Online Transfer Learning: Effect of Prior Knowledge and Negative Transfer

Supplementary Proofs and Results

Xuetong Wu[1], Jonathan H. Manton[1], Uwe Aickelin[2], Jingge Zhu[1]
[1]Department of Electrical and Electronic Engineering
[2]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

CONTENTS

A. Proof of Theorem 1

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

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

Proof. We firstly show that given any prior over Θ_s and Θ_t,

$$\begin{aligned}
I(D_t^n;\Theta_t,\Theta_s|D_s^m) &= I(\Theta_t,\Theta_s;D_t^n,D_s^m) - I(\Theta_s;D_s^m) && (2)\\
&= D(P_{\Theta_t,\Theta_s}(D_t^n,D_s^m)\|Q(D_t^n,D_s^m)) - D(P_{\Theta_s}(D_s^m)\|Q(D_s^m)) && (3)\\
&= \int\left(\mathbb{E}_{\theta_s,\theta_t}\left[\log\frac{P_{\theta_t,\theta_s}(D_t^n,D_s^m)}{Q(D_t^n,D_s^m)}\right] - \mathbb{E}_{\theta_s,\theta_t}\left[\log\frac{P_{\theta_s}(D_s^m)}{Q(D_s^m)}\right]\right)\omega(\theta_s,\theta_t)d\theta_s d\theta_t && (4)\\
&= \int\left(\mathbb{E}_{\theta_s,\theta_t}\left[\log\frac{P_{\theta_t}(D_t^n)}{Q(D_t^n|D_s^m)}\right]\right)\omega(\theta_s,\theta_t)d\theta_s d\theta_t && (5)
\end{aligned}$$

where in the last equality we use the chain rule and the assumption that source data are independent of Θ_t. The mutual information density at $\Theta_s=\theta_s^*$ and $\Theta_t=\theta_t^*$ is then given by,

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

□

B. Proof of Theorem 2

Theorem 2 (Bounds on General Loss)**.** *Assume the loss function is convex and satisfies $|\ell(b,z)-\ell(b^*,z)|\le M$ for any observation z and the predictors b,b^*, at time k, we then choose our predictor to be,*

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

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 ω. The optimal predictor is naturally given by,

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

Then the true expected regrets can be bounded as,

$$\mathcal{R}(n) \le M\sqrt{2nI(D_t^n;\Theta_t=\theta_t^*,\Theta_s=\theta_s^*|D_s^m)} \tag{9}$$

Proof. We can show that the true regrets as,

$$\begin{aligned}
\mathcal{R}(n) &= \mathbb{E}_{\theta_t^*,\theta_s^*}\left[\sum_{k=1}^n \ell(b_k,x_t^{(k)}) - \sum_{k=1}^n \ell(b_k^*,x_t^{(k)})\right] && (10)\\
&= \sum_{k=1}^n \mathbb{E}_{D_s^m,D_t^{k-1}}\mathbb{E}_{x_t^{(k)}}\left[\ell(b_k,x_t^{(k)}) - \ell(b_k^*,x_t^{(k)})|D_s^m,D_t^{k-1}\right] && (11)\\
&= \sum_{k=1}^n \mathbb{E}_{D_s^m,D_t^{k-1}}\int\left(\ell(b_k,x_t^{(k)}) - \ell(b_k^*,x_t^{(k)})\right)P_{\theta_s^*,\theta_t^*}(x_t^{(k)}|D_s^m,D_t^{k-1})dx_t^{(k)} && (12)\\
&= \sum_{k=1}^n \mathbb{E}_{D_s^m,D_t^{k-1}}\int\left(\ell(b_k,x_t^{(k)}) - \ell(b_k^*,x_t^{(k)})\right)(P_{\theta_s^*,\theta_t^*}(x_t^{(k)}|D_s^m,D_t^{k-1}) - Q(x_t^{(k)}|D_s^m,D_t^{k-1}) + Q(x_t^{(k)}|D_s^m,D_t^{k-1}))dx_t^{(k)} && (13)\\
&\overset{(a)}{\le} \sum_{k=1}^n \mathbb{E}_{D_s^m,D_t^{k-1}}\int\left(\ell(b_k,x_t^{(k)}) - \ell(b_k^*,x_t^{(k)})\right)(P_{\theta_s^*,\theta_t^*}(x_t^{(k)}|D_s^m,D_t^{k-1}) - Q(x_t^{(k)}|D_s^m,D_t^{k-1}))dx_t^{(k)} && (14)\\
&\overset{(b)}{\le} \sum_{k=1}^n \mathbb{E}_{D_s^m,D_t^{k-1}}M\int(P_{\theta_s^*,\theta_t^*}(x_t^{(k)}|D_s^m,D_t^{k-1}) - Q(x_t^{(k)}|D_s^m,D_t^{k-1}))dx_t^{(k)} && (15)\\
&\overset{(c)}{\le} \sum_{k=1}^n \mathbb{E}_{D_s^m,D_t^{k-1}}M\sqrt{2D\left(P_{\theta_s^*,\theta_t^*}(x_t^{(k)}|D_s^m,D_t^{k-1})\|Q(x_t^{(k)}|D_s^m,D_t^{k-1})\right)} && (16)
\end{aligned}$$

$$\overset{(d)}{\leq} M\sum_{k=1}^{n}\sqrt{2\mathbb{E}_{D_s^m,D_t^{k-1}}D\left(P_{\theta_s^*,\theta_t^*}(x_t^{(k)}|D_s^m,D_t^{k-1})\|Q(x_t^{(k)}|D_s^m,D_t^{k-1})\right)} \tag{17}$$

$$\overset{(e)}{=} M\sum_{k=1}^{n}\sqrt{2D\left(P_{\theta_t^*}\|Q|D_s^m,D_t^{k-1}\right)} \tag{18}$$

$$\overset{(f)}{\leq} Mn\sqrt{\frac{2}{n}\sum_{k=1}^{n}D\left(P_{\theta_t^*}\|Q|D_s^m,D_t^{k-1}\right)} \tag{19}$$

$$\overset{(g)}{=} M\sqrt{2nD(P_{\theta_t^*}(D_t^n)\|Q(D_t^n|D_s^m))} \tag{20}$$

$$= M\sqrt{2nI(D_t^n;\Theta_t=\theta_t^*,\Theta_s=\theta_s^*|D_s^m)} \tag{21}$$

where in (a) we use the definition of Q, then (b) holds since we assume the loss function is bounded, (c) follows from the Pinsker's inequality, (d) and (f) follows from the Jensen's inequality, (g) holds because of the chain rule of the KL divergence. □

C. Proof of Theorem 3

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

$$I(D_t^n;\Theta_t=\theta_t^*,\Theta_s=\theta_s^*|D_s^m)\to\frac{1}{2}\log\frac{n}{2\pi e}+\frac{1}{2}\log I_t(\theta_t^*)+\log\frac{1}{\omega(\theta_t^*|\theta_s^*)} \tag{22}$$

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

Proof. We give the approximation on the KL divergence to see how the prior will affect the divergence,

$$\mathbb{E}_{\theta_s^*,\theta_t^*}\left[\log\frac{P_{\theta_t^*}(D_t^n)}{Q(D_t^n|D_s^m)}\right]=\mathbb{E}_{\theta_s^*,\theta_t^*}\left[\log\frac{P_{\theta_t^*}(D_t^n)P_{\theta_s^*}(D_s^m)Q(D_s^m)}{Q(D_t^n,D_s^m)P_{\theta_s^*}(D_s^m)}\right] \tag{23}$$

$$=\mathbb{E}_{\theta_s^*,\theta_t^*}\left[\log\frac{P_{\theta_t^*,\theta_s^*}(D_t^n,D_s^m)}{Q(D_t^n,D_s^m)}\right]-\mathbb{E}_{\theta_s^*,\theta_t^*}\left[\log\frac{P_{\theta_s^*}(D_s^m)}{Q(D_s^m)}\right] \tag{24}$$

$$=D(P_{\theta_t^*,\theta_s^*}(D_t^n,D_s^m)\|Q(D_t^n,D_s^m))-D(P_{\theta_s^*}(D_s^m)\|Q(D_s^m)) \tag{25}$$

We can view that source samples and target samples are jointly sampled given the distribution $P_{\theta_s^*}$ and $P_{\theta_t^*}$. Using the results in [1] and [2], with the proper prior $\omega(\theta_s,\theta_t)$ and parametric conditions, the asymptotic normality of the posterior implies that,

$$D(P_{\theta_t^*,\theta_s^*}(D_t^n,D_s^m)\|Q(D_t^n,D_s^m))-\frac{1}{2}\log\det\left(\begin{bmatrix}nI_t(\theta_t^*) & 0\\ 0 & mI_s(\theta_s^*)\end{bmatrix}\right)\to\log\frac{1}{2\pi e}+\log\frac{1}{\omega(\theta_s^*,\theta_t^*)} \tag{26}$$

as both n and m goes to infinity, where the fisher information matrices are denoted by

$$I_t(\theta_t^*)=-\mathbb{E}_{\theta_t^*}\left[\frac{\partial\log P(x|\theta_t^*)}{\partial\theta_t^2}\right] \tag{27}$$

$$I_s(\theta_s^*)=-\mathbb{E}_{\theta_s^*}\left[\frac{\partial\log P(x|\theta_s^*)}{\partial\theta_s^2}\right] \tag{28}$$

