Appendix: Learning Bounds for Open-Set Learning

- Appendix A recalls some important definitions and concepts.
- Appendix B provides the proof for Theorem 1.
- Appendix C provides the proof for Theorem 2.
- Appendix D provides the proof for Theorem 3.
- Appendix E provides the proofs for Theorems 4, 5 and 6.
- Appendix F provides details on datasets and parameter analysis.

1. Appendix A: Notations and Concepts

In this section, we introduce the definition of open-set learning and then introduce important concepts used in this paper.

Let $\mathcal{X} \subset \mathbb{R}^d$ be a feature space and $\mathcal{Y} := \{\mathbf{y}_c\}_{c=1}^{C+1}$ be the label space, where the label $\mathbf{y}_c$ is a one-hot vector whose c-th coordinate is 1 and the other coordinate is 0.

Definition 1 (Domain, Known and Unknown Classes.)**.** *Given random variable $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, a domain is a joint distribution $P_{X,Y}$. The classes from $\mathcal{Y}_k := \{\mathbf{y}_c\}_{c=1}^{C}$ is called known class and $\mathbf{y}_{C+1}$ is called unknown classes.*

The open set learning problem is defined as follows.

Problem 1 (Open-Set Learning)**.** *Given independent and identically distributed (i.i.d.) samples $S = \{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^n$ drawn from $P_{X,Y|Y\in\mathcal{Y}_k}$. Aim of open-set learning is to train a classifier using S such that f can classify 1) the sample from known classes into correct known classes; 2) the sample from unknown classes into unknown classes.*

Table 1. Main notations and their descriptions.

Notation	Description
$\mathcal{X}, \mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^{C+1}, \mathcal{Y}_k = \{\mathbf{y}_i\}_{i=1}^{C}$	feature space, label space, label space for known classes
X, Y	random variables on the feature space $\mathcal{X}$ and $\mathcal{Y}$
$P_{X,Y},\ Q_{X,Y}$	joint distributions
$P_X,\ Q_X$	marginal distributions
$P_{X,Y\|Y\in\mathcal{Y}_k}, Q_{X,Y\|Y\in\mathcal{Y}_k}$	conditional distributions when label belongs to known classes
$P_{X\|Y=\mathbf{y}_{C+1}}, Q_{X\|Y=\mathbf{y}_{C+1}}$	conditional distributions when label belongs to unknown classes
R_P^α, R_Q^α	α-risks corresponding to $P_{X,Y}, Q_{X,Y}$
$R_{P,k}$, $R_{Q,k}$	partial risks for known classes corresponding to $P_{X,Y}, Q_{X,Y}$
$R_{P,u}$, $R_{Q,u}$	partial risks for unknown classes corresponding to $P_{X,Y}, Q_{X,Y}$
$\boldsymbol{h}$	hypothesis function from $\mathcal{X} \rightarrow \mathbb{R}^{C+1}$
$\mathcal{H}$	hypothesis space, a subset of $\{\boldsymbol{h} : \mathcal{X} \rightarrow \mathbb{R}^{C+1}\}$
$\mathcal{H}_K$	RKHS with kernel K
U	auxiliary distribution defined over $\mathcal{X}$
$Q_U^{0,\beta} P_{Y\|X}$	ideal auxiliary domain defined over $\mathcal{X} \times \mathcal{Y}$
$\widehat{Q}_U^{\tau,\beta}$	the approximation of $Q_U^{0,\beta}$
w	weights
S, T	samples drawn from $P_{X,Y}$ and Q_X, respectively
n, m	sizes of samples S and T
$d_{\boldsymbol{h},\mathcal{H}}^{\ell}, \Lambda$	disparity discrepancy, combined risk
$\widehat{R}_{S,T}^{\tau,\beta}$,$\widetilde{R}_{S,T}^{\tau,\beta}$	auxiliary risk, proxy of auxiliary risk

2. Appendix B: Proof of Theorem 1

Proof of Theorem. 1.

$$
\begin{aligned}
&|R_P^\alpha(\boldsymbol{h}) - R_Q^\alpha(\boldsymbol{h})| = |(1-\alpha)R_{P,k}(\boldsymbol{h}) + \alpha R_{P,u}(\boldsymbol{h}) - (1-\alpha)R_{Q,k}(\boldsymbol{h}) - \alpha R_{Q,u}(\boldsymbol{h})|\\
=&\Big|(1-\alpha)\int_{\mathcal{X}\times\mathcal{Y}_k} \ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y})\mathrm{d}P_{X,Y|Y\in\mathcal{Y}_k}(\mathbf{x},\mathbf{y}) + \alpha R_{P,u}(\boldsymbol{h}) - (1-\alpha)\int_{\mathcal{X}\times\mathcal{Y}_k} \ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y})\mathrm{d}Q_{X,Y|Y\in\mathcal{Y}_k}(\mathbf{x},\mathbf{y}) - \alpha R_{Q,u}(\boldsymbol{h})\Big|\\
=&\alpha\big|R_{P,u}(\boldsymbol{h}) - R_{Q,u}(\boldsymbol{h})\big| \quad \text{we have used } Q_{X,Y|Y\in\mathcal{Y}_k} = P_{X,Y|Y\in\mathcal{Y}_k}\\
=&\alpha\big|\int_{\mathcal{X}} \ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}P_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x}) - \int_{\mathcal{X}} \ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}Q_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x})\big|\\
\leq&\alpha\big|\int_{\mathcal{X}} \ell(\boldsymbol{h}(\mathbf{x}),\boldsymbol{h}'(\mathbf{x}))\mathrm{d}P_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x}) - \int_{\mathcal{X}} \ell(\boldsymbol{h}(\mathbf{x}),\boldsymbol{h}'(\mathbf{x}))\mathrm{d}Q_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x})\big|\\
+&\alpha\int_{\mathcal{X}} \ell(\boldsymbol{h}'(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}P_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x}) + \alpha\int_{\mathcal{X}} \ell(\boldsymbol{h}'(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}Q_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x}) \quad \text{the triangle inequality is used}\\
\leq&\alpha d^\ell_{\boldsymbol{h},\mathcal{H}}(P_{X|Y=\mathbf{y}_{C+1}}, Q_{X|Y=\mathbf{y}_{C+1}}) + \alpha\int_{\mathcal{X}} \ell(\boldsymbol{h}'(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}P_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x}) + \alpha\int_{\mathcal{X}} \ell(\boldsymbol{h}'(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}Q_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x}).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&|R_P^\alpha(\boldsymbol{h}) - R_Q^\alpha(\boldsymbol{h})| = \min_{\boldsymbol{h}'\in\mathcal{H}} |R_P^\alpha(\boldsymbol{h}) - R_Q^\alpha(\boldsymbol{h})| \text{ Note that we minimize } \boldsymbol{h}', \text{ but not } \boldsymbol{h}\\
\leq&\min_{\boldsymbol{h}'\in\mathcal{H}} \big(\alpha d^\ell_{\boldsymbol{h},\mathcal{H}}(P_{X|Y=\mathbf{y}_{C+1}}, Q_{X|Y=\mathbf{y}_{C+1}}) + \alpha\int_{\mathcal{X}} \ell(\boldsymbol{h}'(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}P_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x}) + \alpha\int_{\mathcal{X}} \ell(\boldsymbol{h}'(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}Q_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x})\big)\\
\leq&\alpha d^\ell_{\boldsymbol{h},\mathcal{H}}(P_{X|Y=\mathbf{y}_{C+1}}, Q_{X|Y=\mathbf{y}_{C+1}}) + \alpha\Lambda.
\end{aligned}
$$

□

3. Appendix C: Proof of Theorem 2

Proof of Theorem 2. **Step 1.** Note that

$$\int_{\mathcal{X}\times\mathcal{Y}} \ell(\phi\circ\tilde{\boldsymbol{h}}(\mathbf{x}),\phi(\mathbf{y}))\mathrm{d}P_{Y|X}(\mathbf{x})\mathrm{d}\tilde{P}(\mathbf{x})=0,$$

hence, if we set $\tilde{P}_{X,Y}=\tilde{P}P_{Y|X}$, then

$$\int_{\mathcal{X}\times\mathcal{Y}} \ell(\phi\circ\tilde{\boldsymbol{h}}(\mathbf{x}),\phi(\mathbf{y}))\mathrm{d}\tilde{P}_{X,Y|Y\in\mathcal{Y}_k}(\mathbf{x},\mathbf{y})=0,\quad \int_{\mathcal{X}} \ell(\phi\circ\tilde{\boldsymbol{h}}(\mathbf{x}),\phi(\mathbf{y}_{C+1}))\mathrm{d}\tilde{P}_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x})=0.$$

Note that $\ell(\mathbf{y},\mathbf{y}')=0$ iff $\mathbf{y}=\mathbf{y}'$, hence, $\tilde{\boldsymbol{h}}(\mathbf{x})=\mathbf{y}_{C+1}$, for $\mathbf{x}\in\mathrm{supp}\ \tilde{P}_{X|Y=\mathbf{y}_{C+1}}$ a.e. $\tilde{P}$ and $\tilde{\boldsymbol{h}}(\mathbf{x})\neq\mathbf{y}_{C+1}$, for $\mathbf{x}\in\mathrm{supp}\ \tilde{P}_{X|Y\in\mathcal{Y}_k}$ a.e. $\tilde{P}$.

Step 2. Because $P_X\ll Q_X\ll\tilde{P}$, then,

$$\mathrm{supp}\ \tilde{P}_{X|Y\in\mathcal{Y}_k}\supset\mathrm{supp}\ Q_{X|Y\in\mathcal{Y}_k}\supset\mathrm{supp}\ P_{X|Y\in\mathcal{Y}_k}$$

and

$$\mathrm{supp}\ \tilde{P}_{X|Y=\mathbf{y}_{C+1}}\supset\mathrm{supp}\ Q_{X|Y=\mathbf{y}_{C+1}}\supset\mathrm{supp}\ P_{X|Y=\mathbf{y}_{C+1}}.$$

Step 3. We need to check that $\min_{\boldsymbol{h}\in\mathcal{H}}R_P^{\alpha}(\boldsymbol{h})=(1-\alpha)\min_{\boldsymbol{h}\in\mathcal{H}}R_{P,k}(\boldsymbol{h})$. First, it is clear that $\min_{\boldsymbol{h}\in\mathcal{H}}R_P^{\alpha}(\boldsymbol{h})\geq(1-\alpha)\min_{\boldsymbol{h}\in\mathcal{H}}R_{P,k}(\boldsymbol{h})$. If there exists $\boldsymbol{h}_P\in\mathcal{H}$ such that $\min_{\boldsymbol{h}\in\mathcal{H}}R_P^{\alpha}(\boldsymbol{h})>(1-\alpha)\min_{\boldsymbol{h}\in\mathcal{H}}R_{P,k}(\boldsymbol{h}_P)$.

Set

$$\tilde{\boldsymbol{h}}_P(\mathbf{x})=\mathbf{y}_{C+1},\ \text{ if }\ \tilde{\boldsymbol{h}}(\mathbf{x})=\mathbf{y}_{C+1};\ \text{ otherwise, }\ \tilde{\boldsymbol{h}}_P(\mathbf{x})=\boldsymbol{h}_P(\mathbf{x}),$$

hence, using the results of Step 1 and Step 2, we know $\{\mathbf{x}:\tilde{\boldsymbol{h}}(\mathbf{x})=\mathbf{y}_{C+1}\}\supset\mathrm{supp}\ P_{X|Y=\mathbf{y}_{C+1}}$. Then,

$$\begin{aligned}
&(1-\alpha)\int_{\mathcal{X}\times\mathcal{Y}_k}\ell(\boldsymbol{h}_P(\mathbf{x}),\mathbf{y})\mathrm{d}P_{X,Y|Y\in\mathcal{Y}_k}(\mathbf{x},\mathbf{y})\\
=&(1-\alpha)\int_{\{\mathrm{supp}\ P_{X|Y\in\mathcal{Y}_k}\}\times\mathcal{Y}_k}\ell(\boldsymbol{h}_P(\mathbf{x}),\mathbf{y})\mathrm{d}P_{X,Y|Y\in\mathcal{Y}_k}(\mathbf{x},\mathbf{y})\\
=&(1-\alpha)\int_{\{\mathrm{supp}\ P_{X|Y\in\mathcal{Y}_k}\}\times\mathcal{Y}_k}\ell(\tilde{\boldsymbol{h}}_P(\mathbf{x}),\mathbf{y})\mathrm{d}P_{X,Y|Y\in\mathcal{Y}_k}(\mathbf{x},\mathbf{y})\ \text{ have used }\tilde{\boldsymbol{h}}(\mathbf{x})\neq\mathbf{y}_{C+1},\ \text{for }\mathbf{x}\in\mathrm{supp}\ \tilde{P}_{X|Y\in\mathcal{Y}_k}\text{ a.e. }\tilde{P}\\
=&(1-\alpha)\int_{\{\mathrm{supp}\ P_{X|Y\in\mathcal{Y}_k}\}\times\mathcal{Y}_k}\ell(\tilde{\boldsymbol{h}}_P(\mathbf{x}),\mathbf{y})\mathrm{d}P_{X,Y|Y\in\mathcal{Y}_k}(\mathbf{x},\mathbf{y})+0\\
=&(1-\alpha)\int_{\{\mathrm{supp}\ P_{X|Y\in\mathcal{Y}_k}\}\times\mathcal{Y}_k}\ell(\tilde{\boldsymbol{h}}_P(\mathbf{x}),\mathbf{y})\mathrm{d}P_{X,Y|Y\in\mathcal{Y}_k}(\mathbf{x},\mathbf{y})+\alpha\int_{\mathrm{supp}\ P_{X|Y=\mathbf{y}_{C+1}}}\ell(\tilde{\boldsymbol{h}}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}P_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x})\\
=&(1-\alpha)\int_{\{\mathrm{supp}\ P_{X|Y\in\mathcal{Y}_k}\}\times\mathcal{Y}_k}\ell(\tilde{\boldsymbol{h}}_P(\mathbf{x}),\mathbf{y})\mathrm{d}P_{X,Y|Y\in\mathcal{Y}_k}(\mathbf{x},\mathbf{y})+\alpha\int_{\mathrm{supp}\ P_{X|Y=\mathbf{y}_{C+1}}}\ell(\tilde{\boldsymbol{h}}_P(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}P_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x})\\
=&R_P^{\alpha}(\tilde{\boldsymbol{h}}_P)\geq\min_{\boldsymbol{h}\in\mathcal{H}}R_P^{\alpha}(\boldsymbol{h}),
\end{aligned}$$

hence, $\min_{\boldsymbol{h}\in\mathcal{H}}R_P^{\alpha}(\boldsymbol{h})=(1-\alpha)\min_{\boldsymbol{h}\in\mathcal{H}}R_{P,k}(\boldsymbol{h})$. Similarly, we can prove that $\min_{\boldsymbol{h}\in\mathcal{H}}R_Q^{\alpha}(\boldsymbol{h})=(1-\alpha)\min_{\boldsymbol{h}\in\mathcal{H}}R_{Q,k}(\boldsymbol{h})$. Because $Q_{X|Y\in\mathcal{Y}_k}=P_{X|Y\in\mathcal{Y}_k}$, hence, $\min_{\boldsymbol{h}\in\mathcal{H}}R_{Q,k}(\boldsymbol{h})=\min_{\boldsymbol{h}\in\mathcal{H}}R_{P,k}(\boldsymbol{h})$. Using the results of Step 3, we obtain that

$$\min_{\boldsymbol{h}\in\mathcal{H}}R_Q(\boldsymbol{h})=\min_{\boldsymbol{h}\in\mathcal{H}}R_P(\boldsymbol{h}).\tag{1}$$

