, , 1–18 ()

The Complexity of Learning According to Two Models of a Drifting Environment

PHILIP M. LONG plong@comp.nus.edu.sg
Department of Computer Science, National University of Singapore, Singapore 119260, Republic of Singapore.

Editor: Jonathan Baxter and Nicolò Cesa-Bianchi

Abstract. We show that a $\frac{c\epsilon^3}{\mathrm{VCdim}(\mathcal{F})}$ bound on the rate of drift of the distribution generating the examples is sufficient for agnostic learning to relative accuracy ϵ, where $c > 0$ is a constant; this matches a known necessary condition to within a constant factor. We establish a $\frac{c\epsilon^2}{\mathrm{VCdim}(\mathcal{F})}$ sufficient condition for the realizable case, also matching a known necessary condition to within a constant factor.

We provide a relatively simple proof of a bound of $O\left(\frac{1}{\epsilon^2}\left(\mathrm{VCdim}(\mathcal{F}) + \log\frac{1}{\delta}\right)\right)$ on the sample complexity of agnostic learning in a fixed environment.

Keywords: Computational learning theory, concept drift, context-sensitive learning, prediction, PAC learning, agnostic learning, uniform convergence, VC theory.

1. Introduction

Learning often takes place in a gradually changing environment. This phenomenon has been studied theoretically by assuming that the function to be learned, the distribution generating the examples, or both, change at most a certain amount between examples (see [14, 3, 5, 6]).[1]

In this paper, we study the problem of learning functions from some set X to $\{0,1\}$ ("concepts") using two models of a drifting environment. In the first [3], it is assumed that examples $(x_1, y_1), (x_2, y_2), \ldots$ are generated independently at random from a sequence of joint distributions over $X \times \{0,1\}$, and the only constraint is that consecutive pairs of distributions have small total variation distance. If this distance is always at most Δ, then the sequence of distributions is called Δ-gradual. For each t, the learning algorithm must output a hypothesis h_t using only the first $t-1$ examples. For some concept class $\mathcal{F}$ and drift rate Δ, if, for any sequence of Δ-gradual joint distributions, for large enough t, the probability that $h_t(x_t) \neq y_t$ is at most ϵ more than the minimum such probability from among $f \in \mathcal{F}$, then we say that $\mathcal{F}$ is (ϵ, Δ)-*trackable in the agnostic case.*

The second model of learning in a drifting environment [14, 3, 5] is obtained from the above by adding the requirement that each distribution P_t has some $f_t \in \mathcal{F}$ such that the probability that the pair (x_t, y_t) drawn according to P_t has $f_t(x_t) = y_t$ is 1. Here, if, for large enough t, the probability that $h_t(x_t) \neq y_t$ is at most ϵ, we say that $\mathcal{F}$ is (ϵ, Δ)-*trackable in the realizable case.*

In this paper, we show that there is a constant $c > 0$ such that a $\frac{c\epsilon^3}{\mathrm{VCdim}(\mathcal{F})}$ bound on Δ is sufficient for $\mathcal{F}$ to be (ϵ, Δ)-trackable in the agnostic case, and a

$\frac{c\epsilon^2}{\text{VCdim}(\mathcal{F})}$ bound is sufficient for the realizable case. This work continues an existing line of research [14, 3, 5, 6], and matches known necessary conditions for both the agnostic [6] and realizable [3] cases to within a constant factor, closing log-factor gaps. Note that both models allow for variation both in the target and in the marginal distribution on the domain elements; some previous work addressed these two types of changes separately.

The agnostic drift analysis uses a technique called Chaining from Empirical Process Theory (see [23, 24]). We defer a high-level description of this technique until later in the paper when appropriate context is available.

In the realizable case, as in [14, 3, 5], we consider an algorithm based on the one-inclusion graph algorithm [13], which was originally designed for learning concepts in a fixed environment. To determine $h(x_m)$ from some sample

$$(x_1, y_1), ..., (x_{m-1}, y_{m-1}),$$

the original algorithm constructs a graph whose vertices are

$$\{(f(x_1), ..., f(x_m)) : f \in \mathcal{F}\}$$

and has edges between pairs of vertices that differ in only one component (the "one-inclusion graph").[2] The edges of the graph are then directed, and these orientations are used to determine $h(x_m)$. The analysis involves relating the probability of a mistake for some target f to the maximum (over $x_1, ..., x_m$) of the outdegree for the vertex associated with f. Since any one-inclusion graph for $\mathcal{F}$ can be shown to be sparse relative to $\text{VCdim}(\mathcal{F})$, the edges can be directed so that the out-degree of any vertex is at most $\text{VCdim}(\mathcal{F})$ [13]. In [14, 5], the vertex set was expanded to include elements of $\{0,1\}^m$ that are within some Hamming distance of elements of $\{(f(x_1), ..., f(x_m)) : f \in \mathcal{F}\}$; these graphs also can be shown to be sparse. The main new idea in this paper's realizable drift analysis is to show, for each $\mathcal{F}$, how to direct *all* the edges of the m-dimensional hypercube so that the outdegree of each vertex is bounded appropriately in terms of its distance to the closest element of $\{(f(x_1), ..., f(x_m)) : f \in \mathcal{F}\}$ as well as the VC-dimension of $\mathcal{F}$.

1.1. Agnostic learning in a fixed environment

In the standard agnostic learning model [11, 17], random examples

$$(x_1, y_1), ..., (x_m, y_m)$$

are drawn from an arbitrary joint distribution P, and the learner's goal is to output a function h such that probability that $h(x) \neq y$ for another pair (x, y) drawn according to P is nearly as small as that of the best function in $\mathcal{F}$.

We give a proof that, in a fixed environment, for any concept class $\mathcal{F}$,

$$O\left(\frac{1}{\epsilon^2}\left(\text{VCdim}(\mathcal{F}) + \log\frac{1}{\delta}\right)\right)$$

examples are sufficient for an algorithm to, with probability $1-\delta$, output a hypothesis whose error is at most ϵ worse than the best in $\mathcal{F}$. This bound, which also follows from previous work of Talagrand [30], improves on the bound of

$$O\left(\frac{1}{\epsilon^2}\left(\mathrm{VCdim}(\mathcal{F})\log\frac{1}{\epsilon}+\log\frac{1}{\delta}\right)\right)$$

that follows from Vapnik and Chervonenkis' results (see [11]), and matches Simon's general lower bound [29] to within a constant factor for each concept class $\mathcal{F}$. Our constants are greater than Talagrand's, but our proof is simpler and more elementary.

2. Preliminaries

Fix a countable set X. Denote the reals by $\mathbf{R}$, and the natural numbers by $\mathbf{N}$.

An *example* is an element of $X \times \{0,1\}$, and a *sample* is a finite sequence of examples. A *learning algorithm* takes a sample as input, and outputs a *hypothesis*, which is a function from X to $\{0,1\}$. We will also consider randomized learning algorithms, which can be modelled as deterministic functions of another random input along with the sample.

