Tracking a drifting concept in a changing environment

Peter L. Bartlett*
David P. Helmbold†

UCSC-CRL-98-12
March 1998

* Department of Systems Engineering
Australian National University
Canberra 0200, Australia

† Board of Studies in Computer and Information Sciences
University of California, Santa Cruz
Santa Cruz, CA 95064

ABSTRACT

We consider the problem of predicting the labels of randomly chosen points, when both the probability distribution generating the points and the target concept are allowed to change slowly. More precisely, we assume that the total variation distance between consecutive probability distributions is small, and that the probability (under the current distribution) of disagreement between consecutive target concepts is small. We describe a general-purpose algorithm that can cope with combined distribution and concept drift, and has asymptotic prediction error within a log factor of the best that can be achieved by any algorithm, even when either the distribution or the concept is held constant. In addition, we give upper bounds on the cumulative prediction error of a related algorithm when we assume that there is a bound on the average drift over a sequence of trials. We also consider learning in an agnostic setting, with a slowly changing joint probability distribution on the domain and the output space $\{0, 1\}$, and give bounds on the excess mistake probability of a tracking strategy in this setting.

1 Introduction

In this paper we consider the problem of tracking a changing environment. The environment is modeled as a domain from which examples are drawn and a boolean valued target function from some class labeling the examples. At each discrete point in time a random example is drawn and the algorithm predicts the label, which is then revealed to the algorithm. Our drifting model differs from standard learning models in that after each label is revealed to the algorithm both the distribution on the domain and the target function can change.

Of course, if the distribution or target function could change arbitrarily then it would be extremely difficult to do any learning. For our main results we bound the amount that the distribution and target function can change at each point in time, and communicate this bound to the algorithm. We use the total variation distance to measure the change in the distribution (although other measures such as the Kullback-Leibler divergence give similar results) and we use the probability under the current distribution of the symmetric difference of the functions to measure the change in target.

Our main result is a general algorithm whose mistake rate for moderate α is $O(\sqrt{d\alpha})$ where α is the (known) combined drift rate of the target and distribution. For smaller drift rates ($\alpha < 1/e^d$), the VC-dimension gets replaced with a logarithmic factor, becoming $O(\sqrt{\alpha \log 1/\alpha})$. This implies that combined drift rates in $\Theta(\epsilon^2/\ln \epsilon)$ can be tolerated while maintaining a mistake rate less than $\epsilon \leq 1/e^d$. Similarly, if $\epsilon \geq 1/e^d$ then the algorithm has a mistake rate less then ϵ whenever the combined drift rate is less than a constant times ϵ^2/d.

To complement these results, Helmbold and Long [10, 11] give examples showing that no algorithm can have mistake rate ϵ for drift rates as small as a constant times ϵ^2/d, even when the distribution remains fixed and only the target function drifts. We give an additional lower bound showing that for some concept classes, no algorithm can have a mistake rate less than ϵ when the target function is held constant and only the distribution drifts at a rate in $\Theta(\epsilon^2/d)$. It remains open if our algorithm's performance is optimal for very small drift rates.

Thus, for moderate drift rates, our algorithm tolerates a mixture of concept and distribution drift with the same performance (to within a constant factor) of algorithms that must deal only with drifting distributions, or only with drifting targets.

We also examine two variations on this basic problem. It is not surprising that the problem becomes more difficult when only a bound on the average amount of drift is given, since an adversary could "save up" drift in order to make a drastic change in the distribution and/or target. For this problem our algorithm has mistake rate $O(\sqrt[3]{d\alpha})$ or $O(\sqrt[3]{\alpha \ln(1/\alpha)})$ when the average drift is bounded by α.

The second variation is an agnostic setting, where there is a joint probability distribution on $X \times \{0,1\}$ which is slowly drifting. This joint distribution allows one to model noise and errors. Since $\{0,1\}$-valued functions cannot perfectly predict values drawn from an arbitrary joint distribution, we measure the mistake rate of the algorithm minus the mistake rate when, at each time, the best member in the class for that time is used to generate predictions. We give an algorithm where this difference is bounded by $19d^{2/5}\gamma^{1/5}\sqrt{\ln(2d^{-2/5}\gamma^{-1/5})}$, where γ is the drift rate of the joint distribution.

Some results in this paper (in particular, in Sections 5.2 and 6) have been previously presented in [2], which studied the problem of learning with a drifting distribution. There

have been some recent results in this area. Barve and Long [3] study learning with a drifting distribution, and give an improvement on our bounds for agnostic learning. Specifically, they show that there is an algorithm whose mistake rate is within $O((\gamma d)^{1/3})$ of optimal when the distribution drifts by no more than γ between trials.

The problem of tracking a drifting concept with a fixed distribution has been studied previously. Helmbold and Long [10, 11] exhibit a tracking algorithm, and show that it has nearly optimal error for a given amount of concept drift. The algorithm we consider in Section 3 is based on their tracking algorithm. Kuh, Petsche, and Rivest [13] consider learning a drifting interval on the circumference of the unit circle under the uniform distribution. They show that in this case a conservative tracker's mistake rate is $\sqrt{2\alpha/\pi} \pm O(\alpha)$ by analyzing a Markov chain. In addition to experimental validation of this bound, they also consider "benign" adversaries which cause the target concept to drift randomly. In an earlier paper [12], they discuss several tracking variations and define PAC-tracking. There Kuh *et al.* also distinguish between incremental and memory based tracking algorithms. In their terminology, our algorithms are memory based trackers because they use a sliding window containing the most recently seen examples to make their predictions.

In the next section we formalize our learning model and introduce our notation. Section 3 describes the *exception tracking strategy* and bounds its performance in terms of its parameters. Section 4 considers particular settings of the parameters to obtain more meaningful bounds on the performance of the exception tracking strategy. Several lower bounds nearly matching the upper bounds in Section 4 are given in Section 5. Section 6 considers a slightly different "agnostic" model where the learner attempts to perform as well as the current best concept when the examples are generated by a drifting joint probability distribution on $X \times \{0, 1\}$.

2 Definitions and Notation

The learning model described here is similar to the prediction model of learning described by Haussler *et al.* [9]. We have a *domain* X and a *target class* F of functions mapping from X to $\{0, 1\}$. Alternatively, each member of the target class can be viewed as the subset of the domain which it maps to 1. At each time k, an *example* x_k is randomly chosen from X according to an unknown probability distribution P_k on the domain. The learning algorithm tries to predict the value of $f_k(x_k)$ (the *label* of x_k). The algorithm is then told the label, and the process is repeated. In the prediction model described by Haussler *et al.* [9], the distribution P_k and function f_k do not vary with k, remaining constant for all time. In the model considered here, the distribution and function are allowed to vary slowly and the learner is required to track them.

A sequence $\mathbf{x} = (x_1, x_2, \ldots, x_t) \in X^t$ of examples is called a sample. A labeled sample is a sequence $((x_1, f_1(x_1)), \ldots, (x_t, f_t(x_t)))$ of labeled examples. For sample $\mathbf{x} = (x_1, x_2, \ldots, x_t) \in X^t$ and function sequence $\mathbf{f} = (f_1, f_2, \ldots, f_t) \in F^t$, define the labeled sample of $\mathbf{f}$ generated by $\mathbf{x}$ as

$$\mathrm{sam}_t(\mathbf{x}, \mathbf{f}) = ((x_1, f_1(x_1)), \ldots, (x_t, f_t(x_t))) .$$

Throughout, we assume that we have a measurable space $(X, \mathcal{S})$, and that every subset of X that we consider is in $\mathcal{S}$. Blumer *et al.* [4] give conditions on the class F that ensure this is true.

For probability distributions P and Q on a measurable space $(X, \mathcal{S})$, define the total variation distance between P and Q as

$$d_{TV}(P,Q) = \sup\left\{\sum_{E\in R} |P(E) - Q(E)| : R \subseteq \mathcal{S} \text{ partitions } X\right\}.$$

It is easy to see that

$$d_{TV}(P,Q) = 2\sup_{E\in\mathcal{S}} |P(E) - Q(E)|.$$

For a sequence $\mathbf{P} = (P_1, \ldots, P_t)$ of probability distributions on X, define the *drift sequence* of $\mathbf{P}$ as $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{t-1})$ where $\gamma_i = d_{TV}(P_i, P_{i+1})$ for $i = 1, \ldots, t-1$. Define

$$\|\boldsymbol{\gamma}\|_\infty = \max_{1\le i\le t-1} \gamma_i,$$

and

$$\|\boldsymbol{\gamma}\|_1 = \frac{1}{t-1}\sum_{i=1}^{t-1} \gamma_i.$$

We will consider those distribution sequences where the drift $\boldsymbol{\gamma}$ is small, i.e. either $\|\boldsymbol{\gamma}\|_\infty$ or $\|\boldsymbol{\gamma}\|_1$ is less than some small positive number.

Another natural definition of the distance between two distributions is the Kullback-Leibler divergence. If P and Q are probability distributions on X, the Kullback-Leibler divergence of P with respect to Q is

$$d_{KL}(P,Q) = \int_X \log\frac{dP(\omega)}{dQ(\omega)}\, dP(\omega),$$

where dP/dQ is the Radon-Nikodym derivative of P with respect to Q. Kullback [14] has shown that $d_{KL}(P,Q) \ge (d_{TV}(P,Q))^2/2 + (d_{TV}(P,Q))^4/12$, so it is easy to apply the upper bounds described in Sections 4 and 6 when the drift sequence is measured in terms of d_{KL} rather than d_{TV}.

For a sequence $\mathbf{f} = (f_1, \ldots, f_t)$ of concepts from F, define the *drift sequence* of $\mathbf{f}$ as $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{t-1})$ with $\delta_i = P_i(f_i \ne f_{i+1})$ for $i = 1, \ldots, t-1$. Although the drift sequence of $\mathbf{f}$ depends on the sequence of probability distributions, the appropriate $\mathbf{P}$ will always be clear from the context. We will consider distribution sequences with drift $\boldsymbol{\delta}$ where either $\|\delta\|_\infty$ or $\|\delta\|_1$ is small.

Define the space of labeled examples, $X \times \{0,1\}$ and the space of finite length labeled samples, $(X \times \{0,1\})^* = \cup_{m\in\mathcal{N}} (X \times \{0,1\})^m$.

A *deterministic prediction strategy* Q is a function from $(X \times \{0,1\})^* \times X$ to $\{0,1\}$. A *randomized prediction strategy* (Q_r, Z, D) 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 $(X \times \{0,1\})^* \times X \times Z$ to $\{0,1\}$.

