Noise-adaptive Margin-based Active Learning and Lower Bounds under Tsybakov Noise Condition

Yining Wang *and Aarti Singh †

Machine Learning Department, Carnegie Mellon University

March 9, 2015

Abstract

We present a polynomial-time noise-robust margin-based active learning algorithm to find homogeneous (passing the origin) linear separators and analyze its statistical rate of error convergence when labels are corrupted by noise. We show that when the imposed noise satisfies the Tsybakov low noise condition [MT$^+$99, Tsy04] the algorithm is able to adapt to unknown level of noise and achieves optimal statistical rate up to polylogarithmic factors. In addition, the presented algorithm is simple and does not require prior knowledge of the amount of noise in the label distribution.

We also derive lower bounds for margin based active learning algorithms under Tsybakov noise conditions (TNC) for the membership query synthesis scenario [Ang88]. Our result implies lower bounds for the stream based selective sampling scenario [Coh90] 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 [BBL06, CAL94, Das05]. 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 [BBZ07, BL13]. For noisy label distributions, a polynomial improvement in sample complexity is more typical [CN08, BBZ07, BBL06, BL13].

We consider two active learning scenarios in this paper: the *stream-based selective sampling* scenario [Coh90, CAL94], 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 [Ang88] 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 [MT$^+$99, Tsy04]. We show that the algorithm automatically adjusts to

*yiningwa@cs.cmu.edu
†aarti@cs.cmu.edu

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 [BL13, Hanar]. 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 [BBZ07, BL13] 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.

Upper bounds A margin-based active learning algorithm for learning homogeneous linear separators was proposed in [BBZ07] 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 [BL13]. Recently [HY14] 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 2). 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 [Han11] a noise-robust disagreement-based algorithm was proposed for agnostic active learning. However, its error bound is slightly worse under our settings. We discuss their algorithm in Section 5. Under the one-dimensional threshold learning setting, [RS13a] proposed a noise-adaptive active learning algorithm inspired by recent developments of adaptive algorithms for stochastic convex optimization [JN14]. Also, for multiple dimensions, it was shown recently in [ABC$^+$14] 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 this paper we present a margin-based, noise-adaptive active learning algorithm similar to the one independently developed by [ABC$^+$14] and analyze its performance under uniform and log-concave data distributions with Tsybakov low noise conditions imposed on label noise. We slightly adjust previous algorithms by changing the margin shrinking rate in order to relax the constraints imposed on noise level α (previous work like [RS13a] requires $\alpha \geq 1/2$). We show that the noise-adaptive algorithm achieves optimal rate in terms of data dimension d and query budget T up to polylogarithmic factors. It also works for general surrogate loss functions (e.g., hinge loss and logistic loss), which makes the learning algorithm computationally efficient.

Lower bounds 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 [Hanar] 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. [BL13] 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 [Hanar].

[CN08] analyzed the minimax rate of active learning under the membership query synthesis model (cf. Section 2). 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

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

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.

In this paper we prove lower bounds for the membership query synthesis model under the assumption that the Bayes estimator is linear and the label distribution satisfies TNC with parameter $\alpha \in (0, 1)$ along the optimal classification hyperplane. The lower bound is tight up to polylogarithmic factors with previous upper bounds. Our results imply tight lower bounds for the stream-based setting when the underlying data distribution is the uniform distribution on the unit ball. This suggests that previous margin-based algorithm [BBZ07] and noise-robust algorithms presented in this paper are optimal, up to polylogarithmic factors.

Proof techniques Our analysis for noise-adaptive active learning was inspired by recent developments of adaptive algorithms in stochastic convex optimization [JN14] and connections between active learning and stochastic optimization [RS13a, RS13b]. The analysis also built on previous analysis of margin-based active learning algorithms [BBZ07, BL13]. A key observation is that it is not necessary to keep track of the optimal Bayes classifier during the entire course of an algorithm because at later stages the hypothesis space is so small that the per-iteration empirical optimizer cannot deviate much in error performance from previous iterations and hence remains close to the Bayes optimal classifier. This can be shown by smartly shrinking the hypothesis space after each iteration and a careful case analysis on TNC parameters.

The technique for lower bound proofs differs from previous work in, for example, [Hanar, Han11, HY14]. In particular, previous ideas of constructing smart adversarial data distributions fail because under the membership query synthesis scenario the active learning algorithm itself is responsible for synthesizing data points. In this paper, we extend the information-theoretical lower bound for one-dimensional threshold learning in [CN08] to multi-dimensional data with linear classifiers. Our proof is also inspired by analysis of lower bounds for stochastic convex optimization in [ABRW12], At the core of our lower bound proof is a well-separated hypothesis set under Tsybakov noise conditions constructed using previous results on constant-weight codings [GS80].

2 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 $f : \mathcal{X} \to \mathcal{Y}$ such that the generalization error $\mathrm{err}(f) = \mathbb{E}_{(x,y)\sim P}[\ell(f(x), y)]$ 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, $f^*(x) = \mathrm{argmax}_{y\in\{+1,-1\}}\Pr(Y = y|X = x) = \mathrm{sgn}(w^* \cdot x)$ 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 $\mathrm{err}_{0/1}(f) - \mathrm{err}_{0/1}(f^*)$. Without loss of generality, we assume all linear classifiers $f(x) = \mathrm{sgn}(w \cdot x)$ 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 [CN06, RS13a, BBZ07, BL13] 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:

$$\mu \cdot \theta(w, w^*)^{1/(1-\alpha)} \leq \mathrm{err}_{0/1}(w) - \mathrm{err}_{0/1}(w^*), \tag{1}$$

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 [Coh90, CAL94]. 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. 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 $\widehat{f}(x) = \mathrm{sgn}(\widehat{w} \cdot x)$. We use $\mathcal{A}^{\mathrm{str}}_{d,T}$ 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 [BBZ07, BL13]. 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 [BBZ07, BL13, ABC$^+$14] 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 [Ang88] and considered in [CN06, CN08, RS13a]. 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 $\widehat{f}(x) = \mathrm{sgn}(\widehat{w} \cdot x)$, where T is the total number of label queries made. We use $\mathcal{A}^{\mathrm{qs}}_{d,T}$ 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.

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

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

where $L(\widehat{w}, w^*) = \mathrm{err}_{0/1}(\widehat{w}) - \mathrm{err}_{0/1}(w^*)$ is the excess risk of output hypothesis $\widehat{w}$.

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

3 Upper bounds for noise-adaptive active learning

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 polynomial time stream-based active learning algorithm $A \in \mathcal{A}^{\mathrm{str}}_{d,T}$ 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 $\geq 1-\delta$:*

$$\mathrm{err}_{0/1}(\widehat{w}) - \mathrm{err}_{0/1}(w^*) = \widetilde{O}\left(\left(\frac{d+\log(1/\delta)}{T}\right)^{1/2\alpha}\right). \tag{3}$$

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 [BBZ07, BL13] without knowing noise level in the label distribution (characterized by μ and α). To prove Theorem 1, we explicitly construct an efficient 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 [BBZ07, BL13], with the setting of margin thresholds a slight generalization of [ABC$^+$14]. 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 first review properties of surrogate loss functions, which help us design efficient learning algorithms. We then describe the noise-adaptive algorithm we analyzed and provide a proof sketch for Theorem 1. We also show that the presented algorithm can be generalized to log-concave distributions, with details in Section 3.4 and Appendix A.2.

