Competing with stationary prediction strategies

Vladimir Vovk
vovk@cs.rhul.ac.uk
http://vovk.net

February 1, 2008

Abstract

In this paper we introduce the class of stationary prediction strategies and construct a prediction algorithm that asymptotically performs as well as the best continuous stationary strategy. We make mild compactness assumptions but no stochastic assumptions about the environment. In particular, no assumption of stationarity is made about the environment, and the stationarity of the considered strategies only means that they do not depend explicitly on time; we argue that it is natural to consider only stationary strategies even for highly non-stationary environments.

1 Introduction

This paper belongs to the area of learning theory that has been variously referred to as prediction with expert advice, competitive on-line prediction, prediction of individual sequences, and universal on-line learning; see [7] for a review. There are many proof techniques known in this field; this paper is based on Kalnishkan and Vyugin's Weak Aggregating Algorithm [16], but it is possible that some of the numerous other techniques could be used instead.

In Section 2 we give the main definitions and state our main results, Theorems 1–4; their proofs are given in Sections 3–6. In Section 7 we informally discuss the notion of stationarity, and Section 8 concludes.

2 Main results

The *game of prediction* between Predictor and Reality is played according to the following protocol (of *perfect information*, in the sense that either player can see the other player's moves made so far).

PREDICTION PROTOCOL

Reality announces $(\dots, x_{-1}, y_{-1}, x_0, y_0) \in (\mathbf{X} \times \mathbf{Y})^\infty$.
FOR $n = 1, 2, \dots$:

Reality announces $x_n \in \mathbf{X}$.
Predictor announces $\gamma_n \in \Gamma$.
Reality announces $y_n \in \mathbf{Y}$.
END FOR.

After Reality's first move the game proceeds in rounds numbered by the positive integers n. At the beginning of each round $n = 1, 2, \ldots$ Predictor is given some signal x_n relevant to predicting the following observation y_n. The signal is taken from the *signal space* $\mathbf{X}$ and the observations from the *observation space* $\mathbf{Y}$. Predictor then announces his prediction γ_n, taken from the *prediction space* Γ, and the prediction's quality in light of the actual observation is measured by a *loss function* $\lambda : \Gamma \times \mathbf{Y} \to \mathbb{R}$. At the beginning of the game Reality chooses the infinite past, (x_n, y_n) for all $n \le 0$.

In the games of prediction traditionally considered in machine learning there is no infinite past. This situation is modeled in our framework by extending the signal space and observation space by new elements $? \in \mathbf{X}$ and $? \in \mathbf{Y}$, defining $\lambda(\gamma, ?)$ arbitrarily, and making Reality announce the infinite past $(\ldots, x_{-1}, y_{-1}, x_0, y_0) = (\ldots, ?, ?, ?, ?)$ and refrain from announcing $x_n = ?$ or $y_n = ?$ afterwards (intuitively, ? corresponds to "no feedback from Reality").

We will always assume that the signal space $\mathbf{X}$, the prediction space Γ, and the observation space $\mathbf{Y}$ are non-empty topological spaces and that the loss function λ is continuous. Moreover, we are mainly interested in the case where $\mathbf{X}$, Γ, and $\mathbf{Y}$ are locally compact metric spaces, the prime examples being Euclidean spaces and their open and closed subsets. Our first results will be stated for the case where all three spaces $\mathbf{X}$, Γ, and $\mathbf{Y}$ are compact.

Remark Our results can be easily extended to the case where the loss on the nth round is allowed to depend, in addition to γ_n and y_n, on the past $\ldots, x_{n-1}, y_{n-1}, x_n$. This would, however, complicate the notation.

Predictor's strategies in the prediction protocol will be called *prediction strategies* (or *prediction algorithms*, when they are defined explicitly and we want to emphasize this). Mathematically such a strategy is a function $D : (\mathbf{X} \times \mathbf{Y})^\infty \times \mathbf{X} \times \{1, 2, \ldots\} \to \Gamma$; it maps each history $(\ldots, x_{n-1}, y_{n-1}, x_n)$ and the current time n to the chosen prediction. In this paper we will only be interested in continuous prediction strategies D (according to the traditional point of view [22], going back to Brouwer, only continuous prediction strategies can be computable; although it should be mentioned that nowadays there are influential definitions of computability [5, 4] not requiring continuity). An especially natural class of strategies is formed by the *stationary prediction strategies* $D : (\mathbf{X} \times \mathbf{Y})^\infty \times \mathbf{X} \to \Gamma$, which do not depend on time explicitly; since the origin of time is usually chosen arbitrarily, this appears a reasonable restriction (see Section 7 for a further discussion).

Universal prediction strategies: compact deterministic case

In this and next subsections we will assume that the spaces $\mathbf{X}, \Gamma, \mathbf{Y}$ are all compact. A prediction strategy is *CS universal* for a loss function λ if its

predictions γ_n satisfy

$$\limsup_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N}\lambda(\gamma_n,y_n)-\frac{1}{N}\sum_{n=1}^{N}\lambda\big(D(\dots,x_{n-1},y_{n-1},x_n),y_n\big)\right)\le 0 \qquad (1)$$

for any continuous stationary prediction strategy D and any biinfinite $\dots,x_{-1},y_{-1},x_0,y_0,x_1,y_1,\dots$. ("CS" refers to the continuity and stationarity of the prediction strategies we are competing with.)

