Transfer Learning: Theories and Algorithms

Mingsheng Long

School of Software, Tsinghua University
National Engineering Laboratory for Big Data Software

mingsheng@tsinghua.edu.cn
http://ise.thss.tsinghua.edu.cn/~mlong
Chinese Conference on Data Mining, CCDM 2020

Outline

Supervised Learning

Learner: $f : \boldsymbol{x} \to y$ **Distribution:** $(\boldsymbol{x}, y) \sim P(\boldsymbol{x}, y)$

fish

bird

mammal

tree

flower

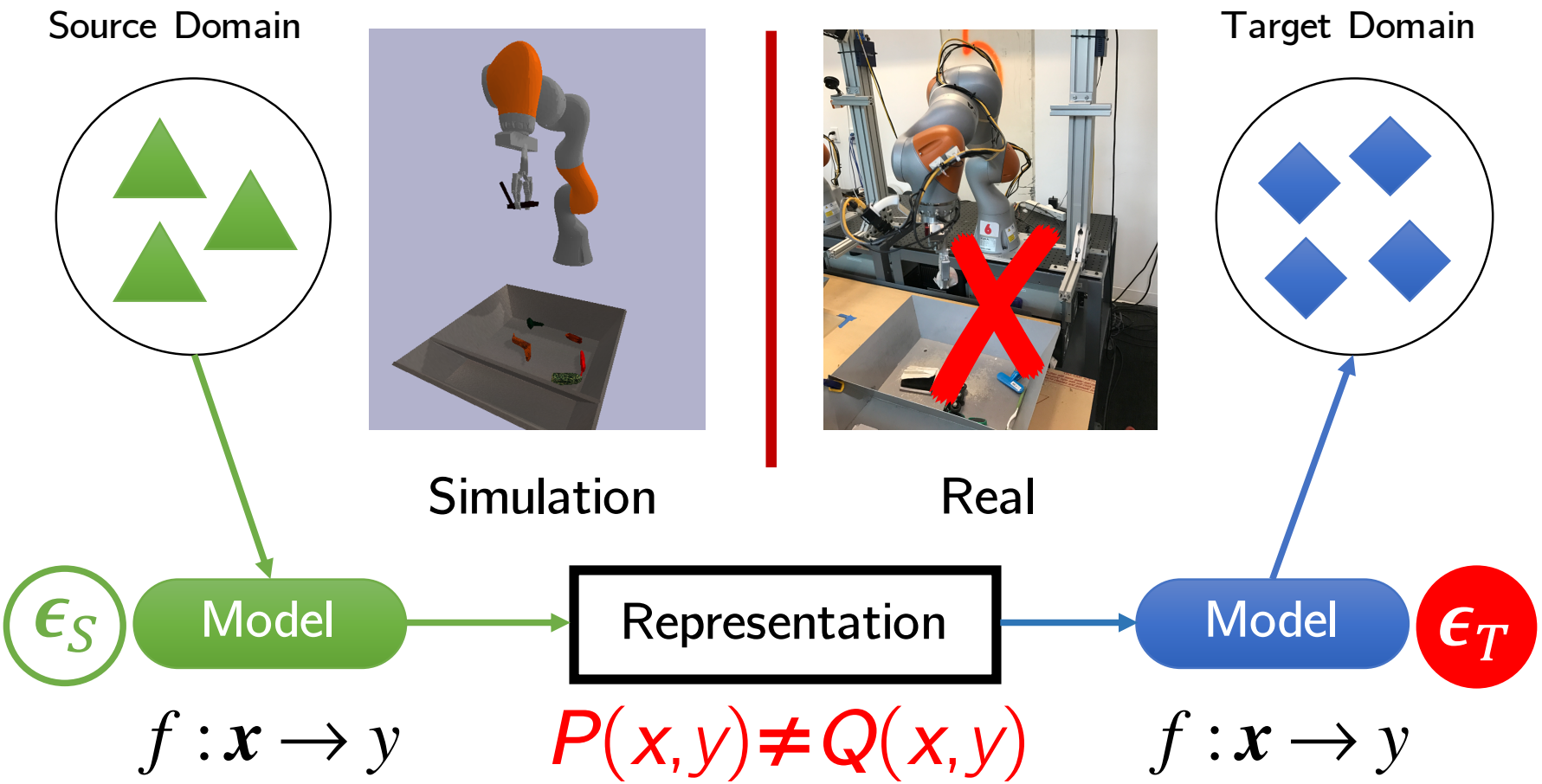
.....

IID Setup

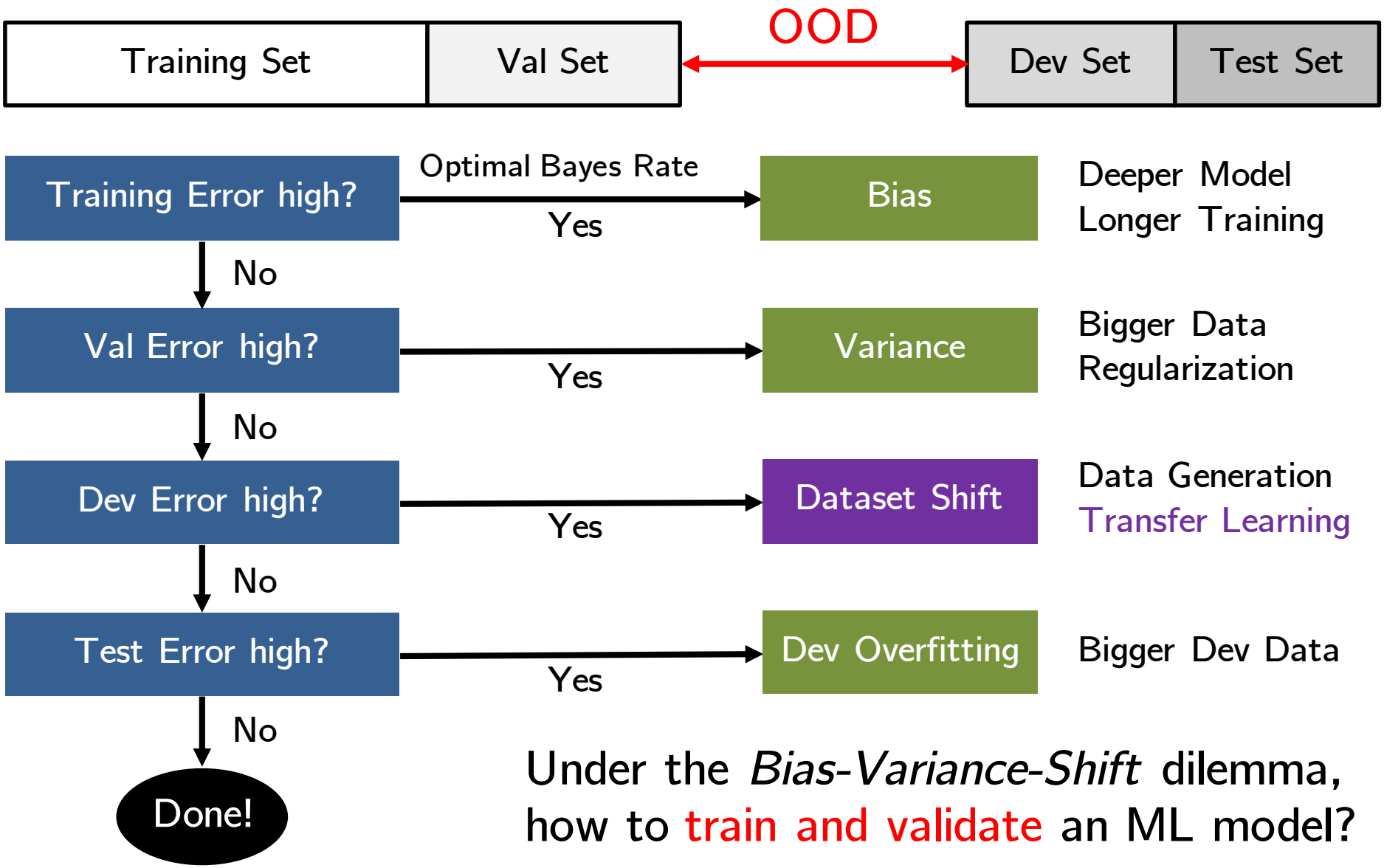
Error Bound: $\epsilon_{\text{test}} \leq \hat{\epsilon}_{\text{train}} + \sqrt{\frac{\text{complexity}}{n}}$

Transfer Learning

- Machine learning across domains of different distributions $P \neq Q$
  - **OOD: Out-of-Distribution** (from IID to Non-IID to OOD)
- How to bound generalization error on target domain for OOD setup?



Bias-Variance-Shift Dilemma

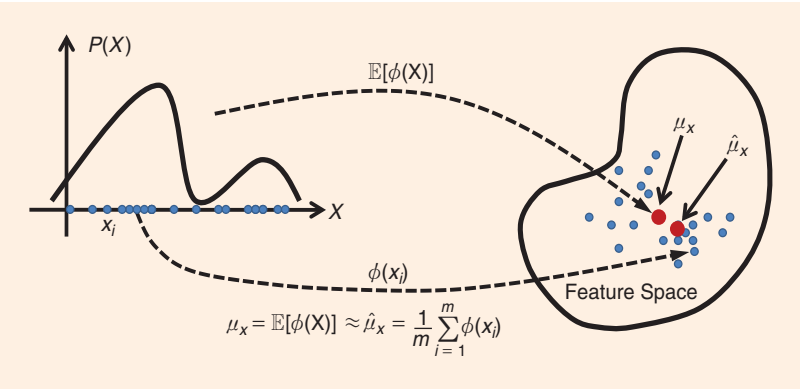


Under the *Bias-Variance-Shift* dilemma, how to train and validate an ML model?

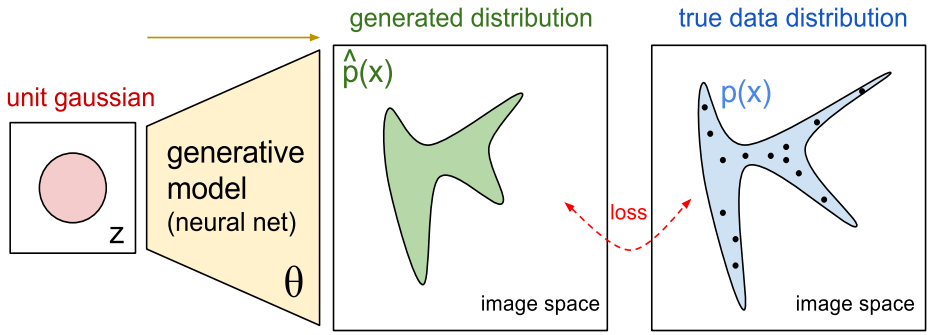
Representative Approaches to Transfer Learning

Learning to **match distributions** across **OOD** domains *s.t.* $P \approx Q$

- Covariate shift: $P(\mathbf{X}) \neq Q(\mathbf{X})$ (mainstream work of this setup)
- Prior shift: $P(\mathbf{Y}) \neq Q(\mathbf{Y})$ (challenging, current hotspot)
- Conditional shift: $P(Y|\mathbf{X}) \neq Q(Y|\mathbf{X})$ (challenging, future research)



Kernel Embedding

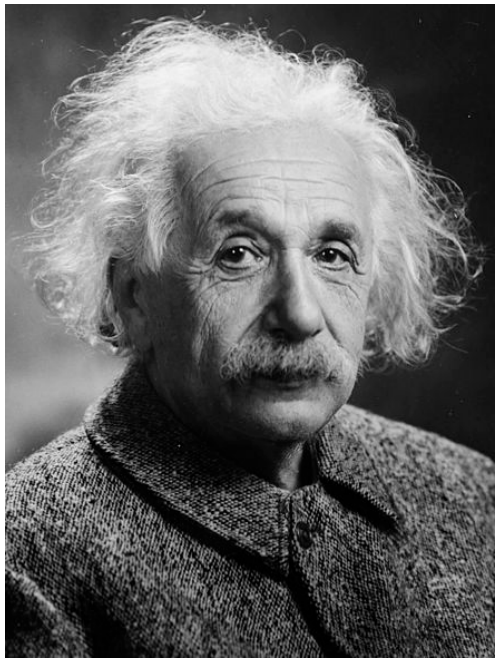


Adversarial Learning

Generally, no theoretical guarantees!

Song et al. Kernel Embeddings of Conditional Distributions. ***IEEE****, 2013.*
Goodfellow et al. Generative Adversarial Networks. ***NIPS*** *2014.*

Principal Problem: Bridging Theory and Algorithm



Everything should be made as simple as possible, but no simpler.

—Albert Einstein

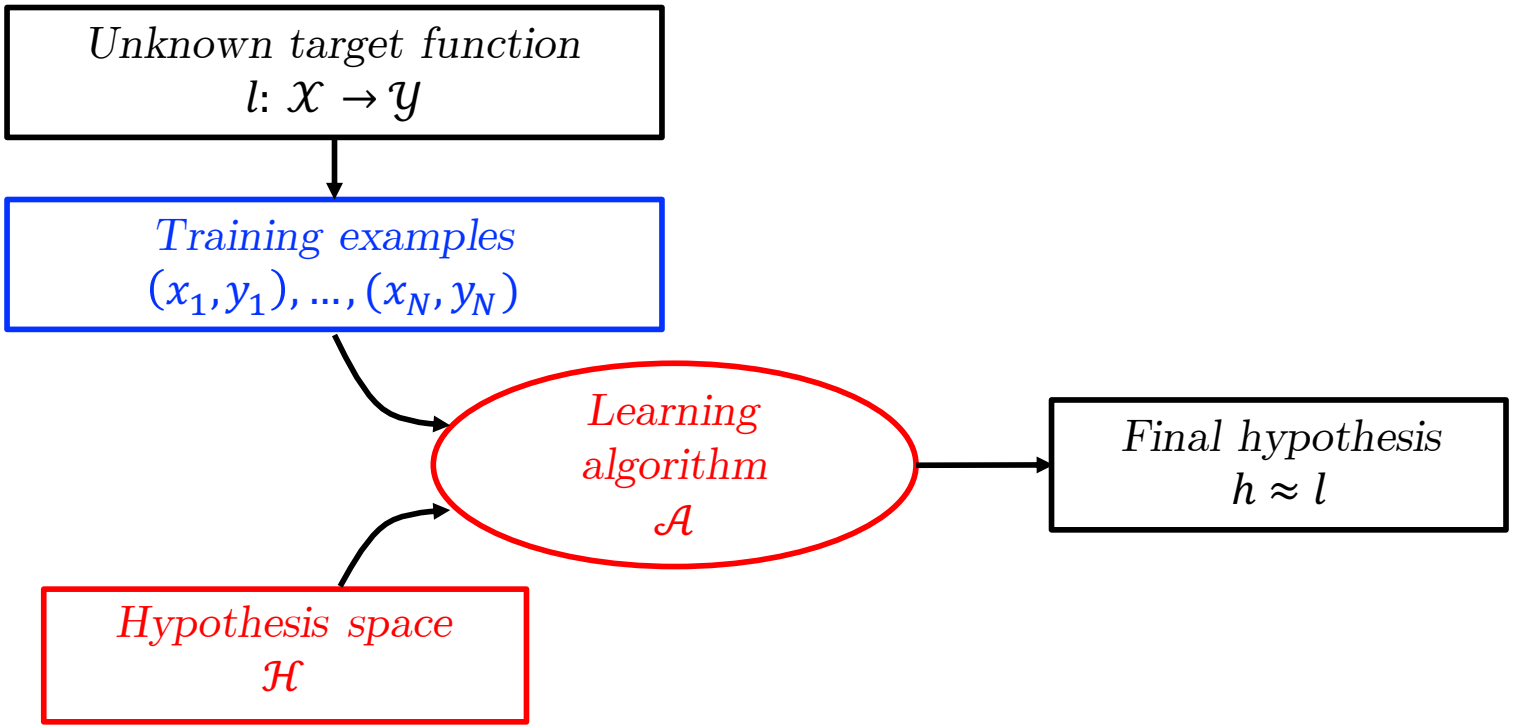
There is nothing more practical than a good theory.

