Noise-Adaptive Margin-Based Active Learning and Lower Bounds under Tsybakov Noise Condition

Yining Wang and **Aarti Singh**

Machine Learning Department, School of Computer Science,
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA 15213

Abstract

We present a simple noise-robust margin-based active learning algorithm to find homogeneous (passing the origin) linear separators and analyze its error convergence when labels are corrupted by noise. We show that when the imposed noise satisfies the Tsybakov low noise condition (Mammen, Tsybakov, and others 1999; Tsybakov 2004) the algorithm is able to adapt to unknown level of noise and achieves optimal statistical rate up to polylogarithmic factors.

We also derive lower bounds for margin based active learning algorithms under Tsybakov noise conditions (TNC) for the membership query synthesis scenario (Angluin 1988). Our result implies lower bounds for the stream based selective sampling scenario (Cohn 1990) under TNC for some fairly simple data distributions. Quite surprisingly, we show that the sample complexity cannot be improved even if the underlying data distribution is as simple as the uniform distribution on the unit ball. Our proof involves the construction of a well-separated hypothesis set on the d-dimensional unit ball along with carefully designed label distributions for the Tsybakov noise condition. Our analysis might provide insights for other forms of lower bounds as well.

1 Introduction

Active learning is an increasingly popular setting in machine learning that makes use of both unlabeled and selectively sampled labeled data (Balcan, Beygelzimer, and Langford 2006; Cohn, Atlas, and Ladner 1994; Dasgupta 2005). In general, an active learning algorithm has access to a large number of unlabeled examples and has the capacity to request labels of specific examples. The hope is that by directing label queries to the most informative examples in a feedback-driven way, we might be able to achieve significant improvements in terms of sample complexity over passive learning algorithms. For instance, in the problem of learning homogeneous (passing the origin) linear separators, an exponential improvement in sample complexity could be achieved under the realizable case, where the labels are consistent with the optimal linear classifier (Balcan, Broder, and Zhang 2007; Balcan and Long 2013). For noisy label distributions, a polynomial improvement in sample complexity is more typical (Castro and Nowak 2008; Balcan, Broder,

and Zhang 2007; Balcan, Beygelzimer, and Langford 2006; Balcan and Long 2013).

We consider two active learning scenarios in this paper: the *stream-based selective sampling* scenario (Cohn 1990; Cohn, Atlas, and Ladner 1994), under which an algorithm has access to a large number of unlabeled data in a stream and can decide whether to query the label of a specific data point, and the *membership query synthesis* scenario (Angluin 1988) under which an algorithm has the capacity of synthesizing data points and obtaining their labels from an oracle. For the stream-based setting, we analyze a noise-robust margin-based active learning algorithm under the Tsybakov noise condition (Mammen, Tsybakov, and others 1999; Tsybakov 2004). We show that the algorithm automatically adjusts to unknown noise levels in the Tsybakov noise condition (TNC) while achieving the same statistical rate (up to polylogarithmic terms) as non-adaptive algorithms. This makes margin-based active learning more practical, as the amount of noise in label distributions is usually unknown in practice.

We also study lower bounds for the membership query synthesis setting under Tsybakov noise conditions. Our lower bound matches previous ones for the stream-based selective sampling setting (Balcan and Long 2013; Hanneke to appear). Quite surprisingly, as a consequence of our lower bound, we show that stream-based active learning algorithms cannot do better even if the underlying data distribution is as simple as the uniform distribution, It also means the previous proposed margin-based active learning algorithms (Balcan, Broder, and Zhang 2007; Balcan and Long 2013) are optimal under their specific problem settings. To the best of our knowledge, such results are not implied by any previous lower bounds on active learning, as we discuss in more detail below.

2 Related work

A margin-based active learning algorithm for learning homogeneous linear separators was proposed in (Balcan, Broder, and Zhang 2007) with its sample complexity analyzed under the Tsybakov low noise condition for the uniform distribution on the unit ball. The algorithm was later extended to log-concave data distributions (Balcan and Long 2013). Recently (Hanneke and Yang 2014) introduced a disagreement-based active learning algorithm that works

for arbitrary underlying data distributions. For all of the above-mentioned algorithms, given data dimension d and query budget T, the excess risk ϵ is upper bounded by [1] $\widetilde{O}((d/T)^{1/2\alpha})$, where α is a parameter characterizing the noise level in TNC (cf. Eq. (1) in Section 3). These algorithms are not noise-adaptive; that is, the selection of key algorithm parameters depend on the noise level α, which may not be available in practice.

In (Hanneke 2011) a noise-robust disagreement-based algorithm was proposed for agnostic active learning. The analysis was further improved in (Zhang and Chauduri 2014) by replacing the disagreement coefficient with a provably smaller quantity. However, their error bounds are slightly worse under our settings, as we discuss in Section 6. Also, in both analysis the desired accuracy ϵ is fixed, while in our setting the number of active queries T is fixed. Under the one-dimensional threshold learning setting, (Ramdas and Singh 2013a) proposed a noise-adaptive active learning algorithm inspired by recent developments of adaptive algorithms for stochastic convex optimization (Juditsky and Nesterov 2014). For multiple dimensions, it was shown recently in (Awasthi et al. 2014) that a noise-robust variant of margin-based active learning achieves near optimal noise tolerance. The authors analyzed the maximum amount of adversarial noise an algorithm can tolerate under the constraints of constant excess risk and polylogarithmic sample complexity, which is equivalent to an exponential rate of error convergence. In contrast, we study the rate at which the excess risk (relative to Bayes optimal classifier) converges to zero with number of samples that are not restricted to be polylogarithmic.