Similarly,

$$D(P_{\theta_s^*}(D_s^m)\|Q(D_s^m))-\frac{1}{2}\log\det(mI_s(\theta_s^*))\to\frac{1}{2}\log\frac{1}{2\pi e}+\log\frac{1}{\omega(\theta_s^*)} \tag{29}$$

as m goes to infinity. Therefore,

$$\lim_{n,m\to\infty}\left(\mathbb{E}_{\theta_s^*,\theta_t^*}\left[\log\frac{P_{\theta_t^*}(D_t^n)}{Q(D_t^n|D_s^m)}\right]\right)=\frac{1}{2}\log\det\left(\begin{bmatrix}nI_t(\theta_t^*) & 0\\ 0 & \alpha nI_s(\theta_s^*)\end{bmatrix}\right)+\log\frac{1}{2\pi e}+\log\frac{1}{\omega(\theta_s^*,\theta_t^*)} \tag{30}$$

$$-\frac{1}{2}\log\det(mI_s(\theta_s^*))-\frac{1}{2}\log\frac{1}{2\pi e}-\log\frac{1}{\omega(\theta_s^*)} \tag{31}$$

$$=\frac{1}{2}\log\det(nI_t(\theta_t^*))+\frac{1}{2}\log\frac{1}{2\pi e}+\log\frac{1}{\omega(\theta_t^*|\theta_s^*)} \tag{32}$$

To conclude, as both n and m goes to infinity, the conditional mutual information will converge to,

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

□

D. Proof of Theorem 4

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

$$I(D_t^n; \Theta_t = \theta_t^*, \Theta_s = \theta_s^* | D_s^m) \to \frac{1}{2}\log\det(nI_t(\theta_{tr}^*)) + \frac{1}{2}\log\det(\mathbf{I}_{j\times j} + \frac{n}{m}\Delta_t\Delta_s^{-1}) + \frac{d-j}{2}\log\frac{1}{2\pi e} + \log\frac{1}{\omega(\theta_t^*|\theta_s^*)} \quad (33)$$

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

$$I_{cs}(\theta_c^*) = -\mathbb{E}_{\theta_s^*}\left[\nabla^2_{\Theta_c}\log P(Z_s|\Theta_c, \theta_{sr}^*)\right]\Big|_{\Theta_c=\theta_c^*} \in \mathbb{R}^{j\times j} \quad (34)$$

$$I_{ct}(\theta_c^*) = -\mathbb{E}_{\theta_t^*}\left[\nabla^2_{\Theta_c}\log P(Z_t|\Theta_c, \theta_{tr}^*)\right]\Big|_{\Theta_c=\theta_c^*} \in \mathbb{R}^{j\times j} \quad (35)$$

$$I_s(\theta_{sr}^*) = -\mathbb{E}_{\theta_s^*}\left[\nabla^2_{\Theta_{sr}}\log P(Z_s|\theta_c^*, \Theta_{sr})\right]\Big|_{\Theta_{sr}=\theta_{sr}^*} \in \mathbb{R}^{(d-j)\times(d-j)} \quad (36)$$

$$I_t(\theta_{tr}^*) = -\mathbb{E}_{\theta_t^*}\left[\nabla^2_{\Theta_{tr,l}}\log P(Z_t|\theta_c^*, \Theta_{tr,l})\right]\Big|_{\Theta_{tr,l}=\theta_{tr}^*} \in \mathbb{R}^{(d-j)\times(d-j)} \quad (37)$$

$$I_{cs}(\theta_c^*, \theta_{sr}^*) = -\mathbb{E}_{\theta_s^*}\left[\frac{\partial\log P(Z_s|\theta_c^*, \theta_{sr}^*)}{\partial\Theta_{c,i}\partial\Theta_{sr,k}}\right]_{\substack{i=1,\cdots,j,\\ k=1,\cdots,d-j}} \in \mathbb{R}^{j\times(d-j)} \quad (38)$$

$$I_{ct}(\theta_c^*, \theta_{tr}^*) = -\mathbb{E}_{\theta_t^*}\left[\frac{\partial\log P(Z_t|\theta_c^*, \theta_{tr}^*)}{\partial\Theta_{c,i}\partial\Theta_{tr,k}}\right]_{\substack{i=1,\cdots,j,\\ k=1,\cdots,d-j}} \in \mathbb{R}^{j\times(d-j)} \quad (39)$$

Proof. By writing $\theta_s = (\theta_c, \theta_{sr})$ and $\theta_t = (\theta_c, \Theta_{tr,l})$, let us rewrite the conditional mutual information as

$$\mathbb{E}_{\theta_s^*,\theta_t^*}\left[\log\frac{P_{\theta_t^*}(D_t^n)}{Q(D_t^n|D_s^m)}\right] = \mathbb{E}_{\theta_s^*,\theta_t^*}\left[\log\frac{P_{\theta_t^*}(D_t^n)P_{\theta_s^*}(D_s^m)Q(D_s^m)}{Q(D_t^n, D_s^m)P_{\theta_s^*}(D_s^m)}\right] \quad (40)$$

$$= \mathbb{E}_{\theta_c^*,\theta_{sr}^*,\theta_{tr}^*}\left[\log\frac{P_{\theta_t^*,\theta_s^*}(D_t^n, D_s^m)}{Q(D_t^n, D_s^m)}\right] - \mathbb{E}_{\theta_c^*,\theta_{sr}^*}\left[\log\frac{P_{\theta_s^*}(D_s^m)}{Q(D_s^m)}\right] \quad (41)$$

$$= D(P_{\theta_c^*,\theta_{sr}^*,\theta_{tr}^*}(D_t^n, D_s^m)\|P(D_t^n, D_s^m)) - D(P_{\theta_c^*,\theta_{sr}^*}(D_s^m)\|Q(D_s^m)) \quad (42)$$

We view that m source samples and n target samples are jointly sampled from the distribution parametrized by the parameters $\boldsymbol{\Theta} = (\Theta_c, \Theta_{sr}, \Theta_{tr,l})$. At time n, we have the asymptotic approximation under the proper prior and assumption 2 using Theorem 2.1 in [1] and [2] as

$$D(P_{\theta_c^*,\theta_{sr}^*,\theta_{tr}^*}(D_t^n, D_s^m)\|P(D_t^n, D_s^m)) - \frac{1}{2}\log\det\left(\mathbf{I}_{\boldsymbol{\theta}^*}\right) \to \frac{2d-j}{2}\log\frac{1}{2\pi e} + \log\frac{1}{\omega(\theta_s^*, \theta_t^*)} \quad (43)$$

where the Fisher information matrix is defined as,

$$\mathbf{I}_{\boldsymbol{\theta}^*} = \begin{bmatrix} mI_{cs}(\theta_c^*) + nI_{ct}(\theta_c^*) & mI_{cs}(\theta_c^*, \theta_{sr}^*) & nI_{ct}(\theta_c^*, \theta_{tr}^*) \\ mI_{cs}^T(\theta_c^*, \theta_{sr}^*) & mI_s(\theta_{sr}^*) & \mathbf{0} \\ nI_{ct}^T(\theta_c^*, \theta_{tr}^*) & \mathbf{0} & nI_t(\theta_{tr}^*) \end{bmatrix} \quad (44)$$

Similarly,

$$D(P_{\theta_c^*,\theta_{sr}^*}(D_s^m)\|Q(D_s^m)) - \frac{1}{2}\log\det\left(\mathbf{I}_{\boldsymbol{\theta}_s^*}\right) \to \frac{d}{2}\log\frac{1}{2\pi e} + \log\frac{1}{\omega(\theta_s^*)} \quad (45)$$

where,

$$\mathbf{I}_{\boldsymbol{\theta}_s^*} = m\begin{bmatrix} I_{cs}(\theta_c^*) & I_{cs}(\theta_c^*, \theta_{sr}^*) \\ I_{cs}^T(\theta_c^*, \theta_{sr}^*) & I_s(\theta_{sr}^*) \end{bmatrix} \quad (46)$$

as m goes to sufficiently large. As a consequence,

$$\lim_{n,m\to\infty}\left(\mathbb{E}_{\theta_s^*,\theta_t^*}\left[\log\frac{P_{\theta_t^*}(D_t^n)}{Q(D_t^n|D_s^m)}\right]\right) = \frac{1}{2}\log\det\left(\mathbf{I}_{\boldsymbol{\theta}^*}\right) + \frac{2d-j}{2}\log\frac{1}{2\pi e} + \log\frac{1}{\omega(\theta_s^*, \theta_t^*)} \quad (47)$$

$$- \frac{1}{2}\log\det(\mathbf{I}_{\boldsymbol{\theta}_s^*}) - \frac{d}{2}\log\frac{1}{2\pi e} - \log\frac{1}{\omega(\theta_s^*)} \quad (48)$$

$$= \frac{1}{2}\log\frac{\det(\mathbf{I}_{\boldsymbol{\theta}^*})}{\det(\mathbf{I}_{\boldsymbol{\theta}_s^*})} + \frac{d-j}{2}\log\frac{1}{2\pi e} + \log\frac{1}{\omega(\theta_t^*|\theta_s^*)} \quad (49)$$