Step 4. Given any $\boldsymbol{h}^*\in\arg\min_{\boldsymbol{h}\in\mathcal{H}}R_P^{\alpha}(\boldsymbol{h})$, then we construct $\tilde{\boldsymbol{h}}^*$ such that

$$\tilde{\boldsymbol{h}}^*(\mathbf{x})=\mathbf{y}_{C+1},\ \text{ if }\ \tilde{\boldsymbol{h}}(\mathbf{x})=\mathbf{y}_{C+1};\ \text{ otherwise, }\ \tilde{\boldsymbol{h}}^*(\mathbf{x})=\boldsymbol{h}^*(\mathbf{x}).$$

It is clear that $\tilde{\boldsymbol{h}}^* \in \mathcal{H}$ according to Assumption 1.

Then,

$$
\begin{aligned}
& R_P^{\alpha}(\boldsymbol{h}^*) \\
\geq & (1-\alpha)\int_{\mathcal{X}\times\mathcal{Y}_k} \ell(\boldsymbol{h}^*(\mathbf{x}),\mathbf{y})\mathrm{d}P_{X,Y|Y\in\mathcal{Y}_k}(\mathbf{x},\mathbf{y}) \\
= & (1-\alpha)\int_{\{\text{supp } P_{X|Y\in\mathcal{Y}_k}\}\times\mathcal{Y}_k} \ell(\boldsymbol{h}^*(\mathbf{x}),\mathbf{y})\mathrm{d}P_{X,Y|Y\in\mathcal{Y}_k}(\mathbf{x},\mathbf{y}) \\
= & (1-\alpha)\int_{\{\text{supp } P_{X|Y\in\mathcal{Y}_k}\}\times\mathcal{Y}_k} \ell(\tilde{\boldsymbol{h}}^*(\mathbf{x}),\mathbf{y})\mathrm{d}P_{X,Y|Y\in\mathcal{Y}_k}(\mathbf{x},\mathbf{y}) \ \text{ have used } \tilde{\boldsymbol{h}}(\mathbf{x}) \neq \mathbf{y}_{C+1}, \text{ for } \mathbf{x} \in \text{supp } \tilde{P}_{X|Y\in\mathcal{Y}_k} \text{ a.e. } \tilde{P} \\
= & (1-\alpha)\int_{\{\text{supp } P_{X|Y\in\mathcal{Y}_k}\}\times\mathcal{Y}_k} \ell(\tilde{\boldsymbol{h}}^*(\mathbf{x}),\mathbf{y})\mathrm{d}P_{X,Y|Y\in\mathcal{Y}_k}(\mathbf{x},\mathbf{y}) + 0 \\
= & (1-\alpha)\int_{\{\text{supp } P_{X|Y\in\mathcal{Y}_k}\}\times\mathcal{Y}_k} \ell(\tilde{\boldsymbol{h}}^*(\mathbf{x}),\mathbf{y})\mathrm{d}P_{X,Y|Y\in\mathcal{Y}_k}(\mathbf{x},\mathbf{y}) + \alpha\int_{\text{supp } P_{X|Y=\mathbf{y}_{C+1}}} \ell(\tilde{\boldsymbol{h}}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}P_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x}) \\
= & (1-\alpha)\int_{\{\text{supp } P_{X|Y\in\mathcal{Y}_k}\}\times\mathcal{Y}_k} \ell(\tilde{\boldsymbol{h}}^*(\mathbf{x}),\mathbf{y})\mathrm{d}P_{X,Y|Y\in\mathcal{Y}_k}(\mathbf{x},\mathbf{y}) + \alpha\int_{\text{supp } P_{X|Y=\mathbf{y}_{C+1}}} \ell(\tilde{\boldsymbol{h}}^*(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}P_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x}) \\
= & R_P^{\alpha}(\tilde{\boldsymbol{h}}^*).
\end{aligned}
$$

Hence, for any $\boldsymbol{h}^* \in \arg\min_{\boldsymbol{h}\in\mathcal{H}} R_P^{\alpha}(\boldsymbol{h})$,

$$\int_{\mathcal{X}} \ell(\boldsymbol{h}^*(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}P_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x}) = \int_{\text{supp } P_{X|Y=\mathbf{y}_{C+1}}} \ell(\boldsymbol{h}^*(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}P_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x}) = 0.$$

Similarly, we can prove that for any $\boldsymbol{h}^* \in \arg\min_{\boldsymbol{h}\in\mathcal{H}} R_Q^{\alpha}(\boldsymbol{h})$,

$$\int_{\mathcal{X}} \ell(\boldsymbol{h}^*(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}Q_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x}) = 0.$$

Step 5. Given any $h_Q \in \arg\min_{\boldsymbol{h}\in\mathcal{H}} R_Q^{\alpha}(\boldsymbol{h})$, we can find that (using result of Step 3)

$$R_Q^{\alpha}(h_Q) = (1-\alpha)R_{Q,k}(h_Q) = (1-\alpha)R_{P,k}(h_Q),$$

and

$$\int_{\mathcal{X}} \ell(\boldsymbol{h}_Q(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}Q_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x}) = 0.$$

Because $P_X \ll Q_X$, we know

$$P_{X|Y=\mathbf{y}_{C+1}} \ll Q_{X|Y=\mathbf{y}_{C+1}},$$

which implies that

$$\int_{\mathcal{X}} \ell(\boldsymbol{h}_Q(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}P_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x}) = 0.$$

Hence,

$$R_Q^{\alpha}(\boldsymbol{h}_Q) = (1-\alpha)R_{Q,k}(\boldsymbol{h}_Q) = (1-\alpha)R_{P,k}(\boldsymbol{h}_Q) + \alpha * 0 = R_P^{\alpha}(\boldsymbol{h}_Q).$$

Using the result (see Eq. (1)) of Step 3,

$$\min_{\boldsymbol{h}\in\mathcal{H}} R_Q^{\alpha}(\boldsymbol{h}) = \min_{\boldsymbol{h}\in\mathcal{H}} R_P^{\alpha}(\boldsymbol{h}).$$

We obtain that

$$\boldsymbol{h}_Q \in \arg\min_{\boldsymbol{h}\in\mathcal{H}} R_P^{\alpha}(\boldsymbol{h}),$$

this implies

$$\arg\min_{\boldsymbol{h}\in\mathcal{H}} R_Q^{\alpha}(\boldsymbol{h}) \subset \arg\min_{\boldsymbol{h}\in\mathcal{H}} R_P^{\alpha}(\boldsymbol{h}).$$

□

4. Appendix D: Proof of Theorem 3

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

$$R_Q^{\alpha}(\boldsymbol{h}) = (1-\alpha)R_{P,k}(\boldsymbol{h}) + \max\{R_Q(\boldsymbol{h}, \mathbf{y}_{C+1}) - (1-\alpha)R_{P,k}(\boldsymbol{h}, \mathbf{y}_{C+1}), 0\},$$

where $\alpha = Q(Y = \mathbf{y}_{C+1})$,

$$R_Q(\boldsymbol{h}, \mathbf{y}_{C+1}) = \int_{\mathcal{X}} \ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1}) \mathrm{d}Q_X(\mathbf{x}),$$

and

$$R_{P,k}(\boldsymbol{h}, \mathbf{y}_{C+1}) = \int_{\mathcal{X}} \ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1}) \mathrm{d}P_{X|Y \in \mathcal{Y}_k}(\mathbf{x}),$$

Proof. **Step 1.** We claim that $R_Q^{\alpha}(\boldsymbol{h}) = (1-\alpha)R_{P,k}(\boldsymbol{h}) + \alpha R_{Q,u}(\boldsymbol{h})$.

First, it is clear that

$$R_Q^{\alpha}(\boldsymbol{h}) = (1-\alpha)R_{Q,k}(\boldsymbol{h}) + \alpha R_{Q,u}(\boldsymbol{h}). \tag{2}$$

Because $Q_{X,Y|Y \in \mathcal{Y}_k} = P_{X,Y|Y \in \mathcal{Y}_k}$, hence,

$$\begin{aligned} R_{Q,k}(\boldsymbol{h}) = & \int_{\mathcal{X} \times \mathcal{Y}_k} \ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}) \mathrm{d}Q_{X,Y|Y \in \mathcal{Y}_k}(\mathbf{x}, \mathbf{y}) \\ = & \int_{\mathcal{X} \times \mathcal{Y}_k} \ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}) \mathrm{d}P_{X,Y|Y \in \mathcal{Y}_k}(\mathbf{x}, \mathbf{y}) \\ = & R_{P,k}(\boldsymbol{h}). \end{aligned} \tag{3}$$

Combining Eq. (2) with Eq. (3), we have that

$$R_Q^{\alpha}(\boldsymbol{h}) = (1-\alpha)R_{P,k}(\boldsymbol{h}) + \alpha R_{Q,u}(\boldsymbol{h}).$$

Step 2. We claim that $\alpha R_{Q,u}(\boldsymbol{h}) = \max\{R_Q(\boldsymbol{h}, \mathbf{y}_{C+1}) - (1-\alpha)R_{P,k}(\boldsymbol{h}, \mathbf{y}_{C+1}), 0\}$.

First, it is clear that

$$\begin{aligned} R_Q(\boldsymbol{h}, \mathbf{y}_{C+1}) = & (1-\alpha) \int_{\mathcal{X}} \ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1}) \mathrm{d}Q_{X|Y \in \mathcal{Y}_k} + \alpha \int_{\mathcal{X}} \ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1}) \mathrm{d}Q_{X|Y = \mathbf{y}_{C+1}} \\ = & (1-\alpha) \int_{\mathcal{X}} \ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1}) \mathrm{d}P_{X|Y \in \mathcal{Y}_k} + \alpha \int_{\mathcal{X}} \ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1}) \mathrm{d}Q_{X|Y = \mathbf{y}_{C+1}} \\ = & (1-\alpha) R_{P,k}(\boldsymbol{h}, \mathbf{y}_{C+1}) + \alpha \int_{\mathcal{X}} \ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1}) \mathrm{d}Q_{X|Y = \mathbf{y}_{C+1}} \\ = & (1-\alpha) R_{P,k}(\boldsymbol{h}, \mathbf{y}_{C+1}) + \alpha R_{Q,u}(\boldsymbol{h}). \end{aligned} \tag{4}$$

Hence,

$$\alpha R_{Q,u}(\boldsymbol{h}) = R_Q(\boldsymbol{h}, \mathbf{y}_{C+1}) - (1-\alpha)R_{P,k}(\boldsymbol{h}, \mathbf{y}_{C+1}).$$

Because $\alpha R_{Q,u}(\boldsymbol{h}) \geq 0$, we obtain that

$$\alpha R_{Q,u}(\boldsymbol{h}) = \max\{R_Q(\boldsymbol{h}, \mathbf{y}_{C+1}) - (1-\alpha)R_{P,k}(\boldsymbol{h}, \mathbf{y}_{C+1}), 0\}.$$

Step 3. Combining the results of Steps 1 and Steps 2, we have that

$$R_Q^{\alpha}(\boldsymbol{h}) = (1-\alpha)R_{P,k}(\boldsymbol{h}) + \max\{R_Q(\boldsymbol{h}, \mathbf{y}_{C+1}) - (1-\alpha)R_{P,k}(\boldsymbol{h}, \mathbf{y}_{C+1}), 0\}.$$

□

Lemma 2. *([Kanamori et al., 2009; 2012]). Assume the feature space $\mathcal{X}$ is compact. Let the RKHS $\mathcal{H}_K$ be the Hilbert space with Gaussian kernel. Suppose that the real density $p/q \in \mathcal{H}_K$ and set the regularization parameter $\lambda = \lambda_{n,m}$ in* KuLSIF *such that*

$$\lim_{n,m\to 0} \lambda_{n,m} = 0, \quad \lambda_{n,m}^{-1} = O(\min\{n,m\}^{1-\delta}),$$

where $0 < \delta < 1$ is any constant, then

$$\sqrt{\int_{\mathcal{X}} (\widehat{w}(\mathbf{x}) - r(\mathbf{x}))^2 \mathrm{d}U(\mathbf{x})} = O_p(\lambda_{n,m}^{\frac{1}{2}}),$$

and

$$\|\widehat{w}\|_{\mathcal{H}_K} = O_p(1),$$

where $\widehat{w}$ is the solution of KuLSIF.