—Vladimir Vapnik

Outline

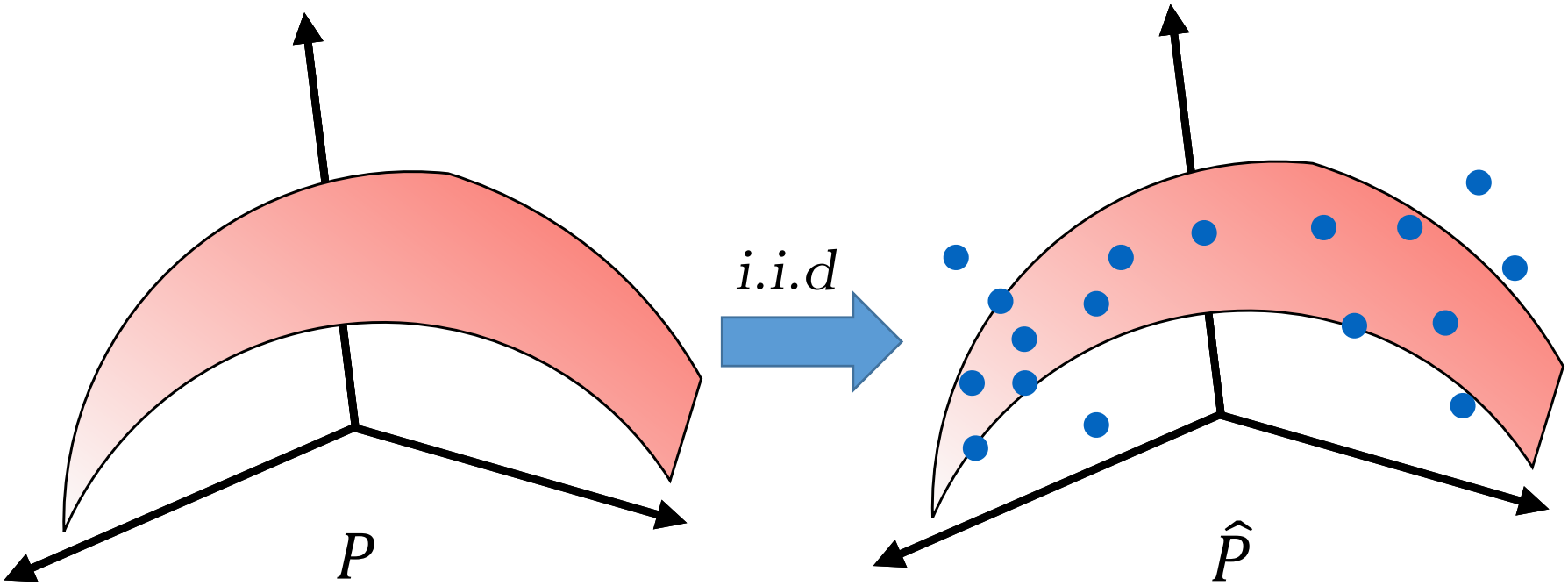
Outline

Machine Learning Framework



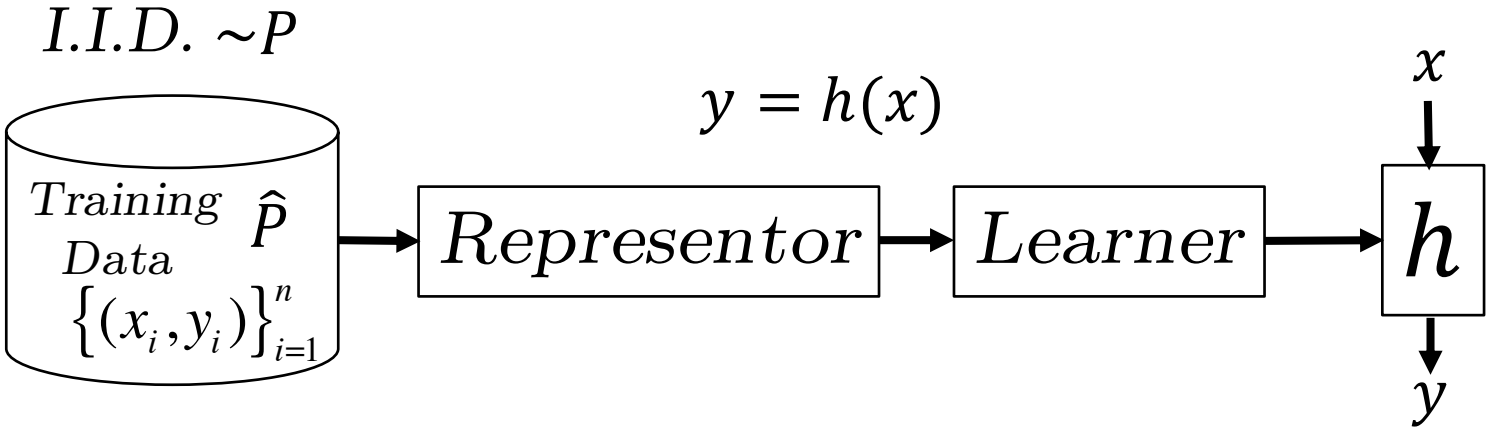
- Algorithms that (automatically) improve their performance (P) at some task (T) with experience (E).
  - Hypothesis space $\mathcal{H}$ — all the possible functions to search from.
  - Learning algorithm $\mathcal{A} : \mathcal{D} \rightarrow \mathcal{H}$ — search for the *best* hypothesis.

Statistical Learning View



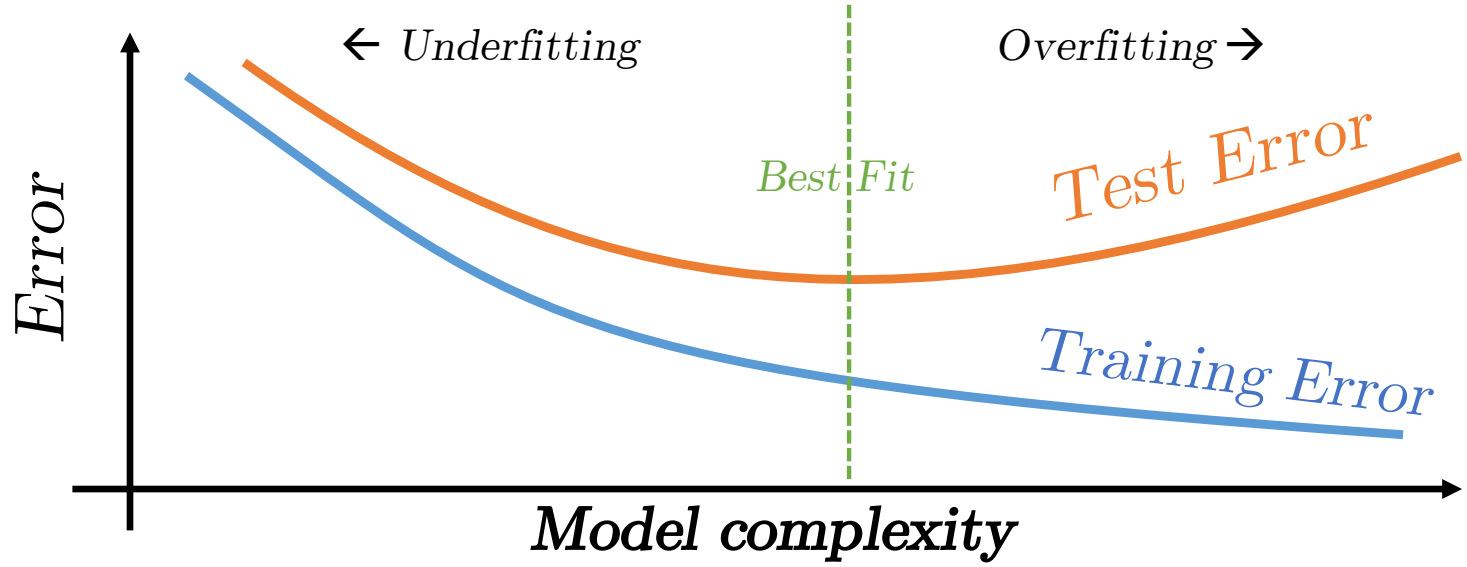
- **All Statistics:** There is a latent data generating distribution $P_{\mathcal{X}\times\mathcal{Y}}$.
- **IID Assumption:** All training and testing pairs $\widehat{P} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ are generated *i.i.d.* from $P_{\mathcal{X}\times\mathcal{Y}}$.

Statistical Learning Formulation



- Formally analyzing the classification problem with **01-loss** $[\cdot \neq \cdot]$.
- Training error: $\epsilon_{\widehat{P}}(h) = \frac{1}{n}\sum_{i=1}^{n}[h(\mathbf{x}_i) \neq y_i] = \mathbb{E}_{(\mathbf{x},y)\sim\widehat{P}}[h(\mathbf{x}) \neq y]$.
- Test error: $\epsilon_P(h) = \mathbb{E}_{(\mathbf{x},y)\sim P}[h(\mathbf{x}) \neq y]$.
- Training error is an *unbiased* estimation of test error.
- Principal problem: Can we control $\epsilon_P(h)$ with observable $\epsilon_{\widehat{P}}(h)$?

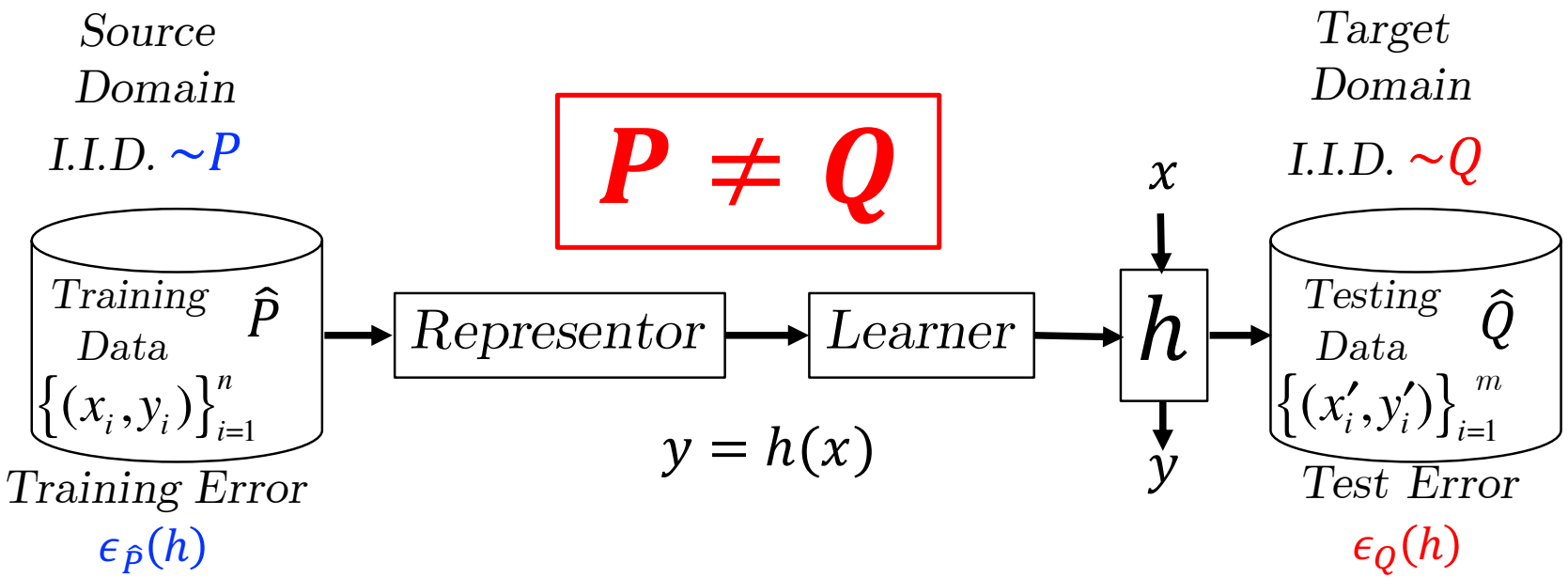
Statistical Learning Theory



- **Generalization error**: The gap between training error and test error.
- Generalization error depends on sample size n and model complexity.
- For hypothesis space $\mathcal{H}$ with VC-dimension d, we have bound:

$$\epsilon_P(h) \leq \epsilon_{\widehat{P}}(h) + O\left(\sqrt{\frac{d \log n + \log \frac{2}{\delta}}{n}}\right)$$

Transfer Learning Formulation



- Only have labeled data sampled from a different source domain P.
- And unlabeled data sampled from a target domain Q. $\epsilon_{\widehat{Q}}(h)$ is not observable!
- Principal problem: Can we control target error $\epsilon_Q(h)$?
- **Disparity on** D is defined by: $\epsilon_D(h_1, h_2) = \mathbb{E}_{(\mathbf{x},y)\sim D}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]$.
- Good news: Computation of disparity does not require (target) label!

Relating Target Risk to Source Risk

Theorem (Bound with Disparity)

For classification tasks of transfer learning, define the ideal joint hypothesis as $h^* = \arg\min_{h \in \mathcal{H}} [\epsilon_P(h) + \epsilon_Q(h)]$, *the target risk* $\epsilon_Q(h)$ *can be bounded by the source risk* $\epsilon_P(h)$, *the ideal joint error, and the disparity difference:*

$$\epsilon_Q(h) \leqslant \epsilon_P(h) + [\epsilon_P(h^*) + \epsilon_Q(h^*)] + |\epsilon_P(h, h^*) - \epsilon_Q(h, h^*)| \tag{1}$$

Proof.