We define the mistake probability of a prediction strategy as follows. Fix a sample $\mathbf{x} = (x_1, \ldots, x_t) \in X^t$ $(t \ge 1)$, function sequence $\mathbf{f} = (f_1, \ldots, f_t)$ from F, and let $\mathbf{x}^- = (x_1, \ldots, x_{t-1})$ and $\mathbf{f}^- = (f_1, \ldots, f_{t-1})$. We can now define the mistake of a deterministic prediction strategy Q on $\mathbf{x}$ with respect to $\mathbf{f}$ as

$$M^t_{Q,\mathbf{f}}(\mathbf{x}) = \begin{cases} 1 & \text{if } Q(\text{sam}_{t-1}(\mathbf{x}^-, \mathbf{f}^-), x_t) \ne f_t(x_t) \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, we define the mistake probability of a randomized prediction strategy (Q_r, Z, D) on $\mathbf{x}$ with respect to $\mathbf{f}$ as

$$M^t_{Q_r,\mathbf{f}}(\mathbf{x}) = D\left\{z \in Z : Q_r\left(\mathrm{sam}_{t-1}(\mathbf{x}^-, \mathbf{f}^-), x_t, z\right) \neq f_t(x_t)\right\}.$$

Finally, we generalize both notations to function sequences of zero drift, writing $M^t_{Q,f}(\mathbf{x})$ for $M^t_{Q,\mathbf{f}}(\mathbf{x})$ where $\mathbf{f}$ is the sequence $(f, f, \ldots, f)$ of length t.

We are interested in both the instantaneous loss, $E\left(M^t_{Q,\mathbf{f}}\right)$, and in the cumulative loss, $E\left(\sum_{i=1}^t M^i_{Q,\mathbf{f}}\right)$. Here the expectations are over both the random draw of $\mathbf{x}$ from the distribution sequence $\mathbf{P}$ and any randomization used by the algorithm.

3 The Exception Tracking Strategy

The exception tracking strategy is a modification of the one-inclusion graph algorithm of Haussler *et al.* [9] in the spirit of Helmbold and Long [10]. In its original form, the one-inclusion graph algorithm is given a labeled set of examples consistent with a function in the class. A permutation argument exploiting the fact that each element in the sample is generated from the same distribution is used to bound the performance of the one-inclusion graph algorithm.

Unfortunately, in our setting the function sequence drifts. It is possible (and often likely) that no single function in the class matches the labeled sample. Therefore we extend the function class (as was done by Helmbold and Long [10]) to include all functions which are "close" to a function in the class on the sample.

Since the distributions are also drifting, all permutations of the example sequence are not equally likely. However, we are still able to use a (more sophisticated) permutation argument to bound the performance of the algorithm. The main idea is to show that for most examples in the sample the probability and function value do not change very much, and so we can apply the permutation argument to those examples.

3.1 Volatile and Uncertain points

For the analysis it is useful to identify the subset of X for which the probability of points varies greatly under the different distributions P_i. The following lemma shows one way to do this. The proof is in the appendix.

Lemma 1: *If P and Q are probability distributions defined on the measurable space $(X, \mathcal{S})$ and α is a positive constant, then there is a set $E \in \mathcal{S}$ such that for all $S \in \mathcal{S}$, we have*

1. *either $Q(S \cap E) = 0$ or $P(S \cap E) > \alpha Q(S \cap E)$, and*
2. *$P(S \cap (X - E)) \leq \alpha Q(S \cap (X - E))$.*

The set E is unique up to measure zero differences (that is, for any other suitable E', we have $P(E \oplus E') = Q(E \oplus E') = 0$, where $\oplus$ denotes symmetric difference).

Furthermore, for any $S \in \mathcal{S}$ and non-negative measurable function $f : X \to \mathcal{R}$,

$$\int_{S\cap(X-E)} f\, dP \leq \alpha \int_{S\cap(X-E)} f\, dQ.$$

We will refer to a set E defined by the theorem as $\textsc{Set}(P > \alpha Q)$. This definition captures the intuitive notion of the set in which the distribution P is more than α times as dense as distribution Q. We extend this notation in the obvious way, letting $\textsc{Set}(P \leq \alpha Q) = X - \textsc{Set}(P > \alpha Q)$, $\textsc{Set}(P < \alpha Q) = \textsc{Set}(Q > \frac{1}{\alpha} P)$, and $\textsc{Set}(P \geq \alpha Q) = X - \textsc{Set}(P < \alpha Q)$.

If $\mathbf{P} = (P_1, \ldots, P_t)$ is a sequence of probability distributions defined on $(X, \mathcal{S})$, define the sets H_i and L_i for $1 \leq i < t$ as $H_i = \textsc{Set}(P_i > 2P_t)$ and $L_i = \textsc{Set}(P_i < \frac{1}{2} P_t)$. Thus H_i is the "heavy" set of points whose density under P_i is twice their density under P_t, and L_i is the "light" points whose density under P_i is half their density under P_t. A point is called *volatile* at time i if it is in either H_i or L_i. The set of volatile points at, or after, time k is denoted by V_k. Formally, $V_k = \bigcup_{i=k}^{t-1} (H_i \cup L_i)$.

The following result shows that volatile points are not very likely if the distribution does not change too quickly.

Lemma 2: *If $\mathbf{P} = (P_1, \ldots, P_t)$ is a sequence of t probability distributions on $(X, \mathcal{S})$, with distribution drift $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{t-1})$, then for all $1 \leq k \leq i \leq t$*

$$P_i(V_k) \leq 3/2 \sum_{j=k}^{t-1} \gamma_j.$$

The proof is given in the Appendix.

Let $\mathbf{f} = (f_1, \ldots, f_t)$ be any sequence of t target functions from F. A point's label changes at time j if it is labeled differently by f_j and f_{j+1}. A point is *uncertain* at time k if its label changes at time k or later. Thus the set of uncertain points at time k (for $1 \leq k < t$), denoted U_k, is defined by $U_k = \bigcup_{k \leq j < t} (f_j \oplus f_{j+1})$.

Lemma 3: *Let $\mathbf{P} = (P_1, \ldots, P_t)$ be a sequence of t probability distributions on $(X, \mathcal{S})$ with drift $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{t-1})$, and let $\mathbf{f} = (f_1, \ldots, f_t)$ be a sequence of t functions on X with drift $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{t-1})$. For all k and i such that $1 \leq k \leq i \leq t$, if U_k is the set of uncertain points defined above, then*

$$P_i(U_k) \leq \sum_{j=k}^{t-1} (\delta_j + \gamma_j/2).$$

The proof is given in the Appendix.

3.2 The one-inclusion tracking strategy, and an upper bound

The one-inclusion graph algorithm of [9] makes a mistake on relatively few permutations of any sample. Let $\mathcal{A}_F$ be the one-inclusion graph algorithm for any concept (function) class F, d be the VC-dimension of F, and for each $1 \leq i \leq t$, let $\sigma_{i,t}$ be the permutation on t elements that swaps elements i and t while leaving the other elements unchanged. Haussler *et al.* show that for any positive integer t, sequence of examples $\mathbf{x} \in X^t$, and single function $f \in F$,

$$\left| \left\{ i : M^t_{\mathcal{A}_F, f}(\mathbf{x}^{\sigma_{i,t}}) = 1 \right\} \right| \leq 2d.$$

Define the class $F^{\oplus u}$ for $u \in \mathcal{N}$ to be F closed under up to u exceptions. Thus $F^{\oplus u} = \{ f \oplus S : f \in F, S \subset X, |S| \leq u \}$.

It has been shown by Auer and Long [1] that the VC-dimension of $F^{\oplus u}$ is at most $4.82(d+u)$ where d is the VC-dimension of F, and results of Cesa-Bianchi *et al.* [5] imply that it is bounded by $4.404(d+u)$. We first analyze the case where our algorithm picks u and then predicts using the one-inclusion graph algorithm for $F^{\oplus u}$ on the examples $(x_1, x_2, \ldots, x_t)$. When t is large, the first examples can be very misleading and we describe later how to adapt the following theorem when only a suffix of the examples is used.

Theorem 4: *Let F be a function class defined on some domain X with VC-dimension d, and let Q be the prediction strategy that applies the one-inclusion graph algorithm for the concept class $F^{\oplus u}$ for some $u \in \mathcal{N}$. For any distribution sequence $\mathbf{P}$ of t distributions on X and function sequence $\mathbf{f}$ of t functions from F, if:*

- *$\boldsymbol{\gamma}$ is the drift sequence of $\mathbf{P}$,*
- *$\boldsymbol{\delta}$ is the drift sequence of $\mathbf{f}$,*
- *d' is the VC-dimension of $F^{\oplus u}$ (at most $4.404(d+u)$), and*
- $u \geq 2t\sum_{i=1}^{t-1}(\delta_i + \gamma_i/2)$,

then for every $v \in \mathcal{N}$ such that

$$v \geq 3t\sum_{i=1}^{t-1}\gamma_i,$$

and $u + v \leq t$, we have

$$E_{\mathbf{P}}\left(M^t_{Q,\mathbf{f}}(\mathbf{x})\right) \leq \sum_{i=1}^{t-1}(2\gamma_i + \delta_i) + \frac{8d'}{t-(u+v)} + e^{-v/6} + e^{-u/6}.$$

Proof Fix a function sequence $\mathbf{f}$ and distribution sequence $\mathbf{P}$, both of length t. Define the volatile set V_1 and the uncertain set U_1 as in Section 3.1. We partition the domain X into four classes, $X_{\mathrm{uv}} = U_1 \cap V_1$, $X_{\bar{\mathrm{u}}\mathrm{v}} = V_1 - U_1$, $X_{\mathrm{u}\bar{\mathrm{v}}} = U_1 - V_1$, and $X_{\bar{\mathrm{u}}\bar{\mathrm{v}}} = X - (U_1 \cup V_1)$. Similarly, X^t is partitioned into 4^t classes, one for each vector $\mathbf{b} \in \{\mathrm{uv}, \bar{\mathrm{u}}\mathrm{v}, \mathrm{u}\bar{\mathrm{v}}, \bar{\mathrm{u}}\bar{\mathrm{v}}\}^t$. A vector $\mathbf{x} \in X^t$ is in the class associated with vector $\mathbf{b} \in \{\mathrm{uv}, \bar{\mathrm{u}}\mathrm{v}, \mathrm{u}\bar{\mathrm{v}}, \bar{\mathrm{u}}\bar{\mathrm{v}}\}^t$ iff each $x_i \in X_{\mathbf{b}_i}$. Let $X^t_{\mathbf{b}}$ denote the set of vectors in the class associated with $\mathbf{b} \in \{\mathrm{uv}, \bar{\mathrm{u}}\mathrm{v}, \mathrm{u}\bar{\mathrm{v}}, \bar{\mathrm{u}}\bar{\mathrm{v}}\}^t$. Clearly the sets $X^t_{\mathbf{b}}$ partition X^t.

We now consider the following subsets of $\{\mathrm{uv}, \bar{\mathrm{u}}\mathrm{v}, \mathrm{u}\bar{\mathrm{v}}, \bar{\mathrm{u}}\bar{\mathrm{v}}\}^t$ whose union is $\{\mathrm{uv}, \bar{\mathrm{u}}\mathrm{v}, \mathrm{u}\bar{\mathrm{v}}, \bar{\mathrm{u}}\bar{\mathrm{v}}\}^t$. In these definitions, the parameters v and u represent thresholds on the acceptable numbers of volatile and uncertain points respectively.

