On the Complexity of Learning from Drifting Distributions

Rakesh D. Barve
Department of Computer Science
Duke University, P.O. Box 90129
Durham, North Carolina 27708 USA

Philip M. Long
ISCS Department
National University of Singapore
Singapore 119260, Republic of Singapore

Proposed running head: Drifting distributions.

Address of corresponding author: Philip M. Long, ISCS Department, National University of Singapore, Singapore 119260, Republic of Singapore.

Abstract

We consider two models of on-line learning of binary-valued functions from drifting distributions due to Bartlett. We show that if each example is drawn from a joint distribution which changes in total variation distance by at most $O(\epsilon^3/(d\log(1/\epsilon)))$ between trials, then an algorithm can achieve a probability of a mistake at most ϵ worse than the best function in a class of VC-dimension d. We prove a corresponding necessary condition of $O(\epsilon^3/d)$. Finally, in the case that a fixed function is to be learned from noise-free examples, we show that if the distributions on the domain generating the examples change by at most $O(\epsilon^2/(d\log(1/\epsilon)))$, then any consistent algorithm learns to within accuracy ϵ.

For a list of the typographical symbols used, please consult *Latex*, by Leslie Lamport.

1 Introduction

In prediction models [Daw84, HLW94] like that studied in this paper, learning proceeds in *trials*, where in the tth trial, the algorithm (1) is given x_t chosen from some set X, (2) is required to output a prediction $\hat{y}_t \in \{0,1\}$, and (3) discovers $y_t \in \{0,1\}$. The focus of these models differs from those like the PAC model [Val84] in that prediction models are tailored to situations where learning is an ongoing process. However, results for one type of model often yield related results for the other.

We will consider two models for prediction problems, both introduced by Bartlett [Bar92]. In the first model, the pairs $(x_1, y_1), (x_2, y_2), ...$ are generated independently from a sequence $P_1, P_2, ...$ of distributions on $X \times \{0,1\}$, and it is assumed that pairs of consecutive distributions are close. If, for a class F of functions from X to $\{0,1\}$, there is a learning algorithm that, for large enough t, whenever the distribution changes by at most γ in total variation distance between trials, the algorithm achieves a probability of mistake that is at most ϵ more than that of the best function (w.r.t. P_t), then we will say that F is (ϵ, γ)-agnostically learnable. For a more formal definition, see Section 2.

Note that this model allows drift both in the marginal distribution on x_t and in the dependence between x_t and y_t. It therefore is a flexible but clean way to capture gradual variation in a learner's environment.

The second model we consider is the analogue of Ben-David, Benedek and Mansour's *solid learnability* [BBM89] with drifting distributions. In this model, there is a fixed function $f \in F$ that maps each x_t to y_t, but the distribution on the domain changes by at most γ between trials. We say that F is (ϵ, γ)-solidly learnable if any algorithm in this setting which at each trial returns a hypothesis consistent with previous trials achieves a probability of mistake of at most ϵ for large enough t.

In this paper, we show that if

$$\gamma = O\left(\frac{\epsilon^3}{\mathrm{VCdim}(F)\ln(1/\epsilon)}\right)$$

then F is (ϵ, γ)-agnostically learnable. This improves on the $O(\epsilon^5/(\mathrm{VCdim}(F)^2 \ln(1/\epsilon)))$ bound that follows from the work of Bartlett [Bar92], as pointed out by Bartlett and Helmbold in [BH95].

We also show that if F is (ϵ, γ)-agnostically learnable, then $\gamma = O(\epsilon^3/\mathrm{VCdim}(F))$, matching our sufficient condition for each F to within a log factor.

Finally, we show that if $\gamma = O(\epsilon^2/(\mathrm{VCdim}(F)\ln(1/\epsilon)))$, then F is (ϵ, γ)-solidly learnable, improving on the $O(\epsilon^3/(\mathrm{VCdim}(F)^2 \ln(1/\epsilon)))$ bound of Bartlett [Bar92].

Bartlett and Helmbold [BH95] have described an algorithm for learning in a drifting environment. Their results imply that when a function f maps each x_t to y_t, then if $\gamma = O(\epsilon^2/(\mathrm{VCdim}(F) + \ln(1/\epsilon)))$, their algorithm achieves a probability of mistake of at most ϵ for large enough t. Their results in general treat the case in which the function mapping x_t to y_t is slowly changing as well. A relative strength of our solid learnability result is that their algorithm requires time that is in general exponential in $\mathrm{VCdim}(F)$ whereas in many concrete cases, efficient algorithms for finding consistent hypotheses are known. For all classes F, both our result and theirs match Bartlett's [Bar92] $\gamma = O(\epsilon^2/\mathrm{VCdim}(F))$ necessary condition up to log factors.

Littlestone and Warmuth [LW94], Kuh, Petsche and Rivest [KPR90, KPR91], Blum and Chalisani

[BC92], Herbster and Warmuth [HW95], and Auer and Warmuth [AW95] also studied learning in a changing environment, but in frameworks substantially different from that considered here.

The main new idea in our proof of the sufficient conditions is in where the assumption that the distributions are close to each other is applied. The proofs use Blumer, Ehrenfeucht, Haussler and Warmuth's idea of learning by estimating the error of all the hypotheses in the class from a single sample [BEHW89]. To bound how hard this is in our setting, we follow the outline of the proof of Vapnik and Chervonenkis [VC71]. In proofs of this type, the probability that the error of some hypothesis is badly estimated is bounded by the probability that two samples give rise to substantially different estimates. Pollard [Pol84] calls this the *symmetrization step*. The "two-sample" probability is then bounded, making use of the resulting symmetry.

Bartlett's [Bar92] analysis proceeded by showing that the product distribution on a suitably small sequence of the most recent examples was close to the product distribution where these examples were drawn from the "current" distribution. Helmbold and Long [HL91], who studied learning drifting concepts from a fixed distribution, applied the fact that the functions were slowly changing in the two-sample step. In their journal version [HL94], they presented an analysis using the fact that, again for a small sequence of the most recent examples, what the learner saw was likely to be close to what it would have seen had the target not been moving. Helmbold and Long [HL94] and Bartlett and Helmbold [BH95] studied modifications of the one-inclusion graph strategy [HLW94].

In our analysis, we apply the fact that the distributions are moving slowly in the symmetrization step. As one can see by examining the proof, this is necessary to ensure that the resulting two-sample problem is indeed symmetric, with both samples being drawn according to the same sequence of drifting distributions. Once this is the case, the resulting two-sample product distribution is invariant with respect to the usual pair-swapping permutations, and the remainder of the VC proof can go through almost without modification.

For our proof of the necessary condition, we make use of techniques due to Simon [Sim93], Ehrenfeucht, Haussler, Kearns and Valiant [EHKV89] and Helmbold and Long [HL94]. The main new idea required to prove this paper's result was how to drift efficiently from a joint distribution with no information to a hard joint distribution of the type useful in arguments of the type of Simon. Our approach was to accomplish this by drifting the conditional distributions of the $\{0,1\}$ labels from $1/2$ to a small amount on either side of $1/2$. Our proof required us to prove a new lower bound on the fatness of a tail of the binomial distribution. For this, we built upon a technique of Littlestone [Lit90], lower bounding the sum of the largest few terms instead of the largest term as he did. A similar bound has been proved by Simon [Sim93], who appealed to the central limit theorem. Our bound has the advantage that it yields specific constants.

In independent work, Bartlett, Ben-David and Kulkarni [BBK96] studied learning a drifting concept with a variety of constraints on the drift, including models which allowed large but infrequent changes in the target concept. In addition to prediction models like those studied here, they also studied estimating the entire trace of target concept positions.

Recently, Freund and Mansour [FM97] studied a model of learning in which, instead of assuming that the *position* of the state of the environment is approximately constant (i.e. that drift is slow), they assume that the *rate and direction* of change is approximately constant (i.e. that drift is "persistent").

2 Preliminaries

2.1 Definitions

Denote the positive integers by $\mathbf{N}$.

For a countable set Z and probability distributions D_1, D_2 over Z, the total variation distance between D_1 and D_2, denoted here by $d_{TV}(D_1, D_2)$, is defined to be the largest difference in the probabilities that D_1 and D_2 assign to any event, i.e. by

$$d_{TV}(D_1, D_2) = \sup_{U \subseteq Z} |D_1(U) - D_2(U)|.$$

Following Bartlett [Bar92], we say a sequence $D_1, D_2, ...$ of probability distributions on Z is γ-*admissible* if for all $t \in \mathbf{N}$, $d_{TV}(D_t, D_{t+1}) \leq \gamma$.

