Proofs of the Theorems

Wei Wang and Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
{wangw,zhouzh}@lamda.nju.edu.cn

Proof of Theorem 1. Let $Q_i = S_1^i \oplus S_2^i$. First we prove that if each view X_v ($v = 1, 2$) satisfies Tsybakov noise condition, i.e., $Pr_{x_v \in X_v}(|\varphi_v(x_v) - 1/2| \leq t) \leq C_3 t^{\lambda_3}$ for some finite $C_3 > 0$, $\lambda_3 > 0$ and all $0 < t \leq 1/2$, Tsybakov noise condition can also be met in Q_i, i.e., $\frac{Pr_{x_v \in Q_i}(|\varphi_v(x_v) - 1/2| \leq t)}{Pr(Q_i)} \leq C_4 t^{\lambda_4}$ for some finite $C_4 > 0$, $\lambda_4 > 0$ and all $0 < t \leq 1/2$. Suppose Tsybakov noise condition cannot be met in Q_i, then for $C_* = \frac{C_3}{Pr(Q_i)}$ and $\lambda_* = \lambda_3$, there exists some $0 < t_* \leq 1/2$ to satisfy that $\frac{Pr_{x_v \in Q_i}(|\varphi_v(x_v) - 1/2| \leq t)}{Pr(Q_i)} > C_* t_*^{\lambda_*}$. So we get

$$Pr_{x_v \in X_v}(|\varphi_v(x_v) - 1/2| \leq t) \geq Pr_{x_v \in Q_i}(|\varphi_v(x_v) - 1/2| \leq t) > C_3 t_*^{\lambda_3}.$$

It is in contradiction with that X_v satisfies Tsybakov noise condition. Thus, we get that Tsybakov noise condition can also be met in Q_i. Without loss of generality, suppose that Tsybakov noise condition in all Q_i and X_v can be met for the same finite C_0 and λ.

Since $m_0 = \frac{256^k C}{C_1^2}\big(V + \log(\frac{16(s+1)}{\delta})\big)$, according to Lemma 1 we know that $d(S_v^0, S^*) \leq \frac{C_1}{16^k}$ with probability at least $1 - \frac{\delta}{16(s+1)}$. With $d(S_v, S_v^*) \geq C_1 d_\Delta^k(S_v, S_v^*)$, we get $d_\Delta(S_v^0, S^*) \leq \frac{1}{16}$. It is easy to find that $d_\Delta(S_1^0 \cap S_2^0, S^*) \leq d_\Delta(S_1^0, S^*) + d_\Delta(S_2^0, S^*) \leq 1/8$ holds with probability at least $1 - \frac{\delta}{8(s+1)}$.

For $i \geq 0$, m_{i+1} number of labels are queried randomly from Q_i. Thus, similarly according to Lemma 1 we have $d_\Delta(S_1^{i+1} \cap S_2^{i+1} \mid Q_i, S^* \mid Q_i) \leq 1/8$ with probability at least $1 - \frac{\delta}{8(s+1)}$. Let $T_v^{i+1} = S_v^{i+1} \cap \overline{Q_i}$ and $\tau_{i+1} = \frac{Pr(T_1^{i+1} \oplus T_2^{i+1} - S^*)}{Pr(T_1^{i+1} \oplus T_2^{i+1})} - \frac{1}{2}$, it is easy to get

$$Pr\big(S^* \cap (S_1^{i+1} \oplus S_2^{i+1})|\overline{Q_i}\big) - Pr\big(\overline{S^*} \cap (S_1^{i+1} \oplus S_2^{i+1})|\overline{Q_i}\big) = -2\tau_{i+1} Pr(S_1^{i+1} \oplus S_2^{i+1}|\overline{Q_i}).$$