3.1 Surrogate loss functions

Finding a linear classifier that minimizes the empirical 0/1 classification error could be difficult because the 0/1 loss function is neither convex nor differentiable. Instead, people often try to minimize a surrogate loss function on the training dataset. Popular choices include hinge loss, logistic loss and exponential loss. In this paper, we consider surrogate loss functions that satisfy the following two properties:

(A1) There exists a passive learning algorithm that, without assuming TNC, achieves an excess surrogate error bounded by ϵ', i.e. with probability at least $1-\delta$, $\mathrm{err}(\widehat{w}) - \mathrm{err}(\widetilde{w}^*) \leq \epsilon'$ for some $0 < \epsilon' < 1$ that depends on δ, n and d. Here $\widehat{w}$ and $\widetilde{w}^*$ denote minimizers of the empirical and generalization surrogate loss error and n is the number of i.i.d. training examples. This holds true, for example, for 0/1 loss (c.f. Theorem 8 in [BBZ07]) as well as hinge and logistic loss (c.f. [RDVC$^+$04]) with $\epsilon' = \widetilde{O}\left(\sqrt{\frac{d+1+\log(1/\delta)}{n}}\right)$.

(A2) $\mathrm{err}_{0/1}(\widehat{w}) - \mathrm{err}_{0/1}(w^*) \leq \mathrm{err}(\widehat{w}) - \mathrm{err}(\widetilde{w}^*)$. That is, the excess risk under the loss we are working with upper bounds the excess risk under 0/1 loss. Here w^* denotes the minimizer of the generalization $0/1$ error, which corresponds to the Bayes classifier under our setting. This is true for many surrogate losses including hinge loss, logistic loss, etc [BJM06].

3.2 A margin based noise-adaptive active learning 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.

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

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

The key difference between the adaptive algorithm and the one presented in [BBZ07] 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 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 [RS13a]. Finally, instead of optimizing an empirical 0/1 error, Algorithm 1 works for any surrogate loss functions with properties (A1) and (A2).

3.3 Proof sketch of Theorem 1

In this section we sketch the proof of Theorem 1. The complete proof is deferred to Appendix A.

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 $S_1^{(k)} = \{x||\widehat{w}_{k-1} \cdot x| \le b_{k-1}\}$ be the acceptance region at the kth iteration. By definition, $D_k \subseteq S_1^{(k)}$. Let $w_k^* = \mathrm{argmax}_{w\in\mathcal{F}_k} \mathrm{err}_{0/1}(w|S_1^{(k)})$ 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 [BBZ07], it can be shown that under surrogate loss properties (A1) and (A2), with probability $\ge 1 - \delta$ the following holds:

$$\mathrm{err}_{0/1}(\widehat{w}_k) - \mathrm{err}_{0/1}(w_k^*) \le \beta_{k-1}\epsilon, \tag{4}$$

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 [BBZ07, BL13] 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 ($k \le k^*$) 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 $k \le k^*$. Formally speaking, we have the following two lemmas which are proved in Appendix A.

Lemma 1. *Suppose $r \in (0, 1/2)$ and $1/(1 + \log(1/r)) \le \alpha < 1$. With probability at least $1 - \delta$, under property (A1)*

$$\mathrm{err}_{0/1}(\widehat{w}_{k^*}) - \mathrm{err}_{0/1}(w_{k^*}^*) \le \beta_{k^*-1}\epsilon \le \frac{\epsilon^{1/\alpha}}{r^{\frac{1+\alpha}{\alpha}}\mu^{\frac{1-\alpha}{\alpha}}}. \tag{5}$$

Lemma 2. *With probability $\ge 1 - \delta E$, under properties (A1) and (A2), $w_k^* = w^*$ for all $k \le k^*$.*

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.

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

$$\mathrm{err}_{0/1}(\widehat{w}_E) - \mathrm{err}_{0/1}(\widehat{w}_{k^*}) \le \frac{r}{1-r}\beta_{k^*-1}\epsilon. \tag{6}$$

Combining Lemma 1,2 and 3 we can upper bound the excess risk $\mathrm{err}_{0/1}(\widehat{w}_E) - \mathrm{err}_{0/1}(w^*)$ 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.

3.4 Extension to log-concave densities

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 [BL13] and is deferred to Appendix A.2.

Theorem 2. *Fix $\delta \in (0, 1), r \in (0, 1/2), d \ge 4$ and $T \ge 4$. 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 $1/(1 + \log(1/r)) \le \alpha < 1$ 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 $1 - \delta$ the following holds:*

$$\mathrm{err}_{0/1}(\widehat{w}) - \mathrm{err}_{0/1}(w^*) = \widetilde{O}\left(\left(\frac{d + \log(1/\delta)}{T}\right)^{1/2\alpha}\right). \tag{7}$$

4 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 [BBZ07] achieve the minimax rate (up to polylogarithmic factors) under the uniform distribution on the unit ball.

[3] C_1 is an absolute constant. See Lemma 7, 8 in Appendix A.2 and Theorem 8 in [BL13] for details.

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:

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

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 [CN06, CN08, RS13a] 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.

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:*

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

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 $\mathcal{A}^{\mathrm{str}}_{d,T}$. We prove Corollary in Appendix B.

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:*

$$\inf_{A \in \mathcal{A}^{\mathrm{qs}}_{d,T}} \sup_{P_{Y|X} \in \mathcal{P}_{\alpha,\mu}} \mathbb{E}[\mathrm{err}_{0/1}(\widehat{w}) - \mathrm{err}_{0/1}(w^*)] = \Omega\left(\left(\frac{d}{T}\right)^{1/2\alpha}\right). \tag{10}$$

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

4.1 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. 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^*

[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 $\int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$ if $P \ll Q$ and $+\infty$ otherwise.

(rigorous mathematical definition of $P_{i,T}$ is given in the appendix). 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 [GS80] and is deferred to Appendix B.

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*

$$t \le \theta(w_i^*, w_j^*) \le 6.5t, \quad \forall i \ne j; \tag{11}$$

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

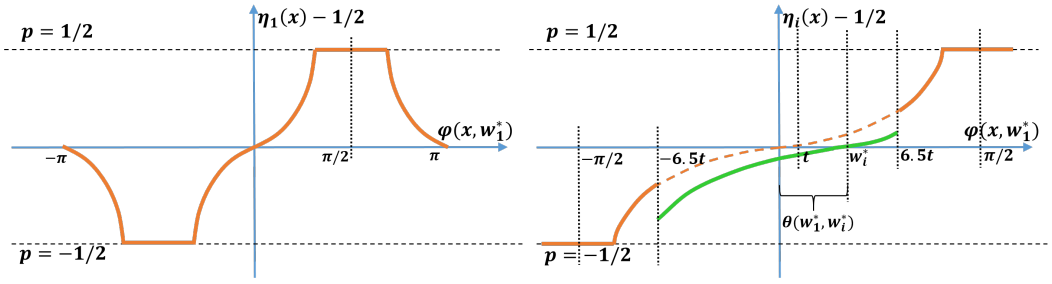
We next tackle the second objective of upper bounding $\mathrm{KL}(P_{i,T} \| P_{j,T})$. This requires designing label distributions $\{P_{Y|X}^{(i)}\}_{i=1}^m$ such that they satisfy the TNC condition in Eq. (8) while having small KL divergence between $P_{Y|X}^{(i)}$ and $P_{Y|X}^{(j)}$ for all distinct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

$$P_{Y|X}^{(i)}(Y = 1|X = x) = \begin{cases} \frac{1}{2} + \mathrm{sgn}(w_i^* \cdot x) \cdot \wp(|\varphi(w_i^*, x)|), & \text{if } |\varphi(w_1^*, x)| \le 6.5t; \\ \frac{1}{2} + \mathrm{sgn}(w_1^* \cdot x) \cdot \wp(|\varphi(w_1^*, x)|), & \text{if } |\varphi(w_1^*, x)| > 6.5t, \end{cases} \tag{12}$$

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

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

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



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

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 $d > 2$.

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 [CN08] 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.

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*

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

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 [TZ09]. A complete proof can be found in Appendix B.