Proof. The result

$$\sqrt{\int_{\mathcal{X}} (\widehat{w}(\mathbf{x}) - r(\mathbf{x}))^2 \mathrm{d}U(\mathbf{x})} = O_p(\lambda_{n,m}^{\frac{1}{2}}),$$

can be found in Theorem 1 of (Kanamori et al., 2009) and Theorem 2 of (Kanamori et al., 2012).

The result

$$\|\widehat{w}\|_{\mathcal{H}_K} = O_p(1)$$

can be found in the proving process (pages 27-28) of Theorem 1 of (Kanamori et al., 2009) and the proving process (pages 354-365) of Theorem 2 of (Kanamori et al., 2012). □

Then, we introduce the Rademacher Complexity.

Definition 2 (Rademacher Complexity)**.** *Let $\mathcal{F}$ be a class of real-valued functions defined in a space $\mathcal{Z}$. Given a distribution P over $\mathcal{Z}$ and sample $\tilde{S} = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_{\tilde{n}}\} \in \mathcal{Z}$ drawn i.i.d. from P, then the Empirical Rademacher Complexity of $\mathcal{F}$ with respect to the sample $\tilde{S}$ is*

$$\widehat{\Re}_{\tilde{S}}(\mathcal{F}) := \mathbb{E}_\sigma[\sup_{f\in\mathcal{F}} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \sigma_i f(\boldsymbol{z}_i)], \tag{5}$$

where $\sigma = (\sigma_1, ..., \sigma_{\tilde{n}})$ are Rademacher variables, with σ_is independent uniform random variables taking values in $-1, +1$.

Then the Rademacher complexity

$$\Re_{\tilde{n},P}(\mathcal{F}) := \mathbb{E}_{\tilde{S}\sim P^{\tilde{n}}} \widehat{\Re}_{\tilde{S}}(\mathcal{F}). \tag{6}$$

With the Rademacher complexity, we have

Lemma 3. ***(Theorem 26.5 in** (Shalev-Shwartz & Ben-David, 2014).) Given a space $\mathcal{Z}$, a function $l : R \times \mathcal{Z} \to \mathbb{R}_+$ and a hypothesis set $\mathcal{H} \subset \{f : \mathcal{Z} \to R\}$, let*

$$\mathcal{F} := l \circ \mathcal{H} = \{l(f(\boldsymbol{z}), \boldsymbol{z}) : f \in \mathcal{H}\},$$

where $l \leq B$. Then for a distribution P on space $\mathcal{Z}$, data $\tilde{S} = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_{\tilde{n}}\} \sim P$ i.i.d, we have with a probability of at least $1 - \delta > 0$, for all $f \in \mathcal{F}$:

$$\widehat{R}(f) - R(f) \leq 2\widehat{\Re}_{\tilde{S}}(\mathcal{F}) + 4B\sqrt{\frac{2\log(4/\delta)}{\tilde{n}}}, \tag{7}$$

where $R(f) := \int_{\mathcal{Z}} l(f(\boldsymbol{z}), \boldsymbol{z}) \mathrm{d}Q(\boldsymbol{z})$ and $\widehat{R}(f) := \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} l(f(\boldsymbol{z}_i), \boldsymbol{z}_i)$.

Using the same technique as in Lemma 3, we have with a probability of at least $1 - 2\delta > 0$, for all $f \in \mathcal{F}$:

$$|R(f) - \widehat{R}(f)| \leq 2\widehat{\Re}_{\tilde{S}}(\mathcal{F}) + 4B\sqrt{\frac{2\log(4/\delta)}{\tilde{n}}}. \tag{8}$$

Definition 3 (Shattering (Shalev-Shwartz & Ben-David, 2014))**.** *Given a feature space $\mathcal{X}$, we say that a set $U \subset \mathcal{X}$ is shattered by $\mathcal{H}$ if there exist two functions $\boldsymbol{h}_0, \boldsymbol{h}_1 : U \rightarrow \mathcal{Y}$, such that*
- *For every $\mathbf{x} \in U$, $\boldsymbol{h}_0(\mathbf{x}) \neq \boldsymbol{h}_1(\mathbf{x})$.*
- *For every $V \subset U$, there exists a function $\boldsymbol{h} \in \mathcal{H}$ such that $\forall \mathbf{x} \in V, \boldsymbol{h}(\mathbf{x}) = \boldsymbol{h}_0(\mathbf{x})$ and $\forall \mathbf{x} \in U \backslash V, \boldsymbol{h}(\mathbf{x}) = \boldsymbol{h}_1(\mathbf{x})$.*

Hence, we can define the Natarajan dimension as follows.

Definition 4 (Natarajan Dimension (Shalev-Shwartz & Ben-David, 2014))**.** *The Natarajan dimension of $\mathcal{H}$, denoted Ndim($\mathcal{H}$), is the maximal size of a shattered set $U \subset X$.*

It is not difficult to see that in the case that there are exactly two classes, Ndim($\mathcal{H}$) = VCdim($\mathcal{H}$). Therefore, the Natarajan dimension generalizes the VC dimension.

Lemma 4. *Assume that $\mathcal{H} \subset \{\boldsymbol{h} : \mathcal{X} \rightarrow \mathcal{Y}\}$ has finite Natarajan dimension and the loss function ℓ has upper bound c, then for any $0 < \delta < 1$,*

$$\sup_{\boldsymbol{h}\in\mathcal{H}} |R_{P,k}(\boldsymbol{h}) - \widehat{R}_S(\boldsymbol{h})| = cO_p(1/n^{\frac{1-\delta}{2}}), \quad \sup_{\boldsymbol{h}\in\mathcal{H}} |R_{P,k}(\boldsymbol{h}, \mathbf{y}_{C+1}) - \widehat{R}_S(\boldsymbol{h}, \mathbf{y}_{C+1})| = cO_p(1/n^{\frac{1-\delta}{2}}),$$

where

$$\widehat{R}_S(\boldsymbol{h}) := \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y})\in S} \ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}), \quad \widehat{R}_S(\boldsymbol{h}, \mathbf{y}_{C+1}) := \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y})\in S} \ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1}).$$

Proof. Assume that the Natarajan dimension is d and the upper bound of ℓ is B.

Let $\mathcal{F} = \{\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}) : \boldsymbol{h} \in \mathcal{H}\}$. Then the Natarajan lemma (Lemma 29.4 of (Shalev-Shwartz & Ben-David, 2014)) tells us that

$$|\{\boldsymbol{h}(\mathbf{x}^1), ..., \boldsymbol{h}(\mathbf{x}^n)|\boldsymbol{h} \in \mathcal{H}\}| \leq n^d (C+1)^{2d}.$$

Denote $A = \{(\ell(\boldsymbol{h}(\mathbf{x}^1), \boldsymbol{h}'(\mathbf{x}^1)), ..., \ell(\boldsymbol{h}(\mathbf{x}^n), \boldsymbol{h}'(\mathbf{x}^n))|\boldsymbol{h}, \boldsymbol{h}' \in \mathcal{H}\}$. This clearly implies that

$$|A| \leq |\{\boldsymbol{h}(\mathbf{x}^1), ..., \boldsymbol{h}(\mathbf{x}^n)|\boldsymbol{h} \in \mathcal{H}\}|^2 \leq (n)^{2d}(C+1)^{4d}.$$

Combining above inequality with Lemma 26.8 of (Shalev-Shwartz & Ben-David, 2014) and inequality (8), we obtain with a probability of at least $1 - 2\delta > 0$,

$$\sup_{\boldsymbol{h}\in\mathcal{H}} |R_{P,k}(\boldsymbol{h}) - \widehat{R}_S(\boldsymbol{h})| \leq 2\widehat{\Re}_S(\mathcal{F}) + 4c\sqrt{\frac{2\log\frac{4}{\delta}}{n}} \leq 2c\sqrt{\frac{4d\log n + 8d\log(C+1)}{n}} + 4c\sqrt{\frac{2\log\frac{4}{\delta}}{n}},$$

hence,

$$\sup_{\boldsymbol{h}\in\mathcal{H}} |R_{P,k}(\boldsymbol{h}) - \widehat{R}_S(\boldsymbol{h})| = cO_p(1/n^{\frac{1-\delta}{2}}).$$

Using the same technique, we can also prove that $\sup_{\boldsymbol{h}\in\mathcal{H}} |R_{P,k}(\boldsymbol{h}, \mathbf{y}_{C+1}) - \widehat{R}_S(\boldsymbol{h}, \mathbf{y}_{C+1})| = cO_p(1/n^{\frac{1-\delta}{2}})$. □

Lemma 5. *Assume the feature space $\mathcal{X}$ is compact and the loss function has an upper bound c. Let the RKHS $\mathcal{H}_K$ is the Hilbert space with Gaussian kernel. Suppose that the real density $p/q \in \mathcal{H}_K$ and set the regularization parameter $\lambda = \lambda_{n,m}$ in* KuLSIF *such that*

$$\lim_{n,m\rightarrow 0} \lambda_{n,m} = 0, \quad \lambda_{n,m}^{-1} = O(\min\{n,m\}^{1-\delta}),$$

where $0 < \delta < 1$ is any constant, then

$$\sup_{\boldsymbol{h}\in\mathcal{H}} |R_Q(\boldsymbol{h}, \mathbf{y}_{C+1}) - \gamma \widehat{R}_T^{\tau,\beta}(\boldsymbol{h}, \mathbf{y}_{C+1})| \leq \gamma\beta c U(\{\mathbf{x} :\ 0 < r(\mathbf{x}) \leq 2\tau\}) + c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big) O_p(\lambda_{n,m}^{\frac{1}{2}}),$$

where

$$R_Q(\boldsymbol{h}, \mathbf{y}_{C+1}) = \int_{\mathcal{X}} \ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1}) \mathrm{d}Q_X(\mathbf{x}), \quad \widehat{R}_T^{\tau,\beta}(\boldsymbol{h}, \mathbf{y}_{K+1}) := \frac{1}{m} \sum_{\mathbf{x}\in T} L_{\tau,\beta}(\widehat{w}(\mathbf{x})) \ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{K+1}),$$

here $Q_X := Q_U^{0,\beta}$, and $\widehat{w}$ is the solution of KuLSIF.

Proof. **Step 1.** We claim that

$$\sup_{\boldsymbol{h}\in\mathcal{H}} |R_Q(\boldsymbol{h},\mathbf{y}_{C+1}) - \gamma R_U^{\tau,\beta}(\boldsymbol{h},\mathbf{y}_{C+1})| \leq \gamma\beta c U(\{\mathbf{x}:~0<r(\mathbf{x})\leq 2\tau\}),$$

where

$$R_U^{\tau,\beta}(\boldsymbol{h},\mathbf{y}_{C+1}) = \int_{\mathcal{X}} L_{\tau,\beta}(r(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x}),$$

here $r(\mathbf{x}) = p(\mathbf{x})/q(\mathbf{x})$.

First, we note that

$$\begin{aligned}
&|\int_{\mathcal{X}} L_{0,\beta}(r(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x}) - \int_{\mathcal{X}} L_{\tau,\beta}(r(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})| \\
\leq&|\int_{\mathcal{X}} L_{0,\beta}(r(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x}) - \int_{\mathcal{X}} L_{\tau,\beta}(r(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})| \\
\leq& c\int_{\mathcal{X}} \big|L_{0,\beta}(r(\mathbf{x})) - L_{\tau,\beta}(r(\mathbf{x}))\big|\mathrm{d}U(\mathbf{x}) \\
\leq& c\int_{\{\mathbf{x}:~0<r(\mathbf{x})\leq 2\tau\}} \beta\mathrm{d}U(\mathbf{x}) = \beta c U(\{\mathbf{x}:~0<r(\mathbf{x})\leq 2\tau\}).
\end{aligned} \tag{9}$$

Because $Q_{X,Y} = Q_U^{0,\beta}P_{Y|X}$, then according to the definition of $Q_U^{0,\beta}$, we know

$$R_Q(\boldsymbol{h},\mathbf{y}_{C+1}) = \gamma\int_{\mathcal{X}} L_{0,\beta}(r(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x}),$$

which implies

$$\sup_{\boldsymbol{h}\in\mathcal{H}} |R_Q(\boldsymbol{h},\mathbf{y}_{C+1}) - \gamma R_U^{\tau,\beta}(\boldsymbol{h},\mathbf{y}_{C+1})| \leq \gamma\beta c U(\{\mathbf{x}:~0<r(\mathbf{x})\leq 2\tau\}).$$

Step 2. We claim that

$$\sup_{\boldsymbol{h}\in\mathcal{H}} |R_U^{\tau,\beta}(\boldsymbol{h},\mathbf{y}_{C+1}) - \int_{\mathcal{X}} L_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})| \leq \max\{c,\frac{c\beta}{\tau}\}O_p(\lambda_{n,m}^{\frac{1}{2}}).$$

First, the Lipschitz constant for $L_{\tau,\beta}$ is smaller than $\max\{1,\frac{\beta}{\tau}\}$.