Considering the non-degradation condition and $d_\Delta(S_1^i \cap S_2^i|\overline{Q_i}, S^*|\overline{Q_i}) = d_\Delta(S_v^i|\overline{Q_i}, S^*|\overline{Q_i})$, we calculate that

$$\begin{aligned}
& d_\Delta(S_1^{i+1} \cap S_2^{i+1}|\overline{Q_i}, S^*|\overline{Q_i}) \\
= \; & \frac{1}{2}\Big(d_\Delta(S_1^{i+1}|\overline{Q_i}, S^*|\overline{Q_i}) + d_\Delta(S_2^{i+1}|\overline{Q_i}, S^*|\overline{Q_i})\Big) + \frac{1}{2}Pr\Big(S^* \cap (S_1^{i+1} \oplus S_2^{i+1})|\overline{Q_i}\Big) \\
& -\frac{1}{2}Pr\Big(\overline{S^*} \cap (S_1^{i+1} \oplus S_2^{i+1})|\overline{Q_i}\Big) \\
\leq \; & \frac{1}{2}\Big(d_\Delta(S_1^i|\overline{Q_i}, S^*|\overline{Q_i}) + d_\Delta(S_2^i|\overline{Q_i}, S^*|\overline{Q_i})\Big) - \tau_{i+1} Pr(S_1^{i+1} \oplus S_2^{i+1}|\overline{Q_i}) \\
= \; & d_\Delta(S_1^i \cap S_2^i|\overline{Q_i}, S^*|\overline{Q_i}) - \tau_{i+1} Pr(S_1^{i+1} \oplus S_2^{i+1}|\overline{Q_i}).
\end{aligned}$$

So we have

$$\begin{aligned}
& d_\Delta(S_1^{i+1} \cap S_2^{i+1}, S^*) \\
= \; & d_\Delta(S_1^{i+1} \cap S_2^{i+1}|Q_i, S^*|Q_i) Pr(Q_i) + d_\Delta(S_1^{i+1} \cap S_2^{i+1}|\overline{Q_i}, S^*|\overline{Q_i}) Pr(\overline{Q_i}) \\
\leq \; & \frac{1}{8}Pr(Q_i) + d_\Delta(S_1^i \cap S_2^i|\overline{Q_i}, S^*|\overline{Q_i}) Pr(\overline{Q_i}) - \tau_{i+1} Pr\big((S_1^{i+1} \oplus S_2^{i+1}) \cap \overline{Q_i}\big).
\end{aligned}$$

Considering $d_\Delta(S_1^i \cap S_2^i | \overline{Q_i}, S^* | \overline{Q_i}) Pr(\overline{Q_i}) = Pr(S_1^i \cap S_2^i - S^*) + Pr(\overline{S_1^i} \cap \overline{S_2^i} - \overline{S^*})$, we have

$$\begin{aligned} & d_\Delta(S_1^{i+1} \cap S_2^{i+1}, S^*) \\ \leq \ & Pr(S_1^i \cap S_2^i - S^*) + Pr(\overline{S_1^i} \cap \overline{S_2^i} - \overline{S^*}) + \frac{1}{8} Pr(S_1^i \oplus S_2^i) - \tau_{i+1} Pr\big((S_1^{i+1} \oplus S_2^{i+1}) \cap \overline{Q_i}\big). \end{aligned}$$

Similarly, we get

$$\begin{aligned} & d_\Delta(S_1^{i+1} \cup S_2^{i+1}, S^*) \\ \leq \ & Pr(S_1^i \cap S_2^i - S^*) + Pr(\overline{S_1^i} \cap \overline{S_2^i} - \overline{S^*}) + \frac{1}{8} Pr(S_1^i \oplus S_2^i) + \tau_{i+1} Pr\big((S_1^{i+1} \oplus S_2^{i+1}) \cap \overline{Q_i}\big). \end{aligned}$$

Let $\gamma_i = \frac{Pr(S_1^i \oplus S_2^i - S^*)}{Pr(S_1^i \oplus S_2^i)} - \frac{1}{2}$, we have

$$\begin{aligned} d_\Delta(S_1^i \cap S_2^i, S^*) &= d_\Delta(S_1^i \cap S_2^i | Q_i, S^* | Q_i) Pr(Q_i) + d_\Delta(S_1^i \cap S_2^i | \overline{Q_i}, S^* | \overline{Q_i}) Pr(\overline{Q_i}) \\ &= (1/2 - \gamma_i) Pr(S_1^i \oplus S_2^i) + Pr(S_1^i \cap S_2^i - S^*) + Pr(\overline{S_1^i} \cap \overline{S_2^i} - \overline{S^*}) \end{aligned}$$

and $d_\Delta(S_1^i \cup S_2^i, S^*) = (1/2 + \gamma_i) Pr(S_1^i \oplus S_2^i) + Pr(S_1^i \cap S_2^i - S^*) + Pr(\overline{S_1^i} \cap \overline{S_2^i} - \overline{S^*})$.

