Active Learning with a Drifting Distribution

Liu Yang

Machine Learning Department
Carnegie Mellon University
liuy@cs.cmu.edu

Abstract. We study the problem of active learning in a stream-based setting, allowing the distribution of the examples to change over time. We prove upper bounds on the number of prediction mistakes and number of label requests for established disagreement-based active learning algorithms, both in the realizable case and under Tsybakov noise. We further prove minimax lower bounds for this problem.

1 Introduction

Most existing analyses of active learning are based on an i.i.d. assumption on the data. In this work, we assume the data are independent, but we allow the distribution from which the data are drawn to shift over time, while the target concept remains fixed. We consider this problem in a stream-based selective sampling model, and are interested in two quantities: the number of mistakes the algorithm makes on the first T examples in the stream, and the number of label requests among the first T examples in the stream.

In particular, we study scenarios in which the distribution may drift within a fixed totally bounded family of distributions. Unlike previous models of distribution drift (Bartlett, 1992; Crammer, Mansour, Even-Dar, and Vaughan, 2010), the minimax number of mistakes (or excess number of mistakes, in the noisy case) can be *sublinear* in the number of samples.

We specifically study the classic CAL active learning strategy in this context, and bound the number of mistakes and label requests the algorithm makes in the realizable case, under conditions on the concept space and the family of possible distributions. We also exhibit lower bounds on these quantities that match our upper bounds in certain cases. We further study a noise-robust variant of CAL, and analyze its number of mistakes and number of label requests in noisy scenarios where the noise distribution remains fixed over time but the marginal distribution on $\mathcal{X}$ may shift. In particular, we upper bound these quantities under Tsybakov's noise conditions. We also prove minimax lower bounds under these same conditions, though there is a gap between our upper and lower bounds.

2 Definition and Notations

As in the usual statistical learning problem, there is a standard Borel space $\mathcal{X}$, called the instance space, and a set $\mathbb{C}$ of measurable classifiers $h : \mathcal{X} \to \{-1, +1\}$, called the concept space. We additionally have a space $\mathbb{D}$ of distributions on $\mathcal{X}$, called the distribution space. Throughout, we suppose that the VC dimension of $\mathbb{C}$, denoted d below, is finite.

For any $\mu_1, \mu_2 \in \mathbb{D}$, let $\|\mu_1 - \mu_2\| = \sup_A \mu_1(A) - \mu_2(A)$ denote the total variation pseudo-distance between μ_1 and μ_2, where the set A in the sup ranges over all measurable subsets of $\mathcal{X}$. For any $\epsilon > 0$, let $\mathbb{D}_\epsilon$ denote a minimal ϵ-cover of $\mathbb{D}$, meaning that $\mathbb{D}_\epsilon \subseteq \mathbb{D}$ and $\forall \mu_1 \in \mathbb{D}, \exists \mu_2 \in \mathbb{D}_\epsilon$ s.t. $\|\mu_1 - \mu_2\| < \epsilon$, and that $\mathbb{D}_\epsilon$ has minimal possible size $|\mathbb{D}_\epsilon|$ among all subsets of $\mathbb{D}$ with this property.

In the learning problem, there is an unobservable sequence of distributions $\mathcal{D}_1, \mathcal{D}_2, \ldots$, with each $\mathcal{D}_t \in \mathbb{D}$, and an unobservable time-independent regular conditional distribution, which we represent by a function $\eta : \mathcal{X} \to [0, 1]$. Based on these quantities, we let $\mathcal{Z} = \{(X_t, Y_t)\}_{t=1}^{\infty}$ denote an infinite sequence of independent random variables, such that $\forall t, X_t \sim \mathcal{D}_t$, and the conditional distribution of Y_t given X_t satisfies $\forall x \in \mathcal{X}, \mathbb{P}(Y_t = +1 | X_t = x) = \eta(x)$. Thus, the joint distribution of (X_t, Y_t) is specified by the pair $(\mathcal{D}_t, \eta)$, and the distribution of $\mathcal{Z}$ is specified by the collection $\{\mathcal{D}_t\}_{t=1}^{\infty}$ along with η. We also denote by $\mathcal{Z}_t =$

$\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_t, Y_t)\}$ the first t such labeled examples. Note that the η conditional distribution is time-independent, since we are restricting ourselves to discussing drifting marginal distributions on $\mathcal{X}$, rather than drifting concepts. Concept drift is an important and interesting topic, but is beyond the scope of our present discussion.

In the active learning protocol, at each time t, the algorithm is presented with the value X_t, and is required to predict a label $\hat{Y}_t \in \{-1, +1\}$; then after making this prediction, it may optionally request to observe the true label value Y_t; as a means of book-keeping, if the algorithm requests a label Y_t on round t, we define $Q_t = 1$, and otherwise $Q_t = 0$.

We are primarily interested in two quantities. The first, $\hat{M}_T = \sum_{t=1}^{T} \mathbb{I}\left[\hat{Y}_t \neq Y_t\right]$, is the cumulative number of mistakes up to time T. The second quantity of interest, $\hat{Q}_T = \sum_{t=1}^{T} Q_t$, is the total number of labels requested up to time T. In particular, we will study the expectations of these quantities: $\bar{M}_T = \mathbb{E}\left[\hat{M}_T\right]$ and $\bar{Q}_T = \mathbb{E}\left[\hat{Q}_T\right]$. We are particularly interested in the asymptotic dependence of $\bar{Q}_T$ and $\bar{M}_T - \bar{M}_T^*$ on T, where $\bar{M}_T^* = \inf_{h \in \mathbb{C}} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left[h(X_t) \neq Y_t\right]\right]$. We refer to $\bar{Q}_T$ as the expected number of label requests, and to $\bar{M}_T - \bar{M}_T^*$ as the expected excess number of mistakes. For any distribution P on $\mathcal{X}$, we define $\mathrm{er}_P(h) = \mathbb{E}_{X \sim P}[\eta(X)\mathbb{I}[h(X) = -1] + (1 - \eta(X))\mathbb{I}[h(X) = +1]]$, the probability of h making a mistake for $X \sim P$ and Y with conditional probability of being $+1$ equal $\eta(X)$. Note that, abbreviating $\mathrm{er}_t(h) = \mathrm{er}_{\mathcal{D}_t}(h) = \mathbb{P}(h(X_t) \neq Y_t)$, we have $\bar{M}_T^* = \inf_{h \in \mathbb{C}} \sum_{t=1}^{T} \mathrm{er}_t(h)$.

Scenarios in which both $\bar{M}_T - \bar{M}_T^*$ and $\bar{Q}_T$ are $o(T)$ (i.e., sublinear) are considered desirable, as these represent cases in which we do "learn" the proper way to predict labels, while asymptotically using far fewer labels than passive learning. Once establishing conditions under which this is possible, we may then further explore the trade-off between these two quantities.

We will additionally make use of the following notions. For $V \subseteq \mathbb{C}$, let $\mathrm{diam}_t(V) = \sup_{h,g \in V} \mathcal{D}_t(\{x : h(x) \neq g(x)\})$. For $h : \mathcal{X} \to \{-1, +1\}$, $\bar{\mathrm{er}}_{s:t}(h) = \frac{1}{t-s+1} \sum_{u=s}^{t} \mathrm{er}_u(h)$, and for finite $S \subseteq \mathcal{X} \times \{-1, +1\}$, $\hat{\mathrm{er}}(h; S) = \frac{1}{|S|} \sum_{(x,y) \in S} \mathbb{I}[h(x) \neq y]$. Also let $\mathbb{C}[S] = \{h \in \mathbb{C} : \hat{\mathrm{er}}(h; S) = 0\}$. Finally, for a distribution P on $\mathcal{X}$ and $r > 0$, define $\mathrm{B}_P(h, r) = \{g \in \mathbb{C} : P(x : h(x) \neq g(x)) \leq r\}$.

2.1 Assumptions

In addition to the assumption of independence of the X_t variables and that $d < \infty$, each result below is stated under various additional assumptions. The weakest such assumption is that $\mathbb{D}$ is *totally bounded*, in the following sense. For each $\epsilon > 0$, let $\mathbb{D}_\epsilon$ denote a minimal subset of $\mathbb{D}$ such that $\forall \mathcal{D} \in \mathbb{D}, \exists \mathcal{D}' \in \mathbb{D}_\epsilon$ s.t. $\|\mathcal{D} - \mathcal{D}'\| < \epsilon$: that is, a minimal ϵ-cover of $\mathbb{D}$. We say that $\mathbb{D}$ is totally bounded if it satisfies the following assumption.

Assumption 1 $\forall \epsilon > 0, |\mathbb{D}_\epsilon| < \infty$.

In some of the results below, we will be interested in deriving specific rates of convergence. Doing so requires us to make stronger assumptions about $\mathbb{D}$ than mere total boundedness. We will specifically consider the following condition, in which $c, m \in [0, \infty)$ are constants.

Assumption 2 $\forall \epsilon > 0, |\mathbb{D}_\epsilon| < c \cdot \epsilon^{-m}$.