For the remainder of this subsection, fix a countable[1] set X.

A *prediction strategy* takes a finite sequence of elements of $X \times \{0,1\}$, and outputs a rule for using the first element of a pair to predict the second. Formally, it is a mapping from $\cup_m (X \times \{0,1\})^m$ to the set of functions from X to $\{0,1\}$.

Choose a class F of functions from X to $\{0,1\}$. We say that F is *agnostically* (ϵ, γ)-*learnable* if there exists a prediction strategy A and $t_0 \in \mathbf{N}$ such that

- for all γ-admissible sequences $P_1, P_2, ...$ of probability distributions on $X \times \{0,1\}$,
- for all $t \in \mathbf{N}, t \geq t_0$

the following holds:

$$\left(\prod_{i=1}^{t} P_i\right) \{\langle (x_i, y_i)\rangle_{i=1}^{t} : (A(\langle (x_i, y_i)\rangle_{i=1}^{t-1}))(x_t) \neq y_t\} \leq \epsilon + \inf_{f \in F} P_t\{(u, v) : f(u) \neq v\},$$

where $\prod_{i=1}^{t} P_i$ denotes the distribution over $(X \times \{0,1\})^t$ obtained by sampling independently from $P_1, ..., P_t$ respectively. This model was studied by Bartlett [Bar92]. The sample complexity of agnostic learning from fixed distributions was first studied by Haussler [Hau92].

A prediction strategy A is *consistent with* F if for any $(x_1, y_1), ..., (x_m, y_m)$, if there exists $f \in F$ such that $y_1 = f(x_1), ..., y_m = f(x_m)$, then $A((x_1, y_1), ..., (x_m, y_m))$ is such an f.

We say that F is (ϵ, γ)-*solidly learnable* if for any prediction strategy A that is consistent with F, there is a $t_0 \in \mathbf{N}$ such that

- for all $f \in F$
- for all γ-admissible sequences $D_1, D_2, ...$ of probability distributions on X,
- for all $t \in \mathbf{N}, t \geq t_0$

[1] We assume that X is countable for convenience, but significantly weaker assumptions, like those of Pollard's [Pol84] Appendix C, are enough.

the following holds:

$$\left(\prod_{i=1}^{t} D_i\right)\{\langle x_i\rangle_{i=1}^{t} : (A(\langle (x_i, f(x_i))\rangle_{i=1}^{t-1}))(x_t) \neq f(x_t)\} \leq \epsilon.$$

This is a natural extension of the definition discussed by Ben-David, Benedek and Mansour [BBM89], which itself extended the PAC model [Val84]. Results for this model follow from the work of Bartlett [Bar92].

A subset $\{x_1, ..., x_d\}$ of some set X is *shattered* by a set F of functions from X to $\{0,1\}$ if

$$\{(f(x_1), ..., f(x_d)) : f \in F\} = \{0,1\}^d.$$

The VC-dimension [VC71] of F is the size of the largest set shattered by F. For examples of the VC-dimension, see [BEHW89, Nat91, AB92].

2.2 Tools

The following is a special case of Fubini's Theorem.

Lemma 1 (see [Roy63]) *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*

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

We also record the standard Hoeffding bound for reference.

Lemma 2 (see [Pol84]) *Choose $a < b$ and a countable set Z. Let D be a probability distribution on Z, and let $f_1, ..., f_m$ be independent random variables taking values in $[a,b]$. Then the probability under D^m of a sequence $(z_1, ..., z_m)$ for which*

$$\left|\left(\frac{1}{m}\sum_{i=1}^{m} f_i(z_i)\right) - \left(\frac{1}{m}\sum_{i=1}^{m}\int_Z f_i(z)D(z)\right)\right| > \epsilon$$

is no more than $2e^{-2\epsilon^2 m/(b-a)^2}$.

3 Sufficient Conditions

The following are the main results of this section.

Theorem 3 *For any set F of at least two functions from X to $\{0,1\}$, for any $\epsilon \leq 1/100$, if $d = \mathrm{VCdim}(F)$, then if*

$$\gamma \leq \frac{\epsilon^3}{100000 d \ln \frac{1}{\epsilon}},$$

then F is agnostically (ϵ, γ)-learnable.

Theorem 4 *For any set F of at least two functions from X to $\{0,1\}$, for any $\epsilon \leq 1/100$, if $d = \mathrm{VCdim}(F)$, then if*

$$\gamma \leq \frac{\epsilon^2}{800 d \ln \frac{1}{\epsilon}},$$

then F is solidly (ϵ, γ)-learnable.

As discussed in the introduction, obtaining uniformly good estimates of the expectations of a family of random variables (in our application, the errors of possible hypothesis) from a single sample plays a key role here. We treat this subject first.

3.1 Uniformly Good Estimates of Expectations

While one can obtain a single unifying bound that yields both Theorem 3 and Theorem 4, the proof is somewhat messy, and we therefore elect to split our analysis into two cases.

Both proofs will make use of the following lemma.

Lemma 5 *Choose a countable set Z, $g : Z \to \{0,1\}$. Choose $\kappa, \beta > 0$. Choose probability distributions $D, D_1, ..., D_m$ on Z such that for each $i \leq m$, $d_{TV}(D, D_i) \leq \kappa$. Then if $m \geq 1/\beta^2$,*

$$\left(\prod_{i=1}^{m} D_i\right)\left\{\vec{z} : \left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \int_Z g(v)\, dD(v)\right| > \beta + \kappa\right\} \leq 1/2.$$

Proof: We have

$$\begin{aligned}
&\left(\prod_{i=1}^{m} D_i\right)\left\{\vec{z} : \left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \int_Z g(v)\, dD(v)\right| > \beta + \kappa\right\} \\
&= \left(\prod_{i=1}^{m} D_i\right)\left\{\vec{z} : \left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \int_Z g(v)\, dD(v) \right.\right. \\
&\qquad \left.\left. + \left(\frac{1}{m}\sum_{i=1}^{m}\int_Z g(v) dD_i(v)\right) - \left(\frac{1}{m}\sum_{i=1}^{m}\int_Z g(v) dD_i(v)\right)\right| > \beta + \kappa\right\} \\
&\leq \left(\prod_{i=1}^{m} D_i\right)\left\{\vec{z} : \left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \left(\frac{1}{m}\sum_{i=1}^{m}\int_Z g(v)\, dD_i(v)\right)\right| \right. \\
&\qquad \left. + \left|\left(\frac{1}{m}\sum_{i=1}^{m}\int_Z g(v)\, dD_i(v)\right) - \int_Z g(v)\, dD(v)\right| > \beta + \kappa\right\} \\
&\leq \left(\prod_{i=1}^{m} D_i\right)\left\{\vec{z} : \left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \left(\frac{1}{m}\sum_{i=1}^{m}\int_Z g(v)\, dD_i(v)\right)\right| \right. \\
&\qquad \left. + \frac{1}{m}\sum_{i=1}^{m}\left|\int_Z g(v)\, dD_i(v) - \int_Z g(v)\, dD(v)\right| > \beta + \kappa\right\} \\
&\leq \left(\prod_{i=1}^{m} D_i\right)\left\{\vec{z} : \left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \left(\frac{1}{m}\sum_{i=1}^{m}\int_Z g(v)\, dD_i(v)\right)\right| \right. \\
&\qquad \left. + \left|\frac{1}{m}\sum_{i=1}^{m}\kappa\right| > \beta + \kappa\right\} \qquad (\text{since } \forall i,\ d_{TV}(D, D_i) \leq \kappa)
\end{aligned}$$

$$= \left(\prod_{i=1}^{m} D_i\right)\left\{\vec{z} : \left|\frac{1}{m}\sum_{i=1}^{m} g(z_i) - \left(\frac{1}{m}\sum_{i=1}^{m}\int_Z g(v)\ dD_i(v)\right)\right| > \beta\right\}$$
$$\leq 2e^{-2\beta^2 m} \qquad \text{(by Lemma 2)}$$
$$\leq 1/2,$$

since $m \geq 1/\beta^2$. □

We begin with a bound that will be applied in the agnostic learning case.