5 Discussion and remarks

Comparison with noise-robust disagreement-based active learning algorithms In [Han11] another noise-robust adaptive learning algorithm was introduced. The algorithm is originally proposed in [DHM07] and is based on the concept of *disagreement coefficient* introduced in [Han07]. The algorithm adapts to different noise level α, and achieves an excess error rate of

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

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 [Han07] $\frac{\pi}{4}\sqrt{d} \leq \vartheta \leq \pi\sqrt{d}$. 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 [BL13]. 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 [JN14]. 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 $\alpha \in [0, 1]$,

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

Furthermore, if $\mu > 0$ we say the function f is *strongly convex*. In [JN14] 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 [RS13a] 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 [RS13b]. 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 [JN14, RS13a] 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 [RS13a].

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

Future work Finally, we mention two directions of future work. First, Algorithm 1 adapts to unknown level of label noise when given the query budget T. It is an interesting problem whether there exist adaptive active learning algorithms when the target error rate ϵ instead of query budget T is given. In addition, Algorithm 1 fails to handle the case when $\alpha = 0$. Therefore, another possible direction of future work would be 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 [BBZ07, BL13, CN08].

Appendix A Proof of Theorem 1

In this section we give complete proof of Theorem 1 in Section 3. We first prove Lemma 6 and Corollary 2. They analyze the statistical rate of error convergence for passive learning (i.e., per-iteration rate for Algorithm 1) under surrogate loss functions with properties (A1) and (A2). The results justify Eq. (18) in Section 3 and can be viewed as a slight generalization of Theorem 4 in [BBZ07].

Lemma 6 (Optimal passive learning)**.** *Fix k. Let $\widehat{w}_{k-1}$ be a linear classifier and b_{k-1} be a margin parameter. Suppose $D_k = \{(x_i, y_i)\}_{i=1}^n$ is a training data set of size n that satisfies $|\widehat{w}_{k-1} \cdot x_i| \le b_{k-1}$ for every data point x_i. For some $0 < \beta_{k-1} < \frac{\pi}{2}$, let $\mathcal{F}_k$ denote the set of all linear classifiers in $B_\theta(\widehat{w}_{k-1}, \beta_{k-1})$ and define $\widehat{w}_k = \operatorname{argmin}_{w \in \mathcal{F}_k} \widehat{\operatorname{err}}(w|D_k)$, $w_k^* = \operatorname{argmin}_{w \in \mathcal{F}_k} \operatorname{err}_{0/1}(w|S_1^{(k)})$, $\widetilde{w}_k^* = \operatorname{argmin}_{w \in \mathcal{F}_k} \operatorname{err}(w|S_1^{(k)})$, where $S_1^{(k)} = \{x||\widehat{w}_{k-1} \cdot x| \le b_{k-1}\}$ is the set of all points within a margin of b_{k-1} with respect to $\widehat{w}_{k-1}$. If $d \ge 4$ and $b_{k-1} \ge \gamma$ with*

$$\gamma := \frac{2 \sin \beta_{k-1}}{\sqrt{d}} \sqrt{\ln C + \ln(1 + \sqrt{\ln \max(1, \cot \beta_{k-1})})}$$

for some constant $C > 0$, then using (A1) and (A2), with probability at least $1 - \delta$,

$$\operatorname{err}_{0/1}(\widehat{w}_k) - \operatorname{err}_{0/1}(w_k^*) \le \epsilon' \cdot \frac{b_{k-1}\sqrt{d}}{2\sqrt{\pi}} + \frac{2 \sin \beta_{k-1}}{C \cdot \cos \beta_{k-1}}. \tag{17}$$

Proof. Define $S_1^{(k)} := \{x||\widehat{w}_{k-1} \cdot x| \le b_{k-1}\}$ and $S_2^{(k)} := \{x||\widehat{w}_{k-1} \cdot x| > b_{k-1}\}$. In Algorithm 1 $D_k \subseteq S_1^{(k)}$. Since $\theta(\widehat{w}_{k-1}, \widehat{w}_k), \theta(\widehat{w}_{k-1}, w_k^*) \le \beta_{k-1}$, by Lemma 7 in [BBZ07], we have

$$\begin{aligned} \Pr[(\widehat{w}_{k-1} \cdot x)(\widehat{w}_k \cdot x) < 0, x \in S_2^{(k)}] &\le \frac{\sin \beta_{k-1}}{C \cos \beta_{k-1}}, \\ \Pr[(\widehat{w}_{k-1} \cdot x)(w_k^* \cdot x) < 0, x \in S_2^{(k)}] &\le \frac{\sin \beta_{k-1}}{C \cos \beta_{k-1}}. \end{aligned}$$

Adding the two inequalities we get

$$\Pr[(\widehat{w}_k \cdot x)(w_k^* \cdot x) < 0, x \in S_2^{(k)}] \le \frac{2 \sin \beta_{k-1}}{C \cos \beta_{k-1}}.$$

Subsequently, using (A2), we have

$$\begin{aligned} &\operatorname{err}_{0/1}(\widehat{w}_k) - \operatorname{err}_{0/1}(w_k^*) \\ \le &(\operatorname{err}_{0/1}(\widehat{w}_k|S_1^{(k)}) - \operatorname{err}_{0/1}(w_k^*|S_1^{(k)})) \Pr[x \in S_1^{(k)}] + \frac{2 \sin \beta_{k-1}}{C \cos \beta_{k-1}} \\ \le &(\operatorname{err}(\widehat{w}_k|S_1^{(k)}) - \operatorname{err}(\widetilde{w}_k^*|S_1^{(k)})) \Pr[x \in S_1^{(k)}] + \frac{2 \sin \beta_{k-1}}{C \cos \beta_{k-1}} \end{aligned}$$

Lemma 4 in [BBZ07] tells us that $\Pr[x \in S_1^{(k)}] \le \frac{b_{k-1}\sqrt{d}}{2\sqrt{\pi}}$. And using (A1) completes the proof.

□

Corollary 2. *Fix k. Let T be the number of samples obtained and E be the number of iterations. Suppose $d \geq 4$, $T \geq 4$, $E = \frac{1}{2}\log T$, $n = T/E$. Assume also that $\beta_k = r^k\pi$ and $b_{k-1} = \frac{2\beta_{k-1}}{\sqrt{d}}\sqrt{E(1+\log(1/r))}$ for some constant $r \in (0, 1/2)$. With probability at least $1-\delta$,*

$$\mathrm{err}_{0/1}(\widehat{w}_k) - \mathrm{err}_{0/1}(w_k^*) \leq \beta_{k-1} \cdot \left[\epsilon'\sqrt{E(1+\log(1/r))} + 2\sqrt{\frac{2}{T}}\right] =: \beta_{k-1}\epsilon. \tag{18}$$

Proof. We first note that Eq. (18) is trivially true for $k = 1$ because $\beta_0 = \pi > 1$, $\sqrt{E(1+\log(1/r))} \geq 1$ and $\mathrm{err}_{0/1}(\widehat{w}_1) - \mathrm{err}_{0/1}(w_1^*) \leq \mathrm{err}(\widehat{w}_1) - \mathrm{err}(\widetilde{w}_1^*) \leq \epsilon'$ due to properties (A1) and (A2). In the remaining part of the proof we will assume that $k \geq 2$, and hence $\beta_{k-1} \in (0, \pi/2)$.

