

Information-theoretic Analysis For Machine Learning and Transfer Learning

Presenter: Xuetong Wu

Principle Supervisor: Jingge Zhu

Co-Supervisor: Jonathan H. Manton, Uwe Aickelin

Chair: Margreta Kuijper

Department of EEE
University of Melbourne

Ph.D. Completion Seminar

March 30, 2025



Contents

Generalization in Machine Learning

Information-Theoretic Bounds for Generalization

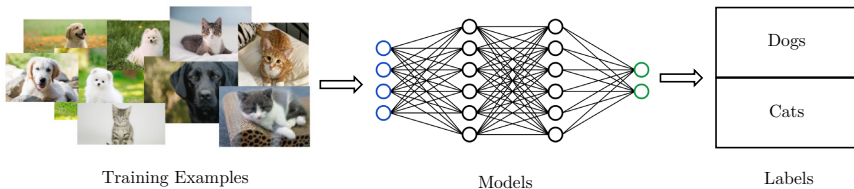
Fast Rate Generalization Error

Information-theoretic Bounds for Transfer Learning

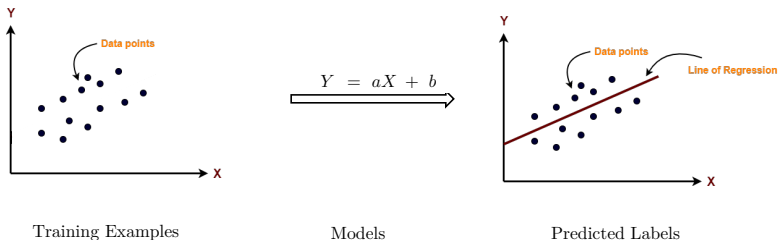
Transfer Learning under Causal Framework

Machine Learning Problems

Classification:



Regression:



Machine Learning

- Machine learning is an application of artificial intelligence (AI) that provides systems with the ability to automatically learn and improve from experience without being explicitly programmed.
- How could we guarantee that the algorithm can yield promising performance?
- The statistical learning theory is not only a tool for theoretical analysis but also a guideline for inspiring practical algorithms.

Statistical Machine Learning Theory

- **Unknown** probability distribution μ over some data space $\mathcal{Z}$
- **Known** training set S consisting of finite n independent samples $z_i \in \mathcal{Z}$ drawn from μ , $i = 1, 2, \dots, n$
- Hypothesis Class $\mathcal{W}$: set of predictors (e.g. classifiers)
- Learning Algorithm $\mathcal{A} : \mathcal{Z}^n \rightarrow \mathcal{W}$
- Loss function $\ell : \mathcal{W} \times \mathcal{Z} \rightarrow \mathbb{R}$

What to achieve?

What to achieve from the sample?

We will use training samples to:

- Learn a predictor
- Certify the predictor's performance

Learning a predictor (hypothesis):

- by an algorithm driven by some learning principle
- depends on hypothesis space

Certifying the performance:

- what happens beyond the training set
- testing on previously unseen data
- generalization ability

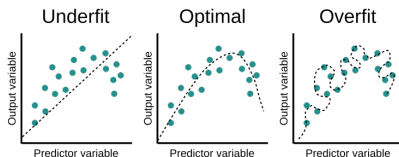


Figure: Overfitting and Underfitting¹

¹Source: www.educative.io/edpresso

Loss Function

- Empirical risk (training):

$$\hat{L}(w) = \frac{1}{n} \sum_{i=1}^n \ell(w, Z_i)$$

- Expected risk (testing):

$$L_{\mu}(w) = \mathbb{E}_{Z \sim \mu} [\ell(w, Z)]$$

Example

- $\ell(W, Z) = \mathbf{1}[W(X) \neq Y]$: 0 - 1 loss (classification)
- $\ell(W, Z) = (1 - YW(X))_+$: hinge loss (classification)
- $\ell(W, Z) = (Y - W(X))^2$: square loss (regression)
- $\ell(W, Z) = -\log(W(X))$: log-loss (density estimation)

ERM Strategy

Because μ is unknown, it is very natural to minimise the **empirical risk** (ERM) instead.

$$w_{\text{ERM}} = \min_{w \in \mathcal{W}} \hat{L}(w) = \frac{1}{n} \sum_{i=1}^n \ell(w, Z_i)$$

Widely used in machine learning:

- Neural networks
- Linear/Logistic Regression
- Support Vector Machine
- Maximum Likelihood
- etc...

If a hypothesis W does well on the training set S with size n , will it still do well on the testing set?

Problem: Generalization Error

- **Generalization error:** $\text{gen}(w, S) = L_\mu(w) - \hat{L}(w)$
- **Excess risk:** $L_\mu(w) - L_\mu(w^*)$ for optimal one minimises $L_\mu(w)$
- Consistency: for w_{ERM} , a small generalization error implies a small excess risk with abundant training data!
- Bounding for generalization error:
 - with probability $1 - \delta$, $\text{gen}(w, S) = L_\mu(w) - \hat{L}(w) \leq \epsilon(\delta, n)$ for all $w \in \mathcal{W}$;
 - or, $\mathbb{E}[\text{gen}(W, S)] \leq \epsilon(n)$

