Multi-View Active Learning in the Non-Realizable Case

Wei Wang, Zhi-Hua Zhou*

National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China

Abstract

The sample complexity of active learning under the realizability assumption has been well-studied. The realizability assumption, however, rarely holds in practice. In this paper, we theoretically characterize the sample complexity of active learning in the non-realizable case under multi-view setting. We prove that, with unbounded Tsybakov noise, the sample complexity of multi-view active learning can be \(\tilde{O}(\log \frac{1}{\epsilon})\), contrasting to single-view setting where the polynomial improvement is the best possible achievement. We also prove that in general multi-view setting the sample complexity of active learning with unbounded Tsybakov noise is \(\tilde{O}(\frac{1}{\epsilon})\), where the order of \(1/\epsilon\) is independent of the parameter in Tsybakov noise, contrasting to previous polynomial bounds where the order of \(1/\epsilon\) is related to the parameter in Tsybakov noise.

Key words: active learning, non-realizable case

1. Introduction

In active learning \([10, 13, 16]\), the learner draws unlabeled data from the unknown distribution defined on the learning task and actively queries some labels from an oracle. In this way, the active learner can achieve good performance with much fewer labels than passive learning. The number of these queried labels, which is necessary and sufficient for obtaining a good learner, is well-known as the sample complexity of active learning.

Many theoretical bounds on the sample complexity of active learning have been derived based on the realizability assumption (i.e., there exists a hypothesis perfectly separating the data in

*Corresponding author. Email: zhouzh@nju.edu.cn

Preprint submitted for review

November 1, 2010
the hypothesis class). The realizability assumption, however, rarely holds in practice. Recently, the sample complexity of active learning in the non-realizable case (i.e., the data cannot be perfectly separated by any hypothesis in the hypothesis class because of the noise) has been studied. It is worth noting that these bounds obtained in the non-realizable case match the lower bound $\Omega(\frac{1}{\eta^2})$, in the same order as the upper bound $O(\frac{1}{\epsilon^2})$ of passive learning ($\eta$ denotes the generalization error rate of the optimal classifier in the hypothesis class and $\epsilon$ bounds how close to the optimal classifier in the hypothesis class the active learner has to get). This suggests that perhaps active learning in the non-realizable case is not as efficient as that in the realizable case. To improve the sample complexity of active learning in the non-realizable case remarkably, the model of the noise or some assumptions on the hypothesis class and the data distribution must be considered. Tsybakov noise model is more and more popular in theoretical analysis on the sample complexity of active learning. However, existing result shows that obtaining exponential improvement in the sample complexity of active learning with unbounded Tsybakov noise is hard.

Inspired by which proved that multi-view setting can help improve the sample complexity of active learning in the realizable case remarkably, we have an insight that multi-view setting will also help active learning in the non-realizable case. In this paper, we present the first analysis on the sample complexity of active learning in the non-realizable case under multi-view setting, where the non-realizability is caused by Tsybakov noise. Specifically:

- We define $\alpha$-expansion, which extends the definition in and to the non-realizable case, and $\beta$-condition for multi-view setting.

- We prove that the sample complexity of active learning with Tsybakov noise under multi-view setting can be improved to $\tilde{O}(\log \frac{1}{\epsilon})$ when the learner satisfies non-degradation condition. This exponential improvement holds no matter whether Tsybakov noise is bounded or not, contrasting to single-view setting where the polynomial improvement is the best possible achievement for active learning with unbounded Tsybakov noise.

- We also prove that, when non-degradation condition does not hold, the sample complexity of active learning with unbounded Tsybakov noise under multi-view setting is $\tilde{O}(\frac{1}{\epsilon^2})$, where the order

---

1The $\tilde{O}$ notation is used to hide the factor $\log \log(\frac{1}{\epsilon})$. 

of $1/\epsilon$ is independent of the parameter in Tsybakov noise, i.e., the sample complexity is always $\tilde{O}(\frac{1}{\epsilon})$ no matter how large the unbounded Tsybakov noise is. While in previous polynomial bounds, the order of $1/\epsilon$ is related to the parameter in Tsybakov noise and is larger than 1 when unbounded Tsybakov noise is larger than some degree (see Section 2). This discloses that, when non-degradation condition does not hold, multi-view setting is still able to lead to a faster convergence rate and our polynomial improvement in the sample complexity is better than previous polynomial bounds when unbounded Tsybakov noise is large.

The rest of this paper is organized as follows. After introducing related work in Section 2 and preliminaries in Section 3, we define $\alpha$-expansion in the non-realizable case in Section 4. Then we analyze the sample complexity of active learning with Tsybakov noise under multi-view setting with and without the non-degradation condition in Section 5 and Section 6, respectively, and verify the improvement in the sample complexity empirically in Section 7. Finally we conclude the paper in Section 8.

2. Related Work

Generally, the non-realizability of learning task is caused by the presence of noise. For learning the task with arbitrary forms of noise, Balcan et al. [2] proposed the agnostic active learning algorithm $A^2$ and proved that its sample complexity is $\tilde{O}(\frac{\eta^2}{\epsilon^2})$. Hoping to get tighter bound on the sample complexity of the algorithm $A^2$, Hanneke [17] defined the disagreement coefficient $\theta$, which depends on the hypothesis class and the data distribution, and proved that the sample complexity of the algorithm $A^2$ is $\tilde{O}(\theta^2 \frac{\eta^2}{\epsilon^2})$. Later, Dasgupta et al. [13] developed a general agnostic active learning algorithm which extends the scheme in [10] and proved that its sample complexity is $\tilde{O}(\theta^2 \frac{\eta^2}{\epsilon^2})$.

Recently, the popular Tsybakov noise model [21] was considered in theoretical analysis on active learning and there have been some bounds on the sample complexity. For some simple cases, where Tsybakov noise is bounded, it has been proved that the exponential improvement in the sample complexity is possible [4, 7, 18]. As for the situation where Tsybakov noise is unbounded,

\footnote{The $\tilde{O}$ notation is used to hide the factor $\text{polylog}(\frac{1}{\epsilon})$.}
only polynomial improvement in the sample complexity has been obtained. Balcan et al. [4] assumed that the samples are drawn uniformly from the unit ball in $\mathbb{R}^d$ and proved that the sample complexity of active learning with unbounded Tsybakov noise is $O(\epsilon^{-\frac{1}{1+\lambda}})$ ($\lambda > 0$ depends on Tsybakov noise). This uniform distribution assumption, however, rarely holds in practice. Castro and Nowak [8] showed that the sample complexity of active learning with unbounded Tsybakov noise is $\hat{O}(\epsilon^{-\frac{2\mu+1+a-2\mu-1}{\mu \omega}})$ ($\mu > 1$ depends on another form of Tsybakov noise, $\omega \geq 1$ depends on the Hölder smoothness and $d$ is the dimension of the data). This result is also based on the strong uniform distribution assumption. Cavallanti et al. [9] assumed that the labels of examples are generated according to a simple linear noise model and indicated that the sample complexity of active learning with unbounded Tsybakov noise is $O\left(\epsilon^{-\frac{2(3+\lambda)}{(1+\lambda)(2+\lambda)}}\right)$. Hanneke [18] proved that the algorithms or variants thereof in [2] and [13] can achieve the polynomial sample complexity $\tilde{O}(\epsilon^{-\frac{2}{1+\lambda}})$ for active learning with unbounded Tsybakov noise. For active learning with unbounded Tsybakov noise, Castro and Nowak [8] also proved that at least $\Omega(\epsilon^{-\rho})$ labels are requested to learn an $\epsilon$-approximation of the optimal classifier ($\rho \in (0, 2)$ depends on Tsybakov noise). This result shows that the polynomial improvement is the best possible achievement for active learning with unbounded Tsybakov noise in single-view setting. Wang [22] introduced smooth assumption to active learning with approximate Tsybakov noise and proved that if the classification boundary and the underlying distribution are smooth to $\xi$-th order and $\xi > d$, the sample complexity of active learning is $\hat{O}(\epsilon^{-\frac{3d}{\xi+\frac{3d}{d}}})$; if the boundary and the distribution are infinitely smooth, the sample complexity of active learning is $O(\text{polylog}(\frac{1}{\epsilon}))$. Nevertheless, this result is for approximate Tsybakov noise and the assumption on large smoothness order (or infinite smoothness order) rarely holds for data with high dimension $d$ in practice.