In terms of negative results, it is well-known that the $\widetilde{O}((d/T)^{1/2\alpha})$ upper bound is tight up to polylogarithmic factors. In particular, Theorem 4.3 in (Hanneke to appear) shows that for any stream-based active learning algorithm, there exists a distribution P_{XY} satisfying TNC such that the excess risk ϵ is lower bounded by $\Omega((d/T)^{1/2\alpha})$. The marginal data distribution P_X is constructed in an adversarial manner and it is unclear whether the same lower bound applies when P_X is some simple (e.g., uniform or Gaussian) distribution. (Balcan and Long 2013) proved lower bounds for stream-based active learning under each log-concave data distribution. However, their proof only applies to the separable case and shows an exponential error convergence. In contrast, we consider Tsybakov noise settings with parameter $\alpha \in (0,1)$, for which polynomial error convergence is expected (Hanneke to appear).

(Castro and Nowak 2008) analyzed the minimax rate of active learning under the membership query synthesis model (cf. Section 3). Their analysis implies a lower bound for stream-based setting when the data distribution is uniform or bounded from below (cf. Proposition 1 and 2). However, their analysis focuses on the nonparametric setting where the Bayes classifier f^* is not assumed to have a parametric form such as linear. Consequently, their is a polynomial gap between their lower bound and the upper bound for linear classifiers.

3 Problem setup and notations

We assume the data points $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are drawn from an unknown joint distribution P_{XY}, where $\mathcal{X}$ is the instance space and $\mathcal{Y}$ is the label space. Furthermore, x are drawn in an i.i.d. manner. In this paper we assume that $\mathcal{X} = \mathcal{S}^d \subseteq \mathbb{R}^d$ is the unit ball in $\mathbb{R}^d$ and $\mathcal{Y} = \{+1, -1\}$.

The goal of active learning is to find a classifier $f : \mathcal{X} \rightarrow \mathcal{Y}$ such that the generalization error $\text{err}(f) = \mathbb{E}_{(x,y)\sim P}[\ell(f(x), y)]$ is minimized. Here $\ell(f(x), y)$ is a loss function between the prediction $f(x)$ and the label y. Under the binary classification setting with $\mathcal{Y} = \{+1, -1\}$, the 0/1 classification loss is of interest, where $\ell(f(x), y) = I[yf(x) > 0]$ with $I[\cdot]$ the indicator function. In this paper we consider the case where the Bayes classifier f^* is linear, that is, $f^*(x) = \text{argmax}_{y \in \{+1,-1\}} \Pr(Y = y | X = x) = \text{sgn}(w^* \cdot x)$ with $w^* \in \mathbb{R}^d$. Note that the Bayes classifier f^* minimizes the generalization 0/1 error $\Pr(Y \neq f^*(X))$. Given the optimal classifier f^*, we define the *excess risk* of a classifier f under 0/1 loss as $\text{err}(f) - \text{err}(f^*)$. Without loss of generality, we assume all linear classifiers $f(x) = \text{sgn}(w \cdot x)$ have norm $\|w\|_2 = 1$. We also use $B_\theta(w, \beta)$ to denote the model class $\{f(x) = w' \cdot x | \theta(w', w) \leq \beta, \|w'\|_2 = 1\}$ consisting of all linear classifiers that are close to w with an angle at most β. Here $\theta(w', w) = \arccos(w' \cdot w)$ is the angle between w' and w. We use $\log$ to denote $\log_2$ and $\ln$ to denote the natural logarithm.

Tsybakov noise condition For the conditional label distribution $P_{Y|X}$, we consider a noise model characterized by the Tsybakov low noise condition (TNC) along the optimal hyperplane. Various forms of the TNC condition for the one-dimensional and multi-dimensional active learning are explored in (Castro and Nowak 2006; Ramdas and Singh 2013a; Balcan, Broder, and Zhang 2007; Balcan and Long 2013) and have been increasingly popular in the active learning literature. In this paper, we use the following version of the TNC condition: there exists a constant $0 < \mu < \infty$ such that for all linear classifiers [2] $w \in \mathbb{R}^d$, $\|w\|_2 = 1$ the following lower bound on excess risk holds:

$$\mu \cdot \theta(w, w^*)^{1/(1-\alpha)} \leq \text{err}(w) - \text{err}(w^*), \qquad (1)$$

with $\alpha \in [0, 1)$ a parameter characterizing the noise level in the underlying label distribution.

Stream-based selective sampling The stream-based selective sampling scheme was proposed in (Cohn 1990; Cohn, Atlas, and Ladner 1994). Under the stream-based setting an algorithm has access to a stream of unlabeled data points and can request labels of selected data points in a feedback-driven manner. Formally speaking, a stream-based active learning algorithm operates in iterations and for iteration t it does the following:

---

[1] In the $\widetilde{O}(\cdot)$ notation we omit dependency on failure probability δ and polylogarithmic dependency on d and T.

[2] To simplify notations, we will interchangeably call w, f and $\text{sgn}(f)$ as linear classifiers.

1. The algorithm obtains an unlabeled data point x_t, sampled from the marginal distribution P_X.

2. The algorithm then decides, based on previous labeled and unlabeled examples, whether to accept x_t and request its label. If a request is made, it obtains label y_t sampled from the conditional distribution $p(\cdot|x_t)$.