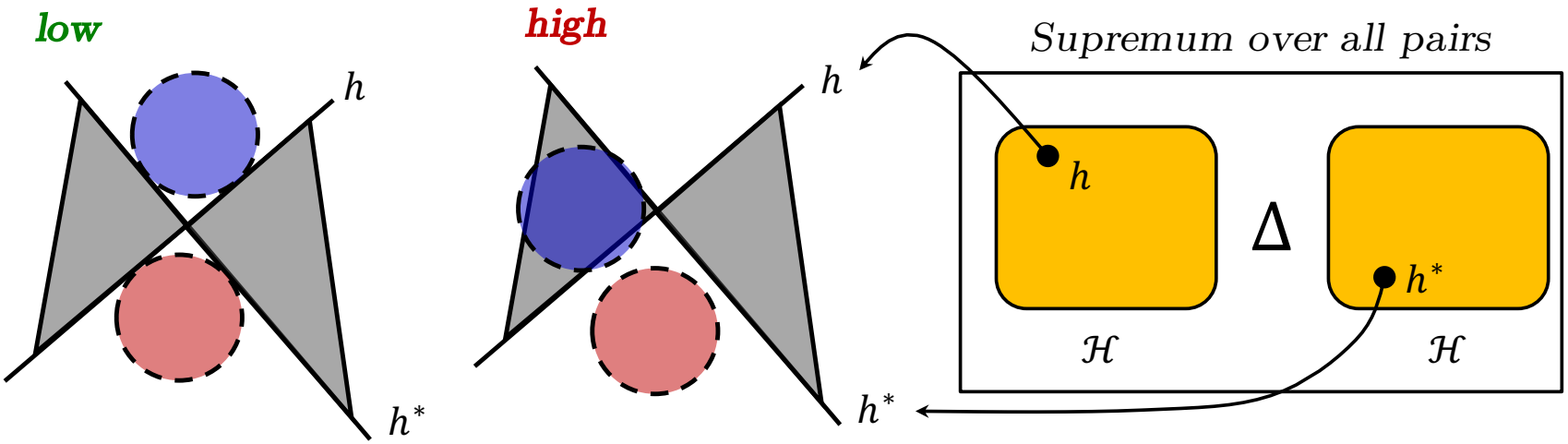
Simply using the *triangle inequalities* of the 01-loss, we have

$$\begin{aligned}
\epsilon_Q(h) &\leqslant \epsilon_Q(h^*) + \epsilon_Q(h, h^*) \\
&= \epsilon_Q(h^*) + \epsilon_P(h, h^*) + \epsilon_Q(h, h^*) - \epsilon_P(h, h^*) \\
&\leqslant \epsilon_Q(h^*) + \epsilon_P(h, h^*) + |\epsilon_Q(h, h^*) - \epsilon_P(h, h^*)| \\
&\leqslant \epsilon_P(h) + [\epsilon_P(h^*) + \epsilon_Q(h^*)] + |\epsilon_P(h, h^*) - \epsilon_Q(h, h^*)|
\end{aligned} \tag{2}$$

□

$\mathcal{H}\Delta\mathcal{H}$-Divergence[1]

- **Assumption:** Small ideal joint error $\epsilon_{ideal} = \epsilon_P(h^*) + \epsilon_Q(h^*)$.
- We can illustrate the disparity difference $|\epsilon_P(h, h^*) - \epsilon_Q(h, h^*)|$:



- However, h^* is unknown and h is undefined!
- $\mathcal{H}\Delta\mathcal{H}$-**Divergence**: $d_{\mathcal{H}\Delta\mathcal{H}}(P, Q) \triangleq \sup_{h,h' \in \mathcal{H}} |\epsilon_P(h, h') - \epsilon_Q(h, h')|$
- Can be estimated from finite unlabeled samples of source and target.

[1] *Ben-David et al. A Theory of Learning from Different Domains. Machine Learning, 2010.*

Bound $\mathcal{H}\Delta\mathcal{H}$-Divergence with Domain Discriminator

Theorem (Generalization Bound with $\mathcal{H}\Delta\mathcal{H}$-Divergence)

Denote by d the VC-dimension of hypothesis space $\mathcal{H}$. We have

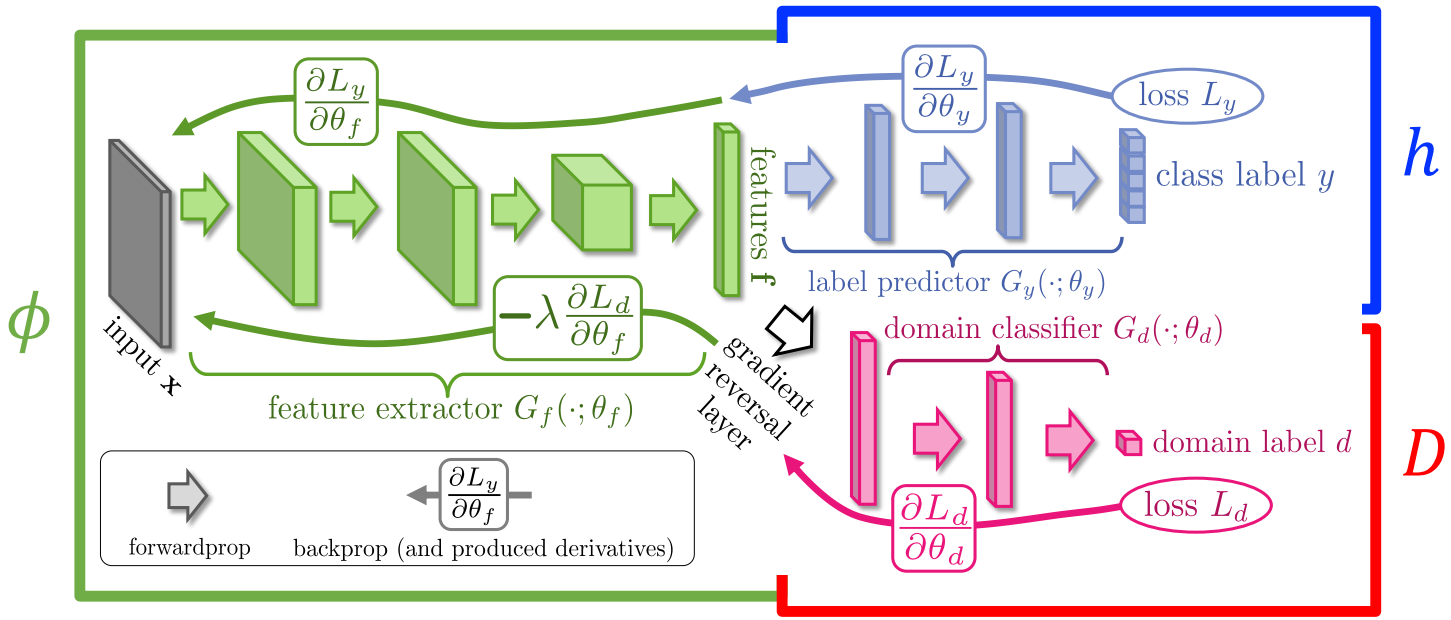
$$\epsilon_Q(h) \leq \epsilon_{\hat{P}}(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{P}, \widehat{Q}) + \epsilon_{ideal} + O\left(\sqrt{\frac{d\log n}{n}} + \sqrt{\frac{d\log m}{m}}\right) \quad (3)$$

- However, $\mathcal{H}\Delta\mathcal{H}$-Divergence is hard to compute and optimize.
- For binary hypothesis h, $\mathcal{H}\Delta\mathcal{H}$-Divergence can be further bounded by

$$\begin{aligned} d_{\mathcal{H}\Delta\mathcal{H}}(P, Q) &\triangleq \sup_{h,h' \in \mathcal{H}} \left|\epsilon_P\left(h, h'\right) - \epsilon_Q\left(h, h'\right)\right| \\ &= \sup_{\delta \in \mathcal{H}\Delta\mathcal{H}} \left|\mathbb{E}_P\left[\delta(\mathbf{x}) \neq 0\right] - \mathbb{E}_Q\left[\delta\left(\mathbf{x}\right) \neq 0\right]\right| \\ &\leqslant \sup_{D \in \mathcal{H}_D} \left|\mathbb{E}_P\left[D(\mathbf{x}) = 1\right] + \mathbb{E}_Q\left[D\left(\mathbf{x}\right) = 0\right]\right| \end{aligned} \quad (4)$$

- This bound can be estimated by training a domain discriminator $D(\mathbf{x})$.
- Under strong assumption that $\mathcal{H}\Delta\mathcal{H} \subset \mathcal{H}_D$ (universal 2-layer DNN).

Domain Adversarial Neural Network (DANN)[2]



Adversarial domain adaptation: learn ϕ to minimize $d_{\mathcal{H}\Delta\mathcal{H}}(\phi(P), \phi(Q))$.

$$\min_{\phi,h}\left\{\mathbb{E}_{(x,y)\sim P}L(h(\phi(x)),y)+\max_{D}\left(\mathbb{E}_{P}L(D(\phi(x)),1)+\mathbb{E}_{Q}L(D(\phi(x)),0)\right)\right\} \quad (5)$$

Supervised Learning on source + Upper-Bound of $d_{\mathcal{H}\Delta\mathcal{H}}$ on source/target

[2] *Ganin et al. Domain Adversarial Training of Neural Networks. JMLR 2016.*

Outline

Theory and Practice: Gap Exists for Decade

- **Theory vs. Practice:**
- Binary Classification vs. Multiclass Classification.
- Discrete Classifier vs. Classifier with Scoring Function.
- $d_{\mathcal{H}\Delta\mathcal{H}}$ does not need label vs. $d_{\mathcal{H}\Delta\mathcal{H}}$ is hard to compute and optimize.
- **Principal problem: How to bridge theory and algorithm?**

Step I: Disparity Discrepancy (DD)[3]

Definition (Disparity Discrepancy (DD))

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

$$d_{h,\mathcal{H}}(P, Q) = \sup_{h' \in \mathcal{H}} \left(\mathbb{E}_Q[h' \neq h] - \mathbb{E}_P[h' \neq h] \right) \tag{6}$$

- The supremum in Disparity Discrepancy (DD) is taken over only one hypothesis space $\mathcal{H}$ without $|\cdot|$, which can be optimized more easily.

Theorem (Bound with Disparity Discrepancy)

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

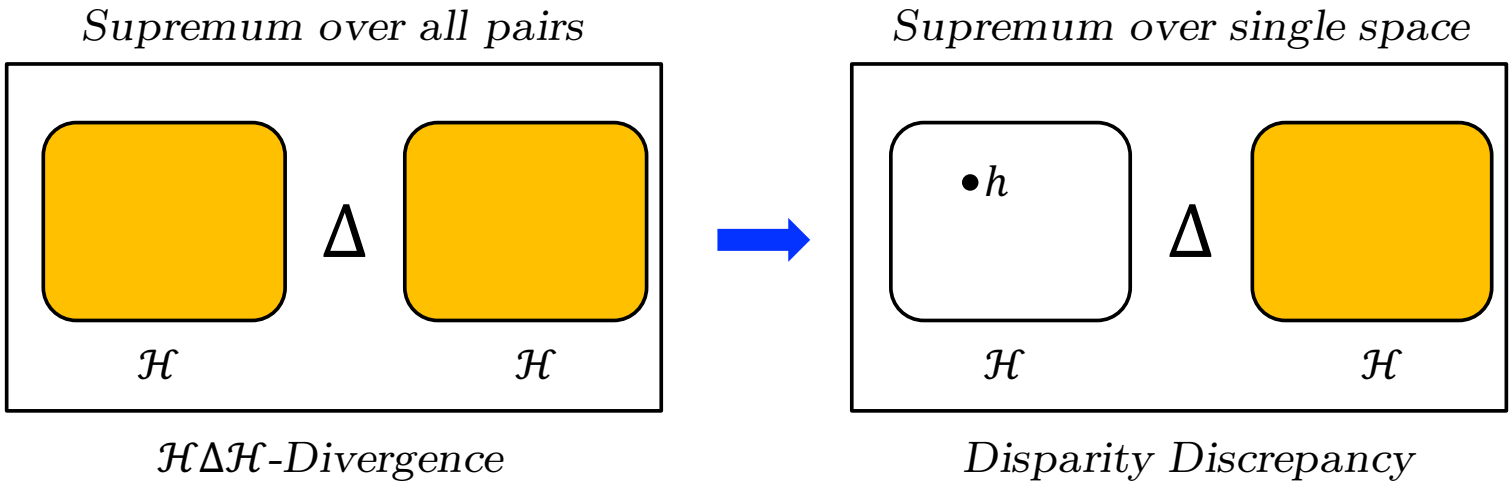
$$\epsilon_Q(h) \leq \epsilon_P(h) + d_{h,\mathcal{H}}(P, Q) + \epsilon_{ideal}, \tag{7}$$

where $\epsilon_{ideal} = \epsilon(\mathcal{H}, P, Q)$ *is the ideal joint error.*

[3] *Zhang & Long. Bridging Theory and Algorithm for Domain Adaptation. ICML 2019.*

Step I: Disparity Discrepancy (DD)

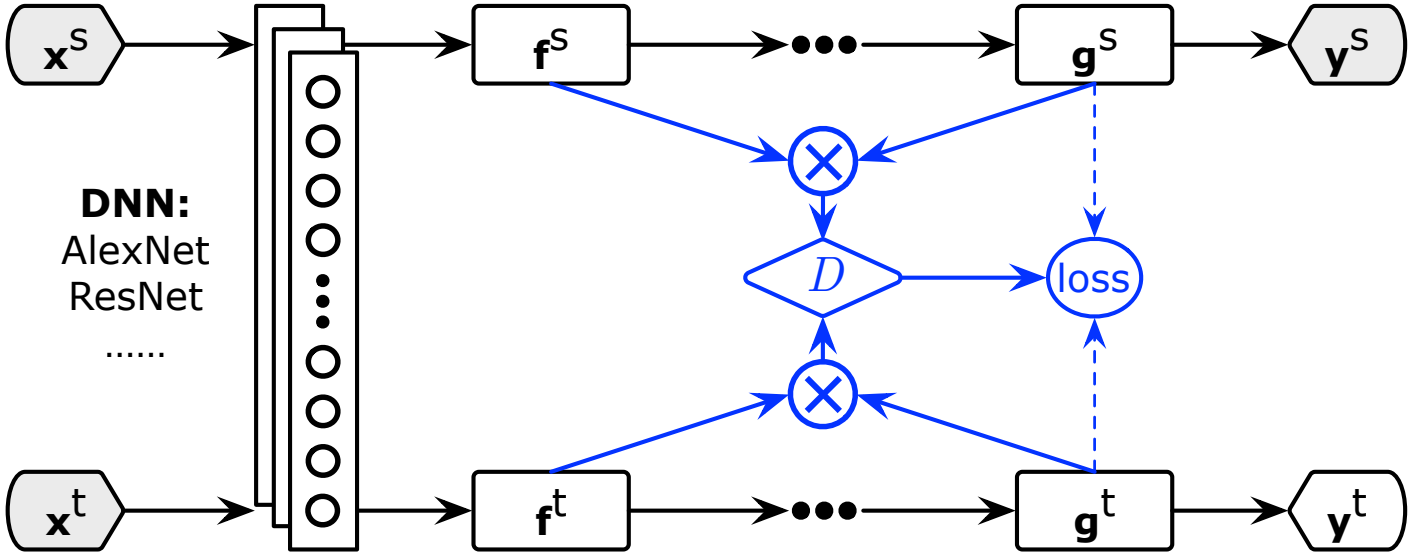
- Disparity Discrepancy (DD) is tighter than $\mathcal{H}\Delta\mathcal{H}$-Divergence.



