Multi-View Active Learning in the Non-Realizable Case

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

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 $\widetilde{O}(\log \frac{1}{\epsilon})$, 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 $\widetilde{O}(\frac{1}{\epsilon})$, 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.

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 leaner, 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 the hypothesis class) [4, 5, 11, 12, 14, 16]. 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 [2, 13, 17]. It is worth noting that these bounds obtained in the non-realizable case match the lower bound $\Omega(\frac{\eta^2}{\epsilon^2})$ [19], in the same order as the upper bound $O(\frac{1}{\epsilon^2})$ of passive learning (η denotes the generalization error rate of the optimal classifier in the hypothesis class and ϵ 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 [21] is more and more popular in theoretical analysis on the sample complexity of active learning. However, existing result [8] shows that obtaining *exponential* improvement in the sample complexity of active learning with unbounded Tsybakov noise is hard.

Inspired by [23] which proved that *multi-view* setting [6] 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 α-*expansion*, which extends the definition in [3] and [23] to the non-realizable case, and β-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 $\widetilde{O}(\log \frac{1}{\epsilon})$ when the learner satisfies non-degradation condition.[1] 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 $\widetilde{O}(\frac{1}{\epsilon})$, where the order of $1/\epsilon$ is independent of the parameter in Tsybakov noise, i.e., the sample complexity is always $\widetilde{O}(\frac{1}{\epsilon})$ 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 α-expansion in the non-realizable case in Section 4. 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. Finally we conclude the paper in Section 7.

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 $\widehat{O}(\frac{\eta^2}{\epsilon^2})$.[2] Hoping to get tighter bound on the sample complexity of the algorithm A^2, Hanneke [17] defined the *disagreement coefficient* θ, which depends on the hypothesis class and the data distribution, and proved that the sample complexity of the algorithm A^2 is $\widehat{O}(\theta^2 \frac{\eta^2}{\epsilon^2})$. 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 $\widehat{O}(\theta \frac{\eta^2}{\epsilon^2})$.

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, only *polynomial* improvement in the sample complexity has been obtained. Balcan et al. [4] assumed that the samples are drawn uniformly from the the unit ball in R^d and proved that the sample complexity of active learning with unbounded Tsybakov noise is $O\big(\epsilon^{-\frac{2}{1+\lambda}}\big)$ ($\lambda > 0$ 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 $\widehat{O}\big(\epsilon^{-\frac{2\mu\omega+d-2\omega-1}{\mu\omega}}\big)$ ($\mu > 1$ 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 $O\big(\epsilon^{-\frac{2(3+\lambda)}{(1+\lambda)(2+\lambda)}}\big)$. Hanneke [18] proved that the algorithms or variants thereof in [2] and [13] can achieve the *polynomial* sample complexity $\widehat{O}\big(\epsilon^{-\frac{2}{1+\lambda}}\big)$ 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

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

[2] The $\widehat{O}$ notation is used to hide the factor $polylog(\frac{1}{\epsilon})$.

an ϵ-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 ξ-th order and $\xi > d$, the sample complexity of active learning is $\widehat{O}\big(\epsilon^{-\frac{2d}{\xi+d}}\big)$; if the boundary and the distribution are infinitely smooth, the sample complexity of active learning is $O\big(polylog(\frac{1}{\epsilon})\big)$. 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 $R(h_v) = R(S_v) = Pr_{(x_1,x_2,y)\in\mathcal{D}}\big(y \neq \mathbf{I}(x_v \in S_v)\big)$. 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^*.

