Statistical Learning under Nonstationary Mixing Processes

Steve Hanneke
Toyota Technological Institute at Chicago
steve.hanneke@gmail.com

Liu Yang
liu.yang0900@outlook.com

Abstract

We study a special case of the problem of statistical learning without the i.i.d. assumption. Specifically, we suppose a learning method is presented with a sequence of data points, and required to make a prediction (e.g., a classification) for each one, and can then observe the loss incurred by this prediction. We go beyond traditional analyses, which have focused on stationary mixing processes or nonstationary product processes, by combining these two relaxations to allow nonstationary mixing processes. We are particularly interested in the case of β-mixing processes, with the sum of changes in marginal distributions growing sublinearly in the number of samples. Under these conditions, we propose a learning method, and establish that for bounded VC subgraph classes, the cumulative excess risk grows sublinearly in the number of predictions, at a quantified rate.

1 Introduction

Our setting is that of stream-based prediction. At each time t, we are given access to data points from times 1 through $t-1$, and are required to produce a predictor f_t, which is then evaluated on a new data point at time t. We study this in the *general learning setting* of (Vapnik, 1982, 1998), which represents the learning objective as an abstract optimization problem. As an example, in the special case of classification, given access to pairs $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$, we would be tasked with producing a function mapping an observed point x_t to a classification $\hat{y}_t$, and we would be evaluated on whether $\hat{y}_t \neq y_t$ (called a *mistake*). We are then interested in characterizing the rate of growth of the cumulative number of mistakes, as we repeat this for increasing values of t.

To study this problem, we suppose the sequence of observations are stochastic, subject to some restrictions on their distribution. Several such restrictions are possible. For instance, the most-common assumption used in the vast majority of the statistical learning literature is that the data are independent and identically distributed (i.i.d.). However, some efforts to relax this assumption have also been explored. There are essentially two main threads of work toward relaxing this assumption: relaxing the independence assumption while maintaining the assumption of identical distributions (or stationarity), or relaxing the assumption of identical distributions while maintaining the independence assumption. In the present work, we are interested in relaxing these assumptions jointly. Before getting into the details, let us first briefly review these two threads of the literature.

Most of the literature on relaxations of the independence assumption focuses on *stationary mixing* processes. At the extreme of this branch, the work of (Adams and Nobel, 2010) reveals that any VC class admits a uniform law of large numbers under stationary ergodic processes. In particular, this implies that the method of *empirical risk minimization* approaches excess risk zero in the limit. However, one cannot establish *rates* of convergence under such general conditions as ergodicity. To establish such rates, other works have therefore introduced stronger conditions, such as the β-mixing condition. Specifically, (Yu, 1994; Karandikar and Vidyasagar, 2002) have proven asymptotic rates of uniform convergence for VC classes under stationary β-mixing processes. One implication of this result is an asymptotic rate of convergence for the excess risk of empirical risk minimization. Other works have established rates of convergence for the excess risk of empirical risk minimization and other learning methods, under related mixing conditions, including α-mixing (Vidyasagar, 2003), η-mixing (Kontorovich, 2007), and ϕ-mixing (Vidyasagar, 2003), all under the

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

stationarity assumption.

The other primary direction in the study of the risk of learning methods under relaxations of the i.i.d. assumption preserves the independence assumption, while allowing the marginal distributions to *drift* over time. This thread in the literature has focused on the specific setting of binary classification. Specifically, (Long, 1999; Helmbold and Long, 1991, 1994; Barve and Long, 1996, 1997; Crammer, Mansour, Even-Dar, and Vaughan, 2010) study a setting in which the marginal distribution of the data point at time t has total variation distance from that of the data point at time $t+1$ at most a given upper bound, called the *drift rate* (see also related work by (Bartlett, 1992; Freund and Mansour, 1997; Bartlett, Ben-David, and Kulkarni, 2000; Yang, 2011; Mohri and Muñoz Medina, 2012)). The data points are still assumed to be independent. The recent works of (Hanneke, Kanade, and Yang, 2015; Mohri and Muñoz Medina, 2012) further explore this problem (in a formulation more-closely paralleling that studied here). In this setting, the learning method produces a sequence of predictors (e.g., classifiers), where the method for choosing the predictor at time t may depend on all of the data up to time $t-1$. The results in these works are expressible as bounds on the risk at each time t (or sometimes averaged over time), as a function of t and the rates of drift of the marginal distributions.

The paper of (Mohri and Muñoz Medina, 2012) also studies a refinement of the notion of "drift" compared to the earlier works, such as (Barve and Long, 1996, 1997). Specifically, rather than measuring the difference between the next and previous distributions by the total variation distance, they instead use a notion of "discrepancy" that depends directly on the function class being used for learning. This discrepancy is sometimes significantly smaller than the total variation distance, yet plays an analogous role in the bounds of (Mohri and Muñoz Medina, 2012) as the total variation distance plays in the bounds of (Helmbold and Long, 1994; Barve and Long, 1997). To allow for this refined notion of drift, our arguments below are phrased generally enough that they can be applied with either notion of drift (discrepancy or total variation).