- A is the set of all $\mathbf{b} \in \{\mathrm{uv}, \bar{\mathrm{u}}\mathrm{v}, \mathrm{u}\bar{\mathrm{v}}, \bar{\mathrm{u}}\bar{\mathrm{v}}\}^t$ where
  $b_t = \bar{\mathrm{u}}\bar{\mathrm{v}}$,
  $|\{i \in \{1,\ldots,t\} : b_i = \bar{\mathrm{u}}\mathrm{v} \text{ or } b_i = \mathrm{uv}\}| \leq v$, and
  $|\{i \in \{1,\ldots,t\} : b_i = \mathrm{u}\bar{\mathrm{v}} \text{ or } b_i = \mathrm{uv}\}| \leq u$.
- B is the set of all $\mathbf{b} \in \{\mathrm{uv}, \bar{\mathrm{u}}\mathrm{v}, \mathrm{u}\bar{\mathrm{v}}, \bar{\mathrm{u}}\bar{\mathrm{v}}\}^t$ where
  $|\{i \in \{1,\ldots,t\} : b_i = \bar{\mathrm{u}}\mathrm{v} \text{ or } b_i = \mathrm{uv}\}| > v$.
- C is the set of all $\mathbf{b} \in \{\mathrm{uv}, \bar{\mathrm{u}}\mathrm{v}, \mathrm{u}\bar{\mathrm{v}}, \bar{\mathrm{u}}\bar{\mathrm{v}}\}^t$ where
  $|\{i \in \{1,\ldots,t\} : b_i = \mathrm{u}\bar{\mathrm{v}} \text{ or } b_i = \mathrm{uv}\}| > u$.
- D is the set of all $\mathbf{b} \in \{\mathrm{uv}, \bar{\mathrm{u}}\mathrm{v}, \mathrm{u}\bar{\mathrm{v}}, \bar{\mathrm{u}}\bar{\mathrm{v}}\}^t$ where
  $b_t \neq \bar{\mathrm{u}}\bar{\mathrm{v}}$.

We now have that

$$\int_{X^t} M^t_{Q,\mathbf{f}}(\mathbf{x})dP_1\cdots dP_t = \sum_{\mathbf{b}}\int_{X^t_{\mathbf{b}}} M^t_{Q,\mathbf{f}}(\mathbf{x})dP_1\cdots dP_t$$
$$\leq \sum_{\mathbf{b}\in A}\int_{X^t_{\mathbf{b}}} M^t_{Q,\mathbf{f}}(\mathbf{x})dP_1\cdots dP_t + \sum_{\mathbf{b}\in B}\int_{X^t_{\mathbf{b}}} M^t_{Q,\mathbf{f}}(\mathbf{x})dP_1\cdots dP_t +$$
$$\sum_{\mathbf{b}\in C}\int_{X^t_{\mathbf{b}}} M^t_{Q,\mathbf{f}}(\mathbf{x})dP_1\cdots dP_t + \sum_{\mathbf{b}\in D}\int_{X^t_{\mathbf{b}}} M^t_{Q,\mathbf{f}}(\mathbf{x})dP_1\cdots dP_t.$$

We will bound each of the four terms separately, starting with the sum over $\mathbf{b}\in D$.

$$\begin{aligned}\sum_{\mathbf{b}\in D}\int_{X^t_{\mathbf{b}}} M^t_{Q,\mathbf{f}}(\mathbf{x})dP_1\cdots dP_t &= \int_{X^{t-1}}\left(\int_{V_1\cup U_1} M^t_{Q,\mathbf{f}}(\mathbf{x})dP_t\right)dP_1\cdots dP_{t-1}\\ &\leq \int_{X^{t-1}}\left(\int_{V_1\cup U_1} dP_t\right)dP_1\cdots dP_{t-1}\\ &\leq \int_{X^{t-1}}\sum_{i=1}^{t-1}(2\gamma_i+\delta_i)\,dP_1\cdots dP_{t-1}\\ &= \sum_{i=1}^{t-1}(2\gamma_i+\delta_i),\end{aligned}$$

where the second inequality follows from Lemmas 2 and 3.

Next we examine the sum over $\mathbf{b}\in B$.

$$\sum_{\mathbf{b}\in B}\int_{X^t_{\mathbf{b}}} M^t_{Q,\mathbf{f}}(\mathbf{x})dP_1\cdots dP_t \leq \sum_{\mathbf{b}\in B}\int_{X^t_{\mathbf{b}}} dP_1\cdots dP_t.$$

This sum is the probability under the distribution $\prod_{i=1}^t P_i$ of a sequence in X^t that has more than v components in V_1. Note that this is bounded by the probability that there will be more than v successes in t Bernoulli trials when probability of success is

$$\max_{1\leq i\leq t-1} P_i(V_1) \leq 3/2\sum_{i=1}^{t-1}\gamma_i.$$

Let m be the expected number of successes of these Bernoulli trials. Since

$$v \geq 3t\sum_{i=1}^{t-1}\gamma_i,$$

v is at least twice m, so the probability of more than v successes is no more than $e^{-(v/m-1)^2m/3} = e^{-(v-m)^2/(3m)} \leq e^{-v/6}$ (see for example [8], Inequality (6)).

We use a similar technique to bound the sum over $\mathbf{b}\in C$.

$$\sum_{\mathbf{b}\in C}\int_{X^t_{\mathbf{b}}} M^t_{Q,\mathbf{f}}(\mathbf{x})dP_1\cdots dP_t \leq \sum_{\mathbf{b}\in C}\int_{X^t_{\mathbf{b}}} dP_1\cdots dP_t.$$

Again, this sum is the probability under the distribution $\prod_{i=1}^t P_i$ of a sequence in X^t that has more than u components in U_1. This is no more than the probability of more than u successes in t Bernoulli trials with probability of success no more than

$$\max_{1\leq i\leq t-1} P_i(U_1) \leq \sum_{i=1}^{t-1}(\delta_i+\gamma_i/2).$$

Since u is at least twice the expected number of successes, this probability is bounded by $e^{-u/6}$.

Finally, we consider the sum over $\mathbf{b} \in A$. For each $\mathbf{b} \in A$, let $\Gamma_{\mathbf{b}}$ be the set of permutations of $\{1, \ldots, t\}$ which swap t with some i in $\{1 \ldots, t\}$ that satisfies $b_i = \bar{\mathrm{u}}\bar{\mathrm{v}}$, and leave the other positions unchanged. Thus for $t = 4$ and $\mathbf{b} = (\bar{\mathrm{u}}\bar{\mathrm{v}}, \bar{\mathrm{u}}\mathrm{v}, \bar{\mathrm{u}}\bar{\mathrm{v}}, \bar{\mathrm{u}}\bar{\mathrm{v}})$, the three permutations in $\Gamma_{\mathbf{b}}$ are $(1, 2, 3, 4)$, $(4, 2, 3, 1)$, and $(1, 2, 4, 3)$.

Fix $\mathbf{b} \in A$ and $i \in \{1, \ldots, t-1\}$ for which $b_i = \bar{\mathrm{u}}\bar{\mathrm{v}}$. We have

$$\int_{X_{\mathbf{b}}^t} M_{Q,\mathbf{f}}^t(\mathbf{x}) dP_1 \cdots dP_t =$$
$$\int_{X_{\mathbf{b}}^{t-2}} \left(\int_{X_{\bar{\mathrm{u}}\bar{\mathrm{v}}}} \int_{X_{\bar{\mathrm{u}}\bar{\mathrm{v}}}} M_{Q,\mathbf{f}}^t(\mathbf{x}) dP_t(x_t)\, dP_i(x_i) \right) dP_1 \cdots dP_{i-1} dP_{i+1} \cdots dP_{t-1},$$

where $X_{\mathbf{b}}^{t-2}$ has the obvious meaning. The inner integral satisfies

$$\begin{aligned} \int_{X_{\bar{\mathrm{u}}\bar{\mathrm{v}}}} \int_{X_{\bar{\mathrm{u}}\bar{\mathrm{v}}}} M_{Q,\mathbf{f}}^t(\mathbf{x}) dP_t(x_t)\, dP_i(x_i) &\leq 2 \int_{X_{\bar{\mathrm{u}}\bar{\mathrm{v}}}} \int_{X_{\bar{\mathrm{u}}\bar{\mathrm{v}}}} M_{Q,\mathbf{f}}^t(\mathbf{x}) dP_i(x_t)\, dP_i(x_i) \\ &\leq 4 \int_{X_{\bar{\mathrm{u}}\bar{\mathrm{v}}}} \int_{X_{\bar{\mathrm{u}}\bar{\mathrm{v}}}} M_{Q,\mathbf{f}}^t(\mathbf{x}) dP_i(x_t)\, dP_t(x_i), \end{aligned}$$

from Theorem 1, so

$$\int_{X_{\mathbf{b}}^t} M_{Q,\mathbf{f}}^t(\mathbf{x}) dP_1 \cdots dP_t \leq 4 \int_{X_{\mathbf{b}}^t} M_{Q,\mathbf{f}}^t(\mathbf{x}^\sigma) dP_1 \cdots dP_t$$

for any σ in $\Gamma_{\mathbf{b}}$. It follows that

$$\begin{aligned} &\sum_{\mathbf{b} \in B} \int_{X_{\mathbf{b}}^t} M_{Q,\mathbf{f}}^t(\mathbf{x}) dP_1 \cdots dP_t \\ &\quad \leq 4 \sum_{\mathbf{b} \in B} \frac{1}{|\Gamma_{\mathbf{b}}|} \sum_{\sigma \in \Gamma_{\mathbf{b}}} \int_{X_{\mathbf{b}}^t} M_{Q,\mathbf{f}}^t(\mathbf{x}^\sigma) dP_1 \cdots dP_t \\ &\quad = 4 \sum_{\mathbf{b} \in B} \int_{X_{\mathbf{b}}^t} \left(\frac{1}{|\Gamma_{\mathbf{b}}|} \sum_{\sigma \in \Gamma_{\mathbf{b}}} M_{Q,\mathbf{f}}^t(\mathbf{x}^\sigma) \right) dP_1 \cdots dP_t \\ &\quad \leq 4 \sum_{\mathbf{b} \in B} \int_{X_{\mathbf{b}}^t} \frac{2d'}{t - (u+v)} dP_1 \cdots dP_t \\ &\quad \leq \frac{8d'}{t - (u+v)}. \end{aligned}$$

Adding the bounds on these four sums completes the proof. □

As noted above, if t is large enough so that the total drift, $\sum_{i=1}^{t-1}(2\gamma_i + \delta_i)$, is greater than one, then the above theorem is trivial. However, the algorithm is able to ignore the initial examples and consider only the last k labeled examples (as well as the unlabeled example on which it predicts). In this case we obtain the following corollary.