Then,

$$\begin{aligned}
&\sup_{\boldsymbol{h}\in\mathcal{H}} |R_U^{\tau,\beta}(\boldsymbol{h},\mathbf{y}_{C+1}) - \int_{\mathcal{X}} L_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})| \\
=&\sup_{\boldsymbol{h}\in\mathcal{H}} |\int_{\mathcal{X}} L_{\tau,\beta}(r(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x}) - \int_{\mathcal{X}} L_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x}) \\
\leq&\sup_{\boldsymbol{h}\in\mathcal{H}} \int_{\mathcal{X}} |L_{\tau,\beta}(r(\mathbf{x})) - L_{\tau,\beta}(\widehat{w}(\mathbf{x}))|\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})| \\
\leq&\sup_{\boldsymbol{h}\in\mathcal{H}} \sqrt{\int_{\mathcal{X}} |L_{\tau,\beta}(r(\mathbf{x})) - L_{\tau,\beta}(\widehat{w}(\mathbf{x}))|^2\mathrm{d}U(\mathbf{x})}\sqrt{\int_{\mathcal{X}} \ell^2(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})}~~\text{Hölder~~Inequality} \\
\leq& c\sup_{\boldsymbol{h}\in\mathcal{H}} \sqrt{\int_{\mathcal{X}} |L_{\tau,\beta}(r(\mathbf{x})) - L_{\tau,\beta}(\widehat{w}(\mathbf{x}))|^2\mathrm{d}U(\mathbf{x})} \\
\leq&\max\{c,\frac{c\beta}{\tau}\}\sup_{\boldsymbol{h}\in\mathcal{H}} \sqrt{\int_{\mathcal{X}} |r(\mathbf{x}) - \widehat{w}(\mathbf{x})|^2\mathrm{d}U(\mathbf{x})}.
\end{aligned}$$

Lastly, using Lemma 2,

$$\sup_{\boldsymbol{h}\in\mathcal{H}} |R_U^{\tau,\beta}(\boldsymbol{h},\mathbf{y}_{C+1}) - \int_{\mathcal{X}} L_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})| \leq \max\{c,\frac{c\beta}{\tau}\}O_p(\lambda_{n,m}^{\frac{1}{2}}).$$

Step 3. We claim that

$$\sup_{\boldsymbol{h}\in\mathcal{H}}\Big|\frac{1}{m}\sum_{\mathbf{x}\in T}L_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})-\int_{\mathcal{X}}L_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})\Big|\leq c\big(\max\{1,\frac{\beta}{\tau}\}+\beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}}).$$

First, we set $\mathcal{F}_B:=\{L_{\tau,\beta}(w)\ell(\boldsymbol{h},\mathbf{y}_{C+1}):w\in\mathcal{H}_K,\|w\|_K\leq B,\boldsymbol{h}\in\mathcal{H}\}$. We consider

$$\sup_{f\in\mathcal{F}_B}\Big|\frac{1}{m}\sum_{\mathbf{x}\in T}f(\mathbf{x})-\int_{\mathcal{X}}f(\mathbf{x})\mathrm{d}U(\mathbf{x})\Big|$$

Using Lemma 3 and inequality 8, it is easy to check that for $1-2\delta>0$, we have

$$\sup_{f\in\mathcal{F}_B}\Big|\frac{1}{m}\sum_{\mathbf{x}\in T}f(\mathbf{x})-\int_{\mathcal{X}}f(\mathbf{x})\mathrm{d}U(\mathbf{x})\Big|\leq 2\widehat{\Re}_T(\mathcal{F}_B)+4(B+\beta)c\sqrt{\frac{2\log(4/\delta)}{m}},\tag{10}$$

here we have used $|f|\leq(B+\beta)c$, for any $f\in\mathcal{F}_B$.

Then, we consider $\widehat{\Re}_T(\mathcal{F}_B)$.

$$\begin{aligned}
&m\widehat{\Re}_T(\mathcal{F}_B)\\
=&\mathbb{E}_\sigma[\sup_{\|w\|_K\leq B,\boldsymbol{h}\in\mathcal{H}}\sum_{i=1}^m\sigma_iL_{\tau,\beta}(w_i)\ell(\boldsymbol{h}_i,\mathbf{y}_{C+1})]\\
=&\mathbb{E}_\sigma[\sup_{\|w\|_K\leq B,\boldsymbol{h}\in\mathcal{H}}\sigma_1L_{\tau,\beta}(w_1)\ell(\boldsymbol{h}_1,\mathbf{y}_{C+1})+\sum_{i=2}^m\sigma_iL_{\tau,\beta}(w_i)\ell(\boldsymbol{h}_i,\mathbf{y}_{C+1})]\\
=&\frac{1}{2}\mathbb{E}_{\sigma_2,...,\sigma_m}[\sup_{\|w\|_K\leq B,\boldsymbol{h}\in\mathcal{H}}L_{\tau,\beta}(w_1)\ell(\boldsymbol{h}_1,\mathbf{y}_{C+1})+\sum_{i=2}^m\sigma_iL_{\tau,\beta}(w_i)\ell(\boldsymbol{h}_i,\mathbf{y}_{C+1})\\
&\qquad+\sup_{\|w\|_K\leq B,\boldsymbol{h}\in\mathcal{H}}-L_{\tau,\beta}(w_1)\ell(\boldsymbol{h}_1,\mathbf{y}_{C+1})+\sum_{i=2}^m\sigma_iL_{\tau,\beta}(w_i)\ell(\boldsymbol{h}_i,\mathbf{y}_{C+1})]\\
=&\frac{1}{2}\mathbb{E}_{\sigma_2,...,\sigma_m}[\sup_{\|w\|_K\leq B,\boldsymbol{h}\in\mathcal{H}}\sup_{\|w'\|_K\leq B,\boldsymbol{h}'\in\mathcal{H}}L_{\tau,\beta}(w_1)\ell(\boldsymbol{h}_1,\mathbf{y}_{C+1})+\sum_{i=2}^m\sigma_iL_{\tau,\beta}(w_i)\ell(\boldsymbol{h}_i,\mathbf{y}_{C+1})\\
&\qquad-L_{\tau,\beta}(w_1')\ell(\boldsymbol{h}_1',\mathbf{y}_{C+1})+\sum_{i=2}^m\sigma_iL_{\tau,\beta}(w_i')\ell(\boldsymbol{h}_i',\mathbf{y}_{C+1})]\\
\leq&\frac{1}{2}\mathbb{E}_{\sigma_2,...,\sigma_m}[\sup_{\|w\|_K\leq B,\boldsymbol{h}\in\mathcal{H}}\sup_{\|w'\|_K\leq B,\boldsymbol{h}'\in\mathcal{H}}|L_{\tau,\beta}(w_1)-L_{\tau,\beta}(w_1')|\ell(\boldsymbol{h}_1',\mathbf{y}_{C+1})+L_{\tau,\beta}(w_1)|\ell(\boldsymbol{h}_1',\mathbf{y}_{C+1})-\ell(\boldsymbol{h}_1,\mathbf{y}_{C+1})|\\
&\qquad+\sum_{i=2}^m\sigma_iL_{\tau,\beta}(w_i)\ell(\boldsymbol{h}_i,\mathbf{y}_{C+1})+\sum_{i=2}^m\sigma_iL_{\tau,\beta}(w_i')\ell(\boldsymbol{h}_i',\mathbf{y}_{C+1})]\\
\leq&\frac{1}{2}\mathbb{E}_{\sigma_2,...,\sigma_m}[\sup_{\|w\|_K\leq B,\boldsymbol{h}\in\mathcal{H}}\sup_{\|w'\|_K\leq B,\boldsymbol{h}'\in\mathcal{H}}Lc|w_1-w_1'|+(B+\beta)|\ell(\boldsymbol{h}_1',\mathbf{y}_{C+1})-\ell(\boldsymbol{h}_1,\mathbf{y}_{C+1})|\\
&\qquad+\sum_{i=2}^m\sigma_iL_{\tau,\beta}(w_i)\ell(\boldsymbol{h}_i,\mathbf{y}_{C+1})+\sum_{i=2}^m\sigma_iL_{\tau,\beta}(w_i')\ell(\boldsymbol{h}_i',\mathbf{y}_{C+1})]\\
=&\mathbb{E}_\sigma[\sup_{\|w\|_K\leq B,\boldsymbol{h}\in\mathcal{H}}Lc\sigma_1w_1+(B+\beta)\sigma_1\ell(\boldsymbol{h}_1,\mathbf{y}_{C+1})+\sum_{i=2}^m\sigma_iL_{\tau,\beta}(w_i)\ell(\boldsymbol{h}_i,\mathbf{y}_{C+1})]
\end{aligned}$$

Repeat the process $m-1$ times for $i=2,...,m$.

$$\begin{aligned}
\leq&\mathbb{E}_\sigma[\sup_{\|w\|_K\leq B,\boldsymbol{h}\in\mathcal{H}}\sum_{i=1}^mLc\sigma_iw_i+\sum_{i=1}^m(B+\beta)\sigma_i\ell(\boldsymbol{h}_i,\mathbf{y}_{C+1})]\\
\leq&mLc\widehat{\Re}_T(\mathcal{H}_{K,B})+m(B+\beta)\widehat{\Re}_T(\mathcal{F}),
\end{aligned}$$

where $w_i = w(\tilde{\mathbf{x}}_i), \boldsymbol{h}_i = \boldsymbol{h}(\tilde{\mathbf{x}}_i)$, $L = \max\{1, \frac{\beta}{\tau}\}$, $\mathcal{H}_{K,B} = \{w : w \in \mathcal{H}_K, \|w\|_K \leq B\}$ and $\mathcal{F} = \{\ell(\boldsymbol{h}, \mathbf{y}_{C+1}) : \boldsymbol{h} \in \mathcal{H}\}$.

According to Theorem 5.5 of (Mohri et al., 2012), we obtain that

$$\widehat{\Re}_T(\mathcal{H}_{K,B}) \leq B\sqrt{\frac{1}{m}}.$$

According to the proving process of Lemma 4, we obtain that

$$\widehat{\Re}_T(\mathcal{F}) \leq c\sqrt{\frac{4d\log m + 8d\log(C+1)}{m}},$$

where d is the Natarajan Dimension of $\mathcal{H}$.

Hence,

$$\widehat{\Re}_T(\mathcal{F}_B) \leq BLc\sqrt{\frac{1}{m}} + (B+\beta)c\sqrt{\frac{4d\log m + 8d\log(C+1)}{m}}.$$

This implies that for $1 - 2\delta > 0$, we have

$$\begin{aligned}
&\sup_{w\in\mathcal{H}_K, \|w\|_K\leq B}\sup_{\boldsymbol{h}\in\mathcal{H}}\Big|\frac{1}{m}\sum_{\mathbf{x}\in T} L_{\tau,\beta}(w(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1}) - \int_{\mathcal{X}} L_{\tau,\beta}(w(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})\Big| \\
\leq& 2B\max\{1, \frac{\beta}{\tau}\}c\sqrt{\frac{1}{m}} + 2(B+\beta)c\sqrt{\frac{4d\log m + 8d\log(C+1)}{m}} + 4(B+\beta)c\sqrt{\frac{2\log(4/\delta)}{m}}.
\end{aligned} \tag{11}$$

Because $\|\widehat{w}\|_K = O_p(1)$, then combining inequality 11, we know that

$$\begin{aligned}
&\sup_{\boldsymbol{h}\in\mathcal{H}}\Big|\frac{1}{m}\sum_{\mathbf{x}\in T} L_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1}) - \int_{\mathcal{X}} L_{\tau,\beta}(w(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})\Big| \\
\leq& 2O_p(1)\max\{1, \frac{\beta}{\tau}\}cO_p(\sqrt{\frac{1}{m}}) + 2(O_p(1)+\beta)cO_p(\sqrt{\frac{4d\log m + 8d\log(C+1)}{m}}) + 4(O_p(1)+\beta)cO_p(\sqrt{\frac{2\log(4/\delta)}{m}}).
\end{aligned}$$

This implies that

$$\sup_{\boldsymbol{h}\in\mathcal{H}}\Big|\frac{1}{m}\sum_{\mathbf{x}\in T} L_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1}) - \int_{\mathcal{X}} L_{\tau,\beta}(w(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})\Big| = c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}}).$$

Step 4. Using the results of Steps 1, 2 and 3, we have

$$\begin{aligned}
&\sup_{\boldsymbol{h}\in\mathcal{H}}|R_Q(\boldsymbol{h}, \mathbf{y}_{C+1}) - \gamma\widehat{R}_T^{\tau,\beta}(\boldsymbol{h}, \mathbf{y}_{C+1})| \\
\leq& \sup_{\boldsymbol{h}\in\mathcal{H}}|R_Q(\boldsymbol{h}, \mathbf{y}_{C+1}) - \gamma R_U^{\tau,\beta}(\boldsymbol{h}, \mathbf{y}_{C+1})| + \sup_{\boldsymbol{h}\in\mathcal{H}}|\gamma R_U^{\tau,\beta}(\boldsymbol{h}, \mathbf{y}_{C+1}) - \gamma\int_{\mathcal{X}} L_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})| \\
&+ \sup_{\boldsymbol{h}\in\mathcal{H}}|\gamma\int_{\mathcal{X}} L_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x}) - \gamma\widehat{R}_T^{\tau,\beta}(\boldsymbol{h}, \mathbf{y}_{C+1})| \\
\leq& \gamma\beta cU(\{\mathbf{x}:\ 0 < r(\mathbf{x}) \leq 2\tau\}) + \gamma\max\{c, \frac{c\beta}{\tau}\}O_p(\lambda_{n,m}^{\frac{1}{2}}) + c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}}).
\end{aligned}$$

Note that $\gamma < 1$, we can write

$$\sup_{\boldsymbol{h}\in\mathcal{H}}|R_Q(\boldsymbol{h}, \mathbf{y}_{C+1}) - \gamma\widehat{R}_T^{\tau,\beta}(\boldsymbol{h}, \mathbf{y}_{C+1})| \leq \gamma\beta cU(\{\mathbf{x}:\ 0 < r(\mathbf{x}) \leq 2\tau\}) + c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}}).$$

□

Proof of Theorem 3. We separate the proof into three steps.