Theorem 1 *Suppose* $\mathbf{X}$ *and* $\mathbf{Y}$ *are compact metric spaces,* Γ *is a compact convex subset of a Banach space, and the loss function* $\lambda(\gamma,y)$ *is continuous in* (γ,y) *and convex in the variable* $\gamma\in\Gamma$*. There exists a CS universal prediction algorithm.*

A CS universal prediction algorithm will be constructed in the next section.

Universal prediction strategies: compact randomized case

When the loss function $\lambda(\gamma,y)$ is not convex in γ, two difficulties appear:

- the conclusion of Theorem 1 becomes false if the convexity requirement is removed ([16], Theorem 2);
- in some cases the notion of a continuous prediction strategy becomes vacuous: e.g., there are no non-constant continuous stationary prediction strategies when $\Gamma=\{0,1\}$ and $(\mathbf{X}\times\mathbf{Y})^\infty\times\mathbf{X}$ is connected (the latter condition is equivalent to $\mathbf{X}$ and $\mathbf{Y}$ being connected—see [11], Theorem 6.1.15).

To overcome these difficulties, we consider randomized prediction strategies. The proof of Theorem 1 will give a universal, in a natural sense, randomized prediction algorithm; on the other hand, there will be a vast supply of continuous stationary prediction strategies.

Remark In fact, the second difficulty is more apparent than real: for example, in the binary case ($\mathbf{Y}=\{0,1\}$) there are many non-trivial continuous prediction strategies in the canonical form of the prediction game [30] with the prediction space redefined as the boundary of the set of superpredictions [16].

A *randomized prediction strategy* is a function $D:(\mathbf{X}\times\mathbf{Y})^\infty\times\mathbf{X}\times\{1,2,\dots\}\to\mathcal{P}(\Gamma)$ mapping the past complemented by the current time to the probability measures on the prediction space; $\mathcal{P}(\Gamma)$ is always equipped with the topology of weak convergence ([3]; this topology is also discussed, in the compact case, in Section 4 below). In other words, this is a prediction strategy in the extended game of prediction with the prediction space $\mathcal{P}(\Gamma)$. Analogously, a *stationary randomized prediction strategy* is a function $D:(\mathbf{X}\times\mathbf{Y})^\infty\times\mathbf{X}\to\mathcal{P}(\Gamma)$.

Let us say that a randomized prediction strategy outputting γ_n is *CS universal* for a loss function λ if, for any continuous stationary randomized prediction strategy D and any biinfinite $\ldots, x_{-1}, y_{-1}, x_0, y_0, x_1, y_1, \ldots$,

$$\limsup_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N}\lambda(g_n, y_n) - \frac{1}{N}\sum_{n=1}^{N}\lambda(d_n, y_n)\right) \le 0 \text{ a.s.}, \tag{2}$$

where $g_1, g_2, \ldots, d_1, d_2, \ldots$ are independent random variables distributed as

$$g_n \sim \gamma_n, \tag{3}$$
$$d_n \sim D(\ldots, x_{n-1}, y_{n-1}, x_n), \tag{4}$$

$n = 1, 2, \ldots$. Intuitively, the "a.s." in (2) refers to the prediction strategies' internal randomization.

Theorem 2 *Let* $\mathbf{X}$, Γ, *and* $\mathbf{Y}$ *be compact metric spaces and* λ *be a continuous loss function. There exists a CS universal randomized prediction algorithm.*

Simple reductions to the compact case

In the following two subsections we will discuss the case where the signal, prediction, and observation spaces are not required to be compact. The goal of this subsection is to show that the compact case is not as special as it may seem, as far as Theorem 2 is concerned. The rest of the paper does not depend on this subsection.

In general, we might consider $\mathbf{X}$, Γ, and $\mathbf{Y}$ together with their fixed compactifications $\overline{\mathbf{X}}$, $\overline{\Gamma}$, and $\overline{\mathbf{Y}}$ (without loss of generality we can and will assume that $\mathbf{X}$, Γ, and $\mathbf{Y}$ are dense in their compactifications, and then the compactifications will be the closures of the original spaces, which explains our notation). Let us suppose that λ is bounded and continuous, and, moreover, can be continuously extended to the product $\overline{\Gamma} \times \overline{\mathbf{Y}}$ of the compactifications; such an extension is then unique and will also be denoted λ.

If $\mathbf{X}$, Γ, and $\mathbf{Y}$ are Euclidean spaces their natural compactifications might be chosen as Aleksandrov's one-point compactification ([11], Theorem 3.5.11), the corresponding projective space (with $\mathbb{R}\mathrm{P}^L$ being the compactification of $\mathbb{R}^L$), or the corresponding closed unit ball (with the interior of the closed unit ball in $\mathbb{R}^L$ identified with $\mathbb{R}^L$ by mapping a vector v of length $l \in [0, 1)$ in the former set to the vector $(\tan(\pi l/2))v$). The Stone–Čech compactification ([11], Section 3.6) will usually be too large: we will want our compactifications to be metrizable.