Corollary 5: *Let Q be the prediction strategy that applies the one-inclusion graph algorithm for the concept class $F^{\oplus u}$ to the last $k+1 \leq t$ examples. For any distribution sequence $\mathbf{P}$ of t distributions on X and function sequence $\mathbf{f}$ of t functions from F, if u and v are positive integers satisfying*

$$u \geq 2k \sum_{i=t-k}^{t-1} (\delta_i + \gamma_i/2),$$

$$v \geq 3k \sum_{i=t-k}^{t-1} \gamma_i,$$

and $u + v \leq k$, *then*

$$E_{\mathbf{P}}\left(M^t_{Q,\mathbf{f}}(\mathbf{x})\right) \leq \sum_{i=t-k}^{t-1} (2\gamma_i + \delta_i) + \frac{8d'}{k+1-(u+v)} + e^{-v/6} + e^{-u/6}.$$

Here d' *is the VC-dimension of* $F^{\oplus u}$, *and is at most* $4.404(d+u)$ *where* d *is the VC-dimension of* F.

4 Upper Bounds

In this section we use Theorem 4 and Corollary 5 from the previous section to bound the expected performance of the one-inclusion graph tracking strategy. This involves assuming bounds on the drift sequences and then finding an appropriate number of exceptions and amount of history to use.

Theorem 6: *Let* F *be a function class on domain* X *and* d *be the VC-dimension of* F. *For* $\hat{\gamma}, \hat{\delta} > 0$, *let* $\alpha = 2\hat{\gamma} + \hat{\delta}$. *If* $\alpha < 1/d$ *and* $d' = d + \lceil 3/2 \ln 1/(d\alpha) \rceil$, *then there is a prediction strategy* Q *such that: for any distribution sequence* $\mathbf{P} = (P_1, \ldots, P_t)$ *on* X *with drift sequence* $\boldsymbol{\gamma}$ *satisfying* $\|\boldsymbol{\gamma}\|_\infty \leq \hat{\gamma}$ *and any function sequence* $\mathbf{f} = (f_1, \ldots, f_t) \in F^t$ *with drift sequence* $\boldsymbol{\delta}$ *satisfying* $\|\boldsymbol{\delta}\|_\infty \leq \hat{\delta}$, *we have*

$$E_{\mathbf{P}}\left(M^t_{Q,\mathbf{f}}\right) \leq \begin{cases} 3\sqrt{d'\alpha} + 126d'/t & \text{if } 8d' < t \leq \sqrt{d'/\alpha} \\ 75\sqrt{d'\alpha} & \text{if } t > \sqrt{d'/\alpha}, \end{cases}$$

and hence

$$E_{\mathbf{P}}\left(\sum_{i=1}^{t} M^i_{Q,\mathbf{f}}\right) \leq \sqrt{\frac{d'}{\alpha}} + 75\sqrt{d'\alpha}t \ .$$

Proof Consider the one-inclusion graph algorithm on $F^{\oplus u}$ where $u = 2d'$. To show that this algorithm meets the bounds of the theorem, we apply Corollary 5 with $v = 2d'$ and $k = \max\{t-1, \lfloor\sqrt{d'/\alpha}\rfloor\}$.

Case 1 : $8d' < t \leq \sqrt{d'/\alpha}$.

For this case, k is set to $t-1$. Note that $u + v = 4d' < t/2$, $2t^2(\hat{\gamma}/2 + \hat{\delta}) \leq u$, and $3t^2\hat{\gamma} \leq v$, so we can apply Corollary 5. This gives

$$\begin{aligned} E_{\mathbf{P}}\left(M^t_{Q,\mathbf{f}}\right) &\leq \sum_{i=1}^{t-1}(2\gamma_i + \delta_i) + \frac{21(d+u)}{t-(u+v)} + e^{-u/6} + e^{-v/6} \\ &\leq t\alpha + \frac{63d'}{t/2} + 2e^{-d'/3} \\ &\leq \sqrt{d'\alpha} + 126d'/t + 2\sqrt{d'\alpha}. \end{aligned}$$

Case 2 : $t > \sqrt{d'/\alpha}$.
Here $k = \left\lfloor \sqrt{d'/\alpha} \right\rfloor$. Assume $\sqrt{d'\alpha} \leq 1/75$, for otherwise the bound is trivial. This implies $\sqrt{d'/\alpha} \geq 75d'$, so

$$\begin{aligned} k - (u+v) &= \left\lfloor \sqrt{d'/\alpha} \right\rfloor - 4d' \\ &\geq (1 - 5/75)\sqrt{d'/\alpha} + 5/75\sqrt{d'/\alpha} - 5d' \\ &\geq 14/15\sqrt{d'/\alpha}. \end{aligned}$$

Also, $2k^2(\hat{\gamma}/2 + \hat{\delta}) \leq u$ and $3k^2\hat{\gamma} \leq v$, so Corollary 5 implies

$$\begin{aligned} E_{\mathbf{P}}\left(M^t_{Q,\mathbf{f}}\right) &\leq \sum_{i=t-k}^{t-1}(2\gamma_i + \delta_i) + \frac{21(d+u)}{k+1-(u+v)} + e^{-u/6} + e^{-v/6} \\ &\leq 3\sqrt{d'\alpha} + \frac{63d'}{14/15\sqrt{d'/\alpha}} \\ &\leq 75\sqrt{d'\alpha}. \end{aligned}$$

The bound on cumulative loss follows immediately. □

The value d' in Theorem 6 is a derived quantity rather than a natural parameter of the problem. However when $\alpha \geq 1/e^d$ then $d' < 5/2d$ and when $\alpha \leq 1/e^d$ then $d' < 5/2 \ln 1/\alpha$. This leads immediately to the following corollary.

Corollary 7: *Under the conditions of Theorem 6, if* $\alpha \leq 1/e^d$ *then*

$$E_{\mathbf{P}}\left(M^t_{Q,\mathbf{f}}\right) \leq 120\sqrt{\alpha \ln(1/\alpha)} \qquad \textit{whenever } t > 1.6\sqrt{\ln(1/\alpha)/\alpha},$$

and if $1/e^d \leq \alpha < 1/d$ *then*

$$E_{\mathbf{P}}\left(M^t_{Q,\mathbf{f}}\right) \leq 120\sqrt{d\alpha} \qquad \textit{whenever } t > 1.6\sqrt{d/\alpha}.$$

Similar results can easily be obtained for the other bounds in Theorem 6.

We now show how bounds on the average drift (as opposed to the maximum drift) can be used to obtain bounds on the one-inclusion graph algorithm's performance. A bound on the average drift gives the algorithm less information since an adversary can "save up" in order to make radical changes in the function and/or distribution.

Theorem 8: *Let* F *be a function class on domain* X *with VC-dimension* $d \geq 1$. *For* $\hat{\gamma}, \hat{\delta} > 0$, *define* $\alpha = 2\hat{\gamma} + \hat{\delta}$.

If $d\alpha < 1$ *and* $d' = d + \lceil \ln(1/d\alpha) \rceil$, *then there is a prediction strategy* Q *such that, for any distribution sequence* $\mathbf{P} = (P_1, \ldots, P_t)$ *with drift sequence* $\boldsymbol{\gamma}$ *satisfying* $\|\boldsymbol{\gamma}\|_1 \leq \hat{\gamma}$ *and any function sequence* $\mathbf{f} = (f_1, \ldots, f_t)$ *with drift sequence* $\boldsymbol{\delta}$ *satisfying* $\|\boldsymbol{\delta}\|_1 \leq \hat{\delta}$, *we have*

$$E\left(\sum_{i=1}^{t} M^i_{Q,\mathbf{f}}\right) \leq \left(d'^2/\alpha\right)^{1/3} + 75t(d'\alpha)^{1/3}.$$

Proof Consider the following algorithm. For the first $(d'^2/\alpha)^{1/3}$ time steps, the algorithm predicts arbitrarily. After that, we use the one-inclusion graph algorithm of Corollary 5 with $k = (d'^2/\alpha)^{1/3}$ and $u = v = 2d'$. Since the constraint on function and distribution drift is not uniform, it is possible that the conditions on u and v of Corollary 5 will be violated at some times. Ignoring the first k time indices, let N_u (N_v) be the number of time indices i for which the condition on u (v) is violated. That is,

$$\begin{aligned} N_u &= \left| \left\{ i \in \{k+1, \ldots, t\} : 2k \sum_{j=i-k}^{i-1} (\gamma_j/2 + \delta_j) > u \right\} \right|, \\ N_v &= \left| \left\{ i \in \{k+1, \ldots, t\} : 3k \sum_{j=i-k}^{i-1} \gamma_j > v \right\} \right|. \end{aligned}$$

Then we have

$$\begin{aligned} N_u u &\leq 2k \sum_{i=k+1}^{t} \sum_{j=i-k}^{i-1} (\gamma_j/2 + \delta_j) \\ &\leq 2k^2 \sum_{j=1}^{t} (\gamma_j/2 + \delta_j) \\ &\leq 2k^2 \alpha t, \end{aligned}$$

and so $N_u \leq (d'\alpha)^{1/3} t$. Similarly, $N_v \leq (d'\alpha)^{1/3} t$. Also, we can assume that $(d'\alpha)^{1/3} < 1/75$, for otherwise the result is trivial. But this implies

$$\begin{aligned} k - (u+v) &= d'/(d'\alpha)^{1/3} - 4d' \\ &\geq 71/75 d'^{2/3}/\alpha^{1/3}. \end{aligned}$$

Applying Corollary 5, we have

$$\begin{aligned} E\left(\sum_{i=1}^{t} M^i_{Q,\mathbf{f}} \right) &\leq k + 2t(d'\alpha)^{1/3} + \\ &\qquad \sum_{i=k}^{t} \left(\sum_{j=i-k}^{i-1} (2\gamma_j + \delta_j) + \frac{21(d+u)}{k-(u+v)} + e^{-u/6} + e^{-v/6} \right) \\ &\leq k + t\left(2(d'\alpha)^{1/3} + k\alpha + \frac{63d'}{71/75 d'^{2/3}/\alpha^{1/3}} + 2(d'\alpha)^{1/3} \right) \\ &\leq k + 75t(d'\alpha)^{1/3}. \end{aligned}$$

□

The bounds of Theorem 8 when given the average drift are clearly weaker than those of Theorem 6 when the maximum drift is bounded. We believe that this gap represents a real difference in difficulty between the two problems. We conjecture that if the total drift is bounded and the actual drift at each time is known to the algorithm then a variant of the one-inclusion graph algorithm has expected cumulative loss $O(t\sqrt{d'\alpha})$ rather than the order $t\sqrt[3]{d'\alpha}$ suggested by Theorem 8. Thus we conjecture that the difference in difficulty between the two problems is not due to the adversary making large changes, but is caused by the algorithm's not knowing when these large changes will occur.

5 Lower Bounds

Here we describe two lower bounds related to the drifting problem. The first lower bound is from Helmbold and Long [10] and uses a fixed distribution so only the target function changes. In the second lower bound, the target is fixed and only the distribution drifts. These lower bounds nearly match (i.e. are within a log factor) of our upper bounds on the one-inclusion tracking strategy, which handles *simultaneous* drifting of functions and distributions.

5.1 Drifting functions

As might be expected, the lower bounds for drifting functions depend greatly on the particular function class. When the function class F has VC dimension d and is suitably rich, $\hat{M}_{Q,F,\gamma}(t) \geq \sqrt{\delta d}/e^2$. However, for the simplest function classes of dimension d, $\hat{M}_{Q,F,\gamma}(t) < \delta d$ for large t.