On the subject of relaxing the independence assumption to allow mixing processes with drift, (Agarwal and Duchi, 2013) made some initial steps by studying the performance of certain learning methods under mixing processes, which may drift over time. However, among other differences, their analysis was restricted to sequences of distributions that are *convergent*, a requirement much stronger than our drift condition below. In recent work, (Kuznetsov and Mohri, 2014) also discusses the problem of learning from nonstationary mixing processes. They derive interesting results bounding the risk at some future time in terms of the empirical risk on all observed data, with clear implications for the performance of methods such as empirical risk minimization. The nature of the results in that work are somewhat different from our results below. However, the spirit of the analysis is similar in many places, and one can conceivably convert some of those results into a more-closely related form with a bit of additional effort.

One significant point of divergence between the present work and that of (Kuznetsov and Mohri, 2014), and all of the above works on product processes (aside from certain special cases discussed by (Hanneke, Kanade, and Yang, 2015)), is that in the general case, these works require access to the sequence of magnitudes of drift of the distribution, or a constant upper bound thereon. The sequence of drift magnitudes is a substantial number of variables to assume we have access to (linear in the number of data points), and relying only on a constant upper bound precludes the possibility of sublinear growth of the cumulative excess risk (Helmbold and Long, 1994; Hanneke, Kanade, and Yang, 2015). The notion of discrepancy studied by (Mohri and Muñoz Medina, 2012; Kuznetsov and Mohri, 2014) (see below) can sometimes be estimated from data, but only under significant further restrictions on the process. In contrast, in the present work, we merely assume an asymptotic bound on the rate of growth of the cumulative amount of drift. Our learning method then depends only on the single parameter that this asymptotic growth rate is described in terms of, and we show that this is enough to achieve sublinear growth of the cumulative excess risk, without needing access to the sequence of drift rates or additional restrictions on the process. We leave as future work the question of whether it is possible to adapt to the value of this parameter. For completeness, we also briefly discuss a case where the drift rates are known, in Section 3.

The present work studies learning under general nonstationary processes, under a condition that allows us to extend the ideas from the above-described literature on learning from product processes with slowly-drifting marginal distributions. Specifically, we replace the independence condition with a β-mixing condition. In addition to this, we suppose that the sum of distances between marginal distributions at adjacent time steps grows only sublinearly (note that this does *not* require that the sequence of distributions be converging). Our objective is then to propose a prediction strategy (for producing the f_t function), and to characterize the rate of growth of the cumulative excess risk over time. The excess risks are calculated relative

to the sequence of *a priori* optimal predictors among functions in a given function class. In particular, for any bounded VC subgraph class, we establish a rate of growth of the cumulative excess risk that is *sublinear* in the number of predictions made.

1.1 Definitions and Summary of Main Result

To formalize this setting, we adopt the abstract perspective of the *general learning setting* of (Vapnik, 1982, 1998). Specifically, fix a measurable space $(\mathcal{Z}, \mathcal{Z})$ and a *function class* $\mathcal{F}$ of measurable functions $f : \mathcal{Z} \to [0,1]$. For instance, in the special case of classification, $\mathcal{Z}$ would be a set of (x, y) pairs, and $\mathcal{F}$ would be a set of functions $f_h((x,y)) = \mathbb{1}[h(x) \neq y]$, where h ranges over a set $\mathcal{H}$ of functions (known as the hypothesis class); see (Koltchinskii, 2006; Shalev-Shwartz, Shamir, Srebro, and Sridharan, 2010) for many other examples. In the general learning setting, the aim of a learning algorithm is to identify a function $f \in \mathcal{F}$ with a relatively small average value, where the average is taken with respect to some unknown probability measure on $\mathcal{Z}$ (as discussed in more detail below). For instance, in the classification setting described above, this average value corresponds to the probability that h makes a "mistake" in predicting the value of y from x.

For simplicity, to avoid the common measurability issues arising in empirical process theory, we will suppose $\mathcal{F}$ is such that the events involved in the proofs below are all measurable (for instance, this is certainly the case if $\mathcal{F}$ is countable; see (van der Vaart and Wellner, 1996) for other sufficient conditions). Let d denote the pseudo-dimension of $\mathcal{F}$ (Pollard, 1984, 1990; Haussler, 1992; Anthony and Bartlett, 1999): that is, d is the largest $k \in \mathbb{N} \cup \{0\}$ such that $\exists (z_1, w_1), \ldots, (z_k, w_k) \in \mathcal{Z} \times \mathbb{R}$ with $|\{(\mathbb{1}[f(z_1) \leq w_1], \ldots, \mathbb{1}[f(z_k) \leq w_k]) : f \in \mathcal{F}\}| = 2^k$, or is ∞ if no such largest k exists. Throughout this article, we suppose $1 \leq d < \infty$ (so that $\mathcal{F}$ is a VC Subgraph class).

We suppose there is a sequence of $\mathcal{Z}$-valued random variables $Z_1, Z_2, \ldots$, called the *data points*, and for each $t \in \mathbb{N}$, we denote by P_t the marginal distribution of the random variable Z_t. Also, generally, for any random variable X, we denote by $\mathbb{P}_X$ the distribution of X (i.e., $\mathbb{P}_X(\cdot) = \mathbb{P}(X^{-1}(\cdot))$). For any probability measures P, Q on a measurable space $(\Omega, \mathcal{B})$, we denote by $\|P - Q\| = \sup_{A \in \mathcal{B}} P(A) - Q(A)$ the total variation distance between P and Q. Additionally, for probability measures P, Q on the measurable space $(\mathcal{Z}, \mathcal{Z})$, we denote by

$$\rho(P,Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{Z \sim P}[f(Z)] - \mathbb{E}_{Z \sim Q}[f(Z)] \right|,$$

