Learning with a Slowly Changing Distribution

Peter L. Bartlett
Department of Electrical Engineering
University of Queensland
Queensland 4072
AUSTRALIA
bartlett@s1.elec.uq.oz.au

Abstract

In this paper, we consider the problem of learning a subset of a domain from randomly chosen examples when the probability distribution of the examples changes slowly but continually throughout the learning process. We give upper and lower bounds on the best achievable probability of misclassification after a given number of examples. If d is the VC-dimension of the target function class, t is the number of examples, and γ is the amount by which the distribution is allowed to change (measured by the largest change in the probability of a subset of the domain), the upper bound decreases as d/t initially, and settles to $O(d^{2/3}\gamma^{1/3})$ for large t. The general lower bound on the probability of misclassification again decreases as d/t initially, but settles to $\Omega(d^{1/2}\gamma^{1/2})$ for large t. These bounds give necessary and sufficient conditions on γ, the rate of change of the distribution of examples, to ensure that some learning algorithm can produce an acceptably small probability of misclassification. We also consider the case of learning a near-optimal subset of the domain when the examples and their labels are generated by a joint probability distribution on the example and label spaces. We give an upper bound on γ that ensures learning is possible from a finite number of examples.

1 INTRODUCTION

In this paper, we examine the problem of learning a subset of a domain from randomly-chosen examples when the distribution of examples changes as learning proceeds. We are interested in how upper bounds on learning curves (graphs of error probability versus number of examples) vary with the amount by which the distribution is allowed to change.

For many learning problems we can expect the distribution of examples (and the target function) to change over time. Consider a learning system in a telecommunications network that aims to avoid network congestion by controlling the admission of calls. The distribution of inputs (and the optimal decision function) for such a system will change with time as the network usage changes, as the channel characteristics (and therefore error rates) change, and as parts of the network fail.

We consider two models of learning. The first is similar to Haussler, Littlestone and Warmuth's prediction model [HLW90]—the aim of learning is to minimize the probability over all sequences of examples of misclassifying the last example. The second is a more general model that allows noise and errors in the examples. In both cases, the distribution is allowed to change slowly throughout the learning process. The amount by which the distribution changes is measured by the largest change in the probability of a subset of the domain.

In [Kra88], Kramer presents a related model of learning, in which the distribution is allowed to drift. However in Kramer's model, when the learning system is presented with an example it can choose to see the classification of the example or to guess its classification (using a hypothesis from a particular class of hypotheses). The aim is for the algorithm to guess the label only if its hypothesis is accurate with high probability (taken over all sequences of random examples, as in Valiant's pac model [Val84]). Kramer is concerned with the minimum number of labelled examples that a successful algorithm of

this type must store. In contrast, the results presented here give bounds on the misclassification probability for an optimal algorithm as a function of the number of examples and the amount of distribution drift.

Helmbold and Long [HL91] consider learning a slowly changing subset of the domain, when the distribution of examples is constant. This problem, and the problem of learning a fixed subset with a changing distribution, are two special cases of a model of learning in which the labelled examples are described by a slowly changing joint distribution on the input and output spaces. We examine this more general model in Section 5.

The paper is organized as follows. In Section 2, we present some notation and formally define the learning model we use. We compare some natural definitions of the distance between distributions. In Section 3 we give upper bounds on the probability of misclassification for two general-purpose algorithms: the one-inclusion graph prediction strategy (presented in [HLW90]), and a consistent hypothesis finder. Section 4 gives a general lower bound on the probability of misclassification. In Section 5, we apply the techniques used in Section 3 to the problem of learning an optimal classification function when the joint distribution on the input and output spaces varies slowly as learning proceeds. In Section 6, we summarize the results and mention some possible extensions.

2 DEFINITIONS AND NOTATION

If D is a distribution on a set X and $P(x)$ is a proposition about $x \in X$, then we denote the probability that $P(x)$ is true when x is chosen according to D by

$$\Pr_{x \in D}(P(x)) = D\{x \in X : P(x)\}.$$

Similarly, if f is a real-valued function defined on X, then $E_{x \in D}(f(x))$ represents the expectation of $f(x)$ when x is chosen according to D,

$$E_{x \in D}(f(x)) = \int_{x \in X} f(x)\, dD(x).$$

We sometimes use $E_D(f) = E_{x \in D}(f(x))$ when the meaning is clear from the context. We assume throughout that every set is measurable ([HL91] gives a supporting argument, claiming that in practice the domain we consider is countable; [BEHW89] gives sufficient conditions for the assumption when the domain is $\mathbb{R}^n$). If $x = (x_1, x_2, \ldots, x_t) \in X^t$ and σ is a permutation on $\{1, 2, \ldots, t\}$, define

$$x^\sigma = \left(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(t)}\right).$$

Given a set X and a set F of functions that map from X to $\{0, 1\}$, we say that F shatters the finite subset $S \subseteq X$ if the functions in F induce all possible dichotomies of S,

$$|\{\{x \in S : f(x) = 1\} : f \in F\}| = 2^{|S|}.$$

The Vapnik-Chervonenkis dimension (VC-dimension) of F is the size of the largest shattered subset of X,

$$\mathrm{VCdim}(F) = \max\{m : \exists S \subseteq X \ |S| = m \text{ and } F \text{ shatters } S\}$$

(see [VC71]).

The learning model described here is similar to the prediction model of learning described in [HLW90]. We have a *domain* X, a class F of functions that map from X to $\{0, 1\}$ (the *target class*), and a *target function* f in F—the function we are trying to learn. At each learning trial, an *example* x is randomly chosen from X. The learning algorithm tries to predict the value of $f(x)$ (the *label* of x). The algorithm is then told the label, and the process is repeated.

A sequence $x = (x_1, x_2, \ldots, x_t) \in X^t$ of examples is called a *sample*. A *labelled sample* is a sequence $((x_1, f(x_1)), \ldots, (x_t, f(x_t)))$ of labelled examples. For sample $x = (x_1, x_2, \ldots, x_t) \in X^t$ and target function f in F, define the labelled sample of f generated by x as

$$\mathrm{sam}_t(x, f) = ((x_1, f(x_1)), \ldots, (x_t, f(x_t))).$$

Instead of assuming that each example is chosen independently from a single distribution on X, we assume that each example x_i is drawn from a (possibly distinct) distribution P_i. The sequence of distributions $\langle P_i \rangle$ is intended to describe the change in the relative frequency of examples as learning proceeds. To quantify that change, we need some definition of the distance between two distributions.