Let us examine the ratio of the determinant, using the block determinant results from [3],

$$\log \frac{\det(\mathbf{I}_{\boldsymbol{\theta}^*})}{\det(\mathbf{I}_{\boldsymbol{\theta}_s^*})} = \log\det\left(mI_{cs}(\theta_c^*) + nI_{ct}(\theta_c^*) - \begin{bmatrix} mI_{cs}(\theta_c^*, \theta_{sr}^*) & nI_{ct}(\theta_c^*, \theta_{tr}^*) \end{bmatrix} \begin{bmatrix} mI_s(\theta_{sr}^*) & \mathbf{0} \\ \mathbf{0} & nI_t(\theta_{tr}^*) \end{bmatrix}^{-1} \begin{bmatrix} mI_{cs}^T(\theta_c^*, \theta_{sr}^*) \\ nI_{ct}^T(\theta_c^*, \theta_{tr}^*) \end{bmatrix} \right)$$
$$+ \log\det\left(\begin{bmatrix} mI_s(\theta_{sr}^*) & \mathbf{0} \\ \mathbf{0} & nI_t(\theta_{tr}^*) \end{bmatrix} \right) - \log\det\left(m(I_{cs}(\theta_c^*) - I_{cs}(\theta_c^*, \theta_{sr}^*) I_s^{-1}(\theta_{sr}^*) I_{cs}^T(\theta_c^*, \theta_{sr}^*)) \right) - \log\det\left(mI_s(\theta_{sr}^*) \right) \tag{50}$$

$$= \log\det(m\Delta_s + n\Delta_t) - \log\det(m\Delta_s) + \log\det(nI_t(\theta_{tr}^*) \tag{51}$$

$$= \log\det(\mathbf{I}_{j\times j} + \frac{n}{m}\Delta_t\Delta_s^{-1}) + \log\det(nI_t(\theta_{tr}^*) \tag{52}$$

where we define $\Delta_s = I_{cs}(\theta_c^*) - I_{cs}(\theta_c^*, \theta_{sr}^*) I_s^{-1}(\theta_{sr}^*) I_{cs}^T(\theta_c^*, \theta_{sr}^*)$ and $\Delta_t = I_{ct}(\theta_c^*) - I_{ct}(\theta_c^*, \theta_{tr}^*) I_t^{-1}(\theta_{tr}^*) I_{ct}^T(\theta_c^*, \theta_{tr}^*)$. Putting everything together, we reach,

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

as both n and m goes to infinity. □

E. Proof of Proposition 1

Proposition 1 (Negative Transfer)**.** *Let $\mathcal{R}_{\omega(\Theta_s,\Theta_t)}(n)$ denote the regret induced by $Q(D_t^n|D_s^m)$ 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 chosen improperly, for any proper $\hat{\omega}(\Theta_t)$, the following inequality holds when both n and m are sufficiently large.*

$$\mathcal{R}_{\omega(\Theta_s,\Theta_t)}(n) > \mathcal{R}_{\hat{\omega}(\Theta_t)}(n) \tag{53}$$

Proof. We need to prove there exists a prior $\omega(\theta_s, \theta_t)$, the expected regrets such that

$$\mathbb{E}_{\theta_s^*,\theta_t^*}\left[\log\frac{P_{\theta_t^*}(D_t^n)}{Q(D_t^n|D_s^m)}\right] > \mathbb{E}_{\theta_t^*}\left[\log\frac{P_{\theta_t^*}(D_t^n)}{\hat{Q}(D_t^n)}\right] \tag{54}$$

It is equivalent to prove that,

$$\mathbb{E}_{\theta_t^*,\theta_s^*}\left[\log\frac{Q(D_s^m)\hat{Q}(D_t^n)}{Q(D_t^n, D_s^m)}\right] > 0 \tag{55}$$

Let us examine the logarithm term,

$$\log\frac{Q(D_s^m)\hat{Q}(D_t^n)}{Q(D_t^n, D_s^m)} = \log\frac{\hat{Q}(D_t^n)\int P_{\theta_t,\theta_s}(D_s^m)\omega(\theta_t,\theta_s)d\theta_t d\theta_s}{\int P_{\theta_t,\theta_s}(D_t^n, D_s^m)\omega(\theta_t,\theta_s)d\theta_t d\theta_s} \tag{56}$$

$$= \log\frac{\hat{Q}(D_t^n)\int P_{\theta_s}(D_s^m)\omega(\theta_s)d\theta_s}{\int P_{\theta_t,\theta_s}(D_t^n, D_s^m)\omega(\theta_t,\theta_s)d\theta_t d\theta_s} \tag{57}$$

$$= \log\frac{1}{\int\int \hat{Q}(\theta_t|D_t^n)\frac{\omega(\theta_t|\theta_s)}{\hat{\omega}(\theta_t)}d\theta_t Q(\theta_s|D_s^m)d\theta_s} \tag{58}$$

When both m and n are sufficient enough and the marginal prior distribution $\omega(\theta_s)$ and $\hat{\omega}(\theta_t)$ are proper, $\hat{Q}(\theta_t|D_t^n)$ and $Q(\theta_s|D_s^m)$ will be concentrated near θ_t^* and θ_s^*. Then the above equation becomes,

$$\log\frac{Q(D_s^m)\hat{Q}(D_t^n)}{Q(D_t^n, D_s^m)} = -\log\int_{\theta_s^*-\delta_s}^{\theta_s^*+\delta_s}\int_{\theta_t^*-\delta_t}^{\theta_t^*+\delta_t}\hat{Q}(\theta_t|D_t^n)\frac{\omega(\theta_t|\theta_s^*)}{\hat{\omega}(\theta_t)}d\theta_t Q(\theta_s|D_s^m)d\theta_s \tag{59}$$

for some small δ_t and δ_s. Since the prior $\omega(\theta_t|\theta_s)$ is imposed improperly, then $\omega(\theta_t|\theta_s)$ has zero density around θ_t^*, then since for any $\hat{\omega}(\theta_t) > 0$, the following inequality holds.

$$\mathbb{E}_{\theta_s^*,\theta_t^*}\left[\log\frac{P(D_t^n)P(D_s^m)}{P(D_t^n, D_s^m)}\right] = -\mathbb{E}_{\theta_s^*,\theta_t^*}\left[\log\int_{\theta_s^*-\delta_s}^{\theta_s^*+\delta_s}\int_{\theta_t^*-\delta_t}^{\theta_t^*+\delta_t}\hat{Q}(\theta_t|D_t^n)\frac{\omega(\theta_t|\theta_s^*)}{\hat{\omega}(\theta_t)}d\theta_t Q(\theta_s|D_s^m)d\theta_s\right] \tag{60}$$

$$> -\mathbb{E}_{\theta_s^*,\theta_t^*}\left[\log\int_{\theta_s^*-\delta_s}^{\theta_s^*+\delta_s}\int_{\theta_t^*-\delta_t}^{\theta_t^*+\delta_t}\hat{Q}(\theta_t|D_t^n)\frac{\hat{\omega}(\theta_t)}{\hat{\omega}(\theta_t)}d\theta_t Q(\theta_s|D_s^m)d\theta_s\right] \tag{61}$$

$$= 0 \tag{62}$$

when the source and target are sufficiently large. Therefore , it implies that,

$$\mathcal{R}_{\omega(\Theta_s,\Theta_t)}(n) > \mathcal{R}_{\hat{\omega}(\Theta_t)}(n) \tag{63}$$

□

F. Proof of Proposition 2

Proposition 2 (Positive Transfer)**.** *If $\omega(\theta_s, \theta_t)$ is chosen properly, and the support of $\omega(\theta_t|\theta_s)$ is a proper subset of Θ for any $\|\theta_s - \theta_s^*\| \leq \delta_s$ with some $\delta_s > 0$, then with sufficient source data, we can always find such a prior that its corresponding prediction $Q(D_t^n|D_s^m)$ will yield strictly better regrets than that the regrets $Q(D_t^n)$ induced by* **any** *proper prior when n is sufficiently large.*

Proof. We need to prove under the certain assumptions, there exists a prior $\omega(\theta_s, \theta_t)$, the expected regrets such that

$$\mathbb{E}_{\theta_s^*,\theta_t^*}\left[\log\frac{P_{\theta_t^*}(D_t^n)}{Q(D_t^n|D_s^m)}\right] < \mathbb{E}_{\theta_t^*}\left[\log\frac{P_{\theta_t^*}(D_t^n)}{\hat{Q}(D_t^n)}\right] \tag{64}$$

Rewrite the expectation and it is equivalent to prove that,

$$\mathbb{E}_{\theta_t^*,\theta_s^*}\left[\log\frac{Q(D_s^m)\hat{Q}(D_t^n)}{Q(D_t^n, D_s^m)}\right] < 0 \tag{65}$$

Similarly, when D_s^m is sufficient enough, the density $P(\theta_s^*|D_s^m)$ will concentrate around θ_s^* , say $P(|\theta_s - \theta_s^*| < \delta_s) \to 0$ as m goes to infinity, furthermore if D_t^n is also very large, $p(\theta_t|D_t^n)$ will be concentrated near θ_t^* such that,

$$\log\frac{Q(D_s^m)\hat{Q}(D_t^n)}{Q(D_t^n, D_s^m)} = -\log\int_{\theta_s^*-\delta_s}^{\theta_s^*+\delta_s}\int_{\theta_t^*-\delta_t}^{\theta_t^*+\delta_t}\hat{Q}(\theta_t|D_t^n)\frac{\omega(\theta_t|\theta_s^*)}{\hat{\omega}(\theta_t)}d\theta_t Q(\theta_s|D_s^m)d\theta_s \tag{66}$$