3. Preliminaries

In multi-view setting, the instances are described with several different disjoint sets of features. For the sake of simplicity, we only consider two-view setting in this paper. Suppose that $X = X_1 \times X_2$ is the instance space, $X_1$ and $X_2$ are the two views, $Y = \{0, 1\}$ is the label space and $\mathcal{D}$ is the distribution over $X \times Y$. Suppose that $c = (c_1, c_2)$ is the optimal Bayes classifier, where $c_1$ and $c_2$ are the optimal Bayes classifiers in the two views, respectively. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be the hypothesis class in each view and suppose that $c_1 \in \mathcal{H}_1$ and $c_2 \in \mathcal{H}_2$. For any instance
$x = (x_1, x_2)$, the hypothesis $h_v \in \mathcal{H}_v$ ($v = 1, 2$) makes that $h_v(x_v) = 1$ if $x_v \in S_v$ and $h_v(x_v) = 0$
 otherwise, where $S_v$ is a subset of $X_v$. In this way, any hypothesis $h_v \in \mathcal{H}_v$ corresponds to a subset $S_v$ of $X_v$ (as for how to combine the hypotheses in the two views, see Section 5). Considering that $x_1$ and $x_2$ denote the same instance $x$ in different views, we overload $S_v$ to denote the instance set \{x = (x_1, x_2) : x_v \in S_v\} without confusion. Let $S_v^*$ correspond to the optimal Bayes classifier $c_v$. It is well-known [15] that $S_v^* = \{x_v : \varphi_v(x_v) \geq \frac{1}{2}\}$, where $\varphi_v(x_v) = P(y = 1|x_v)$. Here, we also overload $S_v^*$ to denote the instances set \{x = (x_1, x_2) : x_v \in S_v^*\}. The error rate of a hypothesis $S_v$ under the distribution $\mathcal{D}$ is $R(h_v) = R(S_v) = Pr((x_1, x_2, y) \in \mathcal{D}(y \neq I(x_v \in S_v)))$. In general, $R(S_v^*) \neq 0$ and the excess error of $S_v$ can be denoted as follows, where $S_v \Delta S_v^* = (S_v - S_v^*) \cup (S_v^* - S_v)$ and $d(S_v, S_v^*)$ is a pseudo-distance between the sets $S_v$ and $S_v^*$.

$$R(S_v) - R(S_v^*) = \int_{S_v \Delta S_v^*} |2\varphi_v(x_v) - 1|p_{x_v} \, dx_v \triangleq d(S_v, S_v^*)$$

Let $\eta_v$ denote the error rate of the optimal Bayes classifier $c_v$ which is also called as the noise rate in the non-realizable case. In general, $\eta_v$ is less than $\frac{1}{2}$. In order to model the noise, we assume that the data distribution and the Bayes decision boundary in each view satisfies the popular Tsybakov noise condition [21] that $Pr_{x_v \in X_v}(|\varphi_v(x_v) - 1/2| \leq t) \leq C_0 t^\lambda$ for some finite $C_0 > 0$, $\lambda > 0$ and all $0 < t < 1/2$, where $\lambda = \infty$ corresponds to the best learning situation and the noise is called bounded [8]; while $\lambda = 0$ corresponds to the worst situation. When $\lambda < \infty$, the noise is called unbounded [8]. According to Proposition 1 in [21], it is easy to know that (2) holds.

$$d(S_v, S_v^*) \geq C_1d^k_\Delta(S_v, S_v^*)$$

Here $k = \frac{1+\lambda}{\lambda}$, $C_1 = 2C_0^{-1/\lambda}(\lambda + 1)^{-1-1/\lambda}$, $d_\Delta(S_v, S_v^*) = Pr(S_v - S_v^*) + Pr(S_v^* - S_v)$ is also a pseudo-distance between the sets $S_v$ and $S_v^*$, and $d(S_v, S_v^*) \leq d_\Delta(S_v, S_v^*) \leq 1$. We will use the following lemma [1] which gives the standard sample complexity for non-realizable learning task.

**Lemma 1** Suppose that $\mathcal{H}$ is a set of functions from $X$ to $Y = \{0, 1\}$ with finite VC-dimension $V \geq 1$ and $\mathcal{D}$ is the fixed but unknown distribution over $X \times Y$. For any $\epsilon, \delta > 0$, there is a positive constant $C$, such that if the size of sample $\{(x^1, y^1), \ldots, (x^N, y^N)\}$ from $\mathcal{D}$ is $N(\epsilon, \delta) = \frac{C}{\epsilon^2}(V + \log(\frac{1}{\delta}))$, then with probability at least $1 - \delta$, for all $h \in \mathcal{H}$, the following holds.

$$\left| \frac{1}{N} \sum_{i=1}^{N} I(h(x^i) \neq y^i) - E_{(x,y) \in \mathcal{D}} I(h(x) \neq y) \right| \leq \epsilon$$
4. \(\alpha\)-Expansion in the Non-realizable Case

Multi-view active learning first described in [20] focuses on the contention points (i.e., unlabeled instances on which different views predict different labels) and queries some labels of them. It is motivated by that querying the labels of contention points may help at least one of the two views to learn the optimal classifier. Let \(S_1 \oplus S_2 = (S_1 - S_2) \cup (S_2 - S_1)\) denote the contention points between \(S_1\) and \(S_2\), then \(Pr(S_1 \oplus S_2)\) denotes the probability mass on the contentions points. “\(\Delta\)” and “\(\oplus\)” mean the same operation rule. In this paper, we use “\(\Delta\)” when referring the excess error between \(S_v\) and \(S^*_v\) and use “\(\oplus\)” when referring the difference between the two views \(S_1\) and \(S_2\). In order to study multi-view active learning, the properties of contention points should be considered. One basic property is that \(Pr(S_1 \oplus S_2)\) should not be too small, otherwise the two views could be exactly the same and two-view setting would degenerate into single-view setting.