- DD can be estimated by a joint domain discriminator $D(\mathbf{x}, h(\mathbf{x}))$.

$$
\begin{aligned}
d_{h,\mathcal{H}}(P,Q) &\triangleq \sup_{h' \in \mathcal{H}} \left(\epsilon_P\left(h,h'\right) - \epsilon_Q\left(h,h'\right)\right) \\
&= \sup_{h' \in \mathcal{H}} \left(\mathbb{E}_P\left[|h(\mathbf{x}) - h'(\mathbf{x})| \neq 0\right] - \mathbb{E}_Q\left[|h(\mathbf{x}) - h'(\mathbf{x})| \neq 0\right]\right) \\
&\leqslant \sup_{D \in \mathcal{H}_D} \left(\mathbb{E}_P\left[D(\mathbf{x}, h(\mathbf{x})) = 1\right] + \mathbb{E}_Q\left[D\left(\mathbf{x}, h(\mathbf{x})\right) = 0\right]\right)
\end{aligned} \tag{8}
$$

Conditional Domain Adversarial Network (CDAN)[4]



Conditional adversarial domain adaptation: minimize $d_{h,\mathcal{H}}(\phi(P), \phi(Q))$.

$$\begin{aligned} &\min_G \ \mathcal{E}(G) - \lambda \mathcal{E}(D, G) \\ &\min_D \ \mathcal{E}(D, G), \end{aligned} \tag{9}$$

$$\mathcal{E}(D, G) = -\mathbb{E}_{\mathbf{x}_i^s \sim \mathcal{D}_s} \log \left[D \left(\mathbf{f}_i^s \otimes \mathbf{g}_i^s \right) \right] - \mathbb{E}_{\mathbf{x}_j^t \sim \mathcal{D}_t} \log \left[1 - D \left(\mathbf{f}_j^t \otimes \mathbf{g}_j^t \right) \right] \tag{10}$$

[4] *Long et al. Conditional Adversarial Domain Adaptation. NIPS 2018.*

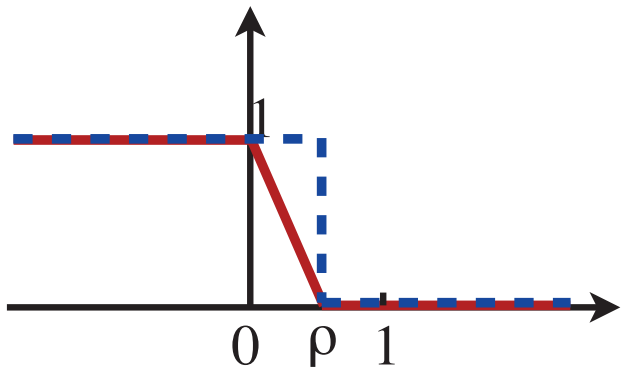
Multiclass Classification Formulation

- Scoring function: $f \in \mathcal{F} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$
- Labeling function induced by f: $h_f : \mathbf{x} \mapsto \arg\max_{y \in \mathcal{Y}} f(\mathbf{x}, y)$
- Labeling function class: $\mathcal{H} = \{h_f | f \in \mathcal{F}\}$
- Margin of a hypothesis f:

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

- Margin Loss:

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



Complexity Measurement

Definition (Rademacher Complexity)

The empirical Rademacher complexity of function class $\mathcal{G}$ w.r.t. sample $\widehat{D}$ is defined as

$$\widehat{\mathfrak{R}}_{\widehat{D}}(\mathcal{G}) = \mathbb{E}_{\sigma} \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i), \tag{11}$$

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

$$\mathfrak{R}_{n,D}(\mathcal{G}) = \mathbb{E}_{\widehat{D} \sim D^n} \widehat{\mathfrak{R}}_{\widehat{D}}(\mathcal{G}). \tag{12}$$

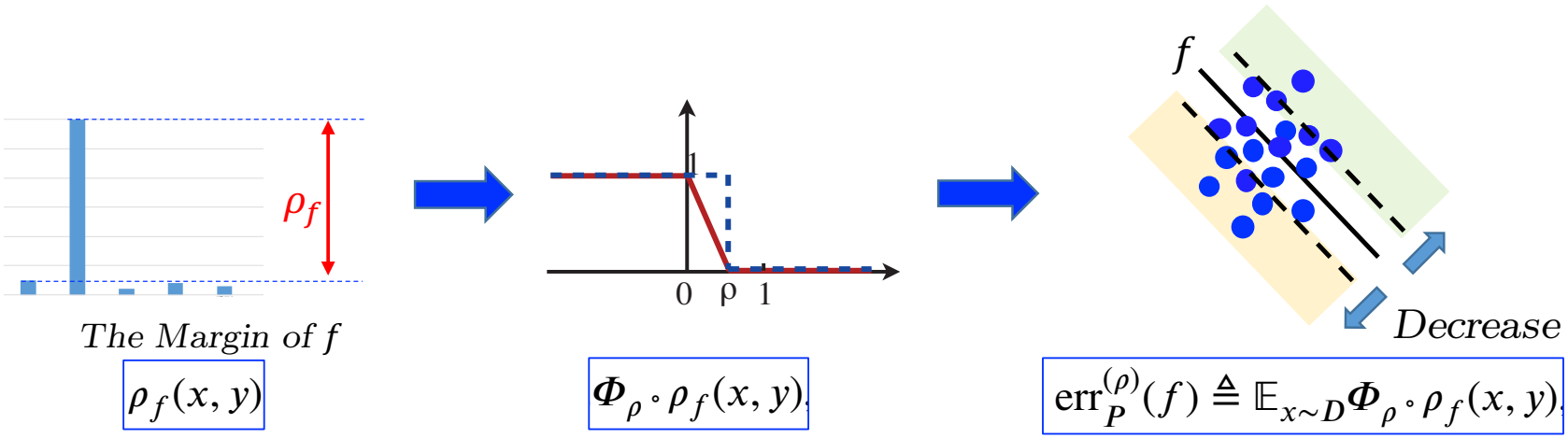
Definition (Covering Number)

(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 $g : \mathcal{X} \to \mathbb{R}$.

These complexity measures can be viewed as extensions of VC-dimensions.

Margin Theory

- Margin error: $\epsilon_D^{(\rho)}(f) = \mathbb{E}_{(\mathbf{x},y)\sim D}\left[\Phi_\rho(\rho_f(\mathbf{x},y))\right]$
- This error takes the margin of the hypothesis f into consideration:



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

$$\Pi_1\mathcal{F} = \{\mathbf{x} \mapsto f(\mathbf{x},y) \big| y \in \mathcal{Y}, f \in \mathcal{F}\}. \tag{13}$$

- Margin Bound for IID setup (generalization error controlled by ρ):

$$\mathrm{err}_P^{(\rho)}(f) \leqslant \mathrm{err}_{\widehat{P}}^{(\rho)}(f) + \frac{2k^2}{\rho}\mathfrak{R}_{n,P}(\Pi_1\mathcal{F}) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}} \tag{14}$$

Step II: Margin Disparity Discrepancy (MDD)[5]

- Margin Disparity: $\epsilon_D^{(\rho)}(f', f) \triangleq \mathbb{E}_{\mathbf{x}\sim D_X}[\Phi_\rho(\rho_{f'}(\mathbf{x}, h_f(\mathbf{x})))]$.
- We further define the margin version of Disparity Discrepancy (DD):

Definition (Margin Disparity Discrepancy (MDD))

Given a hypothesis space $\mathcal{F}$ and a *specific hypothesis* $f \in \mathcal{F}$, the Margin Disparity Discrepancy (MDD) induced by $f' \in \mathcal{F}$ and its empirical version are defined by

$$\begin{aligned} d_{f,\mathcal{F}}^{(\rho)}(P, Q) &\triangleq \sup_{f'\in\mathcal{F}} \left(\epsilon_Q^{(\rho)}(f', f) - \epsilon_P^{(\rho)}(f', f)\right), \\ d_{f,\mathcal{F}}^{(\rho)}(\widehat{P}, \widehat{Q}) &\triangleq \sup_{f'\in\mathcal{F}} \left(\epsilon_{\widehat{Q}}^{(\rho)}(f', f) - \epsilon_{\widehat{P}}^{(\rho)}(f', f)\right). \end{aligned} \tag{15}$$

MDD satisfies $d_{f,\mathcal{F}}^{(\rho)}(P, P) = 0$ as well as nonnegativity and subadditivity.

[5] *Zhang & Long. Bridging Theory and Algorithm for Domain Adaptation. ICML 2019.*

Bound with Margin Disparity Discrepancy

Theorem (Bound with MDD)

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

$$\epsilon_Q(f) \leq \epsilon_P^{(\rho)}(f) + d_{f,\mathcal{F}}^{(\rho)}(P, Q) + \epsilon_{ideal}^{(\rho)}, \tag{16}$$

and $\epsilon_{ideal}^{(\rho)}$ is the ideal joint margin error: $\epsilon_{ideal}^{(\rho)} = \min_{f^* \in \mathcal{F}} \{\epsilon_P^{(\rho)}(f^*) + \epsilon_Q^{(\rho)}(f^*)\}$.

- This upper-bound has a similar form with the classic bound:
  - $\epsilon_P^{(\rho)}(f)$ measures the performance of f on the source domain.
  - MDD bounds the performance gap caused by the domain shift.
  - $\epsilon_{ideal}^{(\rho)}$ is the margin version of the ideal joint error.
- **A new tool for analyzing transfer learning with margin theory.**

Definitions for Generalization Bounds

Two function classes for deriving the generalization bounds with MDD:

Definition (Function Class $\Pi_1\mathcal{F}$, for supervised bound)

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

$$\Pi_1\mathcal{F} = \{\mathbf{x} \mapsto f(\mathbf{x}, y) \big| y \in \mathcal{Y}, f \in \mathcal{F}\}. \tag{17}$$

Definition (Function Class $\Pi_{\mathcal{H}}\mathcal{F}$, for Margin Disparity Discrepancy)

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

$$\Pi_{\mathcal{H}}\mathcal{F} \triangleq \{\mathbf{x} \mapsto f(\mathbf{x}, h(\mathbf{x})) | h \in \mathcal{H}, f \in \mathcal{F}\}. \tag{18}$$

Applying margin loss over $f \in \Pi_{\mathcal{H}}\mathcal{F}$ yields the **"scoring" version** of $\mathcal{H}\Delta\mathcal{H}$

Margin Theory for Transfer Learning

Theorem (Generalization Bound with Rademacher Complexity)

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

$$\begin{aligned}\epsilon_Q(f) \leq & \epsilon^{(\rho)}_{\widehat{P}}(f) + d^{(\rho)}_{f,\mathcal{F}}(\widehat{P}, \widehat{Q}) + \epsilon_{ideal} \\ & + \frac{2k^2}{\rho}\mathfrak{R}_{n,P}(\Pi_1\mathcal{F}) + \frac{k}{\rho}\mathfrak{R}_{n,P}(\Pi_{\mathcal{H}}\mathcal{F}) + 2\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \\ & + \frac{k}{\rho}\mathfrak{R}_{m,Q}(\Pi_{\mathcal{H}}\mathcal{F}) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}.\end{aligned} \tag{19}$$

An expected observation is that the generalization risk is controlled by ρ.

Margin Theory for Transfer Learning

Theorem (Generalization Bound with Covering Numbers)

Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be a hypothesis set with label set $\mathcal{Y} = \{1, \cdots, k\}$ and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be the corresponding $\mathcal{Y}$-valued labeling function 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 $f \in \mathcal{F}$:

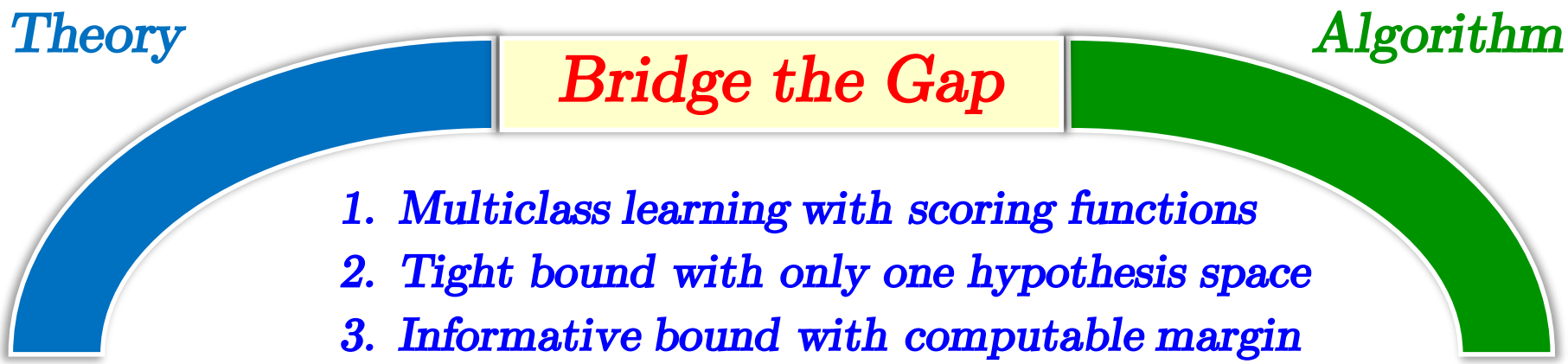
$$\begin{aligned}\epsilon_Q(f) \leq & \epsilon_{\widehat{P}}^{(\rho)}(f) + d_{f,\mathcal{F}}^{(\rho)}(\widehat{P}, \widehat{Q}) + \epsilon_{ideal} + 2\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \\ & + \sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \frac{16k^2\sqrt{k}}{\rho} \inf_{\epsilon\geq 0}\Big\{\epsilon + 3\Big(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}\Big) \\ & \Big(\int_{\epsilon}^{L}\sqrt{\log \mathcal{N}_2(\tau, \Pi_1\mathcal{F})}\mathrm{d}\tau + L\int_{\epsilon/L}^{1}\sqrt{\log \mathcal{N}_2(\tau, \Pi_1\mathcal{H})}\mathrm{d}\tau\Big)\Big\}.\end{aligned} \tag{20}$$

The margin bound for OOD has same order with the margin bound for IID.