Say a function class class F of is *d-suitable* if there are d disjoint subsets of the domain, $X_1, \ldots, X_d$ such that

- for each X_j there is an isomorphisms I_j between X_j and the natural numbers, and
- for each $\mathbf{v} \in \mathcal{N}_0^d$ there is a function $f \in F$ such that for each j, function f assigns 1 to the inverse under I_j of $\{i \in \mathcal{N} : i \leq v_j\}$, and assigns 0 to the inverse under I_j of $\{i \in \mathcal{N} : i > v_j\}$.

Thus a d-suitable function class contains d independent "copies" of the natural numbers and an initial segment of each "copy" can be set to one independently.

For example, consider the class of closed sub-intervals of $[0, 1]$. We focus our attention on two infinite sequences of points: $\langle 1/3, 1/6, 1/9, \ldots \rangle$; and $\langle 2/3, 5/6, 8/9, \ldots \rangle$. For any i and j we can find a closed interval (namely $[1/(3i), 1 - 1/(3j)]$) which contains (only) the first i points from the first sequence and (only) the first j points of the second sequence. Therefore this class is 2-suitable.

The following result, due to Helmbold and Long [10], shows that for general function classes the one-inclusion tracking strategy is nearly optimal.

Theorem 9 (Helmbold and Long): *For $d \in \mathcal{N}$, if F is a d-suitable function class on a set X, $\delta < 1/d$, and $t \geq \lfloor \sqrt{d/\delta} \rfloor$, then for any algorithm Q there is a distribution P on X and a function sequence $\mathbf{f} \in F^t$ with drift $\boldsymbol{\delta}$ such that $\|\boldsymbol{\delta}\|_\infty = \delta$ and*

$$EM_{Q,\mathbf{f}} > \frac{\sqrt{\delta d}}{e^2}.$$

However, simpler function classes with VC-dimension d can be much easier to track. Consider the class of all $\{0, 1\}$-valued functions defined on $X = \{1, \ldots, d\}$. Given a function sequence with drift $\boldsymbol{\delta}$ having $\|\boldsymbol{\delta}\|_\infty = \delta$, the function sequence can change value only on those points which have probability at most δ. Therefore the asymptotic error of the algorithm which always predicts with the last seen value for each point is at most $(d-1)\delta$.

Given only that function class F has VC-dimension d, the best lower bound we can prove is $EM^t_{Q,\mathbf{f}} \geq \delta(d-1)(1 - \delta(d-2)/2)/2 \in \Theta(\delta d)$. This bound can be shown by the simple adversary that takes a shattered set of size d and puts probability δ on the first $d-1$ points and probability $1 - (d-1)\delta$ on the last point. The adversary uses the last $d-1$ trials to randomly reset the values of the first $d-1$ points. No algorithm can do better than random guessing on these $d-1$ points (unless the algorithm was fortunate enough to have seen one of them since its value was reset).

The following theorem shows that the one-inclusion tracking strategy is also nearly optimal with respect to cumulative loss. The proof is closely related to the proof of Theorem 9.

Theorem 10: *If F is a d-suitable function class on domain X, $0 < \delta < 1/16$ with $1/\delta$ an integer, and $t \geq 4\left\lceil \sqrt{d/\delta} \right\rceil$, then for any algorithm Q, there is a fixed distribution on X and a function sequence $\mathbf{f} \in F^t$ with drift sequence $\boldsymbol{\delta}$ satisfying $\|\boldsymbol{\delta}\|_\infty \leq \delta$, and*

$$E\left(\sum_{i=1}^{t} M^i_{Q,\mathbf{f}}\right) \geq \frac{e^{-4}}{64}\sqrt{d\delta}t.$$

Proof We will only consider distributions that have support on the set $\bigcup_{j=1}^{d} X_j$. Without loss of generality, we can assume that $X = \{1,\ldots,d\} \times \mathcal{N}$ and F contains all functions $f_{\mathbf{v}} : X \to \{0,1\}$ for $\mathbf{v} \in \mathcal{N}_0^d$, where

$$f_{\mathbf{v}}(a,b) = \begin{cases} 1 & \text{if } b \leq v_a \\ 0 & \text{otherwise.} \end{cases}$$

Define the distribution P on X as $P(a,b) = \delta/d$ for $a \in \{1,\ldots,d\}$ and $b \in \{1,\ldots,1/\delta\}$. As in [10], we will consider a sequence of functions in $\{f_{\mathbf{v}}\}$ for which $\mathbf{v}$ varies slowly. It is best to view the sequence $\{f_{\mathbf{v}}\}$ as a series of $N = \lfloor t/2k \rfloor$ phases of length $2k$ where $k = \lceil d/\delta \rceil$ (the "left over" functions from time $2\lceil d/\delta \rceil N + 1$ through time t are not important to our argument). Each phase is associated with a $\mathbf{z} \in \{0,1\}^d$, and consists of the function sequence

$$\mathbf{f}_{\mathbf{z}} = \left(f_{0\mathbf{z}}, f_{\mathbf{z}}, f_{2\mathbf{z}}, \ldots, f_{k\mathbf{z}}, f_{(k-1)\mathbf{z}}, \ldots, f_{\mathbf{z}}\right).$$

The particular $\{f_{\mathbf{v}}\}$ we analyze is created by choosing a $\mathbf{z}_i$ uniformly at random from $\{0,1\}^d$ for each phase $i = 1,\ldots,N$, and then concatenating the resulting phases, $\mathbf{f}_{\mathbf{z}_1}$ through $\mathbf{f}_{\mathbf{z}_N}$. Thus

$$\mathbf{f} = \left(\mathbf{f}_{\mathbf{z}_1}, \ldots, \mathbf{f}_{\mathbf{z}_N}\right),$$

and has length $t' = 2Nk$. Since $t \geq 4k$, $t' \geq t/2$. Clearly, for all choices of the $\mathbf{z}_i$, the sequence $\mathbf{f}$ has drift $\boldsymbol{\delta}$ satisfying $\|\boldsymbol{\delta}\|_\infty \leq \delta$.

Consider a sequence $(x_1,\ldots,x_{t'})$ chosen from X according to P. We relabel the sequence to emphasize the phases:

$$\left(x_{1,0},\ldots,x_{1,2k-1},x_{2,0},\ldots,x_{2,2k-1},\ldots,x_{N,2k-1}\right),$$

and relabel $M^l_{Q,\mathbf{f}}$ as $M^{i,j}_{Q,\mathbf{f}}$ in the corresponding way.

Consider for a moment an example $x_{i,j} = (a,b)$ where $b \leq j \leq k$. This example is labeled 1 by $\mathbf{f}$ if and only if bit a of $\mathbf{z}_i$ is 1. Furthermore, the algorithm can do no better than random guessing on the label of $x_{i,j}$ unless there was an earlier example in this phase which has revealed bit a of $\mathbf{z}_i$, namely some $x_{i,j'} = (a,b')$ with $b' \leq j' < j$ The remainder of the proof formalizes this intuition.

For each phase $i \in \{1,\ldots,N\}$ and each $j \in \{\lfloor k/2 \rfloor,\ldots,k\}$, we have

$$E\left(M^{i,j}_{Q,\mathbf{f}}\right) \geq E\left(M^{i,j}_{Q,\mathbf{f}} | E_{i,j}\right)\Pr(E_{i,j}),$$

where the expectations and probability are over random choices of the $\mathbf{z}_i$ and $\mathbf{x}$ sequences, and $E_{i,j}$ is the event that some bit of $\mathbf{z}_i$ is revealed by example $x_{i,j}$. More formally

$$\begin{aligned} E_{i,j} &= \Big\{\mathbf{x} \in X^{N\times 2k} : x_{ij} = (a,b), \text{ with } b \le j \text{ and} \\ &\qquad x_{ij'} \notin \{(a,c) : c \le j'\} \text{ for } j' \in \{0,\ldots,j-1\}\Big\}. \end{aligned}$$

Clearly,

$$E\left(M^{i,j}_{Q,\mathbf{f}}\right) \ge \frac{1}{2}\Pr(E_{i,j}) = \frac{j\delta}{2}\prod_{l=1}^{j-1}(1-l\delta/d) \ge \frac{j\delta}{2}\exp(-j^2\delta/d) \ge \frac{e^{-4}}{8}\sqrt{d\delta}.$$

So

$$E\left(\sum_{j=1}^{t} M^{j}_{Q,\mathbf{f}}\right) \ge \frac{e^{-4}}{8}\sqrt{d\delta}\frac{t'}{4} \ge \frac{e^{-4}}{64}\sqrt{d\delta}t.$$

□

5.2 Drifting distributions

The two results in this section show that both the instantaneous and cumulative loss of the one-inclusion tracking strategy are also nearly optimal with respect to distribution drift.

Theorem 11: *Suppose F is a function class with VC-dimension d such that $3 \le d < \infty$. For any $t \in \mathcal{N}$ and $\gamma > 0$, and any prediction strategy Q, there is a function f in F and a distribution sequence $\mathbf{P}$ with drift sequence $\boldsymbol{\gamma}$ satisfying $\|\boldsymbol{\gamma}\|_\infty \le \gamma$ such that*

$$E\left(M^{t}_{Q,\mathbf{f}}\right) \ge \begin{cases} \dfrac{d-1}{2et} & \text{for all } t \\ \dfrac{\sqrt{\gamma(d-2)}}{4e} & \text{for } t > \sqrt{\dfrac{d-2}{\gamma}}. \end{cases}$$

Proof The first part of the bound is a consequence of the lower bound on the mistake probability for identically distributed examples given in Theorem 3.1 of [9].

The second part of the bound uses a related 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 $\mathbf{P} = (P_1, \ldots, P_t)$ 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

$$\begin{aligned} P_j(z) &= \begin{cases} \dfrac{k}{m} & j = 1,\ldots,t-m \\ \dfrac{(t-j)k}{m^2} & j = t-m+1,\ldots,t \end{cases} \\ P_j(y_0) &= 1 - \frac{k}{m} \\ P_j(y_i) &= \begin{cases} 0 & j = 1,\ldots,t-m \\ \dfrac{j-(t-m)}{m^2} & j = t-m+1,\ldots,t \end{cases} \end{aligned}$$

It is easy to verify that the drift $\boldsymbol{\gamma}$ of the sequence $\mathbf{P}$ satisfies $\|\boldsymbol{\gamma}\|_\infty \leq \max_j |P_{j+1}(X_1) - P_j(X_1)| \leq k/m^2 < \gamma$, where $X_1 = \{y_1, \ldots, y_k\}$.