a general notion of *discrepancy* introduced by (Mansour, Mohri, and Rostamizadeh, 2009; Mohri and Muñoz Medina, 2012). We use ρ below to quantify the magnitude of change in the marginal distribution of Z_{t+1} compared to Z_t. Note that, since every $f \in \mathcal{F}$ is uniformly bounded in $[0,1]$, we clearly have

$$\rho(P,Q) \leq \|P - Q\|.$$

Indeed, readers more comfortable with the familiar total variation distance may feel free to replace $\rho(P,Q)$ with $\|P-Q\|$ in all contexts below, and the results and proofs will remain valid without any further modifications. However, one can construct scenarios in which $\rho(P,Q)$ provides a much smaller value, and generally $\rho(P,Q)$ appears to be more relevant to the learning setting than is the total variation distance. For each $t \geq 2$, let $\Delta_t \in [0,1]$ be a value satisfying

$$\rho(P_t, P_{t-1}) \leq \Delta_t. \tag{1}$$

For completeness, also define $\Delta_1 = 0$.

To obtain nontrivial results, we are interested in restricting the family of processes. Specifically, for our main result below (Theorem 1), we suppose

$$\sum_{t=1}^{T} \Delta_t = O(T^{\alpha}), \tag{2}$$

for a given value $\alpha \in [0,1)$. Note that this does *not* require that the sequence of distributions be converging, only that its average rate of change slows over time. A simple example of a non-convergent sequence satisfying (2) with $\alpha = 1/2$ is P_t as Bernoulli$(\sqrt{t} - \lfloor \sqrt{t} \rfloor)$.

We additionally adopt the standard definition of β-*mixing*, defined as follows. Following (Bradley, 1983) and (Yu, 1994), for each $k \in \mathbb{N}$, define

$$\beta_k = \frac{1}{2} \sup \left\{ \sum_{i=1}^{I} \sum_{j=1}^{J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)| : \{A_i\}_i \in \Pi_\ell, \{B_j\}_j \in \Pi'_{\ell+k}, \ell \geq 1 \right\},$$

where Π_ℓ is defined as the set of $\sigma(\{Z_1, \ldots, Z_\ell\})$-measurable finite partitions, and $\Pi'_{\ell+k}$ is defined as the set of $\sigma(\{Z_{\ell+k}, Z_{\ell+k+1}, \ldots\})$-measurable finite partitions. Then we suppose

$$\beta_k = O(k^{-r}), \tag{3}$$

for some $r \in (0, \infty)$.

Under the assumptions (2) and (3), we propose a learning method, specified as follows. Let $\hat{f}_1$ be arbitrary. For each $t \in \mathbb{N} \setminus \{1\}$, let

$$m_t = \left\lceil (t-1)^{(1-\alpha)\frac{3+2r}{3+3r}} \right\rceil$$

and

$$k_t = \left\lceil (t-1)^{(1-\alpha)\frac{1}{1+r}} \right\rceil$$

and choose as a predictor at time t a function[1]

$$\hat{f}_t = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{s=1}^{\lfloor m_t/k_t \rfloor} f(Z_{t-sk_t}). \tag{4}$$

For $\hat{f}_t$ chosen in this way, we prove the following theorem.

Theorem 1. *If* (2) *and* (3) *are satisfied, then*

$$\sum_{t=1}^{T} \mathbb{E}\left[\hat{f}_t(Z_t)\right] - \sum_{t=1}^{T} \inf_{f \in \mathcal{F}} \mathbb{E}\left[f(Z_t)\right] = O\left(T^{\frac{3+(2+\alpha)r}{3+3r}}\right).$$

In particular, note that the expression on the right hand side grows *sublinearly* in T. To prove this theorem, we first provide two key lemmas from the literature, after which we present the proof of Theorem 1 below. Following this, in Section 3, we conclude the paper by establishing *finite-sample* bounds, and other specialized results, in the special case of *product* processes; this effectively extends to the general learning setting results established by (Barve and Long, 1996, 1997) for binary classification, while also expressing the results in a more general form that allows for a time-varying drift rate.

2 Proof of Theorem 1

The following lemma is a well-known result on β-mixing processes, from (Volkonskii and Rozanov, 1959; Eberlein, 1984) (see also Theorem 2.1 of (Vidyasagar, 2003) or Corollary 2.7 of (Yu, 1994)).

Lemma 1. *For any $t, n, k \in \mathbb{N}$,*

$$\left\| \mathbb{P}_{\{Z_{(j-1)k+t}\}_{j=1}^{n}} - \left(\bigtimes_{j=1}^{n} P_{(j-1)k+t} \right) \right\| \leq (n-1)\beta_k.$$

Additionally, we use the following well-known result (see e.g., (van der Vaart and Wellner, 1996), Theorems 2.14.1 and 2.6.7).

Lemma 2. *There exists a universal constant $c \in [1, \infty)$ such that, for any independent $\mathcal{Z}$-valued random variables $Z'_1, \ldots, Z'_m$,*

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{t=1}^{m} (f(Z'_t) - \mathbb{E}[f(Z'_t)]) \right| \right] \leq c\sqrt{\frac{d}{m}}.$$

[1] For simplicity, we suppose the minimum is actually *achieved* by some $f \in \mathcal{F}$. To handle the general case, all of the results continue to hold, with only minor technical changes to the proofs, if we instead choose $\hat{f}_t \in \mathcal{F}$ with $\sum_{s=1}^{\lfloor m_t/k_t \rfloor} \hat{f}_t(Z_{t-sk_t})$ sufficiently close to $\inf_{f \in \mathcal{F}} \sum_{s=1}^{\lfloor m_t/k_t \rfloor} f(Z_{t-sk_t})$.