Theorem 6 *Choose a countable set Z, and a set G of functions from Z to $\{0,1\}$. Let $d = \mathrm{VCdim}(G)$. Choose $m \in \mathbf{N}$, and a distribution D on Z, and let $D_1, ..., D_m$ be distributions on Z such that for each $1 \leq i \leq m$, $d_{TV}(D_i, D) \leq \kappa$. Then for all $\alpha > 2\kappa$, if $m \geq 4/\alpha^2$,*

$$\left(\prod_{i=1}^{m} D_i\right)\left\{\vec{z} : \exists g \in G,\ \left|\left(\int_Z g(u)\ dD(u)\right) - \frac{1}{m}\sum_{i=1}^{m} g(z_i)\right| > \alpha\right\} \leq 4\left(\frac{2em}{d}\right)^d \exp(-((\alpha - 2\kappa)^2 m/8)).$$

The right hand side of the above inequality is at most δ if

$$m \geq \frac{16}{(\alpha - 2\kappa)^2}\left(2d\ln\left(\frac{10}{\alpha - 2\kappa}\right) + \ln\frac{4}{\delta}\right).$$

We begin with the symmetrization step.

Lemma 7 *Choose a countable set Z, and a set G of functions from Z to $\{0,1\}$. Choose $\alpha > 0$ and $0 < \kappa < \alpha/2$. 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 $1 \leq i \leq m$, $d_{TV}(D_i, D) \leq \kappa$. Suppose*

- *Q is the set of all $\vec{z} \in Z^m$ for which there is a $g \in G$ such that*

$$\left|\int_Z g(v)\ dD(v) - \frac{1}{m}\sum_{i=1}^{m} g(z_i)\right| > \alpha,$$

  and

- *J is the set of all $(\vec{z}, \vec{u}) \in Z^m \times Z^m$ for which there is a $g \in G$ such that*

$$\left|\frac{1}{m}\sum_{i=1}^{m} g(u_i) - \frac{1}{m}\sum_{i=1}^{m} g(z_i)\right| > \alpha/2 - \kappa.$$

Then

$$\left(\prod_{i=1}^{m} D_i\right)(Q) \leq 2\left(\prod_{i=1}^{m} D_i \times \prod_{i=1}^{m} D_i\right)(J).$$

Proof: Choose $\vec{z} \in Q$ and choose a $g \in G$ such that

$$\left|\int_Z g(v)\ dD(v) - \frac{1}{m}\sum_{i=1}^{m} g(z_i)\right| > \alpha. \tag{1}$$

Applying Lemma 5,

$$\left(\prod_{i=1}^{m} D_i\right)\left\{\vec{u} : \left|\frac{1}{m}\sum_{i=1}^{m} g(u_i) - \int_Z g(v)\ dD(v)\right| > \alpha/2 + \kappa\right\} \leq 1/2. \tag{2}$$

By the triangle inequality, for any $\vec{u} \in Z^m$,

$$\begin{aligned}&\left|\int_Z g(v)\ dD(v) - \left(\frac{1}{m}\sum_{i=1}^{m} g(z_i)\right)\right| \\ &\leq \left|\left(\frac{1}{m}\sum_{i=1}^{m} g(u_i)\right) - \left(\frac{1}{m}\sum_{i=1}^{m} g(z_i)\right)\right| + \left|\int_Z g(v)\ dD(v) - \left(\frac{1}{m}\sum_{i=1}^{m} g(u_i)\right)\right|\end{aligned}$$

so

$$\begin{aligned}&\left|\int_Z g(v)\ dD(v) - \left(\frac{1}{m}\sum_{i=1}^{m} g(u_i)\right)\right| \\ &\geq \left|\int_Z g(v)\ dD(v) - \left(\frac{1}{m}\sum_{i=1}^{m} g(z_i)\right)\right| - \left|\left(\frac{1}{m}\sum_{i=1}^{m} g(u_i)\right) - \left(\frac{1}{m}\sum_{i=1}^{m} g(z_i)\right)\right|.\end{aligned}$$

By (1),

$$\left|\int_Z g(v)\ dD(v) - \left(\frac{1}{m}\sum_{i=1}^{m} g(u_i)\right)\right| > \alpha - \left|\left(\frac{1}{m}\sum_{i=1}^{m} g(u_i)\right) - \left(\frac{1}{m}\sum_{i=1}^{m} g(z_i)\right)\right|$$

so

$$\left|\left(\frac{1}{m}\sum_{i=1}^{m} g(u_i)\right) - \left(\frac{1}{m}\sum_{i=1}^{m} g(z_i)\right)\right| \leq \alpha/2 - \kappa \ \Rightarrow\ \left|\int_Z g(v)\ dD(v) - \left(\frac{1}{m}\sum_{i=1}^{m} g(u_i)\right)\right| > \alpha/2 + \kappa.$$

Applying (2),

$$\left(\prod_{i=1}^{m} D_i\right)\left\{\vec{u} \in Z^m : \left|\frac{1}{m}\sum_{i=1}^{m} g(u_i) - \frac{1}{m}\sum_{i=1}^{m} g(z_i)\right| \leq \alpha/2 - \kappa\right\} \leq 1/2.$$

Since $\vec{z}$ was chosen arbitrarily from Q, for any $\vec{z} \in Q$,

$$\left(\prod_{i=1}^{m} D_i\right)\{\vec{u} \in Z^m : (\vec{z}, \vec{u}) \in J\} \geq 1/2. \tag{3}$$

Now, by Lemma 1,

$$\begin{aligned}\left(\prod_{i=1}^{m} D_i \times \prod_{i=1}^{m} D_i\right)(J) &= \int_{Z^m}\left(\left(\prod_{i=1}^{m} D_i\right)\{\vec{u} \in Z^m : (\vec{z}, \vec{u}) \in J\}\right)\ d(\prod_{i=1}^{m} D_i)(\vec{z}) \\ &\geq \int_Q \left(\left(\prod_{i=1}^{m} D_i\right)\{\vec{u} \in Z^m : (\vec{z}, \vec{u}) \in J\}\right)\ d(\prod_{i=1}^{m} D_i)(\vec{z}) \\ &\geq \int_Q 1/2\ d(\prod_{i=1}^{m} D_i)(\vec{z}) \qquad \text{(by (3))} \\ &= \int_Q d(\prod_{i=1}^{m} D_i)(\vec{z})/2. \\ &= \left(\prod_{i=1}^{m} D_i\right)(Q)/2.\end{aligned}$$

Solving completes the proof. □

We will make use of the following version of Sauer's lemma [Sau72], due to Blumer, Ehrenfeucht, Haussler and Warmuth [BEHW89].

Lemma 8 ([Sau72, BEHW89]) *Choose a finite set Z, and a set F of functions from Z to $\{0,1\}$. Let $d = \mathrm{VCdim}(F)$. Then*

$$|F| \leq (e|Z|/d)^d.$$

Now we are ready for the second part of the proof. The proof of this part follows that of Vapnik and Chervonenkis [VC71].

Lemma 9 *Choose a countable set Z, and a set G of functions from Z to $\{0,1\}$. Let $d = \mathrm{VCdim}(G)$. Choose $\eta > 0$. Choose $m \in \mathbf{N}$. Choose distributions $D_1, ..., D_m$ on Z. Let J' be the set of all $(\vec{z}, \vec{u}) \in Z^m \times Z^m$ for which there is a $g \in G$ for which*

$$\left| \frac{1}{m} \sum_{i=1}^{m} g(u_i) - \frac{1}{m} \sum_{i=1}^{m} g(z_i) \right| > \eta$$

Then

$$\left(\prod_{i=1}^{m} D_i \times \prod_{i=1}^{m} D_i \right) (J') \leq 2 \left(\frac{2em}{d} \right)^d e^{-\eta^2 m/2}.$$

Proof: Define T (for "two sample") by

$$T = \prod_{i=1}^{m} D_i \times \prod_{i=1}^{m} D_i.$$

Note that, for each $i \leq m$, u_i and z_i are both drawn independently from D_i. Therefore, by symmetry, for any fixed $\vec{\sigma} \in \{-1,1\}^m$,

$$\begin{aligned}
&T \left\{ (\vec{z}, \vec{u}) : \exists g \in G, \ \left| \frac{1}{m} \sum_{i=1}^{m} g(u_i) - \frac{1}{m} \sum_{i=1}^{m} g(z_i) \right| > \eta \right\} \\
&= T \left\{ (\vec{z}, \vec{u}) : \exists g \in G, \ \left| \frac{1}{m} \sum_{i=1}^{m} g(u_i) - g(z_i) \right| > \eta \right\} \\
&= T \left\{ (\vec{z}, \vec{u}) : \exists g \in G, \ \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i (g(u_i) - g(z_i)) \right| > \eta \right\}.
\end{aligned}$$