Notions for Generalization

■ Hypothesis Space dependent

- VC dimension [Vanpik, 1971, 1999]: $\text{gen}(w, S) \leq \sqrt{\frac{\text{sVC}(\mathcal{W})(\log \frac{2n}{\text{VC}(\mathcal{W})} + 1) + 8 \log \frac{4}{\delta}}{n}}$
- Rademacher Complexity [Bartlett et al., 2002]: $\text{gen}(w, S) \leq 2\mathcal{R}_n(\mathcal{W}) + \sqrt{\frac{\log(2/\delta)}{2n}}$

■ Algorithm dependent

- PAC Learning [McAllester, 1999]: $\text{gen}(w, S) \leq \sqrt{\frac{\log \frac{1}{\pi(w)} + \ln \frac{1}{\delta}}{2n}}$
- Uniform Stability [Bousquet, 2002]: $\text{gen}(w, S) \leq \beta(n) + (n\beta(n) + M)\sqrt{\frac{2 \log(2/\delta)}{n}}$
- ϵ -Differential privacy [Dwork, 2015]: $\text{gen}(w, S) \leq \epsilon + \sqrt{\frac{1}{n}}$ w.p. $1 - 3e^{-n\epsilon^2}$
- Algorithmic Robustness [Xu, 2012]: $\text{gen}(w, S) \leq \epsilon(\mathcal{S}) + M\sqrt{\frac{2N \ln 2 + 2 \ln(1/\delta)}{n}}$

- In general, these various methods for proving generalization are incompatible and incomparable. This raises the question of whether it is possible to provide a unifying framework or language to study generalization.
- Intuitively if one takes specific algorithms, data distribution and hypothesis space into account, the algorithm might generalize better than uniform convergence would suggest.
- **Information-theoretic bounds for Generalization!**

Contents

Generalization in Machine Learning

Information-Theoretic Bounds for Generalization

Fast Rate Generalization Error

Information-theoretic Bounds for Transfer Learning

Transfer Learning under Causal Framework

Definitions

- Shannon entropy:

$$H(X) = - \sum_{i=1}^{|\mathcal{X}|} p_i \log p_i$$

- Differential entropy:

$$h(X) = - \int_{\mathcal{X}} p(x) \log p(x) dx$$

- KL divergence:

$$D(P\|Q) = \mathbb{E}_{X \sim P} \left[\log \frac{P(X)}{Q(X)} \right]$$

- Mutual information:

$$I(X; Y) = D(P_{XY} \| P_X P_Y)$$

- Conditional MI:

$$I(X; Y|Z) = \mathbb{E}_Z \left[D(P_{(X,Y)|Z} \| P_{X|Z} \otimes P_{Y|Z}) \right]$$

Connection between mutual information and generalization

Now we are ready to give an upper bound on **expected** generalization error,

Theorem (Xu & Raginsky, 2017)

Suppose $\ell(w, Z)$ is σ -subgaussian under μ for all $w \in \mathcal{W}$, then

$$|\mathbb{E}_{WS} [\text{gen}(W, S)]| \leq \sqrt{\frac{2\sigma^2}{n}} I(W; S)$$

Single Draw Bound

Theorem (Bassily et al., 2018)

Let $S \sim \mu^n$ and let $\mathcal{A}$ be an algorithm that $W = \mathcal{A}(S)$. Then, for every $\epsilon > 0$,

$$\mathbb{P}_{WS}[|\text{gen}(W)| > \epsilon] < \frac{I(W; S) + 1}{2n\epsilon^2 - 1}$$

where the probability is taken over the randomness in the sample S and the randomness of $\mathcal{A}$.

What do the bounds and the mutual information tell us?

Remark

- Establish data and algorithm-dependent bounds.
- Interpretations on the structure between the output hypothesis and input data.
- $I(W; S)$ is closely related to hypothesis class/distribution, data distribution, algorithm $\mathcal{A}$ (which can be viewed as $P_{W|S}$).
- We hope $I(W; S)$ to be small: an algorithm generalizes better if the output W is "less dependent" of the dataset S .
- A unification for previous notions!

Unification

- Hypothesis Space [Xu & Raginsky, 2017]
- Algorithmic Stability [Raginsky et al., 2016]
- Differential Privacy [Steinke & Zakynthinou, 2020]
- PAC-Bayes [Grünwald et al., 2021]
- They will imply:

$$I(W; S) \leq \varepsilon n$$

for different ε under different assumptions. A small $\varepsilon \rightarrow$ generalization.

Limitations of $I(W; S)$

- $I(W; S)$ can be infinite if W is a continuous random variable and $\mathcal{A}$ is a deterministic algorithm:

$$I(W; S) = h(W) - h(W|S) \rightarrow +\infty$$

- Hard to estimate $I(W; S)$ in practice.
- Mutual information is an expectation and is thus not very sensitive to low-probability failures $\rightarrow$ the error tolerance must grow polynomially in $1/\delta$, whereas polylogarithmic growth would be desirable.
- Suboptimality:

$$|\mathbb{E}_{WS} [\text{gen}(W, S)]| \leq \sqrt{\frac{2\sigma^2}{n}} I(W; S)$$

Can we improve from $\sqrt{\frac{MI}{n}}$ to $\frac{MI}{n}$?

Contents

Generalization in Machine Learning

Information-Theoretic Bounds for Generalization

Fast Rate Generalization Error

Information-theoretic Bounds for Transfer Learning

Transfer Learning under Causal Framework

Improved MI Bounds

- Mutual Information Bound (Xu & Raginsky, 2017):

$$\mathbb{E}_{WS} [\text{gen}(W, S)] \leq \sqrt{\frac{2\sigma^2}{n} I(W; S)}$$

- Under the same condition, sample-wise mutual information bound (Bu et al., 2019):

$$\mathbb{E}_{WS} [\text{gen}(W, S)] \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2\sigma^2 I(W; Z_i)}$$

- Tightened version and tackled the infinity issue, but the suboptimality issue still exists!

