Machine Learning, 41, 153–174, 2000

Learning Changing Concepts by Exploiting the Structure of Change*

PETER L. BARTLETT Peter.bartlett@anu.edu.au
Research School of Information Sciences and Engineering, Australian National University, Canberra, ACT 0200, Australia

SHAI BEN-DAVID shai@cs.technion.ac.il
Department of Computer Science, Technion, Haifa 32000, Israel

SANJEEV R. KULKARNI kulkarni@ee.princeton.edu
Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA

Editor: Lisa Hellerstein

Abstract. This paper examines learning problems in which the target function is allowed to change. The learner sees a sequence of random examples, labelled according to a sequence of functions, and must provide an accurate estimate of the target function sequence. We consider a variety of restrictions on how the target function is allowed to change, including infrequent but arbitrary changes, sequences that correspond to slow walks on a graph whose nodes are functions, and changes that are small on average, as measured by the probability of disagreements between consecutive functions. We first study estimation, in which the learner sees a batch of examples and is then required to give an accurate estimate of the function sequence. Our results provide bounds on the sample complexity and allowable drift rate for these problems. We also study prediction, in which the learner must produce online a hypothesis after each labelled example and the average misclassification probability over this hypothesis sequence should be small. Using a deterministic analysis in a general metric space setting, we provide a technique for constructing a successful prediction algorithm, given a successful estimation algorithm. This leads to sample complexity and drift rate bounds for the prediction of changing concepts.

Keywords: computational learning theory, sample complexity, concept drift, estimation, tracking

1. Introduction

We consider the problem of learning to track a changing subset of a domain from random examples. In many learning problems for which the environment is changing, there is some structure to the change. For example, the daily weather at a given location may be viewed as a changing concept having some basic underlying structure to its change. On a short-term scale changes are of bounded variation, on a larger annual scale changes are roughly cyclic, and there are probably some further subtle rules governing the structure of daily weather changes. Another rather practical example arises in a steel rolling mill, where the efficiency of the mill's operation depends on how accurately the behavior of the rolling surfaces can

*An earlier version of this paper was presented at the Ninth Annual Conference on Computational Learning Theory.

be predicted (Connolly, Chicharo, & Wilbers, 1992). As in many industrial processes, there is an accurate physical model of the target function (relating the measured variables to the desired quantity), but there are several unknown parameters, and these may change over time. The change may be slow (as the rollers wear), or occasionally fast (when something fails). Again, there is a definite structure to the change. Yet another scenario that can be viewed similarly arises in computer vision when one wishes to identify an object from a sequence of photographs taken while the object or the camera are in motion. In this paper we address the question, when can we exploit the structure of change to learn a changing concept?

More formally, we assume that the learner sees, at time t, a random example x_t from some domain X, together with the value of an unknown target function $f_t : X \to \{0, 1\}$ at the point x_t. The function is an element of a known class F. The distribution that generates the examples is assumed to remain fixed, but the function can change between examples, with some structure to the change. We formalize the structure by defining a set of legal function sequences. For instance, cyclic or seasonal changes correspond to a walk on a directed cyclic graph. In the rolling mill example, the legal sequences might be those corresponding to smooth paths in parameter space.

We consider the following two problems of learning in a changing environment.

Estimation When can one estimate a sequence of concepts $(f_1, f_2, \ldots, f_n)$ on the basis of a set of random samples $(x_1, f_1(x_1)), \ldots, (x_n, f_n(x_n))$? This may be thought of as 'understanding the past on the basis of gathered experience.'

Prediction When can one predict the next concept in a sequence of concepts $(f_1, f_2, \ldots, f_n)$, on the basis of random samples of previous concepts? This may be thought of as 'predicting the future on the basis of past experience.'

Note that in the usual PAC model these two issues coincide. In that model, there is only one target concept per learning session (it remains fixed throughout the learning process).

The problem of predicting labels for a changing concept has been considered elsewhere. Helmbold and Long (1994) consider prediction when the concept is allowed to drift slowly between trials. That is, any two consecutive functions f_i and f_{i+1} must have $\Pr(f_i \neq f_{i+1})$ small. This is a natural measure of concept drift, since it can be thought of as the weakest assumption that implies the labels of random examples will not vary much. Their work is in a slightly different setting—they consider prediction strategies that aim to minimize the probability over a long sequence of misclassifying the last example—but the results can be easily converted between settings. The bound on allowable drift that they obtained, $\Omega(\epsilon^2/(d \log(1/\epsilon)))$ (where ϵ is the allowable prediction error and d is the VC-dimension of the class), was subsequently improved by Bartlett and Helmbold (1996) to $\Omega(\epsilon^2/(d + \log(1/\epsilon)))$. This result is also a special case of a later result due to Barve and Long (1997). It has recently been improved by Long (1998) to $\Omega(\epsilon^2/d)$. These papers impose a uniform bound on the probability of disagreement between consecutive functions. One of the function sequence classes considered here contains sequences that satisfy a weaker (time-averaged) version of this bound. The final result of this paper, when converted to a setting analogous to that of the earlier work described above, shows that with this weaker constraint and the correspondingly weaker accuracy criterion, the allowable drift rate decreases by no more than log factors, ϵ^2/d versus $\epsilon^2/(d \log^2(d/\epsilon))$.

Several authors (Bartlett, 1992; Bartlett & Helmbold, 1996; Barve & Long, 1997; Long, 1998) have considered learning problems in which a changing environment is modelled by a slowly changing distribution on the product space $X \times \{0, 1\}$. The allowable drift is restricted by ensuring that consecutive probability distributions are close in total variation distance. Clearly, allowing a changing concept with a bound on the probability of disagreement between consecutive functions is a special case of this model. More recently, Freund and Mansour (1997) have investigated learning when the distribution changes as a linear function of time. They present algorithms that estimate the error of functions, using knowledge of this linear drift.

Blum and Chalasani (1992) consider learning switching concepts. The target concept is allowed to switch between concepts in the class, but with some constraint on the total number of concepts visited, or on the frequency of switches. Their most closely related results concentrate on the computational complexity of predicting switching concepts from particular concept classes. In contrast, we give a general condition on switching frequency for which the estimation and prediction of switching concepts from any class with finite VC-dimension is possible, ignoring computational constraints.

The next section introduces the notation. Section 3 considers the problem of sequence estimation (from a batch of labelled samples). Our main result here is the derivation of a sufficient condition that guarantees estimability of a family of sequences of functions. This result may be viewed as an extension of the basic (Blumer et al., 1989) sufficiency theorem for PAC learnability of classes of (single) functions. We go on and apply this result to provide sample size upper bounds for the estimation of several naturally arising families of function sequences. In Section 4 we discuss the function prediction problem. We show the success of certain kth-order Markovian prediction strategies in the setting of function prediction. We deviate from previous work on prediction via Markovian strategies in that, rather than assuming access to the complete last-k-steps information (and then looking for the best Markovian strategy, or the best one in some computationally restricted family of strategies, as in the work of Merhav and Feder (1993)), we assume that our predictor can only *approximate* the past sequence, $(f_{t-k}, \ldots, f_{t-1})$. We conclude the paper by gluing together our estimation and prediction results to obtain sample size upper bounds for prediction of changing concepts under several types of change constraints.

2. Basic notation

Throughout, we let X be a set, and we consider classes F of $\{0, 1\}$-valued functions defined on X. We fix some σ-algebra of subsets of X and consider probability distributions over X that are defined over this algebra. Furthermore, we shall assume that all functions we consider are measurable with respect to this algebra of sets. This is true if X is countable; for uncountable X, Blumer et al. (1989) give mild conditions on F that suffice. The growth function of F, $\Pi_F : \mathbb{N} \to \mathbb{N}$, is defined as $\Pi_F(n) = \max\{|F_{|x}| : x \in X^n\}$, where

$$F_{|x} = \{(f(x_1), \ldots, f(x_n)) : f \in F\}.$$

The Vapnik-Chervonenkis dimension of F is defined as

$$\mathrm{VCdim}(F) = \max\{n : \Pi_F(n) = 2^n\}.$$

For a sequence of functions, $\bar{f} = (f_1, \ldots, f_n)$ and a sequence of points $x = (x_1, \ldots, x_n) \in X^n$, let $\bar{f}(x)$ denote the sequence $(f_1(x_1), \ldots, f_n(x_n))$. For a set of sequences of functions, $F_n \subseteq F^n$, we denote $\{\bar{f}(x) : \bar{f} \in F_n\}$ by $F_{n|_x}$.