2.1 COMPARING DISTRIBUTIONS

We assume that there is a σ-field $\mathcal{F}$ of subsets of X on which the probability distributions P_i are defined. We define the distance between distributions P_1 and P_2 as the largest change in probability of a subset in $\mathcal{F}$.

Definition 1 *The distance $d(P_1, P_2)$ between two distributions P_1 and P_2 is*

$$d(P_1, P_2) = \sup_{E \in \mathcal{F}} |P_1(E) - P_2(E)|.$$

We can also define this distance using a signed measure μ on the measurable space $(X, \mathcal{F})$, defined as

$$\mu = P_1 - P_2.$$

For this signed measure, choose a partition $\{A, B\}$ of X for which μ is positive in A and negative in B, and define

two measures on $(X, \mathcal{F})$ (the *upper* and *lower variations* of μ),

$$\mu^+(E) = \mu(E \cap A)$$

and

$$\mu^-(E) = -\mu(E \cap B)$$

for $E \in \mathcal{F}$. Clearly, $\mu^+, \mu^- \geq 0$, and $\mu = \mu^+ - \mu^-$. Using this representation (the Jordan decomposition of μ), the distance d is given by

Proposition 2

$$d(P_1, P_2) = \mu^+(X).$$

Proof Since μ^+ and μ^- are measures, $\sup_{E \in \mathcal{F}} \mu^+(E) = \mu^+(X)$ and $\sup_{E \in \mathcal{F}} \mu^-(E) = \mu^-(X)$. But $\mu^+(X) - \mu^-(X) = \mu(X) = 0$, so $\mu^+(X) = \mu^-(X)$. Thus $\sup_{E \in \mathcal{F}} |\mu^+(E) - \mu^-(E)| = \mu^+(X)$. □

The measure $|\mu|$ defined by $|\mu| = \mu^+ + \mu^-$ is called the *total variation* of μ.

To prove upper bounds on the mistake probability under a drifting distribution, we will use the following result. It bounds the difference between the expectations of a $[0, 1]$-valued random variable under two distributions that are close in the distance d.

Lemma 3 *Consider two probability distributions P_1 and P_2 on the measurable space $(X, \mathcal{F})$ that satisfy*

$$d(P_1, P_2) \leq \gamma, \tag{1}$$

where $0 < \gamma \leq 1$. If f is an $\mathcal{F}$-measurable function from X to $[0, 1]$, then

$$|E_{P_1}(f) - E_{P_2}(f)| \leq \gamma. \tag{2}$$

Proof Define the signed measure $\mu = P_2 - P_2$ as above. By definition,

$$\begin{aligned} |E_{P_1}(f) - E_{P_2}(f)| &= \left| \int_X f \, dP_2 - \int_X f \, dP_1 \right| && (3) \\ &= \left| \int_X f \, d\mu \right| && (4) \\ &= \left| \int_X f \, d\mu^+ - \int_X f \, d\mu^- \right| && (5) \end{aligned}$$

Now, $0 \leq \int_X f \, d\mu^+ \leq \mu^+(X)$ for $0 \leq f \leq 1$ (see [Hal50], p124) so

$$|E_{P_1}(f) - E_{P_2}(f)| \leq \mu^+(X) \leq \gamma.$$

□

2.1.1 Other Distances

This section examines two other commonly used ways of defining the distance between distributions, and compares them with d.

In the definitions in this section, let $(X, \mathcal{F}, P)$ and $(X, \mathcal{F}, Q)$ be probability spaces.

Definition 4 *The total variation distance between P and Q is*

$$d_V(P, Q) = |\mu|(X),$$

where $|\mu|$ is the total variation of the signed measure $\mu = P - Q$.

Suppose P and Q are discrete distributions with supports in the set $\{x_1, x_2, \ldots, x_n\} \subseteq X$, with $P(x_i) = p_i$, $Q(x_i) = q_i$ for $i = 1, 2, \ldots, n$. Then the definition reduces to

$$d_V(P, Q) = \sum_{i=1}^{n} |p_i - q_i| .$$

Proposition 5 *The total variation distance d_V is related to d by*

$$d = d_V/2.$$

Proof Using the signed measure μ defined above, we have $d_V = |\mu|(X) = \mu^+(X) + \mu^-(X) = 2\mu^+(X) = 2d$. □

A natural definition of the distance between two distributions is the Kullback-Leibler divergence.

Definition 6 *The Kullback-Leibler divergence of P with respect to Q is*

$$d_{KL}(P, Q) = \int_X p(\omega) \log \frac{p(\omega)}{q(\omega)} \, d\lambda(\omega),$$

where λ is a measure on $(X, \mathcal{F})$ such that P and Q are absolutely continuous with respect to λ, and p and q are the Radon-Nikodym derivatives of P and Q with respect to λ, $p = dP/d\lambda$, $q = dQ/d\lambda$. Notice that $d_{KL}(P, Q)$ is not a symmetric function of its arguments.

If P and Q are the discrete distributions defined above, this definition reduces to

$$d_{KL}(P, Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i},$$

with the conventions $0 \log 0 = 0$ and $\log 0/0 = 1$.

The quantity $d_{KL}(P, Q)$ is also known as the information of order 1 of P with respect to Q. It can be interpreted as the amount of information obtained from observing an event E for which $P(\cdot) = Q(\cdot|E)$ (see [Ren61]).

The following proposition shows that a bound on d_{KL} is a stronger requirement than a bound on d.

Proposition 7 *The Kullback-Leibler divergence d_V is related to d by*

$$d^2 \leq d_{KL}/2.$$

Moreover, there are distributions P and Q for which $d(P,Q) \leq \gamma$ but $d_{KL}(P,Q) = \infty$, for $0 < \gamma \leq 1$.

Proof Kullback [Kul67] shows that $d_{KL} \geq d_V^2/2 + d_V^4/12$. Proposition 5 gives the desired inequality.