For a real-valued function g defined on Z, and $\vec{z} \in Z^m$, define

$$\hat{\mathbf{E}}_{\vec{z}}(g) = \frac{1}{m}\sum_{i=1}^{m} g(z_i).$$

The VC-dimension of a set $G \subseteq \{0,1\}^m$ is the length of the longest sequence $i_1,...,i_d$ of indices such that $\{(g_{i_1},...,g_{i_d}) : g \in G\} = \{0,1\}^d$. The VC-dimension of a set $\mathcal{G}$ of functions from X to $\{0,1\}$ is the maximum, over $m \in \mathbf{N}, \vec{x} \in X^m$, of the VC-dimension of $\{(g(x_1),...,g(x_m)) : g \in \mathcal{G}\}$.

The metric d_{TV} on probability distributions is defined by

$$d_{TV}(P,Q) = 2\sup_E |P(E) - Q(E)|.$$

Say a sequence $P_1, P_2, ...$ of probability distributions is Δ-gradual if for each $t \in \mathbf{N}$, $d_{TV}(P_t, P_{t+1}) \le \Delta$.

For a learning algorithm A, we say that a sample $(x_1,y_1),...,(x_m,y_m)$ and randomization r *cause a mistake for* A if A, given $(x_1,y_1),...,(x_{m-1},y_{m-1})$ and r, outputs a hypothesis h for which $h(x_m) \ne y_m$.

Recall that the Hamming distance, which we will denote by ρ, is defined by $\rho(\vec{v},\vec{w}) = \sum_i |v_i - w_i|$. For $m \in \mathbf{N}$, $F \subseteq \{0,1\}^m, \vec{v} \in \{0,1\}^m$, define $\rho(\vec{v},F) = \min\{\rho(\vec{v},\vec{f}) : \vec{f} \in F\}$. For each $k \in \{0,...,m\}$, define $\rho_k(F) = \{\vec{v} \in \{0,1\}^m : \rho(\vec{v},F) = k\}$.

Both analyses will use Fubini's Theorem.

LEMMA 1 (SEE [26]) *Choose countable sets* Z_1 *and* Z_2, *a function* $f : Z_1 \times Z_2 \to [0,1]$ *and probability distributions* D_1 *over* Z_1 *and* D_2 *over* Z_2. *Then*

$$\int_{Z_1 \times Z_2} f(z_1, z_2)\, d(D_1 \times D_2)(z_1, z_2) = \int_{Z_1} \left(\int_{Z_2} f(z_1, z_2)\, dD_2(z_2) \right) dD_1(z_1)$$
$$= \int_{Z_2} \left(\int_{Z_1} f(z_1, z_2)\, dD_1(z_1) \right) dD_2(z_2).$$

We will also use the standard Hoeffding bound.

LEMMA 2 (SEE [23]) *Let* $Y_1, ..., Y_m$ *be independent random variables taking values in* $[a_1, b_1], ..., [a_m, b_m]$ *respectively. Then*

$$\mathbf{Pr}\left(\left| \left(\sum_{i=1}^m Y_i \right) - \left(\sum_{i=1}^m \mathbf{E}(Y_i) \right) \right| > \eta \right) \leq 2 \exp\left(\frac{-2\eta^2}{\sum_{i=1}^m (b_i - a_i)^2} \right).$$

3. Agnostic Learning

In this section, we consider agnostic learning in both fixed and drifting environments. We begin with a fixed environment.

3.1. Fixed environment

Choose a class $\mathcal{F}$ of functions from X to $\{0,1\}$. For a probability distribution P on $X \times \{0,1\}$ and a function h from X to $\{0,1\}$, the error of h with respect to P, denoted by $\mathbf{er}_P(h)$, is $P\{(x,y) : h(x) \neq y\}$. A learning algorithm A is said to (ϵ, δ)-agnostically learn $\mathcal{F}$ from m examples if for all distributions P on $X \times \{0,1\}$,

$$P^m \left\{ \vec{z} : \mathbf{er}_P(A(\vec{z})) > \epsilon + \inf_{f \in \mathcal{F}} \mathbf{er}_P(f) \right\} \leq \delta.$$

To set the context, we briefly review the work that our analysis builds on [32, 23, 8, 11].

For each $f \in \mathcal{F}$, define $L_f : X \times \{0,1\} \to \{0,1\}$ by $L_f(x,y) = |f(x) - y|$. Define $L_{\mathcal{F}} = \{L_f : f \in \mathcal{F}\}$. The following reduces the learning problem to that of obtaining uniformly good estimates of the errors of possible hypothesis (i.e. expectations of elements of $L_{\mathcal{F}}$).

LEMMA 3 ([11]) *Choose* $\epsilon, \delta > 0$, $m \in \mathbf{N}$. *If for all distributions* P *on* $X \times \{0,1\}$,

$$P^m \left\{ \vec{z} : \exists g \in L_{\mathcal{F}}, \left| \hat{\mathbf{E}}_{\vec{z}}(g) - \int_{X \times \{0,1\}} g(u)\, dP(u) \right| > \epsilon/2 \right\} \leq \delta$$

then $\mathcal{F}$ *is* (ϵ, δ)*-agnostically learnable from* m *examples.*

The following will also be useful.

LEMMA 4 (SEE [8]) $\mathrm{VCdim}(L_{\mathcal{F}}) \leq \mathrm{VCdim}(\mathcal{F})$.

So now we can concentrate on determining distribution-free bounds, in terms on the VC-dimension, on the number of examples required to obtain uniformly good estimates of the expectations of random variables in some set. Choose some countable[3] set Z (in the learning application, Z will be $X \times \{0,1\}$) and some set $\mathcal{G}$ of functions from Z to $\{0,1\}$ (in the learning application, $\mathcal{G}$ will be $L_{\mathcal{F}}$).

The first lemma bounds the probability that any estimate is inaccurate in terms of the probability that two samples yield substantially different estimates.

LEMMA 5 ([32]) *Choose $\eta > 0$ and $m \in \mathbf{N}$ for which $m \geq 2/\eta^2$ and some probability distribution P on Z. Then*

$$\begin{aligned}
&P^m \left\{ \vec{z} : \exists g \in \mathcal{G}, \left| \hat{\mathbf{E}}_{\vec{z}}(g) - \int_Z g(u)\, dP(u) \right| > \eta \right\} \\
&\leq 2P^{2m}\{(\vec{z}, \vec{u}) : \exists g \in \mathcal{G}, |\hat{\mathbf{E}}_{\vec{z}}(g) - \hat{\mathbf{E}}_{\vec{u}}(g)| > \eta/2\} \\
&= 2P^{2m} \left\{ (\vec{z}, \vec{u}) : \exists g \in \mathcal{G}, \left| \sum_{i=1}^m g(z_i) - g(u_i) \right| > \eta m/2 \right\}.
\end{aligned}$$

The next lemma is an example of the "permutation trick": note that setting $\sigma_i = -1$ has the effect of exchanging z_i and u_i.

LEMMA 6 ([32, 23]) *Choose $\eta > 0$, $m \in \mathbf{N}$ and some probability distribution P on Z. Then if U is the uniform distribution on $\{-1,1\}^m$,*

$$\begin{aligned}
&P^{2m} \left\{ (\vec{z}, \vec{u}) : \exists g \in \mathcal{G}, \left| \sum_{i=1}^m g(z_i) - g(u_i) \right| > \eta m \right\} \\
&\leq \sup_{\vec{z}, \vec{u} \in Z^m} U \left\{ \vec{\sigma} : \exists g \in \mathcal{G}, \left| \sum_{i=1}^m \sigma_i (g(z_i) - g(u_i)) \right| > \eta m \right\}.
\end{aligned}$$

The previous lemma allows us to fix some sequence of $2m$ elements of Z, and restrict our attention to the behaviors of elements of $\mathcal{G}$ on those $2m$ elements.