For an example of a class $\mathbb{D}$ satisfying the total boundedness assumption, consider $\mathcal{X} = [0, 1]^n$, and let $\mathbb{D}$ be the collection of distributions that have uniformly continuous density function with respect to the Lebesgue measure on $\mathcal{X}$, with modulus of continuity at most some value $\omega(\epsilon)$ for each value of $\epsilon > 0$, where $\omega(\epsilon)$ is a fixed real-valued function with $\lim_{\epsilon \to 0} \omega(\epsilon) = 0$.

As a more concrete example, when $\omega(\epsilon) = L\epsilon$ for some $L \in (0, \infty)$, this corresponds to the family of Lipschitz continuous density functions with Lipschitz constant at most L. In this case, we have $|\mathbb{D}_\epsilon| \leq O\left(\epsilon^{-n}\right)$, satisfying Assumption 2.

3 Related Work

We discuss active learning under distribution drift, with fixed target concept. There are several branches of the literature that are highly relevant to this, including domain adaptation (Mansour, Mohri, and Rostamizadeh, 2009, 2008), online learning (Littlestone, 1988), learning with concept drift, and empirical processes for independent but not identically distributed data (van de Geer, 2000).

Streamed-based Active Learning with a Fixed Distribution Dasgupta, Kalai, and Monteleoni (2009) show that a certain modified perceptron-like active learning algorithm can achieve a mistake bound $O(d\log(T))$ and query bound $\tilde{O}(d\log(T))$, when learning a linear separator under a uniform distribution on the unit sphere, in the realizable case.

Dekel, Gentile, and Sridharam (2010) also analyze the problem of learning linear separators under a uniform distribution, but allowing Tsybakov noise. They find that with $\bar{Q}_T = \tilde{O}\left(d^{\frac{2\alpha}{\alpha+2}}T^{\frac{2}{\alpha+2}}\right)$ queries, it is possible to achieve an expected excess number of mistakes $\bar{M}_T - M_T^* = \tilde{O}\left(d^{\frac{\alpha+1}{\alpha+2}} \cdot T^{\frac{1}{\alpha+2}}\right)$.

At this time, we know of no work studying the number of mistakes and queries achievable by active learning in a stream-based setting where the distribution may change over time.

Stream-based Passive Learning with a Drifting Distribution There has been work on learning with a drifting distribution and fixed target, in the context of passive learning. Bartlett (1992); Barve and Long (1997) study the problem of learning a subset of a domain from randomly chosen examples when the probability distribution of the examples changes slowly but continually throughout the learning process; they give upper and lower bounds on the best achievable probability of misclassification after a given number of examples. They consider learning problems in which a changing environment is modeled by a slowly changing distribution on the product space. The allowable drift is restricted by ensuring that consecutive probability distributions are close in total variation distance. However, this assumption allows for certain malicious choices of distribution sequences, which shift the probability mass into smaller and smaller regions where the algorithm is uncertain of the target's behavior, so that the number of mistakes grows linearly in the number of samples in the worst case. More recently, Freund and Mansour (1997) have investigated learning when the distribution changes as a linear function of time. They present algorithms that estimate the error of functions, using knowledge of this linear drift.

4 Active Learning in the Realizable Case

Throughout this section, suppose $\mathbb{C}$ is a fixed concept space and $h^* \in \mathbb{C}$ is a fixed target function: that is, $\mathrm{er}_t(h^*) = 0$. The family of scenarios in which this is true are often collectively referred to as the *realizable case*. We begin our analysis by studying this realizable case because it greatly simplifies the analysis, laying bare the core ideas in plain form. We will discuss more general scenarios, in which $\mathrm{er}_t(h^*) \geq 0$, in later sections, where we find that essentially the same principles apply there as in this initial realizable-case analysis.

We will be particularly interested in the performance of the following simple algorithm, due to Cohn, Atlas, and Ladner (1994), typically referred to as CAL after its discoverers. The version presented here is specified in terms of a passive learning subroutine $\mathcal{A}$ (mapping any sequence of labeled examples to a classifier).

CAL
1. $t \leftarrow 0$, $\mathcal{Q}_0 \leftarrow \emptyset$, and let $\hat{h}_0 = \mathcal{A}(\emptyset)$
2. Do
3. $\quad t \leftarrow t + 1$
4. $\quad$ Predict $\hat{Y}_t = \hat{h}_{t-1}(X_t)$
5. $\quad$ If $\max_{y \in \{-1,+1\}} \min_{h \in \mathbb{C}} \hat{\text{er}}(h; \mathcal{Q}_{t-1} \cup \{(X_t, y)\}) = 0$
6. $\quad\quad$ Request Y_t, let $\mathcal{Q}_t = \mathcal{Q}_{t-1} \cup \{(X_t, Y_t)\}$
7. $\quad$ Else let $Y'_t = \operatorname{argmin}_{y \in \{-1,+1\}} \min_{h \in \mathbb{C}} \hat{\text{er}}(h; \mathcal{Q}_{t-1} \cup \{(X_t, y)\})$, and let $\mathcal{Q}_t \leftarrow \mathcal{Q}_{t-1} \cup \{(X_t, Y'_t)\}$
8. $\quad$ Let $\hat{h}_t = \mathcal{A}(\mathcal{Q}_t)$

Below, we let $\mathcal{A}_{1IG}$ denote the one-inclusion graph prediction strategy of Haussler, Littlestone, and Warmuth (1994). Specifically, the passive learning algorithm $\mathcal{A}_{1IG}$ is specified as follows. For a sequence of data points $\mathcal{U} \in \mathcal{X}^{t+1}$, the one-inclusion graph is a graph, where each vertex represents a distinct labeling of $\mathcal{U}$ that can be realized by some classifier in $\mathbb{C}$, and two vertices are adjacent if and only if their corresponding labelings for $\mathcal{U}$ differ by exactly one label. We use the one-inclusion graph to define a classifier based on t training points as follows. Given t labeled data points $\mathcal{L} = \{(x_1, y_1), \ldots, (x_t, y_t)\}$, and one test point x_{t+1} we are asked to predict a label for, we first construct the one-inclusion graph on $\mathcal{U} = \{x_1, \ldots, x_{t+1}\}$; we then orient the graph (give each edge a unique direction) in a way that minimizes the maximum out-degree, and breaks ties in a way that is invariant to permutations of the order of points in $\mathcal{U}$; after orienting the graph in this way, we examine the subset of vertices whose corresponding labeling of $\mathcal{U}$ is consistent with $\mathcal{L}$; if there is only one such vertex, then we predict for x_{t+1} the corresponding label from that vertex; otherwise, if there are two such vertices, then they are adjacent in the one-inclusion graph, and we choose the one toward which the edge is directed and use the label for x_{t+1} in the corresponding labeling of $\mathcal{U}$ as our prediction for the label of x_{t+1}. See (Haussler, Littlestone, and Warmuth, 1994) and subsequent work for detailed studies of the one-inclusion graph prediction strategy.

4.1 Learning with a Fixed Distribution

We begin the discussion with the simplest case: namely, when $|\mathbb{D}| = 1$.

Definition 1. *(Hanneke, 2007, 2011b) Define the disagreement coefficient of h^* under a distribution P as*

$$\theta_P(\epsilon) = \sup_{r > \epsilon} \frac{P\left(\text{DIS}(\text{B}_P(h^*, r))\right)}{r}.$$

Theorem 1. *For any distribution P on $\mathcal{X}$, if $\mathbb{D} = \{P\}$, then running CAL with $\mathcal{A} = \mathcal{A}_{1IG}$ achieves expected mistake bound $\bar{M}_T = O\left(d \log(T)\right)$ and expected query bound $\bar{Q}_T = O\left(\theta_P(\epsilon_T) d \log^2(T)\right)$,*

for $\epsilon_T = d \log(T)/T$.

Proof. First note that, by the assumption that $\forall t, \text{er}_t(h^*) = 0$, with probability 1 we have that $\forall t, \mathcal{Q}_t = \mathcal{Z}_t$. Thus, since the stated bound on $\bar{M}_T$ for the one-inclusion graph algorithm has been established when using the true sequence of labeled examples $\mathcal{Z}_T$ (Haussler, Littlestone, and Warmuth, 1994), it must hold here as well.

The remainder of the proof focuses on the bound on $\bar{Q}_T$. This proof is essentially based on a related proof of Hanneke (2011b), but reformulated for this stream-based model.

Let V_t denote the set of classifiers $h \in \mathbb{C}$ with $\hat{\text{er}}(h; \mathcal{Q}_t) = 0$ (with $V_0 = \mathbb{C}$). Classic results from statistical learning theory (Vapnik, 1982; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989) imply that for $t > d$, with probability at least $1 - \delta$,

$$\text{diam}_t(V_{t-1}) \leq cd \frac{\log(2e(t-1)/d) + \log(4/\delta)}{t-1}, \tag{1}$$

for some universal constant $c \in (1, \infty)$.