To see that d does not provide an upper bound on d_{KL}, consider the distributions P and Q and the set $\{x_1, x_2\} \subseteq X$, with $P(x_1) = 1-\gamma$, $P(x_2) = \gamma$, $Q(x_1) = 1$, and P and Q zero elsewhere. Clearly, $d(P,Q) = \gamma$, but $d_{KL}(P,Q) = \infty$. □

This proposition implies that, if we use d_{KL} instead of d to measure the change in the distribution of examples, the upper bounds described in Sections 3 and 5 are still applicable.

2.2 THE DEFINITION OF LEARNING

We restrict the amount by which the distribution can change between examples by bounding the distance d between consecutive distributions.

Definition 8 (Admissible Distribution Sequence) *Suppose $(X, \mathcal{F}, P_i)$ is a probability space, $i = 1, 2, \ldots, t$, $t > 0$. The sequence $\langle P_i \rangle_{i=1}^t$ is in the class $\mathcal{D}_\gamma^t$ $(0 \leq \gamma < 1)$ of γ-admissible distribution sequences if,*

$$d(P_i, P_{i+1}) \leq \gamma$$

for $i = 1, \ldots, t-1$.

The learning algorithms we consider are prediction strategies (see [HLW90]).

Definition 9 (Prediction Strategy) *Consider an input space X, a class F of functions from X to $\{0,1\}$, and define the space of labelled examples, $S = X \times \{0,1\}$ and the space of finite length labelled samples, $S^* = \cup_{m \in \mathbb{N}} S^m$.*

A deterministic prediction strategy Q for F is a function from $S^* \times X$ to $\{0,1\}$. A randomized prediction strategy (Q_r, Z, D) for F consists of a function Q_r, a space Z, and a distribution D on Z. The strategy chooses a point $z \in Z$ according to D, and passes z to the function Q_r, which maps from $S^* \times X \times Z$ to $\{0,1\}$.

We define the *mistake probability* of a prediction strategy as follows.

Definition 10 (Mistake Probability) *For sample $x = (x_1, \ldots, x_t) \in X^t$ $(t \geq 1)$, $f \in F$, and deterministic prediction strategy Q, define the mistake of Q on x with respect to f as*

$$M_{Q,f}^t(x) = \begin{cases} 1 & Q(\mathrm{sam}_t((x_1, \ldots, x_{t-1}), f), x_t) \neq f(x_t) \\ 0 & \text{otherwise.} \end{cases}$$

For randomized prediction strategy (Q_r, Z, D), define

$$M_{Q,f}^t(x) = D\left\{z \in Z : Q\left(\mathrm{sam}_t((x_1, \ldots, x_{t-1}), f), x_t, z\right) \neq f(x_t)\right\}.$$

For a distribution sequence $\langle P_i \rangle_{i=1}^t$ on X, define the mistake probability of Q with respect to f as

$$E_{x \in \langle P_i \rangle}\left(M_{Q,f}^t(x)\right).$$

We want this probability to be small for all distributions and all target functions.

Definition 11 ((ϵ,γ)-Prediction) *Consider a class F of functions and a prediction strategy Q for F. If f is in F, $\gamma \geq 0$ and $t > 0$, let $\hat{M}_{Q,f,\gamma}(t)$ be the supremum over all γ-admissible distribution sequences $\langle P_i \rangle$ on X of the mistake probability,*

$$\hat{M}_{Q,f,\gamma}(t) = \sup_{\langle P_i \rangle \in \mathcal{D}_\gamma^t} \left\{E_{\langle P_i \rangle}\left(M_{Q,f}^t\right)\right\}.$$

Define the mistake bound,

$$\hat{M}_{Q,F,\gamma}(t) = \sup_{f \in F} \hat{M}_{Q,f,\gamma}(t).$$

We say that Q can (ϵ,γ)-predict F if $\hat{M}_{Q,F,\gamma}(t) < \epsilon$ for some finite t.

3 UPPER BOUNDS

In this section, we give mistake bounds for function classes of finite VC-dimension. To obtain these bounds for the constant-distribution case, we can use the observation that permuting the examples in a sample will not affect the mistake probability, since the distribution on X^t is a product distribution. This allows us to relate the mistake probability to an average of mistakes over a set of permutations (see [BEHW89, HLW90, Vap82]). The proof we use here is similar, but the distribution on X^t is not a product distribution. We proceed by bounding how far the mistake probabilities are from expectations under some product distribution. We can then use the permutation device to bound this expectation. Notice that it does not matter which product distribution we use to bound the mistake probability.

Lemma 12 *If $\langle P_i \rangle_{i=1}^k$ is a γ-admissible distribution sequence on X (where $0 < \gamma \leq 1$) and f is a measurable function from X^k to $[0,1]$ (with $k \geq 1$), then*

$$E_{x \in \langle P_i \rangle_{i=1}^k}(f(x)) \leq E_{x \in P_1^k}(f(x)) + \frac{k(k-1)}{2}\gamma, \qquad (6)$$

and

$$E_{x\in\langle P_i\rangle_{i=1}^k}(f(x)) \le E_{x\in P_k^k}(f(x)) + \frac{k(k-1)}{2}\gamma. \qquad (7)$$

Proof We are interested in the expectation

$$E_{\langle P_i\rangle_{i=1}^k}(f) = \int_{X^k} f(x_1,\ldots,x_k)\, dP_1(x_1)\ldots dP_k(x_k)$$

$$= \int_{X^{k-2}}\int_X\int_X f\, dP_1(x_1)\, dP_2(x_2)\ldots dP_k(x_k).$$

Fix $x_3, x_4, \ldots, x_k$ and consider the integral

$$\int_X\int_X f\, dP_1(x_1)\, dP_2(x_2) = E_{x_2\in P_2}\left(\int_X f\, dP_1(x_1)\right).$$

Call the random variable inside the parentheses $I(x_2)$. Notice that $0 \le I \le 1$, so Lemma 3 gives

$$\begin{aligned} E_{x_2\in P_2}(I(x_2)) &\le E_{x_2\in P_1}(I(x_2)) + \gamma \\ &= \int_{X^2} f\, dP_1^2(x_1,x_2) + \gamma. \end{aligned}$$