For a probability distribution P on X, define the P-induced pseudometric d_P over a class of functions F by $d_P(f, g) = P\{x : f(x) \neq g(x)\}$ (for $f, g \in F$). For sequences $\bar{f} = (f_1, \ldots, f_n)$ and $\bar{g} = (g_1, \ldots, g_n)$ in F^n, extend d_P to the pseudometric $\bar{d}_P$ on F^n by

$$\bar{d}_P(\bar{f}, \bar{g}) = \frac{1}{n} \sum_{i=1}^{n} d_P(f_i, g_i),$$

and for a sequence x in X^n define

$$\hat{d}_x(\bar{f}, \bar{g}) = \frac{1}{n} |\{i : f_i(x_i) \neq g_i(x_i)\}|.$$

We shall consider a variety of constraints on the function sequences, that restrict how much the functions can fluctuate over time. Given a (pseudo)metric d over a class F of functions, a natural measure of these fluctuations is the average distance, in the metric d, between subsequent functions,

$$V_d(\bar{f}) \stackrel{\text{def}}{=} \frac{1}{|\bar{f}| - 1} \sum_{i=1}^{|\bar{f}|-1} d(f_i, f_{i+1}). \tag{1}$$

(where $|\bar{f}|$ stands for the length of the sequence $\bar{f}$).

Note that, for the discrete metric D over F (for which $D(f, g)$ takes value 0 if $f = g$ and 1 otherwise), one gets

$$V_D(\bar{f}) \stackrel{\text{def}}{=} \frac{1}{|\bar{f}| - 1} |\{1 \leq i < |\bar{f}| : f_i \neq f_{i+1}\}| \tag{2}$$

We shall refer to $V_{d_P}(\bar{f})$ as $V_P(\bar{f})$ and to $V_D(\bar{f})$ as $D(\bar{f})$. Note also that, for every sequence $\bar{f}$ and for any probability distribution P over X, $V_P(\bar{f}) \leq D(\bar{f})$.

In the definition of V_d, it is not essential that d be a metric. In particular, it need not be a symmetric function. Consider a directed graph whose nodes are members of the function class F. Such a graph may be used to model a scenario in which, if a system is in some state $f \in F$ at one moment, it may switch at the next moment to a state in a restricted subset of F. In such a case we shall define a digraph by having the edges reflect this 'possible next state' relation. Given a directed graph G over F, let d_G be the 'shortest path' function, so that $d_G(f, g)$ is the length of the shortest path from f to g in G. We shall refer to V_{d_G} as V_G.

Let log denote logarithm to base 2 and ln denote the natural logarithm.

3. Sequence estimation

Definition 1. For $n \in \mathbb{N}$ and any distribution P on X, let $F_n(P) \subseteq F^n$ be a set of function sequences of length n. We call these *legal sequences*.

- An *estimation algorithm* A is a function that maps sequences from $X \times \{0, 1\}$ to sequences of functions from F.
- For $0 < \epsilon, \delta < 1$, we say that A (ϵ, δ)-*estimates* F_n on n examples if, for all distributions P on X, for all $\bar{f} \in F_n(P)$, the probability over $x \in X^n$ that $\bar{d}_P(A(x, \bar{f}(x)), \bar{f}) \geq \epsilon$ is less than δ.

We remark that this definition could be extended to *randomized* estimation algorithms, which return a probability distribution over function sequences.

For brevity, we often write F_n in place of $F_n(P)$. We first consider consistent algorithms, that is, algorithms that choose a function sequence from F_n that agrees with the target sequence on all of the examples. The following theorem gives a uniform convergence result for classes of function sequences. It implies a sufficient condition for a consistent algorithm to be able to estimate F_n. The proof is in Appendix A.

Theorem 2. *For all* $0 < \epsilon < 1$, $n \geq 6/\epsilon$, *and* $\bar{f} \in F_n$, *and for all distributions* P *on* X,

$$P^n \{x \in X^n : \exists \bar{g} \in F_n \text{ s.t. } g_i(x_i) = f_i(x_i) \text{ for all } i, \text{ and } \bar{d}_P(\bar{f}, \bar{g}) \geq \epsilon\} < 2^{-n\epsilon/2+1} E|F_{n|_x}|^2,$$

where the expectation is over x *in* X^n.

In fact, the proof of Theorem 2 does not make use of the fact that the target sequence $\bar{f}$ was in the set F_n of legal sequences. This observation by itself is not useful for learning, since we cannot be sure that there will be a function sequence in the class F_n that is consistent with an arbitrary target sequence. However, we can use a similar argument (together with techniques of Haussler (1992)) to prove the following more general uniform convergence result, which is useful for learning when the target sequence $\bar{f}$ is arbitrary. The proof is in Appendix B.

Theorem 3. *For* $a, b \geq 0$, *define*

$$d_\gamma(a, b) = \frac{|a - b|}{a + b + \gamma}.$$

For all $0 < \alpha, \gamma < 1$, $n \geq 5/(\alpha^2\gamma)$, *all sequences* $\bar{f}$ *of measurable functions, and all distributions* P *on* X,

$$P^n\{x : \exists \bar{g} \in F_n \text{ s.t. } d_\gamma(\hat{d}_x(\bar{f}, \bar{g}), \bar{d}_P(\bar{f}, \bar{g})) \geq \alpha\} < 4E|F_{n|_x}|^2 \exp\left(-\frac{n\gamma\alpha^2}{3}\right),$$

where the expectation is over x *in* X^n.

A slightly weaker version of Theorem 2 follows easily from this result. Choosing appropriate values for α and γ gives the following corollary.

Corollary 4. *Suppose $0 < \epsilon < 1$, $\bar{f}$ is a sequence of measurable functions defined on X, and P is a distribution on X.*
1. *If $n \geq 20/\epsilon$ then*

$$P^n\left\{x : \exists \bar{g} \in F_n \text{ s.t. } \hat{d}_x(\bar{f}, \bar{g}) < \frac{1}{3}\bar{d}_P(\bar{f}, \bar{g}) - \epsilon \text{ or } \hat{d}_x(\bar{f}, \bar{g}) > 3\bar{d}_P(\bar{f}, \bar{g}) + \epsilon\right\}$$
$$< 4E\left|F_{n|_x}\right|^2 \exp\left(-\frac{n\epsilon}{12}\right).$$

2. *If $n \geq 45/\epsilon^2$ then*

$$P^n\{x : \exists \bar{g} \in F_n \text{ s.t. } |\hat{d}_x(\bar{f}, \bar{g}) - \bar{d}_P(\bar{f}, \bar{g})| \geq \epsilon\} < 4E\left|F_{n|_x}\right|^2 \exp\left(-\frac{n\epsilon^2}{27}\right).$$

In both inequalities, the expectation is over x in X^n.

Proof: The first part follows from Theorem 3 with $\alpha = 1/2$ and $\gamma = \epsilon$. For the second part, use $\alpha = \epsilon/3$ and $\gamma = 1$, and notice that $\bar{d}_P(\bar{f}, \bar{g}) \leq 1$ and $\hat{d}_x(\bar{f}, \bar{g}) \leq 1$. □

It is not hard to see that for sets of sequences F_n whose definition does not depend on the underlying distribution P, the slow growth of $E|F_{n|_x}|$ for all distributions is also necessary for uniform convergence. More precisely, if for some $a > 1$, $E|F_{n|_x}| = \Omega(a^n)$, then there exist distributions P relative to which the empirical weighted differences between sequences in F_n do not converge uniformly to $\bar{d}_P$. However the following example shows that there are distributions P and distribution-dependent families of sequences F_n which exhibit uniform convergence as above, despite the exponential growth rate of $E|F_{n|_x}|$. Let F_n be any family of sequences such that for all sequences $\bar{f} \in F_n$, $P(f_i(x) \neq f_{i+1}(x)) = 0$ for all i. Clearly, if the VC-dimension of F is finite, the components of any sequence in F_n are all equal except on a set of measure zero, and so uniform convergence follows from standard results. But if, for instance, $X = [0, 1]$, P is the uniform distribution on X, and F is the set of indicator functions of singletons $\{x\}$, then $E|F_{n|_x}| = 2^n$. (A similar comment was made by Bertoni et al. (1992) in the context of the uniform convergence of relative frequencies of sets to their probabilities.)