In particular, for $d < t \leq T$, since the probability CAL requests the label Y_t is $P(X_t \in \mathrm{DIS}(V_{t-1}))$, (1) implies that this probability satisfies

$$P\left(X_t \in \mathrm{DIS}(V_{t-1})\right) \leq P\left(X_t \in \mathrm{DIS}\left(\mathrm{B}_P\left(h^*, cd\frac{\log(2e(t-1)/d) + \log(4/\delta)}{t-1}\right)\right)\right) + \delta$$
$$\leq \theta_P\left(d\log(T)/T\right) cd\frac{\log(2e(t-1)/d) + \log(4/\delta)}{t-1} + \delta.$$

Taking $\delta = d/(t-1)$, this implies

$$P\left(X_t \in \mathrm{DIS}(V_{t-1})\right) \leq \theta_P\left(d\log(T)/T\right) 2cd\frac{\log(8e(t-1)/d)}{t-1}.$$

Thus, for $T > d$,

$$\bar{Q}_T = \sum_{t=1}^{T} P\left(X_t \in \mathrm{DIS}(V_{t-1})\right) \leq d + 1 + \sum_{t=d+1}^{T-1} \theta_P\left(d\log(T)/T\right) 2cd\frac{\log(8et/d)}{t}$$
$$\leq d + 1 + \theta_P\left(d\log(T)/T\right) 2cd\log(8eT/d)\int_d^T \frac{1}{t}\mathrm{d}t = d + 1 + \theta_P\left(d\log(T)/T\right) 2cd\log(8eT/d)\log(T/d).$$

□

4.2 Learning with a Drifting Distribution

We now generalize the above results to any sequence of distributions from a totally bounded space $\mathbb{D}$. Throughout this section, let

$$\theta_{\mathbb{D}}(\epsilon) = \sup_{P \in \mathbb{D}} \theta_P(\epsilon).$$

First, we prove a basic result stating that CAL can achieve a sublinear number of mistakes, and under conditions on the disagreement coefficient, also a sublinear number of queries.

Theorem 2. *If $\mathbb{D}$ is totally bounded (Assumption 1), then CAL (with $\mathcal{A}$ any empirical risk minimization algorithm) achieves an expected mistake bound $\bar{M}_T = o(T)$, and if $\theta_{\mathbb{D}}(\epsilon) = o(1/\epsilon)$, then CAL makes an expected number of queries $\bar{Q}_T = o(T)$.*

Proof. As mentioned, given that $\mathrm{er}_{\mathcal{Q}_{t-1}}(h^*) = 0$, we have that Y_t' in Step 7 must equal $h^*(X_t)$, so that the invariant $\mathrm{er}_{\mathcal{Q}_t}(h^*) = 0$ is maintained for all t by induction. In particular, this implies $\mathcal{Q}_t = \mathcal{Z}_t$ for all t.

Fix any $\epsilon > 0$, and enumerate the elements of $\mathbb{D}_\epsilon$ so that $\mathbb{D}_\epsilon = \{P_1, P_2, \ldots, P_{|\mathbb{D}_\epsilon|}\}$. For each $t \in \mathbb{N}$, let $k(t) = \mathrm{argmin}_{k \leq |\mathbb{D}_\epsilon|} \|P_k - \mathcal{D}_t\|$, breaking ties arbitrarily. Let

$$L(\epsilon) = \left\lceil \frac{8}{\sqrt{\epsilon}}\left(d\ln\left(\frac{24}{\sqrt{\epsilon}}\right) + \ln\left(\frac{4}{\sqrt{\epsilon}}\right)\right)\right\rceil.$$

For each $i \leq |\mathbb{D}_\epsilon|$, if $k(t) = i$ for infinitely many $t \in \mathbb{N}$, then let T_i denote the smallest value of T such that $|\{t \leq T : k(t) = i\}| = L(\epsilon)$. If $k(t) = i$ only finitely many times, then let T_i denote the largest index t for which $k(t) = i$, or $T_i = 1$ if no such index t exists.

Let $T_\epsilon = \max_{i \leq |\mathbb{D}_\epsilon|} T_i$ and $V_\epsilon = \mathbb{C}[\mathcal{Z}_{T_\epsilon}]$. We have that

$$\forall t > T_\epsilon, \mathrm{diam}_t(V_\epsilon) \leq \mathrm{diam}_{k(t)}(V_\epsilon) + \epsilon.$$

For each i, let $\mathcal{L}_i$ be a sequence of $L(\epsilon)$ i.i.d. pairs (X, Y) with $X \sim P_i$ and $Y = h^*(X)$, and let $V_i = \mathbb{C}[\mathcal{L}_i]$. Then $\forall t > T_\epsilon$,

$$\mathbb{E}\left[\mathrm{diam}_{k(t)}(V_\epsilon)\right] \leq \mathbb{E}\left[\mathrm{diam}_{k(t)}(V_{k(t)})\right] + \sum_{s \leq T_i : k(s) = k(t)} \|\mathcal{D}_s - P_{k(s)}\| \leq \mathbb{E}\left[\mathrm{diam}_{k(t)}(V_{k(t)})\right] + L(\epsilon)\epsilon.$$

By classic results in the theory of PAC learning (Anthony and Bartlett, 1999; Vapnik, 1982),

$$\forall t > T_\epsilon, \mathbb{E}\left[\mathrm{diam}_{k(t)}(V_{k(t)})\right] \leq \sqrt{\epsilon}.$$

Combining the above arguments,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right] &\leq T_\epsilon + \sum_{t=T_\epsilon+1}^{T} \mathbb{E}\left[\mathrm{diam}_t(V_\epsilon)\right] \leq T_\epsilon + \epsilon T + \sum_{t=T_\epsilon+1}^{T} \mathbb{E}\left[\mathrm{diam}_{k(t)}(V_\epsilon)\right] \\
&\leq T_\epsilon + \epsilon T + L(\epsilon)\epsilon T + \sum_{t=T_\epsilon+1}^{T} \mathbb{E}\left[\mathrm{diam}_{k(t)}(V_{k(t)})\right] \\
&\leq T_\epsilon + \epsilon T + L(\epsilon)\epsilon T + \sqrt{\epsilon} T.
\end{aligned}$$

Let ϵ_T be any nonincreasing sequence in $(0,1)$ such that $1 \ll T_{\epsilon_T} \ll T$. Since $|\mathbb{D}_\epsilon| < \infty$ for all $\epsilon > 0$, we must have $\epsilon_T \to 0$. Thus, noting that $\lim_{\epsilon \to 0} L(\epsilon)\epsilon = 0$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right] \leq T_{\epsilon_T} + \epsilon_T T + L(\epsilon_T)\epsilon_T T + \sqrt{\epsilon_T} T \ll T. \tag{2}$$

The result on $\bar{M}_T$ now follows by noting that for any $\hat{h}_{t-1} \in \mathbb{C}[\mathcal{Z}_{t-1}]$ has $\mathrm{er}_t(\hat{h}_{t-1}) \leq \mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])$, so

$$\bar{M}_T = \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{er}_t\left(\hat{h}_{t-1}\right)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right] \ll T.$$

Similarly, for $r > 0$, we have

$$\begin{aligned}
\mathbb{P}(\text{Request } Y_t) &= \mathbb{E}\left[\mathbb{P}(X_t \in \mathrm{DIS}(\mathbb{C}[\mathcal{Z}_{t-1}])|\mathcal{Z}_{t-1})\right] \leq \mathbb{E}\left[\mathbb{P}(X_t \in \mathrm{DIS}(\mathbb{C}[\mathcal{Z}_{t-1}] \cup \mathrm{B}_{\mathcal{D}_t}(h^*, r)))\right] \\
&\leq \mathbb{E}\left[\theta_{\mathbb{D}}(r) \cdot \max\left\{\mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}]), r\right\}\right] \leq \theta_{\mathbb{D}}(r) \cdot r + \theta_{\mathbb{D}}(r) \cdot \mathbb{E}\left[\mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right].
\end{aligned}$$

Letting $r_T = T^{-1}\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right]$, we see that $r_T \to 0$ by (2), and since $\theta_{\mathbb{D}}(\epsilon) = o(1/\epsilon)$, we also have $\theta_{\mathbb{D}}(r_T) r_T \to 0$, so that $\theta_{\mathbb{D}}(r_T) r_T T \ll T$. Therefore,

$$\bar{Q}_T = \sum_{t=1}^{T} \mathbb{P}(\text{Request } Y_t) \leq \theta_{\mathbb{D}}(r_T) \cdot r_T \cdot T + \theta_{\mathbb{D}}(r_T) \cdot \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right] = 2\theta_{\mathbb{D}}(r_T) \cdot r_T \cdot T \ll T.$$

□

We can also state a more specific result in the case when we have some more detailed information on the sizes of the finite covers of $\mathbb{D}$.