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

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

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

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

Using a standard argument (see [7, 9]), it is easy to show that there is an f in F for which $E_{\mathbf{P}}\left(M^t_{Q,\mathbf{f}}\right) \geq \mathbf{P}(B)/2$. □

Theorem 12: *Suppose $d \in \mathcal{N}$, F is a d-suitable function class, $0 < \gamma < 1/4$, and $t \geq 2\sqrt{d/\gamma}$. Then for any algorithm Q, there is a function f in F and a distribution sequence $\mathbf{P}$ with drift $\boldsymbol{\gamma}$ satisfying $\|\boldsymbol{\gamma}\|_\infty \leq \gamma$, and*

$$E\left(\sum_{i=1}^{t} M^i_{Q,\mathbf{f}}\right) \geq \frac{(1 - e^{-1/4})}{4} \sqrt{d\gamma}\, t,$$

Proof We will consider a sequence of distributions that each have support contained in a finite set. Without loss of generality, we may assume that $X = \{1, \ldots, d\} \times [0, 1]$. (This is because we consider only a finite subset of $[0, 1]$; the cardinality of this subset depends on t.) Let $F = \left\{f_{\mathbf{v}} : \mathbf{v} \in [0, 1]^d\right\}$, where

$$f_{\mathbf{v}}(a, b) = \begin{cases} 1 & \text{if } b \geq v_a \\ 0 & \text{otherwise.} \end{cases}$$

Let the target function be $f_{\mathbf{v}}$, where $\mathbf{v}$ is chosen randomly according to the uniform distribution on $[0, 1]^d$. The distribution sequence we will consider begins with $P_1(i, 0) = 1/d$ for $i \in \{1, \ldots, d\}$. The sequence is split into a number of trials, and in trial r the support gradually shifts to points (i, b^r_i) $(i \in \{1, \ldots, d\})$. These points are chosen so that the b^r_i's

approximate the threshold $\mathbf{v}$ progressively more accurately, so it is difficult to predict the label of an unseen point (i, b_i^r). For $r \in \mathcal{N}$ and $i \in \{1, \ldots, d\}$, let

$$b_i^r = \left\lfloor \frac{v_i}{2^{-r}} \right\rfloor 2^{-r} + 2^{-r-1}.$$

From this definition, $f_{\mathbf{v}}(i, b_i^r)$ is the value of the r-th bit of the binary representation of v_i. Since $\mathbf{v}$ is chosen uniformly, if this label has not previously been seen, any algorithm has probability $1/2$ of predicting it correctly.

Now, let $k = \left\lfloor \sqrt{d/\gamma} \right\rfloor$ and $t' = Nk$, where $N = \lfloor t/k \rfloor$. Since $t \geq 2\sqrt{d/\gamma}$, $t' \geq t/2$. The sequence of distributions is split into N trials, each of length k. During trial r, the probability of each point (i, b_i^r) is increased by γ/d, for $i = 1, \ldots, d$. This distribution sequence clearly has drift $\boldsymbol{\gamma}$ satisfying $\|\boldsymbol{\gamma}\|_\infty \leq \gamma$. Consider a balance point (i, b_i^r). The probability that it remains unseen during trial r is no more than

$$\prod_{i=1}^{k} (1 - i\gamma/d) \leq \prod_{i=1}^{k} \exp(-\gamma i/d) \leq \exp(-1/4).$$

So the expected number of balance points $(1, b_1^r), \ldots, (d, b_d^r)$ that are seen in trial r is at least $(1 - e^{-1/4})d$. The first time each balance point occurs, the algorithm has probability $1/2$ of making a mistake. It follows that

$$\begin{aligned} E\left(\sum_{i=1}^{t} M_{Q,\mathbf{f}}^i\right) &\geq \frac{(1 - e^{-1/4})d}{2k} t' \\ &\geq \frac{(1 - e^{-1/4})}{4} \sqrt{d\gamma}\, t, \end{aligned}$$

where the expectation is over random examples and random choice of the function $f_{\mathbf{v}}$. It follows that there exists a suitable f in F. □

6 Agnostic Learning

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, Vapnik [15] and Blumer *et al.* [4] consider learning models in which the relationship is described by a joint probability distribution on $X \times \{0, 1\}$. In this section, we examine a learning model of this kind in which the joint distribution is allowed to change slowly but continually as learning proceeds.

We begin with some notation, analogous to that introduced in Section 2. Suppose that $t \geq 1$, $\xi = ((x_1, y_1), \ldots, (x_t, y_t)) \in (X \times \{0, 1\})^t$ is a labeled sample, and Q is a deterministic prediction strategy. Define the mistake of Q on ξ as

$$M_Q^t(\xi) = \begin{cases} 1 & Q\left(((x_1, y_1), \ldots, (x_{t-1}, y_{t-1})), x_t\right) \neq y_t \\ 0 & \text{otherwise,} \end{cases}$$

and define $M_Q^t(\xi)$ for a randomized prediction strategy $Q = (Q_r, Z, D)$ as

$$M_Q^t(\xi) = D\left\{z \in Z : Q\left(((x_1, y_1), \ldots, (x_{t-1}, y_{t-1})), x_t, z\right) \neq y_t\right\}.$$

Suppose $(S, \mathcal{F}, P_i)$ is a probability space for $i = 1, 2, \ldots, t$, $t > 0$. Define the drift $\boldsymbol{\gamma}$ of the sequence $\mathbf{P} = (P_1, \ldots, P_t)$ as in Section 2.

For a function $f : X \to \{0, 1\}$ and a distribution P on $X \times \{0, 1\}$, define the error of f with respect to P as

$$\mathbf{er}_P(f) = P\{(x, y) : f(x) \neq y\}.$$

For a class F of $\{0, 1\}$-valued functions defined on X, we are interested in the additional instantaneous loss of a prediction strategy Q,

$$E\left(M_Q^t\right) - \inf_{f \in F} \mathbf{er}_{P_t}(f).$$

We first introduce a technical lemma showing that the expectation over any slowly changing distribution sequence is close to the expectation over related sequences of an unchanging distribution.

Lemma 13: *For any $k \geq 1$ and $0 \leq \gamma \leq 1$, if the sequence $\mathbf{P} = (P_1, \ldots, P_k)$ of distributions on X has drift $\boldsymbol{\gamma}$ satisfying $\|\boldsymbol{\gamma}\|_\infty \leq \gamma$ and f is a measurable function from X to $[0, 1]$, then*

$$\int f \, dP_i \leq \int f \, dP_{i+1} + \gamma/2,$$

for $1 \leq i \leq k - 1$. Furthermore, if f is a measurable function from X^k to $[0, 1]$, then

$$\int f \, d\mathbf{P} \leq \int f \, dP_1^k + \frac{k(k-1)}{4}\gamma, \tag{6.1}$$

and

$$\int f \, d\mathbf{P} \leq \int f \, dP_k^k + \frac{k(k-1)}{4}\gamma, \tag{6.2}$$

Proof For the first inequality, define the signed measure $\mu = P_i - P_{i+1}$. 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 the measurable space $(X, \mathcal{F})$ (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, $\mu^+, \mu^- \geq 0$, $\mu = \mu^+ - \mu^-$, and $\mu^+(X) = \mu^-(X) = d_{TV}(P_i, P_{i+1})/2$. By definition,

$$\begin{aligned}
\left|\int f \, dP_i - \int f \, dP_{i+1}\right| &= \left|\int f \, d\mu\right| \\
&= \left|\int f \, d\mu^+ - \int_X f \, d\mu^-\right| \\
&\leq \max\{\mu^+(X), \mu^-(X)\} \leq \gamma/2,
\end{aligned}$$

which is the first inequality.

Now, we are interested in the expectation

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

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 $0 \leq I \leq 1$, so the first inequality gives

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

Therefore

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

Similarly,

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

and

$$\begin{aligned} E_{\langle P_i \rangle_{i=1}^k}(f) &\leq \int_{X^k} f \, dP_1^k(x_1, x_2, \ldots, x_k) + \gamma/2 \sum_{i=1}^{k-1} i \\ &= E_{P_1^k}(f) + \frac{k(k-1)}{4}\gamma, \end{aligned}$$

which is Inequality (6.1). The same argument with the labels for $P_1 \ldots P_k$ reversed gives Inequality (6.2). □

The following theorem is the main result of this section.

Theorem 14: *For any function class F with VC-dimension d ($1 \leq d < \infty$), and any distribution sequence $\mathbf{P} = (P_1, \ldots, P_t)$ with drift $\boldsymbol{\gamma}$ satisfying $\|\gamma\|_\infty \leq \gamma \leq 1/d^2$, there is a prediction strategy Q with*

$$E_{\mathbf{P}}\left(M_Q^t\right) \leq \inf_{f \in F} \mathbf{er}_{P_t}(f) + \left(15\sqrt{\frac{d}{t-1}} + 5d^{2/5}\gamma^{1/5}\right)\left(\ln(2d^{-2/5}\gamma^{-1/5})\right)^{1/2}$$

if $2d \leq t \leq 2d^{1/5}\gamma^{-2/5}$, and

$$E_{\mathbf{P}}\left(M_Q^t\right) \leq \inf_{f \in F} \mathbf{er}_{P_t}(f) + 19d^{2/5}\gamma^{1/5}\left(\ln\left(2d^{-2/5}\gamma^{-1/5}\right)\right)^{1/2}$$

if $t > 2d^{1/5}\gamma^{-2/5}$.

The proof uses the following lemma, which uses a uniform convergence result of Vapnik [15].

Lemma 15: *Let F be a class of functions that map from X to $\{0, 1\}$, with VC-dimension $d \geq 1$. There is a prediction strategy Q such that, for any probability distribution P on $X \times \{0, 1\}$ and any $t \geq 2d$,*

$$E_{P^t}\left(M_Q^t\right) - \inf_{f \in F} \mathbf{er}_P(f) < 10\left(\frac{d}{t-1}\ln\frac{2(t-1)}{d}\right)^{1/2}.$$

Proof If $\xi = (x_1, y_1, \ldots, x_m, y_m) \in S^m$ is a labeled sample and f is a function in F, define the empirical error of f on ξ as the fraction of examples in ξ that f misclassifies,

$$\widehat{\mathbf{er}}_\xi(f) = \frac{1}{m} \left| \{ i \in \{1, 2, \ldots, m\} : f(x_i) \neq y_i \} \right|.$$

Let Q be the prediction strategy that, on input (ξ, x), labels x according to a function $\hat{f}_\xi$ in F that minimizes the empirical error on ξ. That is, $Q(\xi, x) = \hat{f}_\xi(x)$, where $\hat{f}_\xi$ satisfies $\widehat{\mathbf{er}}_\xi\left(\hat{f}_\xi\right) = \min_{f \in F} \widehat{\mathbf{er}}_\xi(f)$.

