

Universally Consistent Online Learning with Arbitrarily Dependent Responses

Steve Hanneke

Purdue University

STEVE.HANNEKE@GMAIL.COM

Abstract

This work provides an online learning rule that is universally consistent under processes on (X, Y) pairs, under conditions only on the X process. As a special case, the conditions admit all processes on (X, Y) such that the process on X is stationary. This generalizes past results which required stationarity for the joint process on (X, Y) , and additionally required this process to be ergodic. In particular, this means that ergodicity is superfluous for the purpose of universally consistent online learning.

Keywords: statistical learning theory, universal consistency, nonparametric estimation, stochastic processes, stationary processes, non-ergodic processes, online learning

1. Introduction

The task of achieving low expected *regret* in online learning is a classic topic in learning theory. Specifically, we consider a sequential setting, where at each time t , a learner observes a point X_t , makes a *prediction* \hat{Y}_t , and then observes a true *response* Y_t : that is, $\hat{Y}_t = f_t(X_{1:(t-1)}, Y_{1:(t-1)}, X_t)$ for some function f_t (possibly randomized). We are then interested in the rate of growth of the long-run cumulative *loss* of the learner: i.e., $\sum_{t=1}^T \ell(\hat{Y}_t, Y_t)$, for a given loss function ℓ . However, as it may sometimes be impossible to achieve low cumulative loss in an absolute sense, we are often interested in understanding the *excess* loss compared to some particular *fixed* predictor f_0 : i.e., $\sum_{t=1}^T \ell(\hat{Y}_t, Y_t) - \sum_{t=1}^T \ell(f_0(X_t), Y_t)$, known as the *regret* (relative to f_0).

Several different formulations of the subject have been proposed, leading to different algorithmic approaches and theoretical analyses of regret. For instance, there is a rich theory of online learning with *arbitrary* sequences $\{(X_t, Y_t)\}_{t=1}^\infty$, but where the reference function f_0 is restricted to belong to some particular function class \mathcal{F} (see e.g., [Cesa-Bianchi and Lugosi, 2006](#); [Ben-David, Pál, and Shalev-Shwartz, 2009](#); [Rakhlin, Sridharan, and Tewari, 2015](#)).

On the other hand, there has also been significant work on theories that allow *any* reference function f_0 , while restricting the sequence $\{(X_t, Y_t)\}_{t=1}^\infty$. This is the subject of the present work. A classic starting point for this line of work is the theory of *universally consistent* predictors under *i.i.d.* sequences. In particular, for binary classification and bounded regression, under mild conditions on the value space \mathcal{X} of the X_t variables, there are simple learning rules \hat{f}_t satisfying a guarantee that, for any *i.i.d.* sequence $\{(X_t, Y_t)\}_{t=1}^\infty$, for every measurable function $f_0 : \mathcal{X} \rightarrow \mathcal{Y}$, it holds that $\sum_{t=1}^T \ell(\hat{Y}_t, Y_t) - \sum_{t=1}^T \ell(f_0(X_t), Y_t) = o(T)$ almost surely (e.g., [Stone, 1977](#); [Devroye, Györfi, and Lugosi, 1996](#); [Györfi, Kohler, Krzyżak, and Walk, 2002](#); [Hanneke, Kontorovich, Sabato, and Weiss, 2021](#)). This result has since been extended to various non-*i.i.d.* conditions on $\{(X_t, Y_t)\}_{t=1}^\infty$, including the case where $\{(X_t, Y_t)\}_{t=1}^\infty$ is any *stationary ergodic* process (e.g., [Morvai, Yakowitz, and Györfi, 1996](#); [Györfi, Lugosi, and Morvai, 1999](#); [Györfi, Kohler, Krzyżak, and Walk, 2002](#)),

or generally satisfies a *law of large numbers* (Morvai, Kulkarni, and Nobel, 1999; Steinwart, Hush, and Scovel, 2009).

In the present work, we are interested in exploring *weakened* assumptions on $\{(X_t, Y_t)\}_{t=1}^\infty$ such that there still exist learners that are consistent for all $\{(X_t, Y_t)\}_{t=1}^\infty$ satisfying the assumption.

The recent work of (Hanneke, 2021a) focuses on the input sequence $\{X_t\}_{t=1}^\infty$, investigating the *minimal* assumption for this sequence to admit consistent learners under the restriction that $Y_t = f^*(X_t)$ for an arbitrary fixed function f^* (or in some cases, more generally with random variables Y_t , where $\mathbb{E}[\ell(f^*(X_t), Y_t)|X_t]$ is minimal, and the Y_t sequence is conditionally independent given the respective X_t variables). While that work identifies such provably-minimal assumptions for related settings (namely, inductive and self-adaptive learning), for the setting of online learning it only establishes a necessary condition and a sufficient condition which are provably distinct, leaving open the problem of exactly identifying the precise minimal condition admitting consistent learners (Hanneke, 2021b).

