



Minerva Access is the Institutional Repository of The University of Melbourne

**Author/s:**

Wu, X;Manton, JH;Aickelin, U;Zhu, J

**Title:**

Online Transfer Learning: Negative Transfer and Effect of Prior Knowledge

**Date:**

2021-07-12

**Citation:**

Wu, X., Manton, J. H., Aickelin, U. & Zhu, J. (2021). Online Transfer Learning: Negative Transfer and Effect of Prior Knowledge. IEEE International Symposium on Information Theory: Proceedings, 2021-July, pp.1540-1545. IEEE. <https://doi.org/10.1109/ISIT45174.2021.9517974>.

**Persistent Link:**

<https://hdl.handle.net/11343/283703>

# Online Transfer Learning: Negative Transfer and Effect of Prior Knowledge

Xuetong Wu<sup>1</sup>, Jonathan H. Manton<sup>1</sup>, Uwe Aickelin<sup>2</sup>, Jingge Zhu<sup>1</sup>

<sup>1</sup>Department of Electrical and Electronic Engineering

<sup>2</sup>Department of Computing and Information Systems

University of Melbourne

Parkville, Victoria, Australia

Email: xuetongw1@student.unimelb.edu, {jmanton, uwe.aickelin, jingge.zhu}.unimelb.edu.au

**Abstract**—Transfer learning is a machine learning paradigm where the knowledge from one task is utilized to resolve the problem in a related task. On the one hand, it is conceivable that knowledge from one task could be useful for solving a related problem. On the other hand, it is also recognized that if not executed properly, transfer learning algorithms could in fact impair the learning performance instead of improving it - commonly known as *negative transfer*. In this paper, we study the online transfer learning problems where the source samples are given in an off-line way while the target samples arrive sequentially. We define the expected regret of the online transfer learning problem, and provide upper bounds on the regret using information-theoretic quantities. We also obtain exact expressions for the bounds when the sample size becomes large. Examples show that the derived bounds are accurate even for small sample sizes. Furthermore, the obtained bounds give valuable insight on the effect of prior knowledge for transfer learning in our formulation. In particular, we formally characterize the conditions under which negative transfer occurs.

## I. INTRODUCTION

Transfer learning is a rising machine learning problem that leverages past knowledge in one or more *source* tasks to resolve the problem (or improve the performance) in a related *target* domain. The key problems are how to use the source intelligently to improve the performance in the target domain, and, how to characterize and avoid negative transfer. Currently, most existing transfer learning methods focus on offline settings where batch target data are available (see [1]–[3] and references therein). As such an assumption may not always hold in some real-time applications such as data transmission, we investigate the *online transfer learning* that is firstly proposed by [4]. Unlike traditional online learning, the framework of OTL is illustrated in Figure 1, where the decision is sequentially made with the aid of source data and historical target data. This framework has been extended to

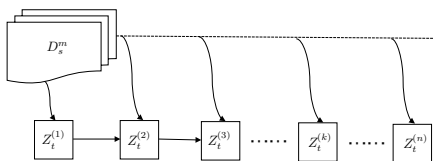


Fig. 1. Online Transfer Learning Framework

multi-task problem [7] and iterative domain adaptation [8]. Yet most of these works do not focus on rigorous theoretical analysis. Moreover, they mostly consider specific learning tasks such as binary classifications with linear models, and the performance is evaluated using a very specific metric (e.g., the number of mistakes). Such a learning framework in general does not exploit the structures (or distributions) of the data or model parameters. Lastly, no prior work has formally studied the problem of negative transfer.

In this work, we propose a more general framework for the online transfer problem that is suitable for general setups from the information-theoretic view. The information-theoretic framework has been established and studied in many online learning and reinforcement learning problems [9]–[14]. One advantage of this framework is that information-theoretic tools are quite useful in studying the asymptotic behaviors as well as deriving learning performance bounds for various statistical problems. This paper is inspired by the universal prediction framework [13]. By universal we mean that no matter which distribution the data are drawn from, the predictor will always yield good performance with theoretical guarantees. Specifically, we formulate the online transfer learning problem under the assumption that the source and target data distributions are parameterized by some unknown parameters  $\theta_s^*, \theta_t^* \in \Lambda$ . Then we define the expected regret and further propose the "mixture" strategy for sequential target predictors. The asymptotic upper bounds are also derived for the expected regret. Practically, the bound can be also applied to the typical transfer learning regime where the abundant source data are available but the target data are lacking. To conclude, the contributions of this paper are listed as follows.

- Consider the online transfer learning framework, at each time  $k$ , we propose the mixture strategy for the predictor  $b_k$  with the prior knowledge over source and target parameters. Then the expected regret is characterized by the conditional mutual information (CMI).
- We give an asymptotic estimation of CMI for  $\Lambda \subseteq \mathbb{R}^d$ , where the bound is captured by the prior knowledge, the number of common parameters that  $\theta_t^*$  and  $\theta_s^*$  share, and the structure of the parametric family. The results can be easily extended to time-variant target domains.

many other problems such as multisource transfer [5], [6],

- Based upon the general asymptotic bounds, we show that the inappropriate prior will lead to the negative transfer. That is, using data from the source domain would hurt the performance on the target domain. To our best knowledge, this is the first theoretical study on the negative transfer perspective of online transfer learning.
- The logistic regression example is examined and simple experiments confirm the effects of the prior and show that the improper prior will lead to negative transfer.