A result of Vapnik gives bounds on the sample size that ensures the error of a function and its empirical error are close for all functions in the class F. Indeed, if $\mathrm{VCdim}(F) = d$ and $m \geq d$, Theorem 6.7 in [15] implies that

$$P^m \left\{ \xi \in S^m : |\mathbf{er}_P(f) - \widehat{\mathbf{er}}_\xi(f)| > \epsilon \right\} < 9e^{-\epsilon^2 m/4} \left(\frac{2m}{d} \right)^{3d}$$

for all f in F. If $\psi(\epsilon, m) = 9e^{-\epsilon^2 m/4} (2m/d)^{3d}$, it follows that

$$P^m \left\{ \xi \in S^m : \widehat{\mathbf{er}}_\xi\left(\hat{f}_\xi\right) < \mathbf{er}_P\left(\hat{f}_\xi\right) - \epsilon/2 \right\} < \psi(\epsilon/2, m),$$

and

$$P^m \left\{ \xi : \exists f \in F, \mathbf{er}_P(f) = \inf_{f \in F} \mathbf{er}_P(f),\ \widehat{\mathbf{er}}_\xi(f) > \mathbf{er}_P(f) + \epsilon/2 \right\} < \psi(\epsilon/2, m).$$

If neither of these events occur, we must have $\mathbf{er}_P\left(\hat{f}_\xi\right) \leq \inf_{f \in F} \mathbf{er}_P(f) + \epsilon$, so that

$$P^m \left\{ \xi \in S^m : \mathbf{er}_P\left(\hat{f}_\xi\right) - \inf_{f \in F} \mathbf{er}_P(f) > \epsilon \right\} < 2\psi(\epsilon/2, m),$$

hence

$$P^m \left\{ \xi \in S^m : \left(\mathbf{er}_P\left(\hat{f}_\xi\right) - \inf_{f \in F} \mathbf{er}_P(f) \right)^2 > \epsilon^2 \right\} < 18e^{-\epsilon^2 m/16} (2m/d)^{3d}.$$

Since $\mathbf{er}_P(f) \leq 1$ for all f in F,

$$E_{\xi \in P^m} \left(\mathbf{er}_P(\hat{f}_\xi) - \inf_{f \in F} \mathbf{er}_P(f) \right)^2 \leq \epsilon^2 + 18e^{-\epsilon^2 m/16} (2m/d)^{3d},$$

for all $\epsilon > 0$. Setting

$$\epsilon^2 = \frac{16}{t-1} \left(\ln \frac{t}{d} + 3d \ln \frac{2(t-1)}{d} \right)$$

and $m = t - 1$ gives

$$\begin{aligned} E_{\xi \in P^{t-1}} \left(\mathbf{er}_P(\hat{f}_\xi) - \inf_{f \in F} \mathbf{er}_P(f) \right)^2 &\leq \frac{16}{t-1} \left(\ln \frac{t}{d} + 3d \ln \frac{2(t-1)}{d} \right) + \frac{18d}{t} \\ &< \left(16 + 48 + \frac{18}{\ln 2} \right) \frac{d}{t-1} \ln \frac{2(t-1)}{d} \end{aligned}$$

provided $t \geq 2d$. Applying Jensen's inequality gives

$$E_{\xi \in P^{t-1}} \left(\mathbf{er}_P(\hat{f}_\xi) - \inf_{f \in F} \mathbf{er}_P(f) \right) < 10 \left(\frac{d}{t-1} \ln \frac{2(t-1)}{d} \right)^{1/2},$$

That is,

$$E_{\xi \in P^t} \left(M_Q^t(\xi) \right) - \inf_{f \in F} \mathbf{er}_P(f) < 10 \left(\frac{d}{t-1} \ln \frac{2(t-1)}{d} \right)^{1/2}.$$

□

Proof (of Theorem 14) The algorithm we analyze is the prediction strategy Q_k that minimizes the empirical error over the last $k = \lceil 2d^{1/5}\gamma^{-2/5} \rceil$ time steps if possible. In other words, if $t \geq k$ then

$$Q_k\left((x_1, y_1, \ldots, x_{t-1}, y_{t-1}), x_t\right) = Q\left((x_{t-k+1}, y_{t-k+1}, \ldots, x_{t-1}, y_{t-1}), x_t\right)$$

where Q is the prediction strategy of Lemma 15. When $t \leq k$, strategy Q_k is identical to Q.

We start with the second bound. Lemma 13 implies that, for all γ-admissible distribution sequences $\langle P_i \rangle_{i=1}^t$ on S,

$$E_{\langle P_i \rangle_{i=t-k+1}^t} \left(M_{Q_k}^k \right) \leq E_{P_t^k} \left(M_{Q_k}^k \right) + \frac{k(k-1)}{2}\gamma$$

when $t \geq k$.

Then

$$\begin{aligned}
L_{F,\gamma}(t) \leq L_{Q_k,F,\gamma}(t) &\leq E_{P_t^k} \left(M_Q^k \right) - \inf_{f \in F} \mathbf{er}_{P_t}(f) + \frac{k(k-1)}{4}\gamma \\
&\leq 10 \left(\frac{d}{k-1} \ln \frac{2(k-1)}{d} \right)^{1/2} + (k-1)^2\gamma \\
&\leq 10 \left(\frac{d}{d^{1/5}\gamma^{-2/5}} \ln \frac{4d^{1/5}\gamma^{-2/5}}{d} \right)^{1/2} + (2d^{1/5}\gamma^{-2/5})^2\gamma \\
&\leq 10 d^{2/5}\gamma^{1/5} \left(2 \ln(2d^{2/5}\gamma^{-1/5}) \right)^{1/2} + 4d^{2/5}\gamma^{1/5} \\
&\leq 19 d^{2/5}\gamma^{1/5} \left(\ln(2d^{2/5}\gamma^{-1/5}) \right)^{1/2},
\end{aligned}$$

giving the second bound.

For the first bound, $2d \leq t < k$. In this case Lemma 13 and the same reasoning gives

$$L_{F,\gamma}(t) \leq 10 \left(\frac{d}{t-1} \ln \frac{2(t-1)}{d} \right)^{1/2} + (t-1)^2\gamma.$$

Substituting in the upper bound $t < 2d^{1/5}\gamma^{-2/5}$ yields

$$\begin{aligned}
L_{F,\gamma}(t) &\leq 10 \left(\frac{d}{t-1} \ln(4d^{-4/5}\gamma^{-2/5}) \right)^{1/2} + 4d^{2/5}\gamma^{1/5} \\
&\leq \left(15\sqrt{\frac{d}{t-1}} + 5d^{2/5}\gamma^{1/5} \right) \left(\ln(2d^{-2/5}\gamma^{-1/5}) \right)^{1/2},
\end{aligned}$$

which is the first bound in the theorem. □

7 Conclusions

We have analyzed how well an algorithm can learn when the target function and distribution are slowly changing at known rates. In particular we give an algorithm based on the 1-inclusion graph algorithm of Haussler *et al.* for this changing setting. For a target class of VC-dimension d and a combined drift rate α of the distribution and target function at least $1/e^{\Omega(d)}$, this algorithm's mistake rate is $O(\sqrt{d\alpha})$. Thus if the drift rate is at most ϵ^2/d then the algorithm's mistake rate is at most ϵ. For very small drift rates, our bound on the algorithm's mistake rate is slightly weaker – $O(\sqrt{d\alpha \log(1/\alpha)})$.

We have (almost) matching lower bounds. We show that when the distribution is held constant but the target function drifts at rate δ then any algorithm learning a suitably rich target class has a mistake rate in $\Omega(\sqrt{d\delta})$. Furthermore, even when the target function remains fixed and only the distribution drifts (at rate γ), any algorithm learning a class of VC-dimension d can be forced to make mistakes at a rate in $\Omega(\sqrt{d\gamma})$. Thus our algorithm's performance when both the function and distribution are drifting is almost the same as the best possible algorithm when only one of the two is drifting.

We have also examined the situation where the drift is uneven – although the total amount of drift is bounded, it may be distributed unevenly over time. In this case the algorithm's mistake rate climbs to $O(\sqrt[3]{d\alpha})$ where α is the average amount of drift. Although we conjecture that knowing how much drift occurred each time would allow a $O(\sqrt{d\alpha})$ mistake rate, this remains an open problem.

Finally, we consider the generalization of drifting to the agnostic model. Here instead of having a distribution over the domain X and examples labeled by a target function, there is a joint distribution over $X \times \{0,1\}$. In this setting the algorithm's goal is to select at each time a hypothesis from the class whose expected error is as small as possible. We give an algorithm whose expected loss (difference in error between the selected hypothesis and the best possible function in the class at that time) is bounded by $O(d^{2/5}\gamma^{1/5}\sqrt{2d^{-2/5}\gamma^{-1/5}})$.

Appendix

Proof of Lemma 1

Proof By the Lebesgue decomposition theorem (see for example [6]), there is a set $T \in \mathcal{S}$ such that the finite measure $P_{|T}$ (defined by $P_{|T}(S) = P(T \cap S)$ for all $S \in \mathcal{S}$) is absolutely continuous with respect to Q (that is, for all $S \in \mathcal{S}$, $Q(S) = 0$ implies $P_{|T}(S) = 0$) and $Q(X - T) = 0$. So there is a function $\frac{dP}{dQ} : T \to \mathcal{R}$ (the Radon-Nikodym derivative) such that, for all $S \in \mathcal{S}$,

$$P_{|T}(S) = \int_{S \cap T} \frac{dP}{dQ}(x)\, dQ(x).$$

Let $W = \left\{ x \in T : \frac{dP}{dQ}(x) > \alpha \right\}$ and $E = (X - T) \cup W$. We will show that E satisfies the conditions of the theorem. For any $S \in \mathcal{S}$ we have $P(S \cap E) = P(S \cap (X - T)) + P(S \cap W)$. But $P(S \cap (X - T)) \geq Q(S \cap (X - T)) = 0$, and

$$P(S \cap W) = \int_{S \cap W} \frac{dP}{dQ}(x)\, dQ(x).$$

Now, if $Q(S \cap W) = Q(S \cap E) \neq 0$, then $P(S \cap W) > \alpha Q(S \cap W)$, so $P(S \cap E) > \alpha Q(S \cap E)$. That is, either $Q(S \cap E) = 0$ or $P(S \cap E) > \alpha Q(S \cap E)$.

Now, $X - E = T - W$, so

$$\begin{aligned} P(S \cap (X - E)) &= P(S \cap (T - W)) \\ &= \int_{S \cap (T-W)} \frac{dP}{dQ}(x)\, dQ(x) \\ &\leq \alpha Q(S \cap (X - E)). \end{aligned}$$

That is, E satisfies conditions 1 and 2. Furthermore, the set T is unique up to measure zero symmetric differences, which implies that E is also essentially unique. If we extend the function $\frac{dP}{dQ}$ to the whole of X by assigning it a value of 0 on $X - T$, then that function is essentially (with respect to P and Q) unique.

Now, if $f : X \to \mathcal{R}$ is a non-negative measurable function, we can write

$$\begin{aligned} \int_{S \cap (X-E)} f\, dP &= \int_{S \cap (T-W)} f(x) \frac{dP}{dQ}\, dQ(x) \\ &\leq \alpha \int_{S \cap (T-W)} f(x)\, dQ(x). \end{aligned}$$

□

Proof of Lemma 2

Recall that Lemma 2 says:

If $\mathbf{P} = (P_1, \ldots, P_t)$ is a sequence of t probability distributions on $(X, \mathcal{S})$, with distribution drift $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{t-1})$, then for all $1 \leq k \leq i \leq t$

$$P_i(V_k) \leq 3/2 \sum_{j=k}^{t-1} \gamma_j.$$

We first give some preliminary definitions and a lemma.