Step 1. We claim that

$$\begin{aligned}&\sup_{\boldsymbol{h}\in\mathcal{H}}|(1-\alpha)\Delta_{S,T}^{\tau,\beta}(\boldsymbol{h})-\max\{R_Q(\boldsymbol{h},\mathbf{y}_{C+1})-(1-\alpha)R_{P,k}(\boldsymbol{h},\mathbf{y}_{C+1}),0\}|\\ \leq&\gamma\beta cU(\{\mathbf{x}:\ 0<r(\mathbf{x})\leq 2\tau\})+c\big(\max\{1,\frac{\beta}{\tau}\}+\beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}}).\end{aligned}$$

First, it is easy to check that

$$\begin{aligned}&\sup_{\boldsymbol{h}\in\mathcal{H}}|(1-\alpha)\Delta_{S,T}^{\tau,\beta}(\boldsymbol{h})-\max\{R_Q(\boldsymbol{h},\mathbf{y}_{C+1})-(1-\alpha)R_{P,k}(\boldsymbol{h},\mathbf{y}_{C+1}),0\}|\\ \leq&\sup_{\boldsymbol{h}\in\mathcal{H}}|(1-\alpha)\widehat{R}_T^{\tau,\beta}(\boldsymbol{h},\mathbf{y}_{K+1})-(1-\alpha)\widehat{R}_S(\boldsymbol{h},\mathbf{y}_{K+1})-R_Q(\boldsymbol{h},\mathbf{y}_{C+1})+(1-\alpha)R_{P,k}(\boldsymbol{h},\mathbf{y}_{C+1})|\\ \leq&\sup_{\boldsymbol{h}\in\mathcal{H}}|(1-\alpha)\widehat{R}_T^{\tau,\beta}(\boldsymbol{h},\mathbf{y}_{K+1})-R_Q(\boldsymbol{h},\mathbf{y}_{C+1})|+(1-\alpha)\sup_{\boldsymbol{h}\in\mathcal{H}}|\widehat{R}_S(\boldsymbol{h},\mathbf{y}_{K+1})-R_{P,k}(\boldsymbol{h},\mathbf{y}_{C+1})|\\ \leq&\gamma\beta cU(\{\mathbf{x}:\ 0<r(\mathbf{x})\leq 2\tau\})+c\big(\max\{1,\frac{\beta}{\tau}\}+\beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}})\quad\text{Use Lemma 5}\\ &+(1-\alpha)\sup_{\boldsymbol{h}\in\mathcal{H}}|\widehat{R}_S(\boldsymbol{h},\mathbf{y}_{K+1})-R_{P,k}(\boldsymbol{h},\mathbf{y}_{C+1})|\\ \leq&\gamma\beta cU(\{\mathbf{x}:\ 0<r(\mathbf{x})\leq 2\tau\})+c\big(\max\{1,\frac{\beta}{\tau}\}+\beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}})\\ &+(1-\alpha)cO_p(\lambda_{n,m}^{\frac{1}{2}})\quad\text{Use Lemma 4.}\end{aligned}$$

Hence, we can write

$$\begin{aligned}&\sup_{\boldsymbol{h}\in\mathcal{H}}|(1-\alpha)\Delta_{S,T}^{\tau,\beta}(\boldsymbol{h})-\max\{R_Q(\boldsymbol{h},\mathbf{y}_{C+1})-(1-\alpha)R_{P,k}(\boldsymbol{h},\mathbf{y}_{C+1}),0\}|\\ \leq&\gamma\beta cU(\{\mathbf{x}:\ 0<r(\mathbf{x})\leq 2\tau\})+c\big(\max\{1,\frac{\beta}{\tau}\}+\beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}}).\end{aligned}$$

Step 2.

$$\begin{aligned}&\sup_{\boldsymbol{h}\in\mathcal{H}}|(1-\alpha)\widehat{R}_S(\boldsymbol{h})+(1-\alpha)\Delta_{S,T}^{\tau,\beta}(\boldsymbol{h})-(1-\alpha)R_{P,k}(\boldsymbol{h})-\max\{R_Q(\boldsymbol{h},\mathbf{y}_{C+1})-(1-\alpha)R_{P,k}(\boldsymbol{h},\mathbf{y}_{C+1}),0\}|\\ \leq&\sup_{\boldsymbol{h}\in\mathcal{H}}|(1-\alpha)\widehat{R}_S(\boldsymbol{h})-(1-\alpha)R_{P,k}(\boldsymbol{h})|+\sup_{\boldsymbol{h}\in\mathcal{H}}|(1-\alpha)\Delta_{S,T}^{\tau,\beta}(\boldsymbol{h})-\max\{R_Q(\boldsymbol{h},\mathbf{y}_{C+1})-(1-\alpha)R_{P,k}(\boldsymbol{h},\mathbf{y}_{C+1}),0\}|\\ \leq&(1-\alpha)cO_p(\lambda_{n,m}^{\frac{1}{2}})\quad\text{Use Lemma 4}\\ &+\gamma\beta cU(\{\mathbf{x}:\ 0<r(\mathbf{x})\leq 2\tau\})+c\big(\max\{1,\frac{\beta}{\tau}\}+\beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}})\quad\text{Use the result of Step 1.}\end{aligned}$$

Hence, we can write

$$\begin{aligned}&\sup_{\boldsymbol{h}\in\mathcal{H}}|(1-\alpha)\widehat{R}_S(\boldsymbol{h})+(1-\alpha)\Delta_{S,T}^{\tau,\beta}(\boldsymbol{h})-(1-\alpha)R_{P,k}(\boldsymbol{h})-\max\{R_Q(\boldsymbol{h},\mathbf{y}_{C+1})-(1-\alpha)R_{P,k}(\boldsymbol{h},\mathbf{y}_{C+1}),0\}|\\ \leq&\gamma\beta cU(\{\mathbf{x}:\ 0<r(\mathbf{x})\leq 2\tau\})+c\big(\max\{1,\frac{\beta}{\tau}\}+\beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}}).\end{aligned}$$

Step 3. Note that

$$R_Q^{\alpha}(\boldsymbol{h})=(1-\alpha)R_{P,k}(\boldsymbol{h})+\max\{R_Q(\boldsymbol{h},\mathbf{y}_{C+1})-(1-\alpha)R_{P,k}(\boldsymbol{h},\mathbf{y}_{C+1}),0\}\quad\text{Use Lemma 1.}$$

Hence,

$$\begin{aligned}&\sup_{\boldsymbol{h}\in\mathcal{H}}|(1-\alpha)\widehat{R}_S(\boldsymbol{h})+(1-\alpha)\Delta_{S,T}^{\tau,\beta}(\boldsymbol{h})-R_Q^{\alpha}(\boldsymbol{h})|\\ \leq&\gamma\beta cU(\{\mathbf{x}:\ 0<r(\mathbf{x})\leq 2\tau\})+c\big(\max\{1,\frac{\beta}{\tau}\}+\beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}}).\end{aligned}$$

□

5. Appendix E: Proofs of Theorem 4, Theorem 5 and Theorem 6

5.1. Proof for Theorem

Proof of Theorem 4. According to Theorem 1, we know that for any $\boldsymbol{h} \in \mathcal{H}$,

$$|R_P^\alpha(\boldsymbol{h}) - R_Q^\alpha(\boldsymbol{h})| \leq \alpha d^\ell_{\boldsymbol{h},\mathcal{H}}(P_{X|Y=\mathbf{y}_{C+1}}, Q_{X|Y=\mathbf{y}_{C+1}}) + \alpha\Lambda. \tag{12}$$

According to Theorem 3, we know that for any $\boldsymbol{h} \in \mathcal{H}$,

$$\begin{aligned} &|(1-\alpha)\widehat{R}_S(\boldsymbol{h}) + (1-\alpha)\Delta^{\tau,\beta}_{S,T}(\boldsymbol{h}) - R_Q^\alpha(\boldsymbol{h})| \\ \leq &\gamma\beta c U(\{\mathbf{x}: \ 0 < r(\mathbf{x}) \leq 2\tau\}) + c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big) O_p(\lambda_{n,m}^{\frac{1}{2}}). \end{aligned} \tag{13}$$

Combining inequalities (12) and (13), we know that for any $\boldsymbol{h} \in \mathcal{H}$,

$$\begin{aligned} &|(1-\alpha)\widehat{R}_S(\boldsymbol{h}) + (1-\alpha)\Delta^{\tau,\beta}_{S,T}(\boldsymbol{h}) - R_P^\alpha(\boldsymbol{h})| \\ \leq &\gamma\beta c U(\{\mathbf{x}: \ 0 < r(\mathbf{x}) \leq 2\tau\}) + c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big) O_p(\lambda_{n,m}^{\frac{1}{2}}) + \alpha d^\ell_{\boldsymbol{h},\mathcal{H}}(P_{X|Y=\mathbf{y}_{C+1}}, Q_{X|Y=\mathbf{y}_{C+1}}) + \alpha\Lambda. \end{aligned}$$

□

5.2. Proof for Theorem 5

Proof of Theorem 5. Assume that

$$\widehat{\boldsymbol{h}} \in \arg\min_{\boldsymbol{h}\in\mathcal{H}} \widehat{R}^{\tau,\beta}_{S,T}(\boldsymbol{h}), \quad \boldsymbol{h}_Q \in \arg\min_{\boldsymbol{h}\in\mathcal{H}} R_Q^\alpha(\boldsymbol{h}).$$

Step 1. It is easy to check that

$$\begin{aligned} R_Q^\alpha(\widehat{\boldsymbol{h}}) - R_Q^\alpha(\boldsymbol{h}_Q) = &R_Q^\alpha(\widehat{\boldsymbol{h}}) - (1-\alpha)\widehat{R}^{\tau,\beta}_{S,T}(\widehat{\boldsymbol{h}}) + (1-\alpha)\widehat{R}^{\tau,\beta}_{S,T}(\widehat{\boldsymbol{h}}) - R_Q^\alpha(\boldsymbol{h}_Q) \\ \leq &R_Q^\alpha(\widehat{\boldsymbol{h}}) - (1-\alpha)\widehat{R}^{\tau,\beta}_{S,T}(\widehat{\boldsymbol{h}}) + (1-\alpha)\widehat{R}^{\tau,\beta}_{S,T}(\boldsymbol{h}_Q) - R_Q^\alpha(\boldsymbol{h}_Q) \\ \leq &2\sup_{\boldsymbol{h}\in\mathcal{H}} |(1-\alpha)\widehat{R}^{\tau,\beta}_{S,T}(\boldsymbol{h}) - R_Q^\alpha(\boldsymbol{h})|, \end{aligned}$$

and

$$\begin{aligned} R_Q^\alpha(\widehat{\boldsymbol{h}}) - R_Q^\alpha(\boldsymbol{h}_Q) = &R_Q^\alpha(\widehat{\boldsymbol{h}}) - (1-\alpha)\widehat{R}^{\tau,\beta}_{S,T}(\boldsymbol{h}_Q) + (1-\alpha)\widehat{R}^{\tau,\beta}_{S,T}(\boldsymbol{h}_Q) - R_Q^\alpha(\boldsymbol{h}_Q) \\ \geq &R_Q^\alpha(\widehat{\boldsymbol{h}}) - (1-\alpha)\widehat{R}^{\tau,\beta}_{S,T}(\widehat{\boldsymbol{h}}) + (1-\alpha)\widehat{R}^{\tau,\beta}_{S,T}(\boldsymbol{h}_Q) - R_Q^\alpha(\boldsymbol{h}_Q) \\ \geq &-2\sup_{\boldsymbol{h}\in\mathcal{H}} |(1-\alpha)\widehat{R}^{\tau,\beta}_{S,T}(\boldsymbol{h}) - R_Q^\alpha(\boldsymbol{h})|, \end{aligned}$$

which implies that

$$|R_Q^\alpha(\widehat{\boldsymbol{h}}) - R_Q^\alpha(\boldsymbol{h}_Q)| \leq 2\sup_{\boldsymbol{h}\in\mathcal{H}} |(1-\alpha)\widehat{R}^{\tau,\beta}_{S,T}(\boldsymbol{h}) - R_Q^\alpha(\boldsymbol{h})|.$$

Using the result of Theorem 3, we obtain that

$$|R_Q^\alpha(\widehat{\boldsymbol{h}}) - R_Q^\alpha(\boldsymbol{h}_Q)| \leq 2c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big) O_p(\lambda_{n,m}^{\frac{1}{2}}) + 2\gamma c\beta U(0 < p/q \leq 2\tau). \tag{14}$$

Then, using the result of Step 3 in the proof of Theorem 2, we obtain that

$$|R_Q^\alpha(\widehat{\boldsymbol{h}}) - R_P^\alpha(\boldsymbol{h}_Q)| \leq 2c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big) O_p(\lambda_{n,m}^{\frac{1}{2}}) + 2\gamma c\beta U(0 < p/q \leq 2\tau). \tag{15}$$

Step 2.

$$\begin{aligned}
\widehat{R}^{\tau,\beta}_{S,T}(\widehat{\boldsymbol{h}}) &= (1-\alpha)\widehat{R}_S(\widehat{\boldsymbol{h}}) + (1-\alpha)\Delta^{\tau,\beta}_{S,T}(\widehat{\boldsymbol{h}}) \\
&\le (1-\alpha)\widehat{R}_S(\boldsymbol{h}_Q) + (1-\alpha)\Delta^{\tau,\beta}_{S,T}(\boldsymbol{h}_Q) \\
&\le R^{\alpha}_Q(\boldsymbol{h}_Q) + c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + \gamma c\beta U(0 < p/q \le 2\tau) \text{ Using Theorem 3} \\
&= (1-\alpha)\min_{\boldsymbol{h}\in\mathcal{H}} R_{Q,k}(\boldsymbol{h}) + c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + \gamma c\beta U(0 < p/q \le 2\tau) \\
&\text{Using the result of Step 3 in proof of Theorem 2} : \min_{\boldsymbol{h}\in\mathcal{H}} R^{\alpha}_Q(\boldsymbol{h}) = (1-\alpha)\min_{\boldsymbol{h}\in\mathcal{H}} R_{Q,k}(\boldsymbol{h}) \\
&\le (1-\alpha)R_{Q,k}(\widehat{\boldsymbol{h}}) + c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + \gamma c\beta U(0 < p/q \le 2\tau).
\end{aligned}$$