Example: Gaussian Mean Estimation

- Data: $Z_i \sim \mathcal{N}(\mu, \sigma_N^2)$
- Loss: $\ell(w, z_i) = (w - z_i)^2$
- $W_{\text{ERM}} = \frac{1}{n} \sum_{i=1}^n Z_i \sim \mathcal{N}(\mu, \frac{\sigma_N^2}{n})$, $w^* = \mu$
- Bu's bound:

$$\mathbb{E}_{WS} [\text{gen}(W, S)] \leq o\left(\frac{1}{\sqrt{n}}\right).$$

- True Generalization Error:

$$\mathbb{E}_{WS} [\text{gen}(W_{\text{ERM}}, S)] = \frac{2\sigma_N^2}{n} \sim o\left(\frac{1}{n}\right)$$

Main Results: Change on the Assumption

Theorem (Bu et al. 2019)

Suppose that $\ell(W, Z)$ is σ_ℓ -sub-Gaussian under the distribution $P_W \otimes \mu$, then:

$$\mathbb{E}_{WS} [\text{gen}(W, S)] \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2\sigma_\ell^2 I(W; Z_i)}$$

Theorem (Wu et al, 2022)

Suppose that $r(W, Z)$ is σ_r -sub-Gaussian under the distribution $P_W \otimes \mu$, then:

$$\mathbb{E}_{WS} [\text{gen}(W, S)] \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2\sigma_r^2 I(W; Z_i)}$$

where: $r(w, z) = \ell(w, z) - \ell(w^*, z)$.

Example: Continued

- Bu's bound:

$$\mathbb{E}_{WS} [\text{gen}(W, S)] \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2 \frac{(n+1)^2 \sigma_N^4}{n^2} \log \frac{n}{n-1}} \sim O\left(\frac{1}{\sqrt{n}}\right).$$

- New bound:

$$\mathbb{E}_{WS} [\text{gen}(W, S)] \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2 \frac{2\sigma_N^4}{n} \log \frac{n}{n-1}} \sim O\left(\frac{1}{n}\right)$$

- True Generalization Error:

$$\mathbb{E}_{WS} [\text{gen}(W_{\text{ERM}}, S)] = \frac{2\sigma_N^2}{n} \sim O\left(\frac{1}{n}\right)$$

$$\mathbb{E}_{WS} [\text{gen}(W, S)] \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2\sigma_r^2 I(W; Z_i)}$$

Intuitively, to achieve the fast rate:

- The output hypothesis of the learning algorithm must be “good” enough compared to the optimal hypothesis w^* ;
- We encode the notion of goodness in terms of the CGF by controlling $r(w, z)$.

Theoretically, we will need to choose a good function for variational representations:

$$I(W; Z_i) = \sup_{f \in \mathcal{F}} \left(\mathbb{E}_{WZ_i} [f(W, Z_i)] - \log \mathbb{E}_{P_W \otimes Z_i} \left[e^{f(W, Z_i)} \right] \right)$$

$f(w, z_i) = -r(w, z_i)$ will lead to a tight characterization!

Expected (η, c) -Central Condition

Definition (Expected (η, c) -Central Condition)

Let $\eta > 0$ and $0 < c \leq 1$ be two constants. We say that $(\mu, \ell, \mathcal{W}, \mathcal{A})$ satisfies the expected (η, c) -central condition if the following inequality holds for the optimal hypothesis w^* :

$$\log \mathbb{E}_{P_W \otimes \mu} \left[e^{-\eta r(W, Z)} \right] \leq c\eta \mathbb{E}_{P_W \otimes \mu} [-r(W, Z)].$$

Closely related conditions:

- Bernstein's Condition [Van Erven et al., 2015]: “easiness” of the learning problem with second moment.
- Central condition and witness condition [Grünwald et al., 2020]: “easiness” of the learning problem with MGF and observability.
- Both imply the (η, c) -central condition!

(η, c) -Central Condition Theorem

Theorem (Fast Rate with (η, c) -central condition)

Assume the learning tuple $(\mu, \ell, \mathcal{W}, \mathcal{A})$ satisfies the (η, c) -central condition for some constants $\eta > 0$ and $0 < c \leq 1$. Then, for all $\eta' \in (0, \eta]$, it holds that for general ERM algorithm that,

$$\mathbb{E}_{WS}[\text{gen}(W, S)] \leq \frac{1}{c\eta'n} \sum_{i=1}^n I(W; Z_i).$$

achieving $O(\frac{1}{n})$ for $I(W; Z_i) \sim O(\frac{1}{n})$.

Extension²

- Regularized ERM
- (v, c) -central condition
- Intermediate Rate
- Other analytical examples

²Wu, X., Manton, J. H., Aickelin, U., & Zhu, J. (2022, November). Fast rate generalization error bounds: Variations on a theme. In 2022 IEEE Information Theory Workshop (ITW) (pp. 43-48). IEEE.

Contents

Generalization in Machine Learning

Information-Theoretic Bounds for Generalization

Fast Rate Generalization Error

Information-theoretic Bounds for Transfer Learning

Transfer Learning under Causal Framework

What is Transfer Learning?

How a learned skill generalises to a different task?