Since we assume the support $\omega(\theta_t|\theta_s)$ is a proper subset of Θ and $\omega(\theta_t|\theta_s)$ is proper over θ_t, that is, this conditional prior has positive density around θ_t^*. Let us define,

$$\hat{\Omega} = \int_{\theta_t^*-\delta_t}^{\theta_t^*+\delta_t}\hat{\omega}(\theta_t)d\theta_t \tag{67}$$

$$\Omega = \int_{\theta_t^*-\delta_t}^{\theta_t^*+\delta_t}\omega(\theta_t|\theta_s)d\theta_t \tag{68}$$

Then there always exists a prior such that for any θ_s around θ_s^*,

$$\Omega - \hat{\Omega} = \Delta > 0$$

with the choice of the prior

$$\omega(\theta_t|\theta_s) = \hat{\omega}(\theta_t) + \frac{\Delta}{2\delta_t} \tag{69}$$

This specific prior will lead to $\log\frac{\hat{\omega}(\theta_t)}{\omega(\theta_t|\theta_s)} < 1$ and the quantity above will be strictly less than zero, which is, positive transfer. □

G. Extended Results on Time-variant Target Domain

We define the problem formally in this subsection. Let the time evolving target data be parametrized by $\theta_{t,l}^*$ where at each index $l \in \mathbb{N}^+$, we will receive n_l target samples $Z_{t,l}^{(i)}$ sequentially drawn from the distribution $P_{\theta_{t,l}^*}$, and it is common to assume that the source parameter θ_s^* shares j common parameters with every $\theta_{t,l}^*$ and $\theta_{t,l}^*$ only depends on the previous parameter $\theta_{t,l-1}^*$. Similarly, we assume given $\theta_{t,l}^*$, $\theta_{t,l}^*$ and θ_s^*, the source samples and target samples are drawn independently as

$$P_{\theta_{t,l}^*,\theta_{t,l}^*,\theta_s^*}(Z_s, Z_{t,l}, Z_{t,l-1}) = P_{\theta_s^*}(Z_s)P_{\theta_{t,l-1}^*}(Z_{t,l-1})P_{\theta_{t,l}^*}(Z_{t,l}) \tag{70}$$

At each index l, let us denote the received target samples $(Z_{t,l}^{(i)})_{i=1,2,\cdots,n_l}$ by D_t^l. We are interested in minimising the expected regret

$$\mathcal{R}(n) = \sum_{l=1}^{k}\mathbb{E}_{\theta_s^*,\theta_{t,l}^*,\theta_{t,l-1}^*}\left[\sum_{i=1}^{n_l}\ell\left(b_i, Z_{t,l}^{(i)}\right) - \sum_{i=1}^{n_l}\ell(b_i^*, Z_{t,l}^{(i)})\right] \tag{71}$$

Here b_i is chosen with the mixture strategy over $\theta_s^*, \theta_{t,l}^*$, and $\theta_{t,l-1}^*$. Let us define the random variables Θ_s, $\Theta_{t,l-1}$ and $\Theta_{t,l}$ as follows.

$$\Theta_s = (\Theta_{c,1}, \Theta_{c,2}, \cdots, \Theta_{c,j}, \underbrace{\Theta_{s,1}, \cdots, \Theta_{s,d-j}}_{\text{source-specific parameters}}) = (\Theta_c, \Theta_{sr})$$

$$\Theta_{t,l-1} = (\Theta_{c,1}, \Theta_{c,2}, \cdots, \Theta_{c,j}, \quad \Theta_{v,1}, \cdots, \Theta_{v,c_l}, \quad \Theta_{t',1}, \cdots, \Theta_{t',d-j-c_l}) = (\Theta_c, \Theta_v, \Theta_{tr,l-1})$$

$$\Theta_{t,l} = (\underbrace{\Theta_{c,1}, \Theta_{c,2}, \cdots, \Theta_{c,j}}_{\text{time-invariant parameters}}, \quad \underbrace{\Theta_{v,1}, \cdots, \Theta_{v,c_l}}_{\text{target common parameters}}, \quad \underbrace{\Theta_{t,1}, \cdots, \Theta_{t,d-j-c_l}}_{\text{target-specific parameters}}) = (\Theta_c, \Theta_v, \Theta_{tr,l})$$

Where we 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. Here Θ_c and Θ_v represent the time-invariant parameters and the target common parameters under index l and the parameters changing from $\Theta_{tr,l-1}$ to $\Theta_{tr,l}$ exhibit the nature of time-varying target domains. Then we use the mixture strategy such that with some prior $\omega(\theta_s, \theta_{t,l-1}, \theta_{t,l})$,

$$Q(D_t^l | D_t^{l-1}, D_s^m) = \frac{Q(D_t^l, D_t^{l-1}, D_s^m)}{Q(D_t^{l-1}, D_s^m)} \tag{72}$$

$$= \frac{\int P_{\theta_s, \theta_{t,l-1}, \theta_{t,l}}(D_s^m, D_t^{l-1}, D_t^l)\omega(\theta_s, \theta_{t,l-1}, \theta_{t,l}) d\theta_s d\theta_{t,l-1} d\theta_{t,l}}{\int P_{\theta_s, \theta_l}(D_s^m, D_t^{l-1})\omega(\theta_s, \theta_{t,l-1}) d\theta_s d\theta_{t,l-1}}$$

$$= \int P_{\theta_t}(D_t^n)\omega(\theta_t | \theta_s, \theta_{t,l-1}) d\theta_t Q(\theta_s, \theta_{t,l-1} | D_s^m, D_t^{l-1}) d\theta_s d\theta_{t,l-1} \tag{73}$$

The posterior $Q(\theta_s, \theta_{t,l-1} | D_s^m, D_t^{l-1})$ firstly gives an estimate of the source parameter and previous target parameter with marginal $\omega(\theta_s, \theta_{t,l-1})$, then the knowledge transfer is reflected on the conditional prior $\omega(\theta_{t,l} | \theta_s, \theta_{t,l-1})$ that may or may not result in a good approximation of θ_t^*. By this, we reach the following theorem.

Theorem 5 (Time-variant Target Regret Bounds)**.** *Given the time-variant target domain described above, suppose that conditions in Therorem 2 and Assumptions 1 and 2 hold for each $\theta_{t,k}^*$ and θ_s^*. For $l = 1, 2, \cdots, k$, 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 $n_l, m \to \infty$, the mixture strategy with proper prior $\omega(\theta_s, \theta_{t,l}, \theta_{t,l-1})$ yields,*

$$\mathcal{R}(n) \leq M \Bigg(k \sum_{l=1}^{k} n_l \Big(\log\det\left(\mathbf{I}_{j\times j} + \frac{n_l}{m + n_{l-1}} \Delta_{ct} \Delta_{cst}^{-1} \right) + \log\det(\mathbf{I}_{c_l \times c_l} + \frac{n_l}{n_{l-1}} \Delta_t \Delta_{t-1}^{-1}) + \log\det(n I_{t,l}(\theta_{tr,l}^*)) \\ + (d - j - c_l) \log \frac{1}{2\pi e} + \frac{2}{\omega(\theta_{t,l}^* | \theta_{t,l-1}^*, \theta_s^*)} \Big) \Bigg)^{\frac{1}{2}} \tag{74}$$

Proof. At each time l, under the conditions from Theorem 2, we give upper of the expectation term as,

$$\mathbb{E}_{\theta_s^*, \theta_{t,l}^*, \theta_{t,l-1}^*} \left[\sum_{i=1}^{n_l} \ell\left(b_i, Z_{t,l}^{(i)}\right) - \sum_{i=1}^{n_l} \ell(b_i^*, Z_{t,l}^{(i)}) \right] \leq M \sqrt{2 n_l I(D_t^l; \Theta_{t,l} = \theta_{t,l}^* | D_s^m, D_t^{l-1})} \tag{75}$$

where we define the conditional mutual information as,

$$I(D_t^l; \Theta_{t,l} = \theta_{t,l}^* | D_s^m, D_t^{l-1}) := \mathbb{E}_{\theta_s^*, \theta_{t,l}^*, \theta_{t,l-1}^*} \left[\log \frac{P_{\theta_{t,l}^*}(D_t^l)}{Q(D_t^l | D_t^{l-1}, D_s^m)} \right] \tag{76}$$

By Cauchy-Schwarz inequality,

$$\mathcal{R}(k) = \sum_{l=1}^{k} \mathbb{E}_{\theta_s^*, \theta_{t,l}^*, \theta_{t,l-1}^*} \left[\sum_{i=1}^{n_l} \ell\left(b_i, Z_{t,l}^{(i)}\right) - \sum_{i=1}^{n_l} \ell(b_i^*, Z_{t,l}^{(i)}) \right] \tag{77}$$

$$\leq \sum_{l=1}^{k} M \sqrt{2 n_l I(D_t^l; \Theta_{t,l} = \theta_{t,l}^* | D_s^m, D_t^{l-1})} \tag{78}$$

$$\leq M \sqrt{2k \sum_{l=1}^{k} n_l I(D_t^l; \Theta_{t,l} = \theta_{t,l}^* | D_s^m, D_t^{l-1})} \tag{79}$$

Let us give an asymptotic estimation on $I(D_t^l; \Theta_{t,l} = \theta_{t,l}^* | D_s^m, D_t^{l-1})$ as all n_l, n_{l-1} and m are large enough shown in the following theorem.