Therefore

$$E_{\langle P_i\rangle_{i=1}^k}(f) \le \int_{X^{k-2}}\int_{X^2} f\, dP_1^2(x_1,x_2)\ldots dP_k(x_k) + \gamma.$$

Similarly,

$$E_{\langle P_i\rangle_{i=1}^k}(f) \le$$
$$\int_{X^{k-3}}\int_{X^3} f\, dP_1^3(x_1,x_2,x_3)\ldots dP_k(x_k) + 2\gamma + \gamma$$

and

$$\begin{aligned} E_{\langle P_i\rangle_{i=1}^k}(f) &\le \int_{X^k} f\, dP_1^k(x_1,x_2,\ldots,x_k) + \gamma\sum_{i=1}^{k-1} i \\ &= E_{P_1^k}(f) + \frac{k(k-1)}{2}\gamma, \end{aligned}$$

which is Inequality (6). The same argument with the labels for $P_1\ldots P_k$ reversed gives Inequality (7). □

3.1 AN UPPER BOUND FOR THE ONE-INCLUSION GRAPH PREDICTION STRATEGY

We can relate the mistake bound to a certain permutation mistake bound. We will use the set of permutations on $\{1,\ldots,t\}$ that swap t with one of the elements of $\{t-(k-1),\ldots,t\}$ and leave the other elements unchanged. Call this class of permutations $\Gamma_{t,k}$. Formally,

$$\Gamma_{t,k} = \{\sigma_i : i = t-k+1,\ldots,t\}$$

where

$$\sigma_i(j) = \begin{cases} i & j = t \\ t & j = i \\ j & \text{otherwise} \end{cases}$$

Now define the permutation mistake bound,

$$\hat{\hat{M}}_{Q,F}(t,k) =$$
$$\sup\left\{\frac{1}{|\Gamma_{t,k}|}\sum_{\sigma\in\Gamma_{t,k}} M_{Q,f}^t(x^\sigma) : f\in F, x\in X^t\right\}$$

for $t = 1,2,\ldots$, and $k = 1,2,\ldots,t$. We can relate this bound to the mistake bound as follows.

Theorem 13 *Consider a prediction strategy Q for function class F, with permutation mistake bound $\hat{\hat{M}}_{Q,F}(t,k)$. For this prediction strategy,*

$$\hat{M}_{Q,F,\gamma}(t) \le \hat{\hat{M}}_{Q,F}(t,k) + \frac{k(k-1)}{2}\gamma \qquad (8)$$

for $k = 1,2,\ldots,t$ and $0 < \gamma \le 1$.

We will use the following lemma ([HLW90], Lemma 2.1) involving permutations of components of a random vector under a product distribution.

Lemma 14 *Consider an input space X, a distribution P on X, a real-valued function χ defined on X^t, and any set Γ of permutations on $\{1,\ldots,t\}$.*

$$E_{P^t}(\chi) = E_{x\in P^t}\left(\frac{1}{|\Gamma|}\sum_{\sigma\in\Gamma}\chi(x^\sigma)\right).$$

Proof (of Theorem 13) For any function f in F,

$$E_{\langle P_i\rangle_{i=1}^t}\left(M_{Q,f}^t\right)$$

$$\begin{aligned} &= \int_{X^{t-k}}\int_{X^k} M_{Q,f}^t(x)\, dP_1(x_1)\ldots dP_t(x_t) \\ &\le \int_{X^{t-k}}\left(\int_{X^k} M_{Q,f}^t(x)\, dP_{t-(k-1)}^k(x_{t-(k-1)},\ldots,x_t)\right. \\ &\qquad \left. + \frac{k(k-1)}{2}\gamma\right) dP_1(x_1)\ldots dP_{t-k}(x_{t-k}) \\ &= \int_{X^{t-k}}\int_{X^k} M_{Q,f}^t(x)\, dP_{t-(k-1)}^k(x_{t-(k-1)},\ldots,x_t) \\ &\qquad dP_1(x_1)\ldots dP_{t-k}(x_{t-k}) + \frac{k(k-1)}{2}\gamma. \end{aligned}$$

Fix $x_1, x_2, \ldots, x_{t-k}$, and consider the inner integral,

$$\begin{aligned} I &= \int_{X^k} M_{Q,f}^t(x)\, dP_{t-(k-1)}^k(x_{t-(k-1)},\ldots,x_t) \\ &= \int_{X^k}\frac{1}{|\Gamma_{t,k}|}\sum_{\sigma\in\Gamma_{t,k}} M_{Q,f}^t(x^\sigma) \end{aligned}$$

$$
\begin{aligned}
& dP^k_{t-(k-1)}(x_{t-(k-1)}, \ldots, x_t) \\
\leq & \int_{X^k} \hat{\hat{M}}_{Q,F}(t,k)\, dP^k_{t-(k-1)}(x_{t-(k-1)}, \ldots, x_t) \\
= & \hat{\hat{M}}_{Q,F}(t,k).
\end{aligned}
$$

Therefore, for any function f in F,

$$
\begin{aligned}
\hat{M}_{Q,f,\gamma}(t) &= \sup_{\langle P_i \rangle} \left(E_{\langle P_i \rangle_{i=1}^t} \left(M^t_{Q,f} \right) \right) \\
&\leq \hat{\hat{M}}_{Q,F}(t,k) + \frac{k(k-1)}{2}\gamma,
\end{aligned}
$$

and so

$$
\begin{aligned}
\hat{M}_{Q,F,\gamma}(t) &= \sup_{f \in F} \left\{ \hat{M}_{Q,f,\gamma}(t) \right\} \\
&\leq \hat{\hat{M}}_{Q,F}(t,k) + \frac{k(k-1)}{2}\gamma,
\end{aligned}
$$

which is Inequality (8). □

In [HLW90], a general purpose deterministic prediction strategy, the one-inclusion graph prediction strategy, is described. Call this strategy Q_1. Using the same argument as the proof of Theorem 2.2 in [HLW90], the one-inclusion graph strategy for a function class F can make a total of no more than $2\,\mathrm{VCdim}(F)$ mistakes for all k permutations in $\Gamma_{t,k}$, so

$$
\hat{\hat{M}}_{Q_1,F}(t,k) \leq \frac{2\,\mathrm{VCdim}(F)}{k}.
$$

This result and Theorem 13 give the mistake bound

$$
\hat{M}_{Q_1,F,\gamma}(t) \leq \frac{2\,\mathrm{VCdim}(F)}{k} + \frac{k(k-1)\gamma}{2} \qquad (9)
$$

for $k = 1, 2, \ldots, t$. By choosing the value of k appropriately, we get the following bounds.