Hence,

$$\begin{aligned}
&(1-\alpha)\Delta^{\tau,\beta}_{S,T}(\widehat{\boldsymbol{h}}) \\
\le&(1-\alpha)R_{Q,k}(\widehat{\boldsymbol{h}}) - (1-\alpha)\widehat{R}_S(\widehat{\boldsymbol{h}}) + c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + \gamma c\beta U(0 < p/q \le 2\tau) \\
=&(1-\alpha)R_{P,k}(\widehat{\boldsymbol{h}}) - (1-\alpha)\widehat{R}_S(\widehat{\boldsymbol{h}}) + c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + \gamma c\beta U(0 < p/q \le 2\tau) \\
\le&(1-\alpha)cO_p(\lambda^{\frac{1}{2}}_{n,m}) + c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + \gamma c\beta U(0 < p/q \le 2\tau) \text{ Using the result of Lemma 4} \\
\le&2c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + \gamma c\beta U(0 < p/q \le 2\tau).
\end{aligned}$$

Then, combining above inequality with the result of Step 1 in the proof of Theorem 3, we obtain that

$$\begin{aligned}
&\max\{R_Q(\widehat{\boldsymbol{h}}, \mathbf{y}_{C+1}) - (1-\alpha)R_{P,k}(\widehat{\boldsymbol{h}}, \mathbf{y}_{C+1}), 0\} \\
\le&3c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + 2\gamma c\beta U(0 < p/q \le 2\tau).
\end{aligned}$$

Because $\max\{R_Q(\widehat{\boldsymbol{h}}, \mathbf{y}_{C+1}) - (1-\alpha)R_{P,k}(\widehat{\boldsymbol{h}}, \mathbf{y}_{C+1}), 0\} = \alpha R_{Q,u}(\widehat{\boldsymbol{h}})$, we obtain that

$$\alpha R_{Q,u}(\widehat{\boldsymbol{h}}) \le 3c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + 2\gamma c\beta U(0 < p/q \le 2\tau).$$

Step 3.

$$\begin{aligned}
&|R^{\alpha}_P(\widehat{\boldsymbol{h}}) - R^{\alpha}_P(\boldsymbol{h}_Q)| \\
\le&|R^{\alpha}_P(\widehat{\boldsymbol{h}}) - R^{\alpha}_Q(\widehat{\boldsymbol{h}})| + |R^{\alpha}_Q(\widehat{\boldsymbol{h}}) - R^{\alpha}_P(\boldsymbol{h}_Q)| \\
=&\alpha|R_{P,u}(\widehat{\boldsymbol{h}}) - R^{\alpha}_{Q,u}(\widehat{\boldsymbol{h}})| + |R^{\alpha}_Q(\widehat{\boldsymbol{h}}) - R^{\alpha}_P(\boldsymbol{h}_Q)| \\
\le&\alpha R_{Q,u}(\widehat{\boldsymbol{h}}) + |R^{\alpha}_Q(\widehat{\boldsymbol{h}}) - R^{\alpha}_P(\boldsymbol{h}_Q)| \\
\le&5c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + 4\gamma c\beta U(0 < p/q \le 2\tau) \text{ Using the results of Step 1 and Step 2.}
\end{aligned}$$

Briefly, we can write (absorbing coefficient 5 into O_p)

$$|R^{\alpha}_P(\widehat{\boldsymbol{h}}) - R^{\alpha}_P(\boldsymbol{h}_Q)| \le c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + 4\gamma c\beta U(0 < p/q \le 2\tau).$$

Combining above inequality with Theorem 2, we obtain that

$$|R^{\alpha}_P(\widehat{\boldsymbol{h}}) - \min_{\boldsymbol{h}\in\mathcal{H}} R^{\alpha}_P(\boldsymbol{h})| \le c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + 4\gamma c\beta U(0 < p/q \le 2\tau).$$

□

5.3. Proof for Theorem 6

Lemma 6. *Assume the feature space $\mathcal{X}$ is compact and the loss function has an upper bound c. Let the RKHS $\mathcal{H}_K$ is the Hilbert space with Gaussian kernel. Suppose that the real density $p/q \in \mathcal{H}_K$ and set the regularization parameter $\lambda = \lambda_{n,m}$ in* KuLSIF *such that*

$$\lim_{n,m\to 0} \lambda_{n,m} = 0, \quad \lambda_{n,m}^{-1} = O(\min\{n,m\}^{1-\delta}),$$

where $0 < \delta < 1$ is any constant, then

$$\sup_{\boldsymbol{h}\in\mathcal{H}} |R_{Q,u}(\boldsymbol{h}) - \gamma' \widehat{R}^{\tau,\beta}_{S,T,u}(\boldsymbol{h})| \le \gamma'\beta c U(\{\mathbf{x}:\ 0 < r(\mathbf{x}) \le 2\tau\}) + c\big(\max\{1, \frac{\beta}{\tau}\} + \beta\big) O_p(\lambda_{n,m}^{\frac{1}{2}}),$$

where $\gamma' = 1/\big(\beta U(\{\mathbf{x}:\ r(\mathbf{x}) = 0\})\big)$, and

$$R_{Q,u}(\boldsymbol{h}) = \int_{\mathcal{X}} \ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1}) \mathrm{d}Q_{X|Y=\mathbf{y}_{C+1}}(\mathbf{x}), \quad \widehat{R}^{\tau,\beta}_{S,T,u}(\boldsymbol{h}) := \frac{1}{m}\sum_{\mathbf{x}\in T} L_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1}),$$

here $Q_{X,Y} := Q_U^{0,\beta} P_{Y|X}$, $\widehat{w}$ is the solution of KuLSIF*, and*

$$L^{-}_{\tau,\beta}(x) = \begin{cases} x+\beta, & x \le \tau; \\ 0, & 2\tau \le x; \\ -\dfrac{\tau+\beta}{\tau}x + 2\tau + 2\beta, & \tau < x < 2\tau. \end{cases} \tag{16}$$

Proof. **Step 1.** We claim that

$$\sup_{\boldsymbol{h}\in\mathcal{H}} |R_{Q,u}(\boldsymbol{h}) - \gamma' R^{\tau,\beta}_{U,u}(\boldsymbol{h})| \le \gamma'\beta c U(\{\mathbf{x}:\ 0 < r(\mathbf{x}) \le 2\tau\}),$$

where

$$R^{\tau,\beta}_{U,u}(\boldsymbol{h}) = \int_{\mathcal{X}} L^{-}_{\tau,\beta}(r(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1}) \mathrm{d}U(\mathbf{x}),$$

here $r(\mathbf{x}) = p(\mathbf{x})/q(\mathbf{x})$.

First, we note that

$$\begin{aligned} &\Big| \int_{\mathcal{X}} L^{-}_{0,\beta}(r(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x}) - \int_{\mathcal{X}} L^{-}_{\tau,\beta}(r(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})\Big| \\ \le& \Big| \int_{\mathcal{X}} L^{-}_{0,\beta}(r(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x}) - \int_{\mathcal{X}} L^{-}_{\tau,\beta}(r(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})\Big| \\ \le& c\int_{\mathcal{X}} \big|L^{-}_{0,\beta}(r(\mathbf{x})) - L^{-}_{\tau,\beta}(r(\mathbf{x}))\big| \mathrm{d}U(\mathbf{x}) \\ \le& c\int_{\{\mathbf{x}:\ 0<r(\mathbf{x})\le 2\tau\}} (\tau+\beta)\mathrm{d}U(\mathbf{x}) = (\tau+\beta)cU(\{\mathbf{x}:\ 0 < r(\mathbf{x}) \le 2\tau\}). \end{aligned} \tag{17}$$

Because $Q_{X,Y} = Q_U^{0,\beta} P_{Y|X}$, then according to the definition of $Q_U^{0,\beta}$, we know

$$R_{Q,u}(\boldsymbol{h}) = \gamma' \int_{\mathcal{X}} L^{-}_{0,\beta}(r(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x}),$$

which implies

$$\sup_{\boldsymbol{h}\in\mathcal{H}} |R_{Q,u}(\boldsymbol{h}) - \gamma' R^{\tau,\beta}_{U,u}(\boldsymbol{h})| \le \gamma'(\tau+\beta) c U(\{\mathbf{x}:\ 0 < r(\mathbf{x}) \le 2\tau\}).$$

Step 2. We claim that

$$\sup_{\boldsymbol{h}\in\mathcal{H}} |R^{\tau,\beta}_{U,u}(\boldsymbol{h}) - \int_{\mathcal{X}} L^{-}_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}), \mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})| \le (c + \frac{c\beta}{\tau}) O_p(\lambda_{n,m}^{\frac{1}{2}}).$$

First, the Lipschitz constant for $L^{-}_{\tau,\beta}$ is smaller than $1+\frac{\beta}{\tau}$.

Then,

$$\begin{aligned}
&\sup_{\boldsymbol{h}\in\mathcal{H}} |R^{\tau,\beta}_{U,u}(\boldsymbol{h}) - \int_{\mathcal{X}} L^{-}_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})| \\
=&\sup_{\boldsymbol{h}\in\mathcal{H}} |\int_{\mathcal{X}} L^{-}_{\tau,\beta}(r(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x}) - \int_{\mathcal{X}} L^{-}_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})| \\
\leq&\sup_{\boldsymbol{h}\in\mathcal{H}} \int_{\mathcal{X}} |L^{-}_{\tau,\beta}(r(\mathbf{x})) - L^{-}_{\tau,\beta}(\widehat{w}(\mathbf{x}))|\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x}) \\
\leq&\sup_{\boldsymbol{h}\in\mathcal{H}} \sqrt{\int_{\mathcal{X}} |L^{-}_{\tau,\beta}(r(\mathbf{x})) - L^{-}_{\tau,\beta}(\widehat{w}(\mathbf{x}))|^2\mathrm{d}U(\mathbf{x})}\sqrt{\int_{\mathcal{X}} \ell^2(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})}\ \text{Hölder Inequality} \\
\leq& c\sup_{\boldsymbol{h}\in\mathcal{H}} \sqrt{\int_{\mathcal{X}} |L^{-}_{\tau,\beta}(r(\mathbf{x})) - L^{-}_{\tau,\beta}(\widehat{w}(\mathbf{x}))|^2\mathrm{d}U(\mathbf{x})} \\
\leq&(c+\frac{c\beta}{\tau})\sup_{\boldsymbol{h}\in\mathcal{H}} \sqrt{\int_{\mathcal{X}} |r(\mathbf{x}) - \widehat{w}(\mathbf{x})|^2\mathrm{d}U(\mathbf{x})}.
\end{aligned}$$

Lastly, using Lemma 2,

$$\sup_{\boldsymbol{h}\in\mathcal{H}} |R^{\tau,\beta}_{U,u}(\boldsymbol{h}) - \int_{\mathcal{X}} L^{-}_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})| \leq (c+\frac{c\beta}{\tau})O_p(\lambda^{\frac{1}{2}}_{n,m}).$$

Step 3. We claim that

$$\sup_{\boldsymbol{h}\in\mathcal{H}} \big|\frac{1}{m}\sum_{\mathbf{x}\in T} L^{-}_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1}) - \int_{\mathcal{X}} L^{-}_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})\big| \leq c\big(1+\frac{\beta}{\tau}+\beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}).$$

First, we set $\mathcal{F}_B := \{L^{-}_{\tau,\beta}(w)\ell(\boldsymbol{h},\mathbf{y}_{C+1}) : w\in\mathcal{H}_K, \|w\|_K \leq B, \boldsymbol{h}\in\mathcal{H}\}$. We consider

$$\sup_{f\in\mathcal{F}_B} \big|\frac{1}{m}\sum_{\mathbf{x}\in T} f(\mathbf{x}) - \int_{\mathcal{X}} f(\mathbf{x})\mathrm{d}U(\mathbf{x})\big|$$

Using Lemma 3 and inequality 8, it is easy to check that for $1-2\delta>0$, we have

$$\sup_{f\in\mathcal{F}_B} \big|\frac{1}{m}\sum_{\mathbf{x}\in T} f(\mathbf{x}) - \int_{\mathcal{X}} f(\mathbf{x})\mathrm{d}U(\mathbf{x})\big| \leq 2\widehat{\mathfrak{R}}_T(\mathcal{F}_B) + 4(\tau+\beta)c\sqrt{\frac{2\log(4/\delta)}{m}}, \tag{18}$$

here we have used $|f| \leq (\tau+\beta)c$, for any $f\in\mathcal{F}_B$.