Theorem 6. *Under Assumptions 1 and 2, with $\Theta_s, \Theta_{t,l-1}.\Theta_{t,l} \in \mathbb{R}^d$ defined in the paper and as $n_{l-1}, n_l, m \to \infty$, the mixture strategy with proper prior $\omega(\Theta_s, \Theta_{t,l-1}, \Theta_{t,l})$ yields,*

$$I(D_t^l; \Theta_{t,l} = \theta_{t,l}^* | D_s^m, D_t^{l-1}) - \log\det\left(\mathbf{I}_{j\times j} + \frac{n_l}{m+n_{l-1}}\Delta_{ct}\Delta_{cst}^{-1}\right) - \log\det(\mathbf{I}_{c_l\times c_l} + \frac{n_l}{n_{l-1}}\Delta_t\Delta_{t-1}^{-1}) - \log\det(n_l I_{t,l}(\theta_{tr,l}^*))$$
$$\to (d-j-c_l)\log\frac{1}{2\pi e} + \frac{2}{\omega(\theta_{t,l}^*|\theta_{t,l-1}^*, \theta_s^*)} \tag{80}$$

Proof. Rewrite the conditional mutual information in terms of the KL divergence as,

$$I(D_t^l; \Theta_{t,l} = \theta_{t,l}^* | D_s^m, D_t^{l-1}) = \mathbb{E}_{\theta_s^*, \theta_{t,l}^*, \theta_{t,l-1}^*}\left[\log\frac{P_{\theta_{t,l}^*}(D_t^l)}{Q(D_t^l|D_t^{l-1}, D_s^m)}\right] \tag{81}$$
$$= D(P_{\theta_s^*, \theta_{t,l}^*, \theta_{t,l-1}^*}(D_s^m, D_t^{l-1}, D_t^l) \| Q(D_s^m, D_t^{l-1}, D_t^l)) - D(P_{\theta_s^*, \theta_{t,l-1}^*}(D_s^m, D_t^{l-1}) \| Q(D_s^m, D_t^{l-1})) \tag{82}$$

Let us align the parameter as $\mathbf{\Theta} = (\Theta_c, \Theta_v, \Theta_{sr}, \Theta_{tr,l-1}, \Theta_{tr,l})$ and $\mathbf{\Theta}_s = (\Theta_c, \Theta_v, \Theta_{sr}, \Theta_{tr,l-1})$ and we define a set of Fisher information matrices as,

$$I_{cs}(\theta_c^*) = -\mathbb{E}_{\theta_s^*}\left[\nabla^2_{\Theta_c}\log P(Z_s|\Theta_c, \theta_{sr}^*)\right]\Big|_{\Theta_c=\theta_c^*} \in \mathbb{R}^{j\times j} \tag{83}$$
$$I_{ct,l-1}(\theta_c^*) = -\mathbb{E}_{\theta_t^*}\left[\nabla^2_{\Theta_c}\log P(Z_{t,l-1}|\Theta_c, \theta_v^*, \theta_{tr,l-1}^*)\right]\Big|_{\Theta_c=\theta_c^*} \in \mathbb{R}^{j\times j} \tag{84}$$
$$I_{ct,l}(\theta_c^*) = -\mathbb{E}_{\theta_t^*}\left[\nabla^2_{\Theta_c}\log P(Z_{t,l}|\Theta_c, \theta_v^*, \theta_{tr,l}^*)\right]\Big|_{\Theta_c=\theta_c^*} \in \mathbb{R}^{j\times j} \tag{85}$$
$$I_s(\theta_{sr}^*) = -\mathbb{E}_{\theta_s^*}\left[\nabla^2_{\Theta_{sr}}\log P(Z_s|\theta_c^*, \Theta_{sr})\right]\Big|_{\Theta_{sr}=\theta_{sr}^*} \in \mathbb{R}^{(d-j)\times(d-j)} \tag{86}$$
$$I_{t,l-1}(\theta_{tr,l-1}^*) = -\mathbb{E}_{\theta_t^*}\left[\nabla^2_{\Theta_{tr,l-1}}\log P(Z_t|\theta_c^*, \theta_v^*, \Theta_{tr,l-1})\right]\Big|_{\Theta_{tr,l-1}=\theta_{tr,l-1}^*} \in \mathbb{R}^{(d-c_l-j)\times(d-c_l-j)} \tag{87}$$
$$I_{t,l}(\theta_{tr,l}^*) = -\mathbb{E}_{\theta_t^*}\left[\nabla^2_{\Theta_{tr,l}}\log P(Z_t|\theta_c^*, \Theta_{tr,l})\right]\Big|_{\Theta_{tr,l}=\theta_{tr}^*} \in \mathbb{R}^{(d-c_l-j)\times(d-c_l-j)} \tag{88}$$
$$I_{v,l-1}(\theta_v^*) = -\mathbb{E}_{\theta_s^*}\left[\nabla^2_{\Theta_v}\log P(Z_s|\theta_c^*, \Theta_v, \theta_{tr,l-1}^*)\right]\Big|_{\Theta_v=\theta_v^*} \in \mathbb{R}^{c_l\times c_l} \tag{89}$$
$$I_{v,l}(\theta_v^*) = -\mathbb{E}_{\theta_s^*}\left[\nabla^2_{\Theta_v}\log P(Z_s|\theta_c^*, \Theta_v, \theta_{tr,l}^*)\right]\Big|_{\Theta_v=\theta_v^*} \in \mathbb{R}^{c_l\times c_l} \tag{90}$$
$$I_{sc}(\theta_c^*, \theta_{sr}^*) = -\mathbb{E}_{\theta_s^*}\left[\frac{\partial\log P(Z_s|\theta_c^*, \theta_{sr}^*)}{\partial\Theta_{c,i}\partial\Theta_{sr,k}}\right]_{\substack{i=1,\cdots,j,\\ k=1,\cdots,d-j}} \in \mathbb{R}^{j\times(d-j)} \tag{91}$$
$$I_{ctr,l-1}(\theta_c^*, \theta_{tr,l-1}^*) = -\mathbb{E}_{\theta_t^*}\left[\frac{\partial\log P(Z_{t,l-1}|\Theta_c=\theta_c^*, \theta_v^*, \Theta_{tr,l-1}=\theta_{tr,l-1}^*)}{\partial\Theta_{c,i}\partial\Theta_{t',k}}\right]_{\substack{i=1,\cdots,j,\\ k=1,\cdots,d-c_l-j}} \in \mathbb{R}^{j\times(d-c_l-j)} \tag{92}$$
$$I_{ctr,l}(\theta_c^*, \theta_{tr,l}^*) = -\mathbb{E}_{\theta_t^*}\left[\frac{\partial\log P(Z_{t,l}|\Theta_c=\theta_c^*, \theta_v^*, \Theta_{tr,l}=\theta_{tr,l}^*)}{\partial\Theta_{c,i}\partial\Theta_{t,k}}\right]_{\substack{i=1,\cdots,j,\\ k=1,\cdots,d-c_l-j}} \in \mathbb{R}^{j\times(d-c_l-j)} \tag{93}$$
$$I_{vt,l-1}(\theta_v^*, \theta_{tr,l-1}^*) = -\mathbb{E}_{\theta_t^*}\left[\frac{\partial\log P(Z_{t,l-1}|\theta_c^*, \Theta_v=\theta_v^*, \Theta_{tr,l-1}=\theta_{tr,l-1}^*)}{\partial\Theta_{v,i}\partial\Theta_{t',k}}\right]_{\substack{i=1,\cdots,c_l,\\ k=1,\cdots,d-c_l-j}} \in \mathbb{R}^{c_l\times(d-c_l-j)} \tag{94}$$
$$I_{vt,l}(\theta_v^*, \theta_{tr,l}^*) = -\mathbb{E}_{\theta_t^*}\left[\frac{\partial\log P(Z_{t,l}|\theta_c^*, \Theta_v=\theta_v^*, \Theta_{tr,l}=\theta_{tr,l}^*)}{\partial\Theta_{v,i}\partial\Theta_{t,k}}\right]_{\substack{i=1,\cdots,c_l,\\ k=1,\cdots,d-c_l-j}} \in \mathbb{R}^{c_l\times(d-c_l-j)} \tag{95}$$
$$I_{cv,l-1}(\theta_c^*, \theta_v^*) = -\mathbb{E}_{\theta_t^*}\left[\frac{\partial\log P(Z_{t,l-1}|\Theta_c=\theta_c^*, \Theta_v=\theta_v^*, \theta_{tr,l-1}^*=\theta_{tr,l-1}^*)}{\partial\Theta_{v,i}\partial\Theta_{t',k}}\right]_{\substack{i=1,\cdots,j,\\ k=1,\cdots,c_l}} \in \mathbb{R}^{j\times c_l} \tag{96}$$
$$I_{cv,l}(\theta_c^*, \theta_v^*) = -\mathbb{E}_{\theta_t^*}\left[\frac{\partial\log P(Z_{t,l}|\Theta_c=\theta_c^*, \Theta_v=\theta_v^*, \theta_{tr,l}^*=\theta_{tr,l}^*)}{\partial\Theta_{v,i}\partial\Theta_{t,k}}\right]_{\substack{i=1,\cdots,j,\\ k=1,\cdots,c_l}} \in \mathbb{R}^{j\times c_l} \tag{97}$$