Figure: Transfer learning of digit recognition

Problem Formulation

- Training Data: $\underbrace{Z_1, \dots, Z_{\beta n}}_{S', \text{each target} \sim \mu'}, \underbrace{Z_{\beta n+1}, \dots, Z_n}_{S, \text{each source} \sim \mu}$ for some $\beta \in [0, 1]$;
- Learning algorithm $\mathcal{A}$: $P_{W|SS'}$, a (stochastic) mapping from data to hypothesis;
- Performance measure on target: $L(w) := \mathbb{E}_{Z \sim \mu'}[\ell(w, Z)]$;
- Excess risk on target:

$$R(w) = L(w) - L(w^*)$$

w^* : the optimal hypothesis that minimises $L(w)$.

Risks Definition

- Empirical Loss [Ben-David et al., 2010]:

$$\hat{L}_\alpha(w) = \frac{\alpha}{\beta n} \sum_{i=1}^{\beta n} \ell(w, z_i) + \frac{1 - \alpha}{(1 - \beta)n} \sum_{i=\beta n+1}^n \ell(w, z_i)$$

- We want to upper bound:

- Excess risk: $R(w_{\text{ERM}}) = L(w_{\text{ERM}}) - L(w^*)$;
- Generalization error: $\text{gen}(w) = L(w) - \hat{L}_\alpha(w)$.

Main Theorem (Informal)

For $\alpha = \beta = 0$ (source only):

- “Slow rate” bound for any W :

$$\mathbb{E}[\text{gen}(W)] \leq \frac{C}{n} \sum_{i=1}^n \sqrt{I(W; Z_i) + D(\mu || \mu')}$$

- “Fast rate” bound for W_{ERM} :

$$\mathbb{E}[\text{gen}(W_{\text{ERM}})] \leq \frac{C'}{n} \sum_{i=1}^n I(W_{\text{ERM}}; Z_i) + D(\mu || \mu')$$

Remark

- Mutual information $I(W; Z_i)$:
 - $\mathcal{W}$ -dependent: $I(W; Z_i) = H(W) - H(W|Z_i)$;
 - Algorithm dependent: as W is induced by the algorithm;
 - Data dependent: $I(W; Z_i) = H(Z_i) - H(Z_i|W)$;
 - Algorithmic Robustness: $I(W; Z_i)$;
- An extra KL-divergence:
 - characterizing the difference between source and target domains;
 - if the divergence is smaller, the generalization error is also smaller, and vice versa;
- A good transfer learning algorithm should have small $I(W; Z_i)$;
- We also hope that the source domain is more similar to the target domain;
- **Best fast rate result known to date.**

Extensions³

- Excess Risk Bound
- Applications on SGD
- Applications on Gibbs algorithm
- ϕ -divergence bound
- Wasserstein distance bound

³Wu, X., Manton, J.H., Aickelin, U. and Zhu, J., 2022. An information-theoretic analysis for transfer learning: Error bounds and applications. arXiv preprint arXiv:2207.05377.

Gaussian Transfer (Source Only)

Data Distribution: $\mu = \mathcal{N}(m, \sigma^2)$, $\mu' = \mathcal{N}(m', \sigma^2)$, $m \neq m'$

Loss function: $\ell(w, z) = (w - z)^2$.

ERM solution & Mutual Information:

$$W_{\text{ERM}} = \frac{1}{n} \sum_{i=1}^n Z_i, \quad I(W_{\text{ERM}}; Z_i) = \frac{1}{2} \log \frac{n}{n-1}$$

True generalization error:

$$\mathbb{E} \left[\hat{L}(W_{\text{ERM}}, S) - L_{\mu'}(W_{\text{ERM}}) \right] = \frac{2\sigma^2}{n} + 2\sigma^2 D(\mu \| \mu')$$

Upper-Bound Slow:

$$\mathbb{E} [\text{gen}(W_{\text{ERM}})] \leq \sqrt{4\sigma^4 D(\mu \| \mu') + 16\sigma^4 D(\mu \| \mu')^2} + O\left(\frac{1}{\sqrt{n}}\right)$$

Fast:

$$\mathbb{E} [\text{gen}(W_{\text{ERM}})] \leq \frac{4\sigma^2}{n} + 4\sigma^2 D(\mu \| \mu')$$

InfoBoost Algorithm

- A novel boosting algorithm
- Recall the idea of boosting:
 - combine many weak learners to form a strong learner
 - Key step: adaptively adjust the weights of samples to train the weaker learners
- InfoBoost: use information measures to determine the weights to minimise the info-theoretic upper bound.