Margin Theory Implied Algorithm (MDD)[6]

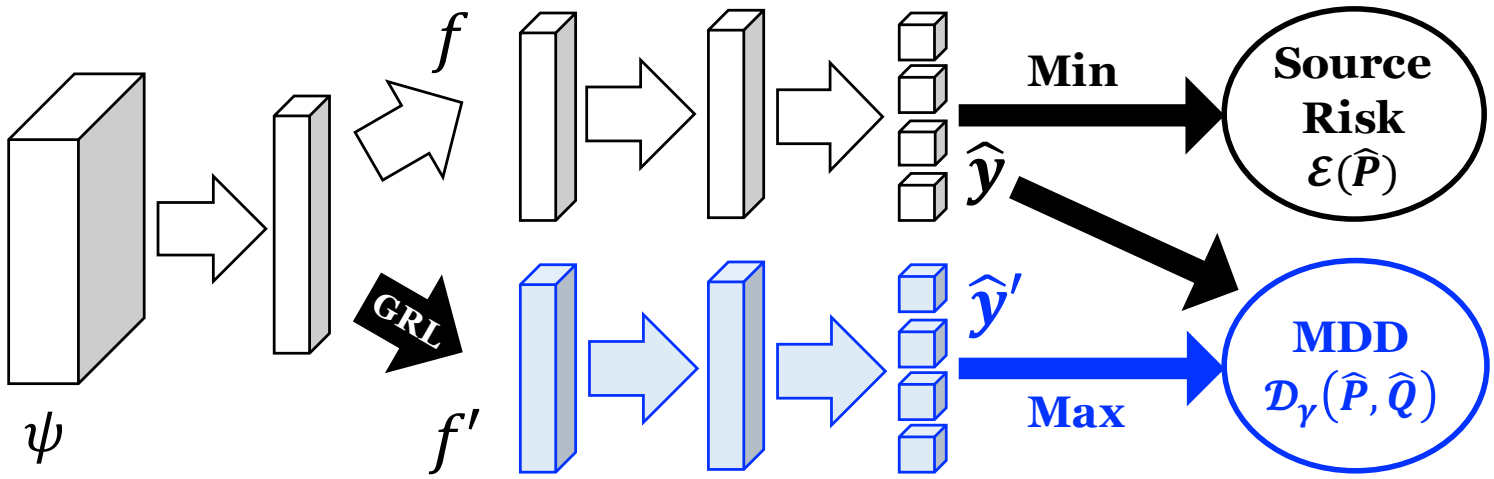
Minimax domain adaptation implied directly through the margin theory

$$\min_{f,\psi} \ \epsilon^{(\rho)}_{\psi(\widehat{P})}(f) + \left(\epsilon^{(\rho)}_{\psi(\widehat{Q})}(f^*, f) - \epsilon^{(\rho)}_{\psi(\widehat{P})}(f^*, f)\right)$$
$$f^* = \max_{f'} \ \left(\epsilon^{(\rho)}_{\psi(\widehat{Q})}(f', f) - \epsilon^{(\rho)}_{\psi(\widehat{P})}(f', f)\right) \tag{21}$$



[6] *Zhang & Long. Bridging Theory and Algorithm for Domain Adaptation. ICML 2019.*

Margin Theory Implied Algorithm (MDD)



$$\mathcal{E}(\widehat{P}) = -\mathbb{E}_{(\mathbf{x}^s, y^s)\sim\widehat{P}} \log[\sigma_{y^s}(f(\psi(\mathbf{x}^s)))]$$
$$\mathcal{D}_\gamma(\widehat{P}, \widehat{Q}) = \mathbb{E}_{\mathbf{x}^t\sim\widehat{Q}} \log[1 - \sigma_{h_f(\psi(\mathbf{x}^t))}(f'(\psi(\mathbf{x}^t)))] \tag{22}$$
$$+ \gamma\mathbb{E}_{\mathbf{x}^s\sim\widehat{P}} \log[\sigma_{h_f(\psi(\mathbf{x}^s))}(f'(\psi(\mathbf{x}^s)))]$$

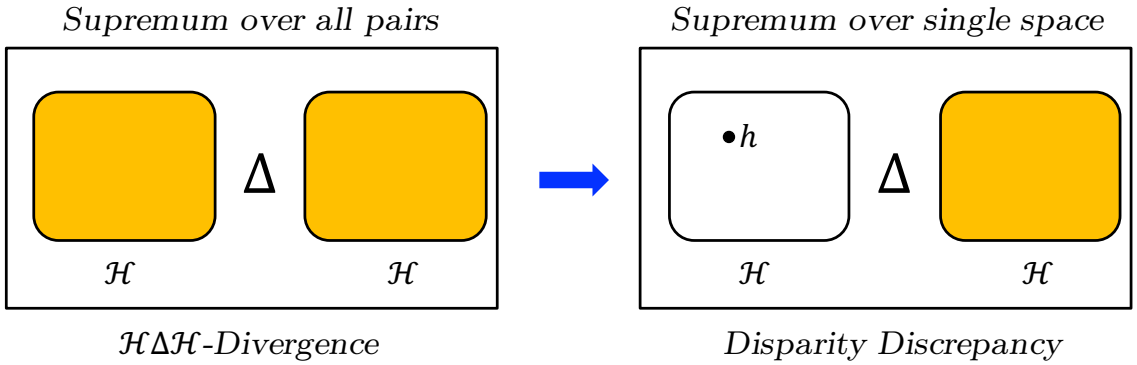
Theorem (Margin Implementation)

(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 $\rho = \log\gamma$.

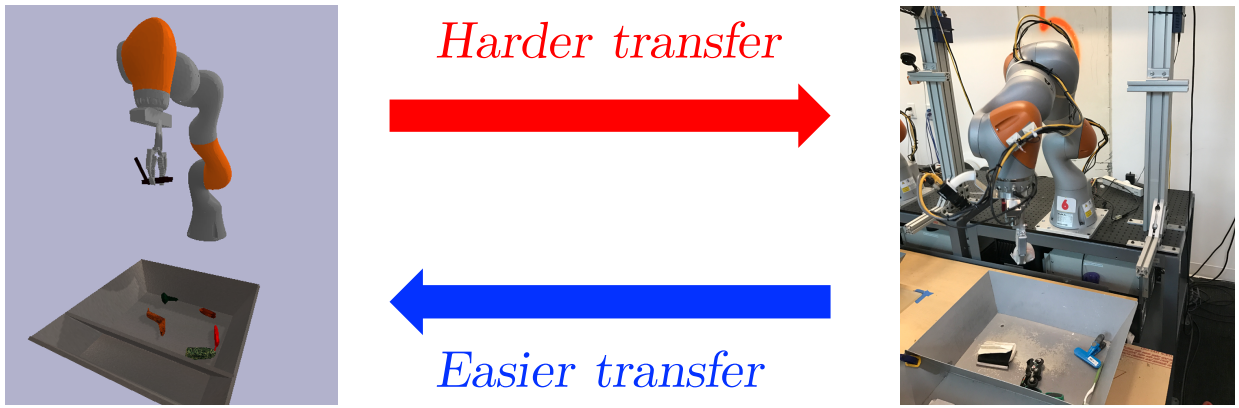
Outline

Theory and Practice: Final Gap to Close

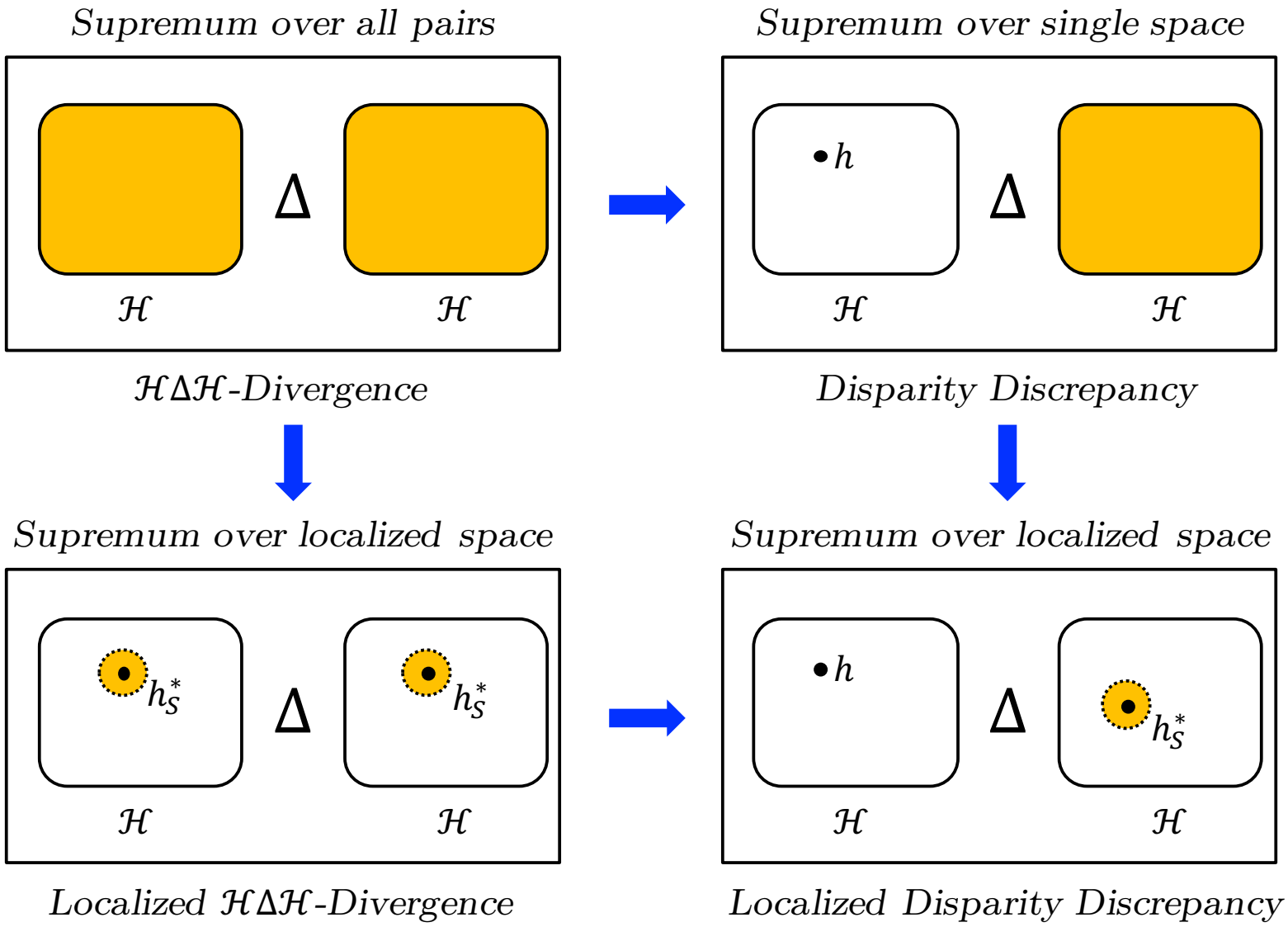
- Previous discrepancies are supremum over the whole hypothesis space — will include bad hypotheses that make the bound excessively large.



- A common observation is that the difficulty of transfer is asymmetric — Previous bounds will remain unchanged after switching P and Q.



Localization for Discrepancies



Step III: Localized Discrepancies

Definition (Localized Hypothesis Space)

For any distributions P and Q on $\mathcal{X} \times \mathcal{Y}$, any hypothesis space $\mathcal{H}$ and any $r \geq 0$, the **localized hypothesis space** $\mathcal{H}_r$ is defined as

$$\mathcal{H}_r = \{h \in \mathcal{H} | \mathbb{E}_P L(h(\mathbf{x}), y) \leq r\}. \tag{23}$$

Definition (Localized $\mathcal{H}\Delta\mathcal{H}$-Discrepancy (LHH))

Based on $\mathcal{H}_r$, the **localized $\mathcal{H}\Delta\mathcal{H}$-discrepancy** from P to Q is defined as

$$d_{\mathcal{H}_r\Delta\mathcal{H}_r}(P, Q) = \sup_{h, h' \in \mathcal{H}_r} \left(\mathbb{E}_Q L(h', h) - \mathbb{E}_P L(h', h)\right). \tag{24}$$

Definition (Localized Disparity Discrepancy (LDD))

Based on $\mathcal{H}_r$, for any $h \in \mathcal{H}$, the **localized disparity discrepancy** from P to Q is

$$d_{h,\mathcal{H}_r}(P, Q) = \sup_{h' \in \mathcal{H}_r} \left(\mathbb{E}_Q L(h', h) - \mathbb{E}_P L(h', h)\right). \tag{25}$$

Localization Theory for Transfer Learning[7]

Recall the generalization bound induced by previous discrepancies:

$$\epsilon_Q(h) \leq \epsilon_{\widehat{P}}(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{P}, \widehat{Q}) + \epsilon_{ideal} + O(\sqrt{\frac{d \log n}{n}} + \sqrt{\frac{d \log m}{m}})$$

Theorem (Generalization Bound with Localized $\mathcal{H}\Delta\mathcal{H}$-Discrepancy)

Set fixed $r > \lambda$. Let $\hat{h}$ be the solution of the source error minimization. Then with probability no less than $1 - \delta$, we have

$$\begin{aligned} \mathrm{err}_Q(\hat{h}) \leq \mathrm{err}_{\widehat{P}}(\hat{h}) + d_{\mathcal{H}_r\Delta\mathcal{H}_r}(\widehat{P}, \widehat{Q}) + \lambda + O(\frac{d \log n}{n} + \frac{d \log m}{m}) \\ + O\left(\sqrt{\frac{2rd \log n}{n}} + \sqrt{\frac{(d_{\mathcal{H}_r\Delta\mathcal{H}_r}(\widehat{P}, \widehat{Q}) + 2r) d \log m}{m}}\right). \end{aligned} \tag{26}$$

To make domain adaptation feasible, we require $d_{\mathcal{H}_r\Delta\mathcal{H}_r}(\widehat{P}, \widehat{Q}) + r \ll 1$.

[7] *Zhang & Long. On Localized Discrepancy for Domain Adaptation. Preprint 2020.*

Localization Theory for Transfer Learning[8]

Recall that Disparity Discrepancy is tighter than $\mathcal{H}\Delta\mathcal{H}$-Discrepancy:

$$\min_{\bar{h}\in\mathcal{H}}\{\mathrm{err}_{\widehat{P}}(\bar{h}) + d_{\bar{h},\mathcal{H}_r}(\widehat{P},\widehat{Q})\} \leq \min_{\hat{h}\in\mathcal{H}} \mathrm{err}_{\widehat{P}}(\hat{h}) + d_{\mathcal{H}_r\Delta\mathcal{H}_r}(\widehat{P},\widehat{Q}) \tag{27}$$

Theorem (Generalization bound with localized disparity discrepancy)

Set fixed $r > \lambda$. Let $\bar{h}$ be the solution of above left objective function. Then with probability no less than $1-\delta$, we have

$$\mathrm{err}_Q(\hat{h}) \leq \mathrm{err}_{\widehat{P}}(\bar{h}) + d_{\bar{h},\mathcal{H}_r}(\widehat{P},\widehat{Q}) + \lambda + O(\frac{d\log n}{n} + \frac{d\log m}{m})$$
$$+ O\left(\sqrt{\frac{(\mathrm{err}_{\widehat{P}}(\bar{h}) + r)d\log n}{n}} + \sqrt{\frac{(\mathrm{err}_{\widehat{P}}(\bar{h}) + d_{\bar{h},\mathcal{H}_r}(\widehat{P},\widehat{Q}) + r)d\log m}{m}}\right). \tag{28}$$

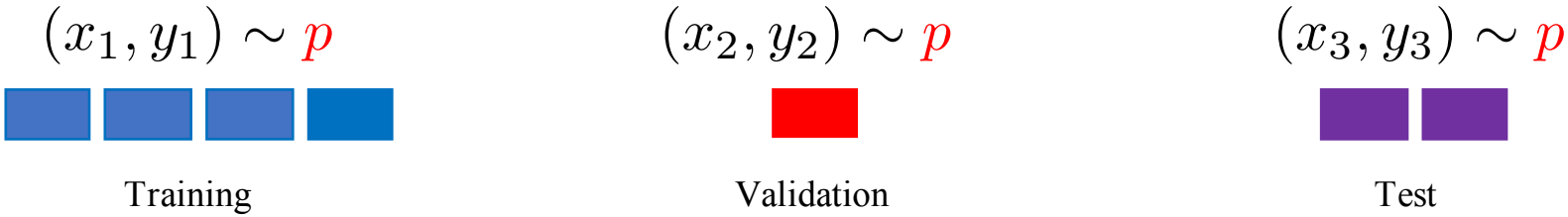
[8] *Zhang & Long. On Localized Discrepancy for Domain Adaptation. Preprint 2020.*