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

Let η_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, η_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 $0 < t \leq 1/2$, 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.

$$d(S_v, S_v^*) \geq C_1 d_\Delta^k(S_v, S_v^*) \tag{2}$$

Here $k = \frac{1+\lambda}{\lambda}$, $C_1 = 2C_0^{-1/\lambda}\lambda(\lambda+1)^{-1-1/\lambda}$, $d_\Delta(S_v, S_v^*) = Pr(S_v - S_v^*) + Pr(S_v^* - S_v)$ 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 lamma [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 ϵ, $\delta > 0$, 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 $N(\epsilon, \delta) = \frac{C}{\epsilon^2}\big(V + \log(\frac{1}{\delta})\big)$, then with probability at least $1 - \delta$, for all $h \in \mathcal{H}$, the following holds.*

$$|\frac{1}{N}\sum\nolimits_{i=1}^{N} \boldsymbol{I}\big(h(x^i) \neq y^i\big) - \boldsymbol{E}_{(x,y)\in\mathcal{D}}\boldsymbol{I}\big(h(x) \neq y\big)| \leq \epsilon$$

4 α-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

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_1^j, x_2^j)
Process:
Query the labels of m_0 instances drawn randomly from $\mathcal{U}$ to compose the labeled data set $\mathcal{L}$
iterate: $i = 0, 1, \cdots, s$
Train the classifier h_v^i $(v = 1, 2)$ by minimizing the empirical risk with $\mathcal{L}$ in each view:
$h_v^i = \arg\min_{h \in \mathcal{H}_v} \sum_{(x_1, x_2, y) \in \mathcal{L}} \mathbf{I}(h(x_v) \neq y)$;
Apply h_1^i and h_2^i to the unlabeled data set $\mathcal{U}$ and find out the contention point set $\mathcal{Q}_i$;
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_+^s and h_-^s

between S_1 and S_2, then $Pr(S_1 \oplus S_2)$ denotes the probability mass on the contentions points. "Δ" and "$\oplus$" mean the same operation rule. In this paper, we use "Δ" 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 α-*expansion* in the non-realizable case.

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

$$Pr\big(S_1 \oplus S_2\big) \geq \alpha\Big(Pr\big(S_1 \cap S_2 - S^*\big) + Pr\big(\overline{S_1} \cap \overline{S_2} - \overline{S^*}\big)\Big) \tag{3}$$

We say that $\mathcal{D}$ is α-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, $Pr(T_1 \oplus T_2) \geq \alpha \min\big[Pr(T_1 \cap T_2), Pr(\overline{T_1} \cap \overline{T_2})\big]$, 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 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 $Pr(\overline{T_1} \cap \overline{T_2})$. 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 $\overline{T_1} \cap \overline{T_2}$. 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:

$$h_+^i(x) = \begin{cases} 1 & \text{if } h_1^i(x_1) = h_2^i(x_2) = 1 \\ 0 & \text{otherwise} \end{cases} \qquad h_-^i(x) = \begin{cases} 0 & \text{if } h_1^i(x_1) = h_2^i(x_2) = 0 \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

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.

$$R(h_+^i) - R(S^*) = R(S_1^i \cap S_2^i) - R(S^*) \leq d_\Delta(S_1^i \cap S_2^i, S^*) \tag{5}$$

$$R(h_-^i) - R(S^*) = R(S_1^i \cup S_2^i) - R(S^*) \leq d_\Delta(S_1^i \cup S_2^i, S^*) \tag{6}$$

Here $S_v^i \subset X_v$ $(v = 1, 2)$ corresponds to the classifier $h_v^i \in \mathcal{H}_v$ 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_v^{i+1} is no larger than that of S_v^i in the region of $\overline{S_1^i \oplus S_2^i}$.

$$Pr\big(S_v^{i+1} \Delta S^* \big| \overline{S_1^i \oplus S_2^i}\big) \leq Pr(S_v^i \Delta S^* \big| \overline{S_1^i \oplus S_2^i}) \tag{7}$$