InfoBoost Algorithm

Algorithm 1: InfoBoost Algorithm

Input : Source, Target Datasets, Iteration T , Loss function

- ```

1 Initialize weights $\gamma_1(i) = \frac{1}{n}$ for all i ;
2 Choose hyper-parameters $\bar{\Gamma}, \zeta, \eta$ and iteration T ;
3 for $t = 1, \dots, T$ do
4 Learn a base hypothesis h_t using $S \cup S'$;
5 Learn h_t^{-i} using $S \cup S'$;
6 Learn domain hypothesis $h_{S'}$ and h_S using S' and S separately;
7 Learn a domain discriminator h_{dis} for the domain features;
8 $\epsilon_t = \sum_{i=1}^{n\beta} \gamma_t(i) \mathbf{1}_{h_t(x_i) \neq y_i} + \sum_{i=\beta n+1}^n \gamma_t(i) \mathbf{1}_{h_t(x_i) \neq y_i}$;
9 $\alpha_t = \log \frac{1-\epsilon_t}{\epsilon_t}$;
10 for $i = 1, \dots, \beta n$ do
11 Incur stability loss $d_i(h_t, h_t^{-i})$;
12 $\gamma_{t+1}(i) = \gamma_t(i) \cdot e^{\alpha_t \ell(h_t, x_i, y_i) - \eta d_i(h_t, h_t^{-i})}$;
13 end
14 for $i = \beta n + 1, \dots, n$ do
15 Incur stability loss $d_i(h_t, h_t^{-i})$;
16 Incur domain discrimination loss $\ell(h_{\text{dis}}(x_i))$;
17 Incur labelling divergence loss $d(h_{S'}(x_i, y_i), h_S(x_i, y_i))$;
18 $\gamma_{t+1}(i) = \gamma_t(i) \cdot e^{\alpha_t \ell(h_t, x_i, y_i) - \eta d_i(h_t, h_t^{-i}) - \zeta(\ell(h_{\text{dis}}(x_i)) + d_i(h_S, h_{S'}))}$;
19 end
20 Let $Z_t = \sum_{i=1}^n \gamma_t(i)$;
21 Normalize $\gamma_t(i) = \frac{\gamma_t(i)}{Z_t}$ for $i = 1, 2, \dots, n$;
22 end
Output: $f(x) = \mathbf{1}_{\sum_{t=1}^T \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^T \alpha_t}$.

```

- Stability loss: for evaluating  $I(W; Z_i)$ ;
- Exponential weighting: for minimising empirical loss  $\hat{L}$ ;
- Domain discrimination loss for evaluating  $D(\mu(X) \parallel \mu'(X))$ ;
- Domain labelling loss for evaluating  $D(\mu(Y|X) \parallel \mu'(Y|X))$ ;

# Example on Real-world Dataset

- Office-Caltech10



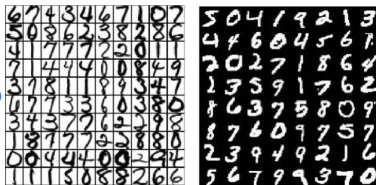
Amazon

DSLR

Webcam

Caltech10

- MNIST/USPS handwritten digits



- 20 Newsgroups: reviews from 20 different topics with various subcategories (science, computer, talk, etc.). Classify the topic.

# Results for OC10

|                   | Adaboost <sub>T</sub> | Adaboost <sub>S&amp;T</sub> | Tradaboost   | Transfer Boost | GapBoost     | InfoBoost    |
|-------------------|-----------------------|-----------------------------|--------------|----------------|--------------|--------------|
| $A \rightarrow W$ | 64.69                 | 75.47                       | 75.88        | 76.42          | 77.78        | <b>78.30</b> |
| $A \rightarrow D$ | 71.60                 | 75.83                       | 76.16        | 75.42          | 76.89        | <b>78.44</b> |
| $A \rightarrow C$ | 57.19                 | 76.88                       | <b>77.24</b> | 74.97          | 75.08        | 74.84        |
| $D \rightarrow A$ | 54.94                 | 56.98                       | 57.12        | 57.68          | 57.87        | <b>58.04</b> |
| $D \rightarrow W$ | 48.91                 | 51.84                       | 51.54        | 52.05          | 52.31        | <b>52.62</b> |
| $D \rightarrow C$ | 54.57                 | 60.97                       | 60.88        | 61.01          | <b>61.11</b> | 61.05        |
| $W \rightarrow A$ | 54.76                 | 62.90                       | 62.75        | 63.04          | 62.86        | <b>63.28</b> |
| $W \rightarrow D$ | 59.41                 | 61.32                       | <b>61.44</b> | 60.66          | 60.96        | 61.20        |
| $W \rightarrow C$ | 55.91                 | 63.07                       | 63.01        | 63.13          | 63.07        | <b>63.27</b> |
| $C \rightarrow A$ | 57.68                 | 76.34                       | <b>78.05</b> | 76.51          | 75.14        | 77.87        |
| $C \rightarrow D$ | 68.24                 | 68.61                       | 74.19        | 70.93          | 70.08        | <b>75.80</b> |
| $C \rightarrow W$ | 65.28                 | 75.01                       | 74.85        | 76.39          | 74.16        | <b>80.49</b> |
| Average           | 59.43                 | 67.10                       | 67.76        | 67.35          | 67.36        | <b>68.77</b> |

