Lecture 2: Generic Analysis of Domain Adaptation by Divergence Notions

Yang Feng[1], Ye Tian[2]

[1]Department of Biostatistics, School of Global Public Health, New York University
[2]Department of Statistics, Columbia University

Overview

§2.1: No target data

- §2.1.1 Learning from the source solution: the 1st bound
- §2.1.2 An improvement: the 2nd bound
- §2.1.3 Further generalization: go beyond classification
- §2.1.4 Summary

§2.1: No target data

- §2.1.1 Learning from the source solution: the 1st bound
- §2.1.2 An improvement: the 2nd bound
- §2.1.3 Further generalization: go beyond classification
- §2.1.4 Summary

No target data

In some cases, it is hard to collect any data in target domain, for example, self-driving test data.

San Francisco (Source)

Self-driving car testing is allowed

New York City (Target)

Self-driving car testing is prohibited

Then we can only rely on the source data to build the model, and we want to understand its performance on the target domain.

Pictures are generated by Gemini Advanced.

Problem setup: No target data

Consider a noiseless classification problem.

- Target domain: $X \sim \mathbb{P}^{(0)}$, $Y = f^{(0)}(X) : \mathcal{X} \to \{0, 1\}$
  Source domain: $X \sim \mathbb{P}^{(1)}$, $Y = f^{(1)}(X) : \mathcal{X} \to \{0, 1\}$
  where $f^{(0)}$, $f^{(1)}$ are deterministic, and $\mathcal{X} \subseteq \mathbb{R}^d$
- Concept drift: $\mathbb{P}^{(0)} \neq \mathbb{P}^{(1)}$, $f^{(0)} \neq f^{(1)}$
- 0-1 loss function: $\ell(y, y') = \mathbb{1}(y \neq y')$, classification error (risk function): $R^{(k)}(h) = \mathbb{E}_{X \sim \mathbb{P}^{(k)}}[\ell(h(X), f^{(k)}(X))]$
- What we observed: only source data $\{\boldsymbol{x}_i^{(1)}, y_i^{(1)}\}_{i=1}^{n_1}$, no target data
- **Goal:** Learn a classifier $\hat{h}$ from a hypothesis class $\mathcal{H}$ with a small target error $R^{(0)}(\hat{h})$

Question: How?

Learning from the source solution

- **The most intuitive way:** use the source solution $\hat{h}^{(1)} = \arg\min_{h\in\mathcal{H}} \hat{R}^{(1)}(h)$, where the empirical risk $\hat{R}^{(1)}(h) = n_1^{-1}\sum_{i=1}^{n_1} \ell(h(\boldsymbol{x}_i^{(1)}), f^{(1)}(\boldsymbol{x}_i^{(1)}))$
- **Question:** How does it perform on the target?

Denote $R^{(k)}(h_1, h_2) = \mathbb{E}_{X\sim\mathbb{P}^{(k)}}[\ell(h_1(X), h_2(X))]$, and $\phi^{(k)}$ is the density of $\mathbb{P}^{(k)}$.

$$
\begin{aligned}
R^{(0)}(\hat{h}) &= R^{(1)}(\hat{h}) + R^{(0)}(\hat{h}) - R^{(1)}(\hat{h}) \\
&= R^{(1)}(\hat{h}) + R^{(0)}(\hat{h}, f^{(0)}) - R^{(1)}(\hat{h}, f^{(1)}) \\
&\le R^{(1)}(\hat{h}) + |R^{(0)}(\hat{h}, f^{(0)}) - R^{(1)}(\hat{h}, f^{(0)})| + |R^{(1)}(\hat{h}, f^{(0)}) - R^{(1)}(\hat{h}, f^{(1)})| \\
&\le R^{(1)}(\hat{h}) + \underbrace{\left|\int [\phi^{(1)}(\boldsymbol{x}) - \phi^{(0)}(\boldsymbol{x})]\mathbb{1}(\hat{h}(\boldsymbol{x}) \neq f^{(0)}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\right|}_{\le \sup_A |\int [\phi^{(1)}(\boldsymbol{x}) - \phi^{(0)}(\boldsymbol{x})]\mathbb{1}(A)\mathrm{d}\boldsymbol{x}|} \\
&\quad + \mathbb{E}_{X\sim\mathbb{P}^{(1)}}|f^{(1)}(X) - f^{(0)}(X)| \\
&\le R^{(1)}(\hat{h}) + \mathbb{E}_{X\sim\mathbb{P}^{(1)}}|f^{(1)}(X) - f^{(0)}(X)| + d_{\mathsf{TV}}(\mathbb{P}^{(1)}, \mathbb{P}^{(0)}).
\end{aligned}
$$

Learning from the source solution

Theorem 2.1.1 (Ben-David et al., 2006, 2010a)

$$R^{(0)}(\hat{h}) \leq R^{(1)}(\hat{h}) + \mathbb{E}_{X\sim\mathbb{P}^{(1)}}|f^{(1)}(X) - f^{(0)}(X)| + d_{\mathsf{TV}}(\mathbb{P}^{(1)}, \mathbb{P}^{(0)})$$

- Total variation $d_{\mathsf{TV}}(\mathbb{P}^{(1)}, \mathbb{P}^{(0)}) = \sup_A |\mathbb{P}^{(1)}(A) - \mathbb{P}^{(0)}(A)|$.
- If $\mathcal{H}$ is not too large, e.g., $\mathcal{H}$ is a VC-class, then we can further bound the first term as

$$\begin{aligned} R^{(1)}(\hat{h}) &\leq \hat{R}^{(1)}(\hat{h}) + C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}}, \\ &\leq \hat{R}^{(1)}(h) + C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}} \\ &\leq R^{(1)}(h) + 2C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}} \quad \text{w.h.p.} \end{aligned}$$

for any h. Plugging in $h = \arg\min_{h\in\mathcal{H}} R^{(1)}(h)$, we have the following result.

Learning from the source solution

Theorem 2.1.2 (Ben-David et al., 2006, 2010a)

If $\mathcal{H}$ is a VC-class, then w.h.p.,

$$R^{(0)}(\hat{h}) \leq \min_{h \in \mathcal{H}} R^{(1)}(h) + 2C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}} + \mathbb{E}_{X \sim \mathbb{P}^{(1)}} |f^{(1)}(X) - f^{(0)}(X)| + d_{\mathsf{TV}}(\mathbb{P}^{(1)}, \mathbb{P}^{(0)}).$$

In fact, we can play the same trick used before on $\min_{h \in \mathcal{H}} R^{(1)}(h)$ and get the following more interpretable result:

$$R^{(0)}(\hat{h}) \leq \min_{h \in \mathcal{H}} R^{(0)}(h) + 2C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}} + 2\mathbb{E}_{X \sim \mathbb{P}^{(1)}} |f^{(1)}(X) - f^{(0)}(X)| + 2d_{\mathsf{TV}}(\mathbb{P}^{(1)}, \mathbb{P}^{(0)}).$$

[1] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.

[2] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.

No target data: the first bound

Theorem 2.1.3 (The first bound)

If $\mathcal{H}$ is a VC-class, then w.h.p.,

$$R^{(0)}(\hat{h}) \leq \underbrace{\min_{h\in\mathcal{H}} R^{(0)}(h)}_{\text{oracle}} + \underbrace{2C\sqrt{\frac{\mathrm{VC}(\mathcal{H})}{n_1}}}_{\text{cost of learning from samples}} + \underbrace{2\mathbb{E}_{X\sim\mathbb{P}^{(1)}}|f^{(1)}(X) - f^{(0)}(X)|}_{\text{posterior drift}} + \underbrace{2d_{\mathrm{TV}}(\mathbb{P}^{(1)}, \mathbb{P}^{(0)})}_{\text{covariate shift}}.$$

A few comments:

- Error decomposition: there is a ``bias-variance'' trade-off (no free lunch)
- A drawback of this bound: the posterior drift and covariate shift terms do not depend on $\mathcal{H}$ $\implies$ might be too loose

§2.1: No target data

- §2.1.1 Learning from the source solution: the 1st bound
- §2.1.2 An improvement: the 2nd bound
- §2.1.3 Further generalization: go beyond classification
- §2.1.4 Summary

No target data: an improvement

Let's go back the proof to see where we might lose something.

Denote $R^{(k)}(h_1, h_2) = \mathbb{E}_{X\sim\mathbb{P}^{(k)}}[\ell(h_1(X), h_2(X))]$, and $\phi^{(k)}$ is the density of $\mathbb{P}^{(k)}$.

$$
\begin{aligned}
R^{(0)}(\hat{h}) &= R^{(1)}(\hat{h}) + R^{(0)}(\hat{h}) - R^{(1)}(\hat{h}) \\
&= R^{(1)}(\hat{h}) + R^{(0)}(\hat{h}, f^{(0)}) - R^{(1)}(\hat{h}, f^{(1)}) \\
&\le R^{(1)}(\hat{h}) + |R^{(0)}(\hat{h}, f^{(0)}) - R^{(1)}(\hat{h}, f^{(0)})| + |R^{(1)}(\hat{h}, f^{(0)}) - R^{(1)}(\hat{h}, f^{(1)})| \\
&\le R^{(1)}(\hat{h}) + \underbrace{\left| \int [\phi^{(1)}(\boldsymbol{x}) - \phi^{(0)}(\boldsymbol{x})] \mathbb{1}(\hat{h}(\boldsymbol{x}) \neq f^{(0)}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} \right|}_{\le \sup_A |\int [\phi^{(1)}(\boldsymbol{x}) - \phi^{(0)}(\boldsymbol{x})] \mathbb{1}(A) \mathrm{d}\boldsymbol{x}|} \\
&\quad + \mathbb{E}_{X\sim\mathbb{P}^{(1)}} |f^{(1)}(X) - f^{(0)}(X)| \\
&\le R^{(1)}(\hat{h}) + \mathbb{E}_{X\sim\mathbb{P}^{(1)}} |f^{(1)}(X) - f^{(0)}(X)| + d_{\mathsf{TV}}(\mathbb{P}_X^{(1)}, \mathbb{P}_X^{(0)}).
\end{aligned}
$$