Nevertheless, the sufficient condition (Condition 1 below) established by (Hanneke, 2021a) is quite general, encompassing all previously-studied conditions on $\{X_t\}_{t=1}^\infty$ admitting universal consistency. It is therefore interesting to investigate whether this condition remains sufficient for universal consistency *without* restricting to deterministic responses $Y_t = f^*(X_t)$ (or conditional independence of Y_t values). In particular, it is most interesting to understand whether there is a family of processes $\{(X_t, Y_t)\}_{t=1}^\infty$ admitting universally consistent learners, and encompassing *all* previously-studied families admitting such consistent learners.

In this work, we propose such a family. Specifically, the family consists of all $\{(X_t, Y_t)\}_{t=1}^\infty$ where $\{X_t\}_{t=1}^\infty$ satisfies the condition suggested by (Hanneke, 2021a) (Condition 1 below). In particular, within this family, the sequence $\{Y_t\}_{t=1}^\infty$ is *completely unrestricted*. We establish this result for all bounded separable metric losses ℓ .

As an interesting implication, while past works established that there are universally consistent learners for the family of all stationary ergodic processes (Morvai, Yakowitz, and Györfi, 1996; Györfi, Lugosi, and Morvai, 1999; Györfi, Kohler, Krzyżak, and Walk, 2002), here we find that the requirement of ergodicity is completely *superfluous*: that is, *stationarity alone* is already sufficient for universally consistent online learning. Indeed, it even suffices if only the *input* sequence $\{X_t\}_{t=1}^\infty$ is stationary, while the response sequence $\{Y_t\}_{t=1}^\infty$ can be essentially *arbitrary*. As discussed below, the proposed family of processes also encompasses interesting families of *non-stationary* processes previously studied in the literature (e.g., Ryabko, 2006).

The algorithm achieving this result essentially uses an infinite variant of the Hedge algorithm for learning with expert advice, applied to a countable set of functions, which were shown by (Hanneke, 2021a) to be *dense*, in an appropriate sense defined relative to long-run averages observed in any process \mathbb{X} satisfying a key condition (Condition 1 below). A related approach has previously been shown to be consistent under stationary ergodic processes (Györfi, Lugosi, and Morvai, 1999).

2. Formal Setup and Main Result

Following the setup from (Hanneke, 2021a), we suppose \mathcal{X} is a non-empty set, equipped with a separable metrizable topology \mathcal{T} , and \mathcal{B} denotes the Borel σ -algebra generated by \mathcal{T} , defining the measurable subsets of \mathcal{X} . We allow (\mathcal{Y}, ℓ) to be any non-empty bounded separable metric space,

where for simplicity we suppose $\sup_{y,y' \in \mathcal{Y}} \ell(y, y') \leq 1$. For instance, this covers the classification setting, where \mathcal{Y} is any non-empty set and $\ell(y, y') = \mathbb{1}[y \neq y']$, as well as regression on $\mathcal{Y} = [0, 1]$ with the absolute loss $\ell(y, y') = |y - y'|$.

We will be interested in stochastic processes $(\mathbb{X}, \mathbb{Y}) := \{(X_t, Y_t)\}_{t=1}^{\infty}$ where each (X_t, Y_t) takes values in $\mathcal{X} \times \mathcal{Y}$. Following (Hanneke, 2021a), for any measurable $A \subseteq \mathcal{X}$, define $\hat{\mu}_{\mathbb{X}}(A) = \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n \mathbb{1}_A(X_t)$. Then we consider the following condition from (Hanneke, 2021a).

Condition 1 (Hanneke, 2021a) *For every monotone sequence $\{A_k\}_{k=1}^{\infty}$ of measurable subsets of \mathcal{X} with $A_k \downarrow \emptyset$,*

$$\lim_{k \rightarrow \infty} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(A_k)] = 0.$$

As discussed by (Hanneke, 2021a), this condition stipulates that $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(\cdot)]$ should behave as a *continuous submeasure*. For instance, (Hanneke, 2021a) shows that every process \mathbb{X} with convergent relative frequencies satisfies Condition 1, which includes all *stationary* processes, or even asymptotically mean stationary processes; indeed, in these cases, Birkhoff's ergodic theorem implies that $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(\cdot)]$ behaves as a *probability measure*, and hence is always continuous in the above sense. (Hanneke, 2021a) also argues that many interesting completely non-stationary processes satisfy the condition as well. Moreover, note that ergodicity is not needed for this to hold. Condition 1 will arise in our analysis as supplying the existence of a countable set $\tilde{\mathcal{F}}$ of functions which are *dense* in an appropriate sense.

