_{h_f(\psi(x^t))}(f'(\psi(x^t)))] \\ &\quad + \mathbb{E}_{x^s\sim\widehat{P}} \log[\sigma_{h_f(\psi(x^s))}(f'(\psi(x^s)))]. \end{aligned} \tag{20}$$

MDD: Theoretically Justified Algorithm

We combine the two terms in $\mathcal{D}(\widehat{P}, \widehat{Q})$ with a coefficient γ.

$$\begin{aligned}\mathcal{E}(\widehat{P}) &= -\mathbb{E}_{(x^s,y^s)\sim\widehat{P}} \log[\sigma_{y^s}(f(\psi(x^s)))], \\ \mathcal{D}_\gamma(\widehat{P}, \widehat{Q}) &= \mathbb{E}_{x^t\sim\widehat{Q}} \log[1 - \sigma_{h_f(\psi(x^t))}(f'(\psi(x^t)))] \\ &\quad + \gamma\mathbb{E}_{x^s\sim\widehat{P}} \log[\sigma_{h_f(\psi(x^s))}(f'(\psi(x^s)))].\end{aligned} \tag{21}$$

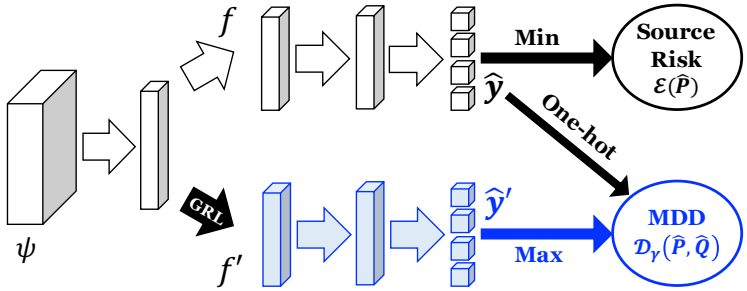
γ is related to the margin of f' when the algorithm reaches equilibrium.

Theorem

(Informal) Assuming that there is no restriction on the choice of f' and $\gamma > 1$, the global minimum of $\mathcal{D}_\gamma(P, Q)$ is $P = Q$. The value of $\sigma_{h_f}(f'(\cdot))$ at equilibrium is $\gamma/(1+\gamma)$ and the corresponding margin of f' is $\rho = \log\gamma$.

We refer to $\gamma = \exp\rho$ as **the margin factor**.

MDD: Theoretically Justified Algorithm



The practical optimization problem in the adversarial learning is stated as

$$\begin{aligned} &\min_{f,\psi} \ \mathcal{E}(\widehat{P}) + \eta \mathcal{D}_\gamma(\widehat{P}, \widehat{Q}), \\ &\max_{f'} \ \mathcal{D}_\gamma(\widehat{P}, \widehat{Q}), \end{aligned} \tag{22}$$

Outline

1. Transfer Learning
2. Previous Theory and Algorithm
3. MDD: Margin Disparity Discrepancy
   - Definition
   - Generalization Bounds
4. MDD: Theoretically Justified Algorithm
5. Experiments

Results

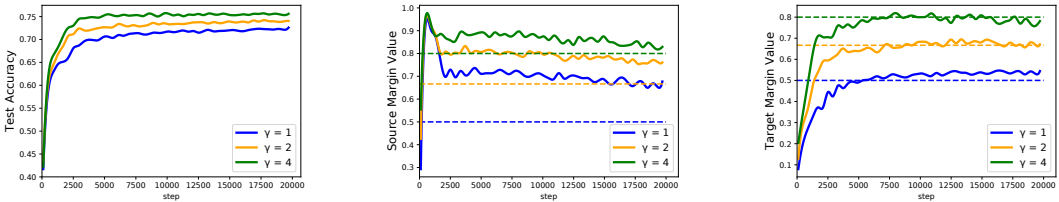
Table: Accuracy (%) on *Office-31* for unsupervised domain adaptation

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
ResNet-50	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
JAN	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	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	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	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
MDD (Proposed)	**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**

Table: Accuracy (%) on *Office-Home* for unsupervised domain adaptation

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	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
JAN	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	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
MDD (Proposed)	**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**

Analysis



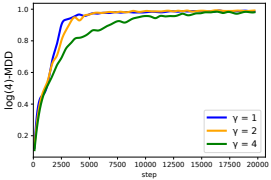
(a) Test Accuracy (b) Equilibrium on Source (c) Equilibrium on Target

Figure: Test accuracy and empirical values of $\sigma_{h_f} \circ f'$ on D $\rightarrow$ A, where dashed lines indicate $\frac{\gamma}{\gamma+1}$.

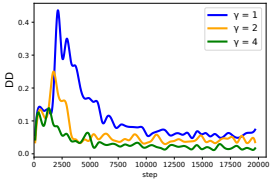
Margin γ	1	2	3	**4**	5	6
A $\rightarrow$ W	92.5	93.7	94.0	**94.5**	93.8	93.5
D $\rightarrow$ A	72.4	73.0	73.7	**74.6**	74.3	74.2
Avg on Office-31	87.6	88.1	88.5	**88.9**	88.7	88.6

Table: Accuracy (%) on Office-31 by different margins.

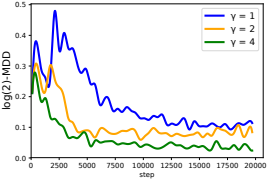
Analysis



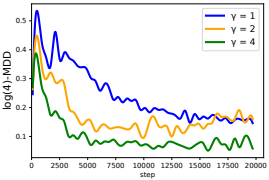
(a) MDD w/o Adv



(b) DD



(c) log 2-MDD



(d) log 4-MDD

Figure: Empirical values of the MDD computed by auxiliary classifier f'.

Summary

- We extend previous theories to multiclass classification in domain adaptation, where classifiers based on the scoring functions and margin loss are standard choices in algorithm design.
- We introduce Margin Disparity Discrepancy, a novel measurement with rigorous generalization bounds, tailored to the distribution comparison with the asymmetric margin loss, and to the minimax optimization for easier training.

Thanks!

Poster: tonight at Pacific Ballroom # 184.

Reference


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


Y. Ganin and V. Lempitsky.
Unsupervised domain adaptation by backpropagation.
In *International Conference on Machine Learning (ICML)*, 2015.


Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh.
Domain adaptation: Learning bounds and algorithms.
In *Conference on Learning Theory (COLT)*, 2009.


Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada.
Maximum classifier discrepancy for unsupervised domain adaptation.
In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3723–3732, 2018.