- $\sup_A |\int [\phi^{(1)}(\boldsymbol{x}) - \phi^{(0)}(\boldsymbol{x})] \mathbb{1}(A) \mathrm{d}\boldsymbol{x}|$ is loose: no need to consider all measurable sets A
- $\mathbb{E}_{X\sim\mathbb{P}^{(1)}} |f^{(1)}(X) - f^{(0)}(X)|$ is also loose: now it is unrelated to the loss ℓ

No target data: an improvement

For any classifier h^*:

$$\begin{aligned} R^{(0)}(\hat{h}) &= R^{(0)}(h^*) + R^{(0)}(\hat{h}, h^*) \\ &= R^{(0)}(h^*) + R^{(1)}(\hat{h}, h^*) + R^{(0)}(\hat{h}, h^*) - R^{(1)}(\hat{h}, h^*) \\ &\leq R^{(0)}(h^*) + R^{(1)}(h^*) + R^{(1)}(\hat{h}) \\ &\quad + |\mathbb{E}_{X\sim\mathbb{P}^{(0)}} \mathbb{1}(\hat{h}(X) \neq h^*(X)) - \mathbb{E}_{X\sim\mathbb{P}^{(1)}} \mathbb{1}(\hat{h}(X) \neq h^*(X))|. \end{aligned}$$

Definition 2.1.1 (Kifer et al., 2004)

- The function class $\mathcal{H}\triangle\mathcal{H} = \{\mathbb{1}(h(\boldsymbol{x}) \neq h'(\boldsymbol{x})) : h, h' \in \mathcal{H}\}$
- The set collection $I(\mathcal{H}\triangle\mathcal{H}) = \{\{\boldsymbol{x} : \mathbb{1}(g(\boldsymbol{x}) = 1)\} : g \in \mathcal{H}\triangle\mathcal{H}\}$
- The $\mathcal{H}\triangle\mathcal{H}$-divergence $d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}, \mathbb{P}') = \sup_{A\in I(\mathcal{H}\triangle\mathcal{H})} |\mathbb{P}^{(0)}(A) - \mathbb{P}^{(1)}(A)|$

Note that
$|\mathbb{E}_{X\sim\mathbb{P}^{(0)}} \mathbb{1}(\hat{h}(X) \neq h^*(X)) - \mathbb{E}_{X\sim\mathbb{P}^{(1)}} \mathbb{1}(\hat{h}(X) \neq h^*(X))| \leq d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)})$.

[1] Kifer, D., Ben-David, S., & Gehrke, J. (2004, August). Detecting change in data streams. In VLDB (Vol. 4, pp. 180-191).

No target data: an improvement

$$R^{(0)}(\hat{h}) \leq R^{(1)}(\hat{h}) + R^{(0)}(h^*) + R^{(1)}(h^*) + d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}).$$

Let's pick $h^* = \arg\min_{h\in\mathcal{H}}[R^{(0)}(h) + R^{(1)}(h)]$ and denote $\lambda^* = R^{(0)}(h^*) + R^{(1)}(h^*)$. Then

$$R^{(0)}(\hat{h}) \leq R^{(1)}(\hat{h}) + \lambda^* + d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}).$$

Similar to before, if $\mathcal{H}$ is a VC-class, then w.h.p.

$$\begin{aligned} R^{(0)}(\hat{h}) &\leq \hat{R}^{(1)}(\hat{h}) + C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}} + \lambda^* + d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) \\ &\leq \min_{h\in\mathcal{H}} R^{(1)}(h) + 2C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}} + \lambda^* + d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}). \end{aligned}$$

No target data: an improvement

Theorem 2.1.2 (Ben-David et al., 2010a)

If $\mathcal{H}$ is a VC-class, then

$$R^{(0)}(\hat{h}) \leq \min_{h \in \mathcal{H}} R^{(1)}(h) + 2C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}} + \lambda^* + d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}).$$

Compared to our previous result in Theorem 2.1.2:

- $d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)})$ only involves sets in $I(\mathcal{H}\triangle\mathcal{H})$, not all measurable ones
- $\lambda^* = R^{(0)}(h^*) + R^{(1)}(h^*)$ depends on the loss ℓ in a more explicit way

In fact, we can derive other forms of bounds that are more interpretable.

No target data: an improvement

Recall that

$$R^{(0)}(\hat{h}) \le R^{(1)}(\hat{h}) + R^{(0)}(h^*) + R^{(1)}(h^*) + d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}).$$

Instead of setting $h^* = \arg\min_{h\in\mathcal{H}}[R^{(0)}(h) + R^{(1)}(h)]$, we set

$$h^* = \arg\min_{h\in\mathcal{H}} R^{(0)}(h)$$

Therefore,

$$\begin{aligned} R^{(0)}(\hat{h}) &\le \hat{R}^{(1)}(\hat{h}) + C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}} + R^{(0)}(h^*) + R^{(1)}(h^*) + d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) \\ &\le \hat{R}^{(1)}(h^*) + C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}} + R^{(0)}(h^*) + R^{(1)}(h^*) + d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) \\ &\le \underbrace{\min_{h\in\mathcal{H}} R^{(0)}(h)}_{\text{oracle}} + \underbrace{2C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}}}_{\text{cost of learning from samples}} + \underbrace{2R^{(1)}(h^*)}_{\text{posterior drift + covariate shift}} + \underbrace{d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)})}_{\text{covariate shift}}. \end{aligned}$$

No target data: an improvement

Theorem 2.1.3 (The second bound)

$$R^{(0)}(\hat{h}) \leq \underbrace{\min_{h\in\mathcal{H}} R^{(0)}(h)}_{\text{oracle}} + \underbrace{2C\sqrt{\frac{\mathrm{VC}(\mathcal{H})}{n_1}}}_{\text{cost of learning from samples}} + \underbrace{2R^{(1)}(h^*)}_{\text{posterior drift + covariate shift}} + \underbrace{d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)})}_{\text{covariate shift}}.$$

- Compared to the result in Theorem 2.1.2:

$$R^{(0)}(\hat{h}) \leq \min_{h\in\mathcal{H}} R^{(1)}(h) + 2C\sqrt{\frac{\mathrm{VC}(\mathcal{H})}{n_1}} + \min_{h\in\mathcal{H}}\{R^{(0)}(h) + R^{(1)}(h)\} + d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}),$$

  this bound in Theorem 2.1.3 looks more like an oracle inequality and it is easier to interpret

- This is an improvement over the previous bound in Theorem 2.1.2 because:
  - $d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)})$ only involves sets in $I(\mathcal{H}\triangle\mathcal{H})$, not all measurable ones
  - $R^{(1)}(h^*)$ depends on the loss ℓ in a more explicit way

§2.1: No target data

- §2.1.1 Learning from the source solution: the 1st bound
- §2.1.2 An improvement: the 2nd bound
- §2.1.3 Further generalization: go beyond classification
- §2.1.4 Summary

Further generalization: go beyond classification

- Target domain: $X \sim \mathbb{P}^{(0)}$, $Y = f^{(0)}(X) : \mathcal{X} \to \mathcal{Y}$
  Source domain: $X \sim \mathbb{P}^{(1)}$, $Y = f^{(1)}(X) : \mathcal{X} \to \mathcal{Y}$
  where $f^{(0)}$, $f^{(1)}$ are deterministic, and $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{Y} \subseteq \mathbb{R}$
- Concept drift: $\mathbb{P}^{(0)} \neq \mathbb{P}^{(1)}$, $f^{(0)} \neq f^{(1)}$
- General loss function: $\ell(y, y')$, risk $R^{(k)}(h) = \mathbb{E}_{X \sim \mathbb{P}^{(k)}}[\ell(h(X), f^{(k)}(X))]$
  - ℓ is symmetric, bounded, and satisfies triangle inequality
  - Examples:
    - Binary classification: $\mathcal{Y} = \{0, 1\}$, 0-1 loss $\ell_{01}(y, y') = \mathbb{1}(y \neq y')$
    - Regression: $\mathcal{Y} =$ a bounded set in $\mathbb{R}$, ℓ_q-loss $\ell_q(y, y') = |y - y'|^q$, $q \geq 1$
- What we observed: only source data $\{\boldsymbol{x}_i^{(1)}, y_i^{(1)}\}_{i=1}^{n_1}$, no target data
- **Goal:** Obtain a learner $\hat{h}$ from a hypothesis class $\mathcal{H}$ with a small target error $R^{(0)}(\hat{h})$

Further generalization: go beyond classification

Let's go back to the derivation of the ``improved" bound. For any classifier h^*:

$$\begin{aligned} R^{(0)}(\hat{h}) &= R^{(0)}(h^*) + R^{(0)}(\hat{h}, h^*) \\ &= R^{(0)}(h^*) + R^{(1)}(\hat{h}, h^*) + R^{(0)}(\hat{h}, h^*) - R^{(1)}(\hat{h}, h^*) \\ &\leq R^{(0)}(h^*) + R^{(1)}(h^*) + R^{(1)}(\hat{h}) + \underbrace{|R^{(0)}(\hat{h}, h^*) - R^{(1)}(\hat{h}, h^*)|}_{(\star)}. \end{aligned}$$

Previously, we bound $(\star)$ by $d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)})$ when ℓ is 0-1 loss.

Definition 2.1.1 (Mansour et al., 2009)

The discrepency distance $\mathrm{disc}_{\ell,\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) = \sup_{h,h' \in \mathcal{H}} |R^{(0)}(h, h') - R^{(1)}(h, h')|$.

When ℓ is 0-1 loss, $\mathrm{disc}_{\ell,\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) = d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)})$.

Therefore, $R^{(0)}(\hat{h}) \leq R^{(0)}(h^*) + R^{(1)}(h^*) + R^{(1)}(\hat{h}) + \mathrm{disc}_{\ell,\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)})$.

[1] Mansour, Y., Mohri, M., & Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. In 22nd Conference on Learning Theory, COLT 2009.