While the proof of this result in (van der Vaart and Wellner, 1996) discusses only i.i.d. random variables, essentially the same proof in fact implies this result, which only assumes independence. As this observation is fairly well known, we do not include a separate proof here.

With these lemmas in hand, we are ready to present the proof of Theorem 1.

Proof of Theorem 1. Let $Z'_1, Z'_2, \ldots$ denote a sequence of independent random variables, also independent from $\{Z_i\}_{i \in \mathbb{N}}$, and with each $Z'_i \sim P_i$. Fix any $t \in \mathbb{N} \setminus \{1\}$. Since $\hat{f}_t$ depends only on $Z_1, \ldots, Z_{t-k_t}$, it follows immediately from the definition of β_{k_t} (see (Yu, 1994), Lemma 2.6) that

$$\left\| \mathbb{P}_{(\hat{f}_t, Z_t)} - \mathbb{P}_{(\hat{f}_t, Z'_t)} \right\| = \left\| \mathbb{P}_{(\hat{f}_t, Z_t)} - \mathbb{P}_{\hat{f}_t} \times \mathbb{P}_{Z_t} \right\| \leq \beta_{k_t}.$$

In particular, this implies

$$\mathbb{E}\left[\hat{f}_t(Z_t)\right] \leq \mathbb{E}\left[\hat{f}_t(Z'_t)\right] + \beta_{k_t}.$$

Additionally, since $\rho(P_{t-ik_t}, P_t) \leq \sum_{q=t-ik_t}^{t-1} \Delta_{q+1}$ for $1 \leq i \leq \lfloor m_t/k_t \rfloor$, and every Z'_j is independent of $\hat{f}_t$, we have that

$$\begin{aligned}
\mathbb{E}\left[\hat{f}_t(Z'_t)\right] &= \mathbb{E}\left[\mathbb{E}\left[\hat{f}_t(Z'_t)\middle|\hat{f}_t\right]\right] \\
&\leq \mathbb{E}\left[\frac{1}{\lfloor m_t/k_t \rfloor} \sum_{i=1}^{\lfloor m_t/k_t \rfloor} \mathbb{E}\left[\hat{f}_t(Z'_{t-ik_t})\middle|\hat{f}_t\right]\right] \\
&\quad + \frac{1}{\lfloor m_t/k_t \rfloor} \sum_{i=1}^{\lfloor m_t/k_t \rfloor} \sum_{q=t-ik_t}^{t-1} \Delta_{q+1}.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\lfloor m_t/k_t \rfloor} \sum_{i=1}^{\lfloor m_t/k_t \rfloor} \mathbb{E}\left[\hat{f}_t(Z'_{t-ik_t})\middle|\hat{f}_t\right]\right] \\
&\leq \mathbb{E}\left[\frac{1}{\lfloor m_t/k_t \rfloor} \sum_{i=1}^{\lfloor m_t/k_t \rfloor} \hat{f}_t(Z_{t-ik_t})\right] \\
&\quad + \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{\lfloor m_t/k_t \rfloor} \sum_{i=1}^{\lfloor m_t/k_t \rfloor} \left(\mathbb{E}[f(Z'_{t-ik_t})] - f(Z_{t-ik_t})\right) \right|\right].
\end{aligned} \tag{5}$$

Now let us bound each term in (5) separately. First,

we have that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\lfloor m_t/k_t\rfloor}\sum_{i=1}^{\lfloor m_t/k_t\rfloor}\hat{f}_t(Z_{t-ik_t})\right] \\
&= \mathbb{E}\left[\inf_{f\in\mathcal{F}}\frac{1}{\lfloor m_t/k_t\rfloor}\sum_{i=1}^{\lfloor m_t/k_t\rfloor}f(Z_{t-ik_t})\right] \\
&\leq \inf_{f\in\mathcal{F}}\frac{1}{\lfloor m_t/k_t\rfloor}\sum_{i=1}^{\lfloor m_t/k_t\rfloor}\mathbb{E}\left[f(Z_{t-ik_t})\right] \\
&\leq \inf_{f\in\mathcal{F}}\mathbb{E}[f(Z_t)] + \frac{1}{\lfloor m_t/k_t\rfloor}\sum_{i=1}^{\lfloor m_t/k_t\rfloor}\sum_{q=t-ik_t}^{t-1}\Delta_{q+1}.
\end{aligned}$$

Next, Lemma 1 implies

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\lfloor m_t/k_t\rfloor}\sum_{i=1}^{\lfloor m_t/k_t\rfloor}\left(\mathbb{E}[f(Z'_{t-ik_t})] - f(Z_{t-ik_t})\right)\right|\right] \\
&\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\lfloor m_t/k_t\rfloor}\sum_{i=1}^{\lfloor m_t/k_t\rfloor}\left(\mathbb{E}[f(Z'_{t-ik_t})] - f(Z'_{t-ik_t})\right)\right|\right] \\
&\quad + \left(\lfloor m_t/k_t\rfloor - 1\right)\beta_{k_t}.
\end{aligned}$$