In multi-view learning, the two views represent the same learning task and generally are consistent with each other, i.e., for any instance \(x = (x_1, x_2)\) the labels of \(x\) in the two views are the same. Hence we first assume that \(S^*_1 = S^*_2 = S^*\). As for the situation where \(S^*_1 \neq S^*_2\), we will discuss on it further in Section 5.2. The instances agreed by the two views can be denoted as \((S_1 \cap S_2) \cup (\overline{S_1} \cap \overline{S_2})\). However, some of these agreed instances may be predicted different label by the optimal classifier \(S^*\), i.e., the instances in \((S_1 \cap S_2 - S^*) \cup (\overline{S_1} \cap \overline{S_2} - \overline{S^*})\). Intuitively, if the contention points can convey some information about \((S_1 \cap S_2 - S^*) \cup (\overline{S_1} \cap \overline{S_2} - \overline{S^*})\), then querying the labels of contention points could help to improve \(S_1\) and \(S_2\). Based on this intuition and that \(Pr(S_1 \oplus S_2)\) should not be too small, we give our definition on \(\alpha\)-expansion in the non-realizable case.

**Definition 1** \(\mathcal{D}\) is \(\alpha\)-expanding if for some \(\alpha > 0\) and any \(S_1 \subseteq X_1, S_2 \subseteq X_2\), (3) holds.

\[
Pr(S_1 \oplus S_2) \geq \alpha \left( Pr(S_1 \cap S_2 - S^*) + Pr(\overline{S_1} \cap \overline{S_2} - \overline{S^*}) \right)
\]

We say that \(\mathcal{D}\) is \(\alpha\)-expanding with respect to hypothesis class \(\mathcal{H}_1 \times \mathcal{H}_2\) if the above holds for all \(S_1 \in \mathcal{H}_1 \cap X_1, S_2 \in \mathcal{H}_2 \cap X_2\) (here we denote by \(\mathcal{H}_v \cap X_v\) the set \(\{h \cap X_v : h \in \mathcal{H}_v\}\) for \(v = 1, 2\)).

Balcan et al. [3] also gave a definition of expansion, \(Pr(T_1 \oplus T_2) \geq \alpha \min[Pr(T_1 \cap T_2), Pr(\overline{T_1} \cap \overline{T_2})]\), for realizable learning task under the assumptions that the learner in each view is never “confident but wrong” and the learning algorithm is able to learn from positive data only. Here \(T_v\) denotes...
Table 1: Multi-view active learning with the non-degradation condition

| Input: | Unlabeled data set $\mathcal{U} = \{x^1, x^2, \cdots, \}$ where each example $x^j$ is given as a pair $(x^j_1, x^j_2)$ |
| Process: | Query the labels of $m_0$ instances drawn randomly from $\mathcal{U}$ to compose the labeled data set $\mathcal{L}$ |
| iterate: | $i = 0, 1, \cdots, s$ |
| iterate: | Train the classifier $h^i_v$ ($v = 1, 2$) by minimizing the empirical risk with $\mathcal{L}$ in each view: $h^i_v = \arg \min_{h \in \mathcal{H}_v} \sum_{(x^1, x^2, y) \in \mathcal{L}} I(h(x^v) \neq y)$; |
| iterate: | Apply $h^i_1$ and $h^i_2$ to the unlabeled data set $\mathcal{U}$ and find out the contention point set $\mathcal{Q}_i$; |
| iterate: | Query the labels of $m_{i+1}$ instances drawn randomly from $\mathcal{Q}_i$, then add them into $\mathcal{L}$ and delete them from $\mathcal{U}$. |
| end iterate | |
| Output: | $h^*_+$ and $h^-_*$ |

the instances which are classified as positive confidently in each view. Generally, in realizable learning tasks, we aim at studying the asymptotic performance and assume that the performance of initial classifier is better than guessing randomly, i.e., $\Pr(T_v) > 1/2$. This ensures that $\Pr(T_1 \cap T_2)$ is larger than $\Pr(T_1^c \cap T_2^c)$. In addition, in [3] the instances which are agreed by the two views but are predicted different label by the optimal classifier can be denoted as $T_1 \cap T_2$. So, it can be found that Definition [1] and the definition of expansion in [3] are based on the same intuition that the amount of contention points is no less than a fraction of the amount of instances which are agreed by the two views but are predicted different label by the optimal classifiers.

5. Multi-view Active Learning with Non-degradation Condition

In this section, we first consider the multi-view learning in Table 1 and analyze whether multi-view setting can help improve the sample complexity of active learning in the non-realizable case remarkably. In multi-view setting, the classifiers are often combined to make predictions and many strategies can be used to combine them. In this paper, we consider the following two combination schemes, $h_+$ and $h_-$, for binary classification:

$$h^i_+(x) = \begin{cases} 1 & \text{if } h^i_1(x_1) = h^i_2(x_2) = 1 \\ 0 & \text{otherwise} \end{cases} \quad h^i_-(x) = \begin{cases} 0 & \text{if } h^i_1(x_1) = h^i_2(x_2) = 0 \\ 1 & \text{otherwise} \end{cases} \quad (4)$$

5.1. The Situation Where $S^*_1 = S^*_2$

With (4), the error rate of the combined classifiers $h^i_+$ and $h^i_-$ satisfy (5) and (6), respectively.
that of

\[ R(h^i_+) - R(S^*) = R(S^i_1 \cap S^i_2) - R(S^*) \leq d_\Delta(S^i_1 \cap S^i_2, S^*) \] (5)

\[ R(h^i_-) - R(S^*) = R(S^i_1 \cup S^i_2) - R(S^*) \leq d_\Delta(S^i_1 \cup S^i_2, S^*) \] (6)

Here \( S^i_v \subset X_v \) \((v = 1, 2)\) corresponds to the classifier \( h^i_v \in \mathcal{H}_v \) in the \( i \)-th round. In each round of multi-view active learning, labels of some contention points are queried to augment the training data set \( \mathcal{L} \) and the classifier in each view is then refined. As discussed in \([23]\), we also assume that the learner in Table 1 satisfies the non-degradation condition as the amount of labeled training examples increases, i.e., (7) holds, which implies that the excess error of \( S^{i+1}_v \) is no larger than that of \( S^i_v \) in the region of \( S^i_1 \oplus S^i_2 \).

\[ Pr(S^{i+1}_v \Delta S^* \mid S^i_1 \oplus S^i_2) \leq Pr(S^i_1 \Delta S^* \mid S^i_1 \oplus S^i_2) \] (7)

To illustrate the non-degradation condition, we give the following example: Suppose the data in \( X_v \) \((v = 1, 2)\) fall into \( n \) different clusters, denoted by \( \pi^v_1, \ldots, \pi^v_n \), and every cluster has the same probability mass for simplicity. The positive class is the union of some clusters while the negative class is the union of the others. Each positive (negative) cluster \( \pi^v_\xi \) in \( X_v \) is associated with only 3 positive (negative) clusters \( \pi^{3-v}_\xi \) \((\xi, \varsigma \in \{1, \ldots, n\}\) in \( X_{3-v} \) (i.e., given an instance \( x_v \) in \( \pi^v_\xi \), \( x_{3-v} \) will only be in one of these \( \pi^{3-v}_\xi \)). Suppose the learning algorithm will predict all instances in each cluster with the same label, i.e., the hypothesis class \( \mathcal{H}_v \) consists of the hypotheses which do not split any cluster. Thus, the cluster \( \pi^v_\xi \) can be classified according to the posterior probability \( P(y = 1 \mid \pi^v_\xi) \) and querying the labels of instances in cluster \( \pi^v_\xi \) will not influence the estimation of the posterior probability for cluster \( \pi^v_\varsigma \) \((\xi \neq \varsigma)\). It is evident that the non-degradation condition holds in this task. Note that the non-degradation assumption may not always hold, and we will discuss on this in Section 6. Now we give Theorem 1.