We consider *online learning rules*: that is, sequences of measurable functions $f_t : \mathcal{X}^{t-1} \times \mathcal{Y}^{t-1} \times \mathcal{X} \rightarrow \mathcal{Y}$, which in general may be *randomized* (where the internal randomness should be independent from the data sequence). In this general setting, for any measurable $f_0 : \mathcal{X} \rightarrow \mathcal{Y}$, for any process $(\mathbb{X}, \mathbb{Y}) = \{(X_t, Y_t)\}_{t=1}^{\infty}$ on $\mathcal{X} \times \mathcal{Y}$, for any online learning rule f_t , we define the long-run average excess loss

$$\hat{\mathcal{L}}_{(\mathbb{X}, \mathbb{Y})}(f; f_0) = \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T (\ell(f_t(X_{1:(t-1)}, Y_{1:(t-1)}, X_t), Y_t) - \ell(f_0(X_t), Y_t)).$$

We are then interested in online learning rules f_t satisfying the guarantee that, for all measurable functions $f_0 : \mathcal{X} \rightarrow \mathcal{Y}$, it holds that $\hat{\mathcal{L}}_{(\mathbb{X}, \mathbb{Y})}(f; f_0) \leq 0$ almost surely, for every (\mathbb{X}, \mathbb{Y}) in some specific family of processes \mathcal{C} . Such an online learning rule f_t is said to be *strongly universally consistent* for \mathcal{C} . The main result is the following.

Theorem 1 *There is an online learning rule that is strongly universally consistent for the set of all processes (\mathbb{X}, \mathbb{Y}) such that \mathbb{X} satisfies Condition 1.*

In particular, this has the following immediate corollary.

Corollary 2 *There is an online learning rule that is strongly universally consistent for the set of all processes (\mathbb{X}, \mathbb{Y}) such that \mathbb{X} is stationary.*

Relation to prior results As discussed above, past works have established universal consistency for the set of all processes (\mathbb{X}, \mathbb{Y}) that are stationary *and ergodic* (e.g., [Morvai, Kulkarni, and Nobel, 1999](#); [Györfi, Lugosi, and Morvai, 1999](#); [Györfi, Kohler, Krzyżak, and Walk, 2002](#)), or otherwise satisfy a law of large numbers ([Morvai, Kulkarni, and Nobel, 1999](#); [Steinwart, Hush, and Scovel, 2009](#)). The work of [Ryabko \(2006\)](#) established universal consistency of certain learning rules for classification, under a family of processes that allow the Y_t sequence to be *arbitrary*, but restrict the X_t sequence to be conditionally i.i.d. given a Y_t value, and respecting that Y_t is determined by X_t . ([Hanneke, 2021a](#)) has shown that any process $\mathbb{X} = \{X_t\}_{t=1}^\infty$ satisfying the condition from [Ryabko \(2006\)](#) also satisfies Condition 1. ([Hanneke, 2021a](#)) established universal consistency for \mathbb{X} satisfying Condition 1, when $Y_t = f^*(X_t)$ for a fixed (arbitrary, unknown) measurable function $f^* : \mathcal{X} \rightarrow \mathcal{Y}$ (or more-generally, when $\mathbb{E}[\ell(f^*(X_t), Y_t) | X_t]$ is minimal, and the Y_t variables are conditionally independent given their respective X_t variables). We note that Theorem 1 unifies and generalizes all of these prior results on families of processes admitting universally consistent learners. Another interesting work is that of [Kulkarni, Posner, and Sandilya \(2002\)](#), which restricts the Y_t sequence to be conditionally independent with *continuous* conditional mean $f^*(X_t)$, but allows the X_t sequence to be *arbitrary*. They establish consistency of certain learning rules under these conditions. That work is, in a certain sense, *dual* to the present work, as we seek restrictions on the X_t sequence while allowing the Y_t sequence to be arbitrary. As such, the results cannot be directly compared. See ([Hanneke, 2021a](#)) for a thorough summary of past work on universal consistency for general families of stochastic processes.

3. Proof of the Theorem

The essential approach is to apply an infinite variant of the Hedge algorithm for learning with expert advice, where the experts are given by a well-chosen countable set of measurable functions which are “dense” in an appropriate sense. This is similar in spirit to certain existing strategies known to be consistent under stationary ergodic processes ([Györfi, Lugosi, and Morvai, 1999](#)).