Thus, if U is the uniform distribution on $\{-1,1\}^m$,

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

Viewing the RHS above as the expectation of the corresponding characteristic function, applying Lemma 1 to express it as a nested integral, and overestimating, we get

$$T\left\{(\vec{z},\vec{u}) : \exists g \in G,\ \left|\frac{1}{m}\sum_{i=1}^{m} g(u_i) - \frac{1}{m}\sum_{i=1}^{m} g(z_i)\right| > \eta\right\}$$
$$\leq \sup_{(\vec{z},\vec{u})\in Z^m\times Z^m} U\left\{\vec{\sigma} : \exists g \in G,\ \left|\frac{1}{m}\sum_{i=1}^{m} \sigma_i(g(u_i) - g(z_i))\right| > \eta\right\}.$$

Fix $(\vec{z},\vec{u}) \in Z^m \times Z^m$. For each $g \in G$, Lemma 2 implies that

$$U\left\{\vec{\sigma} : \left|\frac{1}{m}\sum_{i=1}^{m} \sigma_i(g(u_i) - g(z_i))\right| > \eta\right\} \leq 2e^{-\eta^2 m/2}.$$

Therefore

$$U\left\{\vec{\sigma} : \exists g \in G,\ \left|\frac{1}{m}\sum_{i=1}^{m} \sigma_i(g(u_i) - g(z_i))\right| > \eta\right\}$$
$$\leq 2|\left\{(g(z_1),...,g(z_m),g(u_1),...,g(u_m)) : g \in G\right\}|e^{-\eta^2 m/2},$$

which, applying Lemma 8, implies

$$U\left\{\vec{\sigma} : \exists g \in G,\ \left|\frac{1}{m}\sum_{i=1}^{m} \sigma_i(g(u_i) - g(z_i))\right| > \eta\right\} \leq 2\left(\frac{2em}{d}\right)^d e^{-\eta^2 m/2},$$

completing the proof. □

To prove the second inequality of Theorem 6, we will need the following lemma, written down in this form by Bartlett and Long [BL95], which is implicit in the work of Anthony, Biggs and Shawe-Taylor [ABS90].

Lemma 10 *For any $y_1, y_2, y_4, \delta > 0$ and $y_3 \geq 1$, if*

$$m \geq \frac{2}{y_4}\left(y_2 \ln\left(\frac{2y_2y_3}{y_4}\right) + \ln\frac{y_1}{\delta}\right),$$

then

$$y_1 \exp(y_2 \ln(y_3 m) - y_4 m) \leq \delta.$$

Proof (of Theorem 6): Putting together Lemmas 7 and 9 proves the first inequality. To prove the second, applying Lemma 10 with $y_1 = 4$, $y_2 = d$, $y_3 = 2e/d$, and $y_4 = (\alpha - 2\kappa)^2/8$, we get that

$$m \geq \frac{16}{(\alpha-2\kappa)^2}\left(d\ln\left(\frac{32e}{(\alpha-2\kappa)^2}\right) + \ln\frac{4}{\delta}\right)$$

is sufficient for the right hand side to be at most δ. Simplifying and overapproximating completes the proof. □

Now we turn to a bound that will be applied for solid learning. Its proof is similar to that of Theorem 6.

Theorem 11 *Choose a countable set Z, and a set G of functions from Z to $\{0,1\}$. Let $d = \mathrm{VCdim}(G)$. Choose $m \in \mathbf{N}$, and a distribution D on Z, and let $D_1, ..., D_m$ be distributions on Z such that for each $i, 1 \leq i \leq m$, $d_{TV}(D_i, D) \leq \kappa$. Then for all $\alpha > 2\kappa$, if $m \geq 4/\alpha^2$,*

$$\left(\prod_i D_i\right)\left\{\vec{z} : \exists g \in G,\ \int_Z g(u)\ dD(u) > \alpha \text{ and } \forall 1 \leq i \leq m,\ g(z_i) = 0\right\} \leq 2\left(\frac{2em}{d}\right)^d 2^{-(\alpha/2-\kappa)m}.$$

The right hand side of the above inequality is at most δ if

$$m \geq \frac{6}{\alpha - 2\kappa}\left(d \ln\left(\frac{12e}{\alpha - 2\kappa}\right) + \ln\frac{2}{\delta}\right).$$

Again, we begin with the symmetrization step.

Lemma 12 *Choose a countable set Z, and a set G of functions from Z to $\{0,1\}$. Choose $\alpha > 0$ and $0 < \kappa < \alpha/2$. Choose $m \in \mathbf{N}$ such that $m \geq 4/\alpha^2$. Choose distributions $D_1, ..., D_m$ on Z such that for each $1 \leq i \leq m$, $d_{TV}(D_i, D) \leq \kappa$. Suppose*

- *Q is the set of all $\vec{z} \in Z^m$ for which there is a $g \in G$ such that $\int_Z g(v)\ dD(v) > \alpha$ and for all $i \leq m$, $g(z_i) = 0$, and*
- *J is the set of all $(\vec{z}, \vec{u}) \in Z^m \times Z^m$ for which there is a $g \in G$ such that*

$$\frac{1}{m}\sum_{i=1}^m g(u_i) > \alpha/2 - \kappa$$

and for all $i \leq m$, $g(z_i) = 0$.

Then

$$\left(\prod_{i=1}^m D_i\right)(Q) \leq 2\left(\prod_{i=1}^m D_i \times \prod_{i=1}^m D_i\right)(J).$$

Proof: Choose $\vec{z} \in Q$ and choose a $g \in G$ such that

$$\int_Z g(v)\ dD(v) > \alpha \text{ and } \forall i \leq m,\ g(z_i) = 0. \tag{4}$$

Applying Lemma 5, we have

$$\left(\prod_{i=1}^m D_i\right)\left\{\vec{u} : \left|\frac{1}{m}\sum_{i=1}^m g(u_i) - \int_Z g(v)\ dD(v)\right| > \alpha/2 + \kappa\right\} \leq 1/2.$$

Applying the triangle inequality and (4) in a manner analogous to the proof of Lemma 7, for any $\vec{z} \in Q$,

$$\left(\prod_{i=1}^m D_i\right)\left\{\vec{u} \in Z^m : \frac{1}{m}\sum_{i=1}^m g(u_i) \leq \alpha/2 - \kappa\right\} \leq 1/2,$$

so for any $\vec{z} \in Q$,

$$\left(\prod_{i=1}^m D_i\right)\{\vec{u} \in Z^m : (\vec{z}, \vec{u}) \in J\} \geq 1/2.$$

From here the proof is as in Lemma 7. □

Now we are ready for the second part of the proof. This closely follows the corresponding part of the proof of the main result of [BEHW89].

Lemma 13 *Choose a countable set Z, and a set G of functions from Z to $\{0,1\}$. Let $d = \mathrm{VCdim}(G)$. Choose $\eta > 0$. Choose $m \in \mathbf{N}$. Choose distributions $D_1, ..., D_m$ on Z. Let J' be the set of all $(\vec{z}, \vec{u}) \in Z^m \times Z^m$ for which there is a $g \in G$ such that*

$$\frac{1}{m}\sum_{i=1}^{m} g(u_i) > \eta$$

and for all $i \leq m$, $g(z_i) = 0$. Then

$$\left(\prod_{i=1}^{m} D_i \times \prod_{i=1}^{m} D_i\right)(J') \leq \left(\frac{2em}{d}\right)^d 2^{-\eta m}$$

Proof: Define T by

$$T = \prod_{i=1}^{m} D_i \times \prod_{i=1}^{m} D_i.$$

By Lemma 1 and by symmetry, for any fixed $\vec{\sigma} \in \{0,1\}^m$,

$$T\left\{(\vec{z},\vec{u}) : \exists g \in G,\ \frac{1}{m}\sum_{i=1}^{m} g(u_i) > \eta \text{ and } \forall i \leq m,\ g(z_i) = 0\right\}$$
$$= T\left\{(\vec{z},\vec{u}) : \exists g \in G,\ \frac{1}{m}\sum_{i=1}^{m} g(\sigma_i u_i + (1-\sigma_i)z_i) > \eta \text{ and } \forall i \leq m,\ g(\sigma_i z_i + (1-\sigma_i)u_i) = 0\right\}.$$

Informally, the role of σ_i in the above is to decide whether to "swap" z_i and u_i. Since the z_i's and u_i's are mutually independent and z_i and u_i are identically distributed, if we fix σ, we obtain the same distibution on the "post-swap" sequence pairs as on the original sequence pairs.