Before continuing the proof, we first give a list of elementary inequalities that will be used in this proof (assuming $\beta < \pi/2$):

$$\ln(1+x) \leq x; \quad \sin\beta \leq \beta; \quad \cot\beta \leq 1/\beta; \quad \tan\beta \leq \sqrt{2}\beta.$$

Put $C = 2^E = \sqrt{T}$. Let's first check $b_{k-1} \geq \gamma$:

$$\begin{aligned}
\gamma &= \frac{2\sin\beta_{k-1}}{\sqrt{d}}\sqrt{\ln C + \ln(1+\sqrt{\ln\max(1,\cot\beta_{k-1})})} \\
&\leq \frac{2\beta_{k-1}}{\sqrt{d}}\sqrt{\ln C + \sqrt{\ln\max(1,1/\beta_{k-1})}} \\
&\leq \frac{2\beta_{k-1}}{\sqrt{d}}\sqrt{E + E\log(1/r)} =: b_{k-1}.
\end{aligned}$$

Applying Lemma 6, we have

$$\begin{aligned}
&\mathrm{err}_{0/1}(\widehat{w}_k) - \mathrm{err}_{0/1}(w_k^*) \\
\leq\ & \epsilon'\frac{\beta_{k-1}}{\sqrt{\pi}}\sqrt{E(1+\log(1/r))} + \frac{2\sqrt{2}\beta_{k-1}}{\sqrt{T}} \\
\leq\ & \beta_{k-1}\cdot\left[\epsilon'\sqrt{E(1+\log(1/r))} + 2\sqrt{\frac{2}{T}}\right] = \beta_{k-1}\epsilon.
\end{aligned}$$

□

With Lemma 6 and Corollary 2 we are ready to prove the main theorem. The key component of the proof is a case analysis on the TNC parameter μ. By Eq. (1) and the fact that $0 < \alpha < 1$, μ must satisfy $\mu \leq \pi^{-1/(1-\alpha)} \leq \pi^{-\alpha/(1-\alpha)}$. Since $\alpha \geq 1/(1+\log(1/r))$, we have the following upper bound for μ:

$$\mu \leq \frac{2^{(\log(1/r)\cdot\frac{\alpha}{1-\alpha}-1)E}\cdot r^{-1}}{\pi^{\frac{\alpha}{1-\alpha}}} = \frac{2^{-E}r^{-1}}{(r^E\pi)^{\alpha/(1-\alpha)}} = \frac{2^{-E}r^{-1}}{\beta_E^{\alpha/(1-\alpha)}} \leq \frac{r^{-1}\epsilon}{\beta_E^{\alpha/(1-\alpha)}}. \tag{19}$$

The last step holds due to $\epsilon \geq 1/\sqrt{T} = 2^{-E}$.

When μ is sufficiently small, that is,

$$\mu \leq \frac{r^{-1}\epsilon}{\beta_1^{\alpha/(1-\alpha)}}, \tag{20}$$

we are done after the first iteration because by (A1),

$$\mathrm{err}_{0/1}(\widehat{w}_1) - \mathrm{err}_{0/1}(w_1^*) \leq \beta_0\epsilon \leq r^{-1}\beta_1\epsilon \leq r^{-1}\left(\frac{\epsilon}{r\mu}\right)^{(1-\alpha)/\alpha}\epsilon = \frac{\epsilon^{1/\alpha}}{r^{1/\alpha}\mu^{(1-\alpha)/\alpha}} \tag{21}$$

and with probability $\geq 1 - \delta E$ we have $w_1^* = w^*$ and using arguments akin to Lemma 3 in Section 3, we have [6]

$$\mathrm{err}_{0/1}(\widehat{w}_E) - \mathrm{err}_{0/1}(\widehat{w}_1) \leq \frac{r}{1-r}\pi\epsilon. \tag{22}$$

Combining Eq. (21) and (22) we prove the theorem. Next we consider the case when Eq. (20) does not hold. In this case, by Eq. (19) there must exists $k^* \in \{1, 2, \cdots, E-1\}$ such that

$$\frac{r^{-1}\epsilon}{\beta_{k^*}^{\alpha/(1-\alpha)}} \leq \mu \leq \frac{r^{-1}\epsilon}{\beta_{k^*+1}^{\alpha/(1-\alpha)}}. \tag{23}$$

The k^* in Eq. (23) defines the "tipping-point" of Algorithm 1 and is crucial to the three technical lemmas presented in Section 3.3. Combining Lemma 1, 2 and 3 we get

$$\begin{aligned}
& \mathrm{err}_{0/1}(\widehat{w}_E) - \mathrm{err}_{0/1}(w^*) \\
= \;& \mathrm{err}_{0/1}(\widehat{w}_E) - \mathrm{err}_{0/1}(\widehat{w}_{k^*}) + \mathrm{err}_{0/1}(\widehat{w}_{k^*}) - \mathrm{err}_{0/1}(w^*) \\
= \;& \mathrm{err}_{0/1}(\widehat{w}_E) - \mathrm{err}_{0/1}(\widehat{w}_{k^*}) + \mathrm{err}_{0/1}(\widehat{w}_{k^*}) - \mathrm{err}_{0/1}(w_{k^*}^*) \\
\leq \;& \frac{r}{1-r}\beta_{k^*-1}\epsilon + \beta_{k^*-1}\epsilon \\
= \;& \frac{1}{1-r}\beta_{k^*-1}\epsilon \leq \frac{\epsilon^{1/\alpha}}{r^{\frac{1+\alpha}{\alpha}}(1-r)\mu^{\frac{1-\alpha}{\alpha}}};
\end{aligned}$$

thus, completing proof of the main theorem.

A.1 Proofs of technical lemmas in Section 3.3

Proof of Lemma 1. We know that

$$\beta_{k^*-1} = r^{-2}\beta_{k^*+1} \leq r^{-2}\left(\frac{\epsilon}{r\mu}\right)^{\frac{1-\alpha}{\alpha}}.$$

Using (A1), we get

$$\mathrm{err}_{0/1}(\widehat{w}_{k^*}) - \mathrm{err}_{0/1}(w_{k^*}^*) \leq \beta_{k^*-1}\epsilon \leq r^{-2}\left(\frac{\epsilon}{r\mu}\right)^{\frac{1-\alpha}{\alpha}}\epsilon = \frac{\epsilon^{1/\alpha}}{r^{\frac{1+\alpha}{\alpha}}\mu^{\frac{1-\alpha}{\alpha}}}.$$

□

Proof of Lemma 2. Note that in order to prove $w_k^* = w^*$ we only need to show $\theta(w^*, \widehat{w}_{k-1}) \leq \beta_{k-1}$ because the global optimal classifier w^* is the Bayes estimator and is hence optimal for any region $S \subseteq X$.

We now use induction to prove this lemma. When $k = 1$ the claim $w_1^* = w^*$ is clearly true because $\beta_0 = \pi$. Now assume the claim is true for k, that is, $\theta(\widehat{w}_{k-1}, w^*) \leq \beta_{k-1}$ and $w_k^* = w^*$. We want to prove $\theta(\widehat{w}_k, w^*) \leq \beta_k$, i.e., $w_{k+1}^* = w^*$.

To see this, we apply the TNC condition in Eq. (1) to get

$$\theta(\widehat{w}_k, w^*) \leq \left(\frac{\mathrm{err}_{0/1}(\widehat{w}_k) - \mathrm{err}_{0/1}(w^*)}{\mu}\right)^{1-\alpha}$$

Therefore, we only need to prove $\mu \geq (\mathrm{err}_{0/1}(\widehat{w}_k) - \mathrm{err}_{0/1}(w^*))/\beta_k^{1/(1-\alpha)}$.

[6] Note that in this case $\mu < r^{-1}\epsilon/\beta_1^{\alpha/(1-\alpha)}$ for all k.

Because of the induction assumption $\theta(\widehat{w}_{k-1}, w^*) \leq \beta_{k-1}$, w^* is exactly w_k^*. Applying Corollary 2 we know that with probability at least $1-\delta$,

$$\frac{\mathrm{err}_{0/1}(\widehat{w}_k) - \mathrm{err}_{0/1}(w^*)}{\beta_k^{1/(1-\alpha)}} \leq \frac{\beta_{k-1}\epsilon}{\beta_k^{1/(1-\alpha)}} = \frac{r^{-1}\epsilon}{\beta_k^{\alpha/(1-\alpha)}} \leq \frac{r^{-1}\epsilon}{\beta_{k^*}^{\alpha/(1-\alpha)}}.$$

The last expression is $\leq \mu$ by applying Eq. (23), and hence $\theta(\widehat{w}_k, w^*) \leq \beta_k$. Finally, taking a union bound over all $k \leq k^*$ we complete the proof.