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 $\pi_1^v, \ldots, \pi_n^v$, 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 π_ξ^v in X_v is associated with only 3 positive (negative) clusters π_ς^{3-v} $(\xi, \varsigma \in \{1, \ldots, n\})$ in X_{3-v} (i.e., given an instance x_v in π_ξ^v, x_{3-v} will only be in one of these π_ς^{3-v}). 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 π_ξ^v can be classified according to the posterior probability $P(y = 1|\pi_\xi^v)$ and querying the labels of instances in cluster π_ξ^v will not influence the estimation of the posterior probability for cluster π_ς^v $(\varsigma \neq \xi)$. 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 $\mathcal{D}$ α-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$ according to Definition 1, when the non-degradation condition holds, if $s = \lceil \frac{2 \log \frac{1}{8\epsilon}}{\log \frac{1}{C_2}} \rceil$ and $m_i = \frac{256^k C}{C_1^2}\big(V + \log(\frac{16(s+1)}{\delta})\big)$, the multi-view active learning in Table 1 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 $1 - \delta$.*
Here, $V = \max[VC(\mathcal{H}_1), VC(\mathcal{H}_2)]$ where $VC(\mathcal{H})$ denotes the VC-dimension of the hypothesis class $\mathcal{H}$, $k = \frac{1+\lambda}{\lambda}$, $C_1 = 2C_0^{-1/\lambda}\lambda(\lambda+1)^{-1-1/\lambda}$ and $C_2 = \frac{5\alpha+8}{6\alpha+8}$.

Proof sketch. Let $Q_i = S_1^i \oplus S_2^i$, first with Lemma 1 and (2) we have $d_\Delta(S_1^{i+1} \cap S_2^{i+1} \mid Q_i, S^* \mid Q_i) \leq 1/8$. Let $T_v^{i+1} = S_v^{i+1} \cap \overline{Q_i}$ and $\tau_{i+1} = \frac{Pr(T_1^{i+1} \oplus T_2^{i+1} - S^*)}{Pr(T_1^{i+1} \oplus T_2^{i+1})} - \frac{1}{2}$. Considering (7) and $d_\Delta(S_1^i \cap S_2^i | \overline{Q_i}, S^* | \overline{Q_i}) Pr(\overline{Q_i}) = Pr(S_1^i \cap S_2^i - S^*) + Pr(\overline{S_1^i} \cap \overline{S_2^i} - \overline{S^*})$, then we calculate that

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

As in each round some contention points 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 $Pr(S_1^i \oplus S_2^i)$. Let $\gamma_i = \frac{Pr(S_1^i \oplus S_2^i - S^*)}{Pr(S_1^i \oplus S_2^i)} - \frac{1}{2}$, with Definition 1 and different combinations of τ_{i+1} and γ_i, we can have either $\frac{d_\Delta(S_1^{i+1} \cap S_2^{i+1}, S^*)}{d_\Delta(S_1^i \cap S_2^i, S^*)} \leq \frac{5\alpha+8}{6\alpha+8}$ or $\frac{d_\Delta(S_1^{i+1} \cup S_2^{i+1}, S^*)}{d_\Delta(S_1^i \cup S_2^i, S^*)} \leq \frac{5\alpha+8}{6\alpha+8}$. 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 $d_\Delta(S_1^s \cap S_2^s, S^*) \leq \epsilon$ or $d_\Delta(S_1^s \cup S_2^s, S^*) \leq \epsilon$. Thus, with (5) and (6) we have either $R(h_+^s) \leq R(S^*) + \epsilon$ or $R(h_-^s) \leq R(S^*) + \epsilon$. □

From Theorem 1 we know that we only need to request $\sum_{i=0}^{s} m_i = \widetilde{O}(\log \frac{1}{\epsilon})$ labels to learn h_+^s and h_-^s, at least one of which is with error rate no larger than $R(S^*) + \epsilon$ with probability at least $1 - \delta$. If we choose h_+^s and it happens to satisfy $R(h_+^s) \leq R(S^*) + \epsilon$, 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_+^s and h_-^s, we give Definition 2 at first.

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

$$\Big| \frac{Pr\big(\{x : x \in S_1 \oplus S_2 \wedge y(x) = 1\}\big)}{Pr(S_1 \oplus S_2)} - \frac{Pr\big(\{x : x \in S_1 \oplus S_2 \wedge y(x) = 0\}\big)}{Pr(S_1 \oplus S_2)} \Big| \geq \beta \tag{8}$$

(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^s and S_2^s satisfy β-condition, with the number of $\frac{2\log(\frac{4}{\delta})}{\beta^2}$ 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 $1 - \delta$.*

Theorem 2 *For data distribution $\mathcal{D}$ α-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 β-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 $1 - \delta$.*