Theorem 15 *For any function class F with VC-dimension $1 \leq d < \infty$, there is a prediction strategy Q such that the mistake probability satisfies*

$$
\hat{M}_{Q,F,\gamma}(t) \leq \begin{cases} \dfrac{2d}{t} + \left(\dfrac{d^2\gamma}{2}\right)^{1/3} & t \leq \left(\dfrac{2d}{\gamma}\right)^{1/3} \\ 4(d^2\gamma)^{1/3} & t > \left(\dfrac{2d}{\gamma}\right)^{1/3} \end{cases}
$$

where $0 < \gamma \leq 1$. If $t > 5d/(2\epsilon)$ and $\gamma < \epsilon^3/(64d^2)$, then $\hat{M}_{Q,F,\gamma}(t) < \epsilon$ for this strategy.

Proof We will show that the statement is true for the one-inclusion graph strategy, Q_1. We have

$$
\hat{M}_{Q_1,F,\gamma}(t) \leq \frac{2d}{k} + \frac{k(k-1)}{2}\gamma
$$

for $k = 1, 2, \ldots, t$. The right-hand side of this inequality is less than the function $F(k)$, where

$$
F(k) = \frac{2d}{k} + k^2\frac{\gamma}{2}.
$$

Using elementary calculus, it can be shown that there is a $k < (2d/\gamma)^{1/3}$ such that $F(k) < 4(d^2\gamma)^{1/3}$. This shows that

$$
\hat{M}_{Q,F,\gamma}(t) < 4\left(d^2\gamma\right)^{1/3} \qquad (10)
$$

if $t > (2d/\gamma)^{1/3}$. If $t \leq (2d/\gamma)^{1/3}$, $k = t$ provides the best bound,

$$
\hat{M}_{Q,F,\gamma}(t) < 2d/t + \left(\frac{d^2\gamma}{2}\right)^{1/3}. \qquad (11)
$$

To verify the sufficient conditions for (ϵ, γ)-prediction, suppose $\gamma < \epsilon^3/(64d^2)$. Then

$$
4(d^2\gamma)^{1/3} < \epsilon, \qquad (12)
$$

and

$$
(d^2\gamma/2)^{1/3} < \epsilon/128^{1/3} < \epsilon/5. \qquad (13)
$$

If, in addition, $t > 5d/(2\epsilon)$,

$$
2d/t < 4\epsilon/5. \qquad (14)
$$

So, if the conditions on t and γ in the theorem are satisfied and Q is Q_1, the one-inclusion graph strategy, then either

- $t > (2d/\gamma)^{1/3}$, in which case Inequality (12) and Inquality (10) imply $\hat{M}_{Q,F,\gamma}(t) < \epsilon$, or
- $t \leq (2d/\gamma)^{1/3}$, in which case Inequalities (13) and (14), and Inequality (11) imply $\hat{M}_{Q,F,\gamma}(t) < \epsilon$,

which is the desired result. □

3.2 UPPER BOUNDS FOR CONSISTENT PREDICTION STRATEGIES

While the results in the previous section give general upper bounds on the mistake probability, the prediction strategy for which the bounds were derived (the one-inclusion graph strategy) may be inefficient because its computational complexity can grow as much as exponentially with the VC-dimension of the function class F [HLW90]. In this section, we consider consistent prediction strategies. These strategies make predictions using consistent hypotheses chosen from a particular hypothesis class H. A hypothesis h is consistent with labelled sample $((x_1, y_1), \ldots, (x_t, y_t)) \in S^t$ if $h(x_i) = y_i$ for $i = 1, 2, \ldots, t$. If a function class F is efficiently pac-learnable (that is, learnable in polynomial time),

then there is an efficient randomized consistent hypothesis finder (and hence an efficient randomized consistent prediction strategy) for F ([HKLW88], Theorem 4.1).

We use a bound on the probability that a consistent deterministic strategy makes a mistake on the last example.

Lemma 16 *If H is a set of functions from X to $\{0,1\}$, with $\mathrm{VCdim}(H) = d \geq 1$, P is any distribution on X, and Q is a consistent prediction strategy that uses H, then for any $0 < \gamma \leq 1$ and $k > d+1$,*

$$E_{P^k}\left(M^k_{Q,f}(x_1,\ldots,x_k)\right) \leq \frac{2(d+1)}{k-1}\log_2\frac{4e(k-1)}{d}.$$

□

This result appears in the proof of Theorem 4.1 in [HLW90]. It is based on Theorem A2.1 in [BEHW89] and Sauer's Lemma ([BEHW89], Proposition A2.1). Instead, we could use the corresponding exponential bound in Theorem 3.12 of [ABST90], since it has better constants. However this would complicate the statement and proof of the following theorem.

Theorem 17
For any hypothesis class H with $\mathrm{VCdim}(H) = d$ and $1 \leq d < \infty$, any consistent prediction strategy using H has prediction error satisfying

$$\hat{M}_{Q,F,\gamma}(t) < \begin{cases} \dfrac{4(d+1)}{t}\log_2\dfrac{8e}{(d^2\gamma)^{1/3}} + 2\left(d^2\gamma\right)^{1/3}, \\ \qquad\qquad \text{if } d+2 \leq t < 2\left(\dfrac{d}{\gamma}\right)^{1/3} \\ 19\left(d^2\gamma\right)^{1/3}\log_2\dfrac{16e}{(d^2\gamma)^{1/3}}, \\ \qquad\qquad \text{if } t \geq 2\left(\dfrac{d}{\gamma}\right)^{1/3} \end{cases}$$

where $0 < \gamma < 4/(d+1)^2$.

Proof Obviously, if Q uses a hypothesis that is consistent with all $t-1$ labelled examples, then that hypothesis is also consistent with the last k examples, where $k \leq t-1$. Using this fact, we will find bounds on the probability of a mistake by considering the last k examples in the sample.