Outline

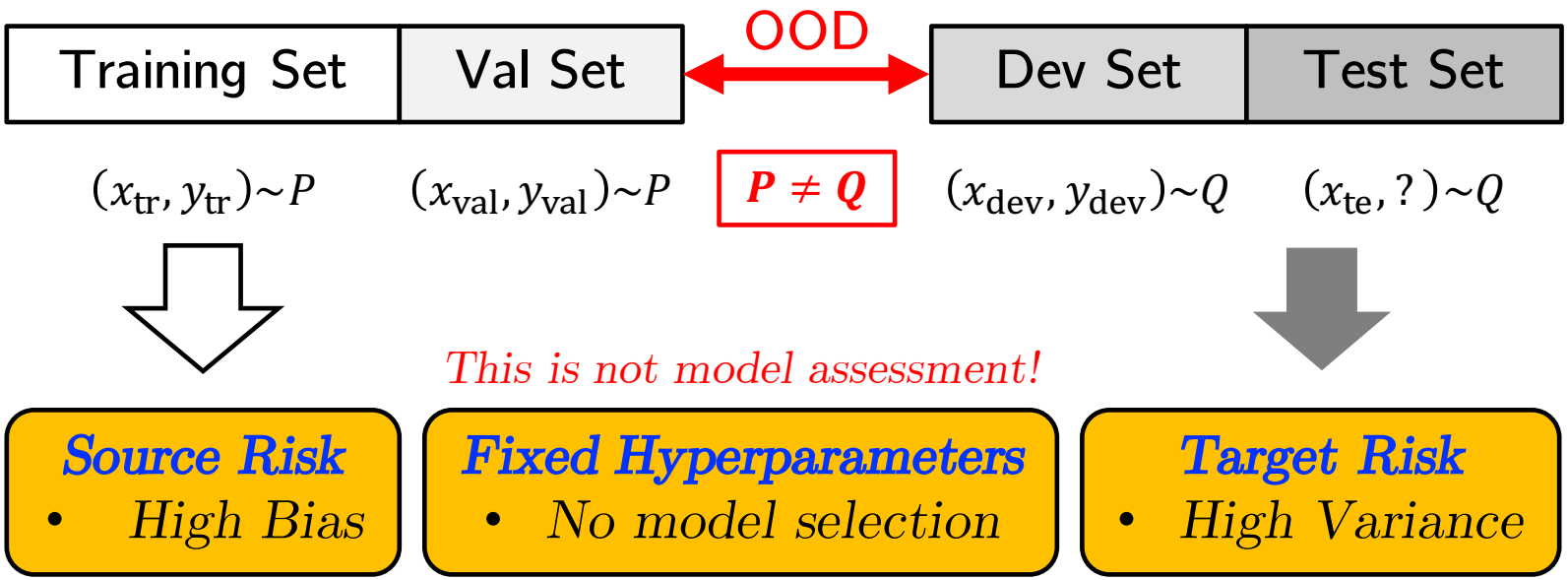
Outline

Model Selection in Transfer Learning

- Supervised Learning



- Bias-Variance-Shift Dilemma of model selection in Transfer Learning



Importance-Weighted Cross-Validation (IWCV)[9]

- Covariate shift assumption: $P(y|\mathbf{x}) = Q(y|\mathbf{x})$
- Model selection by estimating Target Risk $\epsilon_Q(h) = \mathbb{E}_Q[h(\mathbf{x}) \neq y]$
- Importance-Weighted Cross-Validation (IWCV)

$$\mathbb{E}_P w(\mathbf{x}) \cdot [h(\mathbf{x}) \neq y] = \mathbb{E}_P \frac{Q(\mathbf{x})}{P(\mathbf{x})} \cdot [h(\mathbf{x}) \neq y] = \mathbb{E}_Q[h(\mathbf{x}) \neq y] = \epsilon_Q(h) \quad (29)$$

- **Q1**: How to estimate density ratio $w(\mathbf{x})$ given unknown P and Q
- Density ratio $w(\mathbf{x})$ is estimated by discriminative learning (LogReg)

$$\begin{aligned} w(\mathbf{x}) &= \frac{Q(\mathbf{x})}{P(\mathbf{x})} = \frac{J_{\boldsymbol{f}}(\mathbf{x}|d=0)}{J_{\boldsymbol{f}}(\mathbf{x}|d=1)} \\ &= \frac{J_{\boldsymbol{f}}(d=1)}{J_{\boldsymbol{f}}(d=0)} \frac{J_{\boldsymbol{f}}(\mathbf{x}) J_{\boldsymbol{f}}(d=0|\mathbf{x})}{J_{\boldsymbol{f}}(\mathbf{x}) J_{\boldsymbol{f}}(d=1|\mathbf{x})} \\ &= \frac{J_{\boldsymbol{f}}(d=1)}{J_{\boldsymbol{f}}(d=0)} \frac{J_{\boldsymbol{f}}(d=0|\mathbf{x})}{J_{\boldsymbol{f}}(d=1|\mathbf{x})} \approx \frac{n_s}{n_t} \frac{J_{\boldsymbol{f}}(d=0|\mathbf{x})}{J_{\boldsymbol{f}}(d=1|\mathbf{x})} \end{aligned} \quad (30)$$

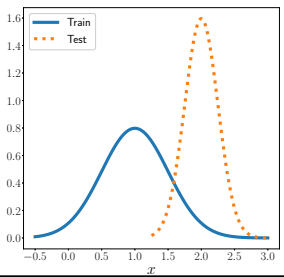
[9] *Sugiyama et al. Covariate Shift Adaptation by Importance Weighted Cross Validation. JMLR 2007.*

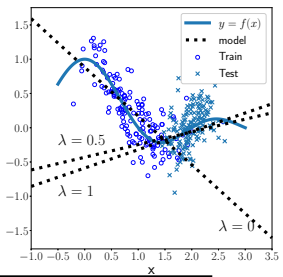
Transferable Validation (TransVal)[10]

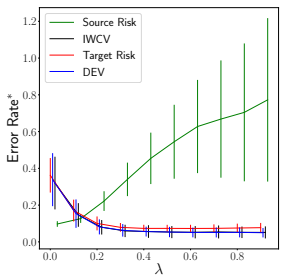
- **Q2**: How to reduce the variance when estimating Target Risk $\epsilon_Q(h)$?
- Variance of IWCV can be bounded by the Rényi divergence:

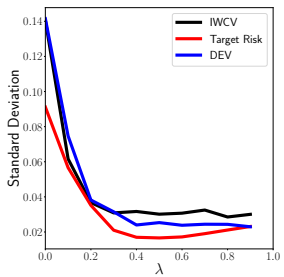
$$\mathrm{Var}_{\mathbf{x}\sim P}[w(\mathbf{x})\cdot[h(\mathbf{x})\neq y]] \leq d_{\alpha+1}(Q\|P)\epsilon_Q(h)^{1-\frac{1}{\alpha}} - \epsilon_Q(h)^2 \quad (31)$$

- Feature matching reduces the distribution discrepancy $d_{\alpha+1}(Q\|P)$
- Control variate reduces the variance of estimating $\mathbb{E}_P w(\mathbf{x})\,[h(\mathbf{x})\neq y]$
  1. Given two unbiased estimators: $\mathbb{E}[z]=\zeta, \mathbb{E}[t]=\tau$
  2. Construct a new estimator: $z^\star = z+\eta(t-\tau)$
  3. $z^\star$ is still unbiased: $\mathbb{E}[z^\star]=\mathbb{E}[z]+\eta\mathbb{E}[t-\tau]=\zeta+\eta(\mathbb{E}[t]-\mathbb{E}[\tau])=\zeta$
  4. $\mathrm{Var}[z^\star]=\mathrm{Var}[z+\eta(t-\tau)]=\eta^2\mathrm{Var}[t]+2\eta\mathrm{Cov}(z,t)+\mathrm{Var}[z]$
  5. $\min \mathrm{Var}[z^\star]=(1-\rho_{z,t}^2)\mathrm{Var}[z]$, when $\hat{\eta}=-\frac{\mathrm{Cov}(z,t)}{\mathrm{Var}[t]}$
  6. Since $0\leq\rho_{z,t}^2\leq 1$, $\mathrm{Var}[z^\star]\leq\mathrm{Var}[z]$, the variance is reduced explicitly









[10] *You & Long. Towards Accurate Model Selection in Deep Unsupervised Domain Adaptation. ICML 2019.*

Outline

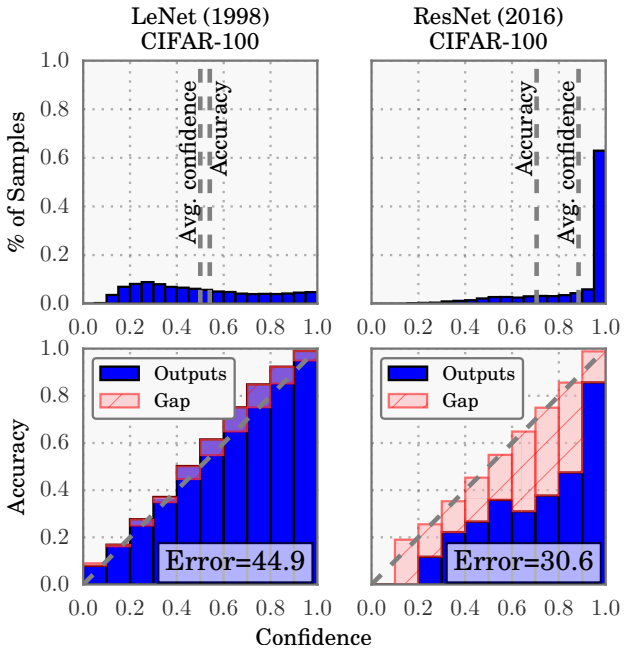
Confidence Calibration in Deep Learning[11]

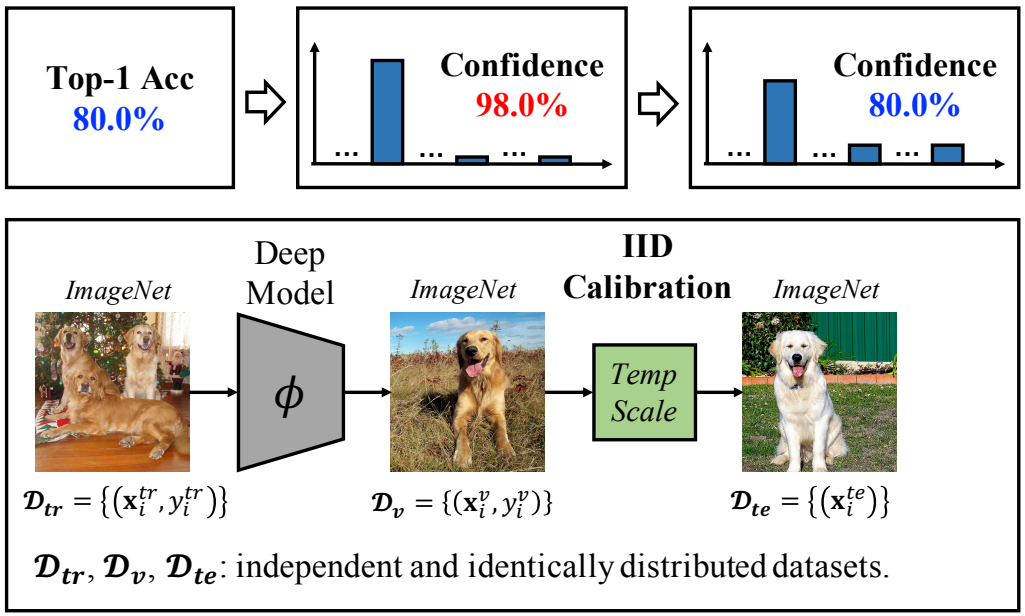
- A model should output a probability reflecting the true frequency:

$$\mathbb{P}(\widehat{Y} = Y | \widehat{P} = c) = c, \ \forall \ c \in [0, 1] \tag{32}$$

where $\widehat{Y}$ is the class prediction and $\widehat{P}$ is its associated confidence.

- Deep networks learn high accuracy at the expense of over-confidence.





[11] *Guo et al. On Calibration of Modern Neural Networks. ICML 2017.*

Temperature Scaling for IID Calibration

- **Calibration Metric**: Expected Calibration Error (ECE)

$$\mathcal{L}_{\mathrm{ECE}} = \sum_{m=1}^{B} \frac{|B_m|}{n} |\mathbb{A}(B_m) - \mathbb{C}(B_m)|$$

$$\mathbb{A}(B_m) = |B_m|^{-1} \sum_{i \in B_m} \mathbf{1}(\widehat{\mathbf{y}}_i = \mathbf{y}_i) \quad \textbf{(Accuracy)} \tag{33}$$

$$\mathbb{C}(B_m) = |B_m|^{-1} \sum_{i \in B_m} p(\widehat{\mathbf{p}}_i | \mathbf{x}_i, \theta) \quad \textbf{(Confidence)}$$

- **IID Calibration**: Temperature Scaling

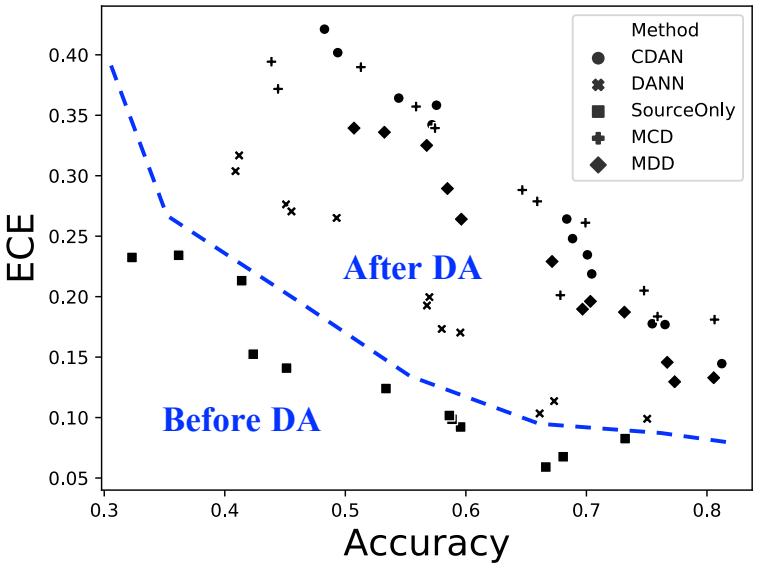
$$T^* = \arg\min_{T} \mathrm{E}_{(\mathbf{x}_v, \mathbf{y}_v) \in \mathcal{D}_v} \mathcal{L}_{\mathrm{NLL}} \left(\sigma(\mathbf{z}_v / T), \mathbf{y}_v \right) \tag{34}$$

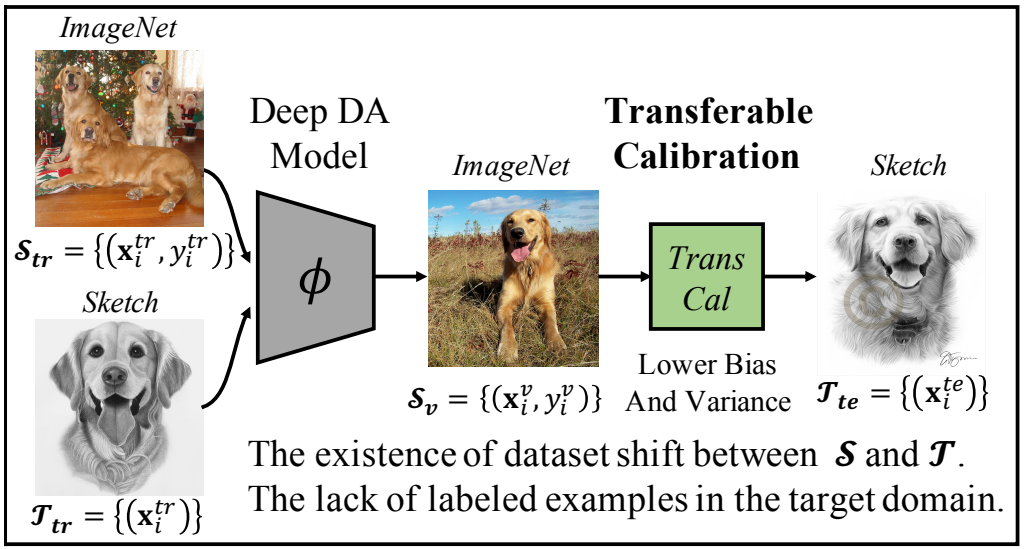
  σ is the *softmax* function, $\mathcal{L}_{\mathrm{NLL}}$ is Negative Log-Likelihood.

