

Online Transfer Learning

Negative Transfer and Effect of Prior Knowledge

Xuetong Wu¹ Jonathan H. Manton¹ Uwe Aickelin² Jingge Zhu¹

¹Department of Electrical and Electronic Engineering

²Department of Computing and Information Systems
University of Melbourne, Australia

IEEE International Symposium on Information Theory
July, 2021



What is Transfer Learning?



Figure: Transfer learning of digit recognition

- Source **similar** Target $\rightarrow$ Successful!
- But if not executed properly, the source data will impair the performance on the target domains, known as **Negative Transfer**.
- The prior knowledge of the source and target domains (e.g., how different the source and target data (distributions) are) known in advance may be helpful for the prediction of target data.
- Further, target data may arrive sequentially.
- **Key Problem: How does the source data help prediction in the target data?**

Online Transfer Learning

- We studied a specific scenario such that introducing a batch of auxiliary (off-line) *source* data may help *target* sequential predictions.

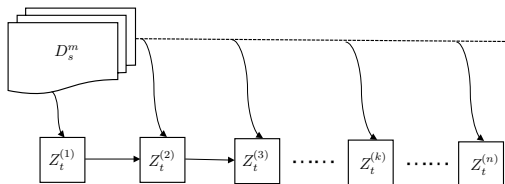


Figure: Online Transfer Learning

- **Problem:** How does the source data affect the learning performance (expected regret)?

Problem Formulation

- Parametric family: P_θ , where $\theta \in \Lambda$
- Target Data: $D_t^n = (Z_t^{(1)}, Z_t^{(2)}, \dots, Z_t^{(n)})$ **arrive sequentially.**
- Target Data distribution: $P_{\theta_t^*}(D_t^n) := \prod_{i=1}^n P_{\theta_t^*}(Z_t^{(i)})$.
- Source Data: $D_s^m = (Z_s^{(1)}, Z_s^{(2)}, \dots, Z_s^{(m)})$ **given in batch.**
- Source Data distribution: $P_{\theta_s^*}(D_s^m) := \prod_{i=1}^m P_{\theta_s^*}(Z_s^{(i)})$.
- Conditions: $\theta_s^*, \theta_t^* \in \Lambda$
- Prior knowledge: $d(\theta_s^*, \theta_t^*)$
- Expected Regret:

$$\mathcal{R}(n) := \mathbb{E}_{\theta_s^*, \theta_t^*} \left[\sum_{k=1}^n \ell(b_k, Z_t^{(k)}) - \sum_{k=1}^n \ell(b_k^*, Z_t^{(k)}) \right]$$

where b_k is learned from both source and historical target data till time $k - 1$.

Main Contributions

- Information-theoretic characterization of $\mathcal{R}(n)$.
- Asymptotic estimation of $\mathcal{R}(n)$ under *mixture strategy*.
- The asymptotic result regret is captured by the prior knowledge over θ_s^* and θ_t^* .
- Identify negative and positive transfer scenarios.
- Time-variant transfer learning and more extensions.
- Examples are demonstrated to validate our findings.

- Log-loss:

$$\ell(b, Z) = -\log(b(Z))$$

- Bounded loss:

$$|\ell(b, Z) - \ell(\hat{b}, Z)| \leq M$$

For any $b, \hat{b}$ and Z .

Mixture Strategy

Under log-loss,

$$\ell(b_k, Z_t^{(k)}) = -\log(Q(Z_t^{(k)} | D_s^m, D_t^{k-1}))$$

The expected regret can be written as,

$$\mathcal{R}(n) = \mathbb{E}_{\theta_s^*, \theta_t^*} \left[\log \frac{1}{Q(D_t^n | D_s^m)} - \log \frac{1}{P_{\theta_t^*}(D_t^n)} \right]$$

We choose the predictor $Q(D_t^n | D_s^m)$ as

$$\begin{aligned} Q(D_t^n | D_s^m) &= \frac{\int P_{\theta_t, \theta_s}(D_t^n, D_s^m) \omega(\theta_t, \theta_s) d\theta_t d\theta_s}{\int P_{\theta_s}(D_s^m) \omega(\theta_s) d\theta_s} \\ &= \int P_{\theta_t}(D_t^n) \omega(\theta_t | \theta_s) d\theta_t Q(\theta_s | D_s^m) d\theta_s \end{aligned}$$

Since $P_{\theta_s^*}$ and $P_{\theta_t^*}$ are unknown, we define Θ_s and Θ_t as random variables over Λ and assign a probability distribution ω over Θ_s and Θ_t . This particular choice is known as "mixture strategy".