Theorem 3. *If Assumption 2 is satisfied, then CAL (with $\mathcal{A}$ any empirical risk minimization algorithm) achieves an expected mistake bound $\bar{M}_T$ and expected number of queries $\bar{Q}_T$ such that*

$$\begin{aligned}
\bar{M}_T &= O\left(T^{\frac{m}{m+1}} d^{\frac{1}{m+1}} \log^2 T\right) \\
\textit{and } \bar{Q}_T &= O\left(\theta_{\mathbb{D}}\left(\epsilon_T\right) T^{\frac{m}{m+1}} d^{\frac{1}{m+1}} \log^2 T\right),
\end{aligned}$$

where $\epsilon_T = (d/T)^{\frac{1}{m+1}}$.

Proof. Fix $\epsilon > 0$, enumerate $\mathbb{D}_\epsilon = \{P_1, P_2, \ldots, P_{|\mathbb{D}_\epsilon|}\}$, and for each $t \in \mathbb{N}$, let $k(t) = \operatorname{argmin}_{1 \le k \le |\mathbb{D}_\epsilon|} \|\mathcal{D}_t - P_k\|$. Let $\{X'_t\}_{t=1}^\infty$ be a sequence of independent samples, with $X'_t \sim P_{k(t)}$, and $\mathcal{Z}'_t = \{(X'_1, h^*(X'_1)), \ldots, (X'_t, h^*(X'_t)\}$. Then

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T \operatorname{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right] &\le \mathbb{E}\left[\sum_{t=1}^T \operatorname{diam}_t(\mathbb{C}[\mathcal{Z}'_{t-1}])\right] + \sum_{t=1}^T \|\mathcal{D}_t - P_{k(t)}| \\ &\le \mathbb{E}\left[\sum_{t=1}^T \operatorname{diam}_t(\mathbb{C}[\mathcal{Z}'_{t-1}])\right] + \epsilon T \le \sum_{t=1}^T \mathbb{E}\left[\operatorname{diam}_{P_{k(t)}}(\mathbb{C}[\mathcal{Z}'_{t-1}])\right] + 2\epsilon T.\end{aligned}$$

The classic convergence rates results from PAC learning Anthony and Bartlett (1999); Vapnik (1982) imply

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}\left[\operatorname{diam}_{P_{k(t)}}(\mathbb{C}[\mathcal{Z}'_{t-1}])\right] &= \sum_{t=1}^T O\left(\frac{d \log t}{|\{i \le t : k(i) = k(t)\}|}\right) \le O(d \log T) \cdot \sum_{t=1}^T \frac{1}{|\{i \le t : k(i) = k(t)\}|} \\ &\le O(d \log T) \cdot |\mathbb{D}_\epsilon| \cdot \sum_{u=1}^{\lceil T/|\mathbb{D}_\epsilon| \rceil} \frac{1}{u} \le O\left(d|\mathbb{D}_\epsilon| \log^2(T)\right).\end{aligned}$$

Thus, we have

$$\sum_{t=1}^T \mathbb{E}\left[\operatorname{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right] \le O\left(d|\mathbb{D}_\epsilon| \log^2(T) + \epsilon T\right) \le O\left(d \cdot \epsilon^{-m} \log^2(T) + \epsilon T\right).$$

Taking $\epsilon = (T/d)^{-\frac{1}{m+1}}$, this is $O\left(d^{\frac{1}{m+1}} \cdot T^{\frac{m}{m+1}} \log^2(T)\right)$.

We therefore have

$$\bar{M}_T \le \mathbb{E}\left[\sum_{t=1}^T \sup_{h \in \mathbb{C}[\mathcal{Z}_{t-1}]} \operatorname{er}_t(h)\right] \le \mathbb{E}\left[\sum_{t=1}^T \operatorname{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right] \le O\left(d^{\frac{1}{m+1}} \cdot T^{\frac{m}{m+1}} \log^2(T)\right).$$

Similarly, letting $\epsilon_T = (d/T)^{\frac{1}{m+1}}$,

$$\begin{aligned}\bar{Q}_T &\le \mathbb{E}\left[\sum_{t=1}^T \mathcal{D}_t(\mathrm{DIS}(\mathbb{C}[\mathcal{Z}_{t-1}]))\right] \le \mathbb{E}\left[\sum_{t=1}^T \mathcal{D}_t\left(\mathrm{DIS}\left(\mathrm{B}_{\mathcal{D}_t}\left(h^*, \max\left\{\operatorname{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}]), \epsilon_T\right\}\right)\right)\right)\right] \\ &\le \mathbb{E}\left[\sum_{t=1}^T \theta_{\mathbb{D}}(\epsilon_T) \cdot \max\left\{\operatorname{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}]), \epsilon_T\right\}\right] \le \mathbb{E}\left[\sum_{t=1}^T \theta_{\mathbb{D}}(\epsilon_T) \cdot \operatorname{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right] + \theta_{\mathbb{D}}(\epsilon_T) \cdot T\epsilon_T \\ &\le O\left(\theta_{\mathbb{D}}(\epsilon_T) \cdot d^{\frac{1}{m+1}} \cdot T^{\frac{m}{m+1}} \log^2(T)\right).\end{aligned}$$

□

We can additionally construct a lower bound for this scenario, as follows. Suppose $\mathbb{C}$ contains a full infinite binary tree for which all classifiers in the tree agree on some point. That is, there is a set of points $\{x_b : b \in \{0,1\}^k, k \in \mathbb{N}\}$ such that, for $b_1 = 0$ and $\forall b_2, b_3, \ldots \in \{0,1\}$, $\exists h \in \mathbb{C}$ such that $h(x_{(b_1,\ldots,b_{j-1})}) = b_j$ for $j \ge 2$. For instance, this is the case for linear separators (and most other natural "geometric" concept spaces).

Theorem 4. *For any $\mathbb{C}$ as above, for any active learning algorithm, $\exists$ a set $\mathbb{D}$ satsifying Assumption 2, a target function $h^* \in \mathbb{C}$, and a sequence of distributions $\{\mathcal{D}_t\}_{t=1}^T$ in $\mathbb{D}$ such that the $\bar{M}_T$ and $\bar{Q}_T$ achieved by the learning algorithm satisfy*

$$\begin{aligned}\bar{M}_T &= \Omega\left(T^{\frac{m}{m+1}}\right), \\ \text{and } \bar{M}_T = O\left(T^{\frac{m}{m+1}}\right) &\implies \bar{Q}_T = \Omega\left(T^{\frac{m}{m+1}}\right).\end{aligned}$$

The proof is analogous to that of Theorem 4 below, and is therefore omitted for brevity.

5 Learning with Noise

In this section, we extend the above analysis to allow for various types of noise conditions commonly studied in the literature. For this, we will need to study a noise-robust variant of CAL, below referred to as Agnostic CAL (or ACAL). We prove upper bounds achieved by ACAL, as well as (non-matching) minimax lower bounds.

5.1 Noise Conditions

The following assumption may be referred to as a *strictly benign noise* condition, which essentially says the model is specified correctly in that $h^* \in \mathbb{C}$, and though the labels may be stochastic, they are not completely random, but rather each is slightly biased toward the h^* label.

Assumption 3 $h^* = \text{sign}(\eta) \in \mathbb{C}$ *and* $\forall x, \eta(x) \neq 1/2$.

A particularly interesting special case of Assumption 3 is given by Tsybakov's noise conditions, which essentially control how common it is to have η values close to $1/2$. This is formally specified by the following assumption.

Assumption 4 η *satisfies Assumption 3 and for some* $c > 0$ *and* $\alpha \geq 0$,
$\forall t > 0, P(|\eta(x) - 1/2| < t) < c \cdot t^{\alpha}$.

In the setting of shifting distributions, we will be interested in conditions for which the above assumptions are satisifed simultaneously for all distributions in $\mathbb{D}$. We formalize this in the following assumption.

Assumption 5 *Assumption 4 is satisfied for all* $\mathcal{D} \in \mathbb{D}$, *with the same* c *and* α *values.*

5.2 Agnostic CAL

The following algorithm is essentially take from (Dasgupta, Hsu, and Monteleoni, 2007; Hanneke, 2011b), adapted here for this stream-based setting. It is based on a subroutine:

$$\textsc{Learn}(\mathcal{L}, \mathcal{Q}) = \begin{cases} \text{argmin}_{h \in \mathbb{C}: \hat{\text{er}}(h;\mathcal{L})=0} \hat{\text{er}}(h; \mathcal{Q}), & \text{if } \min_{h \in \mathbb{C}} \hat{\text{er}}(h; \mathcal{L}) = 0 \\ \varnothing, & \text{otherwise} \end{cases}.$$