## II. PROBLEM FORMULATION AND MAIN RESULTS

Assume the source data  $D_s^m = (Z_s^{(1)}, \dots, Z_s^{(m)}) \in \mathcal{Z}^m$  are given in batch while the target data are received sequentially as  $Z_t^{(1)}, Z_t^{(2)}, \dots, Z_t^{(k)}, \dots$  where each sample takes value in  $\mathcal{Z}$ . Note that  $\mathcal{Z}$  can be discrete or continuous. At each time instant  $k$ , after having seen  $D_t^{k-1} = (Z_t^{(1)}, Z_t^{(2)}, \dots, Z_t^{(k-1)})$ , we predict  $Z_t^{(k)}$  using  $D_s^m$  and  $D_t^{k-1}$  with the predictor  $b_k : \mathcal{Z}^m \times \mathcal{Z}^{k-1} \rightarrow \hat{\mathcal{Z}}$ . Note that  $\hat{\mathcal{Z}}$  could be different from  $\mathcal{Z}$  in general, e.g.,  $\hat{\mathcal{Z}}$  is a quantized version of  $\mathcal{Z}$ . We define the loss function  $\ell : \hat{\mathcal{Z}} \times \mathcal{Z} \rightarrow \mathbb{R}^+$  that evaluates the prediction performance. In this paper, we use the convention that capital letters denote the random variables and small letters as their realizations. We further make the following assumptions.

**Assumption 1** (Parametric Distributions). *We assume that source and target data are generated independently in an i.i.d. fashion. More precisely, the joint distribution of the data sequence pairs  $P_{\theta_s^*, \theta_t^*}(D_t^n, D_s^m)$  can be written as*

$$P_{\theta_s^*, \theta_t^*}(D_t^n, D_s^m) = \prod_{i=1}^n P_{\theta_t^*}(Z_t^{(i)}) \prod_{j=1}^m P_{\theta_s^*}(Z_s^{(j)}), \quad (1)$$

where  $P_{\theta_t^*}$  and  $P_{\theta_s^*}$  are in a parametrized family of distributions  $\mathcal{P} = \{P_{\theta}\}_{\theta \in \Lambda}$ . Here  $\Lambda \subseteq \mathbb{R}^d$  is some measurable space and  $\theta_t^*$  and  $\theta_s^*$  are points in the interior of  $\Lambda$ .

After observing  $n$  target samples, we want to minimise the corresponding expected regret defined as,

$$\mathcal{R}(n) := \mathbb{E}_{\theta_s^*, \theta_t^*} \left[ \sum_{k=1}^n \ell(b_k, Z_t^{(k)}) - \sum_{k=1}^n \ell(b_k^*, Z_t^{(k)}) \right], \quad (2)$$

where  $b_1$  is learned from  $D_s^m$  and  $b_k (k > 1)$  is the decision we made based on both  $D_s^m$  and  $D_t^{k-1}$  but without the knowledge of  $\theta_s^*$  and  $\theta_t^*$ . The predictor  $b_k^*$  is the optimal decision at each time  $k$  that can depend on true target distributions  $P_{\theta_t^*}$ . If not otherwise specified, the notation  $\mathbb{E}_{\theta_s, \theta_t}[\cdot]$  (similar to  $\mathbb{E}_{\theta_t}[\cdot]$  and  $\mathbb{E}_{\theta_s}[\cdot]$ ) means the expectation is taken over all source and target samples that are drawn from  $P_{\theta_s}$  and  $P_{\theta_t}$ .

### A. Expected Regret Bounds

Given the above problem formulation, we begin by considering the *logarithm loss* defined as follows.

**Definition 1** (Logarithm Loss). *Let the predictor  $b_k$  be a probability distribution over the sample  $z_t^{(k)}$  at time  $k$ , the logarithm loss is defined as*

$$\ell(b_k, z_t^{(k)}) = -\log b_k(z_t^{(k)}). \quad (3)$$

At each time  $k$ , we may view the predictor as a conditional probability distribution  $b_k(z_t^{(k)}) = Q(z_t^{(k)} | D_t^{k-1}, D_s^m)$ , and

the optimal predictor  $b_k^*$  is naturally given by the true target distribution over  $z_t^{(k)}$  as  $b_k^*(z_t^{(k)}) = P_{\theta_t^*}(z_t^{(k)})$ . Then the expected regret till time  $n$  can be written explicitly as

$$\mathcal{R}_{\log}(n) = \mathbb{E}_{\theta_s^*, \theta_t^*} \left[ \log \frac{1}{Q(D_t^n | D_s^m)} - \log \frac{1}{P_{\theta_t^*}(D_t^n)} \right]. \quad (4)$$

The effect of source data is reflected in the conditional distribution  $Q(D_t^n | D_s^m)$ . Concerning the choice of  $Q$ , we first define  $\Theta_s$  and  $\Theta_t$  as random variables over  $\Lambda$ . Since  $P_{\theta_s^*}$  and  $P_{\theta_t^*}$  are unknown, we assign a probability distribution  $\omega$  over  $\Theta_s$  and  $\Theta_t$  w.r.t. the Lebesgue measure to represent our prior knowledge and update the posterior with the incoming data to approximate the underlying distributions, which is known as the *mixture strategy* [13], [15]. In particular, we choose the predictor  $Q(D_t^n | D_s^m)$  as