To simplify the notations, we omit the . Then the asymptotic normality implies that,

$$D(P_{\theta_s^*,\theta_{t,l}^*,\theta_{t,l-1}^*}(D_s^m,D_t^{l-1},D_t^l)\|Q(D_s^m,D_t^{l-1},D_t^l))-\frac{1}{2}\log\det\left(\mathbf{I}_{\boldsymbol{\theta}^*}\right)\to\frac{3d-2j-c_l}{2}\log\frac{1}{2\pi e}+\log\frac{1}{\omega(\theta_s^*,\theta_t^*)} \tag{98}$$

where the Fisher information matrix is defined as,

$$\mathbf{I}_{\boldsymbol{\theta}^*}=\begin{bmatrix} mI_{cs}+n_{l-1}I_{ct,l-1}+n_lI_{ct,l} & n_lI_{cv,l}+n_{l-1}I_{cv,l-1} & mI_{sc} & n_{l-1}I_{ctr,l-1} & n_lI_{ctr,l}\\ n_lI_{ct,l}^T+n_{l-1}I_{ct,l-1}^T & n_lI_{v,l}+n_{l-1}I_{v,l-1} & \mathbf{0} & n_{l-1}I_{vt,l-1} & n_lI_{vt,l}\\ mI_{sc}^T & \mathbf{0} & mI_s & \mathbf{0} & \mathbf{0}\\ n_{l-1}I_{ctr,l-1}^T & n_{l-1}I_{vt,l-1}^T & \mathbf{0} & n_{l-1}I_{t,l-1} & \mathbf{0}\\ n_lI_{ctr,l}^T & n_lI_{vt,l}^T & \mathbf{0} & \mathbf{0} & n_lI_{t,l}\end{bmatrix} \tag{99}$$

as all n_l,n_{l-1} and m go to infinity. Similarly,

$$D(P_{\theta_s^*,\theta_{t,l-1}^*}(D_s^m,D_t^{l-1})\|Q(D_s^m,D_t^{l-1}))-\frac{1}{2}\log\det\left(\mathbf{I}_{\boldsymbol{\theta}_s^*}\right)\to\frac{2d-j}{2}\log\frac{1}{2\pi e}+\log\frac{1}{\omega(\theta_s^*,\theta_t^*)} \tag{100}$$

where,

$$\mathbf{I}_{\boldsymbol{\theta}_s^*}=\begin{bmatrix} mI_{cs}+n_{l-1}I_{ct,l-1} & n_{l-1}I_{cv,l-1} & mI_{sc} & n_{l-1}I_{ctr,l-1}\\ n_{l-1}I_{cv,l-1}^T & n_{l-1}I_{v,l-1} & \mathbf{0} & n_{l-1}I_{vt,l-1}\\ mI_{sc}^T & \mathbf{0} & mI_s & \mathbf{0}\\ n_{l-1}I_{ctr,l-1}^T & n_{l-1}I_{vt,l-1}^T & \mathbf{0} & n_{l-1}I_{t,l-1}\end{bmatrix} \tag{101}$$

Then by subtraction, we have,

$$\frac{\log\det(\mathbf{I}_{\boldsymbol{\theta}^*})}{\log\det(\mathbf{I}_{\boldsymbol{\theta}_s^*})}=\log\det\left(\mathbf{I}_{j\times j}+\frac{n_l}{m+n_{l-1}}\Delta_{ct}\Delta_{cst}^{-1}\right)+\log\det(\mathbf{I}_{c_l\times c_l}+\frac{n_l}{n_{l-1}}\Delta_t\Delta_{t-1}^{-1})+\log\det(n_lI_{t,l}(\theta_{tr,l}^*)) \tag{102}$$

where $\Delta_{ct}=I_{ct,l}-I_{ctr,l}I_{t,l}^{-1}I_{ctr,l}^T$, $\Delta_{cst}=\frac{m}{m+n_{l-1}}(I_{cs}-I_{sc}I_s^{-1}I_{sc}^T)+\frac{n_l}{m+n_{l-1}}(I_{ct,l-1}-I_{ctr,l-1}I_{t,l-1}^{-1}I_{ctr,l-1}^T)$, $\Delta_t=I_{v,l}-I_{vt,l}I_{t,l}^{-1}I_{vt,l}^T$, and $\Delta_{t-1}=I_{v,l-1}-I_{vt,l-1}I_{t,l-1}^{-1}I_{vt,l-1}^T$. Then we have the asymptotic estimation as,

$$\begin{aligned}I(D_t^l;\Theta_{t,l}=\theta_{t,l}^*|D_s^m,D_t^{l-1})\to&\log\det\left(\mathbf{I}_{j\times j}+\frac{n_l}{m+n_{l-1}}\Delta_{ct}\Delta_{cst}^{-1}\right)+\log\det(\mathbf{I}_{c_l\times c_l}+\frac{n_l}{n_{l-1}}\Delta_t\Delta_{t-1}^{-1})+\log\det(n_lI_{t,l}(\theta_{tr,l}^*))\\&+(d-j-c_l)\log\frac{1}{2\pi e}+\frac{2}{\omega(\theta_{t,l}^*|\theta_{t,l-1}^*,\theta_s^*)}\end{aligned} \tag{103}$$

□

Based on the theorem above, by putting things together, finally we reach,

$$\begin{aligned}\mathcal{R}(k)\le&M\Bigg(k\sum_{l=1}^k n_l\Big(\log\det\left(\mathbf{I}_{j\times j}+\frac{n_l}{m+n_{l-1}}\Delta_{ct}\Delta_{cst}^{-1}\right)+\log\det(\mathbf{I}_{c_l\times c_l}+\frac{n_l}{n_{l-1}}\Delta_t\Delta_{t-1}^{-1})+\log\det(n_lI_{t,l}(\theta_{tr,l}^*))\\&+(d-j-c_l)\log\frac{1}{2\pi e}+\frac{2}{\omega(\theta_{t,l}^*|\theta_{t,l-1}^*,\theta_s^*)}\Big)\Bigg)^{\frac{1}{2}}\end{aligned} \tag{104}$$

□

H. Extended Experiments on Bernoulli Examples

Let us take Bernoulli case as a simple example, let $P_{\theta_s^*}\sim\text{Ber}(\theta_s^*)$ and $P_{\theta_t^*}\sim\text{Ber}(\theta_t^*)$, that is, $P(Z_t^{(k)}=1)=\theta_t^*$ and $P(Z_s^{(k)}=1)=\theta_s^*$. Given source data D_s^m with m samples and n target samples D_t^n sequentially, the regrets can be explicitly written as,

$$D\left(P_{\theta_t^*,\theta_s^*}(D_t^n)\|Q(D_t^n)|D_s^m\right)=\sum_{D_s^m}\sum_{D_t^n}P_{\theta_t^*}\left(D_t^n\right)\log\left(\frac{P_{\theta_t^*}\left(D_t^n\right)}{Q\left(D_t^n|D_s^m\right)}\right) \tag{105}$$

where we define the distribution Q with the mixture strategy as

$$Q\left(D_t^n|D_s^m\right)=\frac{\int_0^1\int_0^1\omega(\theta_s,\theta_t)P_{\theta_t}(D_t^n)P_{\theta_s}(D_s^m)d\theta_sd\theta_t}{\int_0^1\omega(\theta_s)P_{\theta_s}(D_s^m)d\theta_s} \tag{106}$$

on account of the assumption $P_{\theta_s,\theta_t}(D_t^k,D_s^m)=P_{\theta_t}(D_t^k)P_{\theta_s}(D_s^m)$. Next we will exam the regrets with different types of priors $\omega(\theta_s,\theta_t)$ with different degrees of the correlation.

1) Weak Correlation: First we examine the weak correlation between θ_s and θ_t by assuming that the joint distribution $\omega(\theta_s, \theta_t) = \theta_s + \theta_t$, then the marginals $\omega(\theta_s) = \theta_s + \frac{1}{2}$ and $\omega(\theta_t) = \theta_t + \frac{1}{2}$. By 'weak correlation', we see this from the conditional distribution,

$$\omega(\theta_t|\theta_s) = \frac{\theta_s}{\theta_s + \frac{1}{2}} + \frac{\theta_t}{\theta_s + \frac{1}{2}} \tag{107}$$

for a given θ_s, $\omega(\theta_t|\theta_s)$ is still affine in θ_t and does not provide any structural knowledge over θ_t^*. The predictor distribution $Q(D_t^n|D_s^m)$ can be calculated as,

$$Q\left(D_t^n|D_s^m\right) = \frac{1}{(n+1)} \frac{1}{\binom{n}{k_t}} \frac{2k_s + 2 + (k_t+1)\frac{2(m+2)}{n+2}}{m + 2k_s + 4} \tag{108}$$

where we denote number of 1's received from the source and target by k_t and k_s, respectively. If m, n are sufficiently large and $E_{\theta_t^*, \theta_s^*}[k_s] = \theta_s^* m$ and $E_{\theta_t^*, \theta_s^*}[k_t] = \theta_t^* n$, we expect over a long sequence that the conditional mutual information can be approximated as,

$$E_{\theta_t^*, \theta_s^*}\left[\log \frac{P_{\theta_t^*}\left(D_t^n\right)}{Q\left(D_t^n|D_s^m\right)}\right] \leq \frac{1}{2}\log(n+1) + \frac{1}{2}\log\frac{1}{\pi\theta_t^*(1-\theta_t^*)} + \log\frac{\theta_s^* + \frac{1}{2}}{\theta_t^* + \theta_s^*} \tag{109}$$