Clearly, Theorem 2 implies that if the cardinality of the set F_n is uniformly bounded for all n, then F_n can be estimated. As another example, consider a function class F and suppose that G is a directed graph with nodes in F. For $n \in \mathbb{N}$, let F_n be the set of walks of length n on G. If F has finite VC-dimension and for all nodes f in F the number of walks of length n on G starting at f grows polynomially in n, then F_n can be (ϵ, δ)-estimated for sufficiently large n. By restricting the set of legal sequences to sequences defined by slow walks on the graph, the restriction on the underlying graph may be relaxed.

Theorem 5. *Let $G = (F, E)$ be a directed graph, where F is a function class. For $n \in \mathbb{N}$ and $0 < \Delta < 1/2$, let $F_n(G, \Delta)$ be the set of sequences $\bar{f} = (f_1, \ldots, f_n)$ whose d_G variation is bounded by Δ, i.e., $F_n(G, \Delta) = \{\bar{f} \in F^n : V_G(\bar{f}) \leq \Delta\}$.*

1. *Let p_k denote the number of paths in G of length no more than k. Suppose that, for some $\eta > 0$, $p_k < 2^{\eta k}$ for all $k \geq 1$. Then for any $0 < \delta < 1$, if*

$$\epsilon > 8\Delta\left(\eta + \log\frac{2e}{\Delta}\right)$$

and

$$n \geq \frac{6}{\epsilon}\log\frac{2}{\delta},$$

it follows that any consistent algorithm will (ϵ, δ)-estimate $F_n(G, \Delta)$ (on n examples).

2. *Let O be the outdegree of G, and let d be the VC-dimension of F. Then for $0 < \epsilon, \delta < 1$, if*

$$\epsilon > 8\Delta\log\frac{2eO}{\Delta},$$

and

$$n > \frac{8}{\epsilon}\left(2d\log\frac{16}{\epsilon\ln 2} + \log\frac{2}{\delta}\right),$$

then any consistent algorithm will (ϵ, δ)-estimate $F_n(G, \Delta)$ from n examples.

Proof: We assume that $\lfloor\Delta(n-1)\rfloor \geq 1$; otherwise the result follows trivially from results for learning (constant) functions from F (see, for example, (Blumer et al., 1989; Vapnik, 1982)).

1. Because of the constraint on the number of paths in G,

$$|F_n| < \binom{n-1}{\lfloor\Delta(n-1)\rfloor}2^{\eta\lfloor\Delta(n-1)\rfloor} \leq \left(\frac{e(n-1)}{\lfloor\Delta(n-1)\rfloor}\right)^{\lfloor\Delta(n-1)\rfloor}2^{\eta\lfloor\Delta(n-1)\rfloor}$$
$$\leq \left(\frac{e2^{\eta+1}}{\Delta}\right)^{\Delta(n-1)}.$$

It follows from Theorem 2 that $F_n(G, \Delta)$ can be (ϵ, δ)-estimated by any consistent algorithm whenever

$$\frac{n\epsilon}{2} - 2\Delta n\left(\log\left(\frac{2e}{\Delta}\right) + \eta\right) > \log\frac{2}{\delta}, \tag{3}$$

provided $n \geq 6/\epsilon$. The condition on ϵ in the theorem statement implies that the second term on the left hand side is less than $\epsilon n/4$. The condition on n then implies (3).

2. Since the outdegree of G is bounded by O, the number of function sequences of length k starting from a given function is no more than O^{k-1}. It follows that, for any x in X^n

we have

$$|F_{n|_x}| \leq \sum_{i=0}^{\lfloor \Delta(n-1) \rfloor} \binom{n-1}{i} O^{\lfloor \Delta(n-1) \rfloor} \left(\frac{en}{d}\right)^d \leq \left(\frac{2eO}{\Delta}\right)^{\Delta n} \left(\frac{en}{d}\right)^d.$$

Theorem 2 implies that $F_n(G, \Delta)$ can be (ϵ, δ)-estimated by any consistent algorithm whenever

$$\frac{n\epsilon}{2} - 2\Delta n \log \frac{2eO}{\Delta} - 2d \log \frac{en}{d} > \log \frac{2}{\delta}.$$

So if

$$\epsilon > 8\Delta \log \frac{2eO}{\Delta},$$

it suffices that $n > (4/\epsilon)(2d \log(en/d) + \log(2/\delta))$. Using the fact that $\ln(an) + 1 \leq an$ for all $a > 0$, we have that $(8d/\epsilon) \log n \leq n/2 + (8d/\epsilon) \log(16d/(\epsilon e \ln 2))$. Thus, it suffices that

$$n > \frac{8}{\epsilon}\left(2d \log \frac{16}{\epsilon \ln 2} + \log \frac{2}{\delta}\right). \qquad \square$$

The first part of the next theorem gives a similar result that applies when the functions can change arbitrarily over some class F, but only occasionally. The lower bound on ϵ (the guaranteed proximity of the estimation to the target sequence) now depends on the VC-dimension of F, whereas when graph constraints apply the out-degree of the graph imposed a fixed bound for all (finite VC-dimension) classes F. The second part of the theorem shows that, if there is a slowly changing function sequence that is close to the target $\bar{f}$ with respect to $\bar{d}_P$, an algorithm that chooses any slowly changing sequence with minimal error on the examples will find a good approximation to the target.

Theorem 6. *For a set F of $\{0, 1\}$-valued functions defined on X and $\Delta > 0$, define the sets of Δ-frequently switching function sequences of F as*

$$F_n(D, \Delta) = \{\bar{f} \in F^n : D(\bar{f}) \leq \Delta\},$$

where $D(\bar{f})$ is the average variation of a sequence $\bar{f}$, as defined by Equation (2) above. Suppose that $\mathrm{VCdim}(F) = d \geq 2$.

1. *There are constants c_1 and c_2 such that, for $0 < \epsilon, \delta, \Delta < 1$, if*

$$\epsilon > c_1 \Delta d \log \frac{1}{\Delta}$$

(for which $\Delta < c_1'\epsilon/(d\log(d/\epsilon))$ *suffices) and*

$$n \geq \frac{c_2}{\epsilon}\left(d\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right),$$

then any consistent algorithm will (ϵ, δ)*-estimate* $F_n(D, \Delta)$ *(on* n *examples).*

2. *There are constants* c_1 *and* c_2 *such that, for* $0 < \epsilon, \delta, \Delta < 1$ *and any sequence* $\bar{f}$ *of measurable* $\{0, 1\}$*-valued functions, if*

$$\epsilon > c_1\Delta d\log\frac{1}{\Delta}$$

(for which $\Delta < c_1\epsilon/(d\log(d/\epsilon))$ *suffices) and*

$$n \geq \frac{c_2}{\epsilon}\left(d\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right),$$

then

$$P^n\{x : \exists \bar{g} \in F_n(D, \Delta) \text{ s.t. } \hat{d}_x(\bar{f}, \bar{g}) < \epsilon/6 \text{ and } \bar{d}_P(\bar{f}, \bar{g}) \geq \epsilon \text{ or } \hat{d}_x(\bar{f}, \bar{g}) > 4\epsilon \text{ and } \bar{d}_P(\bar{f}, \bar{g}) \leq \epsilon\} < \delta.$$

Proof: In order to apply Theorem 2 and Corollary 4, we begin by establishing a bound on $E|F_{n|_x}|$. Fix $x \in X^n$. Notice that it does not suffice to argue that

$$|F_{n|_x}| \leq \sum_{i=0}^{\lfloor\Delta(n-1)\rfloor}\binom{n-1}{i}(\Pi_F(n))^{(i+1)},$$

since this grows too quickly with n.