Since the above holds for any σ, if U is the uniform distribution on $\{0,1\}^m$, choosing σ according to U, we get

$$T(J') = (T \times U)\{((\vec{z},\vec{u}),\sigma) : \exists g \in G,\ \frac{1}{m}\sum_{i=1}^{m} g(\sigma_i u_i + (1-\sigma_i)z_i) > \eta$$
$$\text{and } \forall i \leq m,\ g(\sigma_i z_i + (1-\sigma_i)u_i) = 0\}.$$

Applying Lemma 1 and overestimating, we get

$$T(J') \leq \sup_{(\vec{z},\vec{u}) \in Z^m \times Z^m} U\{\vec{\sigma} : \exists g \in G,\ \frac{1}{m}\sum_{i=1}^{m} g(\sigma_i u_i + (1-\sigma_i)z_i) > \eta$$
$$\text{and } \forall i \leq m,\ g(\sigma_i z_i + (1-\sigma_i)u_i) = 0\}.$$

Fix $\vec{z}, \vec{u} \in Z^m$. For each $g \in G$, the above event can only occur if

- for no index i is $g(z_i) = g(u_i) = 1$,
- for at least ηm indices i, either $g(z_i) = 1$ or $g(u_i) = 1$, and
- for each of those, $\sigma_i = 1 \Leftrightarrow g(u_i) = 1$.

Thus, for each $g \in G$,

$$U\{\vec{\sigma} : \frac{1}{m}\sum_{i=1}^{m} g(\sigma_i u_i + (1-\sigma_i)z_i) > \eta \text{ and } \forall i \le m,\ g(\sigma_i z_i + (1-\sigma_i)u_i) = 0\} \le 2^{-m\eta}.$$

Therefore

$$\begin{aligned} &U\{\vec{\sigma} : \exists g \in G,\ \frac{1}{m}\sum_{i=1}^{m} g(\sigma_i u_i + (1-\sigma_i)z_i) > \eta \text{ and } \forall i \le m,\ g(\sigma_i z_i + (1-\sigma_i)u_i) = 0\} \\ &\le |\{(g(z_1), ..., g(z_m), g(u_1), ..., g(u_m)) : g \in G\}| 2^{-\eta m}, \end{aligned}$$

which, applying Lemma 8, implies

$$\begin{aligned} U\left\{\vec{\sigma} : \exists g \in G,\ \frac{1}{m}\sum_{i=1}^{m} g(\sigma_i u_i + (1-\sigma_i)z_i) > \eta \text{ and } \forall i \le m,\ g(\sigma_i z_i + (1-\sigma_i)u_i) = 0\right\} \\ \le \left(\frac{2em}{d}\right)^d 2^{-\eta m}, \end{aligned}$$

completing the proof. □

Proof (of Theorem 11): Here, putting together Lemmas 12 and 13 proves the first inequality of Theorem 11. To obtain the second inequality, rewrite the RHS of the first as

$$2\exp\left(d\ln\frac{2em}{d} - (\ln 2)(\alpha/2-\kappa)m\right).$$

Applying Lemma 10 with $y_1 = 2$, $y_2 = d$, $y_3 = 2e/d$, and $y_4 = (\ln 2)(\alpha/2-\kappa)$, we get that

$$m \ge \frac{2}{(\ln 2)(\alpha/2-\kappa)}\left(d\ln\left(\frac{4e}{(\ln 2)(\alpha/2-\kappa)}\right) + \ln\frac{2}{\delta}\right)$$

is sufficient for the RHS to be at most δ. Applying the fact $\ln 2 \ge 2/3$ completes the proof. □

Now we are ready for the proofs of the sufficient conditions of learning in a drifting environment. These proofs borrow ideas from the work of Blumer, Ehrenfeucht, Haussler and Warmuth [BEHW89] and Haussler [Hau92].

Proof of Theorem 3: Since $|F| \ge 2$, $\mathrm{VCdim}(F) \ge 1$.

Let $m = \lfloor \epsilon/(16\gamma) \rfloor$. Consider the algorithm A which, at each trial $t > m$ returns a hypothesis $h \in F$ that minimizes disagreements with the last m examples. That is, $A((x_1, y_1), ..., (x_{t-1}, y_{t-1}))$ is an $h \in F$ minimizing $|\{t-m \le i < t : h(x_i) \ne y_i\}|$. For each $f \in F$, define $\ell_f : X \times \{0,1\} \to \{0,1\}$ by $\ell_f(x,y) = |f(x) - y|$. Then algorithm A returns a hypothesis on trial t minimizing $\sum_{i=t-m}^{t-1} \ell_h(x_i, y_i)$. Let $\ell_F = \{\ell_f : f \in F\}$. It is known (see [Hau92]) that $\mathrm{VCdim}(\ell_F) \le \mathrm{VCdim}(F)$.

Choose a γ-admissible sequence $P_1, P_2, ...$ of probability distributions, $f \in F$ and a trial $t > m$. Denote the (random) hypothesis output by A on the tth trial by h. Applying Lemma 1 as in [HLW94],

$$\left(\prod_{i=t-m}^{t} P_i\right)\{\langle (x_i, y_i)\rangle : h(x_t) \ne y_t\} = \int_{(X\times\{0,1\})^m} P_t\{(x_t, y_t) : h(x_t) \ne y_t\}\, d(\prod_{i=t-m}^{t-1} P_i).$$

Since

$$\sum_{i=t-m}^{t-1} \ell_h(x_i, y_i) \leq \sum_{i=t-m}^{t-1} \ell_f(x_i, y_i),$$

if

$$P_t\{(x_t, y_t) : h(x_t) \neq y_t\} > P_t\{(x_t, y_t) : f(x_t) \neq y_t\} + \epsilon/2,$$

then either

$$\left| P_t\{(x_t, y_t) : h(x_t) \neq y_t\} - \frac{1}{m} \sum_{i=t-m}^{t-1} \ell_h(x_i, y_i) \right| > \epsilon/4$$

or

$$\left| P_t\{(x_t, y_t) : f(x_t) \neq y_t\} - \frac{1}{m} \sum_{i=t-m}^{t-1} \ell_f(x_i, y_i) \right| > \epsilon/4$$

or both. Since for all $i \leq m$, $d_{TV}(P_{t-i}, P_t) \leq \gamma m = \gamma \lfloor \epsilon/(16\gamma) \rfloor \leq \epsilon/16$, applying Theorem 6 with $Z = X \times \{0, 1\}$ and $G = \ell_F$, the probability that either of these happens is at most $\epsilon/2$ if

$$m \geq \frac{16}{(\epsilon/4 - \epsilon/8)^2} \left(2d \ln \frac{10}{\epsilon/4 - \epsilon/8} + \ln \frac{8}{\epsilon} \right)$$

or equivalently

$$m \geq \frac{1024}{\epsilon^2} \left(2d \ln \frac{80}{\epsilon} + \ln \frac{8}{\epsilon} \right). \tag{5}$$

(Note that both f and h are in F, so if all r.v.'s in ℓ_F are estimated accurately, both ℓ_f and ℓ_h are.) Thus, if (5), the probability that

$$P_t\{(x_t, y_t) : h(x_t) \neq y_t\} > P_t\{(x_t, y_t) : f(x_t) \neq y_t\} + \epsilon/2$$

is at most $\epsilon/2$, and therefore the expectation of

$$P_t\{(x_t, y_t) : h(x_t) \neq y_t\}$$

is at most

$$P_t\{(x_t, y_t) : f(x_t) \neq y_t\} + \epsilon.$$

We have

$$\begin{aligned}
\gamma &\leq \frac{\epsilon^3}{100000 d \ln \frac{1}{\epsilon}} \\
\frac{17\gamma}{\epsilon} &\leq \frac{\epsilon^2}{5883 d \ln \frac{1}{\epsilon}} \\
\frac{\epsilon}{17\gamma} &\geq \frac{5883 d \ln \frac{1}{\epsilon}}{\epsilon^2} \\
\left\lfloor \frac{\epsilon}{16\gamma} \right\rfloor &\geq \frac{5883 d \ln \frac{1}{\epsilon}}{\epsilon^2}
\end{aligned}$$

since $\frac{\epsilon}{16\gamma} > 16$. Substituting, we get $m \geq \frac{5883 d \ln \frac{1}{\epsilon}}{\epsilon^2}$ which immediately implies

$$m \geq \frac{1024d}{\epsilon^2} \left(3 \ln \frac{1}{\epsilon} + \frac{(2 \ln 80 + \ln 8) \ln \frac{1}{\epsilon}}{\ln 100} \right).$$

Since $\epsilon \leq 1/100$, we have

$$\begin{aligned} m &\geq \frac{1024d}{\epsilon^2}\left(3\ln\frac{1}{\epsilon} + 2\ln 80 + \ln 8\right) \\ &= \frac{1024}{\epsilon^2}\left(3d\ln\frac{1}{\epsilon} + (2\ln 80 + \ln 8)d\right) \\ &\geq \frac{1024}{\epsilon^2}\left((2d+1)\ln\frac{1}{\epsilon} + 2d\ln 80 + \ln 8\right) \qquad \text{(since } d \geq 1\text{)} \\ &= \frac{1024}{\epsilon^2}\left(2d\ln\frac{80}{\epsilon} + \ln\frac{8}{\epsilon}\right). \end{aligned}$$

Since this satisfies the requirement of (5), this completes the proof. □

Proof (of Theorem 4): Again, since $|F| \geq 2$, $\mathrm{VCdim}(F) \geq 1$. Set $m = \lfloor \epsilon/(8\gamma) \rfloor$. Choose a consistent algorithm A. Choose $f^* \in F$ (interpreted as a target function), and a γ-admissible sequence $D_1, D_2, \ldots$ of distributions over X.