Before proving Theorem 1, we introduce some useful results from the literature. We first state a well-known result from the literature on prediction with expert advice ([Cesa-Bianchi and Lugosi, 2006](#); [Hannan, 1957](#); [Vovk, 1990, 1992](#); [Littlestone and Warmuth, 1994](#); [Freund and Schapire, 1997](#); [Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire, and Warmuth, 1997](#); [Kivinen and Warmuth, 1999](#); [Singer and Feder, 1999](#); [Györfi and Lugosi, 2002](#)). See Corollary 4.2 of ([Cesa-Bianchi and Lugosi, 2006](#)) for this specific result.

Lemma 3 Fix $N, n \in \mathbb{N}$. Let y_1, y_2, \dots be a sequence in \mathcal{Y} . For each $t \in \{1, \dots, n\}$, let $\{z_{t,i}\}_{i=1}^N$ be a sequence of values in \mathcal{Y} . Let $\eta = \sqrt{(8/n) \ln(N)}$. For each $t \in \mathbb{N}$ and $i \in \{1, \dots, N\}$, define $L_{t,i} = \frac{1}{t} \sum_{s=1}^t \ell(z_{s,i}, y_s)$. Then for each $i \in \{1, \dots, N\}$, define $w_{1,i} = v_{1,i} = 1/N$, and for each $t \in \{2, \dots, n\}$, define $w_{t,i} = (1/N) e^{-\eta(t-1)L_{(t-1),i}}$, and $v_{t,i} = w_{t,i} / \sum_{j=1}^N w_{t,j}$. Finally, let $\{\hat{z}_t\}_{t \in \mathbb{N}}$ be independent \mathcal{Y} -valued random variables, with $\mathbb{P}(\hat{z}_t = z_{t,i}) = v_{t,i}$ for each $i \in \{1, \dots, N\}$ and $t \in \mathbb{N}$: that is, \hat{z}_t is a random sample from the $v_{t,i}$ -weighted distribution on $z_{t,i}$ values. Then, for

any fixed $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds that

$$\sum_{t=1}^n \ell(\hat{z}_t, y_t) \leq \min_{i \in \{1, \dots, N\}} n L_{n,i} + \sqrt{(1/2)n \ln(N)} + \sqrt{(1/2)n \ln\left(\frac{1}{\delta}\right)}.$$

In particular, this has the following implication for learning with an infinite number of experts.

Corollary 4 *Let y_1, y_2, \dots be a sequence in \mathcal{Y} (possibly stochastic). For each $t \in \mathbb{N}$, let $\{z_{t,i}\}_{i=1}^\infty$ be a sequence of values in \mathcal{Y} (possibly stochastic). Let $t_1 = 1$ and $T_1 = \{1\}$, and inductively define $t_{j+1} = t_j + j$ and $T_{j+1} = \{t_{j+1}, \dots, t_{j+1} + j\}$ for each $j \in \mathbb{N}$. Also for each $j \in \mathbb{N}$, define $\eta_j = \sqrt{(8/j) \ln(j)}$. For each $j \in \mathbb{N}$ and $t \in T_j$, for $i \in \{1, \dots, j\}$ define $L_{t,i} = \frac{1}{t-t_j+1} \sum_{s=t_j}^t \ell(z_{s,i}, y_s)$. For each $i \in \{1, \dots, j\}$, define $w_{t_j,i} = v_{t_j,i} = 1/j$, and for each $t \in T_j \setminus \{t_j\}$, define $w_{t,i} = (1/j)e^{-\eta_j(t-t_j)L_{t-1,i}}$, and $v_{t,i} = w_{t,i} / \sum_{i'=1}^j w_{t,i'}$. Finally, let $\{\hat{z}_t\}_{t \in \mathbb{N}}$ be \mathcal{Y} -valued random variables such that \hat{z}_t is conditionally independent of $\{\hat{z}_{t'}\}_{t' \neq t}$ given $\{y_{t'}\}_{t' < t}$ and $\{z_{t',i}\}_{i \in \mathbb{N}, t' \leq t}$, and for each $j \in \mathbb{N}$ and $t \in T_j$, $\mathbb{P}(\hat{z}_t = z_{t,i} | \{y_{t'}\}_{t' < t}, \{z_{t',i}\}_{i \in \mathbb{N}, t' \leq t}) = v_{t,i}$ for each $i \in \{1, \dots, j\}$: that is, \hat{z}_t is a random sample from the $v_{t,i}$ -weighted distribution on $\{z_{t,1}, \dots, z_{t,j}\}$. Then, with probability one, $\exists \hat{n} \in \mathbb{N}$ such that every $n \in \mathbb{N}$ with $n > \hat{n}$ satisfies*

$$\sum_{t=1}^n \ell(\hat{z}_t, y_t) \leq \left(\min_{1 \leq i \leq n^{1/4}} \sum_{t=1}^n \ell(z_{t,i}, y_t) \right) + 19 n^{3/4} \sqrt{\ln(n)} + \hat{n}.$$