Further generalization: go beyond classification

Define the Rademacher complexity of $\mathcal{H}$ as

$$\mathcal{R}_{n_1}^{(1)}(\mathcal{H}) = \frac{1}{n_1}\mathbb{E}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{n_1}\sigma_i h(\boldsymbol{x}_i^{(1)})\right], \quad \{\sigma_i\}_{i=1}^{n_1} \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{\pm 1\}).$$

Then by setting $h^* = \arg\min_{h\in\mathcal{H}} R^{(0)}(h)$,

$$\begin{aligned} R^{(1)}(\hat{h}) &\le \hat{R}^{(1)}(\hat{h}) + \sup_{h\in\mathcal{H}} |R^{(1)}(h) - \hat{R}^{(1)}(h)| \\ &\le \hat{R}^{(1)}(h^*) + \sup_{h\in\mathcal{H}} |R^{(1)}(h) - \hat{R}^{(1)}(h)| \\ &\le R^{(1)}(h^*) + 2\sup_{h\in\mathcal{H}} |R^{(1)}(h) - \hat{R}^{(1)}(h)|. \end{aligned}$$

By bounded difference inequality and symmetrization, w.h.p.,

$$\sup_{h\in\mathcal{H}} |R^{(1)}(h) - \hat{R}^{(1)}(\hat{h})| \le C\sqrt{\frac{1}{n_1}} + \mathcal{R}_{n_1}^{(1)}(\ell(f^{(0)}, \mathcal{H})),$$

where $\ell(f^{(0)}, \mathcal{H}) := \{\ell(f^{(0)}, h) : h \in \mathcal{H}\}$.

Further generalization: go beyond classification

Combining all pieces together, we have

$$
\begin{aligned}
R^{(0)}(\hat{h}) &\leq R^{(0)}(h^*) + R^{(1)}(h^*) + R^{(1)}(\hat{h}) + \mathsf{disc}_{\ell,\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) \\
&\leq \underbrace{\min_{h\in\mathcal{H}} R^{(0)}(h)}_{\text{oracle}} + \underbrace{C\sqrt{\frac{1}{n_1}} + \mathcal{R}^{(1)}_{n_1}(\ell(f^{(0)}, \mathcal{H}))}_{\text{cost of learning from samples}} + \underbrace{2R^{(1)}(h^*)}_{\text{posterior drift+covariate shift}} \\
&\quad + \underbrace{\mathsf{disc}_{\ell,\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)})}_{\text{covariate shift}}
\end{aligned}
$$

This result is adapted from Mansour et al. (2009). Their original results are of a similar flavor as the results in Theorem 2.1.2. The version presented here might be easier to interpret and understand.

[1] Mansour, Y., Mohri, M., & Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. In 22nd Conference on Learning Theory, COLT 2009.

Further generalization: go beyond classification

Theorem 2.1.2 (The third bound)

$$R^{(0)}(\hat{h}) \leq \underbrace{\min_{h\in\mathcal{H}} R^{(0)}(h)}_{\text{oracle}} + \underbrace{C\sqrt{\frac{1}{n_1}} + \mathcal{R}^{(1)}_{n_1}(\ell(f^{(0)},\mathcal{H}))}_{\text{cost of learning from samples}} + \underbrace{2R^{(1)}(h^*)}_{\text{posterior drift+covariate shift}} + \underbrace{\text{disc}_{\ell,\mathcal{H}}(\mathbb{P}^{(0)},\mathbb{P}^{(1)})}_{\text{covariate shift}}$$

A few comments:

- This bound is more general than the VC bound
- If loss function $\ell(y,y')$ is Lipschitz over y and y', through Rademacher contraction inequalities (e.g., Vershynin, 2018; Wainwright, 2019), $\mathcal{R}^{(1)}_{n_1}(\ell(f^{(0)},\mathcal{H}))$ can be further bounded by

$$\mathcal{R}^{(1)}_{n_1}(\ell(f^{(0)},\mathcal{H})) \leq \mathcal{R}^{(1)}_{n_1}(\mathcal{H}) + C\sqrt{\frac{1}{n_1}}, \quad \text{w.h.p.}$$

[1] Vershynin, R. (2018). High-dimensional probability: An introduction with applications in data science (Vol. 47). Cambridge university press.

[2] Wainwright, M. J. (2019). High-dimensional statistics: A non-asymptotic viewpoint (Vol. 48). Cambridge university press.

§2.1: No target data

- §2.1.1 Learning from the source solution: the 1st bound
- §2.1.2 An improvement: the 2nd bound
- §2.1.3 Further generalization: go beyond classification
- §2.1.4 Summary

No target data: summary

- **The first bound:**

$$R^{(0)}(\hat{h}) \le \underbrace{\min_{h\in\mathcal{H}} R^{(0)}(h)}_{\text{oracle}} + \underbrace{2C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}}}_{\text{cost of learning from samples}} + \underbrace{2\mathbb{E}_{X\sim\mathbb{P}^{(1)}}|f^{(1)}(X)-f^{(0)}(X)|}_{\text{posterior drift}} + \underbrace{2d_{\mathsf{TV}}(\mathbb{P}^{(1)},\mathbb{P}^{(0)})}_{\text{covariate shift}}.$$

- **The second bound:** $h^* = \arg\min_{h\in\mathcal{H}} R^{(0)}(h)$

$$R^{(0)}(\hat{h}) \le \underbrace{\min_{h\in\mathcal{H}} R^{(0)}(h)}_{\text{oracle}} + \underbrace{2C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}}}_{\text{cost of learning from samples}} + \underbrace{2R^{(1)}(h^*)}_{\text{posterior drift+covariate shift}} + \underbrace{d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)},\mathbb{P}^{(1)})}_{\text{covariate shift}}$$

- **The third bound:** $h^* = \arg\min_{h\in\mathcal{H}} R^{(0)}(h)$, bounded Lipschitz ℓ

$$R^{(0)}(\hat{h}) \le \underbrace{\min_{h\in\mathcal{H}} R^{(0)}(h)}_{\text{oracle}} + \underbrace{2C\sqrt{\frac{1}{n_1}+\mathcal{R}^{(1)}_{n_1}(\mathcal{H})}}_{\text{cost of learning from samples}} + \underbrace{2R^{(1)}(h^*)}_{\text{posterior drift+covariate shift}} + \underbrace{\mathsf{disc}_{\ell,\mathcal{H}}(\mathbb{P}^{(0)},\mathbb{P}^{(1)})}_{\text{covariate shift}}$$

Key message: We can derive various different bounds, but the trade-off phenomenon (error decomposition) always stands.

No target data: Do (2), (3) really improve (1)?

(1) $R^{(0)}(\hat{h}) \le \underbrace{\min_{h\in\mathcal{H}} R^{(0)}(h)}_{\text{oracle}} + \underbrace{2C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}}}_{\text{cost of learning from samples}} + \underbrace{2\mathbb{E}_{X\sim\mathbb{P}^{(1)}}|f^{(1)}(X)-f^{(0)}(X)|}_{\text{posterior drift}} + \underbrace{2d_{\mathsf{TV}}(\mathbb{P}^{(1)},\mathbb{P}^{(0)})}_{\text{covariate shift}}.$

(2)

$$R^{(0)}(\hat{h}) \le \underbrace{\min_{h\in\mathcal{H}} R^{(0)}(h)}_{\text{oracle}} + \underbrace{2C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}}}_{\text{cost of learning from samples}} + \underbrace{2R^{(1)}(h^*)}_{\text{posterior drift+covariate shift}} + \underbrace{d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)},\mathbb{P}^{(1)})}_{\text{covariate shift}}$$

(3)

$$R^{(0)}(\hat{h}) \le \underbrace{\min_{h\in\mathcal{H}} R^{(0)}(h)}_{\text{oracle}} + \underbrace{2C\sqrt{\frac{1}{n_1}+\mathcal{R}^{(1)}_{n_1}(\mathcal{H})}}_{\text{cost of learning from samples}} + \underbrace{2R^{(1)}(h^*)}_{\text{posterior drift+covariate shift}} + \underbrace{\mathsf{disc}_{\ell,\mathcal{H}}(\mathbb{P}^{(0)},\mathbb{P}^{(1)})}_{\text{covariate shift}}$$

- (2), (3) have sharper (pure) covariate shift terms, but have a ``mixed" term and lose some interpretation
- When $f^{(1)} = f^{(0)}$ and $\mathbb{P}^{(1)} = \mathbb{P}^{(0)}$, i.e. no distribution shift, (2) and (3) have a worse constant 3 in the oracle rate, compared to the constant 1 in (1)
- ``All theories are imperfect, but some are useful!"

§2.2: Few target data + many source data

- §2.2.1 Weighted ERM
- §2.2.2 Extension to multiple sources
- §2.2.3 Summary

§2.2: Few target data + many source data

- §2.2.1 Weighted ERM
- §2.2.2 Extension to multiple sources
- §2.2.3 Summary

Few target data + many source data

In some cases, it is possible to collect data from both target and source domains, but the data from target domain are few since they are hard/expensive to get.

- **Medical research:**
  - a specific target population (target) and general population cohorts (source)
  - a rare medical condition (target) and more common medical conditions (source)
- **Educational research:**
  - underrepresented communities (target) and well-represented communities (source)
  - urban schools (target) and affluent schools in the city (sources)

In these cases, we might want to aggregate the data to enhance the performance on target domain.

Problem setup: Few target data + many source data

Consider a noiseless classification problem.

- Target domain: $X \sim \mathbb{P}^{(0)}$, $Y = f^{(0)}(X) : X \mapsto \{0, 1\}$
  Source domain: $X \sim \mathbb{P}^{(1)}$, $Y = f^{(1)}(X) : X \mapsto \{0, 1\}$
  where $f^{(0)}$, $f^{(1)}$ are deterministic
- Concept drift: $\mathbb{P}^{(0)} \neq \mathbb{P}^{(1)}$, $f^{(0)} \neq f^{(1)}$
- 0-1 loss function: $\ell(y, y') = \mathbb{1}(y \neq y')$, classification error (risk function): $R^{(k)}(h) = \mathbb{E}_{X \sim \mathbb{P}^{(k)}}[\ell(h(X), f^{(k)}(X))]$
- What we observed: source data $\{\boldsymbol{x}_i^{(1)}, y_i^{(1)}\}_{i=1}^{n_1}$, target data $\{\boldsymbol{x}_i^{(0)}, y_i^{(0)}\}_{i=1}^{n_0}$
- Usually, $n_0 \ll n_1$
- **Goal:** Learn a classifier $\hat{h}$ from a hypothesis class $\mathcal{H}$ with a small target error $R^{(0)}(\hat{h})$

Question: How?