$$\begin{aligned} Q(D_t^n | D_s^m) &= \frac{\int P_{\theta_t, \theta_s}(D_t^n, D_s^m) \omega(\theta_t, \theta_s) d\theta_t d\theta_s}{\int P_{\theta_s}(D_s^m) \omega(\theta_s) d\theta_s} \\ &= \int P_{\theta_t}(D_t^n) \omega(\theta_t | \theta_s) d\theta_t Q(\theta_s | D_s^m) d\theta_s, \end{aligned} \quad (5)$$

where  $\omega(\theta_s)$  is the marginal of  $\omega(\theta_s, \theta_t)$ . From Eq (5), the mixture strategy quantitatively explains how transfer learning is implemented via the posterior updates of  $\theta_t$  from a Bayesian perspective. Intuitively speaking, the posterior  $Q(\theta_s | D_s^m)$  firstly gives an estimate of  $\theta_s^*$  from the source data, then the conditional prior  $\omega(\theta_t | \theta_s)$  reflects our belief upon  $\theta_t^*$  given  $\theta_s$  estimated from source data. With the choice of  $Q(D_t^n | D_s^m)$ , the expected regret can be explicitly characterized in the following theorem.

**Theorem 1** (Regret with Log-loss). *With the mixture strategy  $Q(D_t^n | D_s^m)$ , the expected regret in Eq (4) can be written as*

$$\mathcal{R}_{\log}(n) = \mathbb{E}_{\theta_s^*, \theta_t^*} \left[ \log \frac{P_{\theta_t^*}(D_t^n)}{Q(D_t^n | D_s^m)} \right] = I(D_t^n; \Theta_t = \theta_t^*, \Theta_s = \theta_s^* | D_s^m), \quad (6)$$

where  $I(D_t^n; \Theta_t = \theta_t^*, \Theta_s = \theta_s^* | D_s^m)$  denotes the conditional mutual information  $I(D_t^n; \Theta_t, \Theta_s | D_s^m)$  evaluated at  $\Theta_t = \theta_t^*, \Theta_s = \theta_s^*$ .

All proofs in this paper can be found in [16]. In many cases, we need to consider the task-specific loss such as squared loss, 0-1 loss and hinge, etc. For other general bounded loss function  $\ell$ , we define the predictor at time  $k$  to be,

$$b_k = \underset{b}{\operatorname{argmin}} \mathbb{E}_{Q(D_t^k, D_s^m)} \left[ \ell(b, z_t^{(k)}) | D_s^m, D_t^{k-1} \right], \quad (7)$$

with the choice of the mixture strategy  $Q(D_t^k, D_s^m) = \int P_{\theta_t, \theta_s}(D_t^k, D_s^m) \omega(\theta_t, \theta_s) d\theta_t d\theta_s$  for some prior  $\omega$ . The optimal predictor is naturally given by,

$$b_k^* = \underset{b}{\operatorname{argmin}} \mathbb{E}_{P_{\theta_t^*}(D_t^k)} \left[ \ell(b, z_t^{(k)}) | D_t^{k-1} \right]. \quad (8)$$

As a consequence, we arrive at the following theorem.

**Theorem 2** (Bounds on General Loss). *Assume the loss function satisfies  $|\ell(b, z) - \ell(b^*, z)| \leq M$  for any observation  $z$  and the predictors  $b, b^*$ . Then the true expected regret induced by  $b_k$  and  $b_k^*$  in Eq (7) and (8) can be bounded as,*

$$\mathcal{R}(n) \leq M \sqrt{2n I(D_t^n; \Theta_t = \theta_t^*, \Theta_s = \theta_s^* | D_s^m)}. \quad (9)$$

The above theorem implies that if the loss function is bounded, with a certain prior  $\omega$ , the regret induced by the

mixture strategy is also captured by CMI evaluated at  $\Theta = \theta_t^*$  and  $\Theta_s = \theta_s^*$ . However, the bound in its current form is less informative as it does not show what is the effect of the prior  $\omega$  and sample size  $m$  and  $n$ .

### B. Asymptotic Analysis of CMI

To further investigate the effect of prior, we give an asymptotic analysis of CMI. First we make the regular assumptions on parametric conditions [17], [18] and define the proper prior.

**Assumption 2** (Parametric Condition). *Assume the source and target distributions  $P_{\theta_s^*}(Z_s)$  and  $P_{\theta_t^*}(Z_t)$  are twice continuously differentiable at  $\theta_s^*$  and  $\theta_t^*$  for almost every  $Z_s$  and  $Z_t$ . For any  $\theta_t, \theta_s \in \Lambda$ , there exist  $\delta_s, \delta_t > 0$  satisfying,*

$$\mathbb{E}_{\theta_t} \left[ \sup_{\|\theta_t - \theta_t^*\| \leq \delta} \left| \frac{\partial}{\partial \theta_{t,i}} \log P_{\theta_t}(Z_t) \right| \right] < \infty \quad (10)$$

$$\mathbb{E}_{\theta_s} \left[ \sup_{\|\theta_s - \theta_s^*\| \leq \delta} \left| \frac{\partial}{\partial \theta_{s,i}} \log P_{\theta_s}(Z_s) \right| \right] < \infty \quad (11)$$

for  $i = 1, \dots, d$ . In addition, we assume,

$$\mathbb{E}_{\theta_t} \left[ \sup_{\|\theta_t - \theta_t^*\| \leq \delta} \left| \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log P_{\theta_t}(Z_t) \right|^2 \right] < \infty \quad (12)$$

$$\mathbb{E}_{\theta_s} \left[ \sup_{\|\theta_s - \theta_s^*\| \leq \delta} \left| \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log P_{\theta_s}(Z_s) \right|^2 \right] < \infty \quad (13)$$

for any  $i, j = 1, \dots, d$ .