Proof Applying Lemma 3 under the conditional distribution given $\{y_t\}_{t \in \mathbb{N}}$ and $\{z_{t,i}\}_{t,i \in \mathbb{N}}$, for every $j \in \mathbb{N}$, with conditional probability at least $1 - \frac{1}{j^2}$,

$$\sum_{t \in T_j} \ell(\hat{z}_t, y_t) \leq \left(\min_{i \in \{1, \dots, j\}} \sum_{t \in T_j} \ell(z_{t,i}, y_t) \right) + \left(1 + \frac{1}{\sqrt{2}} \right) \sqrt{j \ln(j)}.$$

By the law of total probability, this holds with (unconditional) probability at least $1 - \frac{1}{j^2}$ as well. Since $\sum_{j \in \mathbb{N}} \frac{1}{j^2} < \infty$, the Borel-Cantelli lemma implies that, with probability one, $\exists \hat{j} \in \mathbb{N}$ such that the above inequality holds for every $j \in \mathbb{N}$ with $j > \hat{j}$. For the remainder of the proof, we suppose this event occurs. Without loss of generality, we may suppose $\hat{j} \geq 2$.

For any $n \in \mathbb{N}$, let j_n be the index such that $n \in T_{j_n}$: namely, $j_n = \lceil \frac{1}{2} \sqrt{8n+1} - \frac{1}{2} \rceil$. Define $\hat{n} = \max T_{\hat{j}}$. Then for any $n \in \mathbb{N}$ with $n > \hat{n}$, letting $m(n) = \lceil \sqrt{n} \rceil$, we have

$$\begin{aligned} \sum_{t=1}^n \ell(\hat{z}_t, y_t) &\leq \sum_{j=1}^{j_n} \sum_{t \in T_j} \ell(\hat{z}_t, y_t) \\ &\leq \hat{n} + m(n) + j_{m(n)} + \sum_{j=j_{m(n)}+1}^{j_n} \left(\left(\min_{i \in \{1, \dots, j\}} \sum_{t \in T_j} \ell(z_{t,i}, y_t) \right) + \left(1 + \frac{1}{\sqrt{2}} \right) \sqrt{j \ln(j)} \right) \\ &\leq \hat{n} + m(n) + j_{m(n)} + \min_{i \in \{1, \dots, j_{m(n)}+1\}} \sum_{j=j_{m(n)}+1}^{j_n} \left(\left(\sum_{t \in T_j} \ell(z_{t,i}, y_t) \right) + \left(1 + \frac{1}{\sqrt{2}} \right) \sqrt{j \ln(j)} \right). \end{aligned}$$

We may then note that

$$\sum_{j=j_{m(n)}+1}^{j_n} \sum_{t \in T_j} \ell(z_{t,i}, y_t) \leq \sum_{t=1}^{t_{j_n+1}-1} \ell(z_{t,i}, y_t) \leq j_n + \sum_{t=1}^n \ell(z_{t,i}, y_t)$$

and

$$\begin{aligned} \sum_{j=j_{m(n)}+1}^{j_n} \left(1 + \frac{1}{\sqrt{2}}\right) \sqrt{j \ln(j)} &\leq \left(1 + \frac{1}{\sqrt{2}}\right) \sqrt{\ln(j_n)} \sum_{j=1}^{j_n} \sqrt{j} \\ &\leq \left(1 + \frac{1}{\sqrt{2}}\right) \sqrt{\ln(j_n)} \int_1^{j_n+1} \sqrt{x} dx \leq \left(1 + \frac{1}{\sqrt{2}}\right) \sqrt{\ln(j_n)} \frac{2}{3} (j_n + 1)^{3/2}. \end{aligned}$$

Altogether we have that

$$\begin{aligned} \sum_{t=1}^n \ell(\hat{z}_t, y_t) &\leq \hat{n} + m(n) + j_{m(n)} + \left(\min_{i \in \{1, \dots, j_{m(n)}+1\}} \sum_{t=1}^n \ell(z_{t,i}, y_t) \right) \\ &\quad + j_n + \left(1 + \frac{1}{\sqrt{2}}\right) \sqrt{\ln(j_n)} \frac{2}{3} (j_n + 1)^{3/2}. \end{aligned}$$

Noting that $j_n \leq \sqrt{6n}$ and $n^{1/4} - 1 \leq j_{m(n)} \leq 4n^{1/4}$, we have

$$\begin{aligned} \sum_{t=1}^n \ell(\hat{z}_t, y_t) &\leq \hat{n} + 1 + \sqrt{n} + 4n^{1/4} + \left(\min_{1 \leq i \leq n^{1/4}} \sum_{t=1}^n \ell(z_{t,i}, y_t) \right) \\ &\quad + \sqrt{6n} + \left(1 + \frac{1}{\sqrt{2}}\right) \sqrt{(1/2) \ln(6n)} \frac{2}{3} (\sqrt{6n} + 1)^{3/2} \\ &\leq \left(\min_{1 \leq i \leq n^{1/4}} \sum_{t=1}^n \ell(z_{t,i}, y_t) \right) + 19n^{3/4} \sqrt{\ln(n)} + \hat{n}. \end{aligned}$$

■

We will also use the following result, which is Lemma 35 from [\(Hanneke, 2021a\)](#).