- Transform logits $\mathbf{z}_{te}$ into calibrated probabilities $p_{te} = \sigma(\mathbf{z}_{te} / T^*)$.

Dilemma of Accuracy vs Confidence in OOD Setup[12]

- Transfer models yield high accuracy at the expense of over-confidence.





- Calibration in transfer learning is challenging due to the coexistence:
  - Domain shift — ECE should be unbiased to the target domain
  - Unlabeled target — ECE on the target domain is incomputable
- Bias-Variance-Shift Dilemma of conf. calibration in Transfer Learning

[12] *Wang & Long. Transferable Calibration with Lower Bias and Variance in Domain Adaptation. Preprint 2020.*

Transferable Calibration: Bias Reduction

- Importance-weighting for an unbiased estimate of target ECE
- The **bias** between the estimated ECE and the ground-truth ECE

$$\begin{aligned}&\left|\mathrm{E}^*_{\mathbf{x}\sim q}\left[\mathcal{L}_{\mathrm{ECE}}^{\widehat{w}(\mathbf{x})}\right]-\mathrm{E}_{\mathbf{x}\sim q}\left[\mathcal{L}_{\mathrm{ECE}}^{w(\mathbf{x})}\right]\right| \\ =&\left|\mathrm{E}_{\mathbf{x}\sim p}\left[\widehat{w}(\mathbf{x})\mathcal{L}_{\mathrm{ECE}}(\phi(\mathbf{x}),\ \mathbf{y})\right]-\mathrm{E}_{\mathbf{x}\sim p}\left[w(\mathbf{x})\mathcal{L}_{\mathrm{ECE}}(\phi(\mathbf{x}),\ \mathbf{y})\right]\right| \\ =&\left|\mathrm{E}_{\mathbf{x}\sim p}\left[(w(\mathbf{x})-\widehat{w}(\mathbf{x}))\mathcal{L}_{\mathrm{ECE}}(\phi(\mathbf{x}),\ \mathbf{y})\right]\right|\end{aligned} \tag{35}$$

- The discrepancy between $\widehat{w}(\mathbf{x})$ and $w(\mathbf{x})$ can be bounded by

$$\mathbb{E}_{\mathbf{x}\sim p}\left[(w(\mathbf{x})-\widehat{w}(\mathbf{x}))^2\right]=\mathbb{E}_{\mathbf{x}\sim p}\left[\left(\frac{p(\mathbf{x})-\widehat{p}(\mathbf{x})}{p(\mathbf{x})\widehat{p}(\mathbf{x})}\right)^2\right]\leq(M+1)^4\mathbb{E}_{\mathbf{x}\sim p}\left[(p(\mathbf{x})-\widehat{p}(\mathbf{x}))^2\right] \tag{36}$$

- Use λ $(0\leq\lambda\leq 1)$ to control the bound M of the importance weights

$$\begin{aligned}(T^*,\lambda)&=\arg\min_{T,\lambda}\mathbb{E}_{\mathbf{x}\sim p}\left[\widetilde{w}(\mathbf{x})\mathcal{L}_{\mathrm{ECE}}(\phi(\mathbf{x}),\ \mathbf{y})\right] \\ \widetilde{w}(\mathbf{x}_i)&=\frac{\left(\widehat{w}(\mathbf{x}_i)\right)^{\lambda}}{\sum_{i=1}^{n_s}\left(\widehat{w}(\mathbf{x}_i)\right)^{\lambda}}\end{aligned} \tag{37}$$

Transferable Calibration: Variance Reduction

- **Serial Control Variate:** $\mathrm{Var}[u^{**}] \leq \mathrm{Var}[u^{*}] \leq \mathrm{Var}[u]$

$$\begin{aligned} u^{*} &= u + \eta_1(t_1 - \tau_1) \\ u^{**} &= u^{*} + \eta_2(t_2 - \tau_2) \end{aligned} \tag{38}$$

- First, use importance weight $\widetilde{w}(\mathbf{x}_s)$ as a control covariate

$$\mathbf{E}_q^{*} = \widetilde{\mathbf{E}}_q - \frac{1}{n_s} \frac{\mathrm{Cov}(\mathcal{L}_{\mathrm{ECE}}^{\widetilde{w}}, \widetilde{w}(\mathbf{x}))}{\mathrm{Var}[\widetilde{w}(\mathbf{x})]} \sum_{i=1}^{n_s} [\widetilde{w}(\mathbf{x}_s^i) - 1] \tag{39}$$

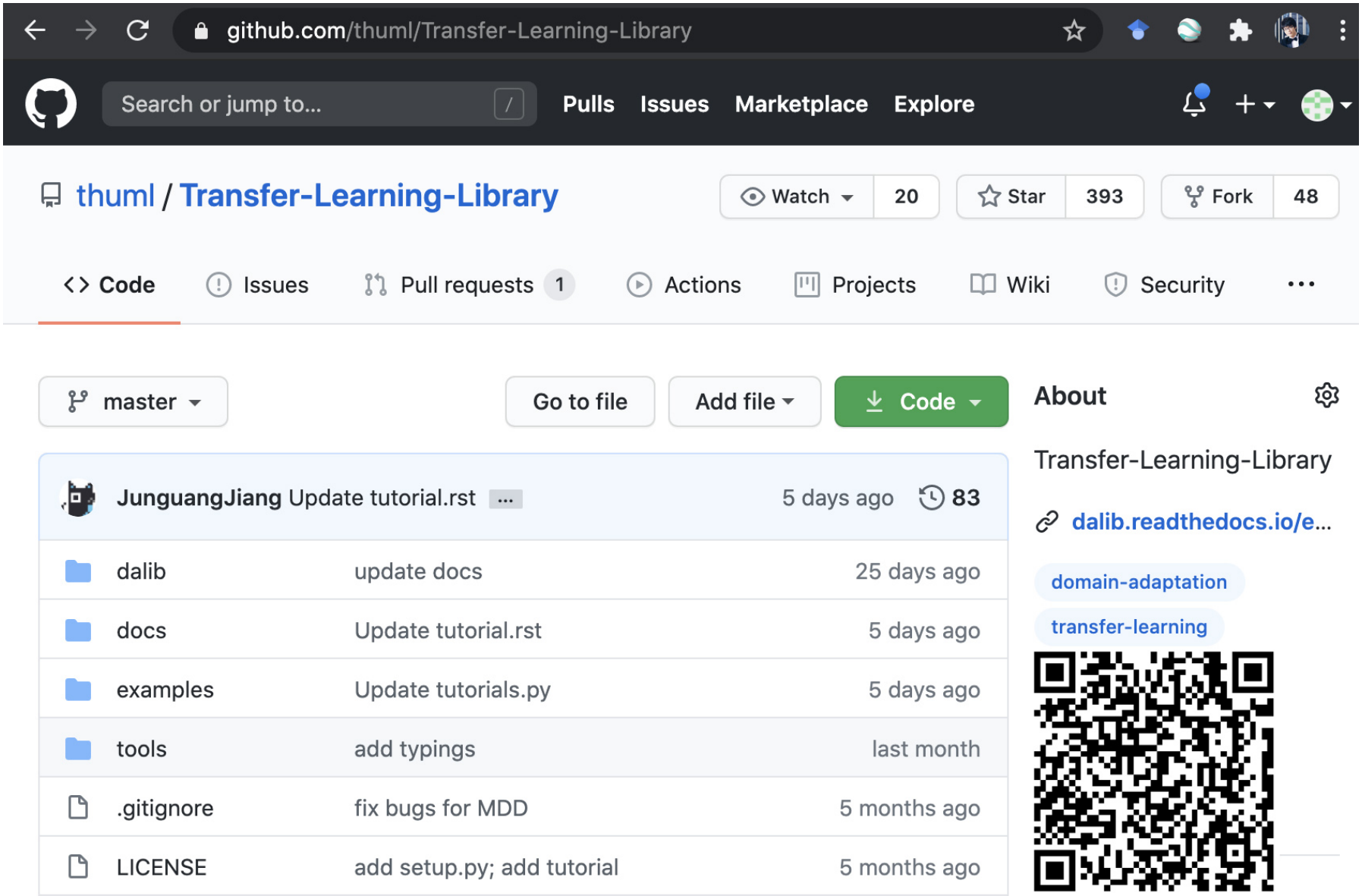
- Second, use the source prediction $r(\mathbf{x}_s)$ as another control variate

$$\mathbf{E}_q^{**} = \mathbf{E}_q^{*} - \frac{1}{n_s} \frac{\mathrm{Cov}(\mathcal{L}_{\mathrm{ECE}}^{\widetilde{w}*}, r(\mathbf{x}))}{\mathrm{Var}[r(\mathbf{x})]} \sum_{i=1}^{n_s} [r(\mathbf{x}_s^i) - c] \tag{40}$$

- Reduce bias, variance, and shift all-in-one for Transferable Calibration

Outline

Transfer Learning Library



Transfer Learning Library: Design Patterns

Reproducible | Stable | Extendible | Ease of Use | TorchVision | Documentation

Docs

- **Examples**
  - ☐ Training codes
  - ☐ Hyperparameters
  - ☐ .....
- **Benchmarks**
  - ☐ Various setups
  - ☐ Reproducible
  - ☐ .....
- **Tutorials**
  - ☐ More data formats
  - ☐ More model backbones
  - ☐ .....

Core

- **Adaptation**
  - ☐ DAN
  - ☐ DANN
  - ☐ MDD
  - ☐ CDAN
  - ☐ .....
- **Module**
  - ☐ Discriminator
  - ☐ GradRevLayer
  - ☐ Kernel
  - ☐ .....
- **Backbone**
  - ☐ ResNet
  - ☐ VGG
  - ☐ Inception
  - ☐ .....
- **Dataset**
  - ☐ Office-31
  - ☐ Office-Home
  - ☐ VisDA-2017
  - ☐ DomainNet
  - ☐ .....
- **Utils**

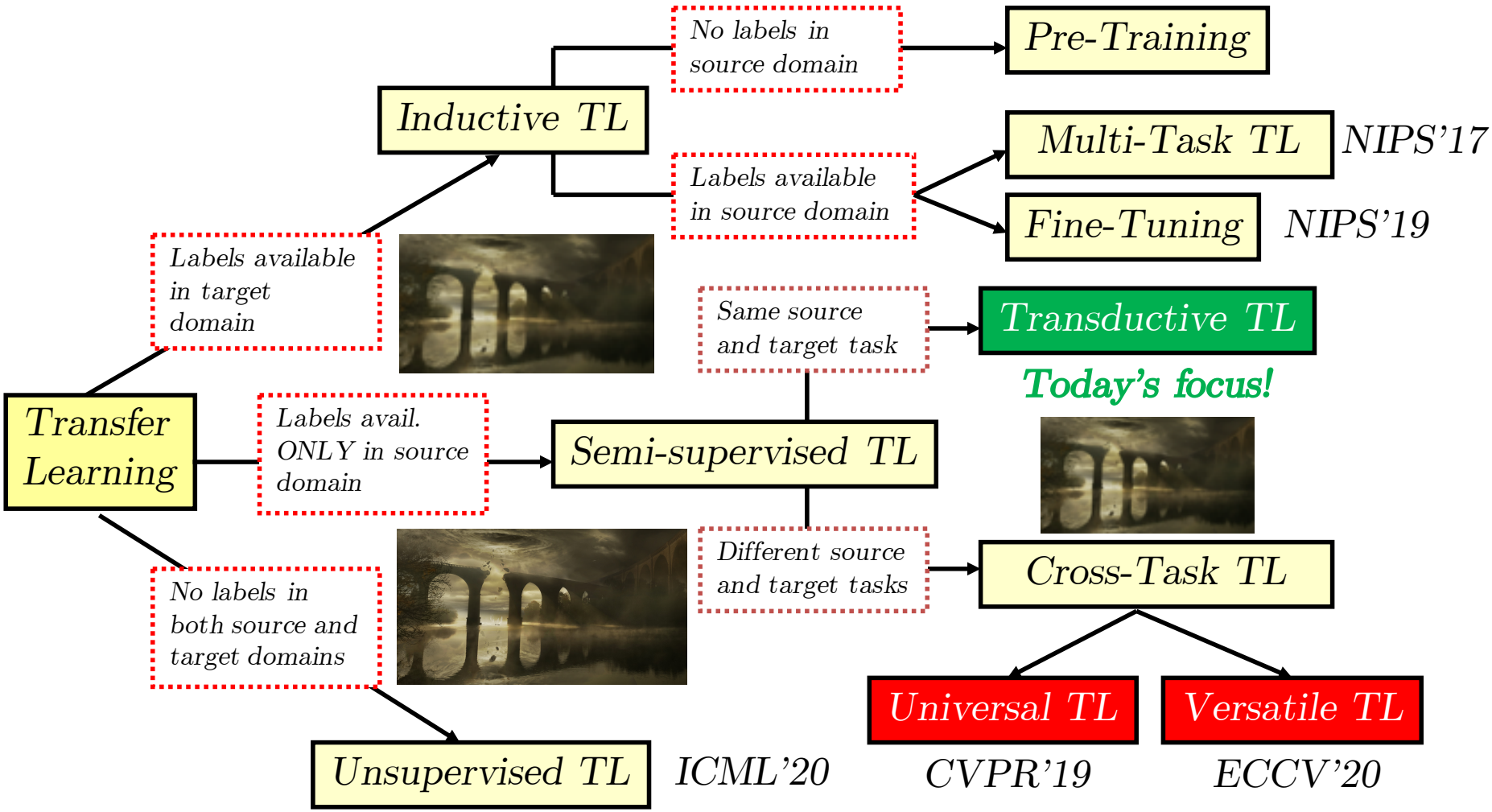
Platform

PyTorch TorchVision Facebook Open Source

.....

Github: `https://github.com/thuml/Transfer-Learning-Library`

Transfer Learning Library: Ongoing Implementations



Final note: Most transfer learning setups are still open for future research!