As in each round of the multi-view active learning some contention points of the two views are queried and added into the training set, the difference between the two views is decreasing, i.e., $Pr(S_1^{i+1} \oplus S_2^{i+1})$ is no larger than $Pr(S_1^i \oplus S_2^i)$.

Case 1: If $|\tau_{i+1}| \leq \gamma_i$, with respect to Definition 1, we have

$$\begin{aligned} \frac{d_\Delta(S_1^{i+1} \cup S_2^{i+1}, S^*)}{d_\Delta(S_1^i \cup S_2^i, S^*)} &\leq \frac{\frac{1}{8} Pr(S_1^i \oplus S_2^i) + |\tau_{i+1}| Pr(S_1^{i+1} \oplus S_2^{i+1}) + \frac{1}{\alpha} Pr(S_1^i \oplus S_2^i)}{(\frac{1}{2} + \gamma_i) Pr(S_1^i \oplus S_2^i) + \frac{1}{\alpha} Pr(S_1^i \oplus S_2^i)} \\ &\leq \frac{(\frac{1}{8} + \gamma_i) Pr(S_1^i \oplus S_2^i) + \frac{1}{\alpha} Pr(S_1^i \oplus S_2^i)}{(\frac{1}{2} + \gamma_i) Pr(S_1^i \oplus S_2^i) + \frac{1}{\alpha} Pr(S_1^i \oplus S_2^i)} \leq \frac{5\alpha + 8}{8\alpha + 8}; \end{aligned}$$

Case 2: If $-|\tau_{i+1}| > \gamma_i$, with respect to Definition 1, we have

$$\begin{aligned} \frac{d_\Delta(S_1^{i+1} \cap S_2^{i+1}, S^*)}{d_\Delta(S_1^i \cap S_2^i, S^*)} &\leq \frac{\frac{1}{8} Pr(S_1^i \oplus S_2^i) + |\tau_{i+1}| Pr(S_1^{i+1} \oplus S_2^{i+1}) + \frac{1}{\alpha} Pr(S_1^i \oplus S_2^i)}{(\frac{1}{2} + |\gamma_i|) Pr(S_1^i \oplus S_2^i) + \frac{1}{\alpha} Pr(S_1^i \oplus S_2^i)} \\ &\leq \frac{5\alpha + 8}{8\alpha + 8}; \end{aligned}$$

Case 3: If $\tau_{i+1} \geq \gamma_i$ and $0 \leq \gamma_i \leq \frac{1}{4}$, with respect to Definition 1, we have

$$\begin{aligned} \frac{d_\Delta(S_1^{i+1} \cap S_2^{i+1}, S^*)}{d_\Delta(S_1^i \cap S_2^i, S^*)} &\leq \frac{\frac{1}{8} Pr(S_1^i \oplus S_2^i) + \frac{1}{\alpha} Pr(S_1^i \oplus S_2^i)}{(\frac{1}{2} - \gamma_i) Pr(S_1^i \oplus S_2^i) + \frac{1}{\alpha} Pr(S_1^i \oplus S_2^i)} \\ &\leq \frac{\alpha + 8}{2\alpha + 8}; \end{aligned}$$

Case 4: If $\tau_{i+1} \geq \gamma_i$ and $\frac{1}{4} < \gamma_i \leq \frac{1}{2}$, with respect to Definition 1, we have

$$\begin{aligned} \frac{d_\Delta(S_1^{i+1} \cup S_2^{i+1}, S^*)}{d_\Delta(S_1^i \cup S_2^i, S^*)} &\leq \frac{\frac{1}{8} Pr(S_1^i \oplus S_2^i) + \tau_{i+1} Pr(S_1^{i+1} \oplus S_2^{i+1}) + \frac{1}{\alpha} Pr(S_1^i \oplus S_2^i)}{(\frac{1}{2} + \gamma_i) Pr(S_1^i \oplus S_2^i) + \frac{1}{\alpha} Pr(S_1^i \oplus S_2^i)} \\ &\leq \frac{5\alpha + 8}{6\alpha + 8}; \end{aligned}$$