The following lemma is an immediate consequence of Lemma 2.

LEMMA 7 *Choose $m \in \mathbf{N}$ and $G \subseteq \{0,1\}^{2m}$. Then if U is the uniform distribution over $\{-1,1\}^m$,*

$$U \left\{ \vec{\sigma} : \exists g \in G, \left| \sum_{i=1}^m \sigma_i (g_i - g_{m+i}) \right| > \eta m \right\} \leq 2|G| e^{-\eta^2 m/2}.$$

By combining Lemmas 3, 4, 5, 6, and 7, and applying a bound on $|G|$ in terms of $\mathrm{VCdim}(G)$ [27, 28, 32] in Lemma 7, one gets a bound of

$$O\left(\frac{1}{\epsilon^2}\left(\mathrm{VCdim}(\mathcal{F}) \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right)\right)$$

on the sample complexity of agnostically learning $\mathcal{F}$ [11].

Our argument will take advantage of the following refinement of a slight generalization of Lemma 7, which also follows directly from Lemma 2.

LEMMA 8 *Choose $m, k \in \mathbf{N}$, and suppose that $H \subseteq \mathbf{R}^m$ has the property that each $h \in H$ has $\sum_{i=1}^m h_i^2 \leq k$. Then if U is the uniform distribution over $\{-1,1\}^m$,*

$$U\left\{\sigma : \exists h \in H, \left|\sum_{i=1}^m \sigma_i h_i\right| > \eta m\right\} \leq 2|H| e^{\frac{-\eta^2 m^2}{2k}}.$$

The idea of Lemma 8 is that if all of the elements of H are small, then the variances of the random terms $\sigma_i h_i$ tend to be small, which means that its less likely that any sum of them will stray far from 0 (its expectation).

The following lemma is the heart of our analysis.

LEMMA 9 *Choose $\eta > 0$, and $d \in \mathbf{N}$. Choose an integer $m \geq \frac{278(d+1)}{\eta^2}$ and $G \subseteq \{0,1\}^{2m}$ for which* $\mathrm{VCdim}(G) = d$. *Then if U is the uniform distribution over $\{-1,1\}^m$, for any $\eta > 0$,*

$$U\left\{\vec{\sigma} : \exists g \in G, \left|\frac{1}{m}\sum_{i=1}^m \sigma_i (g_i - g_{m+i})\right| > \eta\right\} \leq 4 \cdot 41^d e^{-\eta^2 m/400}.$$

The proof is a chaining argument. See Pollard's books [23, 24] for others and for further references. The idea is as follows. First, we form a sequence $G_0, ..., G_n$ of approximations to G. The approximations get successively finer until $G_n = G$. Next, we consider the sets $H_1, H_2, ..., H_n$, where each H_j consists of the adjustments that need to be made to G_{j-1} to get the improved approximation G_j. In particular, H_j consists of the differences between each element of G_j and the closest element of G_{j-1}. (See Figure 1.) If we define $H_0 = G_0$, then each element of G is the sum of an element of H_0, an element of H_1, and so on up to an element of H_n. So, loosely speaking, if things are OK for each of the H_j's, then they're OK for G. We will apply Lemma 8 to analyze each of the H_j's.

For relatively large j, H_j consists of those adjustments needed to make an already fine approximation finer. Thus, the elements of H_j are small, and we can use the fact that Lemma 8 provides a better bound in this case. When j is small, since $|H_j| \leq |G_j|$, and G_j is a relatively coarse approximation to G, H_j does not have many elements, which provides partial compensation for the fact that its elements might be large.

We will use the following result due to Haussler, which bounds the number of significantly different elements of a set G in terms of its VC-dimension. This can be used to bound the size of an approximation to G [18].

LEMMA 10 ([12]) *For all $m \in \mathbf{N}$, for all $k \leq m$, if each pair g, h of elements of $G \subseteq \{0,1\}^m$ has $\rho(g,h) > k$, then*

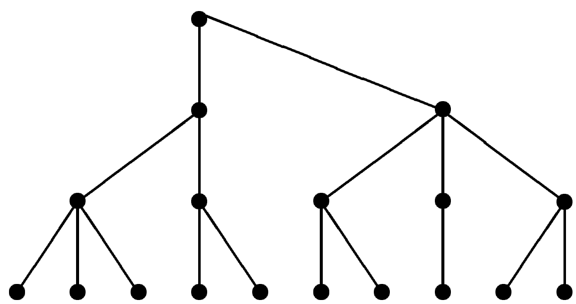

Figure 1. A schematic representation of the G_j's and H_j's from the proof of Lemma 9 in the case $m = 4$. The G_j's, which form increasingly accurate approximations to G, are represented by increasingly dense rows of nodes. For each $j > 0$, an edge is added between the node representing each element of G_j and that representing the closest element of G_{j-1}. If you think of this edge as representing the difference between the two, then each H_j (for $j > 0$) consists of the jth layer of edges.

$$|G| \leq \left(\frac{41m}{k}\right)^{\mathrm{VCdim}(G)}.$$

Proof (of Lemma 9): Let $n = 1 + \lfloor \log_2 m \rfloor$. Construct $G_0, ..., G_n$ as follows. Let G_0 consist of an arbitrary single element of G, and for each $j \in \{1, ..., n\}$, construct G_j by initializing it to G_{j-1}, and as long as there is a $g \in G$ for which $\rho(g, G_j) > m/2^j$, choosing such an g and adding it to G_j. Note that $G_0 \subseteq G_1 \subseteq ... \subseteq G_n = G$. For each $g \in G$ and $j \in \{0, ..., n\}$ choose an element $\psi_j(g)$ of G_j such that $\rho(g, \psi_j(g))$ is minimized. Note that $\rho(g, \psi_j(g)) \leq m/2^j$, since otherwise g would have been added to G_j. Let $H_0 = G_0$, and for each $j \in \{1, ..., n\}$, define H_j to be $\{g - \psi_{j-1}(g) : g \in G_j\}$. Note that since for all $g \in G$, $\rho(g, \psi_{j-1}(g)) \leq m/2^{j-1}$, for each $h \in H_j$, $\sum_{i=1}^{2m} |h_i| \leq m/2^{j-1}$.