If $\langle P_i\rangle_{i=1}^t$ is a γ-admissible distribution sequence, Lemma 12 implies that

$$\begin{aligned} &E_{\langle P_i\rangle_{i=1}^t}\left(M^t_{Q,f}(x_1,\ldots,x_t)\right) \\ &= \int_X \cdots \int_X M^t_{Q,f}(x_1,\ldots,x_t)\, dP_1(x_1)\ldots dP_t(x_t) \\ &\leq \int_X \cdots \int_X E_{P_t^k}\left(M^t_{Q,f}(x_1,\ldots,x_t)\right) \\ &\qquad dP_1(x_1)\ldots dP_{t-k+1}(x_{t-k+1}) + \frac{k(k-1)\gamma}{2}, \end{aligned}$$

for any $1 \leq k \leq t$. Now, for any k satisfying $d+1 < k \leq t$, and any $x_1,\ldots,x_{t-k+1}$, Lemma 16 implies that

$$E_{P_t^k}\left(M^t_{Q,f}(x_1,\ldots,x_t)\right) \leq \frac{2(d+1)}{k-1}\log_2\frac{4e(k-1)}{d},$$

so

$$E_{x\in\langle P_i\rangle_{i=1}^t}\left(M^t_{Q,f}(x)\right) \leq \frac{2(d+1)}{k-1}\log_2\frac{4e(k-1)}{d} + \frac{k(k-1)\gamma}{2},$$

for $k = d+2,\ldots,t$. It follows that

$$\hat{M}_{Q,F,\gamma}(t) \leq \frac{2(d+1)}{k-1}\log_2\frac{4e(k-1)}{d} + \frac{k(k-1)\gamma}{2}.$$

As in the proof of the mistake bounds for Q_1, we choose k to give the best bound. We have

$$\begin{aligned} \hat{M}_{Q,F,\gamma}(t) &\leq \frac{2(d+1)}{k-1}\log_2\frac{4e(k-1)}{d} + \frac{k(k-1)\gamma}{2} \\ &< \frac{4(d+1)}{k}\log_2\frac{4ek}{d} + \frac{k^2\gamma}{2} \qquad (15) \\ &< \left(\frac{4(d+1)}{k} + \frac{k^2\gamma}{2}\right)\log_2\frac{4ek}{d}, \end{aligned}$$

where the last two inequalities hold because $k > d/(2e)$. Using elementary calculus, it can be shown that

$$\hat{M}_{Q,F,\gamma}(t) < 19(d^2\gamma)^{1/3}\log_2\frac{16e}{(d^2\gamma)^{1/3}},$$

provided $t \geq 2(d/\gamma)^{1/3}$ and $\gamma(d+1)^2 < 4$. This gives the second bound in the theorem.

When $d+2 < t < 2(d/\gamma)^{1/3}$, Inequality (15) with $k = t$ gives

$$\hat{M}_{Q,F,\gamma}(t) < \frac{4(d+1)}{t}\log_2\frac{8e}{(d^2\gamma)^{1/3}} + 2(d^2\gamma)^{1/3},$$

which is the first bound in the theorem. □

Notice that this bound has the same shape as the upper bound for the one-inclusion graph strategy (Theorem 15), but with an additional $-\log(d^2\gamma)$ factor.

4 LOWER BOUNDS

To find a lower bound on the mistake probability for any prediction strategy, we construct a 'nasty' admissible distribution sequence.

Theorem 18 *For any function class F with VC-dimension d such that $3 \leq d < \infty$, and for any prediction strategy Q,*

$$\hat{M}_{Q,F,\gamma}(t) \geq \begin{cases} \dfrac{d-1}{2et} & \text{for all } t \\ \dfrac{\sqrt{\gamma(d-2)}}{4e} & t > \sqrt{\dfrac{d-2}{\gamma}} \end{cases}$$

No prediction strategy can (ϵ,γ)-predict F if $\gamma \geq 16e^2\epsilon^2/(d-2)$.

Proof The bound on $\hat{M}_{Q,F,\gamma}(t)$ for all t follows from the general lower bound for constant-distribution prediction ([HLW90], Theorem 3.1), since a constant distribution is always admissible.

The second part of the bound uses a similar proof. Consider the shattered set $X_0 = \{z, y_0, y_1, \ldots y_k\}$ with $d = k+2$ elements. We use a distribution sequence $\langle P_i \rangle_{i=1}^t$ which has a support that drifts from the set $\{y_0, z\}$ to $\{y_0, y_1, \ldots, y_k\}$. The probability of y_0 remains constant throughout; the remainder of the probability shifts from z to $\{y_1, \ldots, y_k\}$, starting at time $t-m$, where

$$m = \left\lceil \sqrt{k/\gamma} \right\rceil .$$

The distribution sequence is given by

$$P_j(z) = \begin{cases} \dfrac{k}{m} & j = 1, \ldots, t-m \\ \dfrac{(t-j)k}{m^2} & j = t-m+1, \ldots, t \end{cases}$$

$$P_j(y_0) = 1 - \frac{k}{m}$$

$$P_j(y_i) = \begin{cases} 0 & j = 1, \ldots, t-m \\ \dfrac{j-(t-m)}{m^2} & j = t-m+1, \ldots, t \end{cases}$$

This distribution sequence is γ-admissible, because the subset of X which experiences the largest increase in probability is the set $X_1 = \{y_1, \ldots, y_k\}$, and

$$\begin{aligned} P_{j+1}(X_1) - P_j(X_1) &= \begin{cases} 0 & j = 1, \ldots, t-m+1 \\ \dfrac{k}{m^2} & j = t-m, \ldots, t-1 \end{cases} \\ &\leq \frac{k}{m^2} \\ &< \gamma. \end{aligned}$$

Let B be the set of samples of length t in which the last example x has not already appeared in $(x_1, \ldots, x_{t-1})$. The probability that a sample is in B is

$$\begin{aligned} &\Pr_{\langle P_i \rangle}(B) \\ &\quad = \Pr_{\langle P_i \rangle}\left((x_1, \ldots, x_t) : x_t \neq x_j, j = 1, \ldots, t-1\right) \\ &\quad \geq \Pr_{\langle P_i \rangle}\left(x_t \neq y_0 \text{ and } x_t \neq x_j, j = 1, \ldots, t-1\right) \\ &\quad = \left(1 - P_t(y_0)\right) \prod_{j=1}^{t-1} \left(1 - P_j(x_t)\right) . \end{aligned}$$