Let $k = \lfloor\Delta(n-1)\rfloor$ and assume that $k \geq 1$. (Otherwise, the result follows from standard results; see, for example, (Blumer et al., 1989; Vapnik, 1982).) Clearly,

$$|F_{n|_x}| \leq \sum\prod_{j=0}^{k}\Pi_F(i_j), \tag{4}$$

where the sum is over all $1 \leq i_j \leq n$ satisfying

$$\sum_{j=0}^{k} i_j = n.$$

Sauer's lemma (see, for example, [15]) implies $\Pi_F(i) \leq 2i^d$. For each term in the sum (4), taking logs gives

$$\log\left(\prod_{j=0}^{k}(2i_j^d)\right) = (k+1) + d\sum_{j=0}^{k}\log i_j$$

$$\leq k+1+d(k+1)\log\left(\sum_{j=0}^{k}\left(\frac{i_j}{k+1}\right)\right)$$

$$= k+1+d(k+1)\log\left(\frac{n}{k+1}\right)$$

$$= \log\left(2\left(\frac{n}{k+1}\right)^d\right)^{k+1},$$

where the inequality follows from Jensen's inequality. It follows that

$$|F_{n|_x}| \leq \sum_{i=0}^{k}\binom{n-1}{i}\left(2\left(\frac{n}{k+1}\right)^d\right)^{k+1} \leq \left(\frac{e(n-1)}{k}\right)^k\left(2\left(\frac{n}{k-1}\right)^d\right)^{k+1}.$$

Theorem 2 implies that any consistent algorithm can (ϵ, δ)-estimate provided that $n \geq 6/\epsilon$ and

$$2^{-n\epsilon/2+1}\left(2e\left(\frac{n}{k+1}\right)^{d+1}\right)^{2(k+1)} < \delta.$$

So it suffices if

$$\frac{n\epsilon}{2} - 2(k+1)\log(2e) - 2(k+1)(d+1)\log\left(\frac{n}{k+1}\right) \geq \log\frac{2}{\delta},$$

which is implied by

$$\frac{n\epsilon}{2} - 2(k+1)(d+1)\log\left(\frac{2en}{k+1}\right) \geq \log\frac{2}{\delta}.$$

Recalling that $k = \lfloor \Delta(n-1) \rfloor$, if

$$\epsilon > 16\Delta(d+1)\log\left(\frac{3e}{\Delta}\right),$$

then the second term in the left hand side is no more than $n\epsilon/4$, so

$$n = \frac{4}{\epsilon}\log\left(\frac{2}{\delta}\right)$$

examples will suffice.

For the second part of the theorem, if $n \geq 20/\epsilon$, Corollary 4 implies that the desired probability is less than

$$4\left(2e\left(\frac{n}{k+1}\right)^{d+1}\right)^{2(k+1)}\exp\left(\frac{-n\epsilon}{72}\right),$$

which is no more than δ when

$$\frac{n\epsilon}{72} - 2(k+1)\ln(2e) - 2(k+1)(d+1)\ln\left(\frac{n}{k+1}\right) \geq \ln\frac{4}{\delta}.$$

This is implied by

$$\frac{n\epsilon}{72} - 2(k+1)(d+1)\ln\left(\frac{2en}{k+1}\right) \geq \ln\left(\frac{4}{\delta}\right).$$

Reasoning as above, it suffices if

$$\frac{\epsilon}{144} \geq 4\Delta(d+1)\ln\left(\frac{3e}{\Delta}\right)$$

and

$$n \geq \frac{144}{\epsilon}\ln\frac{4}{\delta}. \qquad \square$$

Notice that if the class F_n of legal sequences depends on the distribution P, then it is not clear how to construct a consistent algorithm or an algorithm that minimizes error over F_n, since the algorithm does not have access to P. In the following result, we avoid this problem by relating such a sequence class to one that does not depend on the distribution.

Theorem 7. *Suppose F is a set of $\{0, 1\}$-valued functions defined on X with* $\mathrm{VC\ dim}(F) = d$, *P is a probability measure on X, and $\Delta > 0$. Define the set of (P, Δ)-slowly changing sequences in F^n as*

$$F_n(P, \Delta) = \{\bar{f} \in F^n : V_P(\bar{f}) \leq \Delta\}.$$

There are constants c_1 and c_2 such that, for any $0 < \epsilon, \delta < 1$, any

$$\Delta < c_1\epsilon^2 \Big/ \left(d\log\left(\frac{d}{\epsilon}\right)\right),$$

and any

$$n \geq \frac{c_2}{\epsilon}\left(d\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right),$$

there exists an algorithm that (ϵ, δ)-estimates any sequence in $F_n(P, \Delta)$.

Proof: We consider an estimation algorithm that works with the class $F_n(D, \Delta')$ of slowly switching function sequences from F^n, for some appropriate value of Δ'. The algorithm chooses a sequence from that class with minimal error on the training examples. We show that this class (with $\Delta' = 24\Delta/\epsilon$) approximates $F_n(P, \Delta)$, and that this implies the algorithm is successful.

Consider a target sequence $\bar{f} \in F_n(P, \Delta)$. We shall construct a piecewise constant sequence $\bar{g}$ that approximates $\bar{f}$ with respect to $\bar{d}_P$, and does not have too many switches. Let $g_1 = f_1$, and let $i_1 > 1$ denote the first index such that $d_P(f_1, f_{i_1}) \geq \epsilon/24$. Set $g_i = g_1$ for $1 \leq i \leq i_1 - 1$. Let $g_{i_1} = f_{i_1}$, and let $i_2 > i_1$ be the smallest index such that $d_P(f_{i_1}, f_{i_2}) \geq \epsilon/24$. Then set $g_i = f_{i_1}$ for $i_1 \leq i \leq i_2 - 1$. Continue in this manner to form the piecewise constant sequence $\bar{g}$. Note that by the construction of $\bar{g}$ and the triangle inequality, if $D(\bar{g}) = \alpha$ then $V_P(\bar{f}) \geq \alpha\epsilon/24$. Therefore, if $\bar{f} \in F_n(P, \Delta)$ then $D(\bar{g}) < 24\Delta/\epsilon$. Also note that by construction $\bar{d}_P(\bar{f}, \bar{g}) < \epsilon/24$.

Thus, for any target sequence $\bar{f} \in F_n(P, \Delta)$, there is a sequence $\bar{g}$ in the class $F_n(D, 24\Delta/\epsilon)$ with $\bar{d}_P(\bar{g}, \bar{f}) \leq \epsilon/24$. Let $\Delta' = 24\Delta/\epsilon$. Theorem 6 implies that, if

$$\epsilon > c_1 \Delta' d \log \frac{1}{\Delta'}$$

and

$$n \geq \frac{c_2}{\epsilon}\left(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right),$$

then with probability $1 - \delta$ we have both $\hat{d}_x(\bar{g}, \bar{f}) \leq \epsilon/6$ and, for any $\bar{h} \in F_n(D, \Delta')$ with $\bar{d}_P(\bar{h}, \bar{f}) \geq \epsilon$, $\hat{d}_x(\bar{f}, \bar{h}) > \epsilon/6$. That is, with probability $1 - \delta$, the algorithm returns a function sequence $\bar{h}$ that satisfies $\bar{d}_P(\bar{f}, \bar{h}) < \epsilon$.

For some positive constants c_3 and c_4, the condition $\epsilon > c_1 \Delta' d \log(1/\Delta')$ is implied by the condition $\epsilon^2 > c_3 \Delta d \log(\epsilon/\Delta)$, which is implied by $\Delta < c_4 \epsilon^2/(d \log(d/\epsilon))$. □

Note that in contrast with Theorem 6, one cannot in general estimate a sequence in $F_n(P, \Delta)$ using just any consistent algorithm. There are examples for which for every n there are consistent sequences $\bar{h}$ with $V_P(\bar{h}) = 0$ but $\bar{d}_P(\bar{h}, \bar{f}) = 1$. For example, let $X = [0, 1]$ and P be the uniform distribution on $[0, 1]$. Let F contain the indicator functions of the concept $[0, 1]$ and all singletons $\{x\}$ for $x \in [0, 1]$. Note that $\mathrm{VCdim}(F) = 2$. Now, for any n consider the sequence $\bar{f} = (f_1, \ldots, f_n)$ with $f_i = [0, 1]$ for $i = 1, \ldots, n$. For every $x_1, \ldots, x_n$ the sequence $\bar{f}$ labels each x_i as 1. The sequence $\bar{h} = (h_1, \ldots, h_n)$ defined by $h_i = \{x_i\}$ also labels each x_i as 1. Thus, if the true sequence of concepts is $\bar{f}$, then for every $x_1, \ldots, x_n$ the corresponding $\bar{h}$ is consistent and yet $V_P(\bar{h}) = 0$ while $\bar{d}_P(\bar{h}, \bar{f}) = 1$. Clearly, Theorem 2 implies that $E\left|F_n(P, \Delta)_{|_x}\right|$ grows exponentially with n in this case. In fact, it is easy to see directly that $\left|F_n(P, \Delta)_{|_x}\right| = 2^n$ for every x.