Theorem (Regret with Log-loss)

Under log-loss,

$$\mathcal{R}_{\log}(n) = \mathbb{E}_{\theta_s^*, \theta_t^*} \left[\log \frac{P_{\theta_t^*}(D_t^n)}{Q(D_t^n | D_s^m)} \right] = I(D_t^n; \Theta_t = \theta_t^*, \Theta_s = \theta_s^* | D_s^m)$$

Theorem (Regret with Bounded Loss)

Under bounded loss,

$$\mathcal{R}(n) \leq M \sqrt{2n I(D_t^n; \Theta_t = \theta_t^*, \Theta_s = \theta_s^* | D_s^m)}.$$

Theorem (Asymptotic Estimation for Scalar Parameterization)

Under certain regularity, assume $\theta_s^*, \theta_t^* \in \Lambda \subset \mathbb{R}$ and $\theta_s^* \neq \theta_t^*$. As $n, m \rightarrow \infty$, the mixture strategy with proper prior $\omega(\Theta_s, \Theta_t)$ yields,

$$\mathcal{R}_{\log}(n) \approx \frac{1}{2} \log \frac{n}{2\pi e} + \frac{1}{2} \log I_t(\theta_t^*) + \log \frac{1}{\omega(\theta_t^* | \theta_s^*)}$$

Asymptotic Estimation:

$$\mathcal{R}_{\log}(n) \approx \frac{1}{2} \log \frac{n}{2\pi e} + \frac{1}{2} \log l_t(\theta_t^*) + \log \frac{1}{\omega(\theta_t^*|\theta_s^*)}$$

- Prior knowledge $\omega(\theta_t^*|\theta_s^*)$.
- Compared to the typical results without the source,

$$\mathcal{R}_{\log}(n) \approx \frac{1}{2} \log \frac{n}{2\pi e} + \frac{1}{2} \log l_t(\theta_t^*) + \log \frac{1}{\hat{\omega}(\theta_t^*)},$$

- Effect of the Prior?

Definition (Proper Prior)

Given a prior $\omega(\Theta_s, \Theta_t)$, we say,

- the induced marginal density $\omega(\Theta_s)$ is proper if it is continuous and positive over the whole support $\Lambda \subseteq \mathbb{R}^d$.
- the conditional density $\omega(\Theta_t|\Theta_s)$ is proper if there exist some $\delta_s > 0$ and $\delta_t > 0$ such that $\omega(\theta_t|\theta_s) > 0$ for any θ_s and θ_t satisfying $\|\theta_s - \theta_s^*\| \leq \delta_s$ and $\|\theta_t - \theta_t^*\| \leq \delta_t$.
- $\omega(\Theta_s, \Theta_t)$ is proper if $\omega(\Theta_s)$ and $\omega(\Theta_t|\Theta_s)$ are proper.

Negative Transfer

Let

$$\mathcal{R}_{\omega(\Theta_s, \Theta_t)}(n, m) \approx \frac{1}{2} \log \frac{n}{2\pi e} + \frac{1}{2} \log l_t(\theta_t^*) + \log \frac{1}{\omega(\theta_t^* | \theta_s^*)}$$

and

$$\mathcal{R}_{\hat{\omega}(\Theta_t)}(n) \approx \frac{1}{2} \log \frac{n}{2\pi e} + \frac{1}{2} \log l_t(\theta_t^*) + \log \frac{1}{\hat{\omega}(\theta_t^*)}$$