Transfer Learning Library: Reproducible Benchmarks

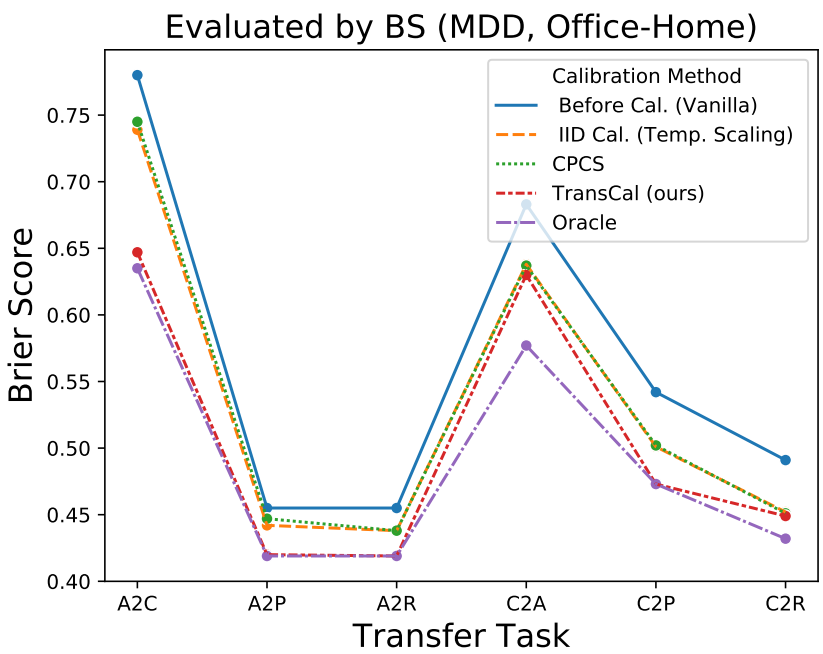
Table: Accuracy (%) on *Office-31* for Unsupervised Domain Adaptation

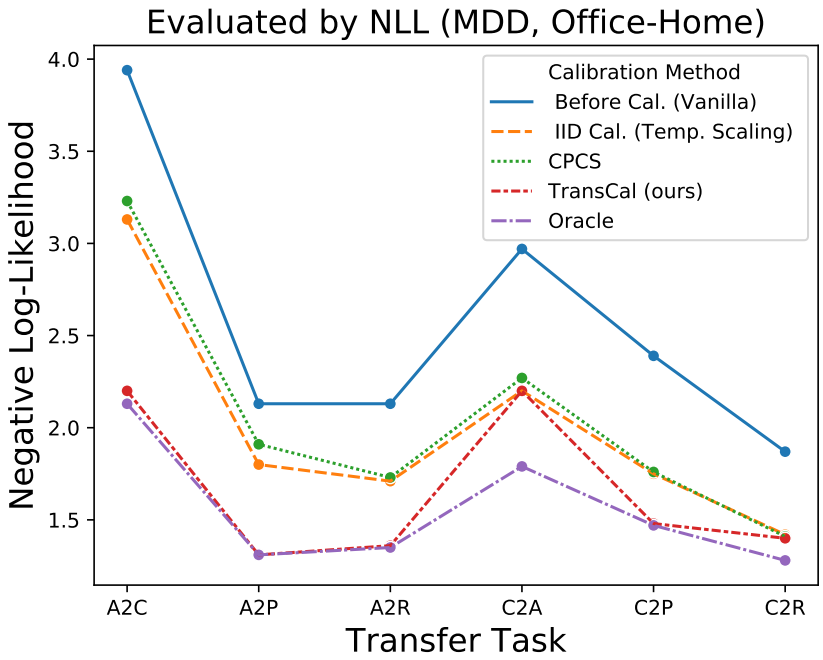
Method	Origin	Ours	Δacc	$A \rightarrow W$	$D \rightarrow W$	$W \rightarrow D$	$A \rightarrow D$	$D \rightarrow A$	$W \rightarrow A$
ResNet-50	76.1	**79.5**	3.4	75.8	95.5	99.0	79.3	63.6	63.8
DANN	82.2	**86.4**	4.2	91.7	97.9	100.0	82.9	72.8	73.3
DAN	80.4	**83.7**	3.3	84.2	98.4	100.0	87.3	66.9	65.2
JAN	84.3	**87.3**	3.0	93.7	98.4	100.0	89.4	71.2	71.0
CDAN	87.7	**88.7**	1.0	93.1	98.6	100.0	93.4	75.6	71.5
MCD	-	**85.9**	-	91.8	98.6	100.0	89.0	69.0	66.9
MDD	88.9	**89.2**	0.3	93.6	98.6	100.0	93.6	76.7	72.9

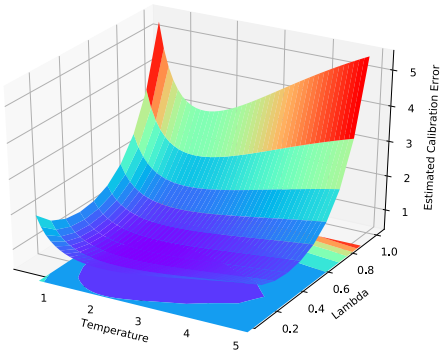
Table: Accuracy (%) on *Office-Home* for Unsupervised Domain Adaptation

Method	Origin	Ours	Δacc	Ar-Cl	Ar-Pr	Ar-Rw	Cl-Ar	Cl-Pr	Cl-Rw	Pr-Ar	Pr-Cl	Pr-Rw	Rw-Ar	Rw-Cl	Rw-Pr
ResNet-50	46.1	**58.4**	12.3	41.1	65.9	73.7	53.1	60.1	63.3	52.2	36.7	71.8	64.8	42.6	75.2
DANN	57.6	**65.2**	7.6	53.8	62.6	74.0	55.8	67.3	67.3	55.8	55.1	77.9	71.1	60.7	81.1
DAN	56.3	**61.4**	5.1	45.6	67.7	73.9	57.7	63.8	66.0	54.9	40.0	74.5	66.2	49.1	77.9
JAN	58.3	**65.9**	7.6	50.8	71.9	76.5	60.6	68.3	68.7	60.5	49.6	76.9	71.0	55.9	80.5
CDAN	65.8	**68.8**	3.0	55.2	72.4	77.6	62.0	69.7	70.9	62.4	54.3	80.5	75.5	61.0	83.8
MCD	-	**67.8**	-	51.7	72.2	78.2	63.7	69.5	70.8	61.5	52.8	78.0	74.5	58.4	81.8
MDD	68.1	**69.6**	1.5	56.4	75.3	78.4	63.2	73.1	73.3	63.9	54.8	79.7	73.2	60.7	83.7

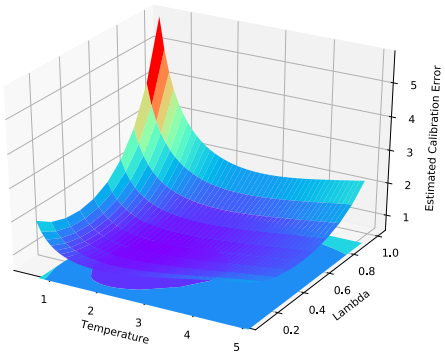
Transfer Learning Library: Validation and Calibration



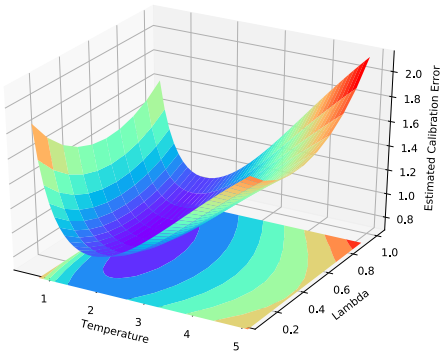




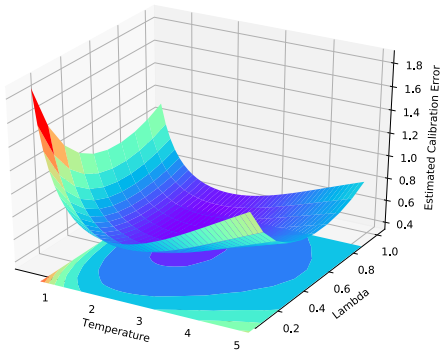
(a) *Art* → *Clipart*



(b) *Art* → *Product*

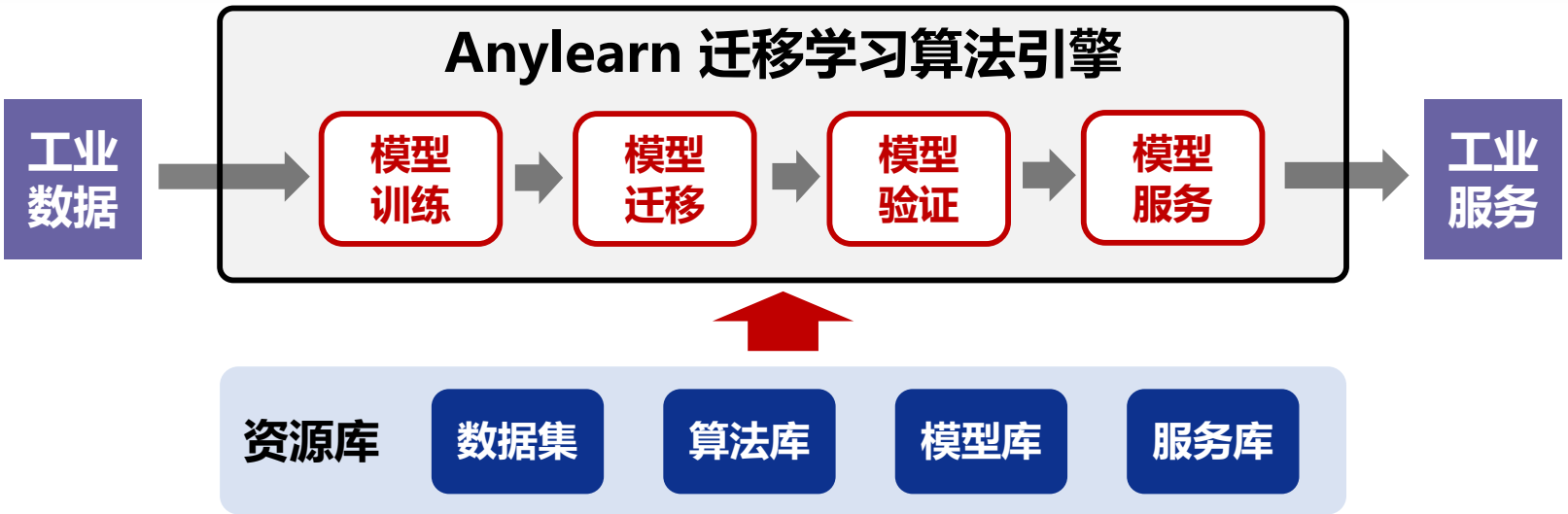


(c) *Art* → *Real-World*



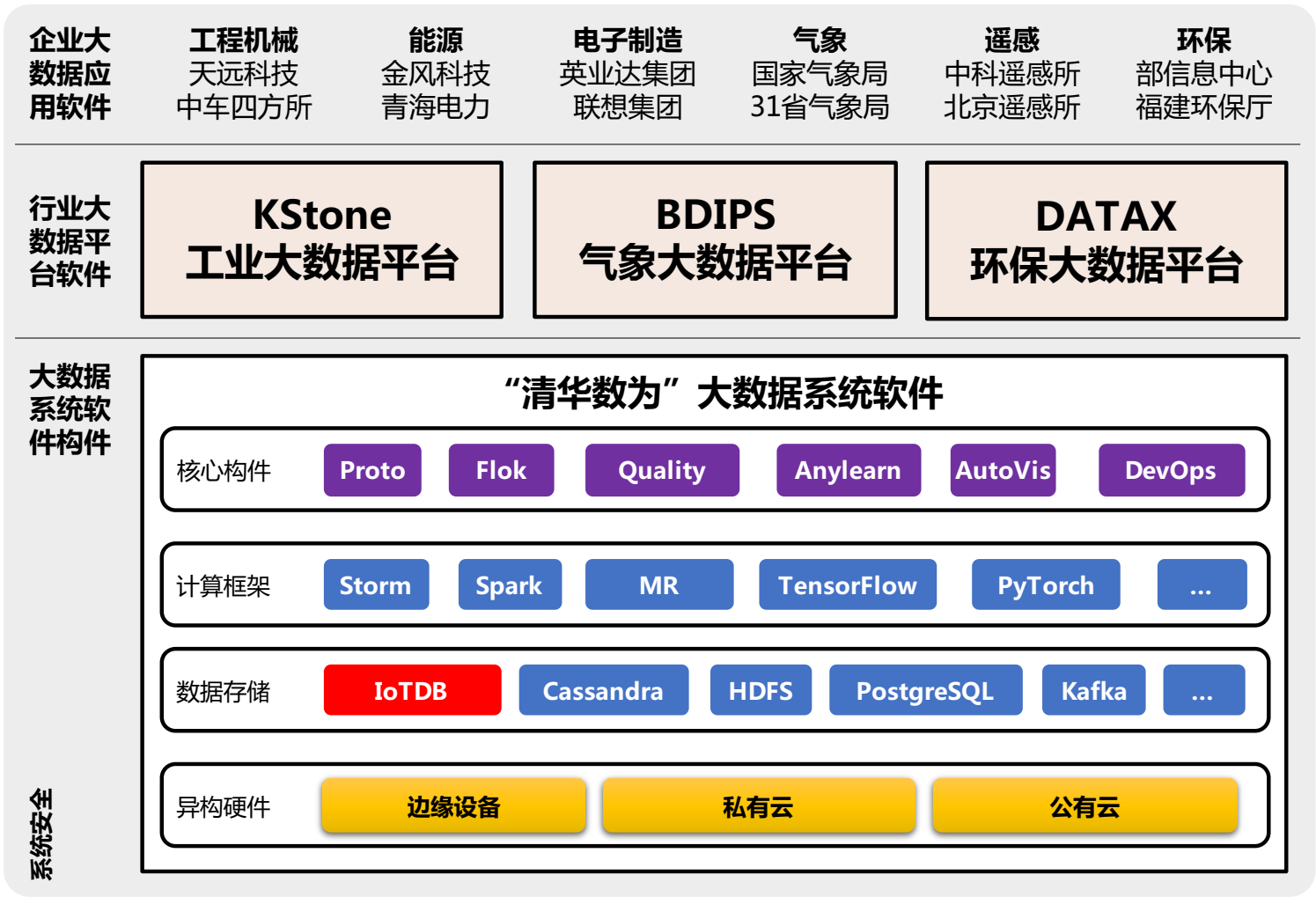
(d) *Clipart* → *Art*

Transfer Learning Software



Anylearn：Anyone and Anywhere Machine Learning

Tsinghua Dataway Big Data Software

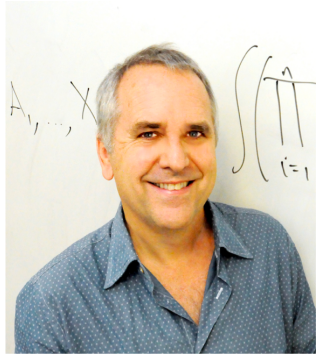


Github: https://github.com/thulab/, https://iotdb.apache.org/

Machine Learning Team @ National Engineering Lab

Jianmin Wang
Professor
Tsinghua University
jimwang@tsinghua.edu.cn

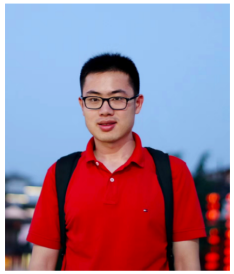
Mingsheng Long
Associate Professor
Tsinghua University
mingsheng@tsinghua.edu.cn



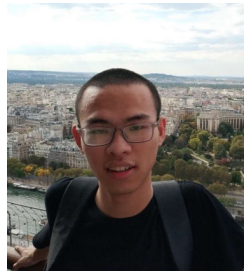
Michael I. Jordan
Professor
UC Berkeley
jordan@cs.berkeley.edu

Yuchen Zhang

Ximei Wang



Zhangjie Cao



Kaichao You

Junguang Jiang

Many thanks for your attention! Any questions?