# Results for M/U and 20News

**Table:** Accuracy in % with 10% training target data for 20 Newsgroups data

| Tasks        | AdaBoost $\mathcal{T}$ | AdaBoost $\mathcal{S \& T}$ | TradaBoost       | TransferBoost                      | GapBoost         | InfoBoost                          |
|--------------|------------------------|-----------------------------|------------------|------------------------------------|------------------|------------------------------------|
| rec vs talk  | 90.30 $\pm$ 3.14       | 87.42 $\pm$ 2.21            | 71.97 $\pm$ 5.89 | 89.42 $\pm$ 2.99                   | 91.77 $\pm$ 3.51 | <b>93.11 <math>\pm</math> 1.71</b> |
| comp vs sci  | 92.87 $\pm$ 1.03       | 84.66 $\pm$ 1.53            | 92.71 $\pm$ 1.78 | 93.35 $\pm$ 1.05                   | 93.55 $\pm$ 2.54 | <b>94.15 <math>\pm</math> 1.04</b> |
| rec vs sci   | 90.17 $\pm$ 1.02       | 86.68 $\pm$ 1.06            | 90.49 $\pm$ 1.36 | <b>93.62 <math>\pm</math> 0.69</b> | 90.70 $\pm$ 2.14 | 92.58 $\pm$ 0.89                   |
| talk vs sci  | 88.66 $\pm$ 2.74       | 71.72 $\pm$ 3.68            | 81.39 $\pm$ 4.98 | 87.78 $\pm$ 3.25                   | 90.14 $\pm$ 2.55 | <b>90.40 <math>\pm</math> 1.78</b> |
| comp vs rec  | 91.44 $\pm$ 1.35       | 87.90 $\pm$ 1.80            | 91.52 $\pm$ 1.12 | <b>94.95 <math>\pm</math> 0.60</b> | 92.62 $\pm$ 2.95 | 94.00 $\pm$ 1.58                   |
| comp vs talk | 93.86 $\pm$ 1.46       | 89.57 $\pm$ 1.32            | 75.78 $\pm$ 1.14 | 94.83 $\pm$ 0.78                   | 94.90 $\pm$ 1.02 | <b>95.08 <math>\pm</math> 0.89</b> |
| Average      | 91.22                  | 84.66                       | 83.98            | 92.33                              | 92.28            | <b>93.22</b>                       |

**Figure:** Accuracy in % with 10% training target data for MNIST/USPS data

| Tasks                    | AdaBoost $\mathcal{T}$ | AdaBoost $\mathcal{S \& T}$ | TradaBoost        | TransferBoost                      | GapBoost                           | InfoBoost                          |
|--------------------------|------------------------|-----------------------------|-------------------|------------------------------------|------------------------------------|------------------------------------|
| U to M <sub>1</sub> vs 7 | 61.52 $\pm$ 4.15       | 57.73 $\pm$ 1.92            | 52.67 $\pm$ 1.08  | 63.47 $\pm$ 2.34                   | 63.58 $\pm$ 1.99                   | <b>64.51 <math>\pm</math> 2.11</b> |
| U to M <sub>2</sub> vs 3 | 59.69 $\pm$ 3.96       | 56.49 $\pm$ 1.87            | 52.18 $\pm$ 3.87  | 63.29 $\pm$ 1.71                   | <b>63.37 <math>\pm</math> 1.87</b> | 62.83 $\pm$ 1.57                   |
| U to M <sub>5</sub> vs 6 | 61.05 $\pm$ 2.87       | 57.22 $\pm$ 1.24            | 55.15 $\pm$ 1.44  | 61.54 $\pm$ 1.74                   | 61.46 $\pm$ 1.70                   | <b>62.66 <math>\pm</math> 1.63</b> |
| U to M <sub>0</sub> vs 8 | 59.78 $\pm$ 3.75       | 57.61 $\pm$ 2.57            | 51.48 $\pm$ 1.29  | 64.50 $\pm$ 1.88                   | 64.77 $\pm$ 1.88                   | <b>65.00 <math>\pm</math> 2.02</b> |
| U to M <sub>4</sub> vs 9 | 57.97 $\pm$ 3.12       | 58.03 $\pm$ 4.24            | 57.15 $\pm$ 4.69  | 61.87 $\pm$ 1.64                   | <b>61.91 <math>\pm</math> 1.71</b> | 61.51 $\pm$ 2.01                   |
| M to U <sub>1</sub> vs 7 | 87.42 $\pm$ 4.94       | 90.44 $\pm$ 12.54           | 72.64 $\pm$ 13.98 | 92.23 $\pm$ 3.66                   | 91.26 $\pm$ 3.50                   | <b>92.43 <math>\pm</math> 3.75</b> |
| M to U <sub>2</sub> vs 3 | 74.38 $\pm$ 5.33       | 74.32 $\pm$ 11.68           | 76.80 $\pm$ 7.34  | <b>77.01 <math>\pm</math> 7.08</b> | 76.94 $\pm$ 5.83                   | 76.29 $\pm$ 6.34                   |
| M to U <sub>5</sub> vs 6 | 71.87 $\pm$ 7.14       | 59.11 $\pm$ 8.62            | 69.53 $\pm$ 3.79  | 69.42 $\pm$ 7.92                   | 75.33 $\pm$ 5.83                   | <b>75.63 <math>\pm</math> 5.60</b> |
| M to U <sub>0</sub> vs 8 | 74.47 $\pm$ 3.85       | 74.84 $\pm$ 4.87            | 71.83 $\pm$ 4.57  | 77.89 $\pm$ 4.08                   | <b>78.43 <math>\pm</math> 4.22</b> | 77.96 $\pm$ 4.43                   |
| M to U <sub>4</sub> vs 9 | 87.88 $\pm$ 4.04       | 88.80 $\pm$ 5.93            | 79.24 $\pm$ 1.43  | 91.07 $\pm$ 4.03                   | 91.21 $\pm$ 3.78                   | <b>91.33 <math>\pm</math> 3.94</b> |
| Average                  | 69.60                  | 67.46                       | 63.87             | 72.23                              | 72.83                              | <b>73.02</b>                       |