For each $f \in F$, define $\ell_f : X \to \{0,1\}$ by $\ell_f(x) = |f(x) - f^*(x)|$. (Note that this is different from the definition used in the proof of Theorem 3.) Then algorithm A returns a hypothesis h on trial t for which $\sum_{i=1}^{t-1} \ell_h(x_i) = 0$. Let $\ell_F = \{\ell_f : f \in F\}$. Clearly, $\mathrm{VCdim}(\ell_F) \leq \mathrm{VCdim}(F)$ (see [BEHW89]).

Fix a trial $t > m$. Denote the (random) hypothesis output by A on the tth trial by h. Since

- $\sum_{i=t-m}^{t-1} \ell_h(x_i) = 0$, and
- for all $i \leq m$, $d_{TV}(D_{t-i}, D_t) \leq \gamma m = \gamma\lfloor \epsilon/(8\gamma) \rfloor \leq \epsilon/8$,

applying Theorem 11 with $Z = X$ and $G = \ell_F$, the probability that

$$D_t\{x_t : h(x_t) \neq f^*(x_t)\} > \epsilon/2,$$

is at most $\epsilon/2$ if

$$m \geq \frac{6}{\epsilon/2 - \epsilon/4}\left(d\ln\left(\frac{12e}{\epsilon/2 - \epsilon/4}\right) + \ln\frac{4}{\epsilon}\right)$$

or equivalently

$$m \geq \frac{24}{\epsilon}\left(d\ln\frac{48e}{\epsilon} + \ln\frac{4}{\epsilon}\right). \tag{6}$$

Thus, if (6), the probability that $D_t\{x_t : h(x_t) \neq f^*(x_t)\} > \epsilon/2$ is at most $\epsilon/2$, and therefore the expectation of $D_t\{x_t : h(x_t) \neq f^*(x_t)\}$ is at most ϵ.

In this case, we have

$$\begin{aligned} \gamma &\leq \frac{\epsilon^2}{800d\ln\frac{1}{\epsilon}} \\ \frac{9\gamma}{\epsilon} &\leq \frac{\epsilon}{85d\ln\frac{1}{\epsilon}} \\ \frac{\epsilon}{9\gamma} &\geq \frac{85d\ln\frac{1}{\epsilon}}{\epsilon} \\ \left\lfloor \frac{\epsilon}{8\gamma} \right\rfloor &\geq \frac{85d\ln\frac{1}{\epsilon}}{\epsilon} \end{aligned}$$

since $\frac{\epsilon}{8\gamma} > 8$. Substituting, we get $m \geq \frac{85d \ln \frac{1}{\epsilon}}{\epsilon}$ which immediately implies

$$m \geq \frac{24d}{\epsilon}\left(2\ln\frac{1}{\epsilon} + \frac{(1+\ln 48+\ln 8)\ln\frac{1}{\epsilon}}{\ln 100}\right).$$

Since $\epsilon \leq 1/100$, we have

$$\begin{aligned} m &\geq \frac{24d}{\epsilon}\left(2\ln\frac{1}{\epsilon} + 1 + \ln 48 + \ln 8\right) \\ &= \frac{24}{\epsilon}\left(2d\ln\frac{1}{\epsilon} + (1+\ln 48+\ln 8)d\right) \\ &\geq \frac{24}{\epsilon}\left((d+1)\ln\frac{1}{\epsilon} + d(1+\ln 48) + \ln 8\right) \qquad \text{(since } d \geq 1\text{)} \\ &= \frac{24}{\epsilon}\left(d\ln\frac{48e}{\epsilon} + \ln\frac{8}{\epsilon}\right). \end{aligned}$$

Since this satisfies the requirement of (6), this completes the proof. □

4 A Necessary Condition

In this section we show that, for small enough ϵ, $\gamma < \frac{1100000\epsilon^3}{\mathrm{VCdim}(F)}$ is a necessary condition for F to be agnostically (ϵ,γ)-learnable. For each class F, if d is the VC-dimension of F, we show that F is not agnostically (ϵ,γ)-learnable if $\gamma = \frac{1100000\epsilon^3}{d}$.

Our proof uses ideas of Ehrenfeucht, Haussler, Kearns and Valiant [EHKV89], Helmbold and Long [HL94], and Simon [Sim93]. At a high level, it proceeds as follows. For some trial t, we will choose a sequence of drifting distributions that drifts only during the last m trials. Each of the distributions will assign equal probability to each of the d shattered points; the conditional probability that the label is 1 is what will change. Before drift, this will be $1/2$ for each of the shattered points. We will then randomly choose a direction for each conditional probability to drift. They will drift just far enough that the learning algorithm must be able to determine which direction most of them drifted to obtain an accurate enough hypothesis. Examples before drift starts yield no information about the drift direction, and if the drift rate is large enough a difficult "test" distribution can be reached with few "useful" examples. We observe that the best way guess the drift direction is to look at the fraction of times each shattered point was accompanied by a label 1, and to guess that the drift was up if and only if this is fraction is at least $1/2$. However, since even during the useful examples the probability of being labelled 1 is close to $1/2$, such information is highly unreliable. Finally, we will argue that since the learner fails on average for a random sequence of directions, there must be some particular sequence of directions that will make it fail.

Given some ϵ and a class F of VC-dimension d, we can define an $m_0 \in \mathbf{N}$, whose value is appropriately chosen in terms of ϵ and d. Consider a subset $X' \subseteq X$, $|X'| = d$ such that the restriction of F to X' contains all functions from X' to $\{0,1\}$ (i.e. a "shattered" set X'). For a large t, we randomly choose a γ-admissible sequence of distributions $P_1, P_2, \ldots, P_t$ on $X' \times \{0,1\}$. The distribution is such that we can characterize the algorithm which minimizes the overall probability of making a mistake on the tth prediction. Call this algorithm A. Moreover, $\inf_{f\in F} P_t\{(x,y) : f(x) \neq y\}$ is obvious, because X' is shattered. This infimum is realized by a function which evaluates to 1 at

exactly those elements of X' for which the corresponding conditional probability of 1 is at least $1/2$ (the behavior outside X' is immaterial since $P_t\{(x,y) : x \notin X'\} = 0$).

Set $c_1 = 1100000$. Given any $\epsilon \leq 1/1100$, let $\gamma = \frac{c_1\epsilon^3}{d}$. Below we describe the chosen γ-admissible distribution sequence $P_1, P_2, \ldots, P_t$ on $X' \times \{0,1\}$. Let $t = m_0 + k$ and $X' = \{x_1, x_2, \ldots, x_d\}$. For each j, $1 \leq j \leq d$, let $C_j = 1$ if an independent flip of an unbiased coin is heads; otherwise let $C_j = 0$. The following constraints uniquely define the target distribution sequence. Note that "drifting" begins only after the kth sample and for each x_j, the drift direction is randomly assigned. We call the kth sample the *drift initiating sample*.

1. $P_i(x_j, 0) + P_i(x_j, 1) = \frac{1}{d}$ for $1 \leq i \leq t, 1 \leq j \leq d$.
2. $P_i(x_j, 0) = P_i(x_j, 1) = \frac{1}{2d}$ for $1 \leq i \leq k, 1 \leq j \leq d$.
3. $P_{k+i}(x_j, C_j) = \frac{1}{2d} + \frac{i\gamma}{2d}$ for $1 \leq i \leq (t-k)$, $1 \leq j \leq d$.

We have the following straightforward characterization of the optimal function for a particular sequence of distributions.