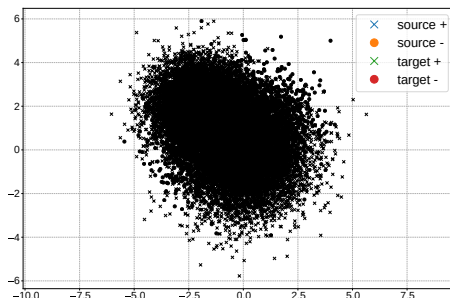
If $\omega(\Theta_t | \Theta_s)$ is improper, then for any proper $\hat{\omega}(\Theta_t)$, the following inequality holds when both n and m are sufficiently large,

$$\mathcal{R}_{\omega(\Theta_s, \Theta_t)}(n, m) > \mathcal{R}_{\hat{\omega}(\Theta_t)}(n).$$

In contrast, for any proper $\hat{\omega}(\Theta_t)$, there always exists a proper prior $\omega(\Theta_s, \Theta_t)$ that leads to the following inequality when both n and m are sufficiently large,

$$\mathcal{R}_{\omega(\Theta_s, \Theta_t)}(n, m) < \mathcal{R}_{\hat{\omega}(\Theta_t)}(n).$$

Logistic Regression Problem



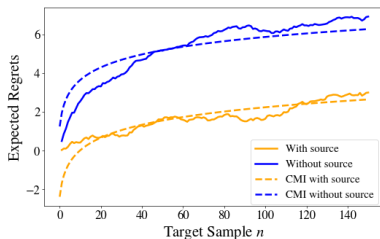
Bernoulli distribution with probability:

$$P_{\theta_t^*}(Y = 1) = \frac{1}{1 + e^{-\theta_t^* x}}, \quad P_{\theta_s^*}(Y = 1) = \frac{1}{1 + e^{-\theta_s^* x}}$$

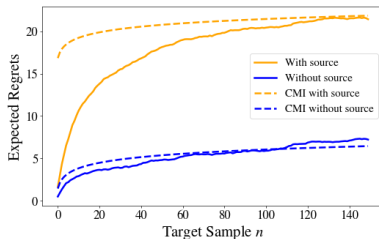
Loss function:

$$\ell(\theta, z_i) := -(y_i \log(\sigma(\theta^T x_i)) + (1 - y_i) \log(1 - \sigma(\theta^T x_i)))$$

Negative and Positive Transfer



(a) Positive Transfer



(b) Negative Transfer

Figure: The comparisons of the expected regret $\mathcal{R}(n)$ of the positive transfer with $\theta_s^* = [0.2, 0.4]$ (left) and the negative transfer (right) with $\theta_s^* = [0.8, 0.2]$ under the common settings where $\theta_t^* = [0.3, 0.5]$ and $c = 0.1$. The results are averaged over 200 experiments.

General Parameterization

$$\begin{aligned}\Theta_s &= (\Theta_{c,1}, \Theta_{c,2}, \dots, \Theta_{c,j}, \quad \Theta_{s,1}, \dots, \Theta_{s,d-j}) = (\Theta_c, \Theta_{sr}) \\ \Theta_t &= (\underbrace{\Theta_{c,1}, \Theta_{c,2}, \dots, \Theta_{c,j}}_{\text{common parameters}}, \quad \underbrace{\Theta_{t,1}, \dots, \Theta_{t,d-j}}_{\text{task-specific parameters}}) = (\Theta_c, \Theta_{tr})\end{aligned}$$

Theorem (Asymptotic Estimation for General Parameterization)

Under certain regularity and general parameterization, as $n, m \rightarrow \infty$, the mixture strategy with proper prior $\omega(\Theta_s, \Theta_t)$ yields,

$$\mathcal{R}_{\log}(n) \approx \frac{1}{2} \log \det(I_{j \times j} + \frac{n}{m} \Delta_t \Delta_s^{-1}) + \frac{1}{2} \log \det(n I_t(\theta_{tr}^*)) + \frac{d-j}{2} \log \frac{1}{2\pi e} + \log \frac{1}{\omega(\theta_t^* | \theta_s^*)}$$

- Instantaneous transfer learning
- Non-parametric transfer learning
- Efficient algorithms for posterior updating
- Time-evolving transfer learning
- etc...

Thank You!