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 $S_1^s \oplus S_2^s$ may be very small, i.e., (9) holds.

$$\Big| \frac{Pr\big(\{x : x \in S_1^s \oplus S_2^s \wedge y(x) = 1\}\big)}{Pr(S_1^s \oplus S_2^s)} - \frac{Pr\big(\{x : x \in S_1^s \oplus S_2^s \wedge y(x) = 0\}\big)}{Pr(S_1^s \oplus S_2^s)} \Big| = O(\epsilon) \tag{9}$$

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 $\mathcal{D}$ α-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$ according to Definition 1, when the non-degradation condition holds, if (9) is satisfied, by requesting $\widetilde{O}(\log \frac{1}{\epsilon})$ labels the multi-view active learning in Table 1 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 $R(h_-^s) \leq R(S^*) + \epsilon$.*

The complete proof of Theorem 1, and the proofs of Lemma 2, Theorem 2 and Theorem 3 are given in the supplementary file.

5.2 The Situation Where $S_1^* \neq S_2^*$

Although the two views represent the same learning task and generally are consistent with each other, sometimes S_1^* may be not equal to S_2^*. Therefore, the α-*expansion* assumption in Definition 1 should be adjusted to the situation where $S_1^* \neq S_2^*$. To analyze this theoretically, we replace S^* by $S_1^* \cap S_2^*$ in Definition 1 and get (10). Similarly to Theorem 1, we get Theorem 4.

$$Pr\big(S_1 \oplus S_2\big) \geq \alpha\Big(Pr\big(S_1 \cap S_2 - S_1^* \cap S_2^*\big) + Pr\big(\overline{S_1} \cap \overline{S_2} - \overline{S_1^* \cap S_2^*}\big)\Big) \tag{10}$$

Theorem 4 *For data distribution $\mathcal{D}$ α-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$ according to (10), when the non-degradation condition holds, if $s = \lceil \frac{2\log\frac{1}{8\epsilon}}{\log\frac{1}{C_2}} \rceil$ and $m_i = \frac{256^k C}{C_1^2}\big(V + \log(\frac{16(s+1)}{\delta})\big)$, the multi-view active learning in Table 1 will generate two classifiers h_+^s and h_-^s, at least one of which is with error rate no larger than $R(S_1^* \cap S_2^*) + \epsilon$ with probability at least $1 - \delta$. (V, k, C_1 and C_2 are given in Theorem 1.)*

Table 2: Multi-view active learning without 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_1^j, x_2^j)
Process:
Query the labels of m_0 instances drawn randomly from $\mathcal{U}$ to compose the labeled data set $\mathcal{L}$;
Train the classifier h_v^0 $(v = 1, 2)$ by minimizing the empirical risk with $\mathcal{L}$ in each view:
$h_v^0 = \arg\min_{h \in \mathcal{H}_v} \sum_{(x_1, x_2, y) \in \mathcal{L}} \mathbf{I}(h(x_v) \neq y)$;
iterate: $i = 1, \cdots, s$
Apply h_1^{i-1} and h_2^{i-1} to the unlabeled data set $\mathcal{U}$ and find out the contention point set $\mathcal{Q}_i$;
Query the labels of m_i instances drawn randomly from $\mathcal{Q}_i$, then add them into $\mathcal{L}$ and delete them from $\mathcal{U}$;
Query the labels of $(2^i - 1)m_i$ instances drawn randomly from $\mathcal{U} - \mathcal{Q}_i$, then add them into $\mathcal{L}$ and delete them from $\mathcal{U}$;
Train the classifier h_v^i by minimizing the empirical risk with $\mathcal{L}$ in each view:
$h_v^i = \arg\min_{h \in \mathcal{H}_v} \sum_{(x_1, x_2, y) \in \mathcal{L}} \mathbf{I}(h(x_v) \neq y)$.
end iterate
Output: h_+^s and h_-^s