□

Proof of Lemma 3. First, decompose the term $\mathrm{err}_{0/1}(\widehat{w}_E) - \mathrm{err}_{0/1}(\widehat{w}_{k^*})$ as follows:

$$\begin{aligned}
&\mathrm{err}_{0/1}(\widehat{w}_E) - \mathrm{err}_{0/1}(\widehat{w}_{k^*}) \\
&= \sum_{k=k^*}^{E-1} \mathrm{err}_{0/1}(\widehat{w}_{k+1}) - \mathrm{err}_{0/1}(\widehat{w}_k) \\
&= \sum_{k=k^*}^{E-1} (\mathrm{err}_{0/1}(\widehat{w}_{k+1}|S_1^{(k)}) - \mathrm{err}_{0/1}(\widehat{w}_k|S_1^{(k)}))\Pr[x \in S_1^{(k)}] \\
&\quad + (\mathrm{err}_{0/1}(\widehat{w}_{k+1}|S_2^{(k)}) - \mathrm{err}_{0/1}(\widehat{w}_k|S_2^{(k)}))\Pr[x \in S_2^{(k)}] \\
&\leq \sum_{k=k^*}^{E-1} (\mathrm{err}_{0/1}(\widehat{w}_{k+1}|S_1^{(k)}) - \mathrm{err}_{0/1}(w_{k+1}^*|S_1^{(k)}))\Pr[x \in S_1^{(k)}] \\
&\quad + \Pr[(\widehat{w}_k \cdot x)(\widehat{w}_{k+1} \cdot x) < 0, x \in S_2^{(k)}],
\end{aligned}$$

where $S_1^{(k)} = \{x||\widehat{w}_k \cdot x| \leq b_k\}$ and $S_2^{(k)} = \{x||\widehat{w}_k \cdot x| \geq b_k\}$. The last inequality is due to the fact that w_{k+1}^* is the optimal classifier with respect to the conditional distribution on $S_1^{(k)}$. Next, taking a union bound over all rounds $k^* \leq k \leq E$ and using arguments akin to Lemma 6 and Corollary 2, we have with probability at least $1 - \delta E$

$$\mathrm{err}_{0/1}(\widehat{w}_E) - \mathrm{err}_{0/1}(\widehat{w}_{k^*}) \leq \sum_{k=k^*}^{E-1} \beta_k \epsilon \leq \beta_{k^*-1}\epsilon(r + r^2 + \cdots) = \frac{r}{1-r}\beta_{k^*-1}\epsilon.$$

□

A.2 Extension to log-concave data distributions

In this section prove Theorem 2. In particular, we show that Algorithm 1 works if the data distribution P_X is isotropic log-concave and the algorithm parameters are chosen as $\beta_k = r^k\pi$ and $b_{k-1} = C_1\beta_{k-1}\log T$ for some absolute constant C_1.

Let $g : \mathbb{R}^d \to (0, +\infty)$ be the pdf of the underlying data distribution P_X. We say P_X is a *log-concave* distribution if $\log g$ is a concave function. Furthermore, P_X is *isotropic* if its mean is zero and its covariance matrix is the identity.

For isotropic log-concave densities, we cite the following lemma from [BL13].

Lemma 7 (Theorem 21, [BL13])**.** *There exist absolute constants $C_1, C_2, C_3 > 0$ such that the following holds. Let u and v be two unit vectors in $\mathbb{R}^d$ and assume $\theta(u, v) = \eta < \pi/2$. Assume that D is isotropic log-concave, then for any $b \geq C_1\eta$ we have*

$$\Pr_{x\sim D}[(u \cdot x)(v \cdot x) < 0, |v \cdot x| \geq b] \leq C_2\eta\exp(-C_3 b/\eta). \tag{24}$$

Using Lemma 7 we can derive a result similar to Corollary 2, as shown below.

Lemma 8. *Fix k and let C_1, C_2, C_3 be absolute constants in Lemma 7. Let T be the number of samples obtained and E be the number of iterations. Suppose $d \geq 2$, $T \geq 4$, $E = \frac{1}{2}\log T$, $n = T/E$. Assume also that $\beta_k = r^k\pi$ for some constant $r \in (0, 1/2)$ and $b_{k-1} = C_1\beta_{k-1}\log T$ for constant C_1 in Lemma 7. Suppose $D_k = \{(x_i, y_i)\}_{i=1}^n$ is a training data set of size n and (x_i, y_i) are i.i.d. sampled from P_X. In addition, all (x_i, y_i) in D_k satisfy $|\widehat{w}_{k-1} \cdot x_i| \leq b_{k-1}$ for some margin classifier $\|\widehat{w}_{k-1}\| = 1$. If P_X is isotropic log-concave, then with probability at least $1 - \delta$ the following holds:*

$$\mathrm{err}_{0/1}(\widehat{w}_k) - \mathrm{err}_{0/1}(w_k^*) \leq \beta_{k-1}\left(\epsilon' \cdot 2C_1 \log T + 2C_2 T^{-C_1C_3}\right) =: \beta_{k-1}\epsilon, \tag{25}$$

where $\widehat{w}_{k-1} = \mathrm{argmin}_{w\in\mathcal{F}_k}\widehat{\mathrm{err}}(w|D_k)$, $w_k^* = \mathrm{argmin}_{w\in\mathcal{F}_k}\mathrm{err}_{0/1}(w|S_1^{(k)})$, $\mathcal{F}_k = B_\theta(\widehat{w}_{k-1}, \beta_{k-1})$, $S_1^{(k)} = \{x||\widehat{w}_{k-1} \cdot x| \leq b_{k-1}\}$.

Proof. Define $S_1^{(k)} := \{x||\widehat{w}_{k-1} \cdot x| \leq b_{k-1}\}$ and $S_2^{(k)} := \{x||\widehat{w}_{k-1} \cdot x| > b_{k-1}\}$. By Algorithm 1 $D_k \subseteq S_1^{(k)}$. Since $\theta(\widehat{w}_{k-1}, \widehat{w}_k), \theta(\widehat{w}_{k-1}, w_k^*) \leq \beta_{k-1}$, by Lemma 7 we have

$$\begin{aligned}
\Pr[(\widehat{w}_{k-1} \cdot x)(\widehat{w}_k \cdot x) < 0, x \in S_2^{(k)}] &\leq C_2\beta_{k-1}T^{-C_1C_3}, \\
\Pr[(\widehat{w}_{k-1} \cdot x)(w_k^* \cdot x) < 0, x \in S_2^{(k)}] &\leq C_2\beta_{k-1}T^{-C_1C_3}.
\end{aligned}$$

Adding the two inequalities we get

$$\Pr[(\widehat{w}_k \cdot x)(w_k^* \cdot x) < 0, x \in S_2^{(k)}] \leq 2C_2\beta_{k-1}T^{-C_1C_3}.$$

Next, let $\widetilde{w}_k^*$ denote the minimizer of the true surrogate loss error err on S_1 over all linear classifiers with $w \in B_\theta(\widehat{w}_{k-1}, \beta_{k-1})$. Using (A2) we get

$$\begin{aligned}
& \mathrm{err}_{0/1}(\widehat{w}_k) - \mathrm{err}_{0/1}(w_k^*) \\
\leq\ & (\mathrm{err}_{0/1}(\widehat{w}_k|S_1^{(k)}) - \mathrm{err}_{0/1}(w_k^*|S_1^{(k)}))\Pr[x \in S_1^{(k)}] + 2C_2\beta_{k-1}T^{-C_1C_3} \\
\leq\ & (\mathrm{err}(\widehat{w}_k|S_1^{(k)}) - \mathrm{err}(\widetilde{w}_k^*|S_1^{(k)}))\Pr[x \in S_1^{(k)}] + 2C_2\beta_{k-1}T^{-C_1C_3} \\
\leq\ & \epsilon' \cdot 2b_k + 2C_2\beta_{k-1}T^{-C_1C_3}.
\end{aligned}$$

The last inequality is due to (A1) and Lemma 2 from [BL13].