Finally, after a finite number of iterations the algorithm outputs a hypothesis $\widehat{f}(x) = \operatorname{sgn}(\widehat{w} \cdot x)$. We use $\mathcal{A}_{d,T}^{\mathrm{str}}$ to denote all stream-based selective sampling algorithms that operate on $\mathcal{X} = \mathcal{S}^d$ and make no more than T label requests.

The stream-based selective sampling setting is slightly weaker than the pool based active learning setting considered in (Balcan, Broder, and Zhang 2007; Balcan and Long 2013). For pool-based active learning, an algorithm has access to the entire pool $(x_1, x_2, \cdots)$ of unlabeled data before it makes any query requests. We remark that all margin-based active learning algorithms proposed in (Balcan, Broder, and Zhang 2007; Balcan and Long 2013; Awasthi et al. 2014) actually work under the stream-based setting.

Membership query synthesis An alternative active learning scenario is the synthetic query setting under which an active learning algorithm is allowed to synthesize queries and ask an oracle to label them. The setting is introduced in (Angluin 1988) and considered in (Castro and Nowak 2006; 2008; Ramdas and Singh 2013a). Formally speaking, a query synthesis active learning algorithm operates in iterations and for iteration t it does the following:

1. The algorithm picks an arbitrary data point $x_t \in \mathcal{X}$, based on previous obtained labeled data.

2. The algorithm is returned with label y_t sampled from the conditional distribution $p(\cdot|x_t)$.

Finally, after T iterations the algorithm outputs a hypothesis $\widehat{f}(x) = \operatorname{sgn}(\widehat{w} \cdot x)$, where T is the total number of label queries made. We use $\mathcal{A}_{d,T}^{\mathrm{qs}}$ to denote all membership query algorithms that operate on $\mathcal{X} = \mathcal{S}^d$ and make no more than T label queries.

We remark that the synthetic query setting is more powerful than stream-based selective sampling. More specifically, we have the following proposition. It can be proved by simple reductions and the proof is deferred to Appendix C in (Wang and Singh 2014).

Proposition 1. *Fix d, T. For any marginal distribution P_X and family of conditional label distributions $\mathcal{P}$ the following holds:*

$$\inf_{A \in \mathcal{A}_{d,T}^{\mathrm{qs}}} \sup_{P_{Y|X} \in \mathcal{P}} \mathbb{E}[L(\widehat{w}, w^*)]$$
$$\leq \inf_{A \in \mathcal{A}_{d,T}^{\mathrm{str}}} \sup_{P_{Y|X} \in \mathcal{P}} \mathbb{E}[L(\widehat{w}, w^*)], \quad (2)$$

where $L(\widehat{w}, w^*) = \operatorname{err}(\widehat{w}) - \operatorname{err}(w^*)$ is the excess risk of output hypothesis $\widehat{w}$.

4 Noise-adaptive upper bounds

In this section we prove the following main theorem, which provides an upper excess-risk bound on stream-based active learning algorithms that adapt to different noise levels under the TNC condition.

Theorem 1. *Fix $\delta \in (0,1), r \in (0, 1/2), d \geq 4$ and $T \geq 4$. Suppose P_X is the uniform distribution on the unit ball $\mathcal{S}^d$. There exists a stream-based active learning algorithm $A \in \mathcal{A}_{d,T}^{\mathrm{str}}$ such that for any label distribution $P_{Y|X}$ that satisfies Eq. (1) with parameters $\mu > 0$ and $1/(1 + \log(1/r)) \leq \alpha < 1$, the following holds with probability $\geq 1 - \delta$:*

$$\operatorname{err}(\widehat{w}) - \operatorname{err}(w^*) = \widetilde{O}\left(\left(\frac{d + \log(1/\delta)}{T}\right)^{1/2\alpha}\right). \quad (3)$$

Here $\widehat{w}$ is the output decision hyperplane of A, w^* is the Bayes classifier and in $\widetilde{O}(\cdot)$ we omit dependency on r, μ and polylogarithmic dependency on T and d.

Theorem 1 shows one can achieve the same error rate (up to polylogarithmic factors) as previously proposed algorithms (Balcan, Broder, and Zhang 2007; Balcan and Long 2013) without knowing noise level in the label distribution (characterized by μ and α). To prove Theorem 1, we explicitly construct an algorithm that is adaptive to unknown noise levels (Algorithm 1). The algorithm is in principle similar to the margin-based active learning algorithms proposed in (Balcan, Broder, and Zhang 2007; Balcan and Long 2013), with the setting of margin thresholds a slight generalization of (Awasthi et al. 2014). However, we analyze it under the noise-adaptive TNC setting, which has not been considered before specifically for margin-based active learning algorithms.

In the remainder of this section we describe the noise-adaptive algorithm we analyzed and provide a proof sketch for Theorem 1. Our analysis can be easily generalized to log-concave densities, with details in Appendix A.2 in (Wang and Singh 2014).

4.1 The algorithm

We present Algorithm 1, a margin-based active learning algorithm that adapts to unknown α and μ values in the TNC condition in Eq. (1). Algorithm 1 admits 4 parameters: d is the dimension of the instance space $\mathcal{X}$; T is the sampling budget (i.e., maximum number of label requests allowed); δ is a confidence parameter; $r \in (0, 1/2)$ is the shrinkage rate of the hypothesis space for every iteration in the algorithm; smaller r allows us to adapt to smaller α values but will result in a larger constant in the excess risk bound. The basic idea of the algorithm is to split T label requests into E iterations, using the optimal passive learning procedure within each iteration and reducing the scope of search for the best classifier after each iteration.