# Contents

Generalization in Machine Learning

Information-Theoretic Bounds for Generalization

Fast Rate Generalization Error

Information-theoretic Bounds for Transfer Learning

Transfer Learning under Causal Framework

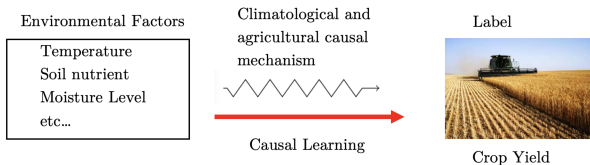
# Supervised Transfer Learning Problems

- IID Labeled source data:  $D_s^m = (X_s^{(1)}, Y_s^{(1)}, \dots, X_s^{(m)}, Y_s^{(m)}) \sim P_S^m(X, Y)$
- IID unlabeled target data:  $D_t^{U,n} = (X_t^{(1)}, \dots, X_t^{(n)}) \sim P_T^n(X)$  where  $P_T(X)$  is the marginal of  $P_T(X, Y)$
- Learning task: given source and target training data, we want to predict  $Y'_t$  for the new data  $X'_t$  in the target domain
- Algorithm  $\mathcal{A} : \mathcal{Z}^m \times \mathcal{X}^n \rightarrow \mathcal{W}$
- Loss function  $\ell : \mathcal{W} \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$
- Excess risk:

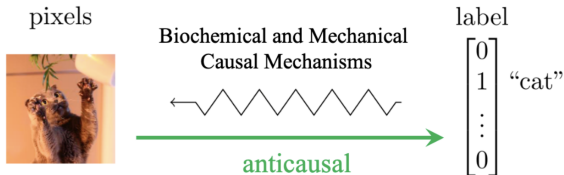
$$\mathcal{R}(n, m) := \mathbb{E}_{D_s^m, D_t^{U,n}, X'_t, Y'_t} [\ell(W, X'_t, Y'_t) - \ell(w^*, X'_t, Y'_t)],$$

- **Question:** When will the source and target data be useful for prediction?

# Learning from Causality



(a) Causal Learning (Bernhard et al., 2012)



(b) Anti-causal Learning (Bernhard et al., 2012)

# Potential Outcome and Causation

- Consider  $X \rightarrow Y$
- Potential Outcome Definition (Imbens & Rubin, 2015):

$Y_x$  is the potential outcome under treatment  $X = x$

- When we assign  $X = 1$ , we can only observe  $Y_1$  but not  $Y_0$ .  $Y_1$  and  $Y_0$  are predefined - which is independent of  $X$  and each other!
- In this sense, we can write the outcome  $Y$  as:

$$Y = \sum_{x \in \mathcal{X}} 1_{X=x} Y_x$$

- This entails:

$$P(X = x, Y = y) = P(X = x)P(Y_x = y)$$

# Causal and Anti-causal Learning

- Feature-label pairs, e.g.,  $(X, Y)$ ;
- Causal Learning:  $X \rightarrow Y$ ;
- Anti-causal learning:  $Y \rightarrow X$ ;
- **The learning task is always predicting  $Y$  from  $X$ !**
- Under causal learning:

$$P(X = x, Y = y) = P(X = x)P(Y_x = y)$$

Knowing  $X$  does **not** help with predicting  $Y$ !

- Under anti-causal learning:

$$P(X = x, Y = y) = P(Y = y)P(X_y = x)$$

Knowing  $X$  does help with predicting  $Y$ !

- Value of labeled source data and unlabeled target data in transfer learning problem?

# Unified Framework

Causal Learning:

$$P(X = x, Y = y) = P(X = x)P(Y_x = y)$$

General Shift:  $P_S(X, Y) \neq P_T(X, Y)$

Concept Drift:  $P_S(Y|X) \neq P_T(Y|X)$

Covariate Shift:  $P_S(X) \neq P_T(X)$

Semi-supervised Learning:  $P_S(X, Y) = P_T(X, Y)$

Anti-causal Learning:

$$P(X = x, Y = y) = P(Y = y)P(X_y = x)$$

General Shift:  $P_S(X, Y) \neq P_T(X, Y)$

Conditional Shift:  $P_S(X|Y) \neq P_T(X|Y)$

Target Shift:  $P_S(Y) \neq P_T(Y)$

Semi-supervised Learning:  $P_S(X, Y) = P_T(X, Y)$

# Main Results

## Theorem (Excess Risk Characterization)

*Under certain assumptions and with the proposed mixture strategy, the excess risk can be expressed as*

$$\mathcal{R}(n, m) = I(Y'_t; \theta_t^*, \theta_s^* | D_s^m, D_t^{U, n}, X'_t).$$

| Causal Model      | Conditions                                   | Unlabeled Target | Source Data | Rate                              |
|-------------------|----------------------------------------------|------------------|-------------|-----------------------------------|
| $X \rightarrow Y$ | $P_S(X) \neq P_T(X), P_S(Y X) \neq P_T(Y X)$ | ✗                | ✗           | -                                 |
|                   | $P_S(X) \neq P_T(X), P_S(Y X) = P_T(Y X)$    | ✗                | ✓           | $O(\frac{k}{m})$                  |
|                   | $P_S(X) = P_T(X), P_S(Y X) \neq P_T(Y X)$    | ✗                | ✗           | -                                 |
|                   | $P_S(X) = P_T(X), P_S(Y X) = P_T(Y X)$       | ✗                | ✓           | $O(\frac{k}{m})$                  |
| $Y \rightarrow X$ | $P_S(Y) \neq P_T(Y), P_S(X Y) \neq P_T(X Y)$ | ✓                | ✗           | $O(\frac{1+k'}{n})$               |
|                   | $P_S(Y) \neq P_T(Y), P_S(X Y) = P_T(X Y)$    | ✓                | ✓           | $O(\frac{1}{n} + \frac{k'}{n+m})$ |
|                   | $P_S(Y) = P_T(Y), P_S(X Y) \neq P_T(X Y)$    | ✓                | ✓           | $O(\frac{k'}{n} + \frac{1}{n+m})$ |
|                   | $P_S(Y) = P_T(Y), P_S(X Y) = P_T(X Y)$       | ✓                | ✓           | $O(\frac{k'+1}{m+n})$             |