□

Proof of Theorem 2. The proof is the same as the proof of Theorem 1 by noting that $\epsilon = \epsilon' \cdot O(\log T)$ is of the order $\widetilde{O}(\sqrt{d/T})$. In particular, Lemma 1,2 and 3 do not require the uniform distribution assumption. □

Appendix B Proof of Theorem 3

In this section we provide a complete proof of Theorem 3. We first prove the main technical lemmas (Lemma 4 and 5) in Section 4.1.

Proof of Lemma 4. We first present the following lemma, which will be proved in Appendix B.1.

Lemma 9. *Assume d is even. Let S be the largest subset of $\{0, 1\}^d$ that satisfies the following conditions:*

1. *$\forall x \in S$, $\|x\|_1 = d/2 =: \omega$.*
2. *$\forall x, x' \in S$, $x \neq x'$, $\Delta_H(x, x') \geq d/16 =: \delta_H$.*

Here $\Delta_H(x,y) = \sum_i |x_i - y_i|$ denotes the Hamming distance between x and y. Then for $d \geq 2$, the following lower bound on the size of S holds:

$$\log |S| \geq 0.0625d. \tag{26}$$

Using Lemma 9 we can construct a well-separated hypothesis set that satisfies Eq. (11) as follows. Suppose $S = \{z_1, \cdots, z_m\} \subseteq \{0,1\}^d$ is the largest subset of $\{0,1\}^d$ that satisfies the two conditions in Lemma 9. Set $a = 4t$. For each $i \in \{1, 2, \cdots, m\}$ define $w_i^* \in \mathbb{R}^d$ as

$$w_i^* = \frac{1}{Z}\left((1, 1, \cdots, 1) - az_i\right), \tag{27}$$

where $Z = \sqrt{d\left(1 - a + \frac{a^2}{2}\right)}$ is a normalization constant to make $\|w_i^*\| = 1$. Note that Z can be upper and lower bounded by $\frac{1}{2}d \leq Z^2 \leq d$ because $\|z_i\|_1 = \frac{d}{2}$. Define $\mathcal{W} = \{w_1^*, \cdots, w_m^*\}$. Since $|\mathcal{W}| = |S|$, by Lemma 9 we have $\log |\mathcal{W}| \geq 0.0625d$ for $d \geq 2$.

Next we prove that $\mathcal{W}$ is a well-separated hypothesis set that satisfies Eq. (11). That is, for every pair $w_i^*, w_j^* \in \mathcal{W}, i \neq j$ one has $t \leq \theta(w_i^*, w_j^*) \leq 6.5t$. First consider the lower bound. A key observation is that two mismatched entries in z_i and z_j reduces the inner product between w_i^* and w_j^* by a^2, as shown in Eq. (28). Note that such "switching" is always possible because both z_i and z_j have the same number of ones and hence the Hamming distance $\Delta_H(z_i, z_j)$ is even.

$$2 - 2a + a^2 = \begin{pmatrix} 1-a & & 1 \\ \times & + & \times \\ 1-a & & 1 \end{pmatrix} \Longrightarrow \begin{pmatrix} 1-a & & 1 \\ \times & + & \times \\ 1 & & 1-a \end{pmatrix} = 2 - 2a. \tag{28}$$

By definition of w_i^* and w_j^* we have

$$\cos\theta(w_i^*, w_j^*) = \langle w_i, w_j \rangle = \frac{d\left(1 - a + \frac{a^2}{2}\right) - \frac{\Delta_H(z_i, z_j)}{2}a^2}{d\left(1 - a + \frac{a^2}{2}\right)} \leq 1 - \frac{a^2}{32}. \tag{29}$$

Subsequently, using the fact that $\cos\theta \geq 1 - \frac{\theta^2}{2}$ we have

$$\theta(w_i^*, w_j^*) \geq \sqrt{2(1 - \cos\theta(w_i^*, w_j^*))} \geq \frac{a}{4} = t. \tag{30}$$

On the other hand, to obtain an upper bound on $\theta(w_i^*, w_j^*)$, we note the following lower bound on $\cos\theta(w_i^*, w_j^*)$:

$$\cos\theta(w_i^*, w_j^*) = \langle w_i, w_j \rangle = \frac{d\left(1 - a + \frac{a^2}{2}\right) - \frac{\Delta_H(z_i, z_j)}{2}a^2}{d\left(1 - a + \frac{a^2}{2}\right)} \geq 1 - \frac{a^2/2}{1/2} = 1 - a^2. \tag{31}$$

The second to last inequality is due to the fact that $\Delta_H(z_i, z_j) \leq d$ and $Z^2 \geq \frac{d}{2}$. By Taylor expansion one has $\cos\theta \leq 1 - \frac{\theta^2}{2} + \frac{\theta^4}{24} \leq 1 - \frac{\theta^2}{2} + \theta^2\frac{\theta^2}{24}$. Since $\langle w_i, w_j \rangle > 0$ the angle θ must be smaller than $\frac{\pi}{2}$. Consequently, $\cos\theta \leq 1 - \frac{7}{18}\theta^2$ and therefore we have the following upper bound on $\theta(w_i^*, w_j^*)$:

$$\theta(w_i^*, w_j^*) \leq \sqrt{\frac{18}{7}(1 - \cos(\theta(w_i^*, w_j^*)))} \leq \sqrt{\frac{18}{7}}a \leq 6.5t. \tag{32}$$

□

Proof of Lemma 5. We first prove that w_i^* is the Bayes classifier for $P_{Y|X}^{(i)}$ and furthermore $P_{Y|X}^{(i)}$ satisfies the TNC condition in Eq. (8) with respect to w_i^*. note that $\eta(x) = \frac{1}{2}$ if and only if $|\varphi(x, w_1^*)| \leq 6.5t$ and $\varphi(x, w_i^*) = 0$. Therefore, w_i^* is the Bayes classifier for $P_{Y|X}^{(i)}$. We also note that if $|\varphi(x, w_1^*)| \geq 6.5t$ then $\operatorname{sgn}(\varphi(x, w_1^*)) = \operatorname{sgn}(\varphi(x, w_i^*))$ because $\theta(w_1^*, w_i^*) \leq 6.5t$. Without loss of generality, we assume in the remainder of the proof that $\varphi(x, w_1^*), \varphi(x, w_i^*) \geq 0$ when $|\varphi(x, w_1^*)| > 6.5t$.

Next we prove that $P_{Y|X}^{(i)}$ satisfies the TNC condition with respect to w_i^*. Consider a data point x with $\varphi(x, w_1^*) = \vartheta$ and $\varphi(x, w_i^*) = \vartheta'$. By definition, if $|\vartheta| \leq 6.5t$ then $P_{Y|X}^{(i)}$ trivially satisfies TNC because $P_{Y|X}^{(i)}(Y = 1|X = x)$ only depends on ϑ'. For $|\vartheta| > 6.5t$ and $\vartheta, \vartheta' > 0$, note that $0.5\vartheta' \leq \vartheta$ because $|\vartheta - \vartheta'| \leq 6.5t$ and hence $\vartheta' \leq \vartheta + 6.5t < 2\vartheta$. Consequently, $P_{Y|X}^{(i)}$ satisfies

$$P_{Y|X}^{(i)}(Y = 1|X = x) - \frac{1}{2} = \min\{1/2, 2^{\frac{\alpha}{1-\alpha}}\mu_0\vartheta^{\frac{\alpha}{1-\alpha}}\} \geq \min\{1/2, \mu_0\vartheta'^{\frac{\alpha}{1-\alpha}}\} \geq \mu_0\vartheta'^{\frac{\alpha}{1-\alpha}}. \tag{33}$$

The last inequality holds due to the fact that Eq. (8) holds for all $\vartheta \in [0, \pi]$. Therefore, $P_{Y|X}^{(i)}$ satisfies the TNC lower bound for $|\vartheta| > 6.5t$ with respect to w_i^*.