Lemma 5 [\(Hanneke, 2021a, Lemma 35\)](#) Suppose $\{\beta_{k,n}\}_{k,n \in \mathbb{N}}$ is an array of values in $[0, \infty)$ such that $\lim_{k \rightarrow \infty} \limsup_{n \rightarrow \infty} \beta_{k,n} = 0$, and that $\{k_n\}_{n \in \mathbb{N}}$ is a sequence in \mathbb{N} with $k_n \rightarrow \infty$. Then there exists a sequence $\{j_n\}_{n \in \mathbb{N}}$ in \mathbb{N} such that $j_n \leq k_n$ for every $n \in \mathbb{N}$, and $\lim_{n \rightarrow \infty} \beta_{j_n, n} = 0$.

Finally, we will make use of one additional result. Following [\(Hanneke, 2021a\)](#), for a process $\mathbb{X} = \{X_t\}_{t=1}^\infty$ and a measurable function $g : \mathcal{X} \rightarrow \mathbb{R}$, define

$$\hat{\mu}_{\mathbb{X}}(g) := \limsup_{m \rightarrow \infty} \frac{1}{m} \sum_{t=1}^m g(X_t).$$

[\(Hanneke, 2021a\)](#) proves the following result.

Lemma 6 (*Hanneke, 2021a, Lemma 24*) *There exists a countable set $\tilde{\mathcal{F}}$ of measurable functions $\mathcal{X} \rightarrow \mathcal{Y}$ such that, for every measurable function $f : \mathcal{X} \rightarrow \mathcal{Y}$, for every process \mathbb{X} satisfying Condition 1,*

$$\inf_{\tilde{f} \in \tilde{\mathcal{F}}} \mathbb{E} \left[\hat{\mu}_{\mathbb{X}} \left(\ell \left(\tilde{f}(\cdot), f(\cdot) \right) \right) \right] = 0.$$

We are now ready for the proof of Theorem 1.

Proof of Theorem 1 Let $\tilde{\mathcal{F}}$ be as in Lemma 6, and enumerate $\tilde{\mathcal{F}} = \{\tilde{f}_1, \tilde{f}_2, \dots\}$.

Define the online learning rule \hat{f}_t as follows. For any sequences $x_{1:t} \in \mathcal{X}^t$ and $y_{1:(t-1)} \in \mathcal{Y}^{t-1}$, let $z_{s,i} = \tilde{f}_i(x_s)$ for each $s \leq t$ and $i \in \mathbb{N}$, and let \hat{z}_t be defined as in Corollary 4 (for these $z_{s,i}$ and y_s values). Note that \hat{z}_t is defined purely in terms of these values and internal randomness of the learner (i.e., the rest of the infinite sequence of y_s values ($s \geq t$) and array of $z_{s,i}$ values ($s > t$) needn't be defined for the purpose of defining \hat{z}_t). Finally, define $\hat{f}_t(x_{1:(t-1)}, y_{1:(t-1)}, x_t) := \hat{z}_t$ in this context.

Fix any process $(\mathbb{X}, \mathbb{Y}) = \{(X_t, Y_t)\}_{t=1}^{\infty}$ with \mathbb{X} satisfying Condition 1, and fix any measurable function $f_0 : \mathcal{X} \rightarrow \mathcal{Y}$. Let i_k be a sequence in \mathbb{N} such that

$$\mathbb{E} \left[\hat{\mu}_{\mathbb{X}} \left(\ell \left(\tilde{f}_{i_k}(\cdot), f_0(\cdot) \right) \right) \right] < 2^{-2k}$$

for all $k \in \mathbb{N}$, guaranteed to exist by the defining property of $\tilde{\mathcal{F}}$ from Lemma 6. By Markov's inequality, for each k , with probability at least $1 - 2^{-k}$,

$$\hat{\mu}_{\mathbb{X}} \left(\ell \left(\tilde{f}_{i_k}(\cdot), f_0(\cdot) \right) \right) < 2^{-k}. \quad (1)$$

Thus, since $\sum_{k \in \mathbb{N}} 2^{-k} < \infty$, by the Borel-Cantelli lemma, on an event E_0 of probability one, there exists $\kappa \in \mathbb{N}$ such that, for all $k \geq \kappa$, (1) holds. For each $n \in \mathbb{N}$, define

$$k_n = \max \{k : k \leq n, \max \{i_1, \dots, i_k\} \leq n^{1/4}\},$$

and note that $k_n \rightarrow \infty$.