The key difference between the adaptive algorithm and the one presented in (Balcan, Broder, and Zhang 2007) is that in Algorithm 1 the number of iterations E as well as other parameters (e.g., b_k, β_k) are either not needed or do not depend on the noise level α, and the number of label queries is divided evenly across the iterations. Another difference is

Algorithm 1 A noise-adaptive margin-based active learning algorithm

---

1: **Parameters**: data dimension d, query budget T, failure probability δ, shrinkage rate r.
2: **Initialize**: $E = \frac{1}{2}\log T$, $n = T/E$, $\beta_0 = \pi$, random $\widehat{w}_0$ with $\|\widehat{w}_0\| = 1$.
3: **for** $k = 1$ to E **do**
4:    $W = \emptyset$. Set $b_{k-1} = \frac{2\beta_{k-1}}{\sqrt{d}}\sqrt{E(1 + \log(1/r))}$ if $k > 1$ and $b_{k-1} = +\infty$ if $k = 1$.
5:    **while** $|W| < n$ **do**
6:       Obtain a sample x from P_X.
7:       If $|\widehat{w}_{k-1} \cdot x| > b_{k-1}$, reject; otherwise, ask for the label of x, and put (x, y) into W.
8:    **end while**
9:    Find $\widehat{w}_k \in B_\theta(\widehat{w}_{k-1}, \beta_{k-1})$ that minimizes the empirical 0/1 error $\sum_{(x,y)\in W} I[yw \cdot x < 0]$.
10:   Update: $\beta_k \leftarrow r \cdot \beta_{k-1}$, $k \leftarrow k + 1$.
11: **end for**
12: **Output**: the final estimate $\widehat{w}_E$.

---

that in our algorithm the sample budget T is fixed while in previous work the error rate ϵ is known. It remains an open problem whether there exists a tuning-free active learning algorithm when a target error rate ϵ instead of query budget T is given (Ramdas and Singh 2013a).

4.2 Proof sketch of Theorem 1

In this section we sketch the proof of Theorem 1. The complete proof is deferred to Appendix A in (Wang and Singh 2014).

We start by defining some notations used in the proof. Let $\mathcal{F}_k = B_\theta(\widehat{w}_{k-1}, \beta_{k-1})$ be the hypothesis space considered in the kth iteration of Algorithm 1. Let D_k be the obtained labeled examples and $S_1^{(k)} = \{x | |\widehat{w}_{k-1} \cdot x| \leq b_{k-1}\}$ be the acceptance region at the kth iteration. By definition, $D_k \subseteq S_1^{(k)}$. Let $w_k^* = \arg\max_{w \in \mathcal{F}_k} \text{err}(w | S_1^{(k)})$ be the optimal classifier in $\mathcal{F}_k$ with respect to the generalization 0/1 loss in the acceptance region $S_1^{(k)}$. Using similar techniques as in (Balcan, Broder, and Zhang 2007), it can be shown that with probability $\geq 1 - \delta$ the following holds:

$$\text{err}(\widehat{w}_k) - \text{err}(w_k^*) \leq \beta_{k-1}\epsilon, \tag{4}$$

where ϵ is of the order $\widetilde{O}(\sqrt{\frac{d + \log(1/\delta)}{T}})$.

Eq. (4) shows that if β_{k-1} is small then we get good excess risk bound. However, β_{k-1} should be large enough so that $\mathcal{F}_k$ contains the Bayes classifier w^* (i.e., $w_k^* = w^*$). In previous analysis (Balcan, Broder, and Zhang 2007; Balcan and Long 2013) the algorithm parameters β_{k-1} and b_{k-1} are carefully selected using the knowledge of α and μ so that $w_k^* = w^*$ for all iterations. This is no longer possible under our setting because the noise parameters α and μ are unknown. Instead, we show that there exists a "tipping point" $k^* \in \{1, 2, \cdots, E - 1\}$ depending and α and μ that divides Algorithm 1 into two phases: in the first phase ($k \leq k^*$) everything behaves the same with previous analysis for non-adaptive margin-based algorithm; that

is, we have per-iteration excess error upper bounded by Eq. (4) and the optimal Bayes classifier w^* is contained in the constrained hypothesis space $\mathcal{F}_k$ (i.e., $w_k^* = w^*$) for all $k \leq k^*$. Formally speaking, we have the following two lemmas which are proved in Appendix A in (Wang and Singh 2014).

Lemma 1. *Suppose $r \in (0, 1/2)$ and $1/(1 + \log(1/r)) \leq \alpha < 1$. With probability at least $1 - \delta$,*

$$\text{err}(\widehat{w}_{k^*}) - \text{err}(w_{k^*}^*) \leq \beta_{k^*-1}\epsilon \leq \frac{\epsilon^{1/\alpha}}{r^{\frac{1+\alpha}{\alpha}}\mu^{\frac{1-\alpha}{\alpha}}}. \tag{5}$$

Lemma 2. *With probability $\geq 1 - \delta E$, $w_k^* = w^*$ for all $k \leq k^*$.*

After iteration k^*, the optimal Bayes classifier w^* diverges from w_k^* and we can no longer apply Eq. (4) directly to bound the excess risk between $\widehat{w}_k$ and w^*. However, for $k > k^*$ the constrained hypothesis space $\mathcal{F}_k$ is quite small and the empirical estimator $\widehat{w}_k$ cannot deviate much from $\widehat{w}_{k-1}$. In particular, we have the following lemma, which is proved in Appendix A in (Wang and Singh 2014).