**Definition 2** (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  $\theta_s$  and  $\theta_t$  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.

We also define the proper prior without the source as having the continuous density  $\hat{\omega}(\theta_t) > 0$  over the whole support  $\Lambda$ .

If the distributions in parametric family  $\mathcal{P}_\theta$  satisfy the Assumption 2 (e.g., the exponential families in [19]), with the proper prior, we ensure that the posterior distribution of  $\Theta_t$  and  $\Theta_s$  given  $D_t^n$  and  $D_s^m$  asymptotically concentrates on neighborhoods of  $\theta_t^*$  and  $\theta_s^*$ , respectively. With definitions in place, for the case where both  $\Theta_s$  and  $\Theta_t$  are scalars, we give the asymptotic estimation for CMI as follows.

**Theorem 3** (Asymptotic Estimation of CMI). *Under Assumptions 1 and 2, for  $\Lambda = \mathbb{R}$  and  $\theta_s^* \neq \theta_t^*$ , as  $n, m \rightarrow \infty$ , the mixture strategy with proper prior  $\omega(\Theta_s, \Theta_t)$  yields,*

$$I(D_t^n; \Theta_t = \theta_t^*, \Theta_s = \theta_s^* | D_s^m) - \frac{1}{2} \log \frac{n}{2\pi e} \rightarrow \frac{1}{2} \log I_t(\theta_t^*) + \log \frac{1}{\omega(\theta_t^*|\theta_s^*)} \quad (14)$$

where we define the Fisher information  $\mathbb{E}_{\Theta_t} [-\nabla_{\Theta_t}^2 \log P_{\Theta_t}(Z_t)]$  evaluated at  $\Theta_t = \theta_t^*$  as  $I_t(\theta_t^*)$ .

**Remark 1.** *Compared to the result without the source data when target sample is abundant [18],*

$$I(D_t^n; \Theta_t = \theta_t^*) - \frac{1}{2} \log \frac{n}{2\pi e} \rightarrow \frac{1}{2} \log I_t(\theta_t^*) + \log \frac{1}{\hat{\omega}(\theta_t^*)} \quad (15)$$

for some prior  $\hat{\omega}(\Theta_t)$ , the difference between Eq (14) and (15) is  $\frac{\hat{\omega}(\theta_t^*)}{\omega(\theta_t^*|\theta_s^*)}$ . It says that if the distribution  $\omega$  can be chosen such that  $\frac{\hat{\omega}(\theta_t^*)}{\omega(\theta_t^*|\theta_s^*)} < 1$ , the source data will help to reduce the regret. However, it should be noted that  $\omega$  is chosen without knowing the exact value of  $\theta_t^*$  and  $\theta_s^*$  so it is not immediately clear if this is always possible. We will show later that if the conditional prior  $\omega(\theta_t|\theta_s)$  is proper, it is always possible to find a distribution such that  $\frac{\hat{\omega}(\theta_t^*)}{\omega(\theta_t^*|\theta_s^*)} < 1$ . On the contrary, if the prior information between the source and target is incorrect, we may always end up with  $\frac{\hat{\omega}(\theta_t^*)}{\omega(\theta_t^*|\theta_s^*)} > 1$ , which is one way to interpret **negative transfer**.

**Remark 2.** *Notice that the source samples change the constant from  $\log \frac{1}{\hat{\omega}(\theta_t)}$  to  $\log \frac{1}{\omega(\theta_t|\theta_s)}$ , which are independent from  $n$ . Hence the effect of the source samples vanishes asymptotically as  $n$  goes to infinity. However, the asymptotic analysis is still useful for two reasons. Firstly, we will show later that when both  $n$  and  $m$  approach infinity, the sample complexity of the regret (i.e., how regret scales in terms of  $m$  and  $n$ ) can change, depending on how fast  $m$  and  $n$  grow relative to each other. Secondly, our numerical results show that the asymptotic bound is in fact very accurate even for relatively small  $m$  and  $n$ .*

Theorem 3 holds when the distributions are parametrized by scalars. We extend to a more typical transfer learning scenario where  $\Theta_t, \Theta_s \in \mathbb{R}^d$  with  $d > 1$  share some common parameters  $\Theta_c \in \mathbb{R}^j$  for  $0 \leq j \leq d$ . To illustrate, we can write the parameters in the following way.

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

where  $\Theta_c \in \mathbb{R}^j$  denotes the common parameter vector and  $\Theta_{sr}, \Theta_{tr} \in \mathbb{R}^{d-j}$  are task-specific parameter vectors. Then we reach the following theorem that gives the asymptotic normality of the conditional mutual information with  $d > 1$ .

**Theorem 4** (Asymptotic Estimation for General Parametrization). *Under Assumptions 1 and 2, with  $\Theta_s, \Theta_t \in \mathbb{R}^d$  defined above and as  $n, m \rightarrow \infty$ , the mixture strategy with proper prior  $\omega(\Theta_s, \Theta_t)$  yields,*

$$\begin{aligned} I(D_t^n; \Theta_t = \theta_t^*, \Theta_s = \theta_s^* | D_s^m) - \frac{1}{2} \log \det(\mathbf{I}_{j \times j} + \frac{n}{m} \Delta_t \Delta_s^{-1}) \\ - \frac{1}{2} \log \det(nI_t(\theta_{tr}^*)) \rightarrow \frac{d-j}{2} \log \frac{1}{2\pi e} + \log \frac{1}{\omega(\theta_t^*|\theta_s^*)}, \quad (16) \end{aligned}$$