Case 5: If $\tau_{i+1} < \gamma_i$ and $-\frac{1}{4} \leq \gamma_i \leq 0$, with respect to Definition 1, we have

$$\begin{aligned} \frac{d_\Delta(S_1^{i+1} \cup S_2^{i+1}, S^*)}{d_\Delta(S_1^i \cup S_2^i, S^*)} &\leq \frac{\frac{1}{8} Pr(S_1^i \oplus S_2^i) + \frac{1}{\alpha} Pr(S_1^i \oplus S_2^i)}{(\frac{1}{2} + \gamma_i) Pr(S_1^i \oplus S_2^i) + \frac{1}{\alpha} Pr(S_1^i \oplus S_2^i)} \\ &\leq \frac{\alpha + 8}{2\alpha + 8}; \end{aligned}$$

Case 6: If $\tau_{i+1} < \gamma_i$ and $-\frac{1}{2} \le \gamma_i < -\frac{1}{4}$, with respect to Definition 1, we have

$$\begin{aligned}\frac{d_\Delta(S_1^{i+1} \cap S_2^{i+1}, S^*)}{d_\Delta(S_1^i \cap S_2^i, S^*)} &\le \frac{\frac{1}{8}Pr(S_1^i \oplus S_2^i) + |\tau_{i+1}|Pr(S_1^{i+1} \oplus S_2^{i+1}) + \frac{1}{\alpha}Pr(S_1^i \oplus S_2^i)}{(\frac{1}{2}+|\gamma_i|)Pr(S_1^i \oplus S_2^i) + \frac{1}{\alpha}Pr(S_1^i \oplus S_2^i)} \\ &\le \frac{5\alpha+8}{6\alpha+8};\end{aligned}$$

Case 7: If $\tau_{i+1} \le -\gamma_i$ and $0 \le \gamma_i \le \frac{1}{2}$, with respect to Definition 1, we have

$$\begin{aligned}\frac{d_\Delta(S_1^{i+1} \cup S_2^{i+1}, S^*)}{d_\Delta(S_1^i \cup S_2^i, S^*)} &\le \frac{\frac{1}{8}Pr(S_1^i \oplus S_2^i) + \frac{1}{\alpha}Pr(S_1^i \oplus S_2^i)}{(\frac{1}{2}+\gamma_i)Pr(S_1^i \oplus S_2^i) + \frac{1}{\alpha}Pr(S_1^i \oplus S_2^i)} \\ &\le \frac{\alpha+8}{4\alpha+8};\end{aligned}$$

Case 8: If $\tau_{i+1} > -\gamma_i$ and $-\frac{1}{2} \le \gamma_i \le 0$, with respect to Definition 1, we have

$$\begin{aligned}\frac{d_\Delta(S_1^{i+1} \cap S_2^{i+1}, S^*)}{d_\Delta(S_1^i \cap S_2^i, S^*)} &\le \frac{\frac{1}{8}Pr(S_1^i \oplus S_2^i) + \frac{1}{\alpha}Pr(S_1^i \oplus S_2^i)}{(\frac{1}{2}+|\gamma_i|)Pr(S_1^i \oplus S_2^i) + \frac{1}{\alpha}Pr(S_1^i \oplus S_2^i)} \\ &\le \frac{\alpha+8}{4\alpha+8}.\end{aligned}$$