By induction, for each $k \in \{0, ..., n\}$ for each $g \in G_k$, there exist

$$h_{g,0} \in H_0, ..., h_{g,k} \in H_k$$

such that $g = \sum_{j=0}^{k} h_{g,j}$. Thus, for each $g \in G = G_n$, there exist $h_{g,0} \in H_0, ..., h_{g,n} \in H_n$ such that $g = \sum_{j=0}^{n} h_{g,j}$. Let

$$p = U\left\{\vec{\sigma} : \exists g \in G, \left|\frac{1}{m}\sum_{i=1}^{m} \sigma_i (g_i - g_{m+i})\right| > \eta\right\}.$$

Then, expressing g as $\sum_{j=0}^{n} h_{g,j}$, we get

$$p = U\left\{\vec{\sigma} : \exists g \in G, \left|\frac{1}{m}\sum_{i=1}^{m} \sigma_i (\sum_{j=0}^{n} (h_{g,j})_i - (h_{g,j})_{m+i}\right| > \eta\right\}.$$

Rearranging the sums yields

$$p = U\left\{\vec{\sigma} : \exists g \in G, \left|\sum_{j=0}^{n} \frac{1}{m} \sum_{i=1}^{m} \sigma_i((h_{g,j})_i - (h_{g,j})_{m+i})\right| > \eta\right\},$$

and applying the triangle inequality, we get

$$p \leq U\left\{\vec{\sigma} : \exists g \in G, \sum_{j=0}^{n} \left|\frac{1}{m} \sum_{i=1}^{m} \sigma_i((h_{g,j})_i - (h_{g,j})_{m+i})\right| > \eta\right\}.$$

For each $j \in \{0, ..., n\}$, let $\eta_j = (\eta/7)\sqrt{(j+1)/2^j}$. Then $\sum_{j=0}^{n} \eta_j \leq \eta$, and therefore

$$p \leq U\left\{\vec{\sigma} : \exists g \in G, \exists j \in \{0, ..., n\}, \left|\frac{1}{m} \sum_{i=1}^{m} \sigma_i((h_{g,j})_i - (h_{g,j})_{m+i})\right| > \eta_j\right\},$$

which implies

$$p \leq \sum_{j=0}^{n} U\left\{\vec{\sigma} : \exists g \in G, \left|\frac{1}{m} \sum_{i=1}^{m} \sigma_i((h_{g,j})_i - (h_{g,j})_{m+i})\right| > \eta_j\right\}.$$

Since each $h_{g,j} \in H_j$, we have

$$p \leq \sum_{j=0}^{n} U\left\{\vec{\sigma} : \exists h \in H_j, \ \left|\frac{1}{m} \sum_{i=1}^{m} \sigma_i(h_i - h_{m+i})\right| > \eta_j\right\}.$$

Choose $j \in \{0, ..., n\}$. For each $h \in H_j$, $\sum_{i=1}^{2m} |h_i| \leq m/2^{j-1}$. Thus, since $h \in \{-1, 0, 1\}^{2m}$,

$$\begin{aligned}
\sum_{i=1}^{m} (h_i - h_{m+i})^2 &= 4|\{i : |h_i - h_{m+i}| = 2\}| + |\{i : |h_i - h_{m+i}| = 1\}| \\
&\leq 2\sum_{i=1}^{2m} |h_i| \\
&\leq m/2^{j-2}.
\end{aligned}$$

Applying Lemma 8, we have

$$p \leq \sum_{j=0}^{n} 2|H_j| \exp\left(\frac{-(\eta_j m)^2}{2\frac{m}{2^{j-2}}}\right).$$

Substituting the value of η_j, we get

$$p \leq \sum_{j=0}^{n} 2|H_j| \exp\left(\frac{-\eta^2 (j+1) m}{400}\right).$$

By construction, each pair of elements of G_j have Hamming distance more than $m/2^j$. Applying Lemma 10, we get

$$|H_j| \leq |G_j| \leq (41 \cdot 2^j)^{\mathrm{VCdim}(G_j)} \leq (41 \cdot 2^j)^d$$

since $G_j \subseteq G$. Therefore

$$\begin{aligned} p &\leq 2\sum_{j=0}^{\infty} \exp\left((\ln 41 + j \ln 2)d - \frac{\eta^2(j+1)m}{400}\right) \\ &= \frac{2 \cdot 41^d e^{-\eta^2 m/400}}{1 - 2^d e^{-\eta^2 m/400}} \\ &\leq 4 \cdot 41^d e^{-\eta^2 m/400}, \end{aligned}$$

since $m \geq \frac{278(d+1)}{\eta^2}$. □

Putting together Lemmas 3, 4, 5, 6, and 9, and solving for m, we get a new proof of the following result due to Talagrand.

THEOREM 1 ([30]) *There is a constant c such that for any class $\mathcal{F}$ of functions from X to $\{0,1\}$, for any $\epsilon, \delta > 0$, there is an algorithm A that (ϵ, δ)-agnostically learns $\mathcal{F}$ from at most $\frac{c}{\epsilon^2}\left(\mathrm{VCdim}(\mathcal{F}) + \ln\frac{1}{\delta}\right)$ examples.*

3.2. Drifting environment

For a class $\mathcal{F}$ of functions from X to $\{0,1\}$, we say a learning algorithm A agnostically (ϵ, Δ)-*tracks* $\mathcal{F}$ if for all Δ-gradual sequences $P_1, P_2, ...$ of distributions over $X \times \{0,1\}$, there is an m_0 such that for all $m \geq m_0$, the probability that a sample drawn according to $\prod_{t=1}^m P_t$ and A's randomization cause a mistake for A is at most $\epsilon + \inf_{f \in \mathcal{F}} P_m\{(x,y) : f(x) \neq y\}$. If there is a prediction strategy that agnostically (ϵ, Δ)-tracks $\mathcal{F}$ then we say $\mathcal{F}$ is (ϵ, Δ)-*trackable in the agnostic case.*

For our analysis of agnostic learning in a drifting environment, we will replace Lemmas 5 and 6 with the following.

LEMMA 11 ([6]) *Choose a countable set Z, and a set $\mathcal{G}$ of functions from Z to $\{0,1\}$. Choose $\alpha > 0$ and $0 \leq \kappa < \alpha$. Choose $m \in \mathbf{N}$ such that $m \geq 4/\alpha^2$. Choose distributions $D, D_1, ..., D_m$ on Z such that for each $1 \leq i \leq m$, $d_{TV}(D_i, D) \leq \kappa$. If U is the uniform distribution over $\{1,-1\}^m$,*

$$\begin{aligned} &\left(\prod_{i=1}^m D_i\right)\left\{\vec{z} \in Z^m : \exists g \in \mathcal{G}, \left|\hat{\mathbf{E}}_{\vec{z}}(g) - \int_Z g(v)\, dD(v)\right| > \alpha\right\} \\ &\leq 2 \sup_{(\vec{z},\vec{u}) \in Z^m \times Z^m} U\left\{\vec{\sigma} : \exists g \in \mathcal{G}, \left|\frac{1}{m}\sum_{i=1}^m \sigma_i(g(u_i) - g(z_i))\right| > (\alpha - \kappa)/2\right\}. \end{aligned}$$

Putting together Lemmas 11 and 9, we get the following.

LEMMA 12 *Choose a countable set Z, and a set $\mathcal{G}$ of functions from Z to $\{0,1\}$. Let $d = \mathrm{VCdim}(\mathcal{G})$. Choose $\alpha > 0$ and $0 \le \kappa < \alpha$. Choose distributions $D, D_1, ..., D_m$ on Z such that for each $1 \le i \le m$, $d_{TV}(D_i, D) \le \kappa$. If $m \ge \frac{1112(d+1)}{(\alpha-\kappa)^2}$ then*

$$(\prod_{i=1}^{d} D_i)\left\{\vec{z} \in Z^m : \exists g \in \mathcal{G}, \left|\hat{\mathbf{E}}_{\vec{z}}(g) - \int_Z g(v)\, dD(v)\right| > \alpha\right\}$$
$$\le 8 \cdot 41^d e^{-(\alpha-\kappa)^2 m/1600}.$$

Next, we record a slight variant of a well-known lemma for converting tail bounds to expectation bounds.