4. Prediction

In this section, we consider the problem of online prediction of function sequences from labelled examples.

Definition 8. Suppose M is a positive integer, P is a distribution on X, and F_M is a set of legal function sequences of length M from a set F of functions. (In some cases, F_M depends on P.) A prediction problem for F_M proceeds as follows. There is an unknown target sequence $\bar{f} = (f_1, \ldots, f_M) \in F_M$. At each time $1 \leq t \leq M$, a prediction strategy receives an example $(x_t, f_t(x_t))$, where x_t is chosen according to P. The strategy then hypothesizes a function h_{t+1}.

For $0 < \epsilon, \delta < 1$, we say that a strategy can (ϵ, δ)-predict F_M online from M examples if, for all probability distributions P on X and all legal sequences $\bar{f}$ in F_M, the probability over x in X^M that $\bar{d}_P(\bar{h}, \bar{f}) \geq \epsilon$ is less than δ, where $\bar{h} = (h_1, \ldots, h_M)$.

Clearly, if there is a strategy that can (ϵ, δ)-predict a class F_M from M examples, then it can be used to construct an algorithm that can (ϵ, δ)-estimate F_M in the sense of Definition 1.

We shall construct prediction strategies that are based on the estimation results of the previous section. Theorem 10 below shows that these strategies predict well whenever the underlying estimation strategy works well. In fact, this result applies more generally to the problem of online prediction of elements of an arbitrary pseudometric space.

Definition 9. Let (F, d) be a pseudometric space. A prediction problem on this space is defined as follows. For $M \in \mathbb{N}$, there is a sequence $\bar{f}$ in F^M. At time $t-1$, a prediction strategy receives partial information $G_{t-1}(f_{t-1})$ about f_{t-1}. The strategy then predicts h_t. At time t, it receives partial information about f_t, and so on. The strategy aims to ensure that

$$\bar{d}(\bar{h}, \bar{f}) \left(\stackrel{\text{def}}{=} \frac{1}{M} \sum_{i=1}^{M} d(f_i, h_i) \right)$$

is small.

A k-th order Markovian prediction strategy H is one that, at time $t-1$, considers only the partial information that it has received about the sequence $(f_{t-k}, \ldots, f_{t-1})$ in forming its prediction h_t.

The pseudometric space that will be of interest to us is that of binary valued functions over some domain set X, with the pseudometric, d_P, induced by some probability measure, P, over X. The prediction strategy receives partial information $G_t(f_t)$ about the target function f_t in the form of the label $f_t(x_t)$ of f_t for a P-randomly drawn example x_t. However, for the statement and proof of Theorem 10 below, we shall remain in the more abstract setting of Definition 9.

Of course the error of a prediction strategy depends on the intrinsic 'fluctuations' of the sequence $\bar{f}$ it comes to estimate. We measure these fluctuations by the 'total variation' of

the function sequence $\bar{f}$ relative to a given metric d over F, $V_d(\bar{f})$ (defined in Eq. (1)). We shall consider separately the case where d is the discrete metric over F, giving rise to the variation $D(\bar{f})$ (defined in Eq. (2)),

Theorem 10. *Let (F, d) be a pseudometric space with $d(f, g) \leq 1$ for all $f, g \in F$. Consider $\bar{f} \in F^M$ for $M \in \mathbb{N}$. Let H be a k-Markovian prediction strategy for (F, d) and, for each $t \leq M$, let $\bar{f}_t^k = (f_{t-k}, \ldots, f_{t-1})$. Suppose that, at time t, H constructs a sequence $\hat{f}_t^k = (\hat{f}_{t,t-k}, \ldots, \hat{f}_{t,t-1}) \in F^k$ that satisfies $\bar{d}(\hat{f}_t^k, \bar{f}_t^k) \leq \epsilon$, and H predicts $h_t = \hat{f}_{t,t-1}$.*

We say that H is D-conservative if, for all t, $L_{\text{tail}}(\hat{f}_t^k) \geq L_{\text{tail}}(\bar{f}_t^k)$, where $L_{\text{tail}}(\bar{g})$ for a sequence $\bar{g} \in F^k$ denotes

$$L_{\text{tail}}(\bar{g}) = \max\{i \in \{1, \ldots, k\} : g_{k+1-i} = g_{k+2-i} = \cdots = g_k\},$$

the length of $\bar{g}$'s "constant tail".

We say that H is V_d-conservative if, for all t, $V_d(\hat{f}_t^k) \leq V_d(\bar{f}_t^k)$.

1. *If H is D-conservative then,*
   - *If $k > 1/(D(\bar{f}) + 1/M)$ then*

$$\bar{d}(\bar{h}, \bar{f}) \leq \frac{k}{M} + 2\epsilon k D(\bar{f}) \log\left(\frac{1}{D(\bar{f})}\right) + V_d(\bar{f}).$$

   - *If $k \leq 1/(D(\bar{f}) + 1/M)$ then*

$$\bar{d}(\bar{h}, \bar{f}) \leq \frac{k}{M} + \epsilon + 2\epsilon k D(\bar{f})(\log(k) - 1) + V_d(\bar{f}).$$

2. *If H is V_d-conservative, then*

$$\bar{d}(\bar{h}, \bar{f}) \leq k/M + \min\{\epsilon + 2kV_d(\bar{f}), k\epsilon\} + V_d(\bar{f}).$$

Furthermore, the bounds are 'local' in the sense that, if H only satisfies

$$\frac{1}{M-k}\left|\left\{k+1 \leq t \leq M : \bar{d}\left(\hat{f}_t^k, \bar{f}_t^k\right) > \epsilon\right\}\right| < \alpha,$$

then the upper bounds on $\bar{d}(\bar{h}, \bar{f})$ increase only by an additive term of α.

Proof: **1.** The proof proceeds in two stages. We start by producing an upper bound on the instantaneous error of a prediction strategy at each time point t. We then apply this bound to derive the desired upper bound for the cumulative error over the full prediction process.

Since $L_{\text{tail}}(\hat{f}_t^k) \geq L_{\text{tail}}(\bar{f}_t^k)$, for all $t - L_{\text{tail}}(\bar{f}_t^k) \leq i \leq t-1$, we have $d(f_i, \hat{f}_{t,i}) = d(f_{t-1}, \hat{f}_{t,t-1})$. By the triangle inequality, the error of the hypothesis $h_t = \hat{f}_{t,t-1}$ is

$$d(h_t, f_t) \leq d\left(\hat{f}_{t,t-1}, f_{t-1}\right) + d(f_{t-1}, f_t).$$

Note that the first term depends upon the strategy H, whereas the second term is an intrinsic property of the target sequence $\bar{f}$, and it sums up to $MV_d(\bar{f})$. Regarding the first term, we

have:

$$\begin{aligned}\bar{d}(\bar{f}_t^k, \hat{f}_t^k) &= \frac{1}{k}\sum_{i=t-k}^{t-1} d(\hat{f}_{t,i}, f_i) \\ &\geq \frac{1}{k}\sum_{i=t-L_{\text{tail}}(\bar{f}_t^k)}^{t-1} d(\hat{f}_{t,i}, f_i) \\ &= \frac{1}{k}L_{\text{tail}}(\bar{f}_t^k) d(\hat{f}_{t,t-1}, f_{t-1}).\end{aligned}$$

Applying the assumption that the hypothesis sequence $\hat{f}_t^k$ is ϵ-close to the target sequence $\bar{f}_t^k$, we conclude that $d(\hat{f}_{t,t-1}, f_{t-1}) \leq k\epsilon / L_{\text{tail}}(\bar{f}_t^k)$. We now turn to the cumulative error over the full sequence $\bar{f}$. We have

$$M\bar{d}(\bar{f}, \bar{h}) = \sum_{t=1}^{M} d(\hat{f}_{t,t-1}, f_t) \leq k + k\epsilon \sum_{t=k+1}^{M} \frac{1}{L_{\text{tail}}(\bar{f}_t^k)} + MV_d(\bar{f}).$$

Denote $\alpha_t = L_{\text{tail}}(\bar{f}_t^k)$. Notice that α_t takes values in the range 1 (when $f_{t-2} \neq f_{t-1}$) up to k (when $f_{t-k} = \cdots = f_{t-1}$). So from a point of change in $\bar{f}$, $1/\alpha_t$ starts at 1 and decays as $1/i$ until it reaches $1/k$, where it remains until the next change in $\bar{f}$. Among the sequences $\bar{f}$ with a fixed number of function switches (i.e. fixed value of $D(\bar{f})$), $\sum_{t=k+1}^{M} 1/\alpha_t$ assumes its maximal value when these switches are equally spaced along the sequence $\bar{f}$. (To see this, suppose that there are N switches in $\bar{f}$. Let s_i denote the index immediately after the ith switch, and let $s_0 = 0$ and $s_{N+1} = M + 1$. Then denote the length of the sequence preceding the ith switch by $l_i = s_i - s_{i-1}$ for $i = 1, \ldots, M + 1$. If we have $l_i \neq l_{i+1}$, moving the switch to decrease the larger (say, l_i) by one and increase the smaller by one will increase the sum by $1/(l_{i+1}) - 1/l_i \geq 0$.)