Lemma 14 *If the ith sample is drawn from the distribution P_i, then the function f_{opt} which minimizes prediction error from among those in F is such that $\forall j, f_{\mathrm{opt}}(x_j) = C_j$. The corresponding probability of error is $1/2 - i'\gamma/2$, where $i' = \max\{0, i-k\}$.*

A similar observation to the following lemma was made by Simon [Sim93].

Lemma 15 *Consider the algorithm A which ignores information from all samples seen prior to and including the kth sample, the drift initiating sample. For $i \leq k$, A behaves arbitrarily. For $i > k$, if the ith sample is the point x_j of X', A outputs a 1 if a majority of the previous instances of x_j that were seen* after the drift initiating sample *were associated with 1; A outputs a 0 otherwise. Then A is an optimal online algorithm. That is, if the sequence of probability distributions is chosen randomly as described above, any other online algorithm will have probability of mistake at least that of A.*

Proof: As is well known (see [DH73]), the optimal algorithm is obtained by choosing the hypothesis to minimize the a posteriori probability of making a mistake.

Since the probability that a hypothesis makes a mistake on the tth example is a linear combination of the probabilities that it makes a mistake given that the various elements of X' are observed on the tth trial, minimizing the overall probability of a mistake can be achieved by separately minimizing the probability of a mistake given that each element of X' is observed last.

Choose x_j. Given that a majority of the time x_j was seen since drift began the corresponding label was 1, the a posteriori probability that $C_j = 1$ is at least $1/2$. This implies that in this case, the a posteriori probability of a mistake given that x_j is seen last is minimized by guessing one. Similarly, if the label was 0 a majority of the drifting time, the a posteriori probability of a mistake is minimized by predicting 0. □

Having established the optimal online algorithms and the optimal element of F for a particular sequence of distributions, we now derive an expression for the difference in the probabilities of prediction errors.

Lemma 16 *Consider the tth sample ($t \geq m_0$) drawn according from P_t. Let p be the probability that the hypothesis output by A and f_{opt} do not agree on their prediction outputs for the tth sample. Then the difference in their probabilities of prediction errors is $p\gamma m_0$.*

Proof: If $e_B = 1/2 - \gamma m_0/2$ represents the probability that f_{opt} errs in its prediction for the tth sample, we note that the probability that A makes a prediction error on the tth sample is $p(1-e_B)+(1-p)e_B$. The required difference in the probabilities of prediction error is then simply $p(1-e_B)+(1-p)e_B - e_B$, which can be simplified to $p\gamma m_0$. □

We have now reduced the problem to one of finding a lower bound on the quantity p above. The following simple observation plays a role in the lower bound.

Lemma 17 *For any odd $m \in \mathbf{N}$,*

$$\sum_{l=(m-1)/2-\lceil\sqrt{m}\rceil}^{(m-1)/2} \binom{m}{l} \geq \frac{1}{2}(1-2e^{-2})2^m.$$

Proof: If we flip an unbiased coin m times Lemma 2 implies that the probability that the number H of heads satifies $|H/m - 1/2| \leq 1/\sqrt{m}$ is at least $1-2e^{-2}$. Since each sequence of outcomes is equally likely,

$$\frac{1}{2^m}\sum_{l=\lfloor m/2-\sqrt{m}\rfloor}^{\lceil m/2+\sqrt{m}\rceil} \binom{m}{l} \geq 1-2e^{-2}.$$

Solving, we get

$$\sum_{l=\lfloor m/2-\sqrt{m}\rfloor}^{\lceil m/2+\sqrt{m}\rceil} \binom{m}{l} \geq (1-2e^{-2})2^m$$

and by symmetry

$$\sum_{l=\lfloor m/2-\sqrt{m}\rfloor}^{(m-1)/2} \binom{m}{l} \geq \frac{1}{2}(1-2e^{-2})2^m$$

which directly implies the lemma. □

Now we are able to show that p is at least a constant, and we do so by choosing an appropriate value for[2] m_0 below.

Theorem 18 *If $\epsilon \leq 1/1100$, a necessary condition for a class F of functions from X to $\{0,1\}$ to be agnostically (ϵ,γ)-learnable is $\gamma < \frac{1100000\epsilon^3}{\text{VCdim}(F)}$.*

Proof: Assume as above that $\gamma = \frac{c_1\epsilon^3}{\text{VCdim}(F)}$. Set $m_0 = \lceil 40d/(c_1\epsilon^2)\rceil$. Note that since $\epsilon \leq 1/1100$,

$$\frac{40d}{c_1\epsilon^2} \leq m_0 \leq \frac{41d}{c_1\epsilon^2}. \tag{7}$$

[2]This proof requires us to lower bound the tail of the binomial distribution. As is apparently required, our lower bound is stronger than what follows from the more general bound of Littlestone [Lit90]. A similar lower bound was proved by Simon [Sim93], who appealed to the central limit theorem. Our bound has the advantage that it yields concrete constants.

Define p as it is earlier in this section.

Consider the time at which the tth sample is drawn as per the distribution P_t of our distribution sequence. Now, since less than m_0 samples were seen after the drift initiating kth sample, at most $d/2$ elements of X' were seen at least $2m_0/d$ times and so at least $d/2$ elements of X' were seen less than $2m_0/d$ times *after* drift initiation. Let m be the least odd integer which is at least $2m_0/d$. We will restrict our attention to the subset $X'' \subset X'$ of the at least $d/2$ elements seen less than m times after drift initiation. The probability that the tth sample is from X'' is thus at least $1/2$.

Moreover, if the tth sample is any particular element $x_j \in X''$, because the probability that more than half of m tosses of a biased coin come up heads is monotone in m, we can assume that x_j was seen exactly m times after drift initiation. Let us recall the assignment of the random variable C_j for $1 \leq j \leq d$ made above and the definition of the γ-admissible sequence and the algorithm A and f_{opt}. Then A disagrees with f_{opt} on the tth sample only if x_j was associated with C_j on l instances and with $1 - C_j$ on $m - l$ instances after drift initiation with $0 \leq l < m/2$.

The probability that this happens is easily seen to be the probability that one gets less than $m/2$ heads when flipping m biased coins having probability of heads between $1/2$ and $1/2 + m_0\gamma/2$. But this is at least the probability that one gets less than $m/2$ heads when sampling m times from the distribution where the probability of heads is $1/2 + m_0\gamma/2$. (It is worth mentioning that substituting into the definitions of m_0 and γ and applying the fact that $\epsilon \leq 1/1100$ implies that $m_0\gamma/2 \leq 1/10$.)

By virtue of the above paragraphs, with $\beta = m_0\gamma/2$, we have that

$$\begin{aligned}
p &\geq \frac{1}{2}\sum_{l=0}^{(m-1)/2}\binom{m}{l}(1/2+\beta)^l(1/2-\beta)^{m-l} \\
&= \frac{1}{2}(1/2-\beta)^m\sum_{l=0}^{(m-1)/2}\binom{m}{l}\left(\frac{1/2+\beta}{1/2-\beta}\right)^l \\
&\geq \frac{1}{2}2^{-m}(1-2\beta)^m\sum_{l=(m-1)/2-\lceil\sqrt{m}\rceil}^{(m-1)/2}\binom{m}{l}\left(\frac{1+2\beta}{1-2\beta}\right)^l \\
&\geq \frac{1}{2}2^{-m}(1-2\beta)^m\sum_{l=(m-1)/2-\lceil\sqrt{m}\rceil}^{(m-1)/2}\binom{m}{l}(1+2\beta)^{2l} \qquad (\text{since } \forall x, 1/(1-x) \geq 1+x) \\
&\geq \frac{1}{2}2^{-m}(1-2\beta)^m(1+2\beta)^{2(m/2-\sqrt{m}-3/2)}\left(\sum_{l=(m-1)/2-\lceil\sqrt{m}\rceil}^{(m-1)/2}\binom{m}{l}\right)
\end{aligned}$$

Applying Lemma 17 and rearranging a little gives

$$p \geq \frac{1}{4}2^{-m}(1-4\beta^2)^m(1+2\beta)^{-2\sqrt{m}-3}2^m(1-2e^{-2}) \tag{8}$$

Recall that $\beta \leq 1/10$, which implies $(1-4\beta^2)^m \geq e^{-5\beta^2 m}$ since $(1-4\beta^2)^{1/(4\beta^2)} \geq (1-1/4)^4 \geq e^{-5/4}$. On the other hand $(1+2\beta)^{-2\sqrt{m}-3} \geq e^{2\beta(-2\sqrt{m}-3)}$ trivially, and since $\beta \leq 1/10$, this implies $(1+2\beta)^{-2\sqrt{m}-3} \geq e^{-4\beta\sqrt{m}}/2$. Thus

$$p \geq \frac{1-2/e^2}{8}e^{-5\beta^2 m}e^{-4\beta\sqrt{m}}.$$

Define the RHS of this bound by c.