LEMMA 13 *For any $[0,1]$-valued random variable Y, if $\varphi : [0,1] \to [0,1]$ is such that for all β, $\mathbf{Pr}(Y > \beta) \le \varphi(\beta)$, then for all $0 = a_0 \le a_1 \le ...a_k \le a_{k+1} = 1$, $\mathbf{E}(Y) \le \sum_{i=0}^{k} \varphi(a_i) a_{i+1}$.*

Proof: The distribution on Y that maximizes its expectation subject to $\forall i, \mathbf{Pr}(Y > a_i) \le \varphi(a_i)$ assigns $\varphi(a_k)$ probability on 1, $\varphi(a_{k-1}) - \varphi(a_k)$ probability on a_k, and so on, until all the probability has been distributed. This can be verified by induction moving from right to left, using a perturbation argument for the induction step. □

THEOREM 2 *There is a constant $c > 0$ such that for any set $\mathcal{F}$ of functions from X to $\{0,1\}$, for any $\epsilon > 0$, if $\Delta \le \frac{c\epsilon^3}{\mathrm{VCdim}(\mathcal{F})}$, then $\mathcal{F}$ is (ϵ, Δ)-trackable in the agnostic case.*

Proof: Choose $\epsilon \le 1$, and $\Delta \le \frac{\epsilon^3}{5000000d}$.

Let $m = \lfloor \epsilon/(16\Delta) \rfloor$. For each $f \in \mathcal{F}$, define $L_f : X \times \{0,1\} \to \{0,1\}$ by $L_f(x,y) = |f(x) - y|$. Consider the algorithm A which, given $(x_1, y_1), ..., (x_{t-1}, y_{t-1})$, returns a hypothesis $h \in \mathcal{F}$ that minimizes $\sum_{i=t-m}^{t-1} L_h(x_i, y_i)$. Let $L_{\mathcal{F}} = \{L_f : f \in \mathcal{F}\}$. Recall that $\mathrm{VCdim}(L_{\mathcal{F}}) \le \mathrm{VCdim}(\mathcal{F})$ (Lemma 4).

Choose a Δ-gradual sequence $P_1, P_2, ...$ of probability distributions, an arbitrary $f_* \in \mathcal{F}$ (to compare h with), and $t > m$. Applying Lemma 1 as in [13], the probability that $(x_1, y_1), ..., (x_t, y_t)$ drawn according to $\prod_{i=1}^{t} P_i$ causes a mistake for A is equal to the expectation, with respect to the first $t-1$ examples, of $P_t\{(x_t, y_t) : h(x_t) \ne y_t\}$ (recall that h is a function of the first $t-1$ examples).

Choose $\beta \ge 6\Delta m$. Since for all $i \le m$, $d_{TV}(P_{t-i}, P_t) \le \Delta m$, applying Lemma 12 with $\alpha = \beta/2$, $Z = X \times \{0,1\}$, and $\mathcal{G} = L_{\mathcal{F}}$, and doing some simple calculations, we get

$$\mathbf{Pr}\left(\exists f \in \mathcal{F}, \left|P_t\{(x_t, y_t) : f(x_t) \ne y_t\} - \frac{1}{m}\sum_{i=t-m}^{t-1} L_f(x_i, y_i)\right| > \beta/2\right)$$
$$\le 8 \cdot 41^d \exp\left(\frac{-\beta^2 m}{14400}\right).$$

Since $\sum_{i=t-m}^{t-1} L_h(x_i, y_i) \le \sum_{i=t-m}^{t-1} L_{f_*}(x_i, y_i)$, for all $\beta > 6\Delta m$,

$$\mathbf{Pr}(P_t\{(x_t, y_t) : h(x_t) \ne y_t\} - P_t\{(x_t, y_t) : f_*(x_t) \ne y_t\} > \beta)$$

$$\leq 8 \cdot 41^d \exp\left(\frac{-\beta^2 m}{14400}\right).$$

Applying Lemma 13 with φ given by the the above bound when $\beta \geq 6\Delta m$ and 1 otherwise, and with $a_1 = 6\Delta m$, and for all relevant $i > 1$, $a_i = \sqrt{\frac{14400(\ln 8 + (\ln 41)d + i \ln 2)}{m}}$, we get

$$\begin{aligned}
&\mathbf{E}(P_t\{(x_t, y_t) : h(x_t) \neq y_t\} - P_t\{(x_t, y_t) : f_*(x_t) \neq y_t\}) \\
&\leq 6\Delta m + \sum_{i=1}^{\infty} \sqrt{\frac{14400(\ln 8 + (\ln 41)d + (i+1)\ln 2)}{m}} 2^{-i} \\
&\leq 6\Delta m + \sqrt{\frac{d}{m}} \sum_{i=1}^{\infty} \sqrt{14400(6 + (i+1)\ln 2)} 2^{-i} \\
&\leq 6\Delta m + 341\sqrt{\frac{d}{m}}.
\end{aligned}$$

Substituting the values of m and Δ and approximating, we get

$$\mathbf{E}(P_t\{(x_t, y_t) : h(x_t) \neq y_t\} - P_t\{(x_t, y_t) : f_*(x_t) \neq y_t\}) \leq \epsilon.$$

As discussed above, this completes the proof. □

4. The realizable case

Say a probability distribution P over $X \times \{0,1\}$ is *consistent* with a function f from X to $\{0,1\}$ if the probability that a pair (x, y) drawn according to P has $f(x) = y$ is 1. For a set $\mathcal{F}$ of functions from X to $\{0,1\}$, say that P is consistent with $\mathcal{F}$ if it is consistent with some member of $\mathcal{F}$. For a class $\mathcal{F}$ of functions from X to $\{0,1\}$, we say a learning algorithm A *(ϵ, Δ)-tracks $\mathcal{F}$ in the realizable case* if for all Δ-gradual sequences $P_1, P_2, ...$ of distributions over $X \times \{0,1\}$ that are consistent with $\mathcal{F}$, there is an m_0 such that for all $m \geq m_0$, the probability that $(x_1, y_1), ..., (x_m, y_m)$ drawn according to $\prod_{t=1}^m P_t$ and A's randomization cause a mistake for A is at most ϵ. If there is a prediction strategy that (ϵ, Δ)-tracks $\mathcal{F}$ in the realizable case then we say $\mathcal{F}$ is *(ϵ, Δ)-trackable in the realizable case.*

Recall that the mth hypercube, which we will denote by H_m, is the undirected graph whose vertex set is $\{0,1\}^m$, and whose edges are all $\vec{v}, \vec{w}$ such that $\rho(\vec{v}, \vec{w}) = 1$.

THEOREM 3 ([13]) *For any $m \in \mathbf{N}$, for any $F \subseteq \{0,1\}^m$, if G is the subgraph of H_m induced by F, the edges of G can be directed so that the maximum outdegree of any node is at most* $\mathrm{VCdim}(F)$.

LEMMA 14 ([28, 27, 8]) *For $m \in \mathbf{N}$, $F \subseteq \{0,1\}^m$, $|F| \leq (em/\mathrm{VCdim}(F))^{\mathrm{VCdim}(F)}$.*

The proof of our next lemma is similar to that of a related result of Roy [25].