To upper bound the cumulative error for sequences of a fixed length M and a fixed $D(\bar{f})$, we therefore consider sequences composed of $(MD(\bar{f}) + 1)$ blocks, each consisting of $1/(D(\bar{f}) + 1/M)$ many identical functions. Let us consider two cases:

Case 1: $k > 1/(D(\bar{f}) + 1/M)$. On each sequence of $1/(D(\bar{f}) + 1/M)$ identical f_t's, the sum of the appropriate terms $1/\alpha_t$ is the harmonic series,

$$\sum_{i=1}^{1/(D(\bar{f})+1/M)} \frac{1}{i} \leq \log \frac{1}{D(\bar{f}) + 1/M}.$$

As there are $MD(\bar{f}) + 1$ many such subsequences in $\bar{f}$, we get

$$\begin{aligned}\bar{d}_P(\bar{f}, H) &\leq \frac{k}{M} + k\epsilon\left(D(\bar{f}) + \frac{1}{M}\right)\log\frac{1}{D(\bar{f}) + 1/M} + V_d(\bar{f}) \\ &\leq \frac{k}{M} + 2k\epsilon D(\bar{f})\log\left(\frac{1}{D(\bar{f})}\right) + V_d(\bar{f}).\end{aligned}$$

Case 2: $k \leq 1/(D(\bar{f}) + 1/M)$. In this case, on each subsequence of identical functions f_t, the corresponding sequence of $1/\alpha_t$'s consists of the harmonic sequence $1, 1/2, \ldots, 1/k$ followed by a sequence of $(1/(D(\bar{f}) + 1/M) - k)$ values equal to $1/k$. A straightforward calculation shows that in this case we get

$$\bar{d}_P(\bar{f}, H) \leq \frac{k}{M} + \epsilon + 2k\epsilon D(\bar{f})(\log(k) - 1) + V_d(\bar{f}).$$

2. Just as in the first part of the proof, at each stage t,

$$d(h_t, f_t) = d(\hat{f}_{t,t-1}, f_t) \leq d(\hat{f}_{t,t-1}, f_{t-1}) + d(f_{t-1}, f_t). \tag{5}$$

The second term sums up to $MV_d(\bar{f})$, so our task is to upper bound the sum $\sum_{t=1}^{M-1} d(\hat{f}_{t,t-1}, f_{t-1})$.

Since H is V_d-conservative, for every $k+1 \leq t \leq M$, $V_d(\hat{f}_t^k) \leq V_d(\bar{f}_t^k)$. This together with the triangle inequality imply that, for every $(t-k) \leq i \leq t-1$,

$$\begin{aligned} d(\hat{f}_{t,t-1}, f_{t-1}) &\leq d(\hat{f}_{t,i}, f_i) + \sum_{j=i}^{t-2} d(f_j, f_{j+1}) + \sum_{j=i}^{t-2} d(\hat{f}_{t,j}, \hat{f}_{t,j+1}) \\ &\leq d(\hat{f}_{t,i}, f_i) + kV_d(\bar{f}_t^k) + kV_d(\hat{f}_t^k) \\ &\leq d(\hat{f}_{t,i}, f_i) + 2kV_d(\bar{f}_t^k). \end{aligned} \tag{6}$$

Since, by assumption,

$$\frac{1}{k} \sum_{i=t-k}^{t-1} d(\hat{f}_{t,i}, f_i) \leq \epsilon,$$

there must be a $t-k \leq i \leq t-1$ with $d(\hat{f}_{t,i}, f_i) \leq \epsilon$. Applying inequality (6), we conclude that

$$d(\hat{f}_{t,t-1}, f_{t-1}) \leq \epsilon + 2kV_d(\bar{f}_t^k).$$

Averaging inequality (5) over all t, one gets,

$$\bar{d}(\bar{h}, \bar{f}) \leq k/M + \epsilon + 2kV_d(\bar{f}) + V_d(\bar{f}).$$

But the assumption $\bar{d}(\hat{f}_t^k, \bar{f}_t^k) \leq \epsilon$ also implies that $d(\hat{f}_{t,t-1}, f_{t-1}) \leq k\epsilon$. Averaging inequality (5) using this last inequality amounts to

$$\bar{d}(\bar{h}, \bar{f}) \leq k/M + k\epsilon + V_d(\bar{f}). \qquad \square$$

We can now combine this theorem with the estimation theorems derived in Section 3 to give a result on the prediction of changing concepts.

Definition 11. For a set F of $\{0, 1\}$-valued functions defined on X, and numbers k, $M \in \mathbb{N}$ and $0 < \Delta < 1$, define the sets of *k-local-Δ-frequently-switching* function M-sequences of F as

$$F_M^k(D, \Delta) = \left\{\bar{f} \in F^M : \text{for } k \leq t < M,\ D\left(\bar{f}_t^k\right) \leq \Delta\right\}.$$

For a probability measure P define the sets of *k-local-(P, Δ)-slowly-changing M-sequences of F* as

$$F_M^k(P, \Delta) = \left\{\bar{f} \in F^M : \text{for } k \leq t < M,\ V_P\left(\bar{f}_t^k\right) \leq \Delta\right\}.$$

Theorem 12. *Suppose that F is a class with VC-dimension d, and P is a probability distribution on X.*

1. *There are constants c_1 and c_2 such that for $0 < \epsilon, \delta, \Delta < 1$, if*

$$\epsilon > c_1 \Delta d \log \frac{1}{\Delta} \log \frac{1}{\Delta\delta},$$

$$k = \frac{c_2}{\epsilon}\left(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta\epsilon}\right),$$

and $M \geq k/\epsilon$, then there is a k-Markovian prediction strategy that can (ϵ, δ)-predict $F_M^k(D, \Delta)$ online from M examples.

2. *There are constants c_1 and c_2 such that for $0 < \epsilon, \delta, \Delta < 1$, if*

$$\epsilon^2 > c_1 \Delta d \log^2 \frac{d}{\delta\Delta},$$

$$k = \frac{c_2}{\epsilon}\left(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta\epsilon}\right),$$

and $M \geq k/\epsilon$, then there is a k-Markovian prediction strategy that can (ϵ, δ)-predict $F_M^k(P, \Delta)$ online from M examples.