Few target data + many source data

Instead of doing empirical risk minimization on source data alone, we will consider a weighted combination of target and source risks:

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} [\alpha \hat{R}^{(0)}(h) + (1-\alpha)\hat{R}^{(1)}(h)], \quad \alpha \in [0,1],$$

where $\hat{R}^{(k)}(h) = n_k^{-1} \sum_{i=1}^{n_k} \ell(h(\boldsymbol{x}_i^{(k)}), y_i^{(k)})$, $k = 0, 1$.

Question: How does $\hat{h}$ perform on the target problem? i.e. $R^{(0)}(\hat{h}) \leq ?$

Few target data + many source data

Denote $\hat{R}^{\alpha}(h) = \alpha\hat{R}^{(0)}(h) + (1-\alpha)\hat{R}^{(1)}(h)$,
$R^{\alpha}(h) = \alpha R^{(0)}(h) + (1-\alpha)R^{(1)}(h)$.

Note that

$$R^{(0)}(\hat{h}) \leq R^{\alpha}(\hat{h}) + (1-\alpha)[R^{(0)}(\hat{h}) - R^{(1)}(\hat{h})].$$

Recall our $\mathcal{H}\triangle\mathcal{H}$-divergence:

$$R^{(0)}(\hat{h}) - R^{(1)}(\hat{h}) \leq R^{(0)}(\hat{h}, h^*) - R^{(1)}(\hat{h}, h^*) + [R^{(0)}(f^{(0)}, h^*) + R^{(1)}(f^{(1)}, h^*)]$$

$$\leq d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) + [R^{(0)}(f^{(0)}, h^*) + R^{(1)}(f^{(1)}, h^*)].$$

Take $h^* = \arg\min_{h\in\mathcal{H}}[R^{(0)}(f^{(0)}, h^*) + R^{(1)}(f^{(1)}, h^*)]$:

$$R^{(0)}(\hat{h}) - R^{(1)}(\hat{h}) \leq d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) + \lambda^*,$$

where $\lambda^* = R^{(0)}(f^{(0)}, h^*) + R^{(1)}(f^{(1)}, h^*)$.

Few target data + many source data

With $h^{(0)} = \arg\min_{h\in\mathcal{H}} R^{(0)}(h)$,

$$\begin{aligned}
\hat{R}^{(0)}(\hat{h}) &\le R^{\alpha}(\hat{h}) + (1-\alpha)[d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) + \lambda^*] \\
&\le \hat{R}^{\alpha}(\hat{h}) + C\sqrt{\left[\frac{\alpha^2}{n_0} + \frac{(1-\alpha)^2}{n_1}\right]\mathsf{VC}(\mathcal{H})} + (1-\alpha)[d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) + \lambda^*] \\
&\le R^{\alpha}(h^{(0)}) + 2C\sqrt{\left[\frac{\alpha^2}{n_0} + \frac{(1-\alpha)^2}{n_1}\right]\mathsf{VC}(\mathcal{H})} + (1-\alpha)[d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) + \lambda^*] \\
&\le R^{(0)}(h^{(0)}) + 2C\sqrt{\left[\frac{\alpha^2}{n_0} + \frac{(1-\alpha)^2}{n_1}\right]\mathsf{VC}(\mathcal{H})} + 2(1-\alpha)[d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) + \lambda^*].
\end{aligned}$$

where $R^{(0)}(h^{(0)}) = \min_{h\in\mathcal{H}} R^{(0)}(h)$, and $\lambda^* = R^{(0)}(f^{(0)}, h^*) + R^{(1)}(f^{(1)}, h^*)$.

Few target data + many source data: choose α

Theorem 2.2.1 (Blitzer et al., 2007)

$$\hat{R}^{(0)}(\hat{h}) \leq R^{(0)}(h^{(0)}) + 2C\sqrt{\left[\frac{\alpha^2}{n_0} + \frac{(1-\alpha)^2}{n_1}\right]\mathsf{VC}(\mathcal{H})} + 2(1-\alpha)[d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) + \lambda^*],$$

w.h.p.

Question: Can we optimize the bound over α?

- Theoretically: Yes.
- Practically:
  - $d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) + \lambda^*$ needs to be estimated from the data
  - Estimating $d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)})$ only requires **unlabeled** data, which is usually cheap to get
  - Estimating λ^* requires labeled data
- If we believe the last term is small, then $\alpha = \frac{n_0}{n_0+n_1}$ minimizes the 2nd term $\Longrightarrow$ weighting risk functions by sample size! This implies a non-adaptive rate:

$$\hat{R}^{(0)}(\hat{h}) \leq \min_{h\in\mathcal{H}} R^{(0)}(h) + 2C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_0+n_1}} + 2\underbrace{\frac{n_1}{n_0+n_1}}_{\approx 1}[d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) + \lambda^*].$$

[1] Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Wortman, J. (2007). Learning bounds for domain adaptation. Advances in neural information processing systems, 20.

Few target data + many source data: choose α

The non-adaptive rate: w.h.p.,

$$\hat{R}^{(0)}(\hat{h}) - \min_{h\in\mathcal{H}} R^{(0)}(h) \le 2C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_0+n_1}} + 2\underbrace{\frac{n_1}{n_0+n_1}}_{\approx 1}[d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)},\mathbb{P}^{(1)}) + \lambda^*].$$

- **The ``target-only'' rate**: Doing ERM on the target data leads to

$$\hat{R}^{(0)}(\hat{h}) - \min_{h\in\mathcal{H}} R^{(0)}(h) \le 2C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_0}}, \quad \text{w.h.p.}$$

- When the 3rd term is large (i.e., severe covariate shift or posterior drift), our non-adaptive rate is **worse** than the target-only rate

Remedy: We can use the same training data to estimate the 3rd term, i.e.,

- $\hat{d}_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)},\mathbb{P}^{(1)}) = \sup_{A\in I(\mathcal{H}\triangle\mathcal{H})} |n_0^{-1}\sum_{i=1}^{n_0} \mathbb{1}(\boldsymbol{x}_i^{(0)}\in A) - n_1^{-1}\sum_{i=1}^{n_1}\mathbb{1}(\boldsymbol{x}_i^{(1)}\in A)|$
- $\hat{\lambda} = \arg\min_{h\in\mathcal{H}}[\hat{R}^{(0)}(h) + \hat{R}^{(1)}(h)]$
- It can be shown that $|\hat{d}_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)},\mathbb{P}^{(1)}) - d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)},\mathbb{P}^{(1)})|, |\hat{\lambda}-\lambda^*|$ $\lesssim \sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_0\wedge n_1}} \asymp \sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_0}}$

Few target data + many source data: choose α

Then we get an adaptive choice of α:

$$\alpha = \begin{cases} \frac{n_0}{n_0+n_1}, & \text{if } \hat{d}_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) + \hat{\lambda} \le C' \sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_0}}, \\ 1, & \text{if } \hat{d}_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) + \hat{\lambda} > C' \sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_0}}, \end{cases}$$

where $C' > 0$ is a large constant.

This leads to an **adaptive rate**: w.h.p.

$$\hat{R}^{(0)}(\hat{h}) - \min_{h\in\mathcal{H}} R^{(0)}(h) \lesssim \sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_0+n_1}} + [d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) + \lambda^*] \wedge \sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_0}}.$$

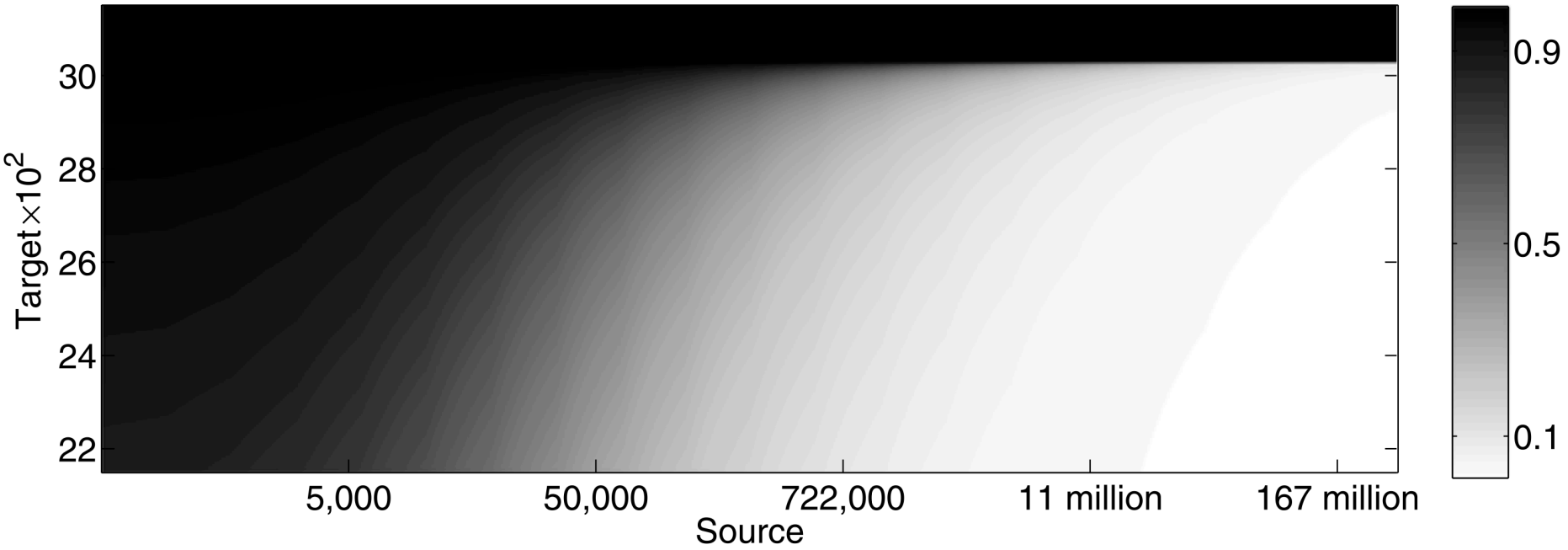
- **Adaptivity:** Even when $d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) + \lambda^*$ is large, $\hat{h}$ performs **no worse than** the target-only ERM classifier
- This rate is already sharp in many cases, but we can further improve it by using

$$\hat{\alpha} = \operatorname*{arg\,min}_{\alpha\in[0,1]} \left\{ 2C \sqrt{\left[\frac{\alpha^2}{n_0} + \frac{(1-\alpha)^2}{n_1}\right] \mathsf{VC}(\mathcal{H})} + 2(1-\alpha)[\hat{d}_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) + \hat{\lambda}] \right\},$$

where we estimate the RHS of Theorem 2.2.1 then choose α to minimize it.