LEMMA 15 *For any $m \in \mathbf{N}$, for any $F \subseteq \{0,1\}^m$, for any $k \in \{1, ..., m\}$,*

$$\text{VCdim}(\rho_{k-1}(F) \cup \rho_k(F)) \leq 5(\text{VCdim}(F) + k).$$

Proof: Assume without loss of generality that $|F| > 1$. Let $d = \text{VCdim}(\rho_{k-1}(F) \cup \rho_k(F))$. Choose a set $i_1, ..., i_d$ such that

$$\{(g_{i_1}, ..., g_{i_d}) : g \in \rho_{k-1}(F) \cup \rho_k(F)\} = \{0,1\}^d.$$

Each element of $\{(g_{i_1}, ..., g_{i_d}) : g \in \rho_{k-1}(F)\}$ can be derived from an element of $\{(f_{i_1}, ..., f_{i_d}) : f \in F\}$ and a subset of $k-1$ elements of $\{1, ..., d\}$, and therefore

$$|\{(g_{i_1}, ..., g_{i_d}) : g \in \rho_{k-1}(F)\}| \leq \binom{d}{k-1} |\{(f_{i_1}, ..., f_{i_d}) : f \in F\}|.$$

Applying a similar observation with regard to $\rho_k(F)$, we get

$$\begin{aligned} & |\{(g_{i_1}, ..., g_{i_d}) : g \in \rho_{k-1}(F) \cup \rho_k(F)\}| \\ & \leq \left(\binom{d}{k-1} + \binom{d}{k}\right) |\{(f_{i_1}, ..., f_{i_d}) : f \in F\}| \\ & = \binom{d+1}{k} |\{(f_{i_1}, ..., f_{i_d}) : f \in F\}| \\ & \leq \binom{d+1}{k} \left(\frac{ed}{\text{VCdim}(F)}\right)^{\text{VCdim}(F)} \end{aligned}$$

by Lemma 14. Thus

$$\begin{aligned} 2^d & \leq \binom{d+1}{k} \left(\frac{ed}{\text{VCdim}(F)}\right)^{\text{VCdim}(F)} \\ & \leq \left(\frac{e(d+1)}{k}\right)^k \left(\frac{ed}{\text{VCdim}(F)}\right)^{\text{VCdim}(F)}. \end{aligned}$$

Taking logs, we get

$$d \ln 2 \leq k \ln\left(\frac{e(d+1)}{k}\right) + \text{VCdim}(F) \ln\left(\frac{ed}{\text{VCdim}(F)}\right).$$

Since for all $x, \lambda > 0$, $1 + \ln x \leq \lambda x + \ln(1/\lambda)$ (see [1]), we have that for all $\lambda > 0$,

$$d \ln 2 \leq \lambda(2d+1) + (\text{VCdim}(F) + k) \ln(1/\lambda).$$

Solving for d and substituting $\lambda = 1/10$ completes the proof. □

LEMMA 16 *Choose $m \in \mathbf{N}$ and $F \subseteq \{0,1\}^m$. Then the edges of H_m can be oriented so that the outdegree of any $\vec{v} \in \{0,1\}^m$ is at most* $15(\text{VCdim}(F) + \rho(\vec{v}, F))$.

Proof: Let $d = \text{VCdim}(F)$. Assume without loss of generality that $|F| > 1$ (and therefore $d > 0$).

Let G_0 be the subgraph of H_m induced by F, and for each $k = 1, ..., m$, let G_k be the subgraph of H_m induced by $\rho_k(F) \cup \rho_{k-1}(F)$. (See Figure 2.) For each k, let

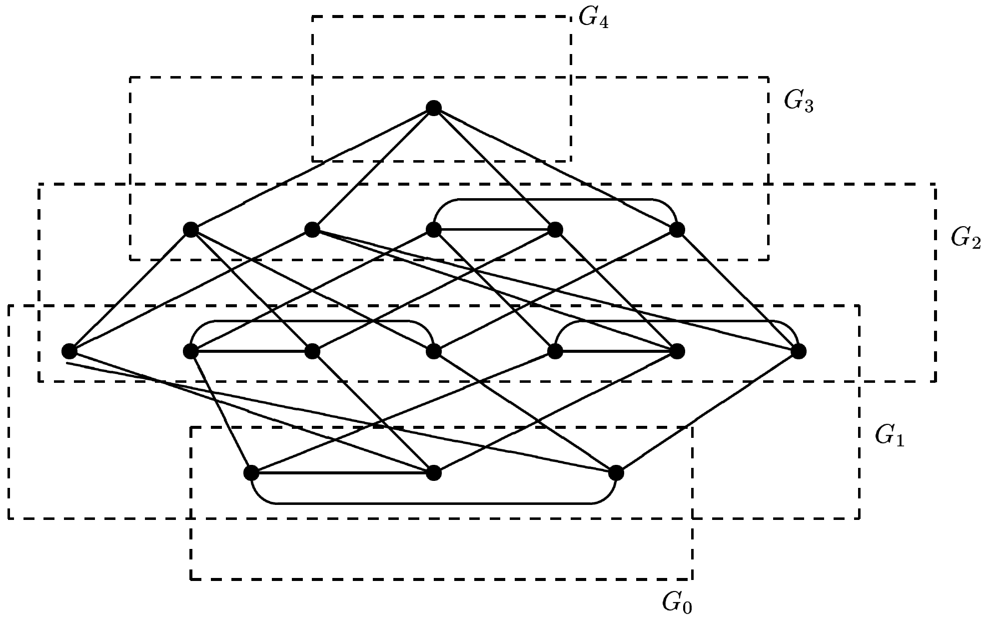

Figure 2. For $m = 4$ and some $F \subseteq \{0,1\}^m$, the m-dimensional hypercube has been diagrammed with F at the bottom, those vertices at a Hamming distance 1 from some element of F in the row above, and so on. The subgraphs $G_0, \ldots, G_4$ from the proof of Lemma 16 are as shown.

G'_k be a directed graph obtained by directing the edges of G_k so that the outdegree of each vertex in G'_k is at most $5(d+k)$.

By the triangle inequality, if $\vec{v}, \vec{w}$ is an edge in H_m, then $|\rho(\vec{v}, F) - \rho(\vec{w}, F)| \leq 1$. Therefore, each edge of H_m is in G_k for at least one k. Form a directed graph H'_m by directing the edges of H_m by choosing the direction for each edge from the graph G'_k with the least k such that the undirected edge is in G_k.

Choose a vertex $\vec{v} \in \{0,1\}^m$. Assume without loss of generality that $\rho(\vec{v}, F) < m$. Then $\vec{v}$ appears in G'_k exactly when $k \in \{\rho(\vec{v}, F), \rho(\vec{v}, F)+1\}$. Hence the outdegree of $\vec{v}$ in H'_m is at most

$$5(d + \rho(\vec{v}, F)) + 5(d + \rho(\vec{v}, F) + 1) \leq 15(d + \rho(\vec{v}, F)),$$

completing the proof. □

For each set $\mathcal{F}$ of possible targets, the tracking algorithm $A'_{\mathcal{F}}$ used to prove Theorem 4 will apply a subalgorithm $A_{\mathcal{F}}$ to a subsequence consisting of the most recent examples. We begin by describing and analyzing $A_{\mathcal{F}}$.