By Corollary 4 and the definition of \hat{f}_t , there is an event E_1 of probability one, on which $\exists \hat{n} \in \mathbb{N}$ such that every $n \in \mathbb{N}$ with $n > \hat{n}$ satisfies

$$\frac{1}{n} \sum_{t=1}^n \ell \left(\hat{f}_t(X_{1:(t-1)}, Y_{1:(t-1)}, X_t), Y_t \right) \leq \left(\min_{1 \leq i \leq n^{1/4}} \frac{1}{n} \sum_{t=1}^n \ell \left(\tilde{f}_i(X_t), Y_t \right) \right) + 19 n^{-1/4} \sqrt{\log(n)} + \frac{\hat{n}}{n}.$$

Since $19 n^{-1/4} \sqrt{\log(n)} \rightarrow 0$ and $\frac{\hat{n}}{n} \rightarrow 0$ as $n \rightarrow \infty$, it suffices to focus on the first term on the right hand side. For this term, by the triangle inequality, we have

$$\min_{1 \leq i \leq n^{1/4}} \frac{1}{n} \sum_{t=1}^n \ell \left(\tilde{f}_i(X_t), Y_t \right) \leq \frac{1}{n} \sum_{t=1}^n \ell \left(f_0(X_t), Y_t \right) + \min_{1 \leq i \leq n^{1/4}} \frac{1}{n} \sum_{t=1}^n \ell \left(\tilde{f}_i(X_t), f_0(X_t) \right).$$

Thus, to complete the proof it suffices to argue that the second term on the right hand side approaches zero almost surely as $n \rightarrow \infty$.

For each $k, n \in \mathbb{N}$, let $\beta_{k,n} = \frac{1}{n} \sum_{t=1}^n \ell \left(\tilde{f}_{i_k}(X_t), f_0(X_t) \right)$. In particular, note that on the event E_0 , $\lim_{k \rightarrow \infty} \limsup_{n \rightarrow \infty} \beta_{k,n} \leq \lim_{k \rightarrow \infty} 2^{-k} = 0$. Therefore, Lemma 5 implies that, on the event E_0 , there exists a sequence $\{j_n\}_{n \in \mathbb{N}}$ in \mathbb{N} with $j_n \leq k_n$ for all $n \in \mathbb{N}$, such that $\lim_{n \rightarrow \infty} \beta_{j_n, n} = 0$.

Thus, on the event E_0 ,

$$\begin{aligned} \limsup_{n \rightarrow \infty} \min_{1 \leq i \leq n^{1/4}} \frac{1}{n} \sum_{t=1}^n \ell(\tilde{f}_i(X_t), f_0(X_t)) &\leq \limsup_{n \rightarrow \infty} \min_{1 \leq k \leq k_n} \frac{1}{n} \sum_{t=1}^n \ell(\tilde{f}_{i_k}(X_t), f_0(X_t)) \\ &\leq \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n \ell(\tilde{f}_{j_n}(X_t), f_0(X_t)) = \limsup_{n \rightarrow \infty} \beta_{j_n, n} = 0. \end{aligned}$$

Altogether, we have that on the event $E_0 \cap E_1$ of probability one (by the union bound),

$$\begin{aligned} \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n \left(\ell(\hat{f}_t(X_{1:(t-1)}, Y_{1:(t-1)}, X_t), Y_t) - \ell(f_0(X_t), Y_t) \right) \\ \leq \limsup_{n \rightarrow \infty} \beta_{j_n, n} + 19 n^{-1/4} \sqrt{\log(n)} + \frac{\hat{n}}{n} = 0. \end{aligned}$$

■

4. Open Problems

This result raises a number of further questions on the subject of universal consistency for online learners with general families of processes (\mathbb{X}, \mathbb{Y}) . One obvious question is whether the above result can be extended beyond metric losses. For instance, it is clearly desirable to at least extend the result to the problem of regression with the squared loss. Another obvious question is whether the learning strategy itself can be simplified, or the result established for other more-familiar learning strategies, such as partition-based learning rules, k -nearest neighbor predictors (when \mathcal{X} is finite-dimensional), or other local averaging predictors, with appropriate use of online regret arguments only in the selection of model complexity (e.g., the partition resolution in partition estimates, or choice of k in k -nearest neighbors), rather than in the selection of the entire function as in the strategy used here. This would extend such results which have been established for these learning rules under stationary ergodic processes (Györfi and Lugosi, 2002; Györfi and Ottucsák, 2012).