Proof: **1.** For all $t \in \{k+1, \ldots, M\}$, $\bar{f}_t^k \in F_k(D, \Delta)$. At time t, our estimation algorithm chooses a consistent sequence $\hat{f}_t^k$. For any t, let M_t represent the set of training samples for which $\bar{d}_P(\hat{f}_t^k, \bar{f}_t^k) \geq \epsilon$. Theorem 6 implies that, if $\epsilon > c_1 \Delta d \log(1/\Delta)$ and

$$k \geq \frac{c_2}{\epsilon}\left(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta\epsilon}\right), \tag{7}$$

then for all t we have $\Pr(M_t) < \delta\epsilon$. It follows that

$$E\left(\frac{1}{M-k} \sum_{t=k+1}^{M} 1_{M_t}\right) < \delta\epsilon,$$

where 1_{M_t} is the indicator function for the set M_t. Markov's inequality implies that, with probability at least $1 - \delta$,

$$\left|\left\{t \in \{k+1, \ldots, M\} : \bar{d}_P\left(\hat{f}_t^k, \bar{f}_t^k\right) \geq \epsilon\right\}\right| < \epsilon(M-k).$$

If our consistent estimation algorithm also maximizes $L_{\text{tail}}(\hat{f}_t^k)$ for its estimate $\hat{f}_t^k$ at each time t, it is also D-conservative. Applying Theorem 10 (taking the maximum of both bounds), if we ensure $k < 1/(2\Delta \log(1/\Delta))$ and $M \geq k/\epsilon$, we have $\bar{d}_P(\bar{h}, \bar{f}) \leq 5\epsilon$. For the first of these inequalities, it suffices if we set k equal to the right hand side of (7), and ensure that

$$\epsilon > c_3 \Delta d \log \frac{1}{\Delta} \log \frac{1}{\delta\epsilon}$$

$$\Leftarrow \quad \epsilon > c_3 \Delta d \log \frac{1}{\Delta} \log \frac{1}{\delta\Delta}.$$

2. The algorithm of Theorem 7 can easily be modified by, instead of choosing the $\bar{h}$ that is closest to the target and does not have too many switches, choosing one that maximizes the length of its "constant tail" but is sufficiently close to the target and does not have too many switches. The proof of Theorem 7 shows that, with probability at least $1 - \delta$, there is a function sequence $\bar{g}$ in $F_n(D, 24\Delta/\epsilon)$ with $\hat{d}_x(\bar{g}, \bar{f}) \leq \epsilon/6$. So if the algorithm returns a function sequence $\bar{h}$ that has the longest "constant tail" of those sequences in $F_n(D, 24\Delta/\epsilon)$ satisfying $\hat{d}_x(\bar{h}, \bar{f}) \leq \epsilon/6$, it will be D-conservative. We can then use an argument identical to the proof above of the first part of the theorem, but with Δ replaced by $24\Delta/\epsilon$. In this case, we get

$$\epsilon > c_3 \Delta d \frac{1}{\epsilon} \log \frac{\epsilon}{\Delta} \log \frac{1}{\delta\epsilon}$$

$$\Leftarrow \quad \epsilon^2 > c_4 \Delta d \log \left(d \log \frac{1}{\delta\Delta} \right) \log \frac{1}{\delta\Delta}$$

$$\Leftarrow \quad \epsilon^2 > c_4 \Delta d \log^2 \frac{d}{\delta\Delta},$$

for some universal constant c_4. □

Appendix A: Proof of Theorem 2

Proof: Define

$$Q = \{x \in X^n : \exists \bar{g} \in F_n, \bar{g}(x) = \bar{f}(x), \bar{d}_P(\bar{f}, \bar{g}) \geq \epsilon\}$$

$$R = \left\{(x, y) \in X^{2n} : \exists \bar{g} \in F_n, \bar{g}(x) = \bar{f}(x), \bar{d}_P(\bar{f}, \bar{g}) \geq \epsilon, \hat{d}_y(\bar{f}, \bar{g}) \geq \frac{\epsilon}{2}\right\}.$$

We shall first show that

$$P^n(Q) \leq \frac{1}{1 - e^{-\epsilon n/8}} P^{2n}(R). \tag{8}$$

To see this, notice that

$$P^{2n}(R) = \int_Q P^n \left\{ y : \exists \bar{g} \in F_n, \bar{g}(x) = \bar{f}(x), \bar{d}_P(\bar{f}, \bar{g}) \right. \\ \left. \geq \epsilon, \hat{d}_y(\bar{f}, \bar{g}) \geq \frac{\epsilon}{2} \right\} dP^n(x). \tag{9}$$

Fix $x \in X^n$ and $\bar{g} \in F_n$ with $\bar{d}(\bar{f}, \bar{g}) \geq \epsilon$ and $\bar{g}(x) = \bar{f}(x)$. Then

$$P^n \left\{ y \in X^n : \hat{d}_y(\bar{f}, \bar{g}) \geq \frac{\epsilon}{2} \right\} = 1 - \Pr\left(\frac{1}{n} \sum_{i=1}^{n} X_i < \frac{\epsilon}{2} \right),$$

where $X_i \in \{0, 1\}$ and $\Pr(X_i = 1) = \Pr(f_i(x_i) = g_i(x_i))$, so that

$$\frac{1}{n} \sum_{i=1}^{n} \Pr(X_i = 1) = \bar{d}(\bar{f}, \bar{g}) \geq \epsilon.$$

Chernoff bounds (see, for example, (Hagerup and Rub, 1990)) imply that

$$\Pr\left(\frac{1}{n} \sum_{i=1}^{n} X_i < \frac{\epsilon}{2} \right) \leq e^{-\epsilon n/8}.$$

It follows that the probability inside the integral in (9) is at least $1 - \exp(-\epsilon n/8)$, which implies (8).

Let U be the uniform distribution on the set of permutations on $\{1, \ldots, 2n\}$ that swap elements from the first to the second half of the sequence (that is, all permutations σ that satisfy $\{\sigma(i), \sigma(i+n)\} = \{i, i+n\}$ for all $i \in \{1, \ldots, n\}$). For such a permutation σ denote the permuted version of a sequence $(x, y) \in X^{2n}$ by (x^σ, y^σ). Then, given the assumption on n, the probability of Q is less than

$$2E_{(x,y)\sim P^{2n}} U \left\{ \sigma : \exists \bar{g} \in F_n \text{ s.t. } \bar{g}(x^\sigma) = \bar{f}(x^\sigma) \right. \\ \left. \text{and } \bar{d}_P(\bar{f}, \bar{g}) > \epsilon \text{ and } \hat{d}_{y^\sigma}(\bar{f}, \bar{g}) > \frac{\epsilon}{2} \right\}$$

$$\leq 2E_{(x,y)\sim P^{2n}} U \left\{ \sigma : \exists (v, w) \in F_{n|_{(x,y)}} \text{ s.t. } \bar{f}(x^\sigma) = v^\sigma \right. \\ \left. \text{and } \frac{1}{n} \left| \left\{ i : f_i(y_{\sigma(i)}) \neq w_{\sigma(i)} \right\} \right| \geq \frac{\epsilon}{2} \right\},$$

where $F_{n|_{(x,y)}}$ is the set

$$\{(h_1(x_1), \ldots, h_n(x_n), h_1(y_1), \ldots, h_n(y_n)) : \bar{h} \in F_n\}.$$

The union bound and a simple counting argument show that this quantity is no more than

$$2E|F_{n|_{(x,y)}}| \sup_{(v,w)\in F_{n|_{(x,y)}}} U\left\{\sigma : \bar{f}(x^\sigma) = v^\sigma \text{ and } \frac{1}{n}\left|\left\{i : f_i\left(y_{\sigma(i)}\right) \neq w_{\sigma(i)}\right\}\right| \geq \frac{\epsilon}{2}\right\}$$
$$\leq 2E\left|F_{n|_{(x,y)}}\right| 2^{-\epsilon n/2}.$$

Notice that $F_{n|_{(x,y)}}$ is contained in

$$\{(h_1(x_1), \ldots, h_n(x_n)) : \bar{h} \in F_n\} \times \{(h_1(y_1), \ldots, h_n(y_n)) : \bar{h} \in F_n\},$$

which gives the desired result. □

Appendix B: Proof of Theorem 3

Proof: Define

$$Q = \{x \in X^n : \exists \bar{g} \in F_n,\ d_\gamma(\hat{d}_x(\bar{f}, \bar{g}), \bar{d}_P(\bar{f}, \bar{g})) \geq \alpha\},$$
$$R = \left\{(x, y) \in X^{2n} : \exists \bar{g} \in F_n,\ d_\gamma(\hat{d}_x(\bar{f}, \bar{g}), \hat{d}_y(\bar{f}, \bar{g})) \geq \frac{\alpha}{2}\right\}.$$

Notice that the triangle inequality for d_γ (see (Haussler, 1992)) implies that

$$P^{2n}(R) \geq \int_Q P^n \left\{y : \exists \bar{g} \in F_n,\ d_\gamma(\hat{d}_x(\bar{f}, \bar{g}), \bar{d}_P(\bar{f}, \bar{g})) \geq \alpha, \text{ and } d_\gamma(\bar{d}_P(\bar{f}, \bar{g}), \hat{d}_y(\bar{f}, \bar{g})) \leq \frac{\alpha}{2}\right\} dP^n(x). \tag{10}$$