Algorithm $A_{\mathcal{F}}$ will make use of an arbitrary order on X. For each $\mathcal{F}$, we will describe the hypothesis h output by $A_{\mathcal{F}}$ on input $(x_1, y_1), ..., (x_{m-1}, y_{m-1})$ by describing a process for generating $h(x_m)$ for each possible x_m. Algorithm $A_{\mathcal{F}}$ first sorts $x_1, ..., x_m$ (let $a_1, ..., a_m$ be the resulting reordering of $x_1, ..., x_m$; let $b_1, ..., b_m$ be the corresponding reordering of $y_1, ..., y_{m-1}, \square$, where $\square$ serves to hold the position corresponding to x_m; and let i^* be the position of x_m in $a_1, ..., a_m$). Next, it sets $F = \{(f(a_1), ..., f(a_m)) : f \in \mathcal{F}\}$, and creates a directed graph H'_m by orienting the edges of H_m so that the outdegree of each vertex $\vec{v}$ is at most $15(\text{VCdim}(F) + \rho(\vec{v}, F))$ as in Lemma 16. Finally, it sets $h(x_m) = 1$ if and only if the edge in H'_m between $(b_1, ..., b_{i^*-1}, 0, b_{i^*+1}, ..., b_m)$ and $(b_1, ..., b_{i^*-1}, 1, b_{i^*+1}, ..., b_m)$ is oriented toward $(b_1, ..., b_{i^*-1}, 1, b_{i^*+1}, ..., b_m)$.

Lemma 17 ([3]) *For any probability distributions P and Q, $d_{TV}(P \times Q, Q \times P) \leq d_{TV}(P, Q)$.*

Lemma 18 *Choose $m \in \mathbf{N}$, a set $\mathcal{F}$ of functions from X to $\{0,1\}$, and a Δ-gradual sequence $P_1, ..., P_m$ of probability distributions on $X \times \{0,1\}$ that are consistent with $\mathcal{F}$. The probability under $\prod_{t=1}^m P_t$ that $(x_1, y_1), ..., (x_m, y_m)$ causes a mistake for $A_{\mathcal{F}}$, is at most*

$$\frac{15\text{VCdim}(F)}{m} + 6\Delta m + \mathbf{Pr}(\exists i, j, x_i = x_j).$$

Proof: Define $\chi((x_1, y_1), ..., (x_m, y_m))$ to indicate whether $(x_1, y_1), ..., (x_m, y_m)$ causes a mistake for $A_{\mathcal{F}}$ and $x_1, ..., x_m$ are distinct. Clearly,

$$\mathbf{Pr}(\text{mistake}) \leq \mathbf{E}(\chi) + \mathbf{Pr}(\text{not distinct}),$$

so we will bound $\mathbf{E}(\chi)$.

Let $Z = X \times \{0,1\}$. For $\vec{z} \in Z^m, j \in \{1, ..., m\}$, define $\varphi(\vec{z}, j)$ to be the result of exchanging z_j and z_m. By the triangle inequality, for all $t \in \{1, ..., m\}$,

$d_{TV}(P_j, P_m) \leq \Delta m$. Choose $j \in \{1, ..., m-1\}$. Repeatedly applying Fubini's Theorem (Lemma 1),

$$\int \chi(\vec{z})\ d(\prod_{t=1}^{m} P_t)(\vec{z})$$
$$= \int \left(\int \chi(\vec{z})\ d(P_j \times P_m)(z_j, z_m) \right) d(\prod_{t \notin \{j,m\}} P_t)(z_1, ..., z_{j-1}, z_{j+1}, ..., z_{m-1}).$$

Applying Lemma 17 and the definition of d_{TV},

$$\int \chi(\vec{z})\ d(\prod_{t=1}^{m} P_t)(\vec{z})$$
$$\leq \int \left(\int \chi(\vec{z})\ d(P_m \times P_j)(z_j, z_m) + \frac{\Delta m}{2} \right) d(\prod_{t \notin \{j,m\}} P_t)(z_1, ..., z_{j-1}, z_{j+1}, ..., z_{m-1})$$
$$= \int \chi(\varphi(\vec{z}, j))\ d(\prod_{t=1}^{m} P_t)(\vec{z}) + \frac{\Delta m}{2},$$

again, because of Fubini's Theorem. Thus

$$\int \chi(\vec{z})\ d(\prod_{t=1}^{m} P_t)(\vec{z}) \leq \frac{\Delta m}{2} + \int \left(\frac{1}{m} \sum_{j=1}^{m} \chi(\varphi(\vec{z}, j)) \right)\ d(\prod_{t=1}^{m} P_t)(\vec{z}). \qquad (1)$$

Fix an arbitrary $\vec{z} = ((x_1, y_1), ..., (x_m, y_m)) \in (X \times \{0,1\})^m$. If $x_1, ..., x_m$ are not distinct, then the definition of χ implies that $\frac{1}{m} \sum_{j=1}^{m} \chi(\varphi(\vec{z}, j)) = 0$. Assume $x_1, ..., x_m$ are distinct. Let $a_1, ..., a_m$ be $x_1, ..., x_m$ in sorted order, and let $v_1, ..., v_m$ be the corresponding reordering of the y_i's. Let

$$F = \{(f(a_1), ..., f(a_m)) : f \in \mathcal{F}\}.$$

Since algorithm $A_{\mathcal{F}}$ sorts the sample, the directed graph H'_m constructed by algorithm $A_{\mathcal{F}}$ using any reordering of the x_i's is the same. Choose $j \in \{1, ..., m\}$. Let j' be the position of x_j when $x_1, ..., x_m$ is sorted. Then $\varphi(\vec{z}, j)$ causes a mistake for $A_{\mathcal{F}}$ if and only if the edge in H'_m between $\vec{v}$ and the vertex obtained by negating the j'th bit of $\vec{v}$ is oriented away from $\vec{v}$. (This is because $\vec{v}$ represents the correct labellings, and $A_{\mathcal{F}}$ predicts according to the direction of the named edge.) Thus $\sum_{j=1}^{m} \chi(\varphi(\vec{z}, j)) \leq \text{outdegree}(\vec{v})$.

For each $t \in \{1, ..., m\}$ choose $f_t \in \mathcal{F}$ such that P_t is consistent with f_t. Then

$$\rho(\vec{v}, F) \leq |\{t : f_t(x_t) \neq f_m(x_t)\}|.$$

Thus,

$$\text{outdegree}(\vec{v}) \leq 15(\text{VCdim}(F) + |\{t : f_t(x_t) \neq f_m(x_t)\}|)$$

and therefore

$$\sum_{j=1}^{m} \chi(\varphi(\vec{z}, j)) \leq 15(\mathrm{VCdim}(F) + |\{t : f_t(x_t) \neq f_m(x_t)\}|).$$

Since $\mathrm{VCdim}(F) \leq \mathrm{VCdim}(\mathcal{F})$, plugging into (1), we have

$$\int \chi(\vec{z}) \; d(\prod_{t=1}^{m} P_t)(\vec{z}) \leq \frac{\Delta m}{2} + \frac{15\mathrm{VCdim}(\mathcal{F})}{m} + \frac{15}{m}\mathbf{E}(|\{t : f_t(x_t) \neq f_m(x_t)\}|). \tag{2}$$

Since P_m is consistent with f_m,

$$P_m\{(x, y) : f_m(x) \neq y\} = 0. \tag{3}$$

For any $t \in \{1, ..., m\}$, since $d_{TV}(P_t, P_m) \leq \Delta m$, (3) implies

$$P_t\{(x, y) : f_t(x) \neq f_m(x)\} = P_t\{(x, y) : f_m(x) \neq y\} \leq \Delta m/2.$$

Thus

$$\mathbf{E}(|\{t : f_t(x_t) \neq f_m(x_t)\}|) \leq \Delta m^2/2.$$

Substituting into (2) completes the proof. □

THEOREM 4 *There is a constant $c > 0$ such that for any set $\mathcal{F}$ of functions from X to $\{0, 1\}$, for any $\epsilon > 0$, if*

$$\Delta \leq \frac{c\epsilon^2}{\mathrm{VCdim}(\mathcal{F})},$$

then $\mathcal{F}$ is (ϵ, Δ)-trackable in the realizable case.