Lemma 3. *Suppose $r \in (0, 1/2)$. With probability at least $1 - \delta E$, we have*

$$\text{err}(\widehat{w}_E) - \text{err}(\widehat{w}_{k^*}) \leq \frac{r}{1-r}\beta_{k^*-1}\epsilon. \tag{6}$$

Combining Lemma 1, 2 and 3 we can upper bound the excess risk $\text{err}(\widehat{w}_E) - \text{err}(w^*)$ by $\widetilde{O}(\epsilon^{1/\alpha})$, which corresponds to $\widetilde{O}((\frac{d + \log(1/\delta)}{T})^{1/2\alpha})$ in Eq. (7). The complete proof is deferred to Appendix A in (Wang and Singh 2014).

4.3 Extension to log-concave densities

Following recent developments in margin-based active learning (Balcan, Broder, and Zhang 2007; Balcan and Long 2013), Theorem 1 can be further generalized to the case when the data distribution P_X has *log-concave densities*, which includes the uniform data distribution. A density function g is said to be *log-concave* if $\log g(\cdot)$ is a concave function. Many popular distributions have log-concave densities, including Gaussian distribution and uniform distribution. We say the data distribution P_X is *isotropic* if the mean of P_X is zero and the covariance matrix of P_X is the identity. Theorem 2 shows that, with slight modifications, Algorithm 1 can be generalized to the case when the data distribution P_X is log-concave and isotropic. Its proof is similar to the one in (Balcan and Long 2013) and is deferred to Appendix A.2 in (Wang and Singh 2014).

Theorem 2. *Fix $\delta \in (0, 1), r \in (0, 1/2), d \geq 4$ and $T \geq 4$. Suppose P_X is an isotropic log-concave distribution on the unit ball $\mathcal{S}^d$ and $P_{Y|X}$ satisfies Eq. (1) with parameters $\mu > 0$ and $1/(1 + \log(1/r)) \leq \alpha < 1$ Let $\widehat{w}$ be the output of Algorithm 1 run with $b_{k-1} = C_1\beta_{k-1}\log T$ and the other parameters unchanged.* [3] *Then with probability at least $1 - \delta$*

---

[3] C_1 is an absolute constant. See Lemma 7, 8 in Appendix A.2 and Theorem 8 in (Balcan and Long 2013) for details.

the following holds:

$$\mathrm{err}(\widehat{w}) - \mathrm{err}(w^*) = \widetilde{O}\left(\left(\frac{d + \log(1/\delta)}{T}\right)^{1/2\alpha}\right). \quad (7)$$

Here $\widehat{w}$ is the output decision hyperplane of A, w^* is the Bayes classifier and in $\widetilde{O}(\cdot)$ we omit dependency on r, μ and polylogarithmic dependency on T and d.

5 Lower bounds

We prove lower bounds for active learning under the membership query synthesis setting. Since the query synthetic setting is more powerful than the stream-based setting as shown in Proposition 1, our result implies a lower bound for stream-based selective sampling. Our lower bound for membership query synthesis setting is for a slightly different version of TNC, which implies TNC in Eq. (1) for distributions that are bounded from below (including the uniform distribution). This shows that both Algorithm 1 and previous margin-based algorithms (Balcan, Broder, and Zhang 2007) achieve the minimax rate (up to polylogarithmic factors) under the uniform distribution on the unit ball.

To facilitate our analysis for the query synthesis setting, in this section we adopt a new formulation of TNC condition in terms of the label distribution function $\eta(x) = \Pr(y = 1|x)$. Formally speaking, we assume that there exist constants $0 < \mu_0 < \infty$ and $\alpha \in [0, 1)$ such that for all $x \in X$ the following holds:

$$\mu_0 \cdot |\varphi(x, w^*)|^{\alpha/(1-\alpha)} \leq |\eta(x) - 1/2|, \quad (8)$$

where w^* is the Bayes classifier with respect to $\eta(\cdot)$ and $\varphi(x, w^*) := \frac{\pi}{2} - \theta(x, w^*) \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ is the signed acute angle between x and the decision hyperplane associated with w^*. Similar formulation was also used in (Castro and Nowak 2006; 2008; Ramdas and Singh 2013a) to analyze active learning algorithms under the query synthesis setting. We also remark that Eq. (8) implies the excess-risk based TNC condition in Eq. (1) for data distributions with densities bounded from below, as shown in Proposition 2. Its proof is deferred to Appendix C in (Wang and Singh 2014).

Proposition 2. *Suppose the density function g associated with the marginal data distribution P_X is bounded from below. That is, there exists a constant $\gamma \in (0, 1)$ such that $g \geq \gamma g_0$, where $g_0 \equiv \pi^{-d/2}\Gamma(1+d/2)$ is the uniform distribution on the unit d-dimensional ball. Then Eq. (8) implies Eq. (1) with $\mu = 2(1 - \alpha)\mu_0\gamma$.*

We now present the main theorem of this section, which establishes a lower bound on the angle between the output classifier $\widehat{w}$ and the Bayes classifier w^* for the membership query synthesis setting, assuming the label distribution $P_{Y|X}$ satisfies TNC condition in Eq. (8).