Proof. Since S_v^* is the optimal Bayes classifier in the v-th view, obviously, $R(S_1^* \cap S_2^*)$ is no less than $R(S_v^*)$, $(v = 1, 2)$. So, learning a classifier with error rate no larger than $R(S_1^* \cap S_2^*) + \epsilon$ is not harder than learning a classifier with error rate no larger than $R(S_v^*) + \epsilon$. 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 $R(S_v^i) > R(S_1^* \cap S_2^*)$ for $i = 0, 1, \ldots, s$. If $R(S_v^i) \leq R(S_1^* \cap S_2^*)$, 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 4 shows that for the situation where $S_1^* \neq S_2^*$, by requesting $\widetilde{O}(\log \frac{1}{\epsilon})$ labels we can learn two classifiers h_+^s and h_-^s, 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 $\mathcal{D}$ α-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 β-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 $R(S_1^* \cap S_2^*) + \epsilon$ with probability at least $1 - \delta$.*

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., $Pr(S_1^* \oplus S_2^*) \leq \epsilon/2$, 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 $\mathcal{D}$ α-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 β-condition and $Pr(S_1^* \oplus S_2^*) \leq \epsilon/2$, by requesting $\widetilde{O}(\log \frac{1}{\epsilon})$ 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 $1 - \delta$.*

The proofs of Theorem 5 and Corollary 1 are given in the supplemental file.

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 (7) 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 $\mathcal{D}$ α-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$ according to Definition 1, if $s = \lceil \frac{2 \log \frac{1}{8\epsilon}}{\log \frac{1}{C_2}} \rceil$ and $m_i = \frac{256^k C}{C_1^2} \big(V + \log(\frac{16(s+1)}{\delta})\big)$, 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 $1 - \delta$. (V, k, C_1 and C_2 are given in Theorem 1.)*

Proof sketch. In the $(i+1)$-th round, we randomly query $(2^{i+1}-1)m_i$ labels from $\overline{Q_i}$ and add them into $\mathcal{L}$. So the number of training examples for S_v^{i+1} $(v=1,2)$ is larger than the number of whole training examples for S_v^i. Thus we know that $d(S_v^{i+1}|\overline{Q_i}, S^*|\overline{Q_i}) \leq d(S_v^i|\overline{Q_i}, S^*|\overline{Q_i})$ holds for any φ_v. Setting $\varphi_v \in \{0,1\}$, the non-degradation condition (7) stands. Thus, with the proof of Theorem 1 we get Theorem 6 proved. □

Theorem 6 shows that we can request $\sum_{i=0}^{s} 2^i m_i = \widetilde{O}(\frac{1}{\epsilon})$ labels to learn 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 $1-\delta$. To guarantee the non-degradation condition (7), 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 $\mathcal{D}$ α-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$ according to Definition 1, if the multi-view classifiers satisfy β-condition, 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 $R(S^*)+\epsilon$ with probability at least $1-\delta$.*

Theorem 7 shows that, without the non-degradation condition, we need to request $\widetilde{O}(\frac{1}{\epsilon})$ 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 $\mathcal{D}$ α-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$ according to Definition 1, if (9) holds, by requesting $\widetilde{O}(\frac{1}{\epsilon})$ 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 $R(h_-^s) \leq R(S^*)+\epsilon$.*

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 $\mathcal{D}$ α-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$ according to (10), if the multi-view classifiers satisfy β-condition, 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 $R(S_1^* \cap S_2^*)+\epsilon$ with probability at least $1-\delta$.*

Corollary 2 *For data distribution $\mathcal{D}$ α-expanding with respect to hypothesis class $\mathcal{H}_1 \times \mathcal{H}_2$ according to (10), if the multi-view classifiers satisfy β-condition and $Pr(S_1^* \oplus S_2^*) \leq \epsilon/2$, by requesting $\widetilde{O}(\frac{1}{\epsilon})$ 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 $1-\delta$.*

The complete proof of Theorem 6, the proofs of Theorem 7 to 9 and Corollary 2 are given in the supplementary file.

7 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 $\widetilde{O}(\log \frac{1}{\epsilon})$, 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 $\widetilde{O}(\frac{1}{\epsilon})$, 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.

Acknowledgments

This work was supported by the NSFC (60635030, 60721002), 973 Program (2010CB327903) and JiangsuSF (BK2008018).

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.