Furthermore, Lemma 2 implies

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\lfloor m_t/k_t\rfloor}\sum_{i=1}^{\lfloor m_t/k_t\rfloor}\left(\mathbb{E}[f(Z'_{t-ik_t})] - f(Z'_{t-ik_t})\right)\right|\right] \\
&\qquad\qquad \leq c\sqrt{\frac{d}{\lfloor m_t/k_t\rfloor}}.
\end{aligned}$$

Together, we have that (5) is at most

$$\begin{aligned}
&\inf_{f\in\mathcal{F}}\mathbb{E}[f(Z_t)] + \left(\frac{1}{\lfloor m_t/k_t\rfloor}\sum_{i=1}^{\lfloor m_t/k_t\rfloor}\sum_{q=t-ik_t}^{t-1}\Delta_{q+1}\right) \\
&\qquad + c\sqrt{\frac{d}{\lfloor m_t/k_t\rfloor}} + \left(\lfloor m_t/k_t\rfloor - 1\right)\beta_{k_t}.
\end{aligned}$$

Altogether, we have established that

$$\begin{aligned}
\mathbb{E}\left[\hat{f}_t(Z_t)\right] &\leq \inf_{f\in\mathcal{F}}\mathbb{E}[f(Z_t)] \\
&\quad + 2\left(\frac{1}{\lfloor m_t/k_t\rfloor}\sum_{i=1}^{\lfloor m_t/k_t\rfloor}\sum_{q=t-ik_t}^{t-1}\Delta_{q+1}\right) \\
&\quad + c\sqrt{\frac{d}{\lfloor m_t/k_t\rfloor}} + \lfloor m_t/k_t\rfloor\beta_{k_t}. \qquad (6)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\hat{f}_t(Z_t)\right] - \sum_{t=1}^{T}\inf_{f\in\mathcal{F}}\mathbb{E}\left[f(Z_t)\right] \\
&\leq 1 + \left(\sum_{t=2}^{T}\frac{2}{\lfloor m_t/k_t\rfloor}\sum_{i=1}^{\lfloor m_t/k_t\rfloor}\sum_{q=t-ik_t}^{t-1}\Delta_{q+1}\right) \\
&\quad + \left(\sum_{t=2}^{T}c\sqrt{\frac{d}{\lfloor m_t/k_t\rfloor}}\right) + \left(\sum_{t=2}^{T}\lfloor m_t/k_t\rfloor\beta_{k_t}\right). \qquad (7)
\end{aligned}$$

All that remains is to bound each of these three terms on the right hand side of (7). The only term presenting a challenge in this regard is the term involving the Δ_{q+1} values, and for that reason we leave this term for last. For the other terms, first note that

$$\begin{aligned}
\sum_{t=1}^{T}t^{-(1-\alpha)\frac{r}{3+3r}} &= O\left(1 + \int_1^T t^{-(1-\alpha)\frac{r}{3+3r}}\mathrm{d}t\right) \\
&= O\left(T^{\frac{3+(2+\alpha)r}{3+3r}}\right).
\end{aligned}$$

Thus, we have that

$$\begin{aligned}
\sum_{t=2}^{T}c\sqrt{\frac{d}{\lfloor m_t/k_t\rfloor}} &= O\left(\sum_{t=1}^{T}t^{-(1-\alpha)\frac{r}{3+3r}}\right) \\
&= O\left(T^{\frac{3+(2+\alpha)r}{3+3r}}\right). \qquad (8)
\end{aligned}$$

Also, we have

$$\begin{aligned}
\sum_{t=2}^{T}\lfloor m_t/k_t\rfloor\beta_{k_t} &= O\left(\sum_{t=2}^{T}m_t/k_t^{1+r}\right) \\
&= O\left(\sum_{t=1}^{T}t^{-(1-\alpha)\frac{r}{3+3r}}\right) = O\left(T^{\frac{3+(2+\alpha)r}{3+3r}}\right). \qquad (9)
\end{aligned}$$

The remaining term,

$$\sum_{t=2}^{T}\frac{2}{\lfloor m_t/k_t\rfloor}\sum_{i=1}^{\lfloor m_t/k_t\rfloor}\sum_{q=t-ik_t}^{t-1}\Delta_{q+1},$$

requires more work to bound. First note that

$$\sum_{t=2}^{T}\frac{2}{\lfloor m_t/k_t\rfloor}\sum_{i=1}^{\lfloor m_t/k_t\rfloor}\sum_{q=t-ik_t}^{t-1}\Delta_{q+1} \leq 2\sum_{t=2}^{T}\sum_{q=t-m_t}^{t-1}\Delta_{q+1}.$$

We will focus on bounding the right hand side. Now note that every value of $t\in\mathbb{N}$ for which $q\in\{t-m_t,\ldots,t-1\}$ satisfies

$$\begin{aligned}
2q &\geq 2t - 2m_t = 2t\left(1 - \frac{m_t}{t}\right) \\
&\geq 2t\left(1 - 2(t-1)^{(1-\alpha)\frac{3+2r}{3+3r}-1}\right) \\
&= 2t\left(1 - 2(t-1)^{-\frac{3\alpha+2r\alpha+r}{3+3r}}\right) \\
&\geq 2t\left(1 - 2q^{-\frac{3\alpha+2r\alpha+r}{3+3r}}\right) \geq 2t\left(1 - 2q^{-\frac{r}{3+3r}}\right).
\end{aligned}$$