ACAL
1. $t \leftarrow 0$, $\mathcal{L}_t \leftarrow \emptyset$, $\mathcal{Q}_t \leftarrow \emptyset$, let $\hat{h}_t$ be any element of $\mathbb{C}$
2. Do
3. $\quad t \leftarrow t + 1$
4. $\quad$ Predict $\hat{Y}_t = \hat{h}_{t-1}(X_t)$
5. $\quad$ For each $y \in \{-1, +1\}$, let $h^{(y)} = \textsc{Learn}(\mathcal{L}_{t-1}, \mathcal{Q}_{t-1})$
6. $\quad$ If either y has $h^{(-y)} = \varnothing$ or
$\qquad \hat{\text{er}}(h^{(-y)}; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) - \hat{\text{er}}(h^{(y)}; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) > \hat{\mathcal{E}}_{t-1}(\mathcal{L}_{t-1}, \mathcal{Q}_{t-1})$
7. $\quad\quad \mathcal{L}_t \leftarrow \mathcal{L}_{t-1} \cup \{(X_t, y)\}$, $\mathcal{Q}_t \leftarrow \mathcal{Q}_{t-1}$
8. $\quad$ Else Request Y_t, and let $\mathcal{L}_t \leftarrow \mathcal{L}_{t-1}$, $\mathcal{Q}_t \leftarrow \mathcal{Q}_{t-1} \cup \{(X_t, Y_t)\}$
9. $\quad$ Let $\hat{h}_t = \textsc{Learn}(\mathcal{L}_t, \mathcal{Q}_t)$
10. $\quad$ If t is a power of 2
11. $\quad\quad \mathcal{L}_t \leftarrow \emptyset$, $\mathcal{Q}_t \leftarrow \emptyset$

The algorithm is expressed in terms of a function $\hat{\mathcal{E}}_t(\mathcal{L}, \mathcal{Q})$, defined as follows. Let δ_i be a nonincreasing sequence of values in $(0, 1)$. Let $\xi_1, \xi_2, \ldots$ denote a sequence of independent Rademacher random variables (i.e., uniform in $\{-1, +1\}$), also independent from the data. Then for any set $V \subseteq \mathbb{C}$, define

$$\hat{R}_t(V) = \sup_{h_1, h_2 \in V} \frac{1}{t - 2^{\lfloor \log_2(t-1) \rfloor}} \sum_{m=2^{\lfloor \log_2(t-1) \rfloor}+1}^{t} \xi_m \cdot (h_1(X_m) - h_2(X_m)),$$

$$\hat{D}_t(V) = \sup_{h_1, h_2 \in V} \frac{1}{t - 2^{\lfloor \log_2(t-1) \rfloor}} \sum_{m=2^{\lfloor \log_2(t-1) \rfloor}+1}^{t} |h_1(X_m) - h_2(X_m)|,$$

$$\hat{U}_t(V, \delta) = 12\hat{R}_t(V) + 34\sqrt{\hat{D}_t(V)\frac{\ln(32t^2/\delta)}{t}} + \frac{752 \ln(32t^2/\delta)}{t}. \tag{3}$$

Also, for any finite sets $\mathcal{L}, \mathcal{Q} \subseteq \mathcal{X} \times \mathcal{Y}$, let $\mathbb{C}[\mathcal{L}] = \{h \in \mathbb{C} : \hat{\text{er}}(h; \mathcal{L}) = 0\}$, $\hat{\mathbb{C}}(\epsilon; \mathcal{L}, \mathcal{Q}) = \{h \in \mathbb{C}[\mathcal{L}] : \hat{\text{er}}(h; \mathcal{L} \cup \mathcal{Q}) - \min_{g \in \mathbb{C}[\mathcal{L}]} \hat{\text{er}}(g; \mathcal{L} \cup \mathcal{Q}) \leq \epsilon\}$. Then define

$$\hat{U}_t(\epsilon, \delta; \mathcal{L}, \mathcal{Q}) = \hat{U}_t(\hat{\mathbb{C}}_t(\epsilon; \mathcal{L}, \mathcal{Q}), \delta),$$

and (letting $\mathbb{Z}_\epsilon = \{j \in \mathbb{Z} : 2^j \geq \epsilon\}$)

$$\hat{\mathcal{E}}_t(\mathcal{L}, \mathcal{Q}) = \inf \left\{ \epsilon > 0 : \forall j \in \mathbb{Z}_\epsilon, \min_{m \in \mathbb{N}} \hat{U}_t(\epsilon, \delta_{\lfloor \log(t) \rfloor}; \mathcal{L}, \mathcal{Q}) \leq 2^{j-4} \right\}.$$

5.3 Learning with a Fixed Distribution

The following results essentially follow from the analysis of Hanneke (2011a), adapted to this stream-based setting.

Theorem 5. *For any strictly benign (P, η), if $2^{-2^i} \ll \delta_i \ll 2^{-i}/i$, ACAL achieves an expected excess number of mistakes $\bar{M}_T - M_T^* = o(T)$, and if $\theta_P(\epsilon) = o(1/\epsilon)$, then ACAL makes an expected number of queries $\bar{Q}_T = o(T)$.*

Theorem 6. *For any (P, η) satisfying Assumption 4, if $\mathbb{D} = \{P\}$, ACAL achieves an expected excess number of mistakes*

$$\bar{M}_T - M_T^* = \tilde{O}\left(d^{\frac{1}{\alpha+2}} \cdot T^{\frac{\alpha+1}{\alpha+2}} \log\left(\frac{1}{\delta_{\lfloor \log(T) \rfloor}}\right) + \sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i \right).$$

and and expected number of queries

$$\bar{Q}_T = \tilde{O}\left(\theta_P(\epsilon_T) \cdot d^{\frac{2}{\alpha+2}} \cdot T^{\frac{\alpha}{\alpha+2}} \log\left(\frac{1}{\delta_{\lfloor \log(T) \rfloor}}\right) + \sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i \right).$$

where $\epsilon_T = T^{-\frac{\alpha}{\alpha+2}}$.

Corollary 1. *For any (P, η) satisfying Assumption 4, if $\mathbb{D} = \{P\}$ and $\delta_i = 2^{-i}$ in ACAL, the algorithm achieves an expected excess number of mistakes $\bar{M}_T - M_T^* = \tilde{O}\left(d^{\frac{1}{\alpha+2}} \cdot T^{\frac{\alpha+1}{\alpha+2}}\right)$. and and expected number of queries $\bar{Q}_T = \tilde{O}\left(\theta_P(\epsilon_T) \cdot d^{\frac{2}{\alpha+2}} \cdot T^{\frac{\alpha}{\alpha+2}}\right)$. where $\epsilon_T = T^{-\frac{\alpha}{\alpha+2}}$.*

5.4 Learning with a Drifting Distribution

The following lemma is similar to a result proven by Hanneke (2011b), based on the work of Koltchinskii (2006), except here we have adapted the result to the present setting with changing distributions. The proof is essentially identical to the proof of the original result of Hanneke (2011b), and is therefore omitted here.

Lemma 1. *(Hanneke, 2011b) Suppose η satisfies Assumption 3. For every $i \in \mathbb{N}$, on an event E_i with $\mathbb{P}(E_i) \geq 1 - \delta_i$, $\forall t \in \{2^i + 1, \ldots, 2^{i+1}\}$, letting $t(i) = t - 2^i$,*

- $\hat{\text{er}}(h^*; \mathcal{L}_{t-1}) = 0$,
- $\forall h \in \mathbb{C}$ *s.t.* $\hat{\text{er}}(h; \mathcal{L}_{t-1}) = 0$ *and* $\hat{\text{er}}(h; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) - \hat{\text{er}}(h^*; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) \leq \hat{\mathcal{E}}_{t-1}(\mathcal{L}_{t-1}, \mathcal{Q}_{t-1})$,
  we have $\bar{\text{er}}_{2^i+1:t-1}(h) - \bar{\text{er}}_{2^i+1:t-1}(h^*) \leq 2\hat{\mathcal{E}}_{t-1}(\mathcal{L}_{t-1}, \mathcal{Q}_{t-1})$,
- *if Assumption 5 is satisifed,* $\hat{\mathcal{E}}_{t-1}(\mathcal{L}_{t-1}, \mathcal{Q}_{t-1}) \leq \tilde{K} \cdot \left(\frac{d \log(t(i)/\delta_i)}{t(i)}\right)^{\frac{\alpha+1}{\alpha+2}}$,

for some (c, α)-dependent constant $\tilde{K} \in (1, \infty)$.

We can now state and prove our results concerning ACAL, which are analogous to Theorems 2 and 3 proved earlier for CAL in the realizable case.

Theorem 7. *If $\mathbb{D}$ is totally bounded (Assumption 1) and η satisfies Assumption 3, then ACAL with $\delta_i = 2^{-i}$ achieves an excess expected mistake bound $\bar{M}_T - M_T^* = o(T)$, and if additionally $\theta_{\mathbb{D}}(\epsilon) = o(1/\epsilon)$, then ACAL makes an expected number of queries $\bar{Q}_T = o(T)$.*

The proof of Theorem 7 essentially follows from a combination of the reasoning for Theorem 2 and Theorem 8 below. Its proof is omitted.

Theorem 8. *If Assumptions 2 and 5 are satisfied, then ACAL achieves an expected excess number of mistakes*

$$\bar{M}_T - M_T^* = \tilde{O}\left(T^{\frac{(\alpha+2)m+1}{(\alpha+2)(m+1)}} \log\left(\frac{1}{\delta_{\lfloor \log(T) \rfloor}}\right) + \sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i\right),$$

and an expected number of queries

$$\bar{Q}_T = \tilde{O}\left(\theta_{\mathbb{D}}(\epsilon_T) T^{\frac{(\alpha+2)(m+1)-\alpha}{(\alpha+2)(m+1)}} \log\left(\frac{1}{\delta_{\lfloor \log(T) \rfloor}}\right) + \sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i\right),$$

where $\epsilon_T = T^{-\frac{\alpha}{(\alpha+2)(m+1)}}$.