Few target data + many source data: choose α

$$R^{(0)}(\hat{h}) = A\sqrt{\frac{\alpha^2}{n_0} + \frac{(1-\alpha)^2}{n_1}} + B(1-\alpha), \quad A, B \text{ are some constants.}$$



This seems to verify that our first adaptive choice of $\alpha = \frac{n_0}{n_0+n_1}$ or 1 is nearly optimal.

Picture source: Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.

§2.2: Few target data + many source data

- §2.2.1 Weighted ERM
- §2.2.2 Extension to multiple sources
- §2.2.3 Summary

Extension to multiple sources

Consider a noiseless classification problem.

- Target domain: $X \sim \mathbb{P}^{(0)}$, $Y = f^{(0)}(X) : X \mapsto \{0, 1\}$
  Source domains: $X \sim \mathbb{P}^{(k)}$, $Y = f^{(k)}(X) : X \mapsto \{0, 1\}$, $k = 1 : K$
  where $\{f^{(k)}\}_{k=0}^{K}$ are deterministic
- 0-1 loss function: $\ell(y, y') = \mathbb{1}(y \neq y')$, classification error (risk function): $R^{(k)}(h) = \mathbb{E}_{X \sim \mathbb{P}^{(k)}}[\ell(h(X), f^{(k)}(X))]$
- What we observed:
  - Source data $\{\boldsymbol{x}_i^{(k)}, y_i^{(k)}\}_{i=1}^{n_k}$ for $k = 1 : K$;
  - Target data $\{\boldsymbol{x}_i^{(0)}, y_i^{(0)}\}_{i=1}^{n_0}$.
- **Goal:** Learn a classifier $\hat{h}$ from a hypothesis class $\mathcal{H}$ with a small target error $R^{(0)}(\hat{h})$

Extension to multiple sources

Given a weight vector $\boldsymbol{\alpha} = \{\alpha_k\}_{k=0}^K \in \mathcal{S}^K$ (i.e. $\alpha_k \geq 0$ and $\sum_k \alpha_k = 1$), we set

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \left\{ \sum_{k=0}^{K} \alpha_k \hat{R}^{(k)}(h) \right\}.$$

Ben-David et al. (2010a) proves two different bounds of $R^{(0)}(\hat{h})$.

Extension to multiple sources

Theorem 2.2.1 (Pairwise divergence, Ben-David et al., 2010a)

$$R^{(0)}(\hat{h}) \leq \min_{h\in\mathcal{H}} R^{(0)}(h) + C\sqrt{\sum_{k=0}^{K} \frac{\alpha_k^2}{n_k}\mathsf{VC}(\mathcal{H})} + \sum_{k=1}^{K} \alpha_k[2\lambda_k^* + d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(k)}, \mathbb{P}^{(0)})],$$

w.h.p., where $\lambda_k^* = \min_{h\in\mathcal{H}}\{R^{(0)}(h) + R^{(k)}(h)\}$.

Theorem 2.2.2 (Combined divergence, Ben-David et al., 2010a)

$$R^{(0)}(\hat{h}) \leq \min_{h\in\mathcal{H}} R^{(0)}(h) + C'\sqrt{\sum_{k=0}^{K} \frac{\alpha_k^2}{n_k}\mathsf{VC}(\mathcal{H})} + 2\lambda_{\boldsymbol{\alpha}}^* + d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}_{\boldsymbol{\alpha}}, (1-\alpha_0)\mathbb{P}^{(0)}),$$

w.h.p., where $\lambda_{\boldsymbol{\alpha}}^* = \min_{h\in\mathcal{H}}\{(1-\alpha_0)R^{(0)}(h) + \sum_{k=1}^{K}\alpha_k R^{(k)}(h)\}$,
$\mathbb{P}_{\boldsymbol{\alpha}} = \sum_{k=1}^{K}\alpha_k\mathbb{P}^{(k)}$.

- Theorem 2.2.1 reduces to one of the previous single-source results when $K = 1$.
- Similar to our previous discussions, we can find the optimal $\boldsymbol{\alpha}$ by estimating the RHS then minimize it.

[1] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.

Weighting risk functions or weighting hypotheses?

- We have studied the ERM hypothesis derived from **weighted risk functions**:

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \left\{ \sum_{k=0}^{K} \alpha_k \hat{R}^{(k)}(h) \right\}.$$

- What about the ERM hypothesis derived from **weighted hypotheses**?

$$\hat{h}(\boldsymbol{x}) = \sum_{k=0}^{K} \alpha_k \hat{h}^{(k)}(\boldsymbol{x}),$$

  where $\hat{h}^{(k)} = \arg\min_{h \in \mathcal{H}} \hat{R}^{(k)}(h)$. [1]

- **Facts:**
  - For some specific hypothesis classes $\mathcal{H}$ and loss functions ℓ (e.g., linear regressions with square loss), weighting hypotheses can deliver a similar rate.
  - But in general, two weighting strategies can be quite different.

[1] For classification, a post-processing step is needed to ensure that $\hat{h}$ is a classifier.

Weighting risk functions or weighting hypotheses?

- For weighted hypotheses, the proof used in Ben-David et al. (2010a) usually leads to an extra term $\sim \sqrt{(K+1)/\sum_{k=0}^{K} n_k}$ due to the simultaneous control of $(K+1)$ hypotheses $h \in \mathcal{H}$, which is sub-optimal in many cases
- Mansour et al. (2008) creates the following example where any convex combination of source hypotheses fails on the target domain

A regression example of $K = 2$ sources with no target data: $\mathcal{X} = \{a, b\}$, $\mathbb{P}_X^{(1)} = \delta_a$, $\mathbb{P}_{Y|X}^{(2)} = \delta_0$, $\mathbb{P}_X^{(1)} = \delta_b$, $\mathbb{P}_{Y|X}^{(2)} = \delta_1$. Consider absolute loss ℓ and zero-error hypotheses $h^{(1)}(x) \equiv 0$, $h^{(2)}(x) \equiv 1$ on two source domains. On target domain, $\mathbb{P}_{X,Y}^{(0)} = \frac{1}{2}\mathbb{P}_{X,Y}^{(1)} + \frac{1}{2}\mathbb{P}_{X,Y}^{(2)}$.

- For any $\lambda \in [0, 1]$, $h_\lambda := \lambda h^{(0)} + (1-\lambda) h^{(1)}$ has target risk $1/2$
- Instead, for any $\lambda \in (0, 1)$, $\tilde{h}_\lambda := \underset{h:\{a,b\}\to\{0,1\}}{\arg\min} \{\lambda R^{(0)}(h) + (1-\lambda) R^{(1)}(h)\}$ has target risk 0

Why this happens: Learning single source hypotheses might be unstable!

[1] Mansour, Y., Mohri, M., & Rostamizadeh, A. (2008). Domain adaptation with multiple sources. Advances in neural information processing systems, 21.

Weighting risk functions or weighting hypotheses?

- Mansour et al. (2008) shows that the weighted hypotheses may generalize well to the target domain, but the weight depends on the $\mathbb{P}_X^{(k)}$ and differs for different test points
- Moreover, if we fix a hypothesis class $\mathcal{H}$ to learn source hypotheses, the weighted average of them may not belong to $\mathcal{H}$

Overall, weighting the risk functions might be easier to analyze and more reliable in some cases.

[1] Mansour, Y., Mohri, M., & Rostamizadeh, A. (2008). Domain adaptation with multiple sources. Advances in neural information processing systems, 21.

§2.2: Few target data + many source data

- §2.2.1 Weighted ERM
- §2.2.2 Extension to multiple sources
- §2.2.3 Summary

Few target data + many source data: a summary

Weighted estimator: $\hat{h} \in \arg\min_{h\in\mathcal{H}}[\alpha\hat{R}^{(0)}(h) + (1-\alpha)\hat{R}^{(1)}(h)], \quad \alpha \in [0,1].$

- **A general bound:** w.h.p.

  $$\hat{R}^{(0)}(\hat{h}) \le R^{(0)}(h^{(0)}) + 2C\sqrt{\left[\frac{\alpha^2}{n_0} + \frac{(1-\alpha)^2}{n_1}\right]\mathsf{VC}(\mathcal{H})} + 2(1-\alpha)[d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) + \lambda^*],$$

  where $\lambda^* = \min_{h\in\mathcal{H}}\{R^{(0)}(h) + R^{(1)}(h)\}$.
- Some choices of α:
  - $\alpha = n_0/(n_0+n_1)$: good choice for weak distribution shift, potential negative transfer $\Rightarrow$ non-adaptive rate
  - Estimate α from the data $\Rightarrow$ non-adaptive rate
- Extension to the multi-source situation

§2.3: Posterior drift

Posterior drift

Examples:

- Social science: The population of a county does not change a lot, but the social policy changes over time, which leads to changes of the response
- Movie/Book rating: The underlying rating mechanisms for different genres of movies/books might be different.
- ...

In this section, we will focus on posterior drift in a general set-up, and we will see that the generalization error under **posterior drift** is simpler than the previous result under **concept drift**.

In the Lecture 4, we will come back to posterior drift and discuss it under more specific models.