Figure: Results On Performance

| Causal Model      | Conditions                                   | Unlabeled Target | Source Data | Rate                              |
|-------------------|----------------------------------------------|------------------|-------------|-----------------------------------|
| $X \rightarrow Y$ | $P_S(X) \neq P_T(X), P_S(Y X) \neq P_T(Y X)$ | ✗                | ✗           | -                                 |
|                   | $P_S(X) \neq P_T(X), P_S(Y X) = P_T(Y X)$    | ✗                | ✓           | $O(\frac{k}{m})$                  |
|                   | $P_S(X) = P_T(X), P_S(Y X) \neq P_T(Y X)$    | ✗                | ✗           | -                                 |
|                   | $P_S(X) = P_T(X), P_S(Y X) = P_T(Y X)$       | ✗                | ✓           | $O(\frac{k}{m})$                  |
| $Y \rightarrow X$ | $P_S(Y) \neq P_T(Y), P_S(X Y) \neq P_T(X Y)$ | ✓                | ✗           | $O(\frac{1+k'}{n})$               |
|                   | $P_S(Y) \neq P_T(Y), P_S(X Y) = P_T(X Y)$    | ✓                | ✓           | $O(\frac{1}{n} + \frac{k'}{n+m})$ |
|                   | $P_S(Y) = P_T(Y), P_S(X Y) \neq P_T(X Y)$    | ✓                | ✓           | $O(\frac{k'}{n} + \frac{1}{n+m})$ |
|                   | $P_S(Y) = P_T(Y), P_S(X Y) = P_T(X Y)$       | ✓                | ✓           | $O(\frac{k'+1}{m+n})$             |

## ■ Causal Learning:

- Unlabeled target data does not help improve the learning rate
- The source data is helpful when the conditional distribution  $P(Y|X)$  is the same
- The convergence rate only depends on the source sample size

## ■ Anticausal Learning:

- Unlabeled target data is always helpful
- The source data is partially helpful depending on the conditions whether  $P(Y)$  or  $P(Y|X)$  are identical

## ■ Coincide with the empirical results in [Bernhard et al., 2012]

## ■ Unified framework for excess risk in transfer learning, recovering other works under parametric setup

# Summary

- Traditional generalization error bounds
- Information-theoretic bounds for the generalization error
- Improved generalization error bound - fast rate results
- Information-theoretic analysis for transfer learning - first attempt on the practical algorithm
- Information-theoretic bounds under the causal framework - the value of the data
- **\*Bounds on the online/instantaneous/time-variant transfer learning - negative transfer**
- **\*Transfer learning algorithms on benchmark datasets and health-care applications - practical side**

# Research Outcomes

- **Wu, X.**, Akbarzadeh Khorshidi, H., Aickelin, U., Edib, Z., & Peate, M. (2019). Imputation techniques on missing values in breast cancer treatment and fertility data. *Health Information Science and Systems*, 7(1), 1-8.
- **Wu, X.**, Khorshidi, H. A., Aickelin, U., Edib, Z., & Peate, M. (2020). Transfer learning to enhance amenorrhea status prediction in cancer and fertility data with missing values. In *Artificial Intelligence: Applications in Healthcare Delivery* (pp. 233-260). Productivity Press.
- **Wu, X.**, Manton, J. H., Aickelin, U., & Zhu, J. (2020). Information-theoretic analysis for transfer learning. In *2020 IEEE ISIT*.
- **Wu, X.**, Manton, J. H., Aickelin, U., & Zhu, J. (2021). Online Transfer Learning: Negative Transfer and Effect of Prior Knowledge. In *2021 IEEE ISIT*.
- **Wu, X.**, Manton, J. H., Aickelin, U., & Zhu, J. (2021). A Bayesian Approach to (Online) Transfer Learning: Theory and Algorithms. *arXiv preprint arXiv:2109.01377*, to appear in *Journal of Artificial Intelligence*
- **Wu, X.**, Gong, M., Manton, J. H., Aickelin, U., & Zhu, J. (2022). On Causality in Domain Adaptation and Semi-Supervised Learning: an Information-Theoretic Analysis. *arXiv preprint arXiv:2205.04641*, submitted to *JMLR*, under (equivalently) major review
- **Wu, X.**, Manton, J. H., Aickelin, U., & Zhu, J. (2022). Fast Rate Generalization Error Bounds: Variations on a Theme. In *2022 IEEE ITW*.
- **Wu, X.**, Manton, J. H., Aickelin, U., & Zhu, J. (2022). An Information-Theoretic Analysis for Transfer Learning: Error Bounds and Applications. *arXiv preprint arXiv:2207.05377*, submitted to *TIT*.
- Jia, J., **Wu, X.**, Evans, J., Z., & Zhu, J. (2023). On the Value of Stochastic Side Information in Online Learning, submitted to *ICASSP 2023*
- **Wu, X.**, Manton, J. H., Aickelin, U., & Zhu, J. (2023). Fast Rate Information-theoretic Generalization Error Bounds, On Preparation.

Thanks For Your Listening!