Thus, after the $(i+1)$-th round, either $\frac{d_\Delta(S_1^{i+1} \cap S_2^{i+1}, S^*)}{d_\Delta(S_1^i \cap S_2^i, S^*)} \le \frac{5\alpha+8}{6\alpha+8}$ or $\frac{d_\Delta(S_1^{i+1} \cup S_2^{i+1}, S^*)}{d_\Delta(S_1^i \cup S_2^i, S^*)} \le \frac{5\alpha+8}{6\alpha+8}$ holds. Hence, we have $d_\Delta(S_1^s \cap S_2^s, S^*) \le \frac{1}{8}\left(\frac{5\alpha+8}{6\alpha+8}\right)^{s/2}$ or $d_\Delta(S_1^s \cup S_2^s, S^*) \le \frac{1}{8}\left(\frac{5\alpha+8}{6\alpha+8}\right)^{s/2}$ with probability at least $1-\delta$. When $s = \lceil \frac{2\log\frac{1}{8\epsilon}}{\log\frac{1}{C_2}} \rceil$, where $C_2 = \frac{5\alpha+8}{6\alpha+8}$ is a constant less than 1, we have either $d_\Delta(S_1^s \cap S_2^s, S^*) \le \epsilon$ or $d_\Delta(S_1^s \cup S_2^s, S^*) \le \epsilon$ with probability at least $1-\delta$. Thus, considering $R(h_+^i) - R(S^*) = R(S_1^i \cap S_2^i) - R(S^*) \le d_\Delta(S_1^i \cap S_2^i, S^*)$ and $R(h_-^i) - R(S^*) = R(S_1^i \cup S_2^i) - R(S^*) \le d_\Delta(S_1^i \cup S_2^i, S^*)$, we have either $R(h_+^s) \le R(S^*) + \epsilon$ or $R(h_-^s) \le R(S^*) + \epsilon$. □

Proof of Lemma 2. We apply S_1^s and S_2^s to the unlabeled instances set and identify the contention point set. Then we query for labels of $\frac{2\log(\frac{4}{\delta})}{\beta^2}$ instances drawn randomly from the contention points set. With these labels we estimate the empirical value $\widehat{P}_1$ of $\frac{Pr(\{x: x\in S_1^s \oplus S_2^s \wedge y(x)=1\})}{Pr(S_1^s \oplus S_2^s)}$ and the empirical value $\widehat{P}_2$ of $\frac{Pr(\{x: x\in S_1^s \oplus S_2^s \wedge y(x)=0\})}{Pr(S_1^s \oplus S_2^s)}$. By Chernoff bound, with number of $\frac{2\log(\frac{4}{\delta})}{\beta^2}$ labels we have the following two equations with probability at least $1-\delta$.

$$\widehat{P}_1 \in \Big[\frac{Pr\big(\{x: x\in S_1^s \oplus S_2^s \wedge y(x)=1\}\big)}{Pr(S_1^s \oplus S_2^s)} - \frac{\beta}{2}, \frac{Pr\big(\{x: x\in S_1^s \oplus S_2^s \wedge y(x)=1\}\big)}{Pr(S_1^s \oplus S_2^s)} + \frac{\beta}{2}\Big]$$

$$\widehat{P}_2 \in \Big[\frac{Pr\big(\{x: x\in S_1^s \oplus S_2^s \wedge y(x)=0\}\big)}{Pr(S_1^s \oplus S_2^s)} - \frac{\beta}{2}, \frac{Pr\big(\{x: x\in S_1^s \oplus S_2^s \wedge y(x)=0\}\big)}{Pr(S_1^s \oplus S_2^s)} + \frac{\beta}{2}\Big]$$

If $\widehat{P}_1 \le \widehat{P}_2$, we get $Pr\big(\{x: x\in S_1^s \oplus S_2^s \wedge y(x)=1\}\big) \le Pr\big(\{x: x\in S_1^s \oplus S_2^s \wedge y(x)=0\}\big)$ with probability at least $1-\delta$; otherwise, we get $Pr\big(\{x: x\in S_1^s \oplus S_2^s \wedge y(x)=1\}\big) > Pr\big(\{x: x\in S_1^s \oplus S_2^s \wedge y(x)=0\}\big)$ with probability at least $1-\delta$. □

Proof of Theorem 2. According to Theorem 1, by requesting $\widetilde{O}(\log\frac{1}{\epsilon})$ labels the multi-view active learning in Table 1 can get either $R(h_+^s) \le R(S^*) + \epsilon$ or $R(h_-^s) \le R(S^*) + \epsilon$ with probability at least $1-\frac{\delta}{2}$. According to Lemma 2, by requesting $\frac{2\log(\frac{8}{\delta})}{\beta^2}$ labels we can decide correctly whether $Pr\big(\{x: x\in S_1^s \oplus S_2^s \wedge y(x)=1\}\big)$ or $Pr\big(\{x: x\in S_1^s \oplus S_2^s \wedge y(x)=0\}\big)$ is smaller with probability at least $1-\frac{\delta}{2}$.