Posterior drift: problem setup

- Target domain: $X \sim \mathbb{P}^{(0)}$, $Y = f^{(0)}(X) : \mathcal{X} \to \mathcal{Y}\}$
  Source domains: $X \sim \mathbb{P}^{(k)}$, $Y = f^{(k)}(X) : \mathcal{X} \to \mathcal{Y}$, $k = 1 : K$
  where $\{f^{(k)}\}_{k=0}^{K}$ are deterministic, and $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{Y} \subseteq \mathbb{R}$
- **Posterior drift:** $\mathbb{P}^{(k)}$ all equal $:= \mathbb{P}$, $f^{(k)}$ are not equal
- General loss function: $\ell(y, y')$, risk $R^{(k)}(h) = \mathbb{E}_{X \sim \mathbb{P}}[\ell(h(X), f^{(k)}(X))]$
  - ℓ is symmetric, bounded, and satisfies triangle inequality
  - Examples:
    - Binary classification: $\mathcal{Y} = \{0, 1\}$, 0-1 loss $\ell_{01}(y, y') = \mathbb{1}(y \neq y')$
    - Regression: $\mathcal{Y} =$ a bounded set in $\mathbb{R}$, ℓ_q-loss $\ell_q(y, y') = |y - y'|^q$, $q \geq 1$
- What we observed:
  - Source data $\{\boldsymbol{x}_i^{(k)}, y_i^{(k)}\}_{i=1}^{n_k}$ for $k = 1 : K$;
  - Target data $\{\boldsymbol{x}_i^{(0)}, y_i^{(0)}\}_{i=1}^{n_0}$.
- **Goal:** Obtain a learner $\hat{h}$ from a hypothesis class $\mathcal{H}$ with a small target error $R^{(0)}(\hat{h})$

Posterior drift: problem setup

- Target domain: $X \sim \mathbb{P}^{(0)}$, $Y = f^{(0)}(X) : \mathcal{X} \to \mathcal{Y}\}$
  Source domains: $X \sim \mathbb{P}^{(k)}$, $Y = f^{(k)}(X) : \mathcal{X} \to \mathcal{Y}$, $k = 1 : K$
  where $\{f^{(k)}\}_{k=0}^{K}$ are deterministic, and $\mathcal{X} \subseteq \mathbb{R}^d$, bounded $\mathcal{Y} \subseteq \mathbb{R}$
- We will characterize the similarity between target and the k-th source by $R(f^{(0)}, f^{(k)}) := \mathbb{E}_{X \sim \mathbb{P}}[\ell(f^{(0)}(X), f^{(k)}(X))]$.
- WLOG, assume $R(f^{(0)}, f^{(1)}) \leq R(f^{(0)}, f^{(2)}) \ldots \leq R(f^{(0)}, f^{(K)})$.
- For ℓ, we assume:
  - It is symmetric;
  - It satisfies β-triangle inequality with $\beta \geq 1$:
    $$\ell(h_1, h_2) \leq \beta[\ell(h_1, h_3) + \ell(h_3, h_2)].$$
    Examples: 0-1 loss ℓ_{01}, ℓ_q-loss ($q \geq 1$), square-root loss ($\sqrt{x - x'}$)

Posterior drift: problem setup

- Consider the weighted ERM $\hat{h} = \arg\min_{h \in \mathcal{H}} \left\{ \sum_{k=0}^{K} \alpha_k \hat{R}^{(k)}(h) \right\}$, where $\sum_{k=0}^{K} \alpha_k = 1$.
- Assume that the following uniform concentration holds:

$$|\hat{R}_{\boldsymbol{\alpha}}(h) - R_{\boldsymbol{\alpha}}(h)| \leq \text{rate}(\boldsymbol{n}, \boldsymbol{\alpha}, \mathcal{H}, \ell), \quad \forall h \in \mathcal{H}, \text{ w.h.p.}$$

  where $\hat{R}_{\boldsymbol{\alpha}}(h) = \sum_{k=0}^{K} \alpha_k \hat{R}^{(k)}(h)$, $R_{\boldsymbol{\alpha}}(h) = \sum_{k=0}^{K} \alpha_k R^{(k)}(h)$, $\boldsymbol{n} = \{n_k\}_{k=0}^{K}$.

  - In general: $\text{rate}(\boldsymbol{n}, \boldsymbol{\alpha}, \mathcal{H}, \ell) \leq 2\mathcal{R}(\boldsymbol{n}, \boldsymbol{\alpha}, \mathcal{H}, \ell) + C\sqrt{\sum_{k=0}^{K}(\alpha_k^2/n_k)}$, where the Rademacher complexity $\mathcal{R}(\boldsymbol{n}, \boldsymbol{\alpha}, \mathcal{H}, \ell) := \mathbb{E}\big[\sup_{h \in \mathcal{H}} \sum_{k=0}^{K} \frac{\alpha_k}{n_k} \sum_{i=1}^{n_k} \sigma_i^{(k)} \ell(h(\boldsymbol{x}_i^{(k)}), f^{(k)}(\boldsymbol{x}_i^{(k)}))\big]$
  - For ℓ_{01} loss and a VC-class $\mathcal{H}$, similar to before: $\text{rate}(\boldsymbol{n}, \boldsymbol{\alpha}, \mathcal{H}, \ell) \lesssim \sqrt{\text{VC}(\mathcal{H}) \sum_{k=0}^{K}(\alpha_k^2/n_k)}$
  - For ℓ_q loss and many d-dimensional parametric classes $\mathcal{H}$: $\text{rate}(\boldsymbol{n}, \boldsymbol{\alpha}, \mathcal{H}, \ell) \lesssim \sqrt{d \sum_{k=0}^{K}(\alpha_k^2/n_k)}$

Posterior drift: target excess risk

The analysis is quite similar to our previous one. Let $h^{(0)} = \arg\min_{h\in\mathcal{H}} R^{(0)}(h)$.

$$
\begin{aligned}
R(\hat{h}, f^{(0)}) &\le \sum_{k=0}^{K} \alpha_k \beta [R(\hat{h}, f^{(k)}) + R(f^{(k)}, f^{(0)})] \\
&= \beta R_{\boldsymbol{\alpha}}(\hat{h}) + \beta \sum_{k=1}^{K} \alpha_k R(f^{(k)}, f^{(0)}) \\
&\le \beta \hat{R}_{\boldsymbol{\alpha}}(\hat{h}) + \beta \cdot \mathsf{rate}(\boldsymbol{n}, \boldsymbol{\alpha}, \mathcal{H}, \ell) + \beta \sum_{k=1}^{K} \alpha_k R(f^{(k)}, f^{(0)}) \\
&\le \beta R_{\boldsymbol{\alpha}}(h^{(0)}) + 2\beta \cdot \mathsf{rate}(\boldsymbol{n}, \boldsymbol{\alpha}, \mathcal{H}, \ell) + \beta \sum_{k=1}^{K} \alpha_k R(f^{(k)}, f^{(0)}) \\
&\le \beta^2 R^{(0)}(h^{(0)}) + 2\beta \cdot \mathsf{rate}(\boldsymbol{n}, \boldsymbol{\alpha}, \mathcal{H}, \ell) + (\beta + \beta^2) \sum_{k=1}^{K} \alpha_k R(f^{(k)}, f^{(0)})
\end{aligned}
$$

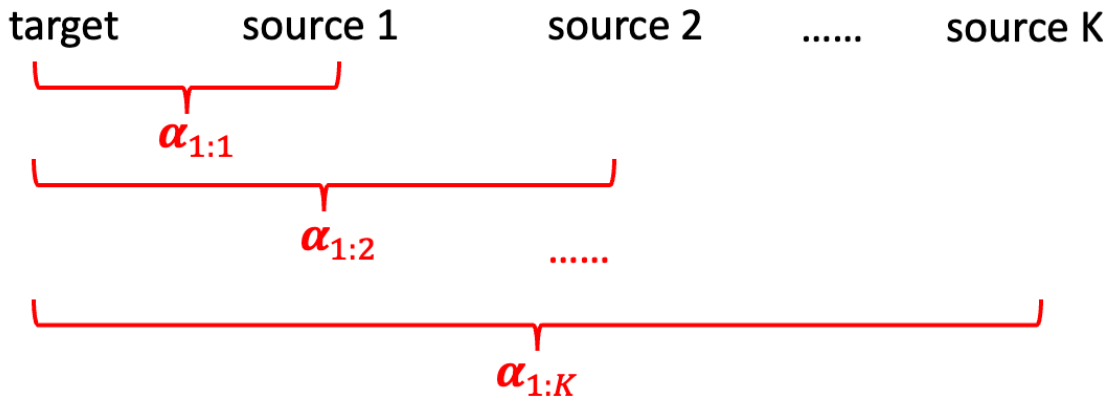
Posterior drift: target excess risk

Theorem 2.3.1 (An extension of the result in Crammer et al., 2008)

w.h.p.

$$R(\hat{h}, f^{(0)}) \leq \underbrace{\beta^2 \min_{h \in \mathcal{H}} R^{(0)}(h)}_{\text{oracle}} + \underbrace{2\beta \cdot \text{rate}(\boldsymbol{n}, \boldsymbol{\alpha}, \mathcal{H}, \ell)}_{\text{cost of learning from samples}} + \underbrace{(\beta + \beta^2) \sum_{k=1}^{K} \alpha_k R(f^{(k)}, f^{(0)})}_{\text{posterior drift}}.$$

- For ℓ_{01} loss and a VC-class $\mathcal{H}$ of dimension d, or for ℓ_q loss and many d-dimensional parametric classes $\mathcal{H}$: $\text{rate}(\boldsymbol{n}, \boldsymbol{\alpha}, \mathcal{H}, \ell) \lesssim \sqrt{d \sum_{k=0}^{K} (\alpha_k / n_k)^2}$
- If we pick $\alpha_k = n_k / N$ with $N = \sum_{k=0}^{K} n_k$: $\text{rate}(\boldsymbol{n}, \boldsymbol{\alpha}, \mathcal{H}, \ell) \lesssim \sqrt{d/N}$
- Crammer et al. (2008) considered a progressive source inclusion with K $\boldsymbol{\alpha}$-choices:



[1] Crammer, K., Kearns, M., & Wortman, J. (2008). Learning from Multiple Sources. Journal of Machine Learning Research, 9(8).

Posterior drift: target excess risk

Theorem 2.3.2 (Crammer et al., 2008)

Consider $\hat{h}_{0:k} = \arg\min_{h\in\mathcal{H}}\{\sum_{j=0}^{k} \frac{n_j}{n_{0:k}} \hat{R}^{(k)}(h)\}$. *Then w.h.p., for all* $k = 0 : K$,

$$R(\hat{h}_{1:k}, f^{(0)}) \le \underbrace{\beta^2 \min_{h\in\mathcal{H}} R^{(0)}(h)}_{\text{oracle}} + \underbrace{2\beta \text{rate}(\{n_j\}_{j=0}^{k}, \mathcal{H}, \ell)}_{\text{cost of learning from samples}} + \underbrace{(\beta+\beta^2)\sum_{j=1}^{k} \frac{n_j}{n_{0:k}} R(f^{(j)}, f^{(0)})}_{\text{posterior drift}}.$$

- For ℓ_{01} loss and a VC-class $\mathcal{H}$ of dimension d, or for ℓ_q loss and many d-dimensional parametric classes $\mathcal{H}$:
  $\text{rate}(\{n_j\}_{j=0}^{k}, \mathcal{H}, \ell) \lesssim \sqrt{\frac{d+\log K}{n_{0:k}}}$, $n_{0:k} = \sum_{j=0}^{k} n_j$
- We can choose the optimal hypothesis from $\{\hat{h}_{0:k}\}_{k=0}^{K}$ based on the in-sample or hold-out evaluation on the RHS, and analyze its excess risk

[1] Crammer, K., Kearns, M., & Wortman, J. (2008). Learning from Multiple Sources. Journal of Machine Learning Research, 9(8).