where  $\Delta_s = I_{cs}(\theta_c^*) - I_{cs}(\theta_c^*, \theta_{sr}^*) I_{sr}^{-1}(\theta_{sr}^*) I_{cs}^T(\theta_c^*, \theta_{sr}^*)$  and  $\Delta_t = I_{ct}(\theta_c^*) - I_{ct}(\theta_c^*, \theta_{tr}^*) I_{tr}^{-1}(\theta_{tr}^*) I_{ct}^T(\theta_c^*, \theta_{tr}^*)$ ,  $\mathbf{I}_{j \times j}$  denotes the identity matrix with size  $j$  and  $\theta^* = (\theta_c^*, \theta_{sr}^*, \theta_{tr}^*)$  denotes the true parameters. With a little abuse of notation, we define the fisher information matrix as

$$\begin{aligned} I_{cs}(\theta_c^*) &= -\mathbb{E}_{\theta_s^*} [\nabla_{\Theta_c}^2 \log P(Z_s | \Theta_c, \theta_{sr}^*)] \Big|_{\Theta_c = \theta_c^*} \in \mathbb{R}^{j \times j} \\ I_{ct}(\theta_c^*) &= -\mathbb{E}_{\theta_t^*} [\nabla_{\Theta_c}^2 \log P(Z_t | \Theta_c, \theta_{tr}^*)] \Big|_{\Theta_c = \theta_c^*} \in \mathbb{R}^{j \times j} \\ I_{s}(\theta_{sr}^*) &= -\mathbb{E}_{\theta_s^*} [\nabla_{\Theta_{sr}}^2 \log P(Z_s | \theta_c^*, \Theta_{sr})] \Big|_{\Theta_{sr} = \theta_{sr}^*} \in \mathbb{R}^{(d-j) \times (d-j)} \\ I_{t}(\theta_{tr}^*) &= -\mathbb{E}_{\theta_t^*} [\nabla_{\Theta_{tr}}^2 \log P(Z_t | \theta_c^*, \Theta_{tr})] \Big|_{\Theta_{tr} = \theta_{tr}^*} \in \mathbb{R}^{(d-j) \times (d-j)} \end{aligned}$$

$$I_{cs}(\theta_c^*, \theta_{sr}^*) = -\mathbb{E}_{\theta_s^*} \left[ \frac{\partial \log P(Z_s | \theta_c^*, \theta_{sr}^*)}{\partial \Theta_{c,i} \partial \Theta_{sr,k}} \right]_{\substack{i=1, \dots, j, \\ k=1, \dots, d-j}} \in \mathbb{R}^{j \times (d-j)}$$

$$I_{ct}(\theta_c^*, \theta_{tr}^*) = -\mathbb{E}_{\theta_t^*} \left[ \frac{\partial \log P(Z_t | \theta_c^*, \theta_{tr}^*)}{\partial \Theta_{c,i} \partial \Theta_{tr,k}} \right]_{\substack{i=1, \dots, j, \\ k=1, \dots, d-j}} \in \mathbb{R}^{j \times (d-j)}$$

**Remark 3.** In the above expression, we can intuitively interpret the term  $\frac{1}{2} \log \det(\mathbf{I}_{j \times j} + \frac{n}{m} \Delta_t \Delta_s^{-1})$  as the "learning cost" of  $\theta_c$ , which is captured by the ratio  $\frac{n}{m}$ . If  $m$  is

- sublinear in  $n$ , the rate is  $O(\log(nj))$  and source samples do not improve the learning performance
- linear in  $n$ , the cost reduces to  $O(1)$ .
- superlinear in  $n$ , the rate is  $o(1)$ , and abundant source samples indeed improve the performance and the cost vanishes in this case.

While the learning cost of  $\theta_{tr}$  is relied on  $\frac{1}{2} \log \det(nI_t(\theta_{tr}^*))$  and the prior  $\omega(\theta_t^* | \theta_s^*)$ , whereas the prior knowledge can only change the constant but does not change the rate.

**Remark 4.** As a special case, if there is no common parameters ( $j = 0$ ), then as both  $m$  and  $n$  are sufficiently large,

$$I(D_t^n; \Theta_t = \theta_t^*, \Theta_s = \theta_s^* | D_s^m) - \frac{d}{2} \log \frac{n}{2\pi e} \rightarrow \frac{1}{2} \log \det(I_t(\theta_{tr}^*)) + \log \frac{1}{\omega(\theta_t^* | \theta_s^*)}.$$

Let  $d = 1$ , we can recover the results in Theorem 3 and the knowledge transfer is only reflected on the prior knowledge  $\omega(\theta_t^* | \theta_s^*)$ . If the number of the common parameters is  $d$  ( $j = d$ ), that is, the source and target distributions are characterized by the same parameters, which yields the asymptotic estimation as,

$$I(D_t^n; \Theta_t = \theta_t^*, \Theta_s = \theta_s^* | D_s^m) - \frac{1}{2} \log \det(\mathbf{I}_{d \times d} + \frac{n}{m} I_{ct}(\theta_c^*) I_{cs}^{-1}(\theta_c^*)) \rightarrow \log \frac{1}{\omega(\theta_t^* | \theta_s^*)}.$$

Under this case, the regret depends on the ratio  $\frac{n}{m}$  and prior  $\omega(\theta_t^* | \theta_s^*)$  as discussed in Remark 3.