Lastly, we prove the upper bound on $\mathrm{KL}(P_{i,T}\|P_{j,T})$ in Eq. (14). Here $P_{i,T}$ represents the data/label distribution for an active learning algorithm that is allowed for T label queries under $P_{Y|X}^{(i)}$. Mathematically, $P_{i,T}$ is a distribution over $(\mathcal{X} \times \mathcal{Y})^T$ and can be decomposed as

$$\begin{aligned}
P_{i,T}(x_1, y_1, \cdots, x_T, y_T) &= P^{(i)}_{X_1,Y_1,\cdots,X_T,Y_T}(x_1, y_1, \cdots, x_T, y_T) \\
&= \prod_{t=1}^{T} P^{(i)}_{Y_t|X_t}(y_t|x_t) P_{X_t|X_1,Y_1,\cdots,X_{t-1},Y_{t-1}}(x_t|x_1, y_1, \cdots, x_{t-1}, y_{t-1})
\end{aligned}$$

because a query synthetic algorithm $A \in \mathcal{A}^{\mathrm{qs}}_{d,T}$ proposes data points and requests labels in a sequential, feedback-driven manner. Note that $P_{X_t|X_1,Y_1,\cdots,X_{t-1},Y_{t-1}}$ does not depend on i because the underlying label distribution $P_{Y|X}^{(i)}$ is unknown to A. Subsequently, we establish an upper bound on $\mathrm{KL}(P_{i,T}\|P_{j,T})$ following analysis in [CN08]: ($\mathbb{E}_i$ and $\mathbb{E}_j$ denote the expectation taking with respect to $P_{i,T}$ and $P_{j,T}$)

$$\begin{aligned}
\mathrm{KL}(P_{i,T}\|P_{j,T}) &= \mathbb{E}_i\left[\log \frac{P^{(i)}_{X_1,Y_1,\cdots,X_T,Y_T}(x_1, y_1, \cdots, x_T, y_T)}{P^{(j)}_{X_1,Y_1,\cdots,X_T,Y_T}(x_1, y_1, \cdots, x_T, y_T)}\right] \\
&= \mathbb{E}_i\left[\log \frac{\prod_{t=1}^{T} P^{(i)}_{Y_t|X_t}(y_t|x_t) P_{X_t|X_1,Y_1,\cdots,X_{t-1},Y_{t-1}}(x_t|x_1, y_1, \cdots, x_{t-1}, y_{t-1})}{\prod_{t=1}^{T} P^{(j)}_{Y_t|X_t}(y_t|x_t) P_{X_t|X_1,Y_1,\cdots,X_{t-1},Y_{t-1}}(x_t|x_1, y_1, \cdots, x_{t-1}, y_{t-1})}\right] \\
&= \mathbb{E}_i\left[\log \frac{\prod_{t=1}^{T} P^{(i)}_{Y_t|X_t}(y_t|x_t)}{\prod_{t=1}^{T} P^{(j)}_{Y_t|X_t}(y_t|x_t)}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_i\left[\mathbb{E}_i\left[\log \frac{P^{(i)}_{Y|X}(y_t|x_t)}{P^{(j)}_{Y|X}(y_t|x_t)}\middle| X_1 = x_1, \cdots, X_T = x_T\right]\right] \\
&\leq T \cdot \sup_{x \in \mathcal{X}} \mathrm{KL}(P^{(i)}_{Y|X}(\cdot|x)\|P^{(j)}_{Y|X}(\cdot|x)).
\end{aligned} \tag{34}$$

Eq. (34) shows that the upper bound on $\mathrm{KL}(P_{i,T}\|P_{j,T})$ does not depend on which active learning algorithm is used, though both $P_{i,T}$ and $P_{j,T}$ are defined in an algorithm-dependent way. Note that both $P_{Y|X}^{(i)}(\cdot|x)$ and $P_{Y|X}^{(j)}(\cdot|x)$ are Bernoulli random variables. To bound their KL divergence we cite the following result from [CN08]:

Lemma 10 ([CN08], Lemma 1)**.** *Let P and Q be Bernoulli random variables with parameters $1/2-p$ and $1/2-q$, respectively. If $|p|,|q| \le 1/4$ then $\mathrm{KL}(P\|Q) \le 8(p-q)^2$.*

Fix $x \in \mathcal{X} = \mathcal{S}^d$ and let $\vartheta_i = \varphi(x, w_i^*)$, $\vartheta_j = \varphi(x, w_j^*)$ and $\vartheta_1 = \varphi(x, w_1^*)$. Suppose $P_{Y|X}^{(i)}(\cdot|x) = 1/2 - p_i$ and $P_{Y|X}^{(j)}(\cdot|x) = 1/2 - p_j$. A simple case study on ϑ_1 reveals that $|p_i - p_j| = 0$ when $|\vartheta_1| > 6.5t$ and

$$\left|p_i - p_j\right| \le c \cdot \left(\left|\vartheta_i\right|^{\alpha/(1-\alpha)} + \left|\vartheta_j\right|^{\alpha/(1-\alpha)}\right) \le C \cdot t^{\alpha/(1-\alpha)} \tag{35}$$

when $|\vartheta_1| \le 6.5t$ (which implies $|\vartheta_i|, |\vartheta_j| \le 13t$). Here $c = 2^{\alpha/(1-\alpha)}\mu_0$ and $C = 2 \times 13^{\alpha/(1-\alpha)}c$ are constants that do not depend on t or T. Furthermore, for sufficiently small t one has $|p_i|, |p_j| \le 1/4$. Therefore, by Lemma 10 the following holds:

$$\mathrm{KL}(P_{i,T}\|P_{j,T}) \le T \cdot \sup_{x \in \mathcal{S}^d} \mathrm{KL}(P_{Y|X}^{(i)}(\cdot|x)\|P_{Y|X}^{(j)}(\cdot|x)) \le 8C^2 \cdot Tt^{2\alpha/(1-\alpha)}. \tag{36}$$

□

To prove Theorem 3, we cite the following information-theoretical lower bound from [TZ09].

Theorem 4 ([TZ09])**.** *Let $\mathcal{F}$ be a set of models and $\mathcal{F}_0 = \{f_1, \cdots, f_M\} \subseteq \mathcal{F}$ be a finite subset of $\mathcal{F}$. We have a probability measure P_f defined on a common probability space associated with each model $f \in \mathcal{F}$. Let $D : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ be a collection of semi-distances. If there exist constants $\rho > 0$ and $0 < \gamma < 1/8$ such that the following holds:*

1. *$D(f_j, f_k) \ge 2\rho > 0$ for every $j, k \in \{1, \cdots, M\}$, $j \ne k$.*
2. *$P_{f_j} \ll P_{f_1}$ for every $j \in \{1, \cdots, M\}$.* [7]
3. *$\frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(P_{f_j}\|P_{f_0}) \le \gamma \log M$.*

Then the following bound holds:

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}} P_f\left(D(\widehat{f}, f) \ge \rho\right) \ge \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\gamma - 2\sqrt{\frac{\gamma}{\log M}}\right). \tag{37}$$

With Lemma 4, 5 and Theorem 4 we can prove Theorem 3 and Corollary 1 easily.

Proof of Theorem 3. We take $\mathcal{F}_0 = \{P_{Y|X}^{(1)}, \cdots, P_{Y|X}^{(M)}\}$ and $D(w, w') = \theta(w, w')$ in Theorem 4. By Lemma 4, $D(w_i^*, w_j^*) \ge t$ and $\log M = \log|\mathcal{W}| \ge 0.0625d$ for $d \ge 2$. In addition, Lemma 5 yields $\gamma \log M = 8C^2 \cdot Tt^{2\alpha/(1-\alpha)}$. Put $t = \kappa \cdot (d/T)^{(1-\alpha)/(2\alpha)}$ for some sufficiently small constant $\kappa > 0$. We then have $\gamma = O(1)$ and $\sqrt{\gamma/\log M} = O(1)$. Consequently,

$$\inf_{A \in \mathcal{A}_{d,T}^{\mathrm{qs}}} \sup_{P_{Y|X} \in \mathcal{P}_{\alpha,\mu_0}} \Pr\left(\theta(\widehat{w}, w^*) \ge \frac{t}{2}\right) = \Omega(1).$$

Finally, applying Markov's inequality we obtain

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

□

[7] For two distributions P and Q, $P \ll Q$ means the support of P is contained in the support of Q.