§2.4: Hardness results

Recall the general upper bound

In this section, we will understand where the hardness comes from in domain adaptation.

Let's recall our generalization bound in §2.1 on target domain when a single source data set of size n_1 is present and no target data is available.

The second bound of target risk in §2.1: If $\mathcal{H}$ is a VC-class, then

$$R^{(0)}(\hat{h}) \leq \underbrace{\min_{h\in\mathcal{H}} R^{(0)}(h)}_{\text{oracle}} + \underbrace{2C\sqrt{\frac{\mathrm{VC}(\mathcal{H})}{n_1}}}_{\text{cost of learning from samples}} + \underbrace{2R^{(1)}(h^*)}_{\text{posterior drift + covariate shift}} + \underbrace{d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)})}_{\text{covariate shift}},$$

where $h^* = \arg\min_{h\in\mathcal{H}} R^{(0)}(h)$.

Question: Are the last two terms necessary?

Answering this question can help me understand the fundamental difficulty of domain adaptation.

Hardness results: covariate shift

Consider the noiseless classification setting with covariate shift:

- Target domain: $X \sim \mathbb{P}^{(0)}$, $Y = f(X) : \mathcal{X} \to \{0, 1\}$
- Source domain: $X \sim \mathbb{P}^{(1)}$, $Y = f(X) : \mathcal{X} \to \{0, 1\}$
- Covariate drift: $\mathbb{P}^{(0)} \neq \mathbb{P}^{(1)}$
- 0-1 loss function: $\ell(y, y') = \mathbb{1}(y \neq y')$, classification error (risk function): $R^{(k)}(h) = \mathbb{E}_{X \sim \mathbb{P}^{(k)}}[\ell(h(X), f(X))]$
- What we observed: source data $S = \{\boldsymbol{x}_i^{(1)}, y_i^{(1)}\}_{i=1}^{n_1}$, unlabeled target data $T = \{\boldsymbol{x}_i^{(0)}\}_{i=1}^{n_0}$
- Hypothesis class $\mathcal{H}$ is a VC-class, $h^* = \arg\min_{h \in \mathcal{H}} R^{(0)}(h)$

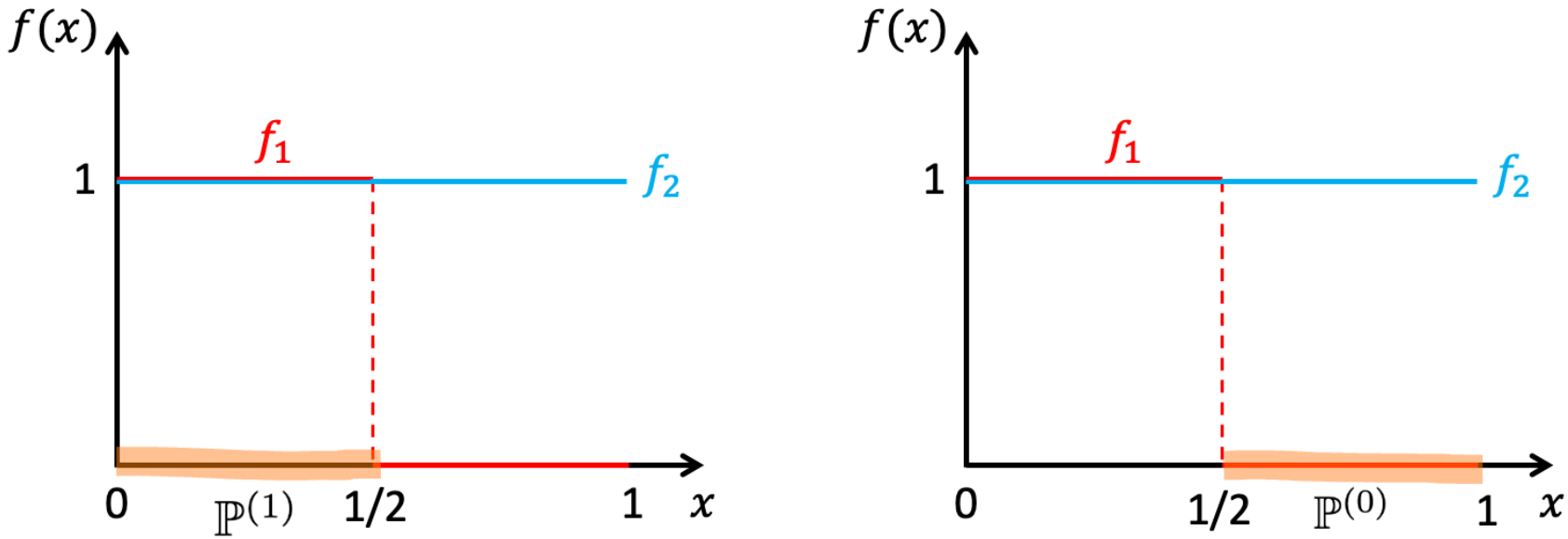
Hardness results: covariate shift

Theorem 2.4.1 (Necessity of a small $d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)})$, Ben-David et al., 2010b)

For any $\epsilon > 0$ and learner $\hat{h} = \hat{h}(\cdot; S) : \mathcal{X} \to \{0, 1\}$, $\exists$ a labeling function f, s.t.

(1) $R^{(1)}(h^*) \leq \epsilon$;

(2) $\mathbb{P}_{S,T}[R^{(0)}(\hat{h}) \geq 1/2] \geq 1/2$.



$\mathcal{H} = \{f_1, f_2\}$, $d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) = 1$. No algorithm can distinguish f_1, f_2!

[1] David, S. B., Lu, T., Luu, T., & Pál, D. (2010, March). Impossibility theorems for domain adaptation. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (pp. 129-136). JMLR Workshop and Conference Proceedings.

Hardness results: covariate shift

Theorem 2.4.2 (Necessity of a small $R^{(1)}(h^*)$, Ben-David et al., 2010b)

For any $\epsilon > 0$ and learner $\hat{h} = \hat{h}(\cdot; S) : \mathcal{X} \to \{0, 1\}$, $\exists$ a labeling function f, s.t.

(1) $d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) \leq \epsilon$;

(2) $\mathbb{P}_{S,T}[R^{(0)}(\hat{h}) \geq 1/2] \geq 1/2$.

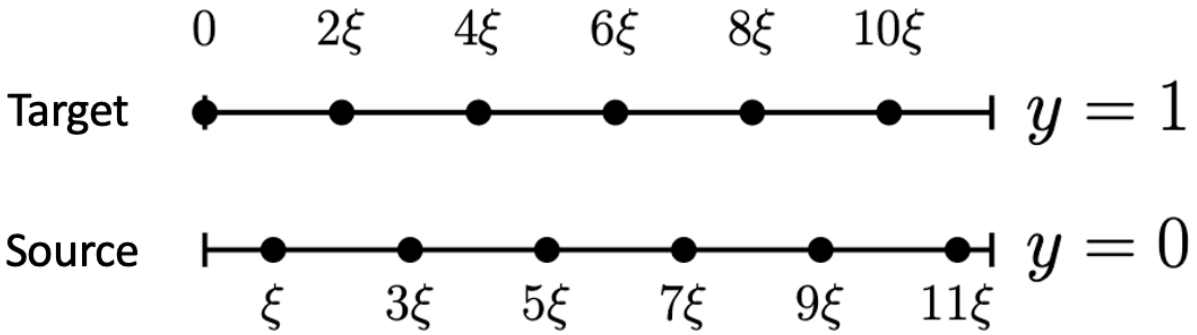
A handy example:

- The true $f(x) = \begin{cases} 0, & \text{if } 0 \leq x \leq 1/2; \\ 1, & \text{if } 1/2 < x \leq 1. \end{cases}$ $\mathcal{H} = \{\text{constant function } 1\}$, $\mathbb{P}^{(0)} = \mathbb{P}^{(1)}$.
- In fact, in this example, f is not learnable by $\mathcal{H}$ (i.e. $\min_{h \in \mathcal{H}} R^{(0)} = 1/2$)

Question: Is there a more ``beautiful" example, where transferring from the source fails but $\min_{h \in \mathcal{H}} R^{(0)}$ is small?

Hardness results: covariate shift

Consider the following example in Ben-David et al. (2010b).



- $\mathbb{P}^{(0)} = \text{Unif}(\{0, 2\epsilon, 4\epsilon, \ldots, 1\})$, $\mathbb{P}^{(1)} = \text{Unif}(\{\epsilon, 3\epsilon, \ldots, 1-\epsilon\})$, $\mathcal{H} = \{h(x) = \mathbb{1}(x \le t) : t \in [0,1]\}$, $\mathbb{P}^{(0)} = \mathbb{P}^{(1)}$.

  The true $f_1(x) = \begin{cases} 0, & \text{if } x = \epsilon, 3\epsilon, \ldots, 1-\epsilon; \\ 1, & \text{if } x = 0, 2\epsilon, 4\epsilon, \ldots, 1. \end{cases}$ or $f_2(x) \equiv 0$

- Check:
  - $\mathcal{H} \triangle \mathcal{H} = \{(a, b] : 0 \le a \le b \le 1\}$, $d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) = \epsilon$
  - $\min_{h\in\mathcal{H}} R^{(0)}(h) = \min_{h\in\mathcal{H}} R^{(1)}(h) = 0$
  - $R^{(1)}(h^*) \ge 1 - \epsilon$, $\forall h^* \in \arg\min_{h\in\mathcal{H}} R^{(0)}(h)$
  - No algorithm can distinguish f_1 and f_2!

[1] David, S. B., Lu, T., Luu, T., & Pál, D. (2010, March). Impossibility theorems for domain adaptation. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (pp. 129-136). JMLR Workshop and Conference Proceedings.