### C. Time-variant Target Domains

In the above problem, we assume that the target parameter  $\theta_t^*$  stays fixed for all time. However, in some applications, the target distribution may change over time, and this motivates us to consider the time-variant transfer learning scenarios. Let the time evolving target data be parameterized by  $\theta_{t,l}^*$  where at each index  $l \in \mathbb{N}^+$ , we will receive  $n_l$  target samples  $Z_{t,l}^{(i)}$  drawn from the distribution  $P_{\theta_{t,l}^*}$ . It is common to assume that  $\theta_{t,l}^*$  only depends on the previous parameter  $\theta_{t,l-1}^*$ . At index  $k$ , we are interested in minimising the expected regret

$$\mathcal{R}(k) = \sum_{l=1}^k \mathbb{E}_{\theta_s^*, \theta_{t,l}^*, \theta_{t,l-1}^*} \left[ \sum_{i=1}^{n_l} \ell(b_i, Z_{t,l}^{(i)}) - \sum_{i=1}^{n_l} \ell(b_i^*, Z_{t,l}^{(i)}) \right]. \quad (17)$$

Here  $b_i$  is chosen to be the mixture strategy over  $\theta_s^*, \theta_{t,l}^*$ , and  $\theta_{t,l-1}^*$ . Combining Theorem 2 and 4, one can easily reach the asymptotic estimation of the expected regret.

**Theorem 5** (Time-variant Target Regret Bounds). *Given the time-variant target domain described above, suppose that conditions in Theorem 2 and Assumptions 1 and 2 hold for each  $\theta_{t,k}^*$  and  $\theta_s^*$ . For  $l = 1, 2, \dots, k$ , we further assume that*

source parameters will share  $j$  parameters with every  $\theta_{t,l}^*$ , and  $\theta_{t,l}^*, \theta_{t,l-1}^*$  have  $c_l$  common parameters. As  $n_l, m \rightarrow \infty$ , the mixture strategy with proper prior  $\omega(\theta_s, \theta_{t,l}, \theta_{t,l-1})$  yields,

$$\begin{aligned} \mathcal{R}(k) \leq & M \left( k \sum_{l=1}^k n_l \left( \log \det \left( \mathbf{I}_{j \times j} + \frac{n_l}{m + n_{l-1}} \Delta_{ct} \Delta_{cs}^{-1} \right) \right. \right. \\ & + \log \det(\mathbf{I}_{c_l \times c_l} + \frac{n_l}{n_{l-1}} \Delta_t \Delta_{t-1}^{-1}) + \log \det(n_l I_{t,l}(\theta_{tr,l}^*)) \\ & \left. \left. + (d - j - c_l) \log \frac{1}{2\pi e} + \frac{2}{\omega(\theta_{t,l}^* | \theta_{t,l-1}^*, \theta_s^*)} \right) \right)^{\frac{1}{2}}. \end{aligned}$$

In this case, the prior knowledge  $\omega(\theta_{t,l}^* | \theta_{t,l-1}^*, \theta_s^*)$  and the common parameters among which determine the prediction performance. Due to the space limit, we omit some analogous definitions and settings here, but readers can refer to the supplementary proof [16] for more details and insights.

### D. Improper Prior and Negative Transfer

As previously discussed,  $\omega(\Theta_s, \Theta_t)$  should be chosen properly so that the posterior updating will asymptotically converge to the true parameter  $\theta_s^*$  and  $\theta_t^*$ . However, if the prior distribution (particularly  $\omega(\Theta_t | \Theta_s)$ ) is imposed improperly, the extra source data do not necessarily mean that our prediction for target data can always be improved. Roughly speaking, if our prior knowledge on  $\theta_s^*$  and  $\theta_t^*$  is incorrect, under our scheme, this would translate to an improper prior distribution for the mixture strategy. We will show that with an improper prior, the extra source data will in fact cause a higher regret (i. e. worse performance) compared to the case without source data.

**Proposition 1** (Negative Transfer). *Let  $\mathcal{R}_{\omega(\Theta_s, \Theta_t)}(n, m)$  denote the regret induced by the mixture strategy  $Q(D_t^n | D_s^m)$  with the prior  $\omega(\Theta_s, \Theta_t)$  and  $\mathcal{R}_{\hat{\omega}(\Theta_t)}(n)$  denote the regret induced by  $\hat{Q}(D_t^n)$  with the prior  $\hat{\omega}(\Theta_t)$ . If  $\omega(\Theta_t | \Theta_s)$  is improper<sup>1</sup>, then for any proper  $\hat{\omega}(\Theta_t)$ , the following inequality holds when both  $n$  and  $m$  are sufficiently large,*

$$\mathcal{R}_{\omega(\Theta_s, \Theta_t)}(n, m) > \mathcal{R}_{\hat{\omega}(\Theta_t)}(n). \quad (18)$$

**Proof Sketch.** • By subtraction, we need to prove that,

$$\mathbb{E}_{\theta_s^*, \theta_t^*} \left[ \log \frac{Q(D_s^m) \hat{Q}(D_t^n)}{Q(D_t^n, D_s^m)} \right] > 0,$$

- Let us examine the logarithm term in the expectation as,

$$\log \frac{Q(D_s^m) \hat{Q}(D_t^n)}{Q(D_t^n, D_s^m)} = \log \frac{1}{\int \int \hat{Q}(\theta_t | D_t^n) \frac{\omega(\theta_t | \theta_s)}{\hat{\omega}(\theta_t)} d\theta_t Q(\theta_s | D_s^m) d\theta_s}.$$

- It can be found that the difference is characterized by the ratio  $\frac{\omega(\theta_t | \theta_s)}{\hat{\omega}(\theta_t)}$  and improper  $\omega$  leads to zero mass near  $\theta_t^*$  compared to proper  $\hat{\omega}$ , thus a higher regret.