Proof. Fix any $i \in \mathbb{N}$, and we will focus on bounding the expected excess number of mistakes and expected number of queries for the values $t \in \{2^i + 1, \ldots, 2^{i+1}\}$. The result will then follow from this simply by summing this over values of $i \leq \log(T)$.

The predictions for $t \in \{2^i + 1, \ldots, 2^{i+1}\}$ are made by $\hat{h}_{t-1}$. Lemma 1 implies that with probability at least $1 - \delta_i$, every $t \in \{2^i + 1, \ldots, 2^{i+1}\}$ has $\forall h \in \mathbb{C}[\mathcal{L}_{t-1}]$ with $\hat{\text{er}}(h; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) - \hat{\text{er}}(h^*; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) \leq \hat{\mathcal{E}}_{t-1}(\mathcal{L}_{t-1}, \mathcal{Q}_{t-1})$ (and therefore in particular for $\hat{h}_{t-1}$)

$$\begin{aligned}\sum_{s=2^i+1}^{t-1} \text{er}_s(h) - \text{er}_s(h^*) &\leq K_1 \cdot (t - 2^i) \cdot \left(\frac{d \log((t-2^i)/\delta_i)}{t - 2^i}\right)^{\frac{\alpha+1}{\alpha+2}} \\ &\leq K_1 \cdot t^{\frac{1}{\alpha+2}} \cdot (d \log(t/\delta_i))^{\frac{\alpha+1}{\alpha+2}}. \end{aligned} \tag{4}$$

for some finite constant K_1.

Fix some value $\epsilon > 0$, and enumerate the elements of $\mathbb{D}_\epsilon = \{P_1, P_2, \ldots, P_{|\mathbb{D}_\epsilon|}\}$. Then let $\mathbb{D}_{\epsilon,k} = \{P \in \mathbb{D} : k = \operatorname{argmin}_{j \le |\mathbb{D}_\epsilon|} \|P_j - P\|\}$, breaking ties arbitrarily in the argmin. This induces a (Voronoi) partition $\{\mathbb{D}_{\epsilon,k} : k \le |\mathbb{D}_\epsilon|\}$ of $\mathbb{D}$.

Rewriting (4) in terms of this partition, we have

$$\sum_{k=1}^{|\mathbb{D}_\epsilon|} \sum_{\substack{s \in \{2^i+1,\ldots,t-1\}: \\ \mathcal{D}_s \in \mathbb{D}_{\epsilon,k}}} \operatorname{er}_s(h) - \operatorname{er}_s(h^*) \le K_1 \cdot (t)^{\frac{1}{\alpha+2}} \cdot (d \log(t/\delta_i)) \,.$$

This means that, for any $k \le |\mathbb{D}_\epsilon|$, we have

$$\begin{aligned}
&(\operatorname{er}_{P_k}(h) - \operatorname{er}_{P_k}(h^*)) \cdot \left|\left\{s \in \{2^i+1,\ldots,t-1\} : \mathcal{D}_s \in \mathbb{D}_{\epsilon,k}\right\}\right| \\
&+ \sum_{s=2^i+1}^{t-1} (\operatorname{er}_s(h) - \operatorname{er}_s(h^*)) \cdot \mathbb{I}_{\mathbb{D}\setminus\mathbb{D}_{\epsilon,k}}(\mathcal{D}_s) \\
&\le K_1 \cdot (t)^{\frac{1}{\alpha+2}} \cdot (d \log(t/\delta_i)) + 2\epsilon \left|\left\{s \in \{2^i+1,\ldots,t-1\} : \mathcal{D}_s \in \mathbb{D}_{\epsilon,k}\right\}\right| .
\end{aligned}$$

Abbreviating by $k(s)$ the value of $k \le |\mathbb{D}_\epsilon|$ with $\mathcal{D}_s \in \mathbb{D}_{\epsilon,k}$, we have that

$$\begin{aligned}
&\operatorname{er}_t(h) - \operatorname{er}_t(h^*) \\
&\le 2\epsilon + \operatorname{er}_{P_{k(t)}}(h) - \operatorname{er}_{P_{k(t)}}(h^*) \\
&\le 2\epsilon + \frac{2\epsilon \left|\left\{s \in \{2^i+1,\ldots,t-1\} : k(s) = k(t)\right\}\right| + K_1 \cdot (t)^{\frac{1}{\alpha+2}} \cdot (d \log(t/\delta_i))}{\max\left\{1, |\{s \in \{2^i+1,\ldots,t-1\} : k(s) = k(t)\}|\right\}} \\
&\le 4\epsilon + \frac{2K_1 \cdot (t)^{\frac{1}{\alpha+2}} \cdot (d \log(t/\delta_i))}{|\{s \in \{2^i+1,\ldots,t\} : k(s) = k(t)\}|} .
\end{aligned} \tag{5}$$

Applying (5) simultaneously for all $t \in \{2^i+1,\ldots,2^{i+1}\}$ for $h = \hat{h}_{t-1}$, we have

$$\begin{aligned}
&\bar{M}_T - M_T^* \le 4\epsilon T + \sum_{i=0}^{\lfloor \log(T) \rfloor} 2^i \delta_i + \\
&2K_1 \cdot T^{\frac{1}{\alpha+2}} \cdot \log(T) \left(d \log(T/\delta_{\lfloor \log(T) \rfloor})\right) \sum_{i=0}^{\lfloor \log(T) \rfloor} \sum_{k=1}^{|\mathbb{D}_\epsilon|} \sum_{u=1}^{|\{t \in \{2^i+1,\ldots,2^{i+1}\} : k(t) = k\}|} \frac{1}{u} \\
&\le 4\epsilon T + \sum_{i=0}^{\lfloor \log(T) \rfloor} 2^i \delta_i + \\
&2K_1 \cdot T^{\frac{1}{\alpha+2}} \cdot \log(T) \left(d \log(T/\delta_{\lfloor \log(T) \rfloor})\right) \log^2(2T) |\mathbb{D}_\epsilon| . \\
&= O\left(\epsilon T + \epsilon^{-m} T^{\frac{1}{\alpha+2}} d \log^3(T) \log(1/\delta_{\lfloor \log(T) \rfloor}) + \sum_{i=0}^{\lfloor \log(T) \rfloor} 2^i \delta_i\right) .
\end{aligned}$$

Taking $\epsilon = T^{-\frac{\alpha+1}{(\alpha+2)(m+1)}}$, this shows that

$$\bar{M}_T - M_T^* = O\left(T^{\frac{(\alpha+2)m+1}{(\alpha+2)(m+1)}} d \log^3(T) \log(1/\delta_{\lfloor \log(T) \rfloor}) + \sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i\right) .$$

We can bound $\bar{Q}_T$ in a similar fashion as follows. Fix any $i \leq \log(T)$. Lemma 1 implies that with probability at least $1-\delta_i$, for every $t \in \{2^i+1,\ldots,2^{i+1}\}$, letting $\bar{\mathcal{E}}_t = 4\epsilon + \frac{2K_1 \cdot t^{\frac{1}{\alpha+2}} d\log(t/\delta_{\lfloor \log(t) \rfloor})}{|\{s \in \{2^i+1,\ldots,t\} : k(s)=k(t)\}|}$, we have

$$
\begin{aligned}
&\mathbb{P}(\text{request } Y_t | \mathcal{L}_{t-1}, \mathcal{Q}_{t-1}) \\
&\leq \mathbb{P}\left(X_t \in \text{DIS}\left(\{h \in \mathbb{C}[\mathcal{L}_{t-1}] : \hat{\text{er}}(h; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) - \hat{\text{er}}(h^*; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) \leq \hat{\mathcal{E}}_{t-1}(\mathcal{L}_{t-1}, \mathcal{Q}_{t-1})\}\right) \middle| \mathcal{L}_{t-1}, \mathcal{Q}_{t-1}\right) \\
&\leq \mathbb{P}\left(X_t \in \text{DIS}\left(\{h \in \mathbb{C} : \text{er}_t(h) - \text{er}_t(h^*) \leq \bar{\mathcal{E}}_t\}\right)\right) \\
&\leq \mathbb{P}\left(X_t \in \text{DIS}\left(\left\{h \in \mathbb{C} : P_t(x : h(x) \neq h^*(x)) \leq K_2 \cdot \bar{\mathcal{E}}_t^{\frac{\alpha}{\alpha+1}}\right\}\right)\right) \\
&\leq \theta_{\mathbb{D}}\left(\bar{\mathcal{E}}_t^{\frac{\alpha}{\alpha+1}}\right) \cdot K_3 \cdot \bar{\mathcal{E}}_t^{\frac{\alpha}{\alpha+1}},
\end{aligned}
$$

where the third inequality above is due to Assumption 5.