**Theorem 1** For data distribution \( \mathcal{D} \) \( \alpha \)-expanding with respect to hypothesis class \( \mathcal{H}_1 \times \mathcal{H}_2 \) according to Definition 7, when the non-degradation condition holds, if \( s = \lceil \frac{2 \log \frac{1}{\delta}}{\log \frac{C_1}{C_2}} \rceil \) and \( m_i = \frac{256kC}{C_1} \left( V + \log(\frac{16(s+1)}{\delta}) \right) \), the multi-view active learning in Table 1 will generate two classifiers \( h^*_+ \) and \( h^*_- \), at least one of which is with error rate no larger than \( R(S^*) + \epsilon \) with probability at least \( 1 - \delta \).

Here, \( V = \max[VC(\mathcal{H}_1), VC(\mathcal{H}_2)] \) where \( VC(\mathcal{H}) \) denotes the VC-dimension of the hypothesis class \( \mathcal{H} \), \( k = \frac{1+\lambda}{\lambda} \), \( C_1 = 2C_0^{-1/\lambda} \lambda^{-1} \) and \( C_2 = \frac{5+8}{6\alpha+8} \).
Proof: 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 $C_s = \frac{C_4}{Pr(Q_i)}$ and $\lambda_s = \lambda_3$, there exists some $0 < t_s \leq 1/2$ to satisfy that $Pr_{x_v \in Q_i}(|\varphi_v(x_v) - 1/2| \leq t) > C_s t_s^{\lambda_s}$. So we get

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

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 $\lambda$.

Since $m_0 = \frac{256C}{C^2_1}(V + \log(\frac{16(s+1)}{\delta}))$, according to Lemma 1 we know that $d(S_v^0, S^*) \leq \frac{C_0}{10}$ with probability at least $1 - \frac{\delta}{16(s+1)}$. With $d(S_v, S_v^*) \geq C_1 d_\Delta(S_v, S_v^*)$, we get $d_\Delta(S_v^0, S^*) \leq \frac{1}{10}$. It is easy to find that $d_\Delta(S_1^i \cap S_2^i, S^*) \leq d_\Delta(S_1^i, S^*) + d_\Delta(S_2^i, S^*) \leq 1/8$ 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 $T_{v+1}^i = S_{v+1}^i \cap \overline{Q_i}$ and $\tau_{i+1} = \frac{Pr(T_{v+1}^i \oplus T_{v+1}^i \mid S_v^*, S^*)}{Pr(T_{v+1}^i \oplus T_{v+1}^i)} - \frac{1}{2}$, it is easy to get

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

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

$$d_\Delta(S_1^{i+1} \cap S_2^{i+1} \mid \overline{Q_i}, S^* \mid \overline{Q_i})$$

$$= \frac{1}{2}(d_\Delta(S_1^{i+1} \mid \overline{Q_i}, S^* \mid \overline{Q_i}) + d_\Delta(S_2^{i+1} \mid \overline{Q_i}, S^* \mid \overline{Q_i})) + \frac{1}{2} Pr(S^* \cap (S_1^{i+1} \oplus S_2^{i+1}) \mid \overline{Q_i})$$

$$- \frac{1}{2} Pr(S^* \cap (S_1^{i+1} \oplus S_2^{i+1}) \mid \overline{Q_i})$$

$$\leq \frac{1}{2}(d_\Delta(S_1^i \mid \overline{Q_i}, S^* \mid \overline{Q_i}) + d_\Delta(S_2^i \mid \overline{Q_i}, S^* \mid \overline{Q_i}) - \tau_{i+1} Pr(S_1^{i+1} \oplus S_2^{i+1} \mid \overline{Q_i}))$$

$$= d_\Delta(S_1^i \cap S_2^i \mid \overline{Q_i}, S^* \mid \overline{Q_i}) - \tau_{i+1} Pr(S_1^{i+1} \oplus S_2^{i+1} \mid \overline{Q_i}).$$
So we have

\[ d_\Delta(S_1^{i+1} \cap S_2^{i+1}, S^*) = d_\Delta(S_1^{i+1} \cap S_2^{i+1}, Q_i, S^*|Q_i)Pr(Q_i) + d_\Delta(S_1^{i+1} \cap S_2^{i+1}, \overline{Q_i}, S^*|\overline{Q_i})Pr(\overline{Q_i}) \]
\[ \leq \frac{1}{8}Pr(Q_i) + d_\Delta(S_1^i \cap S_2^i, \overline{Q_i}, S^*|Q_i)Pr(\overline{Q_i}) - \tau_{i+1}Pr((S_1^{i+1} \oplus S_2^{i+1}) \cap \overline{Q_i}). \]

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

\[ d_\Delta(S_1^{i+1} \cap S_2^{i+1}, S^*) \leq Pr(S_1^i \cap S_2^i - S^*) + Pr(\overline{S_1^i \cap S_2^i - S^*}) + \frac{1}{8}Pr(S_1^i \oplus S_2^i) - \tau_{i+1}Pr((S_1^{i+1} \oplus S_2^{i+1}) \cap \overline{Q_i}).\]

Similarly, we get

\[ d_\Delta(S_1^{i+1} \cup S_2^{i+1}, S^*) \leq Pr(S_1^i \cap S_2^i - S^*) + Pr(\overline{S_1^i \cap S_2^i - S^*}) + \frac{1}{8}Pr(S_1^i \oplus S_2^i) + \tau_{i+1}Pr((S_1^{i+1} \oplus S_2^{i+1}) \cap \overline{Q_i}).\]

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

\[ d_\Delta(S_1^i \cap S_2^i, S^*) = d_\Delta(S_1^i \cap S_2^i|Q_i, S^*|Q_i)Pr(Q_i) + d_\Delta(S_1^i \cap S_2^i, \overline{Q_i}, S^*|Q_i)Pr(\overline{Q_i}) \]
\[ = (1/2 - \gamma_i)Pr(S_1^i \oplus S_2^i) + Pr(S_1^i \cap S_2^i - S^*) + Pr(\overline{S_1^i \cap S_2^i - S^*}) \]

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

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

\[ \frac{d_\Delta(S_1^{i+1} \cup S_2^{i+1}, S^*)}{d_\Delta(S_1^i \cup S_2^i, S^*)} \leq \frac{\frac{1}{8}Pr(S_1^i \oplus S_2^i) + |\tau_{i+1}|Pr(S_1^{i+1} \oplus S_2^{i+1}) + \frac{1}{8}Pr(S_1^i \oplus S_2^i)}{(1/2 + \gamma_i)Pr(S_1^i \oplus S_2^i) + \frac{1}{8}Pr(S_1^i \oplus S_2^i)} \]
\[ \leq \frac{(1/2 + \gamma_i)Pr(S_1^i \oplus S_2^i) + \frac{1}{8}Pr(S_1^i \oplus S_2^i)}{(1/2 + \gamma_i)Pr(S_1^i \oplus S_2^i) + \frac{1}{8}Pr(S_1^i \oplus S_2^i)} \leq \frac{5\alpha + 8}{8\alpha + 8}; \]