For $1 \leq k \leq t$ and $t - k \leq i \leq t - 1$, let $T_i = H_i - \bigcup_{i<j<t} E_j^+$ and $W_{i,k} = L_i - \bigcup_{i<j<t} E_j^- - V_k^+$. Thus the T_i and $W_{i,k}$ for $t - k \leq i \leq t - 1$ partition V_k. Furthermore, for all $t - k \leq i \leq t - 1$, $P_i(T_i) \geq 2P_t(T_i)$ and $P_i(W_{i,k}) \leq \frac{1}{2} P_t(W_{i,k})$.

Lemma 16: *For all* $1 \leq k \leq t$,

$$\sum_{i=t-k}^{t-1} P_t(T_i) \leq \sum_{j=t-k}^{t-1} \gamma_j/2,$$

and

$$\sum_{i=t-k}^{t-1} P_i(W_{i,k}) \leq \sum_{j=t-k}^{t-1} \gamma_j/2.$$

Proof For the first inequality we have

$$\begin{aligned} 2 \sum_{i=t-k}^{t-1} P_t(T_i) &\leq \sum_{i=t-k}^{t-1} P_i(T_i) \\ \Leftrightarrow P_t(V_k^+) &\leq \sum_{i=t-k}^{t-1} [P_i(T_i) - P_t(T_i)] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=t-k}^{t-1}\sum_{j=i}^{t-1}[P_j(T_i) - P_{j+1}(T_i)] \\
&= \sum_{j=t-k}^{t-1}\sum_{i=t-k}^{j}[P_j(T_i) - P_{j+1}(T_i)] \\
&= \sum_{j=t-k}^{t-1}\left[P_j\left(\bigcup_{k\le i\le j} T_i\right) - P_{j+1}\left(\bigcup_{k\le i\le j} T_i\right)\right] \\
&\le \sum_{j=t-k}^{t-1}\gamma_j/2.
\end{aligned}
$$

For the second inequality we have

$$
\begin{aligned}
2\sum_{i=t-k}^{t-1} P_i(W_{i,k}) &\le \sum_{i=t-k}^{t-1} P_t(W_{i,k}) \\
\Leftrightarrow \sum_{i=t-k}^{t-1} P_i(W_{i,k}) &\le \sum_{i=t-k}^{t-1}[P_t(W_{i,k}) - P_i(W_{i,k})] \\
&= \sum_{i=t-k}^{t-1}\sum_{j=i}^{t-1}[P_{j+1}(W_{i,k}) - P_j(W_{i,k})] \\
&= \sum_{j=t-k}^{t-1}\sum_{i=t-k}^{j}[P_{j+1}(W_{i,k}) - P_j(W_{i,k})] \\
&= \sum_{j=t-k}^{t-1}\left[P_{j+1}\left(\bigcup_{k\le i\le j} W_{i,k}\right) - P_j\left(\bigcup_{k\le i\le j} W_{i,k}\right)\right] \\
&\le \sum_{j=t-k}^{t-1}\gamma_j/2.
\end{aligned}
$$

□

We now return to the proof of Lemma 2.

Proof For each i in $\{t-k,\ldots,t\}$, $P_i(V_k) = \sum_{j=t-k}^{t-1}[P_i(T_j) + P_i(W_{j,k})]$. From Lemma 16 we get the following inequality.

$$
\begin{aligned}
P_i(V_k) &\le \sum_{j=t-k}^{t-1}[P_i(T_j) + P_i(W_{j,k})] - \sum_{j=t-k}^{t-1} P_t(T_j) - \sum_{j=t-k}^{t-1} P_j(W_{j,k}) + \\
&\qquad \sum_{j=t-k}^{t-1}\gamma_j \\
&= \sum_{j=t-k}^{t-1}[P_i(T_j) + P_i(W_{j,k}) - P_t(T_j) - P_j(W_{j,k})] + \sum_{j=t-k}^{t-1}\gamma_j \\
&= \sum_{j=t-k}^{t-1}[P_i(T_j) - P_t(T_j)] + \sum_{j=t-k}^{i-1}[P_i(W_{j,k}) - P_j(W_{j,k})] + \\
&\qquad \sum_{j=i+1}^{t-1}[P_i(W_{j,k}) - P_j(W_{j,k})] + \sum_{j=t-k}^{t-1}\gamma_j
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{j=t-k}^{t-1}\sum_{l=i}^{t-1}[P_l(T_j)-P_{l+1}(T_j)] + \sum_{j=t-k}^{i-1}\sum_{l=j}^{i-1}[P_{l+1}(W_{j,k})-P_l(W_{j,k})] \\
&\quad + \sum_{j=i+1}^{t-1}\sum_{l=i}^{j-1}[P_l(W_{j,k})-P_{l+1}(W_{j,k})] + \sum_{j=t-k}^{t-1}\gamma_j \\
&= \sum_{l=i}^{t-1}\sum_{j=t-k}^{t-1}[P_l(T_j)-P_{l+1}(T_j)] + \sum_{l=t-k}^{i-1}\sum_{j=t-k}^{l}[P_{l+1}(W_{j,k})-P_l(W_{j,k})] \\
&\quad + \sum_{l=i}^{t-2}\sum_{j=l+1}^{t-1}[P_l(W_{j,k})-P_{l+1}(W_{j,k})] + \sum_{j=t-k}^{t-1}\gamma_j \\
&= \sum_{l=i}^{t-1}[P_l(V_k^+)-P_{l+1}(V_k^+)] + \\
&\qquad \sum_{l=t-k}^{i-1}\left[P_{l+1}\left(\bigcup_{t-k\le j\le l} W_{j,k}\right) - P_l\left(\bigcup_{t-k\le j\le l} W_{j,k}\right)\right] \\
&\quad + \sum_{l=i}^{t-2}\left[P_l\left(\bigcup_{l+1\le j\le t} W_{j,k}\right) - P_{l+1}\left(\bigcup_{l+1\le j\le t} W_{j,k}\right)\right] + \sum_{j=t-k}^{t-1}\gamma_j \\
&\le P_{t-1}(V_k^+) - P_t(V_k^+) + \\
&\quad \sum_{l=i}^{t-2}\left[P_l\left(V_k^+ \cup \bigcup_{l+1\le j\le t-1} W_{j,k}\right) - P_{l+1}\left(V_k^+ \cup \bigcup_{l+1\le j\le t-1} W_{j,k}\right)\right] \\
&\quad + \sum_{l=t-k}^{i-1}\gamma_l/2 + \sum_{j=t-k}^{t-1}\gamma_j \\
&\le \gamma_{t-1}/2 + \sum_{l=i}^{t-2}\gamma_l/2 + + \sum_{l=t-k}^{i-1}\gamma_l/2 + \sum_{j=t-k}^{t-1}\gamma_j \\
&= 3/2\sum_{l=t-k}^{t-1}\gamma_l.
\end{aligned}
$$

□

Proof of Lemma 3

Proof

$$
\begin{aligned}
P_i(U_{t-k}) &= \sum_{j=t-k}^{t-1} P_i(D_j) \\
&\le \sum_{j=t-k}^{t-1} P_i(D_j) - \sum_{j=t-k}^{t-1} P_j(D_j) + \sum_{j=t-k}^{t-1}\delta_j \\
&= \sum_{j=t-k}^{i-1}[P_i(D_j)-P_j(D_j)] + \sum_{j=i+1}^{t-1}[P_i(D_j)-P_j(D_j)] + \sum_{j=t-k}^{t-1}\delta_j \\
&= \sum_{j=t-k}^{i-1}\sum_{l=j}^{i-1}[P_{l+1}(D_j)-P_l(D_j)] +
\end{aligned}
$$

$$
\begin{aligned}
& \sum_{j=i+1}^{t-1} \sum_{l=i+1}^{j} [P_{l-1}(D_j) - P_l(D_j)] + \sum_{j=t-k}^{t-1} \delta_j \\
= & \sum_{l=t-k}^{i-1} \sum_{j=t-k}^{l} [P_{l+1}(D_j) - P_l(D_j)] + \\
& \sum_{l=i+1}^{t-1} \sum_{j=l}^{t-1} [P_{l-1}(D_j) - P_l(D_j)] + \sum_{j=t-k}^{t-1} \delta_j \\
= & \sum_{l=t-k}^{i-1} \left[P_{l+1}\left(\bigcup_{t-k \le j \le l} D_j \right) - P_l\left(\bigcup_{t-k \le j \le l} D_j \right) \right] + \\
& \sum_{l=i+1}^{t-1} \left[P_{l-1}\left(\bigcup_{l \le j \le t-1} D_j \right) - P_l\left(\bigcup_{l \le j \le t-1} D_j \right) \right] + \\
& \sum_{j=t-k}^{t-1} \delta_j \\
\le & \sum_{l=t-k}^{t-1} \gamma_l/2 + \sum_{j=t-k}^{t-1} \delta_j .
\end{aligned}
$$

□

References

[1] P. Auer and P. M. Long. Structural results about on-line learning models with and without queries. *Machine Learning*, 1994. To appear.

[2] P. L. Bartlett. Learning with a slowly changing distribution. In *Proc. 5th Annu. Workshop on Comput. Learning Theory*, pages 243–252. ACM Press, New York, NY, 1992.

[3] R. D. Barve and P. M. Long. On the complexity of learning from drifting distributions. Manuscript, January 1996.

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

[5] N. Cesa-Bianchi, Y. Freund, D. P. Helmbold, and M. Warmuth. On-line prediction and conversion strategies. In *Computational Learning Theory: Eurocolt '93*, volume New Series Number 53 of *The Institute of Mathematics and its Applications Conference Series*, pages 205–216, Oxford, 1994. Oxford University Press.

[6] R. M. Dudley. *Real Analysis and Probability.* Wadsworth and Brooks/Cole, Belmont, CA, 1989.

[7] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. G. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82:247–261, 1989. First appeared in Proc. 1st Annu. Workshop on Comput. Learning Theory, 1988.

[8] T. Hagerup and C. Rub. A guided tour of Chernov bounds. *Inform. Proc. Lett.*, 33:305–308, 1990.

[9] D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting {0,1} functions on randomly drawn points. In *Proceedings of the 29th Annual IEEE Symposium on Foundations of Computer Science*, pages 100–109. IEEE Computer Society Press, 1988.

[10] D. P. Helmbold and P. M. Long. Tracking drifting concepts using random examples. In *Proc. 4th Annu. Workshop on Comput. Learning Theory*, pages 13–23, San Mateo, CA, 1991. Morgan Kaufmann.

[11] D. P. Helmbold and P. M. Long. Tracking drifting concepts by minimizing disagreements. *Machine Learning*, 14(1):27–45, 1994.

[12] T. Kuh, T. Petsche, and R. Rivest. Mistake bounds of incremental learners when concepts drift with applications to feedforward networks. In *NIPS 4*. Morgan Kaufmann, 1991.

[13] T. Kuh and R. L. Rivest. Incrementally learning time-varying half-planes. In *Advances in Neural Information Processing Systems 4*, pages 920–927. Morgan Kaufmann, 1992.

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

[15] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.