Theorem 3. *Fix $d \geq 2$, T, $\mu_0 > 0$ and $\alpha \in (0, 1)$. Suppose $\mathcal{X} = \mathcal{S}^d$ and $\mathcal{Y} = \{+1, -1\}$. Let $\mathcal{P}_{\alpha,\mu_0}$ denote the class of all conditional label distributions that satisfy the label distribution based TNC condition in Eq. (8) with parameters*

α, μ_0. *Then the following excess risk lower bound holds:*

$$\inf_{A \in \mathcal{A}_{d,T}^{qs}} \sup_{P_{Y|X} \in \mathcal{P}_{\alpha,\mu_0}} \mathbb{E}[\theta(\widehat{w}, w^*)] = \Omega\left(\left(\frac{d}{T}\right)^{(1-\alpha)/2\alpha}\right). \quad (9)$$

Here in the $\Omega(\cdot)$ notation we omit dependency on μ_0.

Theorem 3 implies a lower bound for excess-risk based TNC in Eq. (1) when the data distribution P_X is uniform or bounded from below, as shown in Corollary 1. By Proposition 1, Eq. (10) holds also for stream-based algorithms $\mathcal{A}_{d,T}^{str}$. We prove Corollary in Appendix B in (Wang and Singh 2014).

Corollary 1. *Fix $d \geq 2, T, \mu, \gamma > 0$ and $\alpha \in (0, 1)$. Suppose $\mathcal{X} = \mathcal{S}^d$, $\mathcal{Y} = \{+1 - 1\}$ and the density of P_X is bounded from below with constant γ. Let $\mathcal{P}_{\alpha,\mu}$ denotes the class of all label distributions that satisfy the excess-risk based TNC condition in Eq. (1) with parameters α, μ. Then the following lower bound holds:*

$$\inf_{A \in \mathcal{A}_{d,T}^{qs}} \sup_{P_{Y|X} \in \mathcal{P}_{\alpha,\mu}} \mathbb{E}[\mathrm{err}(\widehat{w}) - \mathrm{err}(w^*)]$$

$$= \Omega\left(\left(\frac{d}{T}\right)^{1/2\alpha}\right). \quad (10)$$

Here in the $\Omega(\cdot)$ notation we omit dependency on μ and γ.

5.1 Proof sketch of Theorem 3

In this section we sketch a proof for Theorem 3. The complete proof is deferred to Appendix B due to space constraints in (Wang and Singh 2014). We assume the data dimension $d \geq 2$ is even. This does not lose any generality because the lower bounds in Eq. (9) and (10) remain asymptotically the same if d is replaced with $(d + 1)$.

The main idea of the proof is the construction of a hypothesis set $\mathcal{W} = \{w_1^*, \cdots, w_m^*\} \subseteq \mathbb{R}^d$ with $\log |\mathcal{W}| = \Omega(d)$ such that for any hypothesis pair (w_i^*, w_j^*) the angle $\theta(w_i^*, w_j^*)$ is large while $\mathrm{KL}(P_{i,T}\|P_{j,T})$ is small [4]. Here $P_{i,T}$ denotes the distribution of T labels under the label distribution associated with w_i^* (rigorous mathematical definition of $P_{i,T}$ is given in the appendix of (Wang and Singh 2014)). Intuitively, we want w_i^* and w_j^* to be well separated in terms of the loss function (i.e., $\theta(w_i^*, w_j^*)$) while being hard to distinguish by any active learning algorithm under a fixed query budget T (implied by the KL divergence condition).

The following lemma accomplishes the first objective by lower bounding $\theta(w_i^*, w_j^*)$. Its proof is based on the construction of constant-weight codings (Graham and Sloane 1980) and is deferred to Appendix B in (Wang and Singh 2014).

Lemma 4. *Assume d is even. Fix a parameter $t \in (0, 1/4)$. There exists a hypothesis set $\mathcal{W} = \{w_1^*, \cdots, w_m^*\} \subseteq \mathbb{R}^d$ such that*

$$t \leq \theta(w_i^*, w_j^*) \leq 6.5t, \quad \forall i \neq j; \quad (11)$$

---

[4]For two continuous distributions P and Q with densities p and q, their Kullback-Leibler (KL) divergence $\mathrm{KL}(P\|Q)$ is defined as $\int p(x) \log \frac{p(x)}{q(x)} dx$ if $P \ll Q$ and $+\infty$ otherwise.

(a) Illustration of $P_{Y|X}^{(1)}$

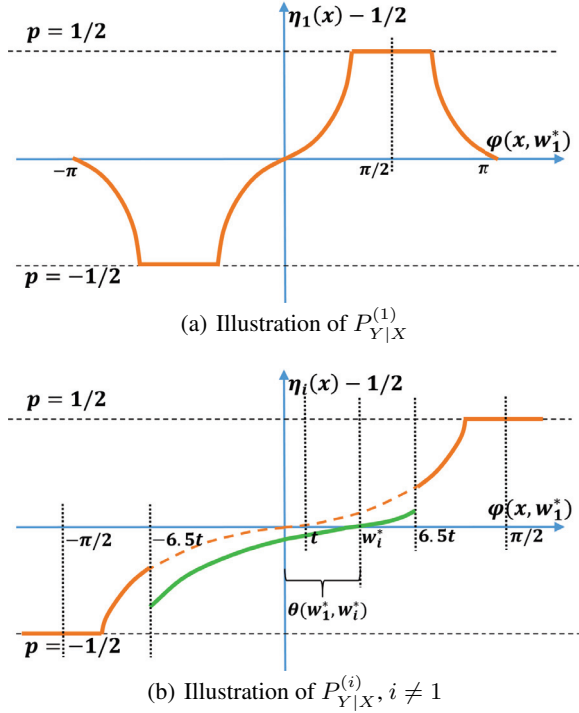(b) Illustration of $P_{Y|X}^{(i)}, i \neq 1$