**Case 2:** If \(|\tau_{i+1}| > \gamma_i\), with respect to Definition [1] we have

\[ \frac{d_\Delta(S_1^{i+1} \cap S_2^{i+1}, S^*)}{d_\Delta(S_1^i \cap S_2^i, S^*)} \leq \frac{\frac{1}{8}Pr(S_1^i \oplus S_2^i) + |\tau_{i+1}|Pr(S_1^{i+1} \oplus S_2^{i+1}) + \frac{1}{8}Pr(S_1^i \oplus S_2^i)}{(1/2 + \gamma_i)Pr(S_1^i \oplus S_2^i) + \frac{1}{8}Pr(S_1^i \oplus S_2^i)} \]
\[ \leq \frac{5\alpha + 8}{8\alpha + 8}; \]
Case 3: If \( \tau_{i+1} \geq \gamma_i \) and \( 0 \leq \gamma_i \leq \frac{1}{4} \), with respect to Definition \( \| \), we have

\[
\frac{d_{\Delta}(S_i^{i+1} \cap S_2^{i+1}, S^*)}{d_{\Delta}(S_i^1 \cap S_2^2, S^*)} \leq \frac{\frac{1}{8} Pr(S_i^1 \oplus S_2^1) + \frac{1}{\alpha} Pr(S_i^1 \oplus S_2^1)}{(\frac{1}{2} - \gamma_i) Pr(S_i^1 \oplus S_2^1) + \frac{1}{\alpha} Pr(S_i^1 \oplus S_2^1)} \leq \frac{\alpha + 8}{2\alpha + 8}.
\]

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

\[
\frac{d_{\Delta}(S_i^{i+1} \cup S_2^{i+1}, S^*)}{d_{\Delta}(S_i^1 \cup S_2^2, S^*)} \leq \frac{\frac{1}{8} Pr(S_i^1 \oplus S_2^1) + \frac{1}{\alpha} Pr(S_i^1 \oplus S_2^1)}{(\frac{1}{2} + \gamma_i) Pr(S_i^1 \oplus S_2^1) + \frac{1}{\alpha} Pr(S_i^1 \oplus S_2^1)} \leq \frac{5\alpha + 8}{6\alpha + 8};
\]

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

\[
\frac{d_{\Delta}(S_i^{i+1} \cup S_2^{i+1}, S^*)}{d_{\Delta}(S_i^1 \cup S_2^2, S^*)} \leq \frac{\frac{1}{8} Pr(S_i^1 \oplus S_2^1) + \frac{1}{\alpha} Pr(S_i^1 \oplus S_2^1)}{(\frac{1}{2} + |\gamma_i|) Pr(S_i^1 \oplus S_2^1) + \frac{1}{\alpha} Pr(S_i^1 \oplus S_2^1)} \leq \frac{\alpha + 8}{2\alpha + 8};
\]

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

\[
\frac{d_{\Delta}(S_i^{i+1} \cap S_2^{i+1}, S^*)}{d_{\Delta}(S_i^1 \cap S_2^2, S^*)} \leq \frac{\frac{1}{8} Pr(S_i^1 \oplus S_2^1) + \frac{1}{\alpha} Pr(S_i^1 \oplus S_2^1)}{(\frac{1}{2} + |\gamma_i|) Pr(S_i^1 \oplus S_2^1) + \frac{1}{\alpha} Pr(S_i^1 \oplus S_2^1)} \leq \frac{5\alpha + 8}{6\alpha + 8};
\]

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

\[
\frac{d_{\Delta}(S_i^{i+1} \cup S_2^{i+1}, S^*)}{d_{\Delta}(S_i^1 \cup S_2^2, S^*)} \leq \frac{\frac{1}{8} Pr(S_i^1 \oplus S_2^1) + \frac{1}{\alpha} Pr(S_i^1 \oplus S_2^1)}{(\frac{1}{2} + |\gamma_i|) Pr(S_i^1 \oplus S_2^1) + \frac{1}{\alpha} Pr(S_i^1 \oplus S_2^1)} \leq \frac{\alpha + 8}{4\alpha + 8};
\]

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

\[
\frac{d_{\Delta}(S_i^{i+1} \cap S_2^{i+1}, S^*)}{d_{\Delta}(S_i^1 \cap S_2^2, S^*)} \leq \frac{\frac{1}{8} Pr(S_i^1 \oplus S_2^1) + \frac{1}{\alpha} Pr(S_i^1 \oplus S_2^1)}{(\frac{1}{2} + |\gamma_i|) Pr(S_i^1 \oplus S_2^1) + \frac{1}{\alpha} Pr(S_i^1 \oplus S_2^1)} \leq \frac{\alpha + 8}{4\alpha + 8}.
\]

Thus, after the \((i+1)\)-th round, either \( \frac{d_{\Delta}(S_i^{i+1} \cap S_2^{i+1}, S^*)}{d_{\Delta}(S_i^1 \cap S_2^2, S^*)} \leq \frac{5\alpha + 8}{6\alpha + 8} \) or \( \frac{d_{\Delta}(S_i^{i+1} \cup S_2^{i+1}, S^*)}{d_{\Delta}(S_i^1 \cup S_2^2, S^*)} \leq \frac{5\alpha + 8}{6\alpha + 8} \) holds.

Hence, we have \( d_{\Delta}(S_i^1 \cap S_2^2, S^*) \leq \frac{1}{8} \left( \frac{5\alpha + 8}{6\alpha + 8} \right)^{g/2} \) or \( d_{\Delta}(S_i^1 \cup S_2^2, S^*) \leq \frac{1}{8} \left( \frac{5\alpha + 8}{6\alpha + 8} \right)^{g/2} \) with probability...
at least $1 - \delta$. When $s = \left\lceil \frac{2 \log \frac{1}{\delta^2}}{\log \frac{1}{\epsilon^2}} \right\rceil$, where $C_2 = \frac{5\alpha + \epsilon}{\beta_0 + \delta}$ is a constant less than 1, we have either $d_\Delta(S^*_1 \cap S^*_2, S^*) \leq \epsilon$ or $d_\Delta(S^*_1 \cup S^*_2, S^*) \leq \epsilon$ with probability at least $1 - \delta$. Thus, considering $R(h^*_1) - R(S^*) = R(S^*_1 \cap S^*_2) - R(S^*) \leq d_\Delta(S^*_1 \cap S^*_2, S^*)$ and $R(h^*_1) - R(S^*) = R(S^*_1 \cup S^*_2) - R(S^*) \leq d_\Delta(S^*_1 \cup S^*_2, S^*)$, we have either $R(h^*_1) \leq R(S^*) + \epsilon$ or $R(h^*_1) \leq R(S^*) + \epsilon$.

From Theorem 1 we know that we only need to request $\sum_i R_i$ whose error rate is no larger than $1 - \delta$ to negative class in the contention region of (8) implies the difference between the examples belonging to positive class and that belonging to negative class at least $1 - \delta$. If we choose $h^*_1$ and it happens to satisfy $R(h^*_1) \leq R(S^*) + \epsilon$, we can get a classifier whose error rate is no larger than $R(S^*) + \epsilon$. Fortunately, there are only two classifiers and the probability of getting the right classifier is no less than $\frac{1}{2}$. To study how to choose between $h^*_1$ and $h^*_2$, we give Definition 2 at first.