Denote $q_r = \left\lceil 4^{\frac{3+3r}{r}} \right\rceil$, and note that for any $q \geq q_r$ we have $2t\left(1 - 2q^{-\frac{r}{3+3r}}\right) \geq t$. Thus, for any $q \geq q_r$, every $t \in \mathbb{N}$ with $q \in \{t - m_t, \ldots, t-1\}$ has $t \leq 2q$, so that (by monotonicity of m_t) we also have $q \in \{t - m_{2q}, \ldots, t-1\}$, or equivalently $t \in \{q+1, \ldots, q+m_{2q}\}$. In particular, this means any such q has at most m_{2q} appearances of the quantity Δ_{q+1} in the summation $\sum_{t=2}^{T}\sum_{q=t-m_t}^{t-1}\Delta_{q+1}$. Also, clearly the largest q with Δ_{q+1} appearing in this summation is $q = T-1$. Additionally, since m_t is sublinear in t, we have $t - m_t \to \infty$ as $t \to \infty$, so that there is some finite t_0 such that every $t > t_0$ has $t - m_t \geq q_r$. Thus, every $q < q_r$ has Δ_{q+1} appearing at most t_0 times in the summation $\sum_{t=2}^{T}\sum_{q=t-m_t}^{t-1}\Delta_{q+1}$. Altogether, we have that

$$\begin{aligned}
2\sum_{t=2}^{T}\sum_{q=t-m_t}^{t-1}\Delta_{q+1} &\leq 2t_0\sum_{q=1}^{q_r-1}\Delta_{q+1} + 2\sum_{q=q_r}^{T-1} m_{2q}\Delta_{q+1} \\
&= O\left(m_{2T}\sum_{q=1}^{T}\Delta_q\right) \\
&= O\left(T^{(1-\alpha)\frac{3+2r}{3+3r}+\alpha}\right) = O\left(T^{\frac{3+(2+\alpha)r}{3+3r}}\right),
\end{aligned}$$

where we have used the assumption (2) on the Δ_t sequence.

Plugging this bound into (7) along with (8) and (9), we have established that

$$\sum_{t=1}^{T}\mathbb{E}\left[\hat{f}_t(Z_t)\right] - \sum_{t=1}^{T}\inf_{f\in\mathcal{F}}\mathbb{E}\left[f(Z_t)\right] = O\left(T^{\frac{3+(2+\alpha)r}{3+3r}}\right),$$

which completes the proof. □

3 Product Processes

In this section, unlike above, we suppose the algorithm has direct access to the Δ_t sequence. Our objective is then to derive more-explicit (non-asymptotic) bounds under the assumption that $\{Z_t\}_{t=1}^{\infty}$ is a product process. The results here are already known in the special case of binary classification, in the case that Δ_t is bounded by a t-invariant *constant* for all t (Barve and Long, 1997). Thus, this section represents a generalization of these classic results to the general learning setting, and to general time-varying drift rates. That said, we note that the results here would also readily follow from the classic analysis of (Barve and Long, 1997) and the more-recent work of (Mohri and Muñoz Medina, 2012), with only minor additional work to apply those results to a recent history of data points trailing the prediction time t; there is nevertheless some value in stating the results explicitly here, particularly since they follow directly from our analysis above.

Throughout this section, for any functions $f, g : A \to [0,\infty)$, for any set A, we write $f(a) \lesssim g(a)$ to express the claim that there exists a numerical constant $c \in (0,\infty)$ such that $f(a) \leq cg(a)$ for all $a \in A$; this allows us to express non-asymptotic bounds (in terms of T, d, and the Δ_t sequence), without concerning ourselves with precise numerical constant factors. For each $t \in \mathbb{N}\setminus\{1\}$, define

$$\tilde{m}_t = \operatorname*{argmin}_{m\in\{1,\ldots,t-1\}}\left(\sum_{q=t-m}^{t-1}\Delta_{q+1} + \sqrt{\frac{d}{m}}\right)$$

and

$$\tilde{f}_t = \operatorname*{argmin}_{f\in\mathcal{F}}\sum_{s=t-\tilde{m}_t}^{t-1} f(Z_s).$$

For completeness, define $\tilde{f}_1$ as an arbitrary element of $\mathcal{F}$.

Theorem 2. *If $\{Z_t\}_{t=1}^{\infty}$ is a product process, then for $T \in \mathbb{N}\setminus\{1\}$,*

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\tilde{f}_t(Z_t)\right] - \sum_{t=1}^{T}\inf_{f\in\mathcal{F}}\mathbb{E}[f(Z_t)] \\
&\qquad \lesssim \sum_{t=2}^{T}\min_{m\in\{1,\ldots,t-1\}}\left(\sum_{q=t-m}^{t-1}\Delta_{q+1} + \sqrt{\frac{d}{m}}\right).
\end{aligned}$$

Proof. We begin by noting that, in the proof of Theorem 1, the argument leading to (7) in fact more generally holds for *any* β-mixing process $\{Z_t\}_{t\in\mathbb{N}}$ (regardless of whether (2) and (3) are satisfied for the corresponding Δ_t and β_k sequences), and for *any* sequence $\hat{f}_t$ defined as in (4), where the values $m_t, k_t \in \mathbb{N}$ can be specified *arbitrarily*, subject to $k_t \leq m_t \leq t-1$. In particular, substituting $k_t = 1$ and $m_t = \tilde{m}_t$, the corresponding $\hat{f}_t$ from (4) is precisely $\tilde{f}_t$. Then since $\beta_1 = 0$ for product processes, (7) implies