Now, if $x_t \neq y_0$,

$$P_j(x_t) = \begin{cases} 0 & j = 1, \ldots, t-m \\ \dfrac{j-(t-m)}{m^2} & j = t-m+1, \ldots, t \end{cases}$$

So

$$\begin{aligned} \Pr_{\langle P_i \rangle}(B) &\geq \frac{k}{m} \prod_{j=t-m+1}^{t-1} \left(1 - \frac{j-(t-m)}{m^2}\right) \\ &= \frac{k}{m} \prod_{l=1}^{m-1} \left(1 - \frac{l}{m^2}\right) \\ &\geq \frac{k}{m} \prod_{l=1}^{m-1} \left(1 - \frac{m-1}{m^2}\right) \\ &> \frac{k}{m} \left(1 - \frac{1}{m}\right)^{m-1} \\ &> \frac{k}{em} \\ &\geq \frac{\sqrt{k\gamma}}{2e}. \end{aligned}$$

Now, using the same argument as in the proof of Theorem 3.1 in [HLW90] (picking a set of 2^d functions that shatters X_0 and finding the expected error under the uniform distribution on this set of functions), we can show that there is a function in F such that $E_{\langle P_i \rangle}(M^t_{Q,f}) \geq \Pr(B)/2 > \sqrt{k\gamma}/(4e)$.

Rearranging the bound for large t shows that

$$\gamma \geq \frac{16e^2\epsilon^2}{d-2}$$

implies that $\hat{M}_{Q,F,\gamma}(t) \geq \epsilon$. □

Notice that this necessary condition on γ for (ϵ,γ)-prediction is a factor of ϵ/d from the sufficient condition given in Theorem 15.

5 LEARNING CHANGING NOISY PROBLEMS

In the prediction model of learning (and the pac model), we assume that the relationship between examples and their labels is a deterministic function in a known function class. This is an optimistic assumption, since it forbids noise and errors, and it assumes a great deal of knowledge about the function. To dispense with these assumptions, Blumer *et al.* [BEHW89] proposed a learning model in which the relationship is described by a joint probability distribution on $X \times \{0,1\}$. In this section, we consider a learning model of this kind in which the joint distribution is allowed to change slowly but continually as learning proceeds. This is a more general problem than either learning with a slowly changing distribution of examples or learning with a slowly changing target function.

We begin with some notation.

Definition 19 *Let S be the space of labelled examples, $S = X \times \{0,1\}$. If $\xi = ((x_1,y_1),\ldots,(x_t,y_t)) \in S^t$ and h is a function from X to $\{0,1\}$, define the empirical error of h as*

$$\widehat{\mathbf{er}}_\xi(h) = \frac{1}{t}\left|\{i \in \{1,\ldots,t\} : h(x_i) \neq y_i\}\right|.$$

Define the expected error of h with respect to the distribution D on S as

$$\mathbf{er}_D(h) = D\left(\{(x,y) \in S : h(x) \neq y\}\right).$$

For the set H of functions from X to $\{0,1\}$, the distribution D on S, and parameters $0 < \beta \leq 1$, and $0 < \epsilon < 1$, define the set $B_D = B_D(H,t,\beta,\epsilon)$ of misleading labelled samples as

$$B_D = \{\xi \in S^t : \exists h \in H, \widehat{\mathbf{er}}_\xi(h) \leq (1-\beta)\epsilon \text{ and } \mathbf{er}_D(h) > \epsilon\}.$$

The following theorem gives conditions on γ and t that ensure that the empirical error for part of a labelled sample of length t is an accurate indication of the expected error for the next example, when the labelled examples are generated according to a γ-admissible distribution sequence.

Theorem 20 *Consider a hypothesis class H of functions from X to $\{0,1\}$ with $\mathrm{VCdim}(H) = d < \infty$, parameters $0 < \epsilon,\delta < 1$, $0 < \beta \leq 1$, and $0 \leq \gamma < 1$, and a γ-admissible distribution sequence $\langle P_i\rangle_{i=1}^t$ on $S = X \times \{0,1\}$. If*

$$\gamma \leq \frac{\delta}{(k+1)^2}$$

and $t \geq k$, then

$$\Pr_{\langle P_i\rangle_{i=t-k+1}^t}\left(B_{P_{t+1}}(H,k,\beta,\epsilon)\right) \leq \delta,$$

where

$$k = \left\lceil \frac{1}{\beta^2\epsilon(1-\sqrt{\epsilon})}\left(4\log\frac{8}{\delta} + 6d\log\frac{4}{\beta^{2/3}\epsilon}\right)\right\rceil.$$

and $B_{P_{t+1}}(H,k,\beta,\epsilon) \subseteq S^k$ is the set of misleading labelled samples defined in Definition 19.

We will use the following results.

Lemma 21 *Let $\langle P_i\rangle_{i=1}^t$ be a γ-admissible distribution sequence on $S = X \times \{0,1\}$, $0 \leq \gamma < 1$. If f is a $[0,1]$-valued, measurable function defined on $S^* = \cup_{m\in\mathbb{N}} S^m$,*

$$E_{P_i^n}(f) \leq E_{P_{i+1}^n}(f) + n\gamma, \tag{16}$$

for $n = 1,2,\ldots$ For an event $A \subseteq S$,

$$P_i^n(A) \leq P_{i+1}^n(A) + n\gamma, \tag{17}$$

for $n = 1,2,\ldots$

Proof We prove Inequality (16) by induction. By Lemma 3,

$$E_{P_i}(f) \leq E_{P_{i+1}}(f) + \gamma.$$

If Inequality (16) is true for n, we have

$$\begin{aligned} E_{P_i^{n+1}}(f) &= \int_S \int_{S^n} f\, dP_i^n(x_1,\ldots,x_n)\, dP_i(x_{n+1}) \\ &= \int_S E_{P_i^n}(f)\, dP_i(x_{n+1}) \\ &\leq \int_S E_{P_{i+1}^n}(f)\, dP_i(x_{n+1}) + n\gamma. \end{aligned}$$

The integrand is $[0,1]$-valued, so Lemma 3 gives

$$E_{P_i^{n+1}}(f) \leq E_{P_{i+1}^{n+1}}(f) + (n+1)\gamma.$$

Inequality (17) follows immediately, using $D(A) = E_D(1_A)$ for any distribution D, where 1_A is the indicator function for A ($1_A(x)$ is 1 when $x \in A$ and 0 otherwise). □

The following Lemma is due to Anthony and Shawe-Taylor ([AST90], Proposition 3.2). It improves on a similar result presented by Blumer *et al.* ([BEHW89], Theorem A3.1).