A more formal question in this context is whether there is a *largest* set $\mathcal{C}_{\mathcal{X}}$ of processes \mathbb{X} such that there exists an online learning rule that is strongly universally consistent for the family $\{(\mathbb{X}, \mathbb{Y}) : \mathbb{X} \in \mathcal{C}_{\mathcal{X}}\}$. More concretely, (Hanneke, 2021a) defines a set SUOL (strong universal online learning), the set of all processes \mathbb{X} such that there exists an online learning rule that is universally consistent for the set of all processes (\mathbb{X}, \mathbb{Y}) such that Y_t is a deterministic function of X_t : that is, there exists a measurable $f^* : \mathcal{X} \rightarrow \mathcal{Y}$ for which $Y_t = f^*(X_t)$. Hanneke (2021a,b) has posed the question of whether there exists an online learning rule that is strongly universally consistent for the family of all (\mathbb{X}, \mathbb{Y}) with $\mathbb{X} \in \text{SUOL}$ and Y_t a deterministic function of X_t (Hanneke, 2021b), known as an *optimistically* universal online learning rule (Hanneke, 2021a,b). Very recently, Blanchard (2022) has proposed a positive solution to this problem. In the context of the present work, we may instead pose a stronger version of this question:

Open Problem 1 *Does there exist an online learning rule that is strongly universally consistent for the family $\{(\mathbb{X}, \mathbb{Y}) : \mathbb{X} \in \text{SUOL}\}$?*

We note that a positive resolution of this question would be strictly stronger than the open problem of (Hanneke, 2021b) regarding the question of the existence of optimistically universal online learners, to which a solution has recently been proposed by Blanchard (2022). Moreover, if the above question is answered positively, then it would identify the *largest possible* set of processes \mathbb{X} for which, without any restrictions on \mathbb{Y} , there is an algorithm that is universally consistent for the family.

References

S. Ben-David, D. Pál, and S. Shalev-Shwartz. Agnostic online learning. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

M. Blanchard. Universal online learning: An optimistically universal learning rule. *arXiv:2201.05947*, 2022.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the Association for Computing Machinery*, 44(3):427–485, 1997.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag New York, 1996.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

L. Györfi and G. Lugosi. Strategies for sequential prediction of stationary time series. In M. Dror, P. L’Ecuyer, and F. Szidarovszky, editors, *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, pages 225–248. Kluwer Academic Publishers, 2002.

L. Györfi, G. Lugosi, and G. Morvai. A simple randomized algorithm for sequential prediction of ergodic time series. *IEEE Transactions on Information Theory*, 45(7):2642–2650, 1999.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag New York, 2002.

László Györfi and György Ottucsák. Nonparametric sequential prediction of stationary time series. *Machine Learning For Financial Engineering*, pages 179–226, 2012.

J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3: 97–139, 1957.

S. Hanneke. Learning whenever learning is possible: Universal learning under general stochastic processes. *Journal of Machine Learning Research*, 22:1–116, 2021a.

S. Hanneke. Open problem: Is there an online learning algorithm that learns whenever online learning is possible? In *Proceedings of the 34th Conference on Learning Theory*, 2021b.

S. Hanneke, A. Kontorovich, S. Sabato, and R. Weiss. Universal Bayes consistency in metric spaces. *The Annals of Statistics*, To appear, 2021.

J. Kivinen and M. K. Warmuth. Averaging expert predictions. In *Proceedings of the 4th European Conference on Computational Learning Theory*, 1999.

S. R. Kulkarni, S. E. Posner, and S. Sandilya. Data-dependent k_n -nn and kernel estimators consistent for arbitrary processes. *IEEE Transactions on Information Theory*, 48(10), 2002.

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

G. Morvai, S. Yakowitz, and L. Györfi. Nonparametric inference for ergodic, stationary time series. *The Annals of Statistics*, 24(1):370–379, 1996.

G. Morvai, S. R. Kulkarni, and A. B. Nobel. Regression estimation from an individual stable sequence. *Statistics*, 33:99–118, 1999.

A. Rakhlin, K. Sridharan, and A. Tewari. Online learning via sequential complexities. *Journal of Machine Learning Research*, 16(2):155–186, 2015.

D. Ryabko. Pattern recognition for conditionally independent data. *Journal of Machine Learning Research*, 7(4):645–664, 2006.

A. Singer and M. Feder. Universal linear prediction by model order weighting. *IEEE Transactions on Signal Processing*, 47(10):2685–2699, 1999.

I. Steinwart, D. Hush, and C. Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194, 2009.

C. J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620, 1977.

V. Vovk. Aggregating strategies. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, 1990.

V. Vovk. Universal forecasting algorithms. *Information and Computation*, 96(2):245–277, 1992.