Case 1: If $Pr\big(\{x: x\in S_1^s \oplus S_2^s \wedge y(x)=1\}\big) \le Pr\big(\{x: x\in S_1^s \oplus S_2^s \wedge y(x)=0\}\big)$, we have $R(h_-^s) \le R(h_+^s)$. Thus, we get $R(h_-^s) \le R(S^*) + \epsilon$ with probability at least $1-\delta$.

Case 2: If $Pr\big(\{x : x \in S_1^s \oplus S_2^s \wedge y(x) = 1\}\big) > Pr\big(\{x : x \in S_1^s \oplus S_2^s \wedge y(x) = 0\}\big)$, we have $R(h_+^s) < R(h_-^s)$. Thus, we get $R(h_+^s) \leq R(S^*) + \epsilon$ with probability at least $1 - \delta$.

The total number of labels to be requested is $\widetilde{O}(\log \frac{1}{\epsilon}) + \frac{2\log(\frac{8}{\delta})}{\beta^2} = \widetilde{O}(\log \frac{1}{\epsilon})$. □

Proof of Theorem 3. Since $Pr(S_1^s \oplus S_2^s) \leq 1$, with the following equation

$$\Big|\frac{Pr\big(\{x : x \in S_1^s \oplus S_2^s \wedge y(x) = 1\}\big)}{Pr(S_1^s \oplus S_2^s)} - \frac{Pr\big(\{x : x \in S_1^s \oplus S_2^s \wedge y(x) = 0\}\big)}{Pr(S_1^s \oplus S_2^s)}\Big| = O(\epsilon)$$

we have $|Pr\big(\{x : x \in S_1^s \oplus S_2^s \wedge y(x) = 1\}\big) - Pr\big(\{x : x \in S_1^s \oplus S_2^s \wedge y(x) = 0\}\big)| = O(\epsilon)$. So it is easy to get $|R(h_+^s) - R(h_-^s)| = O(\epsilon)$. According to Theorem 1, by requesting $\widetilde{O}(\log \frac{1}{\epsilon})$ labels we can get either $R(h_+^s) \leq R(S^*) + \epsilon$ or $R(h_-^s) \leq R(S^*) + \epsilon$ with probability at least $1 - \delta$. Thus, we get that h_+^s and h_-^s satisfy either (a) or (b) with probability at least $1 - \delta$. □

Proof of Theorem 5. According to Theorem 4, by requesting $\widetilde{O}(\log \frac{1}{\epsilon})$ labels the multi-view active learning in Table 1 can get either $R(h_+^s) \leq R(S_1^* \cap S_2^*) + \epsilon$ or $R(h_-^s) \leq R(S_1^* \cap S_2^*) + \epsilon$ with probability at least $1 - \frac{\delta}{2}$. According to Lemma 2, by requesting $\frac{2\log(\frac{8}{\delta})}{\beta^2}$ labels we can decide correctly whether $Pr\big(\{x : x \in S_1^s \oplus S_2^s \wedge y(x) = 1\}\big)$ or $Pr\big(\{x : x \in S_1^s \oplus S_2^s \wedge y(x) = 0\}\big)$ is smaller with probability at least $1 - \frac{\delta}{2}$.

Case 1: If $Pr\big(\{x : x \in S_1^s \oplus S_2^s \wedge y(x) = 1\}\big) \leq Pr\big(\{x : x \in S_1^s \oplus S_2^s \wedge y(x) = 0\}\big)$, we have $R(h_-^s) \leq R(h_+^s)$. Thus, we get $R(h_-^s) \leq R(S_1^* \cap S_2^*) + \epsilon$ with probability at least $1 - \delta$.

Case 2: If $Pr\big(\{x : x \in S_1^s \oplus S_2^s \wedge y(x) = 1\}\big) > Pr\big(\{x : x \in S_1^s \oplus S_2^s \wedge y(x) = 0\}\big)$, we have $R(h_+^s) < R(h_-^s)$. Thus, we get $R(h_+^s) \leq R(S_1^* \cap S_2^*) + \epsilon$ with probability at least $1 - \delta$.