Figure 1: Graphical illustrations of $P_{Y|X}^{(1)}$ (left) and $P_{Y|X}^{(i)}$ (right) constructed as in Eq. (12). Solid lines indicate the actual shifted probability density functions $\eta(x) - 1/2$ where $\eta(x) = \Pr[Y = 1 | X = x]$. In Figure 1(b), the orange curve (both solid and dashed) satisfies TNC with respect to w_1^* and the green curve satisfies TNC with respect to w_i^*. Note the two discontinuities at $\varphi(x, w_1^*) = \pm 6.5t$. Figure 1(b) is not 100% accurate because it assumes that $\varphi(x, w_1^*) = \varphi(x, w_i^*) + \theta(w_1^*, w_i^*)$, which may not hold for $d > 2$.

furthermore, $\log |\mathcal{W}| \geq 0.0625d$ *for* $d \geq 2$.

We next tackle the second objective of upper bounding $\mathrm{KL}(P_{i,T} \| P_{j,T})$. This requires designing label distributions $\{P_{Y|X}^{(i)}\}_{i=1}^m$ such that they satisfy the TNC condition in Eq. (8) while having small KL divergence between $P_{Y|X}^{(i)}$ and $P_{Y|X}^{(j)}$ for all distinct pairs (i, j). We construct the label distribution for the ith hypothesis as below:

$$P_{Y|X}^{(i)}(Y = 1|x) = \begin{cases} \frac{1}{2} + \mathrm{sgn}(w_i^* \cdot x)\wp(|\varphi(w_i^*, x)|), \\ \quad \text{if } |\varphi(w_1^*, x)| \leq 6.5t; \\ \frac{1}{2} + \mathrm{sgn}(w_1^* \cdot x)\wp(|\varphi(w_1^*, x)|), \\ \quad \text{if } |\varphi(w_1^*, x)| > 6.5t; \end{cases}$$

$$(12)$$

where $\varphi(w, x) = \frac{\pi}{2} - \theta(x, w) \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ and $\wp$ is defined as

$$\wp(\vartheta) := \min\{2^{\alpha/(1-\alpha)}\mu_0 \cdot \vartheta^{\alpha/(1-\alpha)}, 1/2\}. \quad (13)$$

A graphical illustration of $P_{Y|X}^{(1)}$ and $P_{Y|X}^{(i)}$ constructed in Eq. (12) is depicted in Figure 1. We use the same distribution

when data points are far from the optimal classification hyperplane (i.e., $|\varphi(w^*, x)| > 6.5t$) in order to maximize the "indistinguishability" of the constructed conditional distributions. On the other hand, by TNC assumption $P_{Y|X}^{(i)}$ must have $f_i^*(x) = \mathrm{sgn}(w_i^* \cdot x)$ as its Bayes classifier and TNC condition along the hyperplane w_i^* must hold. As a result, when a data point is close to the hyperplane represented by w_i^* the label distribution differs for each hypothesis w_i^* in $\mathcal{W}$. Similar construction of adversarial distributions was also adopted in (Castro and Nowak 2008) to prove lower bounds for one-dimensional active threshold learners.

Lemma 5 summarizes key properties of the label distributions $\{P_{Y|X}^{(i)}\}_{i=1}^m$ constructed as in Eq. (12). It is proved in Appendix B of (Wang and Singh 2014).

Lemma 5. *Suppose* $\mathcal{W} = \{w_1^*, \cdots, w_m^*\} \subseteq \mathbb{R}^d$ *satisfies Eq. (11) and* $\{P_{Y|X}^{(i)}\}_{i=1}^m$ *is constructed as in Eq. (12). Then for every i the hypothesis $f_i^*(x) = \mathrm{sgn}(w_i^* \cdot x)$ is the Bayes estimator of $P_{Y|X}^{(i)}$ and the TNC condition in Eq. (8) holds with respect to w_i^*. In addition, for every $i \neq j$ the KL divergence between $P_{i,T}$ and $P_{j,T}$ is upper bounded by*

$$\mathrm{KL}(P_{i,T} \| P_{j,T}) \leq C \cdot Tt^{2\alpha/(1-\alpha)}, \quad (14)$$

where C is a positive constant that does not depend on T or t.

With Lemma 12 lower bounding $\theta(w_i^*, w_j^*)$ and Lemma 5 upper bounding $\mathrm{KL}(P_{i,T} \| P_{j,T})$, Theorem 3 and Corollary 1 can be proved by applying standard information theoretical lower bounds (Tsybakov and Zaiats 2009). A complete proof can be found in Appendix B in (Wang and Singh 2014).

6 Discussion and remarks

Comparison with noise-robust disagreement-based active learning algorithms In (Hanneke 2011) another noise-robust adaptive learning algorithm was introduced. The algorithm is originally proposed in (Dasgupta, Hsu, and Monteleoni 2007) and is based on the concept of *disagreement coefficient* introduced in (Hanneke 2007). The algorithm adapts to different noise level α, and achieves an excess error rate of

$$O\left(\left(\frac{\vartheta(d \log T + \log(1/\delta))}{T}\right)^{\frac{1}{2\alpha}}\right) \quad (15)$$

with probability $1 - \delta$, where d is the underlying dimensionality, T is the sample query budget and ϑ is the disagreement coefficient. Under our scenario where X is the origin-centered unit ball in $\mathbb{R}^d$ for $d > 2$, the hypothesis class $\mathbb{C}$ contains all linear separators whose decision surface passes passing the origin and P_X is the uniform distribution, the disagreement coefficient ϑ satisfies (Hanneke 2007) $\frac{\pi}{4}\sqrt{d} \leq \vartheta \leq \pi\sqrt{d}$. As a result, the algorithm presented in this paper achieves a polynomial improvement in d in terms of the convergence rate. Such improvements show the advantage of margin-based active learning and were also observed in (Balcan and Long 2013). Also, our algorithm is considerably much simpler and does not require computing lower and upper confidence bounds on the classification performance.