Using Lemma 16 and substituting $\gamma = c_1\epsilon^3/d$ implies that the difference in probabilities of prediction error of A and f_{opt} is *at least* $c\gamma m_0$ with c as defined above. Plugging in the value of c, we have

$$c\gamma m_0 = \frac{1-2/e^2}{8} e^{-5\beta^2 m} e^{-4\beta\sqrt{m}} \gamma m_0 \geq \frac{e^{-5\beta^2 m} e^{-4\beta\sqrt{m}} \gamma m_0}{11}.$$

By (7), since $\epsilon \leq 1/1100$, we have $2m_0/d \geq 80/(c_1\epsilon^2) > 80$, so m, the least odd integer which is at least $2m_0/d$, is at most $3m_0/d$. Thus

$$c\gamma m_0 \geq \frac{e^{-15\beta^2 m_0/d} e^{-7\beta\sqrt{m_0/d}} \gamma m_0}{11}.$$

Substituting $m_0\gamma/2$ for β, we get

$$c\gamma m_0 \geq \frac{e^{-4m_0^3\gamma^2/d} e^{-4\gamma\sqrt{m_0^3/d}} \gamma m_0}{11}.$$

Substituting $c_1\epsilon^3/d$ for γ, we get

$$c\gamma m_0 \geq \frac{e^{-4m_0^3 c_1^2\epsilon^6/d^3} e^{-4c_1\sqrt{m_0^3\epsilon^6/d^3}} c_1\epsilon^3 m_0}{11d}.$$

Applying (7), we get

$$c\gamma m_0 \geq \frac{e^{-4\cdot 41^3/c_1} e^{-4\sqrt{41^3/c_1}} 40\epsilon}{11} > \epsilon.$$

Thus the difference in probabilities of prediction error of A and f_{opt} is more than ϵ.

Applying Lemmas 15 and 14, we can see that if $\gamma \geq \frac{1100000\epsilon^3}{d}$, then there is a distribution over $P_1, ..., P_t$ such that the difference between the probability that the Bayes optimal algorithm makes a mistake on the tth prediction and $\inf_{f\in F} P_t\{(x,y) : f(x) \neq y\}$ is greater than ϵ.

Now we apply the trick of Ehrenfeucht, Haussler, Kearns and Valiant [EHKV89]. The above lower bound for the Bayes optimal algorithm implies that for any on-line algorithm A', the difference between the probability that A' makes a mistake on the tth prediction and $\inf_{f\in F} P_t\{(x,y) : f(x) \neq y\}$ is greater than ϵ. This implies that there is a particular sequence $P_1, ..., P_t$ such that the probability that A' makes a mistake on its tth prediction is greater than $\inf_{f\in F} P_t\{(x,y) : f(x) \neq y\} + \epsilon$, completing the proof. □

5 Conclusion

In this paper, we have determined, to within a log factor, the complexity of learning according to two models which capture a drifting environment.

The algorithms for agnostic learning analyzed in this paper work by minimizing disagreements with some of the most recent examples. For many simple concept classes, such as monomials [AL88] and halfspaces [HS92], this has been shown to be NP-hard, and even a nonapproximability result for the latter is known [ABSS93].

Helmbold and Long [HL94] showed that approximation algorithms for minimizing disagreements could be applied for "noise-free" learning of drifting concepts, but their techniques do not apparently

extend to the agnostic case, at least not to bound the *difference* between the learner's error and that of the best function in F, as in the model considered here. One direction for future research would be to search for efficient algorithms for agnostic learning in a drifting environment.

Another obvious problem is to try to close the log factor gaps that remain in these models. It seems possible that modifying the proof of a result of Talagrand [Tal94] in a manner analogous to the modification of the proof of Vapnik and Chervonenkis given here might improve the general bound on the rate of drift sufficient for agnostic learning by a log factor. It also seems possible that the lower bound of Haussler, Littlestone and Warmuth [HLW94] for solid learning in a fixed environment can be adapted to show that the bound given here for solid learning in a drifting environment cannot be improved in general by more than a constant factor.

Acknowledgements

We are very grateful to an anonymous referee for many helpful comments on an earlier version of this paper.

Rakesh Barve was supported by an IBM Fellowship. Phil Long was supported by Office of Naval Research grant N00014–94–1–0938 and National University of Singapore Academic Research Fund Grant RP960625.

References

[AB92] Martin Anthony and Norman Biggs. *Computational learning theory: an introduction*, volume 30 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 1992.

[ABS90] M. Anthony, N. Biggs, and J. Shawe-Taylor. The learnability of formal concepts. *The 1990 Workshop on Computational Learning Theory*, pages 246–257, 1990.

[ABSS93] S. Arora, L. Babai, J. Stern, and Z. Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. *Proceedings of the 34th Annual Symposium on Foundations of Computer Science*, pages 724–733, 1993.

[AL88] D. Angluin and P.D. Laird. Learning from noisy examples. *Machine Learning*, 2:343–370, 1988.

[AW95] P. Auer and M. K. Warmuth. Tracking the best disjunction. *Proceedings of the 36th Annual Symposium on the Foundations of Computer Science*, 1995.

[Bar92] P.L. Bartlett. Learning with a slowly changing distribution. *The 1992 Workshop on Computational Learning Theory*, pages 243–252, 1992.

[BBK96] P.L. Bartlett, S. Ben-David, and S.R. Kulkarni. Learning changing concepts by exploiting the structure of change. *The 1995 Conference on Computational Learning Theory*, pages 131–139, 1996.

[BBM89] S. Ben-David, G. M. Benedek, and Y. Mansour. A parametrization scheme for classifying models of learnability. In *Proc. 2nd Annu. Workshop on Comput. Learning Theory*, pages 285–302, San Mateo, CA, 1989. Morgan Kaufmann.

[BC92] A. Blum and P. Chalasani. Learning switching concepts. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages pages 231–242, 1992.

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *JACM*, 36(4):929–965, 1989.

[BH95] P.L. Bartlett and D.P. Helmbold, 1995. Manuscript.

[BL95] P. L. Bartlett and P. M. Long. Prediction, learning, uniform convergence, and scale-sensitive dimensions, 1995. Submitted.

[Daw84] A. Dawid. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society (Series A)*, pages 278–292, 1984.

[DH73] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.

[EHKV89] A. Ehrenfeucht, D. Haussler, M. Kearns, and L.G. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–251, 1989.

[FM97] Y. Freund and Y. Mansour. Learning under persistent drift. *Proceedings of the 1997 European Conference on Computational Learning Theory*, 1997.

[Hau92] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inform. Comput.*, 100(1):78–150, 1992.

[HL91] D.P. Helmbold and P.M. Long. Tracking drifting concepts using random examples. *The 1991 Workshop on Computational Learning Theory*, pages 13–23, 1991.

[HL94] D.P. Helmbold and P.M. Long. Tracking drifting concepts by minimizing disagreements. *Machine Learning*, 14(1):27–46, 1994.

[HLW94] D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):129–161, 1994.

[HS92] K. Höffgen and H. Simon. Robust trainability of single neurons. *The 1992 Workshop on Computational Learning Theory*, pages 428–439, 1992.

[HW95] M. Herbster and M.K. Warmuth. Tracking the best expert. *Proceedings of of the Twelvth International Conference on Machine Learning*, 1995.

[KPR90] T. Kuh, T. Petsche, and R. Rivest. Learning time varying concepts. In *NIPS 3*. Morgan Kaufmann, 1990.

[KPR91] T. 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.

[Lit90] N. Littlestone. On the derivation and quality of Chernoff bounds, 1990. Submitted.

[LW94] N. Littlestone and M.K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

[Nat91] B.K. Natarajan. *Machine Learning: A Theoretical Approach*. Morgan Kaufmann, San Mateo, California, 1991.

[Pol84] D. Pollard. *Convergence of Stochastic Processes*. Springer Verlag, 1984.

[Roy63] H.L. Royden. *Real Analysis*. Macmillan, 1963.

[Sau72] N. Sauer. On the density of families of sets. *J. Combinatorial Theory (A)*, 13:145–147, 1972.

[Sim93] H.U. Simon. General lower bounds on the number of examples needed for learning probabilistic concepts. *The 1993 Conference on Computational Learning Theory*, pages 402–412, 1993.

[Tal94] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, 22:28–76, 1994.

[Val84] L.G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[VC71] 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.