Fix $x \in Q$ and $\bar{g} \in F_n$ with $d_\gamma(\hat{d}_x(\bar{f}, \bar{g}), \bar{d}_P(\bar{f}, \bar{g})) \geq \alpha$. We shall show that, for $n \geq 5/(\alpha^2\gamma)$,

$$P^n\left\{y : d_\gamma(\bar{d}_P(\bar{f}, \bar{g}), \hat{d}_y(\bar{f}, \bar{g})) > \frac{\alpha}{2}\right\} \leq \frac{1}{2}. \tag{11}$$

To see this, notice that $\hat{d}_y(\bar{f}, \bar{g}) \geq 0$ implies that

$$P^n\left\{\frac{|\bar{d}_P(\bar{f}, \bar{g}) - \hat{d}_y(\bar{f}, \bar{g})|}{\bar{d}_P(\bar{f}, \bar{g}) + \hat{d}_y(\bar{f}, \bar{g}) + \gamma} > \frac{\alpha}{2}\right\}$$
$$\leq P^n\left\{|\bar{d}_P(\bar{f}, \bar{g}) - \hat{d}_y(\bar{f}, \bar{g})| > \frac{\alpha}{2}(\bar{d}_P(\bar{f}, \bar{g}) + \gamma)\right\}$$
$$= P^n\left\{\left|\mu - \frac{1}{n}\sum_{i=1}^n X_i\right| > \frac{\alpha}{2}(\mu + \gamma)\right\}$$

where the independent random variables $X_i \in \{0, 1\}$ satisfy $\Pr(X_i = 1) = \Pr(f_i(x_i) \neq g_i(x_i))$, and μ is defined as $\mu = (1/n)\sum_{i=1}^{n} \Pr(X_i = 1)$. Chernoff bounds imply that this probability is no more than

$$2\exp\left(-\frac{n\alpha^2(\mu+\gamma)^2}{12\mu}\right).$$

Elementary calculus shows that $n\alpha^2(\mu+\gamma)^2/(12\mu)$ is minimized when $\mu = \gamma$, so the probability is no more than $2\exp(-n\alpha^2\gamma/3)$. This is no more than $1/2$ for $n \geq 4/(\alpha^2\gamma)$, which implies (11). It follows that, for any $x \in Q$, the probability inside the integral in (10) is at least $1/2$, and so $P^n(Q) \leq 2P^{2n}(R)$.

Now,

$$\begin{aligned}
&P^n\{x : \exists \bar{g} \in F_n \text{ s.t. } d_\gamma(\hat{d}_x(\bar{f}, \bar{g}), \bar{d}_P(\bar{f}, \bar{g})) \geq \alpha\} \\
&\quad \leq 2P^{2n}\left\{(x, y) : \exists \bar{g} \in F_n \text{ s.t. } d_\gamma(\hat{d}_x(\bar{f}, \bar{g}), \hat{d}_y(\bar{f}, \bar{g})) \geq \frac{\alpha}{2}\right\} \\
&\quad = 2E_{(x,y)\sim P^{2n}} U\left\{\sigma : \exists \bar{g} \in F_n \text{ s.t. } d_\gamma(\hat{d}_{x^\sigma}(\bar{f}, \bar{g}), \hat{d}_{y^\sigma}(\bar{f}, \bar{g})) \geq \frac{\alpha}{2}\right\},
\end{aligned}$$

where U is the uniform distribution on the group of swapping permutations on the set $\{1, \ldots, 2n\}$, as in the proof of Theorem 2. Taking the union bound as in that proof, we have that

$$\begin{aligned}
&P^n\{x : \exists \bar{g} \in F_n \text{ s.t. } d_\gamma(\hat{d}_x(\bar{f}, \bar{g}), \bar{d}_P(\bar{f}, \bar{g})) \geq \alpha\} \\
&\quad \leq 2E\left|F_{n|(x,y)}\right| \sup_{(v,w)\in X^{2n}} U\left\{\sigma : d_\gamma(\hat{d}_{v^\sigma}(\bar{f}, \bar{g}), \hat{d}_{w^\sigma}(\bar{f}, \bar{g})) \geq \frac{\alpha}{2}\right\}.
\end{aligned}$$

Now, $d_\gamma(\hat{d}_{x^\sigma}(\bar{f}, \bar{g}), \hat{d}_{y^\sigma}(\bar{f}, \bar{g})) \geq \alpha/2$ if and only if

$$\left|\frac{1}{n}\sum_{i=1}^{n} \beta_i(a_i - b_i)\right| \geq \frac{\alpha}{2}\left(\frac{1}{n}\sum_{i=1}^{n}(a_i + b_i) + \gamma\right),$$

where $a_i = |f_i(v_i) - g_i(v_i)|$, $b_i = |f_i(w_i) - g_i(w_i)|$, and the independent random variables $\beta_i \in \{-1, 1\}$ satisfy $\Pr(\beta_i = 1) = 1/2$. Then Bernstein's inequality (see, for example, (Anthony and Bartlett, 1999, p. 363)) implies that this occurs with probability no more than

$$2\exp\left(-\frac{(\alpha^2/4)\left((1/n)\sum_{i=1}^{n}(a_i + b_i) + \gamma\right)^2 n}{(2/n)\sum_{i=1}^{n}(a_i - b_i)^2 + (\alpha/3)\left((1/n)\sum_{i=1}^{n}(a_i + b_i) + \gamma\right)}\right).$$

Since $a_i, b_i \in \{0, 1\}$, $(a_i - b_i)^2 \leq a_i + b_i$, so this probability is no more than

$$2\exp\left(-\frac{(\alpha^2/4)(S+\gamma)^2 n}{2S + (\alpha/3)(S+\gamma)}\right),$$

where $S = (1/n)\sum_{i=1}^{n}(a_i + b_i)$. Since $(S+\gamma)^2/(2S + (\alpha/3(S+\gamma))$ is minimized at $S = \gamma(2-\alpha/3)/(2+\alpha/3)$, this is no more than

$$2\exp\left(-\frac{2\alpha^2\gamma^2 n}{(2+\alpha/3)^2\gamma}\right) \leq 2\exp\left(-\frac{\alpha^2\gamma n}{3}\right),$$

since $\alpha \leq 1$. □

Acknowledgments

This research was partially supported by the Australian Research Council, and by a DIST Bilateral Science and Technology Collaboration Program Travel Grant. Thanks to the anonymous reviewers for helpful comments on an earlier version of this paper. In particular, thanks to the reviewer who suggested an improvement to the proof of Theorem 6.

References

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

Bartlett, P. L. (1992). Learning with a slowly changing distribution. In *Proceedings of the 1992 Workshop on Computational Learning Theory* (pp. 243–252).

Bartlett, P. L. & Helmbold, D. P. (1996). Learning changing problems. Technical report, Australian National University.

Barve, R. D. & Long, P. M. (1997). On the complexity of learning from drifting distributions. *Information and Computation, 138*(2), 101–123.

Bertoni, A., Campadelli, P., Morpurgo, A., & Panizza, S. (1992). Polynomial uniform convergence and polynomial-sample learnability. In *Proceedings of the 1992 Workshop on Computational Learning Theory* (pp. 265–271).

Blum, A. & Chalasani, P. (1992). Learning switching concepts. In *Proceeding of the fifth Annual Workshop on Computer Learning Theory* (pp. 231–242). New York, NY: ACM Press.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis Dimension. *J. ACM, 36*(4), 929–965.

Connolly, A. J., Chicharo, J. F., & Wilbers, P. (1992). Temperature modelling and prediction of steel strip in a HSM. In *Control 92 Conference Proceedings*. Institute of Engineers Australia.

Freund, Y. & Mansour, Y. (1997). Learning under persistent drift. In S. Ben-David, (Ed.), *Computational Learning Theory: Third European Conference, Euro-COLT'97* (pp. 109–118). Springer.

Hagerup, T. & Rub, C. (1990). A guided tour of Chernov bounds. *Inform. Proc. Lett., 33*, 305–308.

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

Helmbold, D. P. & Long, P. M. (1994). Tracking drifting concepts by minimizing disagreements. *Machine Learning, 14*, 27.

Long, P. M. (1998). The complexity of learning according to two models of a drifting environment. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory* (pp. 116–125). ACM.

Merhav, N. & Feder, M. (1993). Universal schemes for sequential decision from individual data sequences. *IEEE Transactions on Information Theory, 39*(4), 1280–1292.

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

Received May 13, 1998
Revised October 6, 1999
Final manuscript September 15, 1999