**Definition 2** The multi-view classifiers $S_1$ and $S_2$ satisfy $\beta$-condition if (8) holds for some $\beta > 0$.

$$\left| \frac{Pr\{x : x \in S_1 \oplus S_2 \land y(x) = 1\}}{Pr(S_1 \oplus S_2)} - \frac{Pr\{x : x \in S_1 \oplus S_2 \land y(x) = 0\}}{Pr(S_1 \oplus S_2)} \right| \geq \beta$$

(8) implies the difference between the examples belonging to positive class and that belonging to negative class in the contention region of $S_1 \oplus S_2$. Based on Definition 2, we give Lemma 2 which provides information for deciding how to choose between $h_+$ and $h_-$. This helps to get Theorem 2.

**Lemma 2** If the multi-view classifiers $S_1^*$ and $S_2^*$ satisfy $\beta$-condition, with the number of $\frac{2 \log(\frac{4}{\beta^2})}{\beta^2}$ labels we can decide correctly whether $Pr\{x : x \in S_1^* \oplus S_2^* \land y(x) = 1\}$ or $Pr\{x : x \in S_1^* \oplus S_2^* \land y(x) = 0\}$ is smaller with probability at least $1 - \delta$.

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

$$\tilde{P}_1 \in \left[ \frac{Pr\{x : x \in S_1^* \oplus S_2^* \land y(x) = 1\}}{Pr(S_1^* \oplus S_2^*)} - \frac{\beta}{2}, \frac{Pr\{x : x \in S_1^* \oplus S_2^* \land y(x) = 1\}}{Pr(S_1^* \oplus S_2^*)} + \frac{\beta}{2} \right]$$

$$\tilde{P}_2 \in \left[ \frac{Pr\{x : x \in S_1^* \oplus S_2^* \land y(x) = 0\}}{Pr(S_1^* \oplus S_2^*)} - \frac{\beta}{2}, \frac{Pr\{x : x \in S_1^* \oplus S_2^* \land y(x) = 0\}}{Pr(S_1^* \oplus S_2^*)} + \frac{\beta}{2} \right]$$
If \( \tilde{P}_1 \leq \tilde{P}_2 \), we get \( \Pr(\{x : x \in S_1^s + S_2^s \land y(x) = 1\}) \leq \Pr(\{x : x \in S_1^s + S_2^s \land y(x) = 0\}) \) with probability at least \( 1 - \delta \); otherwise, we get \( \Pr(\{x : x \in S_1^s + S_2^s \land y(x) = 1\}) > \Pr(\{x : x \in S_1^s + S_2^s \land y(x) = 0\}) \) with probability at least \( 1 - \delta \).

**Theorem 2** For data distribution \( \mathcal{D} \) \( \alpha \)-expanding with respect to hypothesis class \( \mathcal{H}_1 \times \mathcal{H}_2 \) according to Definition 1, when the non-degradation condition holds, if the multi-view classifiers satisfy \( \beta \)-condition, 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 \( R(S^*) + \epsilon \) with probability at least \( 1 - \delta \).

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

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

**Case 2:** If \( \Pr(\{x : x \in S_1^s + S_2^s \land y(x) = 1\}) > \Pr(\{x : x \in S_1^s + S_2^s \land y(x) = 0\}) \), 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 \( \widetilde{O}(\log \frac{1}{\epsilon}) + \frac{2\log(\frac{4}{\delta})}{\beta^2} = \widetilde{O}(\log \frac{1}{\epsilon}) \).

From Theorem 2 we know that we only need to request \( \widetilde{O}(\log \frac{1}{\epsilon}) \) labels to learn a classifier with error rate no larger than \( R(S^*) + \epsilon \) with probability at least \( 1 - \delta \). Thus, we achieve an exponential improvement in sample complexity of active learning in the non-realizable case under multi-view setting. Sometimes, the difference between the examples belonging to positive class and that belonging to negative class in \( S_1^s + S_2^s \) may be very small, i.e., (9) holds.

\[
\left| \frac{\Pr(\{x : x \in S_1^s + S_2^s \land y(x) = 1\})}{\Pr(S_1^s + S_2^s)} - \frac{\Pr(\{x : x \in S_1^s + S_2^s \land y(x) = 0\})}{\Pr(S_1^s + S_2^s)} \right| = O(\epsilon) \tag{9}
\]

If so, we need not to estimate whether \( R(h_+^s) \) or \( R(h_-^s) \) is smaller and Theorem 3 indicates that both \( h_+^s \) and \( h_-^s \) are good approximations of the optimal classifier.
Theorem 3 For data distribution $\mathcal{D}$ $\alpha$-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$ according to Definition 4, when the non-degradation condition holds, if (9) is satisfied, by requesting $O(\log \frac{1}{\delta})$ labels the multi-view active learning in Table 1 will generate two classifiers $h^+_\ast$ and $h^-\ast$ which satisfy either (a) or (b) with probability at least $1 - \delta$. (a) $R(h^+_\ast) \leq R(S^\ast) + \epsilon$ and $R(h^-\ast) \leq R(S^\ast) + O(\epsilon)$; (b) $R(h^+_\ast) \leq R(S^\ast) + O(\epsilon)$ and $R(h^-\ast) \leq R(S^\ast) + \epsilon$.

Proof: Since $Pr(S^1_i \oplus S^2_i) \leq 1$, with the following equation

$$\left| \frac{Pr\{x : x \in S^1_i \oplus S^2_i \land y(x) = 1\}}{Pr(S^1_i \oplus S^2_i)} - \frac{Pr\{x : x \in S^1_i \oplus S^2_i \land y(x) = 0\}}{Pr(S^1_i \oplus S^2_i)} \right| = O(\epsilon)$$

we have $|Pr\{x : x \in S^1_i \oplus S^2_i \land y(x) = 1\} - Pr\{x : x \in S^1_i \oplus S^2_i \land y(x) = 0\}| = O(\epsilon)$. So it is easy to get $|R(h^+_\ast) - R(h^-\ast)| = O(\epsilon)$. According to Theorem 4 by requesting $O(\log \frac{1}{\delta})$ labels we can get either $R(h^+_\ast) \leq R(S^\ast) + \epsilon$ or $R(h^-\ast) \leq R(S^\ast) + \epsilon$ with probability at least $1 - \delta$. Thus, we get that $h^+_\ast$ and $h^-\ast$ satisfy either (a) or (b) with probability at least $1 - \delta$. \hfill \Box

5.2. The Situation Where $S^1_i \neq S^2_i$

Although the two views represent the same learning task and generally are consistent with each other, sometimes $S^1_i$ may be not equal to $S^2_i$. Therefore, the $\alpha$-expansion assumption in Definition 4 should be adjusted to the situation where $S^1_i \neq S^2_i$. To analyze this theoretically, we replace $S^\ast$ by $S^1_i \cap S^2_i$ in Definition 4 and get (10). Similarly to Theorem 4, we get Theorem 5

$$Pr(S_1 \oplus S_2) \geq \alpha \left( Pr(S_1 \cap S_2 - S^1_i \cap S^2_i) + Pr(S^1_i \cap S^2_i - S^1_i \cap S^2_i) \right)$$

(10)

Theorem 4 For data distribution $\mathcal{D}$ $\alpha$-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$ according to (10), when the non-degradation condition holds, if $s = \left\lceil \frac{2 \log \frac{k}{\delta}}{\log \frac{1}{\epsilon}} \right\rceil$ and $m_i = \frac{256kC}{C^2}(V + \log \left(\frac{16(\alpha+1)}{\delta}\right))$, the multi-view active learning in Table 1 will generate two classifiers $h^+_\ast$ and $h^-\ast$, at least one of which is with error rate no larger than $R(S^1_i \cap S^2_i) + \epsilon$ with probability at least $1 - \delta$. ($V$, $k$, $C_1$ and $C_2$ are given in Theorem 4.)