Proof: Let $d = \mathrm{VCdim}(\mathcal{F})$. Consider the algorithm $A'_{\mathcal{F}}$ defined as follows. First, it sets $R = \{1, ..., \lceil 11560d^2/\epsilon^3 \rceil\}$, and for each t, it draws r_t uniformly at random from R.

Given $(x_1, y_1), ..., (x_m, y_m)$, if[4] $m > 33d/\epsilon$, then $A'_{\mathcal{F}}$ gives the last $m' = \lceil 33d/\epsilon \rceil$ elements of $((x_1, r_1), y_1), ..., ((x_m, r_m), y_m)$ to $A_{\mathcal{F}}$.

Let U be the uniform distribution over R. For some $\Delta \geq 0$, choose a Δ-gradual sequence $P_1, P_2, ...$ of distributions over X. Then $P_1 \times U, P_2 \times U, ...$ is also Δ-gradual. Also, if for each $f \in \mathcal{F}$, we define a function f_R from $X \times R$ to $\{0, 1\}$ by $f_R(x, r) = f(x)$, then, straight from the definitions, $\mathrm{VCdim}(\{f_R : f \in \mathcal{F}\}) = d$. So applying Lemma 18, if $m > 33d/\epsilon$, the probability that $A'_{\mathcal{F}}$ makes a mistake is at most $15d/m' + 6\Delta m' + (m')^2/|R|$. Substituting the definitions of m' and R and observing that $33d/\epsilon \leq m' \leq 34d/\epsilon$, if $\Delta \leq \frac{\epsilon^2}{458d}$, this probability is at most ϵ, completing the proof. □

Acknowledgments

We thank the COLT'98 program committee and anonymous referees for valuable comments, including pointing out some mistakes in earlier drafts of the paper. We gratefully acknowledge the support of National University of Singapore Academic Research Fund Grant RP960625.

Notes

1. Recently, other constraints on the drift have been examined (e.g., [4, 9]). In this paper we restrict our attention to the simplest drift models, but direct application of a slight variant of Lemma 12 of this paper leads to a small improvement in the analysis of [9]. Models of a changing environment that are more dissimilar to that studied here were considered in [22, 20, 21, 7, 10, 15, 2, 19, 31, 16].
2. Their statement of their algorithm is slightly different; we describe an equivalent algorithm to facilitate comparison with our modification.
3. We assume that Z is countable for convenience. Considerably weaker measurability assumptions suffice for the results mentioned in this paper [23, 11].
4. The behavior of $A'_{\mathcal{F}}$ for small m is immaterial.

References

1. M. Anthony, N. Biggs, and J. Shawe-Taylor. The learnability of formal concepts. *Proceedings of the 1990 Workshop on Computational Learning Theory*, pages 246–257, 1990.
2. P. Auer and M. K. Warmuth. Tracking the best disjunction. *Proceedings of the 36th Annual Symposium on the Foundations of Computer Science*, 1995.
3. P. L. Bartlett. Learning with a slowly changing distribution. *Proceedings of the 1992 Workshop on Computational Learning Theory*, pages 243–252, 1992.
4. P. L. Bartlett, S. Ben-David, and S. R. Kulkarni. Learning changing concepts by exploiting the structure of change. *Proceedings of the 1996 Conference on Computational Learning Theory*, pages 131–139, 1996.
5. P. L. Bartlett and D. P. Helmbold, 1995. Manuscript.
6. R. D. Barve and P. M. Long. On the complexity of learning from drifting distributions. *Information and Computation*, 138(2):101–123, 1997.
7. A. Blum and P. Chalasani. Learning switching concepts. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages pages 231–242, 1992.
8. A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *JACM*, 36(4):929–965, 1989.
9. Y. Freund and Y. Mansour. Learning under persistent drift. *Proceedings of the 1997 European Conference on Computational Learning Theory*, 1997.
10. Y. Freund and D. Ron. Learning to model sequences generated by switching distributions. *Proceedings of the 1995 Conference on Computational Learning Theory*, pages 41–50, 1995.
11. D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.
12. D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.
13. D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):129–161, 1994.
14. D. P. Helmbold and P. M. Long. Tracking drifting concepts by minimizing disagreements. *Machine Learning*, 14(1):27–46, 1994.
15. M. Herbster and M. K. Warmuth. Tracking the best expert. *Proceedings of of the Twelvth International Conference on Machine Learning*, 1995.

16. M. Herbster and M. K. Warmuth. Tracking the best regressor. *Proceedings of the 1998 Conference on Computational Learning Theory*, 1998.
17. M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.
18. A. N. Kolmogorov and V. M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *American Mathematical Society Translations (Ser. 2)*, 17:277–364, 1961.
19. A. Kuh. Comparison of tracking algorithms for single layer threshold networks in the presence of random drift. *IEEE Trans. on Signal Processing*, 45(3):640–650, 1997.
20. A. Kuh, T. Petsche, and R. Rivest. Learning time varying concepts. In *NIPS 3*. Morgan Kaufmann, 1990.
21. A. Kuh, T. Petsche, and R. Rivest. Mistake bounds of incremental learners when concepts drift with applications to feedforward networks. In *NIPS 4*. Morgan Kaufmann, 1991.
22. N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.
23. D. Pollard. *Convergence of Stochastic Processes*. Springer Verlag, 1984.
24. D. Pollard. *Empirical Processes : Theory and Applications*, volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Math. Stat. and Am. Stat. Assoc., 1990.
25. S. Roy. Semantic complexity of relational queries and data independent data partitioning. *Proceedings of the ACM SIGACT-SIGART-SIGMOD Annual Symposium on Principles of Database Systems*, 1991.
26. H. L. Royden. *Real Analysis*. Macmillan, 1963.
27. N. Sauer. On the density of families of sets. *J. Combinatorial Theory (A)*, 13:145–147, 1972.
28. S. Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific J. Math.*, 41:247–261, 1972.
29. H. U. Simon. General lower bounds on the number of examples needed for learning probabilistic concepts. *Journal of Computer and System Sciences*, 52(2):239–254, 1996.
30. M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, 22:28–76, 1994.
31. X. Tian and A. Kuh. Performance bounds for single layer threshold networks when tracking a drifting adversary. *Neural Networks*, 10(5):897–906, 1997.
32. V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.