Compared to the mixture distribution Q by the marginal $\omega(\theta_t) = \theta_t + \frac{1}{2}$ without introducing the source,

$$Q\left(D_t^n\right) = \int_0^1 (\theta_t + \frac{1}{2})(\theta_t)^{k_t}(1-\theta_t)^{n-k_t} d\theta_t = \frac{1}{(n+1)} \frac{1}{\binom{n}{k_t}} (\frac{1}{2} + \frac{k_t+1}{n+2}) \tag{110}$$

and the expected regrets induced by this predictor can be bounded as,

$$E_{\theta_t^*, \theta_s^*}\left[\log \frac{P_{\theta_t^*}\left(D_t^n\right)}{Q\left(D_t^n\right)}\right] \leq \frac{1}{2}\log(n+1) + \frac{1}{2}\log\frac{1}{\pi\theta_t^*(1-\theta_t^*)} + \log\frac{1}{\theta_t^* + \frac{1}{2}} \tag{111}$$

The difference of the bounds is $\log\frac{\omega(\theta_t)}{\omega(\theta_t|\theta_s)} = \log\frac{(\theta_s^*+\frac{1}{2})(\theta_s^*+\frac{1}{2})}{\theta_t^*+\theta_s^*}$, from this example, we do not take any specific assumptions over θ_t^* and θ_s^*, so the prior distribution $\omega(\theta_t|\theta_s)$ may not interpret the dependence correctly and the bound with the source can be either looser or tighter. Additionally, even when m goes to infinity and the source data does not change the convergence rate w.r.t. n, this result confirms the Theorem 3 numerically.

2) Strong Correlation: If we consider another extreme case where prior distribution $\omega(\theta_t|\theta_s) = \delta(\theta_s)$ and $\omega(\theta_s) = 1$, that is, θ_t and θ_s are totally dependent. Then the mixture is calculated as,

$$Q\left(D_t^n|D_s^m\right) = \frac{\frac{1}{n+m+1}\frac{1}{\binom{m+n}{k_s+k_t}}}{\frac{1}{m+1}\frac{1}{\binom{m}{k_s}}} \tag{112}$$

Analogously we expect over a large m, where $k_s = \theta_s^* m$ and $k_t = \theta_t^* n$ and the source samples are abundant, e.g., $m \gg n$ and we have,

$$\mathbb{E}_{\theta_s^*, \theta_t^*}\left[\log \frac{P_{\theta_s^*, \theta_t^*}(D_t^n)}{Q\left(D_t^n|D_s^m\right)}\right] \leq \log(1 + \frac{n}{m+1}) + n\left(\theta_t^* \log\frac{\theta_t^*}{\theta_s^*} + (1-\theta_t^*)\log\frac{1-\theta_t^*}{1-\theta_s^*}\right) + C \tag{113}$$

where C is some constant that depends on θ_t^* and θ_s^*. By introducing abundant source data, $\log(1+\frac{n}{m+1})$ term will vanish with the rate $O(\frac{n}{m})$ but it will also introduce the term $n\left(\theta_t^* \log\frac{\theta_t^*}{\theta_s^*} + (1-\theta_t^*)\log\frac{1-\theta_t^*}{1-\theta_s^*}\right)$ that grows linearly with n. This specific choice of prior is improper since $\omega(\theta_t^*|\theta_s^*) = 0$ whereas $\theta_t^* \neq \theta_s^*$, and the improper prior will finally lead to the inaccurate estimation of θ_t^* while both m and n are sufficiently large. This result is unsurprising since we can estimate θ_s^* accurately and $\omega(\theta_t\|\theta_s^*)$ enforces $\theta_t = \theta_s^*$ for all target data predictions, and the regret for each sample is $D(P_{\theta_t^*}\|P_{\theta_s^*})$. To see this, if m goes to infinity, the true regrets then can be calculated explicitly by,

$$\mathbb{E}_{\theta_t^*}\left[\log \frac{P_{\theta_t^*}(D_t^n)}{Q(D_t^n|D_s^m)}\right] = n\left(\theta_t^* \log\frac{\theta_t^*}{\theta_s^*} + (1-\theta_t^*)\log\frac{1-\theta_t^*}{1-\theta_s^*}\right) = nD(P_{\theta_t^*}\|P_{\theta_s^*}) \tag{114}$$

3) Correlation with Prior Knowledge: From both weak and strong correlation cases, we do not made any beneficial assumptions on θ_s and θ_t, which leads to less informative results or wrong estimation. Then how to choose appropriate ω for better predictions? Considering the true regrets are captured by the conditional prior $\omega(\theta_t|\theta_s)$, although we do not have the access to the true parameters, we may know the interrelation of the source and target parameters. In particular, for one dimension parameter, we make the following assumption.

Assumption 1 (Prior knowledge with ℓ_1 norm)**.** *For $\theta_s^*, \theta_t^* \in [0,1]$ and $c > 0$,*

$$|\theta_s^* - \theta_t^*| \leq c \tag{115}$$

This assumption implies that given θ_s^*, θ_t^* is not far away, and if θ_s^* can be approximated accurately, θ_t^* can be estimated more accurately with tighter support. We encode this particular relationship in terms of the conditional prior distribution $\omega(\theta_t|\theta_s)$, say, given any θ_s, θ_t is uniformly distributed over $[\theta_s - c, \theta_s + c]$ with density $\frac{1}{2c}$. Additionally we assume the marginal $\omega(\theta_s)$ is proper, e.g., θ_s^* can be estimated accurately with sufficient source samples. As a result, we can give the explicit expression of the mixture distribution $Q(D_t^n|D_s^m)$,

Lemma 1. *Under Assumption 1, given any θ_s, θ_t is uniformly distributed over $[\theta_s - c, \theta_s + c]$ with density $\frac{1}{2c}$, and θ_s^* can be estimated accurately via the source samples D_s^m, then*

$$Q(D_t^n|D_s^m) = \frac{1}{2c}\binom{n}{k_t}^{-1}\left(\sum_{i=1}^{k_t}\binom{n}{i}\frac{(\theta_s^*-c)^i(1-\theta_s^*+c)^{n-i+1}}{n-i+1} - \frac{(\theta_s^*+c)^i(1-\theta_s^*-c)^{n-i+1}}{n-i+1} + \frac{(1-\theta_s^*+c)^{n+1}-(1-\theta_s^*-c)^{n+1}}{n+1}\right) \tag{116}$$

Proof.

$$Q(D_t^n|D_s^m) = \int_{\theta_s}\int_{\theta_t|\theta_s} P(D_t^n|\theta_t)P(\theta_t|\theta_s)d\theta_t P(\theta_s|D_s^m)d\theta_s \tag{117}$$

$$= \int_{\theta_s^*-c}^{\theta_s^*+c}\frac{1}{2c}P(D_t^n|\theta_t)d\theta_t \tag{118}$$

$$= \frac{1}{2c}\int_{\theta_s^*-c}^{\theta_s^*+c}(\theta_t)^{k_t}(1-\theta_t)^{n-k_t}d\theta_t \tag{119}$$

We denote,

$$I(n,k) = \binom{n}{k}\int_0^a x^k(1-x)^{n-k}dx \tag{120}$$

Then we have,

$$I(n,k) = \binom{n}{k}\left(\left[\frac{-x^k(1-x)^{n-k+1}}{n-k+1}\right]_0^a + \frac{k}{n-k+1}\int_0^a x^{k-1}(1-x)^{n-k+1}dx\right) \tag{121}$$

$$= \binom{n}{k}\frac{-a^k(1-a)^{n-k+1}}{n-k+1} + \binom{n}{k}\frac{k}{n-k+1}\binom{n}{k-1}^{-1}I(n,k-1) \tag{122}$$

$$= \binom{n}{k}\frac{-a^k(1-a)^{n-k+1}}{n-k+1} + I(n,k-1) \tag{123}$$

By induction,

$$I(n,k) = \sum_{i=1}^{k}\binom{n}{i}\frac{-a^i(1-a)^{n-i+1}}{n-i+1} + I(a,0) \tag{124}$$

$$= \sum_{i=1}^{k}\binom{n}{i}\frac{-a^i(1-a)^{n-i+1}}{n-i+1} + \frac{1-(1-a)^{n+1}}{n+1} \tag{125}$$

Hence,

$$\int_0^a x^k(1-x)^{n-k}dx = \binom{n}{k}^{-1}\left(\sum_{i=1}^{k}\binom{n}{i}\frac{-a^i(1-a)^{n-i+1}}{n-i+1} + \frac{1-(1-a)^{n+1}}{n+1}\right) \tag{126}$$

and for any $b > a$,

$$\int_0^b x^k(1-x)^{n-k}dx = \binom{n}{k}^{-1}\left(\sum_{i=1}^{k}\binom{n}{i}\frac{-b^i(1-b)^{n-i+1}}{n-i+1} + \frac{1-(1-b)^{n+1}}{n+1}\right) \tag{127}$$

By subtraction,

$$\frac{1}{b-a}\int_a^b x^k(1-x)^{n-k}dx = \frac{1}{b-a}\binom{n}{k}^{-1}\left(\sum_{i=1}^{k}\binom{n}{i}\frac{a^i(1-a)^{n-i+1}-b^i(1-b)^{n-i+1}}{n-i+1}+\frac{(1-a)^{n+1}-(1-b)^{n+1}}{n+1}\right) \tag{128}$$

□

Remark 1. *It is relatively hard to directly tell the effect of c in the prediction, in consideration of this closed form distribution, we can calculate $Q(D_t^n|D_s^m)$ explicitly if all the parameters θ_s^*, c, n, k_t are known, thereby we shall verify the intuition by conducting the experiments with different settings, which are presented in the next section.*

4) Toy Experiments: To confirm the numerical calculations of Bernoulli cases under Assumption 1, some experiments are conducted and results are presented. Consider the following specific settings, the true parameter $\theta_t^* = \frac{1}{3}$ is unknown, but we know the prior knowledge such that the source parameter and target parameter satisfy the relationship $|\theta_s - \theta_t| \leq c$ for some $c > 0$, this prior knowledge may or may not be correct. Further we assume $\omega(\theta_s) = 1$ over $[0, 1]$ and given θ_s, θ_t is distributed uniformly over $[\theta_s - c, \theta_s + c]$. As the target arrives sequentially and we have sufficient source data with $m = 100000$, we plot the predicted probability $P(z_t^{(k)}|D_t^{k-1}, D_s^m)$ in the figure below for a single trial with different θ_s^* and c.