Hardness results: summary

The second bound of target risk in §2.1: If $\mathcal{H}$ is a VC-class, then

$$R^{(0)}(\hat{h}) \leq \underbrace{\min_{h\in\mathcal{H}} R^{(0)}(h)}_{\text{oracle}} + \underbrace{2C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}}}_{\text{cost of learning from samples}} + \underbrace{2R^{(1)}(h^*)}_{\text{posterior drift + covariate shift}} + \underbrace{d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)})}_{\text{covariate shift}},$$

where $h^* = \arg\min_{h\in\mathcal{H}} R^{(0)}(h)$.

Question: Are the last two terms necessary?

Answer:

- Yes, they are, if no labeled target data is provided.
- In our constructed hard cases, there is no posterior drift. $\Longrightarrow$ Even the covariate shift itself can cause lots of trouble.
- Similarly, in the case of posterior drift , it is easy to show that a small $R^{(1)}(h^*)$ is necessary.
- Access to **labeled** target data is helpful. E.g., with some labeled target data, the target problem is at least learnable, despite of a potential slow rate.

Hardness results: summary

- Ben-David and Urner (2012) has more discussions on the hardness of domain adaptation

[1] Ben-David, S., & Urner, R. (2012). On the hardness of domain adaptation and the utility of unlabeled target samples. In Algorithmic Learning Theory: 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings 23 (pp. 139-153). Springer Berlin Heidelberg.

§2.5: Other similarity notions: go beyond divergence

Problems of previous divergence notions

We have seen that how different divergence notions can be used to bound the target risk.

The second bound of target risk in §2.1: If $\mathcal{H}$ is a VC-class, then w.h.p.

$$R^{(0)}(\hat{h}) \le \underbrace{\min_{h\in\mathcal{H}} R^{(0)}(h)}_{\text{oracle}} + \underbrace{2C\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{n_1}}}_{\text{cost of learning from samples}} + \underbrace{2R^{(1)}(h^*)}_{\text{posterior drift + covariate shift}} + \underbrace{d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)})}_{\text{covariate shift}},$$

where $h^* = \arg\min_{h\in\mathcal{H}} R^{(0)}(h)$.

There are some issues with the divergence notions we discussed before.

- They might be **over-pessimistic**, i.e. $\exists$ cases where $d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) = \Omega(1)$ while the target excess risk is $o_{\mathbb{P}}(1)$
- $d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)})$ is **symmetric**, while transfer learning is **asymmetric**.

Question: Are there other notions to characterize how much information can be transferred from source to target?

Go beyond divergence

Hanneke and Kpotufe (2019) proposes a notion called **transfer exponent** under the non-parametric classification setting (with 0-1 loss).

Definition 2.5.1 (Hanneke and Kpotufe, 2019)

We call $\rho > 0$ a **transfer exponent** from $\mathbb{P}^{(1)}$ to $\mathbb{P}^{(0)}$ w.r.t. $\mathcal{H}$, if

$$\mathcal{E}^{\rho}_{\mathbb{P}^{(0)}}(h) \leq C_\rho \mathcal{E}_{\mathbb{P}^{(1)}}(h), \ \forall h \in \mathcal{H},$$

where C_ρ is a universal constant and $\mathcal{E}^{\rho}_{\mathbb{P}^{(k)}}$ is the excess risk under distribution $\mathbb{P}^{(k)}$.

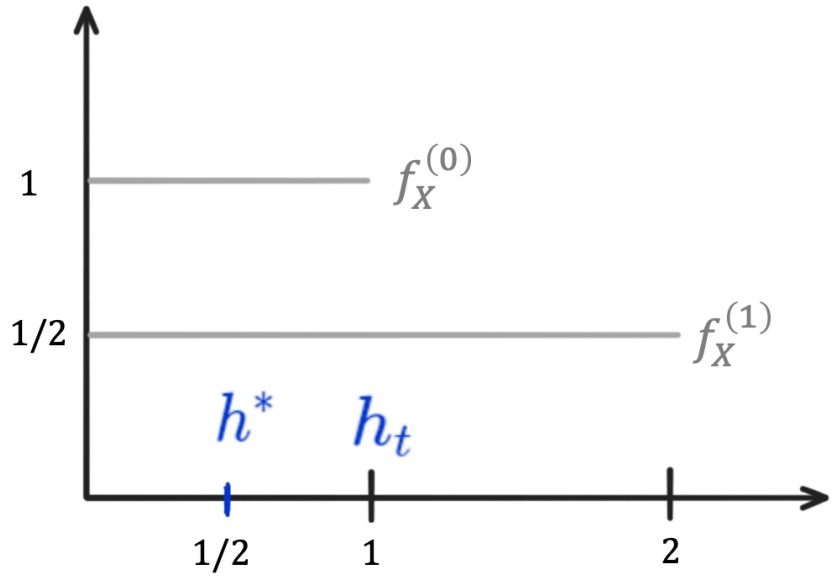
- Smaller $\rho \Rightarrow$ more information to transfer
- This notion is **asymmetric**
- Hanneke and Kpotufe (2019) derives minimax rate for the target excess risk using ρ

[1] Hanneke, S., & Kpotufe, S. (2019). On the value of target data in transfer learning. Advances in Neural Information Processing Systems, 32.

Go beyond divergence

Let's see an example where $d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) = \Omega(1)$ while $\rho > 0$ (hence leads to a shrinking target excess risk).

$\mathbb{P}_X^{(0)} = \text{Unif}(0,1)$, $\mathbb{P}_X^{(1)} = \text{Unif}(0,2)$, $\mathcal{H}$ is the class of one-sided thresholds on the line. Suppose Y is deterministic given X. The optimal classifier $h^*(x) = \mathbb{1}(x \le 1/2)$.



- $d_{\mathcal{H}\triangle\mathcal{H}}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) = \sup_{h,h' \in \mathcal{H}} |\mathbb{P}^{(0)}(h \neq h') - \mathbb{P}^{(1)}(h \neq h')| = 1/2$
- $\mathcal{E}_{\mathbb{P}^{(0)}}(h) = \mathbb{P}^{(0)}(h \neq h^*) \le 2\mathbb{P}^{(1)}(h \neq h^*) = 2\mathcal{E}_{\mathbb{P}^{(1)}}(h) \;\Rightarrow\; \rho = 1$

Picture adapted from: Hanneke, S., & Kpotufe, S. (2019). On the value of target data in transfer learning. Advances in Neural Information Processing Systems, 32.

Go beyond divergence

Cai and Wei (2021) also considers the setting of non-parametric classification, and proposes a notion called **reltative signal exponent** γ to describe the relation between source and target:

$$(\eta_{\mathbb{P}^{(1)}}(\boldsymbol{x}) - 1/2)(\eta_{\mathbb{P}^{(0)}}(\boldsymbol{x}) - 1/2) \geq 0,$$
$$|\eta_{\mathbb{P}^{(1)}}(\boldsymbol{x}) - 1/2| \geq C_\gamma |\eta_{\mathbb{P}^{(0)}}(\boldsymbol{x}) - 1/2|^\gamma,$$

where $\eta_{\mathbb{P}^{(k)}}(\boldsymbol{x}) = \mathbb{P}^{(k)}(Y = 1|X = \boldsymbol{x})$, $k = 0, 1$.

- This notion is mainly used to characterize the impact of posterior drift, because Cai and Wei (2021) imposes some conditions on $\mathbb{P}_X^{(0)}$ and $\mathbb{P}_X^{(1)}$ (e.g. share the same support, have bounded Lebesgue densities etc.) and does not focus on the distributions of covariate
- Smaller $\gamma \Rightarrow$ more information to transfer
- Target and source Bayes classifiers align with each other
- This notion is unrelated to the hypothesis class, which might be a drawback

---

[1] Cai, T. T., & Wei, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. The Annals of Statistics, 49(1).

Go beyond divergence

- At the end of the next lecture, we will introduce more ``metrics" beyond the symmetric divergence notions to characterize the covariate shift, which allows for the unbounded density ratio.
- For more about this line of research, see Redko et al. (2019, 2020).
- We have seen in some hardness results in $\S 2.4$ that terms in the bound we derived with symmetric divergences are **necessary**.

Question: Is this contradicted with the discussion in this section?

No! In general, the bound we derived with symmetric divergences might be tight, under the minimax sense, which is quite conservative. The development in Hanneke and Kpotufe (2019) and Cai and Wei (2021) shows that the symmetric divergence notions might be loose in some situations.

---

[1] Redko, I., Morvant, E., Habrard, A., Sebban, M., & Bennani, Y. (2019). Advances in domain adaptation theory. Elsevier.

[2] Redko, I., Morvant, E., Habrard, A., Sebban, M., & Bennani, Y. (2020). A survey on domain adaptation theory: learning bounds and theoretical guarantees. arXiv preprint arXiv:2004.11829.

[3] Hanneke, S., & Kpotufe, S. (2019). On the value of target data in transfer learning. Advances in Neural Information Processing Systems, 32.

[4] Cai, T. T., & Wei, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. The Annals of Statistics, 49(1).

References I

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

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2006). Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19.

Ben-David, S., Lu, T., Luu, T., and Pál, D. (2010b). Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129--136. JMLR Workshop and Conference Proceedings.

Ben-David, S. and Urner, R. (2012). On the hardness of domain adaptation and the utility of unlabeled target samples. In *Algorithmic Learning Theory: 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings 23*, pages 139--153. Springer.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Wortman, J. (2007). Learning bounds for domain adaptation. *Advances in neural information processing systems*, 20.

References II

Cai, T. T. and Wei, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *Annals of Statistics*, 49(1):100--128.

Crammer, K., Kearns, M., and Wortman, J. (2008). Learning from multiple sources. *Journal of Machine Learning Research*, 9(8).

Hanneke, S. and Kpotufe, S. (2019). On the value of target data in transfer learning. *Advances in Neural Information Processing Systems*, 32.

Kifer, D., Ben-David, S., and Gehrke, J. (2004). Detecting change in data streams. In *VLDB*, volume 4, pages 180--191. Toronto, Canada.

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2008). Domain adaptation with multiple sources. *Advances in neural information processing systems*, 21.

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. In *22nd Conference on Learning Theory, COLT 2009*.

Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. (2019). *Advances in domain adaptation theory*. Elsevier.

References III

Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. (2020). A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.