Applying this simultaneously to all $i \leq \log(T)$ and $t \in \{2^i+1,\ldots,2^{i+1}\}$, we have, for $\bar{\epsilon}_T = \epsilon + T^{-\frac{\alpha+1}{\alpha+2}}$,

$$
\begin{aligned}
\bar{Q}_T &\leq \sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i + \theta_{\mathbb{D}}\left(\bar{\epsilon}_T^{\frac{\alpha}{\alpha+1}}\right) K_4 d\log(T/\delta_{\lfloor \log(T) \rfloor}) \sum_{i=0}^{\lfloor \log(T) \rfloor} \sum_{k=1}^{|\mathbb{D}_\epsilon|} \sum_{u=1}^{|\{t \in \{2^i+1,\ldots,2^{i+1}\} : k(t)=k\}|} \left(\max\left\{\epsilon, T^{\frac{1}{\alpha+2}} \frac{1}{u}\right\}\right)^{\frac{\alpha}{\alpha+1}} \\
&\leq \sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i + \theta_{\mathbb{D}}\left(\bar{\epsilon}_T^{\frac{\alpha}{\alpha+1}}\right) \cdot K_5 \cdot d\log(1/\delta_{\lfloor \log(T) \rfloor}) \log^2(T) \cdot \left(\epsilon^{\frac{\alpha}{\alpha+1}} T + |\mathbb{D}_\epsilon| T^{\frac{\alpha}{(\alpha+2)(\alpha+1)}} \left(\frac{T}{|\mathbb{D}_\epsilon|}\right)^{\frac{1}{\alpha+1}}\right) \\
&= O\left(\sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i + \theta_{\mathbb{D}}\left(\bar{\epsilon}_T^{\frac{\alpha}{\alpha+1}}\right) \log(1/\delta_{\lfloor \log(T) \rfloor}) \log^2(T) \cdot \left(\epsilon^{\frac{\alpha}{\alpha+1}} T + \epsilon^{-m\frac{\alpha}{\alpha+1}} T^{\frac{2}{\alpha+2}}\right)\right).
\end{aligned}
$$

Taking $\epsilon = \epsilon_T^{\frac{\alpha+1}{\alpha}} = T^{-\frac{\alpha+1}{(\alpha+2)(m+1)}}$, we have

$$
\bar{Q}_T = O\left(\sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i + \theta_{\mathbb{D}}\left(\epsilon_T\right) \log(1/\delta_{\lfloor \log(T) \rfloor}) \log^2(T) \cdot T^{\frac{(\alpha+2)(m+1)-\alpha}{(\alpha+2)(m+1)}}\right).
$$

□

We immediately have the following corollary for a specific δ_i sequence.

Corollary 2. *With $\delta_i = 2^{-i}$ for all i in ACAL, the algorithm achieves expected excess number of mistakes $\bar{M}_T - M_T^* = \tilde{O}\left(T^{\frac{(\alpha+2)m+1}{(\alpha+2)(m+1)}}\right)$, and expected number of label requests $\bar{Q}_T = \tilde{O}\left(\theta_{\mathbb{D}}(\epsilon_T) \cdot T^{\frac{(\alpha+2)(m+1)-\alpha}{(\alpha+2)(m+1)}}\right)$, where $\epsilon_T = T^{-\frac{\alpha}{(\alpha+2)(m+1)}}$.*

5.5 Minimax Lower Bounds

Theorem 9. *For any $\mathbb{C}$ as in Theorem 4, for any active learning algorithm, $\exists$ a set $\mathbb{D}$ satisfying Assumption 2, a conditional distribution η, such that Assumption 5 is satisfied, and a sequence of distributions $\{\mathcal{D}_t\}_{t=1}^T$ in $\mathbb{D}$ such that the $\bar{M}_T$ and $\bar{Q}_T$ achieved by the learning algorithm satisfy*

$$
\begin{aligned}
\bar{M}_T - M_T^* &= \Omega\left(T^{\frac{1+m\alpha}{\alpha+2+m\alpha}}\right) \\
\text{and } \bar{M}_T - M_T^* &= O\left(T^{\frac{1+m\alpha}{\alpha+2+m\alpha}}\right) \implies \bar{Q}_T = \Omega\left(T^{\frac{2+m\alpha}{\alpha+2+m\alpha}}\right).
\end{aligned}
$$

Proof. Fix any $T \in \mathbb{N}$, and any particular active learning algorithm $\mathcal{A}$. We construct a set of distributions tailored for these, as follows. Let $\kappa = (\alpha+1)/\alpha$. Let $\epsilon = T^{-\frac{\kappa}{2\kappa-1+m}}$, $M = T^{\frac{m}{2\kappa+m-1}} = \epsilon^{-m/\kappa}$, and $K = T^{\frac{2\kappa-1}{2\kappa+m-1}} = T/M$.

Inductively define a sequence $\{b_k\}_{k=1}^{\infty}$ as follows. Let $b_1 = 0$, $b_2 = 1$. For any integer $k \geq 3$, given that values of $b_1, b_2, \ldots, b_{k-1}$, $\eta_3, \ldots, \eta_{k-1}$, $D_3, \ldots D_{k-1}$, and $X_1, X_2, \ldots, X_{(k-3)K}$ have already been defined, it is known (Hanneke, 2011b) that for any active learning algorithm (possibly randomized) there exists a value b_k such that, for the distribution D_k with $D_k(\{x_{b_1,b_2,\ldots,b_{k-1}}\}) = \epsilon^{1/\kappa} = 1 - D_k(\{x_{b_1}\})$, there is a label distribution $\eta_k(x) = P(Y = 1|X = x)$ having $\eta_k(x_{b_1}) = 1$ and inducing $h^*(x_{b_1,b_2,\ldots,b_{k-1}}) = b_k$, which also satisfies Tsybakov noise with parameters c and α under distribution D_k: namely, $\eta_k(x_{b_1,b_2,\ldots,b_{k-1}}) = \frac{1}{2}\left(1 + (2b_k - 1)\epsilon^{\frac{\kappa-1}{\kappa}}\right)$. Furthermore, Hanneke (2011b) shows that this b_k can be chosen so that, for some $N = \Omega\left(\epsilon^{\frac{2}{\kappa}-2}\right)$, after observing any number fewer than N random labeled observations (X, Y) with $X = x_{b_1,b_2,\ldots,b_{k-1}}$, if $\hat{h}_n$ is the algorithm's hypothesis, then $\mathbb{E}[\mathrm{er}(\hat{h}_n) - \mathrm{er}(h^*)] > \epsilon$, where the error rate is evaluated under η_k and D_k. In particular, this means that if the unlabeled samples are distributed according to D_k, then with any fewer than N label requests, the expected excess error rate will be greater than ϵ. But this also means that with any fewer than $\Omega(\epsilon^{-1/\kappa}N) = \Omega(\epsilon^{\frac{1}{\kappa}-2}) = \Omega(K)$ unlabeled examples sampled according to D_k, the expected excess error rate will be greater than ϵ.

Thus, to define the value b_k given the already-defined values $b_1, b_2, \ldots, b_{k-1}$, we consider $X_{(k-3)K+1}$, $X_{(k-3)K+2}$, …, $X_{(k-2)K}$ i.i.d. D_k, independent from the other $X_1, \ldots, X_{(k-3)K}$ variables, and consider the values of b_k and η_k mentioned above, but defined for the active learning algorithm that feeds the stream $X_1, X_2, \ldots, X_{(k-3)K}$ into $\mathcal{A}$ before feeding in the samples from D_k. Thus, in this perspective, these $X_1, X_2, \ldots, X_{(k-3)K}$ random variables, and their labels (which $\mathcal{A}$ may request), are considered *internal* random variables in this active learning algorithm we have defined. This completes the inductive definition.

Now for the original learning problem we are interested in, we take as our fixed label distribution an η with $\eta(x_{b_1}) = 1$ and $\forall k \geq 2$, $\eta(x_{b_1,b_2,\ldots,b_{k-1}}) = \eta_k(x_{b_1,b_2,\ldots,b_{k-1}})$, and defined arbitrariliy elsewhere. Thus, for any D_k, this satisfies Tsybakov noise with the given c and α parameters.

We define the family $\mathbb{D}$ of distributions as $\{D_3, , D_4, \ldots, D_{M+2}\}$ for $M = T^{\frac{m}{2\kappa+m-1}} = \epsilon^{-m/\kappa}$ as above. Since these D_i are each separated by distance exactly $\epsilon^{1/\kappa}$, $\mathbb{D}$ satisfies the constraint on its cover sizes.

The sequence of data points will be the $X_1, X_2, \ldots, X_T$ sequence defined above, and the corresponding sequence of distributions has $\mathcal{D}_1 = \mathcal{D}_2 = \cdots = \mathcal{D}_K = D_3$, $\mathcal{D}_{K+1} = \mathcal{D}_{K+2} = \cdots = \mathcal{D}_{2K} = D_4$, and so on, up to $\mathcal{D}_{(M-1)K+1} = \mathcal{D}_{(M-1)K+2} = \cdots \mathcal{D}_T = D_{M+2}$.