For example, let  $\mathcal{Z} = \{0, 1\}$  and assume  $\Theta_s$  and  $\Theta_t$  are the probabilities that the source and target samples take value in 1. Also assume that our (incorrect) prior knowledge on the parameters is that  $|\theta_s - \theta_t| \leq 0.1$  given any  $\theta_s \in \Lambda$ . Suppose the true underlying parameters are  $\theta_t^* = 0.6$  and

<sup>1</sup>We say  $\omega(\Theta_t | \Theta_s)$  is improper if it does not satisfy conditions in Def 2.

$\theta_s^* = 0.8$ . In this case, even if knowing  $\theta_s^*$  precisely, one can never end up with the correct estimation for  $\theta_t^*$  even with abundant target samples if the prior  $\omega(\theta_t|\theta_s)$  is improper. As a consequence, the regret becomes higher compared to the case without knowing such prior. For those who are interested in detailed analysis, we refer to [16] for more theoretical and experimental results.

In contrast, if  $\omega(\Theta_s, \Theta_t)$  is chosen properly, we can always find a prior such that the knowledge transfer from source data encourages lower regret, namely, the positive transfer.

**Proposition 2** (Positive Transfer). *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,*

$$\mathcal{R}_{\omega(\Theta_s, \Theta_t)}(n, m) < \mathcal{R}_{\hat{\omega}(\Theta_t)}(n). \quad (19)$$

In our claim, we can always find a proper prior  $\omega(\Theta_s, \Theta_t)$  whose marginal  $\omega(\Theta_t|\Theta_s)$  encourages a tighter support over  $\Theta_t$ . In other words, making use of source data appropriately can narrow down the uncertainty range over  $\Theta_t$ . It then follows that such prior assigns more concentrated mass around  $\theta_t^*$ , which reduces the expected regret.

### III. EXAMPLES

Consider a logistic regression problem in a 2-dimensional space. For the given parameter  $\theta \in [0, 1]^2$  and  $Z_i = (X_i, Y_i) \in \mathbb{R}^2 \times \{0, 1\}$ , each label  $Y_i \in \{0, 1\}$ , is generated from a Bernoulli distribution with probability  $p(Y_i = 1) = \frac{1}{1 + e^{-\theta^T X_i}}$ . Suppose that the source and target input features  $X_s^{(k)}$  and  $X_t^{(k)}$  are drawn from the same normal distribution  $\mathcal{N}\left(\begin{bmatrix} 5 \\ -5 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$ . The loss function is then given by

$$\ell(\theta, Z_i) := -(Y_i \log(\sigma(\theta^T X_i)) + (1 - Y_i) \log(1 - \sigma(\theta^T X_i))),$$

where  $\sigma(x) = \frac{1}{1 + e^{-x}}$ . Let  $\theta_t^* = (0.3, 0.5)$  and  $\theta_s^* = (0.2, 0.4)$  denote the true parameters for the target and source domains. Given  $m = 5000$ , let the marginal prior  $\omega(\Theta_s)$  be uniformly distributed over  $[0, 1]^2$  and our prior knowledge  $\omega(\Theta_t|\Theta_s)$  assumes that  $\Theta_t$  is normally distributed with the mean of  $\Theta_s$  and covariance of  $\begin{bmatrix} c^2 & 0 \\ 0 & c^2 \end{bmatrix}$ , here  $c$  represents the prior belief on  $\Theta_t$  such that smaller  $c$  implies  $\Theta_t$  is closer to  $\Theta_s$  and vice versa. To show the usefulness of the source data, we compare with the target only case ( $m = 0$ ) where we assume the prior  $\hat{\omega}(\Theta_t)$  is uniformly distributed over  $[0, 1]^2$ .

After receiving  $n$  target samples, we plot different posteriors to see the effect of the mixture strategy induced by the chosen prior. From Figure 2, given sufficient source data, the posterior of  $\Theta_s$  will give a precise estimation of  $\theta_s^*$  and the density will mostly concentrate around  $[0.2, 0.4]$ . While there is a lack of target samples ( $n$  is small), the posterior  $\hat{Q}(\Theta_t|D_t^n)$  without the source is relatively scattered and the density around  $\theta_t^*$  is quite low. On the contrary, with the prior knowledge  $\omega(\Theta_t|\Theta_s)$  and small  $c = 0.1$ , the posterior  $Q(\Theta_t|D_s^m, D_t^n)$  will be concentrated more around  $\theta_t^*$  as source and target parameters

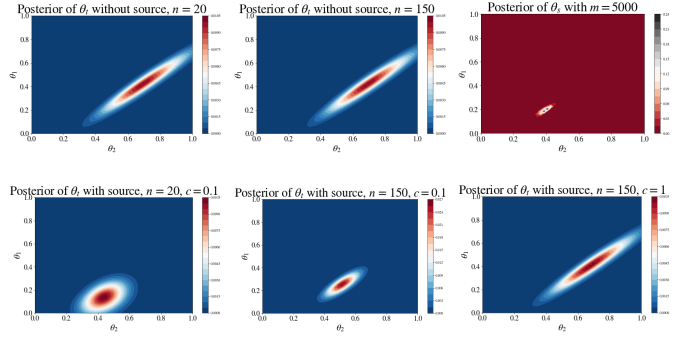


Fig. 2. The posterior of  $\theta_s$  and  $\theta_t$  given  $D_s^m$  and  $D_t^n$  under different prior belief  $c$  and target sample size  $n$

are particularly close. When  $c$  increases to 1, the source data is no longer helpful as  $\Theta_t$  is roughly distributed uniformly on  $[0, 1]^2$  and the posterior behaves similarly to target only case.