The total number of labels to be requested is $\widetilde{O}(\log \frac{1}{\epsilon}) + \frac{2\log(\frac{8}{\delta})}{\beta^2} = \widetilde{O}(\log \frac{1}{\epsilon})$. □

Proof of Corollary 1. According to Theorem 5 we know that by requesting $\widetilde{O}(\log \frac{1}{\epsilon})$ labels the multi-view active learning in Table 1 will generate a classifier whose error rate is no larger than $R(S_1^* \cap S_2^*) + \frac{\epsilon}{2}$ with probability at least $1 - \delta$. Considering that

$$R(S_1^* \cap S_2^*) - R(S_v^*) = \int_{(S_1^* \cap S_2^*) \Delta S_v^*} |2\varphi_v(x_v) - 1| p_{x_v} d_{x_v} \leq Pr(S_1^* \oplus S_2^*),$$

we have $R(S_1^* \cap S_2^*) \leq R(S_v^*) + \frac{\epsilon}{2}$. Thus, we get that $R(S_1^* \cap S_2^*) + \frac{\epsilon}{2}$ is no larger than $R(S_v^*) + \epsilon$. □

Proof of Theorem 6. After the i-th round in Table 2, the number of training examples in $\mathcal{L}$ is $\sum_{b=0}^{i} 2^b m_i = (2^{i+1} - 1)m_i$. While in the $(i+1)$-th round, we randomly query $(2^{i+1} - 1)m_i$ labels from the region of $\overline{Q_i}$ and add them into $\mathcal{L}$. So in the $(i+1)$-th round, the number of training examples for S_v^{i+1} $(v = 1, 2)$ drawn randomly from region of $\overline{Q_i}$ is larger than the number of whole training examples for S_v^i. Since the optimal Bayes classifier c_v belongs to $\mathcal{H}_v$, according to the standard PAC-model, it is easy to know that $d(S_v^{i+1}|\overline{Q_i}, S^*|\overline{Q_i}) \leq d(S_v^i|\overline{Q_i}, S^*|\overline{Q_i})$ can be met for any φ_v, where $d(S_v|\overline{Q_i}, S^*|\overline{Q_i})$ is defined as

$$d(S_v|\overline{Q_i}, S^*|\overline{Q_i}) \triangleq R(S_v|\overline{Q_i}) - R(S^*|\overline{Q_i}) = \int_{(S_v \cap \overline{Q_i}) \Delta (S^* \cap \overline{Q_i})} |2\varphi_v(x_v) - 1| p_{x_v} d_{x_v} / Pr(\overline{Q_i}).$$

So, by setting $\varphi_v \in \{0, 1\}$, we get $d_\Delta(S_v^{i+1}|\overline{Q_i}, S^*|\overline{Q_i}) \leq d_\Delta(S_v^i|\overline{Q_i}, S^*|\overline{Q_i})$, which implies the non-degradation condition. Thus, with the proof of Theorem 1, we get Theorem 6 proved. □

Proof of Theorem 9. Similarly to the proof of Theorem 4 and Theorem 6, we know that by requesting $\widetilde{O}(\frac{1}{\epsilon})$ labels the multi-view active learning in Table 2 can get either $R(h_+^s) \leq R(S_1^* \cap S_2^*) + \epsilon$ or $R(h_-^s) \leq R(S_1^* \cap S_2^*) + \epsilon$ with probability at least $1 - \frac{\delta}{2}$. According to Lemma 2, by requesting

$\frac{2\log(\frac{8}{\delta})}{\beta^2}$ labels we can decide correctly whether $R(h^s_+)$ or $R(h^s_-)$ is smaller with probability at least $1-\frac{\delta}{2}$. Thus, we can get a classifiers whose error rate is no larger than $R(S_1^* \cap S_2^*) + \epsilon$ with probability at least $1-\delta$. The total number of labels to be requested is $\widetilde{O}(\frac{1}{\epsilon}) + \frac{2\log(\frac{8}{\delta})}{\beta^2} = \widetilde{O}(\frac{1}{\epsilon})$. □

Proof of Corollary 2. According to Theorem 9 we know that by requesting $\widetilde{O}(\frac{1}{\epsilon})$ labels the multi-view active learning in Table 2 will generate a classifier whose error rate is no larger than $R(S_1^* \cap S_2^*) + \frac{\epsilon}{2}$ with probability at least $1-\delta$. With the proof of Corollary 1, we get that $R(S_1^* \cap S_2^*) + \frac{\epsilon}{2}$ is no larger than $R(S_v^*) + \epsilon$. □