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\tilde{f}_t(Z_t)\right] - \sum_{t=1}^{T}\inf_{f\in\mathcal{F}}\mathbb{E}\left[f(Z_t)\right] \\
&\lesssim \sum_{t=2}^{T}\left(\sum_{q=t-\tilde{m}_t}^{t-1}\Delta_{q+1} + \sqrt{\frac{d}{\tilde{m}_t}}\right) \\
&= \sum_{t=2}^{T}\min_{m\in\{1,\ldots,t-1\}}\left(\sum_{q=t-m}^{t-1}\Delta_{q+1} + \sqrt{\frac{d}{m}}\right).
\end{aligned}$$

□

It remains an interesting open problem to determine whether the above guarantee is achievable by a learning rule that has no direct dependence on the Δ_t values: that is, a method that is *adaptive* to variations in the rates of drift. Resolution of this question seems

an important step toward applicability of these ideas in practice. Of course, as established in Theorem 1, if we instead assume that the asymptotic bound (2) holds, then it is possible to replace the direct dependence on Δ_t with a mere dependence on a single parameter α; however, the price for this is that the finite-sample bound in Theorem 2 would be replaced by an asymptotic guarantee. An alternative option is to suppose the drift rates Δ_t are *bounded* by a value γ, and then provide an algorithm depending only on γ; this coarse condition on Δ_t precludes the possibility of a sublinear cumulative excess risk guarantee, but it can nonetheless be interesting to study the dependence of the achieved excess risk on γ. This is the subject of the next subsection.

3.1 Constant Drift Rate

In the context of binary classification, (Long, 1999; Helmbold and Long, 1991, 1994; Barve and Long, 1996, 1997; Crammer, Mansour, Even-Dar, and Vaughan, 2010; Hanneke, Kanade, and Yang, 2015; Mohri and Muñoz Medina, 2012) have derived bounds on the sequence of risks (or the number of mistakes) achieved by various methods, under the assumptions that $\{Z_t\}_{t=1}^{\infty}$ is a product process, and that $\Delta_t \leq \gamma$, for some fixed constant $\gamma \in (0,1)$. Here we briefly note that some of these results (and in particular, those of (Barve and Long, 1997)) can be generalized to the general learning setting, where we find analogous results on the average of the $\hat{f}_t(Z_t)$ function values. We note that a similar type of result is also immediate from the analysis of (Mohri and Muñoz Medina, 2012) with minor additional effort to convert to our sequential setting.

Let $\bar{m} = \left\lceil d^{1/3}\gamma^{-2/3} \right\rceil$. For each integer $t > \bar{m}$, let

$$\bar{f}_t = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{s=t-\bar{m}}^{t-1} f(Z_s).$$

For completeness, for $t \leq \bar{m}$ define $\bar{f}_t$ as an arbitrary element of $\mathcal{F}$.

Theorem 3. *If $\{Z_t\}_{t=1}^{\infty}$ is a product process, then for $T > 1/\gamma$,*

$$\sum_{t=1}^{T} \mathbb{E}\left[\bar{f}_t(Z_t)\right] - \sum_{t=1}^{T} \inf_{f \in \mathcal{F}} \mathbb{E}[f(Z_t)] \lesssim (d\gamma)^{1/3} T.$$

It is worth noting that the bound in Theorem 3 would also hold for the predictor $\tilde{f}_t$ from Theorem 2; indeed, this follows immediately from plugging in γ for the values of Δ_t, in which case $\tilde{f}_t$ itself is quite similar to $\bar{f}_t$. However, as $\bar{f}_t$ admits the above simplified explicit form in this special case, we include a brief direct proof of this result as follows.

Proof. As in the proof of Theorem 2, the proof is based on the general validity of (6). In particular, taking $k_t = 1$ and $m_t = \min\{\bar{m}, (t-1)\}$, the corresponding $\hat{f}_t$ is equal $\bar{f}_t$ for all $t > \bar{m}$. Thus, (6) implies

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[\bar{f}_t(Z_t)\right] - \sum_{t=1}^{T} \inf_{f \in \mathcal{F}} \mathbb{E}\left[f(Z_t)\right] \\
&\lesssim \bar{m} + \sum_{t=\bar{m}+1}^{T} \left(\sum_{q=t-\bar{m}}^{t-1} \Delta_{q+1} + \sqrt{\frac{d}{\bar{m}}} \right) \\
&\leq \bar{m} + \sum_{t=\bar{m}+1}^{T} \left(\bar{m}\gamma + (d\gamma)^{1/3} \right) \lesssim d^{1/3}\gamma^{-2/3} + (d\gamma)^{1/3} T.
\end{aligned}$$

The proof is completed by noting that, for $T > 1/\gamma$, we have $(d\gamma)^{1/3} T > d^{1/3}\gamma^{-2/3}$, so that $d^{1/3}\gamma^{-2/3} + (d\gamma)^{1/3} T < 2(d\gamma)^{1/3} T$. □

