Utilitarian Online Learning from Open-World Soft Sensing

I. Proof of Theorem

To proof our theorem, we first introduce the triangle inequality for classification error [1], [2] which implies that $\epsilon(h_1, h_2) \leq \epsilon(h_1, h_3) + \epsilon(h_2, h_3)$. Then, we have:

$$\begin{aligned}
\epsilon_{\mathbb{R}^k_{t+1}}(h) &\leq \epsilon_{\mathbb{R}^k_{t+1}}(h^*) + \epsilon_{\mathbb{R}^k_{t+1}}(h, h^*), \\
&= \epsilon_{\mathbb{R}^k_{t+1}}(h^*) + \epsilon_{\mathbb{R}^k_{t+1}}(h, h^*) \\
&\qquad + \epsilon_{\mathbb{R}^k_t}(h, h^*) - \epsilon_{\mathbb{R}^k_t}(h, h^*), \\
&\leq \epsilon_{\mathbb{R}^k_{t+1}}(h^*) + \epsilon_{\mathbb{R}^k_t}(h, h^*) \\
&\qquad + \left|\epsilon_{\mathbb{R}^k_{t+1}}(h, h^*) - \epsilon_{\mathbb{R}^k_t}(h, h^*)\right|.
\end{aligned} \tag{1}$$

To proceed with the proof, we adapt the definition and inequality suggested by [3] as follows:

Definition 1. *For a hypothesis space $\mathcal{H}$, the* symmetric difference hypothesis space $\mathcal{H}\Delta\mathcal{H}$ *is the set of hyperspheres*

$$g \in \mathcal{H}\Delta\mathcal{H} \iff g(\mathbf{x}) = h(\mathbf{x}) \oplus h'(\mathbf{x}) \quad \text{for some } h, h' \in \mathcal{H},$$

where $\oplus$ is the XOR function, determining whether the outcomes of two functions h and h' are equal.

If the maximum discrepancy between two functions across two spaces is founded, then this value defines the H-divergence distance of two spaces as follows:

Lemma 1. *For any hyperspheres $h, h' \in \mathcal{H}$,*

$$\left|\epsilon_{\mathbb{R}^k_t}(h, h') - \epsilon_{\mathbb{R}^k_{t+1}}(h, h')\right| \leq \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}\left(\mathbb{R}^k_{t+1}, \mathbb{R}^k_t\right).$$

So, by Lemma 1, we have:

$$\begin{aligned}
\epsilon_{\mathbb{R}^k_{t+1}}(h) &\leq \epsilon_{\mathbb{R}^k_{t+1}}(h^*) + \epsilon_{\mathbb{R}^k_t}(h, h^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathbb{R}^k_{t+1}, \mathbb{R}^k_t), \\
&\leq \epsilon_{\mathbb{R}^k_{t+1}}(h^*) + \epsilon_{\mathbb{R}^k_t}(h) + \epsilon_{\mathbb{R}^k_t}(h^*) \\
&\qquad + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathbb{R}^k_{t+1}, \mathbb{R}^k_t), \\
&= \epsilon_{\mathbb{R}^k_t}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathbb{R}^k_{t+1}, \mathbb{R}^k_t) + \gamma.
\end{aligned} \tag{2}$$

With adapting Lemma 2 proposed by [3], the H-divergence distance between two spaces $\mathbb{R}^k_{t+1}$ and $\mathbb{R}^k_t$ can be estimated using a finite number of samples extracted separately from each space as follows:

Lemma 2. *Let $\mathcal{H}$ be a hypothesis space on data $\mathcal{X}$ with VC dimension d. $|\mathbb{R}^k_t|$ and $|\mathbb{R}^k_{t+1}|$ are samples of size n from two spaces $\mathbb{R}^k_t$ and $\mathbb{R}^k_{t+1}$ respectively and $d_{\mathcal{H}}\left(|\mathbb{R}^k_t|, |\mathbb{R}^k_{t+1}|\right)$ is the $\mathcal{H}$-divergence between samples, then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$d_{\mathcal{H}}\left(\mathbb{R}^k_t, \mathbb{R}^k_{t+1}\right) \leq d_{\mathcal{H}}\left(|\mathbb{R}^k_t|, |\mathbb{R}^k_{t+1}|\right) + 4\sqrt{\frac{d\log(2n) + \log\left(\frac{2}{\delta}\right)}{n}}.$$

combining Lemma 2 with Eq. (2), we arrive at:

$$\begin{aligned}
\epsilon_{\mathbb{R}^k_{t+1}}(h) &\leq \epsilon_{\mathbb{R}^k_t}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}\left(|\mathbb{R}^k_t|, |\mathbb{R}^k_{t+1}|\right) \\
&\qquad + 4\sqrt{\frac{d\log(2n) + \log\left(\frac{2}{\delta}\right)}{4n}} + \gamma,
\end{aligned} \tag{3}$$

as desired. □

References

[1] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *NeuIPS*, vol. 19. MIT Press, 2006.

[2] K. Crammer, M. Kearns, and J. Wortman, "Learning from multiple sources." *Journal of Machine Learning Research*, vol. 9, no. 8, 2008.

[3] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, pp. 151–175, 2010.