Proof: Since $S^1_i$ is the optimal Bayes classifier in the $v$-th view, obviously, $R(S^1_i \cap S^2_i)$ is no less than $R(S^v_i)$, ($v = 1, 2$). So, learning a classifier with error rate no larger than $R(S^1_i \cap S^2_i) + \epsilon$ is not harder than learning a classifier with error rate no larger than $R(S^v_i) + \epsilon$. Now we aim at
learning a classifier with error rate no larger than \( R(S_1^* \cap S_2^*) + \epsilon \). Without loss of generality, we assume \( R(S_i^*) > R(S_i^* \cap S_2^*) \) for \( i = 0, 1, \ldots, s \). If \( R(S_i^*) \leq R(S_1^* \cap S_2^*) \), we get a classifier with error rate no larger than \( R(S_1^* \cap S_2^*) + \epsilon \). Thus, we can neglect the probability mass on the hypothesis whose error rate is less than \( R(S_1^* \cap S_2^*) \) and regard \( S_1^* \cap S_2^* \) as the optimal. Replacing \( S^* \) by \( S_1^* \cap S_2^* \) in the discussion of Section 5.1, with the proof of Theorem 1 we get Theorem 4 proved.

Theorem 5 shows that for the situation where \( S_1^* \neq S_2^* \), by requesting \( \tilde{O}(\log \frac{1}{\epsilon}) \) labels we can learn two classifiers \( h_+^* \) and \( h_-^* \), at least one of which is with error rate no larger than \( R(S_1^* \cap S_2^*) + \epsilon \) with probability at least \( 1 - \delta \). With Lemma 2 we get Theorem 5 from Theorem 4.

**Theorem 5** For data distribution \( D \) \( \alpha \)-expanding with respect to hypothesis class \( \mathcal{H}_1 \times \mathcal{H}_2 \) according to (10), when the non-degradation condition holds, if the multi-view classifiers satisfy \( \beta \)-condition, by requesting \( \tilde{O}(\log \frac{1}{\epsilon}) \) labels the multi-view active learning in Table 1 will generate a classifier whose error rate is no larger than \( R(S_1^* \cap S_2^*) + \epsilon \) with probability at least \( 1 - \delta \).

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

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

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

The total number of labels to be requested is \( \tilde{O}(\log \frac{1}{\epsilon}) + \frac{2\log \left( \frac{2}{\epsilon} \right)}{\beta^2} = \tilde{O}(\log \frac{1}{\epsilon}). \)

Generally, \( R(S_1^* \cap S_2^*) \) is larger than \( R(S_1^*) \) and \( R(S_2^*) \). When \( S_1^* \) is not too much different from \( S_2^* \), i.e., \( Pr(S_1^* \oplus S_2^*) \leq 2/3 \), we have Corollary 1 which indicates that the exponential improvement in the sample complexity of active learning with Tsybakov noise is still possible.
Corollary 1 For data distribution $D \alpha$-expanding with respect to hypothesis class $H_1 \times H_2$ according to \cite{10}, when the non-degradation condition holds, if the multi-view classifiers satisfy $\beta$-condition and $\Pr(S^*_1 \oplus S^*_2) \leq \epsilon/2$, by requesting $\tilde{O}(\log \frac{1}{\delta})$ labels the multi-view active learning in Table 1 will generate a classifier with error rate no larger than $R(S^*_v) + \epsilon (v = 1, 2)$ with probability at least $1 - \delta$.

**Proof:** According to Theorem 5 we know that by requesting $\tilde{O}(\log \frac{1}{\delta})$ 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

$$R(S^*_1 \cap S^*_2) - R(S^*_v) = \int_{(S^*_1 \cap S^*_2) \Delta S^*_v} |2\varphi_v(x_v) - 1| p_{x_v} dx_v \leq \Pr(S^*_1 \oplus S^*_2),$$

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 $R(S^*_v) + \epsilon$.

$\square$

6. Multi-view Active Learning without Non-degradation Condition

Section 5 considers situations when the non-degradation condition holds, there are cases, however, the non-degradation condition \cite{17} does not hold. In this section we focus on the multi-view active learning in Table 2 and give an analysis with the non-degradation condition waived. Firstly, we give Theorem 6 for the sample complexity of multi-view active learning in Table 2 when $S^*_1 = S^*_2 = S^*$.

**Theorem 6** For data distribution $D \alpha$-expanding with respect to hypothesis class $H_1 \times H_2$ according to Definition \cite{11}, if $s = \left[ \frac{2 \log \frac{1}{\delta}}{\log \frac{1}{2}} \right]$ and $m_i = \frac{256kC_1}{C_2}(V + \log(\frac{16(s+1)}{\delta}))$, the multi-view active learning in Table 2 will generate two classifiers $h^*_+$ and $h^*_-$, at least one of which is with error rate no larger than $R(S^*) + \epsilon$ with probability at least $1 - \delta$. ($V$, $k$, $C_1$ and $C_2$ are given in Theorem \cite{17})

**Proof:** After the $i$-th round in Table 2, the number of training examples in $L$ 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 $Q_i$ and add them into $L$. So in the $(i + 1)$-th round, the number of training examples
Proof: According to Theorem 6, by requesting $d$ labels in the $R_i$ than the multi-view active learning in Table 2 will generate a classifier whose error rate is no larger than $\alpha$ with probability at least $1 - \delta$. Thus, with the proof of Theorem 1, we get Theorem 6 proved.

So, by setting $\phi = \epsilon$, we get $d(S^i + 1 | \overline{Q_i}, S^* | \overline{Q_i}) \leq d(S^i | \overline{Q_i}, S^* | \overline{Q_i})$ can be met for any $\varphi_v$, where $d(S_v | \overline{Q_i}, S^* | \overline{Q_i})$ is defined as

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

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

Theorem 6 shows that we can request $\sum_{i=0}^{n} 2^i m_i = O(\frac{1}{\epsilon})$ labels to learn two classifiers $h_+^*$ and $h_*^*$, at least one of which is with error rate no larger than $R(S^*) + \epsilon$ with probability at least $1 - \delta$. To guarantee the non-degradation condition, we only need to query $(2^i - 1)m_i$ more labels in the $i$-th round. With Lemma 2, we get Theorem 7.

**Theorem 7** For data distribution $D$ $\alpha$-expanding with respect to hypothesis class $H_1 \times H_2$ according to Definition 4, if the multi-view classifiers satisfy $\beta$-condition, by requesting $O(\frac{1}{\epsilon})$ labels the multi-view active learning in Table 2 will generate a classifier whose error rate is no larger than $R(S^*) + \epsilon$ with probability at least $1 - \delta$.

**Proof:** According to Theorem 6 by requesting $O(\frac{1}{\epsilon})$ labels the multi-view active learning in Table 2 will generate two classifiers $h_+^*$ and $h_*^*$, at least one of which is with error rate no larger
than $R(S^*) + \epsilon$ with probability at least $1 - \delta$. Similarly to the proof of Theorem 2, we get Theorem 7 proved.