4 Discussion and Open Problems

There remains an interesting question of whether the rate established in Theorem 1 is optimal. In the case of stationary β-mixing processes, the best known result is $O\left(T^{\frac{3+r}{3+2r}}\right)$ (Karandikar and Vidyasagar, 2002). This result can be recovered with our technique by setting $m_t = t-1$ and $k_t = \left\lceil (t-1)^{\frac{3}{3+2r}} \right\rceil$, noting that the term in (7) depending on the Δ_t values is equal 0 in the stationary case; indeed, to achieve this rate we required only that $\Delta_t = 0$ for all t, which is a strictly weaker requirement than stationarity. Stationary processes are a special case of $\alpha = 0$ in (2). However, the result given in Theorem 1 for $\alpha = 0$ obtains a somewhat faster growth of $O\left(T^{\frac{3+2r}{3+3r}}\right)$. Since the general case of $\alpha = 0$ includes many nonstationary processes as well, it is not clear whether Theorem 1 can be improved to provide a rate $O\left(T^{\frac{3+r}{3+2r}}\right)$ for general processes having $\alpha = 0$. If so, it would seem to require a different approach to the analysis, since if we were to take $m_t = t-1$ and $k_t = \left\lceil (t-1)^{\frac{3}{3+2r}} \right\rceil$ for a general process with $\alpha = 0$, the summation involving the Δ_t sequence in (7) might then potentially grow faster than $T^{\frac{3+r}{3+2r}}$. Complementary to this question is the problem of establishing lower bounds on the minimax rates, which seems to require development of novel techniques for constructing nonstationary mixing processes for which the learning problem is challenging.

Acknowledgments

We thank Tommi Jaakkola for several helpful discussions.

References

Adams, T. M. and Nobel, A. B. (2010). Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling. *Annals of Probability*, **38**(4), 1345–1367. 1

Agarwal, A. and Duchi, J. C. (2013). The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, **59**(1), 573–587. 1

Anthony, M. and Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press. 1.1

Bartlett, P. L. (1992). Learning with a slowly changing distribution. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pages 243–252. 1

Bartlett, P. L., Ben-David, S., and Kulkarni, S. R. (2000). Learning changing concepts by exploiting the structure of change. *Machine Learning*, **41**, 153–174. 1

Barve, R. D. and Long, P. M. (1996). On the complexity of learning from drifting distributions. In *Proceedings of the 9th Conference on Computational Learning Theory*, pages 122–130. 1, 1.1, 3.1

Barve, R. D. and Long, P. M. (1997). On the complexity of learning from drifting distributions. *Information and Computation*, **138**(2), 170–193. 1, 1.1, 3, 3.1

Bradley, R. C. (1983). Absolute regularity and functions of Markov chains. *Stochastic Processes and their Applications*, **14**, 67–77. 1.1

Crammer, K., Mansour, Y., Even-Dar, E., and Vaughan, J. W. (2010). Regret minimization with concept drift. In *Proceedings of the 23rd Conference on Learning Theory*, pages 168–180. 1, 3.1

Eberlein, E. (1984). Weak convergence of partial sums of absolutely regular sequences. *Statistics & Probability Letters*, **2**, 291–293. 2

Freund, Y. and Mansour, Y. (1997). Learning under persistent drift. In *Proceedings of the 3rd European Conference on Computational Learning Theory*, pages 109–118. 1

Hanneke, S., Kanade, V., and Yang, L. (2015). Learning with a drifting target concept. In *Proceedings of the 26th International Conference on Algorithmic Learning Theory*. 1, 3.1

Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, **100**, 78–150. 1.1

Helmbold, D. P. and Long, P. M. (1991). Tracking drifting concepts using random examples. In *Proceedings of the 4th Annual Workshop on Computational Learning Theory*, pages 13–23. 1, 3.1

Helmbold, D. P. and Long, P. M. (1994). Tracking drifting concepts by minimizing disagreements. *Machine Learning*, **14**(1), 27–45. 1, 3.1

Karandikar, R. L. and Vidyasagar, M. (2002). Rates of uniform convergence of empirical means with mixing processes. *Statistics & Probability Letters*, **58**(3), 297–307. 1, 4

Koltchinskii, V. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, **34**(6), 2593–2656. 1.1

Kontorovich, L. (2007). *Measure Concentration of Strongly Mixing Processes with Applications*. Ph.D. thesis, Carnegie Mellon University. 1

Kuznetsov, V. and Mohri, M. (2014). Generalization bounds for time series prediction with nonstationary processes. In *Proceedings of The 25th International Conference on Algorithmic Learning Theory*. 1

Long, P. M. (1999). The complexity of learning according to two models of a drifting environment. *Machine Learning*, **37**(3), 337–354. 1, 3.1

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. In *Proceedings of the 22nd Conference on Learning Theory*. 1.1

Mohri, M. and Muñoz Medina, A. (2012). New analysis and algorithm for learning with drifting distributions. In *Proceedings of The 23rd International Conference on Algorithmic Learning Theory*. 1, 1.1, 3, 3.1

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, Berlin / New York. 1.1

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 2, Institute of Mathematical Statistics and American Statistical Association. 1.1

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2010). Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, **11**, 2635–2670. 1.1

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer. 1.1, 2, 2

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. Springer-Verlag New York. 1, 1.1

Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley & Sons, Inc. 1, 1.1

Vidyasagar, M. (2003). *Learning and Generalization with Applications to Neural Networks*. Springer-Verlag, 2nd edition. 1, 2

Volkonskii, V. A. and Rozanov, Y. A. (1959). Some limit theorems for random functions. I. *Theory of Probability and its Applications*, **4**, 178–197. 2

Yang, L. (2011). Active learning with a drifting distribution. In *Advances in Neural Information Processing Systems 24*. 1

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, **22**(1), 94–116. 1, 1.1, 2, 2