To further demonstrate our theoretical results, we plot the expected regrets in Figure 3 for positive and negative transfer cases, and we also plot the asymptotic estimation of CMI in dashed lines from Theorem 1 and 4 to numerically evaluate the difference. From the left figure, it is observed that introducing the source indeed yields lower regret, which fits our intuition from the posteriors. Even for small  $n$  ( $\approx 40$ ), CMI captures the regret quite well and the gap is roughly  $\log \frac{\omega(\theta_t^*|\theta_s^*)}{\omega(\theta_t^*)}$  as noted in Remark 1. In contrast, we also examine the negative transfer case with  $\theta_s^* = [0.8, 0.2]$  where the results are shown in the right figure. With this specific choice of  $\theta_s^*$ , the prior distribution  $\omega(\theta_t^*|\theta_s^*)$  in this case has an extremely low density and the estimation will hardly approach the true parameters. As a result, the negative transfer happens and source samples will hurt the performance instead. It also appears that CMI captures this trend well when  $n$  goes reasonably large ( $\approx 80$ ).

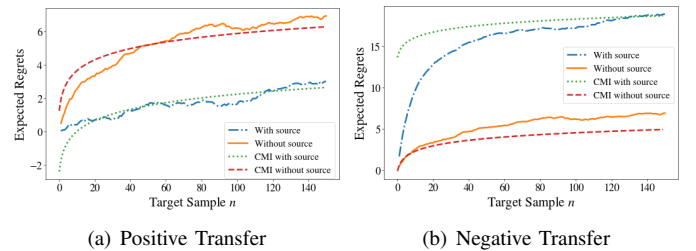


Fig. 3. The comparisons of the expected regret  $\mathcal{R}(n)$  of the positive transfer with  $\theta_s^* = [0.2, 0.4]$  (left) and the negative transfer (right) with  $\theta_s^* = [0.8, 0.2]$  under the common settings where  $\theta_t^* = [0.3, 0.5]$  and  $c = 0.1$ . The results are averaged over 200 experiments.

Overall, in both positive and negative transfer cases, the gaps between the regrets are mainly reflected on the prior knowledge  $\omega(\theta_t^*|\theta_s^*)$  when  $n$  is reasonably large as mentioned in Remark 1 and 3, which experimentally confirms Theorem 4. Moreover, it shows that the asymptotic bounds are still reasonably accurate in the case when  $n$  and  $m$  are small.

## REFERENCES

- [1] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
- [2] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, p. 9, 2016.
- [3] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, 2020.
- [4] P. Zhao, S. C. Hoi, J. Wang, and B. Li, "Online transfer learning," *Artificial Intelligence*, vol. 216, pp. 76–102, 2014.
- [5] Q. Wu, X. Zhou, Y. Yan, H. Wu, and H. Min, "Online transfer learning by leveraging multiple source domains," *Knowledge and Information Systems*, vol. 52, no. 3, pp. 687–707, 2017.
- [6] Z. Kang, B. Yang, S. Yang, X. Fang, and C. Zhao, "Online transfer learning with multiple source domains for multi-class classification," *Knowledge-Based Systems*, vol. 190, p. 105149, 2020.
- [7] J. He and R. Lawrence, "A graphbased framework for multi-task multi-view learning," in *ICML*, 2011, pp. 25–32.
- [8] H. S. Bhatt, A. Rajkumar, and S. Roy, "Multi-source iterative adaptation for cross-domain classification," in *IJCAI*, 2016, pp. 3691–3697.
- [9] A. Lazaric, M. Restelli, and A. Bonarini, "Transfer of samples in batch reinforcement learning," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 544–551.
- [10] A. Lazaric, "Transfer in reinforcement learning: a framework and a survey," in *Reinforcement Learning*. Springer, 2012, pp. 143–173.
- [11] Y. Zhan and M. E. Taylor, "Online transfer learning in reinforcement learning domains," *arXiv preprint arXiv:1507.00436*, 2015.
- [12] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *Journal of Machine Learning Research*, vol. 10, no. 7, 2009.
- [13] N. Merhav and M. Feder, "Universal prediction," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.
- [14] D. Russo and B. Van Roy, "An information-theoretic analysis of thompson sampling," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2442–2471, 2016.
- [15] Q. Xie and A. R. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 431–445, 2000.
- [16] "Supplementary proofs and results." [Online]. Available: [https://github.com/wfyitf/Online-Transfer-Learning/blob/main/ISIT2021\\_Proof.pdf](https://github.com/wfyitf/Online-Transfer-Learning/blob/main/ISIT2021_Proof.pdf)
- [17] B. S. Clarke, "Asymptotic normality of the posterior in relative entropy," *IEEE Transactions on Information Theory*, vol. 45, no. 1, pp. 165–176, 1999.
- [18] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of bayes methods," *IEEE Transactions on Information Theory*, vol. 36, no. 3, pp. 453–471, 1990.
- [19] J.-I. Takeuchi and A. Barron, "Asymptotically minimax regret for exponential and curved exponential families," *Proc. IEEE*, 1998.