Then, we consider $\widehat{\Re}_T(\mathcal{F}_B)$.

$$
\begin{aligned}
& m\widehat{\Re}_T(\mathcal{F}_B) \\
=& \mathbb{E}_\sigma[\sup_{\|w\|_K\le B, \boldsymbol{h}\in\mathcal{H}} \sum_{i=1}^m \sigma_i L^-_{\tau,\beta}(w_i)\ell(\boldsymbol{h}_i, \mathbf{y}_{C+1})] \\
=& \mathbb{E}_\sigma[\sup_{\|w\|_K\le B, \boldsymbol{h}\in\mathcal{H}} \sigma_1 L^-_{\tau,\beta}(w_1)\ell(\boldsymbol{h}_1, \mathbf{y}_{C+1}) + \sum_{i=2}^m \sigma_i L^-_{\tau,\beta}(w_i)\ell(\boldsymbol{h}_i, \mathbf{y}_{C+1})] \\
=& \frac{1}{2}\mathbb{E}_{\sigma_2,...,\sigma_m}[\sup_{\|w\|_K\le B, \boldsymbol{h}\in\mathcal{H}} L^-_{\tau,\beta}(w_1)\ell(\boldsymbol{h}_1, \mathbf{y}_{C+1}) + \sum_{i=2}^m \sigma_i L^-_{\tau,\beta}(w_i)\ell(\boldsymbol{h}_i, \mathbf{y}_{C+1}) \\
& \quad + \sup_{\|w\|_K\le B, \boldsymbol{h}\in\mathcal{H}} -L^-_{\tau,\beta}(w_1)\ell(\boldsymbol{h}_1, \mathbf{y}_{C+1}) + \sum_{i=2}^m \sigma_i L^-_{\tau,\beta}(w_i)\ell(\boldsymbol{h}_i, \mathbf{y}_{C+1})] \\
=& \frac{1}{2}\mathbb{E}_{\sigma_2,...,\sigma_m}[\sup_{\|w\|_K\le B, \boldsymbol{h}\in\mathcal{H}} \sup_{\|w'\|_K\le B, \boldsymbol{h}'\in\mathcal{H}} L^-_{\tau,\beta}(w_1)\ell(\boldsymbol{h}_1, \mathbf{y}_{C+1}) + \sum_{i=2}^m \sigma_i L^-_{\tau,\beta}(w_i)\ell(\boldsymbol{h}_i, \mathbf{y}_{C+1}) \\
& \qquad - L^-_{\tau,\beta}(w'_1)\ell(\boldsymbol{h}'_1, \mathbf{y}_{C+1}) + \sum_{i=2}^m \sigma_i L^-_{\tau,\beta}(w'_i)\ell(\boldsymbol{h}'_i, \mathbf{y}_{C+1})] \\
\le& \frac{1}{2}\mathbb{E}_{\sigma_2,...,\sigma_m}[\sup_{\|w\|_K\le B, \boldsymbol{h}\in\mathcal{H}} \sup_{\|w'\|_K\le B, \boldsymbol{h}'\in\mathcal{H}} |L^-_{\tau,\beta}(w_1) - L^-_{\tau,\beta}(w'_1)|\ell(\boldsymbol{h}'_1, \mathbf{y}_{C+1}) + L^-_{\tau,\beta}(w_1)|\ell(\boldsymbol{h}'_1, \mathbf{y}_{C+1}) - \ell(\boldsymbol{h}_1, \mathbf{y}_{C+1})| \\
& \qquad + \sum_{i=2}^m \sigma_i L^-_{\tau,\beta}(w_i)\ell(\boldsymbol{h}_i, \mathbf{y}_{C+1}) + \sum_{i=2}^m \sigma_i L^-_{\tau,\beta}(w'_i)\ell(\boldsymbol{h}'_i, \mathbf{y}_{C+1})] \\
\le& \frac{1}{2}\mathbb{E}_{\sigma_2,...,\sigma_m}[\sup_{\|w\|_K\le B, \boldsymbol{h}\in\mathcal{H}} \sup_{\|w'\|_K\le B, \boldsymbol{h}'\in\mathcal{H}} Lc|w_1 - w'_1| + (B+\beta)|\ell(\boldsymbol{h}'_1, \mathbf{y}_{C+1}) - \ell(\boldsymbol{h}_1, \mathbf{y}_{C+1})| \\
& \qquad + \sum_{i=2}^m \sigma_i L^-_{\tau,\beta}(w_i)\ell(\boldsymbol{h}_i, \mathbf{y}_{C+1}) + \sum_{i=2}^m \sigma_i L^-_{\tau,\beta}(w'_i)\ell(\boldsymbol{h}'_i, \mathbf{y}_{C+1})] \\
=& \mathbb{E}_\sigma[\sup_{\|w\|_K\le B, \boldsymbol{h}\in\mathcal{H}} Lc\sigma_1 w_1 + (B+\beta)\sigma_1\ell(\boldsymbol{h}_1, \mathbf{y}_{C+1}) + \sum_{i=2}^m \sigma_i L^-_{\tau,\beta}(w_i)\ell(\boldsymbol{h}_i, \mathbf{y}_{C+1})]
\end{aligned}
$$

Repeat the process $m-1$ times for $i = 2, ..., m$.

$$
\begin{aligned}
\le& \mathbb{E}_\sigma[\sup_{\|w\|_K\le B, \boldsymbol{h}\in\mathcal{H}} \sum_{i=1}^m Lc\sigma_i w_i + \sum_{i=1}^m (B+\beta)\sigma_i\ell(\boldsymbol{h}_i, \mathbf{y}_{C+1})] \\
\le& mLc\widehat{\Re}_T(\mathcal{H}_{K,B}) + m(B+\beta)\widehat{\Re}_T(\mathcal{F}),
\end{aligned}
$$

where $w_i = w(\tilde{\mathbf{x}}_i), \boldsymbol{h}_i = \boldsymbol{h}(\tilde{\mathbf{x}}_i)$, $L = 1 + \frac{\beta}{\tau}$, $\mathcal{H}_{K,B} = \{w : w \in \mathcal{H}_K, \|w\|_K \le B\}$ and $\mathcal{F} = \{\ell(\boldsymbol{h}, \mathbf{y}_{C+1}) : \boldsymbol{h} \in \mathcal{H}\}$.

According to Theorem 5.5 of Mohri et al. (2012), we obtain that

$$\widehat{\Re}_T(\mathcal{H}_{K,B}) \le B\sqrt{\frac{1}{m}}.$$

According to the proving process of Lemma 4, we obtain that

$$\widehat{\Re}_T(\mathcal{F}) \le c\sqrt{\frac{4d\log m + 8d\log(C+1)}{m}},$$

where d is the Natarajan Dimension of $\mathcal{H}$.

Hence,

$$\widehat{\Re}_T(\mathcal{F}_B) \le BLc\sqrt{\frac{1}{m}} + (B+\beta)c\sqrt{\frac{4d\log m + 8d\log(C+1)}{m}}.$$

This implies that for $1-2\delta>0$, we have

$$\begin{aligned}&\sup_{w\in\mathcal{H}_K,\|w\|_K\leq B}\sup_{\boldsymbol{h}\in\mathcal{H}}\big|\frac{1}{m}\sum_{\mathbf{x}\in T}L^{-}_{\tau,\beta}(w(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})-\int_{\mathcal{X}}L^{-}_{\tau,\beta}(w(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})\big|\\ \leq&2B(1+\frac{\beta}{\tau})c\sqrt{\frac{1}{m}}+2(B+\beta)c\sqrt{\frac{4d\log m+8d\log(C+1)}{m}}+4(\tau+\beta)c\sqrt{\frac{2\log(4/\delta)}{m}}.\end{aligned}\tag{19}$$

Because $\|\widehat{w}\|_K=O_p(1)$, then combining inequality 19, we know that

$$\begin{aligned}&\sup_{\boldsymbol{h}\in\mathcal{H}}\big|\frac{1}{m}\sum_{\mathbf{x}\in T}L^{-}_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})-\int_{\mathcal{X}}L^{-}_{\tau,\beta}(w(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})\big|\\ \leq&2O_p(1)(1+\frac{\beta}{\tau})cO_p(\sqrt{\frac{1}{m}})+2(O_p(1)+\beta)cO_p(\sqrt{\frac{4d\log m+8d\log(C+1)}{m}})+4(\tau+\beta)cO_p(\sqrt{\frac{2\log(4/\delta)}{m}}).\end{aligned}$$

This implies that

$$\sup_{\boldsymbol{h}\in\mathcal{H}}\big|\frac{1}{m}\sum_{\mathbf{x}\in T}L^{-}_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})-\int_{\mathcal{X}}L^{-}_{\tau,\beta}(w(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})\big|\leq c\big(1+\frac{\beta}{\tau}+\tau+\beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}}).$$

Step 4. Using the results of Steps 1, 2 and 3, we have

$$\begin{aligned}&\sup_{\boldsymbol{h}\in\mathcal{H}}|R_{Q,u}(\boldsymbol{h})-\gamma'\widehat{R}^{\tau,\beta}_{S,T,u}(\boldsymbol{h})|\\ \leq&\sup_{\boldsymbol{h}\in\mathcal{H}}|R_{Q,u}(\boldsymbol{h})-\gamma' R^{\tau,\beta}_{U,u}(\boldsymbol{h})|+\sup_{\boldsymbol{h}\in\mathcal{H}}|\gamma' R^{\tau,\beta}_{U,u}(\boldsymbol{h})-\gamma'\int_{\mathcal{X}}L^{-}_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}))\mathrm{d}U(\mathbf{x})|\\ &+\sup_{\boldsymbol{h}\in\mathcal{H}}|\gamma'\int_{\mathcal{X}}L^{-}_{\tau,\beta}(\widehat{w}(\mathbf{x}))\ell(\boldsymbol{h}(\mathbf{x}),\mathbf{y}_{C+1})\mathrm{d}U(\mathbf{x})-\gamma'\widehat{R}^{\tau,\beta}_{S,T,u}(\boldsymbol{h},\mathbf{y}_{C+1})|\\ \leq&\gamma'\beta cU(\{\mathbf{x}:\ 0<r(\mathbf{x})\leq 2\tau\})+\gamma'(c+\frac{c\beta}{\tau})O_p(\lambda_{n,m}^{\frac{1}{2}})+c\gamma'\big(1+\frac{\beta}{\tau}+\tau+\beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}}).\end{aligned}$$

We can write

$$\sup_{\boldsymbol{h}\in\mathcal{H}}|R_{Q,u}(\boldsymbol{h})-\gamma'\widehat{R}^{\tau,\beta}_{S,T,u}(\boldsymbol{h})|\leq\gamma'\beta cU(\{\mathbf{x}:\ 0<r(\mathbf{x})\leq 2\tau\})+c\gamma'\big(1+\frac{\beta}{\tau}+\tau+\beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}}).$$

□

Proof of Theorem 6. Assume that

$$\widetilde{\boldsymbol{h}} \in \arg\min_{\boldsymbol{h}\in\mathcal{H}} \widetilde{R}^{\tau,\beta}_{S,T}(\boldsymbol{h}), \quad \boldsymbol{h}_Q \in \arg\min_{\boldsymbol{h}\in\mathcal{H}} R^{\alpha}_{Q}(\boldsymbol{h}).$$

Step 1. It is easy to check that

$$\begin{aligned} R^{\alpha}_{Q}(\widetilde{\boldsymbol{h}}) - R^{\alpha}_{Q}(\boldsymbol{h}_Q) =& R^{\alpha}_{Q}(\widetilde{\boldsymbol{h}}) - \widetilde{R}^{\tau,\beta}_{S,T}(\widetilde{\boldsymbol{h}}) + \widetilde{R}^{\tau,\beta}_{S,T}(\widetilde{\boldsymbol{h}}) - R^{\alpha}_{Q}(\boldsymbol{h}_Q) \\ \leq& R^{\alpha}_{Q}(\widetilde{\boldsymbol{h}}) - \widetilde{R}^{\tau,\beta}_{S,T}(\widetilde{\boldsymbol{h}}) + \widetilde{R}^{\tau,\beta}_{S,T}(\boldsymbol{h}_Q) - R^{\alpha}_{Q}(\boldsymbol{h}_Q) \\ \leq& 2\sup_{\boldsymbol{h}\in\mathcal{H}} |\widetilde{R}^{\tau,\beta}_{S,T}(\boldsymbol{h}) - R^{\alpha}_{Q}(\boldsymbol{h})|, \end{aligned}$$

and

$$\begin{aligned} R^{\alpha}_{Q}(\widetilde{\boldsymbol{h}}) - R^{\alpha}_{Q}(\boldsymbol{h}_Q) =& R^{\alpha}_{Q}(\widetilde{\boldsymbol{h}}) - \widetilde{R}^{\tau,\beta}_{S,T}(\boldsymbol{h}_Q) + \widetilde{R}^{\tau,\beta}_{S,T}(\boldsymbol{h}_Q) - R^{\alpha}_{Q}(\boldsymbol{h}_Q) \\ \geq& R^{\alpha}_{Q}(\widetilde{\boldsymbol{h}}) - \widetilde{R}^{\tau,\beta}_{S,T}(\widetilde{\boldsymbol{h}}) + \widetilde{R}^{\tau,\beta}_{S,T}(\boldsymbol{h}_Q) - R^{\alpha}_{Q}(\boldsymbol{h}_Q) \\ \geq& -2\sup_{\boldsymbol{h}\in\mathcal{H}} |\widetilde{R}^{\tau,\beta}_{S,T}(\boldsymbol{h}) - R^{\alpha}_{Q}(\boldsymbol{h})|, \end{aligned}$$

which implies that

$$|R^{\alpha}_{Q}(\widetilde{\boldsymbol{h}}) - R^{\alpha}_{Q}(\boldsymbol{h}_Q)| \leq 2\sup_{\boldsymbol{h}\in\mathcal{H}} |\widetilde{R}^{\tau,\beta}_{S,T}(\boldsymbol{h}) - R^{\alpha}_{Q}(\boldsymbol{h})|.$$

Using the result of Lemma 6 and Lemma 4, we obtain that

$$|R^{\alpha}_{Q}(\widetilde{\boldsymbol{h}}) - R^{\alpha}_{Q}(\boldsymbol{h}_Q)| \leq 2c\gamma'\big(1+\tau+\frac{\beta}{\tau}+\beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + 2\gamma' c\alpha\beta U(0<p/q\leq 2\tau). \tag{20}$$

Then, using the result of Step 3 in the proof of Theorem 2, we obtain that

$$|R^{\alpha}_{Q}(\widetilde{\boldsymbol{h}}) - R^{\alpha}_{P}(\boldsymbol{h}_Q)| \leq 2c\gamma'\big(1+\tau+\frac{\beta}{\tau}+\beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + 2\gamma' c\alpha\beta U(0<p/q\leq 2\tau). \tag{21}$$

Step 2.

$$\begin{aligned} \widetilde{R}^{\tau,\beta}_{S,T}(\widetilde{\boldsymbol{h}}) &= (1-\alpha)\widehat{R}_S(\widetilde{\boldsymbol{h}}) + \alpha\gamma'\widehat{R}^{\tau,\beta}_{S,T,u}(\widetilde{\boldsymbol{h}}) \\ &\leq (1-\alpha)\widehat{R}_S(\boldsymbol{h}_Q) + \alpha\gamma'\widehat{R}^{\tau,\beta}_{S,T,u}(\boldsymbol{h}_Q) \\ &\leq R^{\alpha}_{Q}(\boldsymbol{h}_Q) + c\gamma'\big(1+\tau+\frac{\beta}{\tau}+\beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + \gamma' c\alpha\beta U(0<p/q\leq 2\tau) \text{ Using Lemma 6 and Lemma 4} \\ &= (1-\alpha)\min_{\boldsymbol{h}\in\mathcal{H}} R_{Q,k}(\boldsymbol{h}) + c\gamma'\big(1+\tau+\frac{\beta}{\tau}+\beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + \gamma' c\alpha\beta U(0<p/q\leq 2\tau) \\ &\quad \text{Using the result of Step 3 in proof of Theorem 2}: \min_{\boldsymbol{h}\in\mathcal{H}} R^{\alpha}_{Q}(\boldsymbol{h}) = (1-\alpha)\min_{\boldsymbol{h}\in\mathcal{H}} R_{Q,k}(\boldsymbol{h}) \\ &\leq (1-\alpha)R_{Q,k}(\widetilde{\boldsymbol{h}}) + c\gamma'\big(1+\tau+\frac{\beta}{\tau}+\beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + \gamma' c\alpha\beta U(0<p/q\leq 2\tau). \end{aligned}$$