Lemma 22 *Define $B_D, H, t, \beta, \epsilon, d, \delta$ as in Theorem 20. For any distribution D on S,*

$$D^t\left(B_D(H,t,\beta,\epsilon)\right) \leq \delta$$

if

$$t \geq \frac{1}{\beta^2\epsilon(1-\sqrt{\epsilon})}\left(4\log\frac{4}{\delta} + 6d\log\frac{4}{\beta^{2/3}\epsilon}\right).$$

Proof (of Theorem 20) Let f in Lemma 3 be the indicator function of $B_{P_{j+1}}(H,j,\beta,\epsilon)$ for $j \in \mathbb{N}$. Then

$$\begin{aligned} &\Pr_{\langle P_i\rangle_{i=1}^j}\left(B_{P_{j+1}}(H,j,\beta,\epsilon)\right) \\ &\quad\leq P_j^j\left(B_{P_{j+1}}(H,j,\beta,\epsilon)\right) + \frac{j(j-1)}{2}\gamma \\ &\quad\leq P_{j+1}^j\left(B_{P_{j+1}}(H,j,\beta,\epsilon)\right) + \frac{j(j+1)}{2}\gamma, \end{aligned}$$

where the second inequality follows from Lemma 21. Now,

$$P_{j+1}^j\left(B_{P_{j+1}}(H,j,\beta,\epsilon)\right) \leq \frac{\delta}{2}$$

if $j = k$ (by Lemma 22), and

$$\frac{k(k+1)}{2}\gamma \leq \frac{\delta}{2}$$

if $\gamma \leq \delta/(k+1)^2$.

Since the labelled examples are independent, the probability of a labelled sample of length t in which the last k

elements possess some property is the same as the probability that a labelled sample of length k possesses the property. $\square$

This theorem suggests the following learning procedure: an algorithm considers the most recent $k(\beta, \epsilon, \delta, d)$ labelled examples, and attempts to find a hypothesis $h \in H$ that minimizes disagreements with the examples. With probability $1 - \delta$, the empirical error for that hypothesis will be an accurate estimate of its expected error. Notice that this algorithm does not need to know the bound γ on the change in the distribution. Of course, the choice of the parameters β, δ and ϵ imposes an upper bound on γ.

6 CONCLUSIONS

We have presented two models of learning from random examples that allow the distribution of the examples to change slowly but continually as learning proceeds. The first model assumes that there is a target function that defines the label of each example. If γ is the amount by which the distribution of examples is allowed to drift and d is the VC-dimension of the target function class, we showed that an upper bound on the probability that a prediction strategy misclassifies the last example in a sequence of t examples decreases as d/t at first (as in the constant-distribution case), but that this probability can reach a steady-state value between $\Omega(d^{1/2}\gamma^{1/2})$ and $O(d^{2/3}\gamma^{1/3})$. Using these bounds, we gave necessary and sufficient conditions that (ϵ, γ)-prediction is possible ($\gamma = O(\epsilon^2/d)$ and $\gamma = O(\epsilon^3/d^2)$, respectively). Obviously, it would be desirable to remove the ϵ/d factor separating these bounds.

Section 5 investigated the problem of learning when the labelled examples are generated by a slowly changing joint distribution on $X \times \{0, 1\}$. We gave an upper bound on γ that ensures that the empirical error of a hypothesis is close to its expected error (provided there are enough training examples). Since the most recent examples contain the most relevant information (and the earliest examples might be misleading), it may be possible to improve on this result by using a weighting scheme (see [HL91]), in which a hypothesis that is consistent with most of the recent examples would be rated more highly than one that is consistent with earlier examples.

Acknowledgements

This research was supported by OTC Australia, by the Australian Telecommunications and Electronics Research Board, and through an Australian Postgraduate Research Award. I thank D. Lovell and R. Williamson for helpful comments, and a reviewer for suggesting alternative definitions of distance between distributions.

References

[ABST90] M. Anthony, N. Biggs, and J. Shawe-Taylor. Learnability and formal concept analysis. Technical Report CSD-TR-624, UCL, 1990.

[AST90] M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. Technical Report CSD-TR-628, UCL, 1990.

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

[Hal50] P. R. Halmos. *Measure Theory*. Van Nostrand, 1950.

[HKLW88] D. Haussler, M. Kearns, N. Littlestone, and M. K. Warmuth. Equivalence of models for polynomial learnability. In *Proceedings of the 1988 Workshop on Computational Learning Theory*, pages 42–55. Morgan Kaufmann, San Mateo, CA, 1988.

[HL91] D. P. Helmbold and P. M. Long. Tracking drifting concepts using random examples. In *Proceedings of the Fourth Annual Workshop on Computational Learning Theory*, pages 13–23. Morgan Kaufmann, San Mateo, CA, 1991.

[HLW90] D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. Technical Report UCSC CRL-90-54, Baskin Center for Computer Engineering and Information Sciences, University of California Santa Cruz, 1990.

[Kra88] A. H. Kramer. Learning despite distribution drift. In *Proceedings of the Connectionist Models Summer School*, pages 201–210. Morgan Kaufmann, San Mateo, CA, 1988.

[Kul67] S. Kullback. A lower bound for discrimination information in terms of variation. *IEEE Transactions on Information Theory*, IT-13:126–127, 1967.

[Ren61] A. Renyi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561. University of California Press, 1961.

[Val84] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1143, 1984.

[Vap82] V. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, 1982.

[VC71] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, XVI(2):264–280, 1971.