Now applying the stated result of Hanneke (2011b) used in the definition of the sequence, for any $1 \leq t \leq \min\{\epsilon^{-1/\kappa}N, K\}$, and any $k < M$, denoting by $\hat{h}_{kK+t-1}$ the classifier produced by $\mathcal{A}$ after processing $kK + t - 1$ examples from this stream, $\mathbb{E}\left[\mathrm{er}_{\mathcal{D}_{kK+t}}(\hat{h}_{kK+t-1})\right] - \mathrm{er}_{\mathcal{D}_{kK+t}}(h^*) > \epsilon = T^{-\frac{\kappa}{2\kappa+m-1}}$.

Since $\min\{\epsilon^{-1/\kappa}N, K\} = \Omega(K)$, the expected excess number of mistakes is

$$\begin{aligned}
\hat{M}_T - M_T^* &= \sum_{k=0}^{M-1}\sum_{t=1}^{K}\mathbb{E}\left[\mathrm{er}_{\mathcal{D}_{kK+t}}(\hat{h}_{kK+t-1})\right] - \mathrm{er}_{\mathcal{D}_{kK+t}}(h^*) \\
&\geq \sum_{k=0}^{M-1}\sum_{t=1}^{\min\{\epsilon^{-1/\kappa}N,K\}}\mathbb{E}\left[\mathrm{er}_{\mathcal{D}_{kK+t}}(\hat{h}_{kK+t-1})\right] - \mathrm{er}_{\mathcal{D}_{kK+t}}(h^*) \\
&\geq \sum_{k=0}^{M-1}\sum_{t=1}^{\min\{\epsilon^{-1/\kappa}N,K\}}\epsilon \\
&= \Omega\left(M \cdot K \cdot \epsilon\right) = \Omega\left(M \cdot (T/M) \cdot T^{-\frac{\kappa}{2\kappa+m-1}}\right) = \Omega\left(T^{\frac{\kappa+m-1}{2\kappa+m-1}}\right).
\end{aligned}$$

Similarly, applying the stated result of Hanneke (2011b) regarding the number of samples of labels for the point $x_{b_1,b_2,\ldots,b_{k-1}}$ to achieve excess error ϵ being larger than N, we see that in order to achieve this $\hat{M}_T - M_T^* = O\left(T^{\frac{\kappa+m-1}{2\kappa+m-1}}\right)$, we need that at least some constant fraction of these M segments receive an expected number of queries $\Omega(N)$, so that we will need $\hat{Q}_T = \Omega(M \cdot N) = \Omega\left(T^{\frac{2\kappa+m-2}{2\kappa+m-1}}\right)$. □

6 Querying before Predicting

One interesting alternative to the above framework is to allow the learner to make a label request *before* making its label predictions. From a practical perspective, this may be more desirable and in many cases quite realistic. From a theoretical perspective, analysis of this alternative framework essentially separates out the mistakes due to over-confidence from the mistakes due to recognized uncertainty. In some sense, this is related to the KWIK model of learning of (Li, Littman, and Walsh, 2008).

Analyzing the above procedures in this alternative model yields several interesting details. Specifically, consider the following natural modifications to the above procedures. We refer to the algorithm LAC as the same sequence of steps as CAL, except with Step 4 removed, and an additional step added after Step 8 as follows. In the case that we requested the label Y_t, we predict Y_t, and otherwise we predict $\hat{h}_t(X_t)$. Similarly, we define the algorithm ALAC as having the same sequence of steps as ACAL, except with Step 4 removed, and an additional step added after Step 11 as follows. In the case that we requested the label Y_t, we predict Y_t, and otherwise we predict $\hat{h}_t(X_t)$.

The analysis of the number of queries made by LAC in this setting remains essentially unchanged. However, if we consider running LAC in the realizable case, then the total number of mistakes in the entire sequence will be *zero*. As above, for any example for which LAC does not request the label, every classifier in the version space agrees with the target function's label, and therefore the inferred label will be correct. For any example that LAC requests the label of, in the setting where queries are made *before* predictions, we simply use the label itself as our prediction, so that LAC certainly does not make a mistake in this case.

On the other hand, the analysis of ALAC in this alternative setting when we have noisy labels can be far more subtle. In particular, because the version space is only guaranteed to contain the best classifier *with high confidence*, there is still a small probability of making a prediction that disagrees with the best classifier h^* on each round that we do not request a label. So controlling the number of mistakes in this setting comes down to controlling the probability of removing h^* from the version space. However, this confidence parameter appears in the analysis of the number of queries, so that we have a natural trade-off between the number of mistakes and the number of label requests.

Formally, for any given nonincreasing sequence δ_i in $(0,1)$, under Assumptions 2 and 5, ALAC achieves an expected excess number of mistakes $\bar{M}_T - M_T^* \leq \sum_{i=1}^{\lfloor \log(T) \rfloor} \delta_i 2^i$, and an expected number of queries $\bar{Q}_T = \tilde{O}\left(\theta_{\mathbb{D}}(\epsilon_T) \cdot T^{\frac{(\alpha+2)(m+1)-\alpha}{(\alpha+2)(m+1)}} \log\left(\frac{1}{\delta_{\lfloor \log(T) \rfloor}}\right) + \sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i\right)$, where $\epsilon_T = T^{-\frac{\alpha}{(\alpha+2)(m+1)}}$. In particular, given any nondecreasing sequence M_T, we can set this δ_i sequence to maintain $\bar{M}_T - M_T^* \leq M_T$ for all T.

7 Discussion

What is not implied by the results above is any sort of *trade-off* between the number of mistakes and the number of queries. Intuitively, such a trade-off should exist; however, as CAL lacks any parameter to adjust the behavior with respect to this trade-off, it seems we need a different approach to address that question.

In the batch setting, the analogous question is the trade-off between the number of label requests and the number of unlabeled examples needed. In the realizable case, that trade-off is tightly characterized by Dasgupta's *splitting index* analysis (Dasgupta, 2005). It would be interesting to determine whether the splitting index tightly characterizes the mistakes-vs-queries trade-off in this stream-based setting as well.

In the batch setting, in which unlabeled examples are considered free, and performance is only measured as a function of the number of label requests, Balcan, Hanneke, and Vaughan (2010) have found that there is an important distinction between the *verifiable* label complexity and the *unverifiable* label complexity. In particular, while the former is sometimes no better than passive learning, the latter can always provide improvements for VC classes. Is there such a thing as unverifiable performance measures in the stream-based setting? To be concrete, we have the following open problem. Is there a method for every VC class that achieves $O(\log(T))$ mistakes and $o(T)$ queries in the realizable case?

Bibliography

Anthony, M. and Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press.

Balcan, M.-F., Hanneke, S., and Vaughan, J. W. (2010). The true sample complexity of active learning. *Machine Learning*, **80**(2–3), 111–139.

Bartlett, P. L. (1992). Learning with a slowly changing distribution. In *Proceedings of the fifth annual workshop on Computational learning theory*, COLT '92, pages 243–252.

Barve, R. D. and Long, P. M. (1997). On the complexity of learning from drifting distributions. *Inf. Comput.*, **138**(2), 170–193.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. (1989). Learnability and the vapnik-chervonenkis dimension. *Journal of the Association for Computing Machinery*, **36**(4), 929–965.

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

Crammer, K., Mansour, Y., Even-Dar, E., and Vaughan, J. W. (2010). Regret minimization with concept drift. In *COLT*, pages 168–180.

Dasgupta, S. (2005). Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems 18*.

Dasgupta, S., Hsu, D., and Monteleoni, C. (2007). A general agnostic active learning algorithm. Technical Report CS2007-0898, Department of Computer Science and Engineering, University of California, San Diego.

Dasgupta, S., Kalai, A., and Monteleoni, C. (2009). Analysis of perceptron-based active learning. *Journal of Machine Learning Research*, **10**, 281–299.

Dekel, O., Gentile, C., and Sridharam, K. (2010). Robust selective sampling from single and multiple teachers. In *Conference on Learning Theory*.

Freund, Y. and Mansour, Y. (1997). Learning under persistent drift. In *Proceedings of the Third European Conference on Computational Learning Theory*, EuroCOLT '97, pages 109–118.

Hanneke, S. (2007). A bound on the label complexity of agnostic active learning. In *Proceedings of the 24^{th} International Conference on Machine Learning*.

Hanneke, S. (2011a). Activized learning: Transforming passive to active with improved label complexity. *Manuscript*.

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

Haussler, D., Littlestone, N., and Warmuth, M. (1994). Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, **115**, 248–292.

Koltchinskii, V. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, **34**(6), 2593–2656.

Li, L., Littman, M. L., and Walsh, T. J. (2008). Knows what it knows: A framework for self-aware learning. In *International Conference on Machine Learning*.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, **2**, 285–318.

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2008). Domain adaptation with multiple sources. In *In Advances in Neural Information Processing Systems (NIPS)*, pages 1041–1048.

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. In *COLT*.

van de Geer, S. (2000). *Empirical Processes in M-Estimation (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press.

Vapnik, V. (1982). *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York.