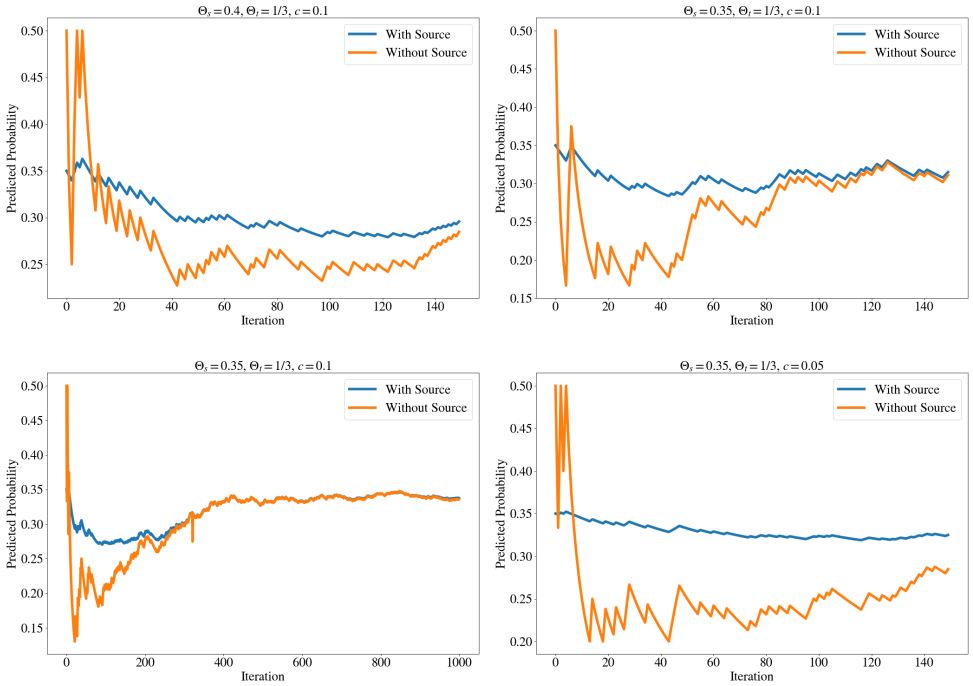

Fig. 1. Predicted probability $P(z_t^{(k)}|D_t^{k-1}, D_s^m)$ with single trial, the prediction the curves that are closer to $\theta_t^* = \frac{1}{3}$ entails more accurate estimation

From the figures, we observe that when k is relatively small, the prior knowledge about the source and the target distribution can help estimate of the true value of θ_t better than without introducing the source data. It can also be seen from the third figure, when k is relatively large, both estimates of θ_t^* with or without the source are already fairly close. The single trial does not fully reflect the usefulness of the source data, we next examine the posterior distribution $P(\theta_t|D_t^k, D_s^m)$ with source and $P(\theta_t|D_t^k)$ without source after receiving 100 target data with different θ_s and c to see the effect of prior knowledge.

From the comparison, the posterior distribution of θ_t with the source data is more concentrated and moreover, closer θ_s^* and smaller c will yield more concentrated density around θ_t^*, which fits in line with our intuition. To evaluate the expected regrets, we repeat the experiments 2000 times and take the average for different number of target samples. The results are shown as follows.

The result reflects the influence of c and the gaps between the true parameter θ_s^* and θ_t^*. When $\omega(\theta_t|\theta_s^*)$ is proper, e.g., the density is concentrated around θ_t^*, then smaller c will yield lower regrets. However, if the conditional prior is improper ($[\theta_s^* - c, \theta_s^* + c]$ does not cover θ_t^*), the regrets are determined by both c and the distance $|\theta_s^* - \theta_c^*|$, for example, compare the case $\theta_s^* = 0.4, c = 0.1$ with $\theta_s^* = 0.4, c = 0.05$, the former case is the proper, while the latter does not cover the true θ_t^*, so

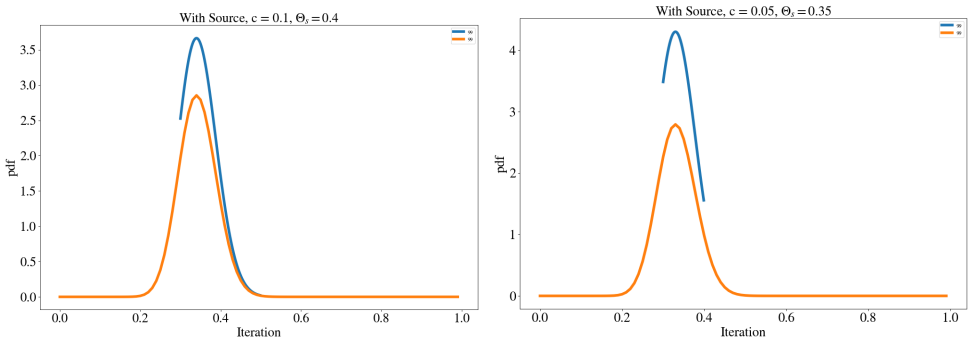

Fig. 2. Posterior distribution of θ_t with (blue) and without (orange) the source data

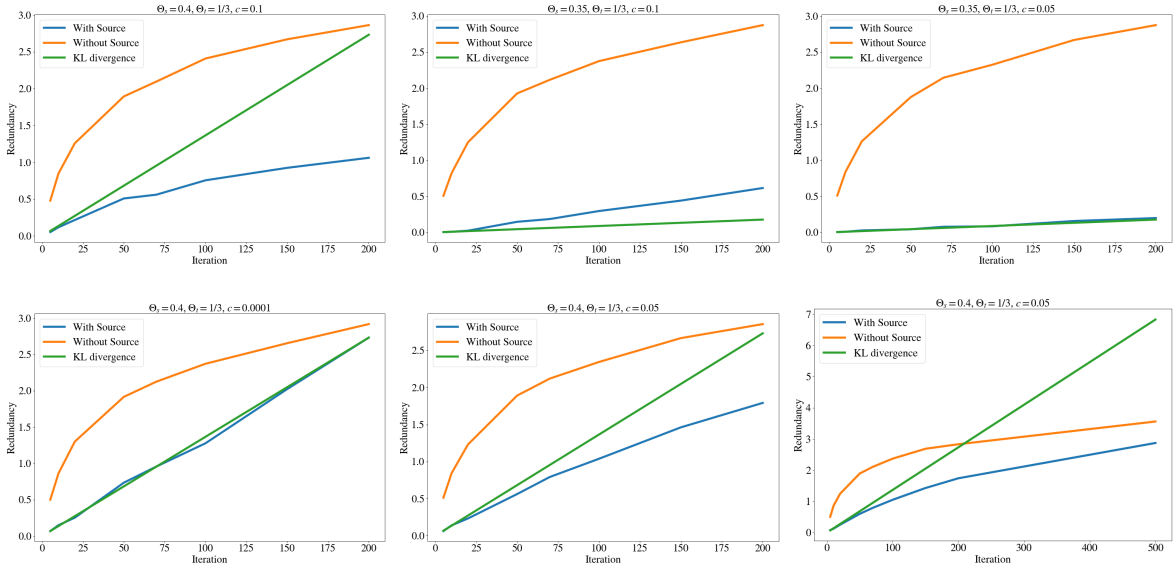

Fig. 3. Expected regrets by 2000 repeats with different c and θ_s^*. Orange, blue and green curves represents the expected regrets without the source, with the source, and $kD(P_{\theta_t^*}\|P_{\theta_s^*})$, respectively.

the worse regrets. In addition, if c is small enough ($c = 0.001$), the estimation of θ_t will be centered at θ_s^* thus the regrets will coincide with the KL divergence $kD(\theta_t^*\|\theta_s^*)$, which confirms the case we discussed in section -H2. Overall, once the prior information $\omega(\theta_t|\theta_s)$ is located around θ_t^* and the target samples are inadequate to make accurate prediction, the knowledge transfer is sensible and indeed the regrets can be further optimized.

REFERENCES

[1] 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.
[2] B. S. Clarke, "Asymptotic normality of the posterior in relative entropy," *IEEE Transactions on Information Theory*, vol. 45, no. 1, pp. 165–176, 1999.
[3] P. D. Powell, "Calculating determinants of block matrices," *arXiv preprint arXiv:1112.4379*, 2011.