Connection to adaptive convex optimization Algorithm 1 is inspired by an adaptive algorithm for stochastic convex optimization presented in (Juditsky and Nesterov 2014). A function f is called *uniformly convex* on a closed convex set Q if there exists $\rho \geq 2$ and $\mu \geq 0$ such that for all $x, y \in Q$ and $\alpha \in [0, 1]$,

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) \\ - \frac{1}{2}\mu\alpha(1 - \alpha)\|x - y\|^{\rho}. \quad (16)$$

Furthermore, if $\mu > 0$ we say the function f is *strongly convex*. In (Juditsky and Nesterov 2014) an adaptive stochastic optimization algorithm for uniformly and strongly convex functions was presented. The algorithm adapts to unknown convexity parameters ρ and μ in Eq. (16).

In (Ramdas and Singh 2013a) a connection between multi-dimensional stochastic convex optimization and one-dimensional active learning was established. The TNC condition in Eq. (1) and the strongly convex condition in Eq. (16) are closely related, and the exponents α and ρ are tied together in (Ramdas and Singh 2013b). Based on this connection, a one-dimensional active threshold learner that adapts to unknown TNC noise levels was proposed.

In this paper, we extend the algorithms presented in (Juditsky and Nesterov 2014; Ramdas and Singh 2013a) to build an adaptive margin-based active learning for multi-dimensional data. Furthermore, the presented algorithm adapts to all noise level parameters $\alpha \in (0, 1)$ with appropriate setting of r, which corresponds to convexity parameters $\rho > 1$. [5] Therefore, we conjecture the existence of similar stochastic optimization algorithms that can adapt to a notion of degree of convexity $\rho < 2$ as introduced in (Ramdas and Singh 2013a).

Future work Algorithm 1 fails to handle the case when $\alpha = 0$. We feel it is an interesting direction of future work to design active learning algorithms that adapts to $\alpha = 0$ while still retaining the exponential improvement on convergence rate for this case, which is observed in previous active learning research (Balcan, Broder, and Zhang 2007; Balcan and Long 2013; Castro and Nowak 2008).

Acknowledgement

References

Angluin, D. 1988. Queries and concept learning. *Machine learning* 2(4):319–342.

Awasthi, P.; Bandeira, A.; Charikar, M.; Krishnaswamy, R.; Villar, S.; and Ward, R. 2014. The power of localization for efficiently learning linear separators with noise. In *STOC*.

Balcan, M.-F., and Long, P. 2013. Active and passive learning of linear separators under log-concave distributions. In *COLT*.

Balcan, M.-F.; Beygelzimer, A.; and Langford, J. 2006. Agnostic active learning. In *ICML*.

Balcan, M.-F.; Broder, A.; and Zhang, T. 2007. Margin based active learning. In *COLT*.

Castro, R., and Nowak, R. 2006. Upper and lower error bounds for active learning. In *The 44th Annual Allerton Conference on Communication, Control and Computing*.

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

Cohn, D.; Atlas, L.; and Ladner, R. 1994. Improving generalization with active learning. *Machine Learning* 15(2):201–221.

Cohn, D. 1990. Neural network exploration using optimal experiment design. In *NIPS*.

Dasgupta, S.; Hsu, D.; and Monteleoni, C. 2007. A general agnostic active learning algorithm. In *NIPS*.

Dasgupta, S. 2005. Coarse sample complexity bounds for active learning. In *NIPS*.

Graham, R., and Sloane, N. 1980. Lower bounds for constant weight codes. *IEEE Transactions on Information Theory* 26(1):37–43.

Hanneke, S., and Yang, L. 2014. Minimax analysis of active learning. *arXiv:1410.0996*.

Hanneke, S. 2007. A bound on the label complexity of agnostic active learning. In *ICML*.

Hanneke, S. 2011. Rates of convergence in active learning. *The Annals of Statistics* 39(1):333–361.

Hanneke, S. to appear. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*.

Juditsky, A., and Nesterov, Y. 2014. Primal-dual subgradient methods for minimizing uniformly convex functions. *arXiv:1401.1792*.

Mammen, E.; Tsybakov, A. B.; et al. 1999. Smooth discrimination analysis. *The Annals of Statistics* 27(6):1808–1829.

Ramdas, A., and Singh, A. 2013a. Algorithmic connections between active learning and stochastic convex optimization. In *ALT*.

Ramdas, A., and Singh, A. 2013b. Optimal rates for stochastic convex optimization under tsybakov noise condition. In *ICML*.

Tsybakov, A. B., and Zaiats, V. 2009. *Introduction to nonparametric estimation*, volume 11. Springer.

Tsybakov, A. B. 2004. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics* 135–166.

Wang, Y., and Singh, A. 2014. Noise-adaptive margin-based active learning and lower bounds under tsybakov noise condition. *arXiv:1406.5383*.

Zhang, C., and Chaudhuri, K. 2014. Beyond disagreement-based agnostic active learning. In *NIPS*.

---

[5]The relationship between α and ρ can be made explicitly by noting $\alpha = 1 - 1/\rho$.