Theorem 7 shows that, without the non-degradation condition, we need to request $\tilde{O}(1/\epsilon)$ labels to learn a classifier with error rate no larger than $R(S^*) + \epsilon$ with probability at least $1 - \delta$. The order of $1/\epsilon$ is independent of the parameter in Tsybakov noise. Similarly to Theorem 3, we get Theorem 8 which indicates that both $h^+_s$ and $h^-_s$ are good approximations of the optimal classifier.

**Theorem 8** For data distribution $D$ $\alpha$-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$ according to Definition 1 if (9) holds, by requesting $\tilde{O}(1/\epsilon)$ labels the multi-view active learning in Table 2 will generate two classifiers $h^+_s$ and $h^-_s$ which satisfy either (a) or (b) with probability at least $1 - \delta$. (a) $R(h^+_s) \leq R(S^*) + \epsilon$ and $R(h^-_s) \leq R(S^*) + O(\epsilon)$; (b) $R(h^+_s) \leq R(S^*) + O(\epsilon)$ and $R(h^-_s) \leq R(S^*) + \epsilon$.

**Proof:** According to Theorem 6 by requesting $\tilde{O}(1/\epsilon)$ labels the multi-view active learning in Table 2 will generate two classifiers $h^+_s$ and $h^-_s$, at least one of which is with error rate no larger than $R(S^*) + \epsilon$ with probability at least $1 - \delta$. Similarly to the proof of Theorem 3 we get Theorem 8 proved.

As for the situation where $S^*_1 \neq S^*_2$, similarly to Theorem 5 and Corollary 1 we have Theorem 9 and Corollary 2.

**Theorem 9** For data distribution $D$ $\alpha$-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$ according to (10), if the multi-view classifiers satisfy $\beta$-condition, by requesting $\tilde{O}(1/\epsilon)$ labels the multi-view active learning in Table 2 will generate a classifier whose error rate is no larger than $R(S^*_1 \cap S^*_2) + \epsilon$ with probability at least $1 - \delta$.

**Proof:** Similarly to the proof of Theorem 4 and Theorem 6 we know that by requesting $\tilde{O}(1/\epsilon)$ 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 $1 - \delta$. According to Lemma 2 by requesting $\tilde{O}(1/\epsilon)$ 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 classifier 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 $\tilde{O}\left(\frac{1}{\epsilon}\right) + \frac{2\log\left(\frac{2}{\delta}\right)}{\beta^2} = \tilde{O}\left(\frac{1}{\epsilon}\right)$.

\section*{Corollary 2}
For data distribution $\mathcal{D}$ $\alpha$-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$ according to (10), if the multi-view classifiers satisfy $\beta$-condition and $\Pr(S_1^* \oplus S_2^*) \leq \frac{\epsilon}{2}$, by requesting $\tilde{O}\left(\frac{1}{\epsilon}\right)$ labels the multi-view active learning in Table 2 will generate a classifier with error rate no larger than $R(S_v^*) + \epsilon$ ($v = 1, 2$) with probability at least $1 - \delta$.

\textbf{Proof:} According to Theorem 9 we know that by requesting $\tilde{O}\left(\frac{1}{\epsilon}\right)$ 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 $R(S_v^*) + \epsilon$.

\section*{7. Empirical Verification}

In this section we empirically verify that whether multi-view setting can improve the sample complexity of active learning in the non-realizable case remarkably.

In the experiment we use the semi-artificial data set \cite{20} and the course data set \cite{6}. The semi-artificial data set has two artificial views which are created by randomly pairing two examples from the same class and contains 800 examples. In order to control the correlation between the two views, the number of clusters per class can be set as a parameter. We use 1 cluster, 2 clusters and 4 clusters in the experiments, respectively. The course data set has two natural views: pages view (i.e., the text appearing on the page) and links view (i.e., the anchor text attached to hyper-links pointing to the page) and contains 1,051 examples. We randomly use 25% data as the test set and use the remaining 75% data to generate the unlabeled data set $\mathcal{U}$. We use Random Sampling as the baseline. In each round, we fix the number of examples to be queried in Multi-View Active Learning and that in Random Sampling. Thus, we can study their performances under the same number of queried examples. In the experiments, we query two
Figure 1: Multi-view setting improves the sample complexity of active learning in the non-realizable case remarkably.

examples in each round of the two methods and implement the classifiers with NaiveBayes in WEKA. The experiments are repeated for 20 runs and Figure 1 plots the average error rates of the two methods against the number of examples that have been queried. From Figure 1 it can be found that the performance of Multi-View Active Learning is far better than the performance of Random Sampling with the same number of queried examples. In other words, multi-view setting can help improve the sample complexity of active learning in the non-realizable case remarkably.

8. Conclusion

We present the first study on active learning in the non-realizable case under multi-view setting in this paper. We prove that the sample complexity of multi-view active learning with unbounded Tsybakov noise can be improved to $\tilde{O}(\log \frac{1}{\epsilon})$, contrasting to single-view setting where only polynomial improvement is proved possible with the same noise condition. In general multi-view
setting, we prove that the sample complexity of active learning with unbounded Tsybakov noise is $\tilde{O}(\frac{1}{\epsilon})$, where the order of $1/\epsilon$ is independent of the parameter in Tsybakov noise, contrasting to previous polynomial bounds where the order of $1/\epsilon$ is related to the parameter in Tsybakov noise. Generally, the non-realizability of learning task can be caused by many kinds of noise, e.g., misclassification noise and malicious noise. It would be interesting to extend our work to more general noise model.

References

[1] M. Anthony and P. L. Bartlett, editors. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, UK, 1999.

[2] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *ICML*, pages 65–72, 2006.

[3] M.-F. Balcan, A. Blum, and K. Yang. Co-training and expansion: Towards bridging theory and practice. In *NIPS 17*, pages 89–96. 2005.

[4] M.-F. Balcan, A. Z. Broder, and T. Zhang. Margin based active learning. In *COLT*, pages 35–50, 2007.

[5] M.-F. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. In *COLT*, pages 45–56, 2008.

[6] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, pages 92–100, 1998.

[7] R. M. Castro and R. D. Nowak. Upper and lower error bounds for active learning. In *Allerton Conference*, pages 225–234, 2006.

[8] R. M. Castro and R. D. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

[9] G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Linear classification and selective sampling under low noise conditions. In *NIPS 21*, pages 249–256. 2009.
[10] D. A. Cohn, L. E. Atlas, and R. E. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[11] S. Dasgupta. Analysis of a greedy active learning strategy. In *NIPS 17*, pages 337–344. 2005.

[12] S. Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS 18*, pages 235–242. 2006.

[13] S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *NIPS 20*, pages 353–360. 2008.

[14] S. Dasgupta, A. T. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *COLT*, pages 249–263, 2005.

[15] L. Devroye, L. Györfi, and G. Lugosi, editors. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

[16] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[17] S. Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, pages 353–360, 2007.

[18] S. Hanneke. Adaptive rates of convergence in active learning. In *COLT*, 2009.

[19] M. Kääriäinen. Active learning in the non-realizable case. In *ACL*, pages 63–77, 2006.

[20] I. Muslea, S. Minton, and C. A. Knoblock. Active + semi-supervised learning = robust multi-view learning. In *ICML*, pages 435–442, 2002.

[21] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[22] L. Wang. Sufficient conditions for agnostic active learnable. In *NIPS 22*, pages 1999–2007. 2009.

[23] W. Wang and Z.-H. Zhou. On multi-view active learning and the combination with semi-supervised learning. In *ICML*, pages 1152–1159, 2008.