Proof of Corollary 1. Suppose the density g associated with P_X is bounded from below with parameter $\gamma \in (0,1]$ (see Proposition 2). Let $\mathcal{Q}_{\alpha,\mu_0}$ denote the class of all conditional distributions $P_{Y|X}$ that satisfy Eq. (8) with parameter $\mu_0 = \mu/(2(1-\alpha)\gamma)$. Note that μ_0 does not depend on d or T when γ is a constant. By Proposition 2, $\mathcal{Q}_{\alpha,\mu_0} \subseteq \mathcal{P}_{\alpha,\mu}$. We then have

$$\inf_{A \in \mathcal{A}_{d,T}^{\mathrm{qs}}} \sup_{P_{Y|X} \in \mathcal{P}_{\alpha,\mu}} \mathbb{E}[\mathrm{err}_{0/1}(\widehat{w}) - \mathrm{err}_{0/1}(w^*)]$$

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

The first inequality is by Eq. (1). The second one is due to Jensen's inequality and Theorem 3. □

B.1 Proof of technical lemmas in Appendix B

In this section we prove Lemma 9 used in the proof of Lemma 4 in Appendix B. We first cite a result from [GS80] concerning the size of the largest separable set of constant-weight codes.

Lemma 11 (Theorem 7, [GS80])**.** *Fix d and $\omega \leq d$. Let $X_{d,\omega}$ denote all ω-weight binary codes of length d, that is, $X_{d,\omega} = \{x \in \{0,1\}^d | \sum_{i=1}^d x_i = \omega\}$. For any $\delta_H \leq d$, there exists a subset S of $X_{d,\omega}$ such that*

$$\Delta_H(x,x') \geq \delta_H, \quad \forall x, x' \in S, x \neq x' \tag{38}$$

and

$$|S| \geq \frac{\binom{d}{\omega}}{\sum_{i=0}^{\delta_H/2} \binom{\omega}{i}\binom{d-\omega}{i}}. \tag{39}$$

Here $\Delta_H(\cdot,\cdot)$ denotes the Hamming distance.

We are now ready to prove Lemma 9.

Proof of Lemma 9. We first comment that it is always possible to select $x_1, x_2 \in \{0,1\}^d$ with $\|x_1\|_1 = \|x_2\|_1 = \Delta_H(x_1,x_2)/2 = d/2$ whenever $d \geq 2$ and d is even. Consequently, for $2 \leq d < 16$ the bound $\log|S| \geq 0.0625d$ always holds. In the remainder of the proof we shall focus on the case when $d \geq 16$.

The following lower and upper bounds for binomial coefficient $\binom{n}{k}$ are well-known:

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k. \tag{40}$$

Applying Eq. (40) and the lower bound in Lemma 11 we have

$$\begin{aligned}
\ln|S| &\geq \frac{d}{2}\ln 2 - \ln\left(\frac{d}{32}\right) - \ln\binom{\omega}{\delta_H/2} - \ln\binom{d-\omega}{\delta_H/2} \\
&\geq \frac{d}{2}\ln 2 - \ln\left(\frac{d}{32}\right) - \frac{d}{16}\ln(16e) \\
&= d\cdot\left(\frac{1}{2}\ln 2 - \frac{1}{16}\ln(16e)\right) - \ln\left(\frac{d}{32}\right) \\
&\geq 0.11d - 0.02d = 0.09d.
\end{aligned} \tag{41}$$

The last inequality is due to the fact that $0.02x - \ln(x/32) \geq 0$ for all $x > 0$. Therefore, $\log|S| = \ln|S|/\ln 2 \geq 0.13d \geq 0.0625d$. □

Appendix C Proofs of some technical propositions

Proof of Proposition 1. Suppose we have a stream based algorithm $A \in \mathcal{A}^{\mathrm{str}}_{d,T}$. An equivalent query synthetic algorithm $B \in \mathcal{A}^{\mathrm{qs}}_{d,T}$ can be constructed based on A as follows: at iteration t the algorithm B repeatedly samples data points from P_X until A accepts a sample x_t; it then picks x_t and requests its label. Clearly A and B are equivalent and they have the same worst-case expected excess risk. □

Proof of Proposition 2. Fix $w, w^* \in \mathbb{R}^d$ with $\|w\| = \|w^*\| = 1$. Denote $\Delta \subseteq \mathcal{X}$ as the set of data points on which w and w^* disagree; that is, $\Delta = \{x \in \mathcal{X} : \mathrm{sgn}(w \cdot x) \neq \mathrm{sgn}(w^* \cdot x)\}$. Suppose $\theta(w, w^*) = \theta$. We have the following:

$$\mathrm{err}_{0/1}(w) - \mathrm{err}_{0/1}(w^*) = \int_\Delta 2\left|\eta(x) - \frac{1}{2}\right| g(x)\mathrm{d}x \geq 2\int_0^\theta \mu_0 \varphi^{\frac{\alpha}{1-\alpha}} \cdot \gamma \mathrm{d}\varphi = 2(1-\alpha)\mu_0\gamma \cdot \theta^{\frac{1}{1-\alpha}}. \quad (42)$$

□

References

[ABC+14] Pranjal Awasthi, Afonso Bandeira, Moses Charikar, Ravishankar Krishnaswamy, Soledad Villar, and Rachel Ward. The power of localization for efficiently learning linear separators with noise. In *STOC*, 2014.

[ABRW12] Alekh Agarwal, Peter Bartlett, Pradeep Ravikumar, and Martin Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transaction on Information Theory*, 58(5), 2012.

[Ang88] Dana Angluin. Queries and concept learning. *Machine learning*, 2(4):319–342, 1988.

[BBL06] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *ICML*, 2006.

[BBZ07] Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *COLT*, 2007.

[BJM06] Peter Bartlett, Michael Jordan, and Jon McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[BL13] Maria-Florina Balcan and Philip Long. Active and passive learning of linear separators under log-concave distributions. In *COLT*, 2013.

[CAL94] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[CN06] Rui Castro and Robert Nowak. Upper and lower error bounds for active learning. In *The 44th Annual Allerton Conference on Communication, Control and Computing*, 2006.

[CN08] Rui M Castro and Robert D Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

[Coh90] David Cohn. Neural network exploration using optimal experiment design. In *NIPS*, 1990.

[Das05] Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS*, 2005.

[DHM07] Sanjoy Dasgupta, Daniel Hsu, and Claire Monteleoni. A general agnostic active learning algorithm. In *NIPS*, 2007.

[GS80] R Graham and N Sloane. Lower bounds for constant weight codes. *IEEE Transactions on Information Theory*, 26(1):37–43, 1980.

[Han07] Steve Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, 2007.

[Han11] Steve Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

[Hanar] Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, to appear.

[HY14] Steve Hanneke and Liu Yang. Minimax analysis of active learning. *arXiv:1410.0996*, 2014.

[JN14] Anatoli Juditsky and Yuri Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. *arXiv:1401.1792*, 2014.

[MT+99] Enno Mammen, Alexandre B Tsybakov, et al. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

[RDVC+04] Lorenzo Rosasco, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004.

[RS13a] Aaditya Ramdas and Aarti Singh. Algorithmic connections between active learning and stochastic convex optimization. In *ALT*, 2013.

[RS13b] Aaditya Ramdas and Aarti Singh. Optimal rates for stochastic convex optimization under tsybakov noise condition. In *ICML*, 2013.

[Tsy04] Alexandre B Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, pages 135–166, 2004.

[TZ09] Alexandre B Tsybakov and Vladimir Zaiats. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.