Hence,

$$\begin{aligned} &\alpha\gamma'\widehat{R}^{\tau,\beta}_{S,T,u}(\widetilde{\boldsymbol{h}}) \\ \leq& (1-\alpha)R_{Q,k}(\widetilde{\boldsymbol{h}}) - (1-\alpha)\widehat{R}_S(\widetilde{\boldsymbol{h}}) + c\gamma'\big(1+\tau+\frac{\beta}{\tau}+\beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + \gamma' c\alpha\beta U(0<p/q\leq 2\tau) \\ =& (1-\alpha)R_{P,k}(\widetilde{\boldsymbol{h}}) - (1-\alpha)\widehat{R}_S(\widetilde{\boldsymbol{h}}) + c\gamma'\big(1+\tau+\frac{\beta}{\tau}+\beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + \gamma' c\alpha\beta U(0<p/q\leq 2\tau) \\ \leq& (1-\alpha)cO_p(\lambda^{\frac{1}{2}}_{n,m}) + c\gamma'\big(1+\tau+\frac{\beta}{\tau}+\beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + \gamma' c\alpha\beta U(0<p/q\leq 2\tau) \text{ Using the result of Lemma 4} \\ \leq& 2c\gamma'\big(1+\tau+\frac{\beta}{\tau}+\beta\big)O_p(\lambda^{\frac{1}{2}}_{n,m}) + \gamma' c\alpha\beta U(0<p/q\leq 2\tau). \end{aligned}$$

Then, combining the above inequality with the result of Lemma 6, we obtain that

$$\alpha R_{Q,u}(\widetilde{\boldsymbol{h}}) \leq 3c\gamma'\big(1+\tau+\frac{\beta}{\tau}+\beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}})+2\gamma' c\alpha\beta U(0<p/q\leq 2\tau).$$

Step 3.

$$\begin{aligned}
&|R_P^\alpha(\widetilde{\boldsymbol{h}})-R_P^\alpha(\boldsymbol{h}_Q)|\\
\leq&|R_P^\alpha(\widetilde{\boldsymbol{h}})-R_Q^\alpha(\widetilde{\boldsymbol{h}})|+|R_Q^\alpha(\widetilde{\boldsymbol{h}})-R_P^\alpha(\boldsymbol{h}_Q)|\\
=&\alpha|R_{P,u}(\widetilde{\boldsymbol{h}})-R_{Q,u}^\alpha(\widetilde{\boldsymbol{h}})|+|R_Q^\alpha(\widetilde{\boldsymbol{h}})-R_P^\alpha(\boldsymbol{h}_Q)|\\
\leq&\alpha R_{Q,u}(\widetilde{\boldsymbol{h}})+|R_Q^\alpha(\widetilde{\boldsymbol{h}})-R_P^\alpha(\boldsymbol{h}_Q)|\\
\leq&5c\gamma'\big(1+\tau+\frac{\beta}{\tau}+\beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}})+4\gamma' c\alpha\beta U(0<p/q\leq 2\tau) \text{ Using the results of Step 1 and Step 2.}
\end{aligned}$$

Briefly, we can write (absorbing coefficient 5 into O_p)

$$|R_P^\alpha(\widetilde{\boldsymbol{h}})-R_P^\alpha(\boldsymbol{h}_Q)| \leq c\gamma'\big(1+\tau+\frac{\beta}{\tau}+\beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}})+4\gamma' c\alpha\beta U(0<p/q\leq 2\tau).$$

Combining the above inequality with Theorem 2, we obtain that

$$|R_P^\alpha(\widetilde{\boldsymbol{h}})-\min_{\boldsymbol{h}\in\mathcal{H}} R_P^\alpha(\boldsymbol{h})| \leq c\gamma'\big(1+\tau+\frac{\beta}{\tau}+\beta\big)O_p(\lambda_{n,m}^{\frac{1}{2}})+4\gamma' c\alpha\beta U(0<p/q\leq 2\tau).$$

□

6. Appendix F: Details on Experiments

6.1. Datasets

• MNIST dataset (LeCun & Cortes, 2010). The MNIST[1] database of handwritten digits, has a training set of $60,000$ samples, and a testing set of $10,000$ samples. The digits have been size-normalized and centered in a fixed-size image. Following the set up in Yoshihashi et al. (2019), we use MNIST (LeCun & Cortes, 2010) as the training samples and use Omniglot (Ager, 2008), MNIST-Noise, and Noise these datasets as unknown classes. Omniglot contains alphabet characters. Noise is synthesized by sampling each pixel value from a uniform distribution on $[0, 1]$ (i.i.d). MNIST-Noise is synthesized by adding noise on MNIST testing samples. Each dataset has $10,000$ testing samples.

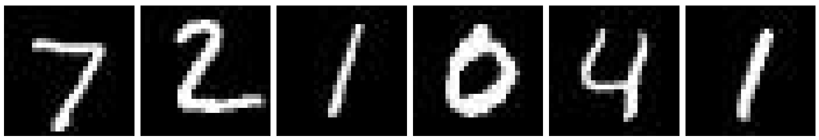


Figure 1. MNIST.

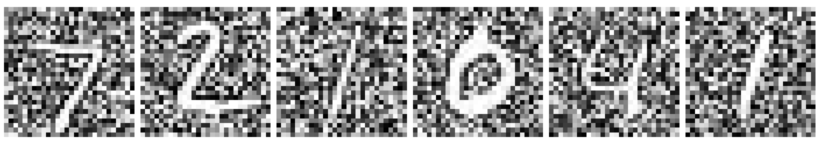


Figure 2. MNIST-Noise.

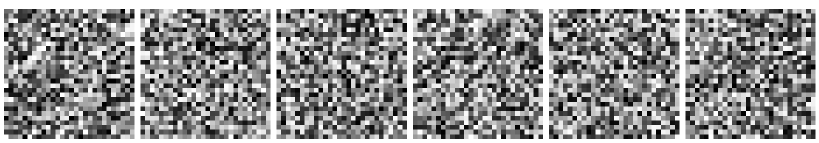


Figure 3. Noise.

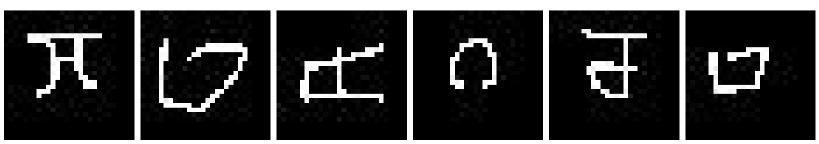


Figure 4. Omniglot.

Table 2. Introduction of MNIST Dataset in Open-set learning.

Dataset	#Sample	#Class	Known/Unknown	Train/Test
MNIST	60,000	10	Known Classes	Train
MNIST	10,000	10	Known Classes	Test
MNIST-Noise	10,000	10	Unknown Classes	Test
Omniglot	10,000	1,623	Unknown Classes	Test
Noise	10,000	1	Unknown Classes	Test

[1] http://yann.lecun.com/exdb/mnist/

Table 3. Introduction of CIFAR-10 Dataset in Open-set learning.

Dataset	#Sample	#Class	Known/Unknown	Train/Test
CIFAR-10	50,000	10	Known Classes	Train
CIFAR-10	10,000	10	Known Classes	Test
ImageNet-crop	10,000	1,000	Unknown Classes	Test
ImageNet-resize	10,000	1,000	Unknown Classes	Test
LSUN-crop	10,000	10	Unknown Classes	Test
LSUN-resize	10,000	10	Unknown Classes	Test

• CIFAR-10 dataset. The CIFAR-10 dataset consists of $60,000$ 32×32 colour images in 10 classes, with $6,000$ images per class. There are $50,000$ training images and $10,000$ testing images. Following the set up in Yoshihashi et al. (2019), we use the training samples from CIFAR-10 (Krizhevsky & Hinton, 2009) as training samples in open-set learning problem. We collect unknown samples from datasets ImageNet and LSUN. Similar to Yoshihashi et al. (2019), we resized or cropped them so that they would have the same sizes with known samples. Hence, we generated four datasets ImageNet-crop, ImageNet-resize, LSUN-crop and LSUN-resize as unknown classes.

Figure 5. CIFAR-10.

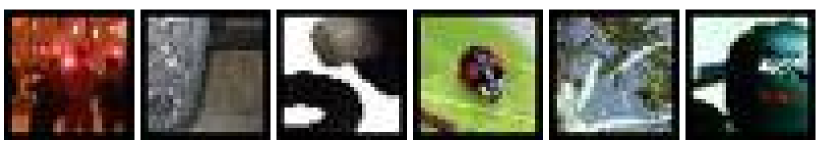


Figure 6. ImageNet-crop.

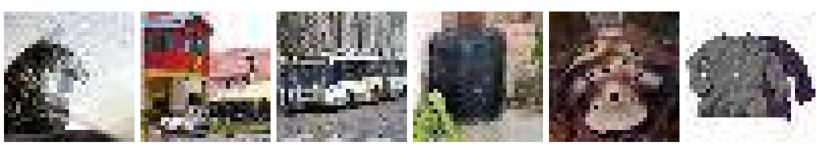


Figure 7. ImageNet-resize.

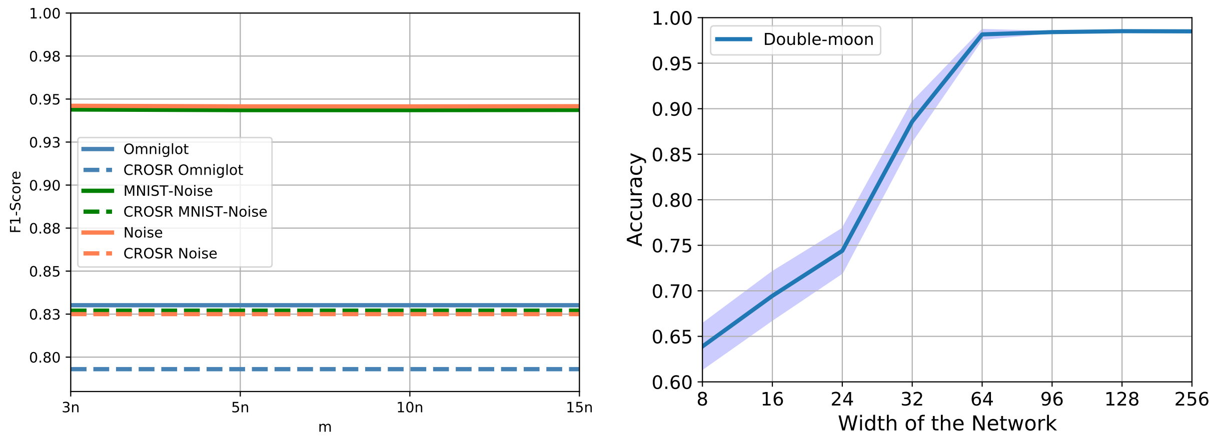
Figure 8. LSUN-crop.

Figure 9. LSUN-resize.

6.2. Network Architecture and Experimental Setup

All details can be found in github.com/Anjin-Liu/Openset_Learning_AOSR.

6.3. Parameter Analysis and Influence of Model Capacity



(a) Parameter Analysis for m (b) Influence of Model Capacity

Figure 10. Parameter Analysis and Influence of Model Capacity

Experiment results on parameter m are shown in Figure 10 (a). m is the size of generated samples T. We set $m = 3n, 5n, 10n$ and $15n$. By changing m in the range of $3n, 5n, 10n, 15n$, AOSR achieves consistent performance. This result can be explained by our theory. Because when $m > n$, the increases of m does not influence the error bound in Theorem 6.

Experiment results on the width of the network are shown in Figure 10 (b). We generate $2,000$ training samples and adjust the width for the second to the last layer from 8 to 256. For different width, we run 100 times and report the mean accuracy and standard error. As increasing the network's width from 8 to 256, the accuracy of double-moon increases. When the width is larger than 64, the performance achieves a stable performance. This means the model capacity has a profound impact on the performance of OSL. Generally, the larger the model capacity is, the better the model's performance is. This is because a larger hypothesis space $\mathcal{H}$ has a greater possibility to meet the conditions of Assumption 1 (realization assumption for unknown classes).

References

Ager, S. Omniglot-writing systems and languages of the world. *Retrieved January*, 27:2008, 2008.

Kanamori, T., Suzuki, T., and Sugiyama, M. Condition number analysis of kernel-based density ratio estimation. *Technical Report TR09-0006, Department of Computer Science, Tokyo Institute of Technology*, 2009.

Kanamori, T., Suzuki, T., and Sugiyama, M. Statistical analysis of kernel-based least-squares density-ratio estimation. *Mach. Learn.*, pp. 335–367, 2012.

Krizhevsky, A. and Hinton, G. Convolutional deep belief networks on cifar-10. *Technical report, Citeseer*, 2009.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. 2012.

Shalev-Shwartz, S. and Ben-David, S. Understanding machine learning: From theory to algorithms. In *Cambridge university press*, 2014.

Yoshihashi, R., Shao, W., Kawakami, R., You, S., Iida, M., and Naemura, T. Classification-reconstruction learning for open-set recognition. In *CVPR*, pp. 4016–4025, 2019.