Theorem 2 will remain true if instead of assuming $\mathbf{X}$, Γ, and $\mathbf{Y}$ to be metric compacts we assume that $\overline{\mathbf{X}}$, $\overline{\Gamma}$, and $\overline{\mathbf{Y}}$ are metric compacts and if in the definition of CS universality (2) we only consider continuous stationary prediction strategies that have a continuous extension to $(\overline{\mathbf{X}} \times \overline{\mathbf{Y}})^\infty \times \overline{\mathbf{X}}$.

Remark An elegant way to avoid considering compactifications would be to assume that $\mathbf{X}$, Γ, and $\mathbf{Y}$ are metrizable proximity spaces (see [11], Section 8.4, or [23], where [11]'s "proximity spaces" are called "separated proximity spaces")

and to consider only proximity prediction strategies. By Smirnov's theorem ([11], Theorem 8.4.13 and also Theorem 8.4.9; [23], Theorem 7.7) a proximity space can be identified with the corresponding topological space equipped with a compactification. Assuming that the loss function λ is a bounded proximity function, it can be uniquely continuously extended to the compactification $\overline{\Gamma}\times\overline{\mathbf{Y}}$ ([23], Theorem 7.10), and every proximity stationary prediction strategy can be identified with a continuous function on the compactification $(\overline{\mathbf{X}}\times\overline{\mathbf{Y}})^\infty\times\overline{\mathbf{X}}$ (by the same theorem). To ensure that the compactifications are metrizable, it is sufficient to assume that the proximity spaces are second-countable (i.e., have countable proximity weights; see [23], Theorem 8.14, and [11], Theorem 4.2.8). We chose the slightly clumsier language of compactifications because the notion of a topological space is much more familiar than that of a proximity space.

Universal prediction strategies: deterministic case

Let us say that a set in a topological space is *precompact* if its closure is compact. In Euclidean spaces, precompactness means boundedness. In this and next subsections we drop the assumption of compactness of $\mathbf{X}$, Γ, and $\mathbf{Y}$, and so we have to redefine the notion of CS universality.

A prediction strategy outputting $\gamma_n\in\mathcal{P}(\Gamma)$ is *CS universal* for a loss function λ if, for any continuous stationary prediction strategy D and for any biinfinite $\dots,x_{-1},y_{-1},x_0,y_0,x_1,y_1,\dots$,

$$\begin{aligned}&\big(\{\dots,x_{-1},x_0,x_1,\dots\}\text{ and }\{\dots,y_{-1},y_0,y_1,\dots\}\text{ are precompact}\big)\\&\Longrightarrow\limsup_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^N\lambda(\gamma_n,y_n)-\frac{1}{N}\sum_{n=1}^N\lambda\big(D(\dots,x_{n-1},y_{n-1},x_n),y_n\big)\right)\le 0.\end{aligned}\tag{5}$$

The intuition behind the antecedent of (5), in the Euclidean case, is that the prediction algorithm knows that $\|x_n\|$ and $\|y_n\|$ are bounded but does not know an upper bound in advance.

Let us say that the loss function λ is *large at infinity* if, for all $y^*\in\mathbf{Y}$,

$$\lim_{\substack{y\to y^*\\ \gamma\to\infty}}\lambda(\gamma,y)=\infty$$

(in the sense that for each constant M there exists a neighborhood $O_{y^*}\ni y^*$ and compact $C\subseteq\Gamma$ such that $\lambda\left(\Gamma\setminus C,O_{y^*}\right)\subseteq(M,\infty)$). Intuitively, we require that faraway $\gamma\in\Gamma$ should be poor predictions for nearby $y^*\in\mathbf{Y}$. This assumption is satisfied for most of the usual loss functions used in competitive on-line prediction.

Theorem 3 *Suppose $\mathbf{X}$ and $\mathbf{Y}$ are locally compact metric spaces, Γ is a convex subset of a Banach space, and the loss function $\lambda(\gamma,y)$ is continuous, large at infinity, and convex in the variable $\gamma\in\Gamma$. There exists a CS universal prediction algorithm.*

To have a specific example in mind, the reader might check that $\mathbf{X} = \mathbb{R}^K$, $\Gamma = \mathbf{Y} = \mathbb{R}^L$, and $\lambda(\gamma, y) := \|y - \gamma\|$ satisfy the conditions of the theorem.

Universal prediction strategies: randomized case

We say that a randomized prediction strategy outputting randomized predictions γ_n is *CS universal* if, for any continuous stationary randomized prediction strategy D and for any biinfinite $\ldots, x_{-1}, y_{-1}, x_0, y_0, x_1, y_1, \ldots$,

$$\bigl(\{\ldots, x_{-1}, x_0, x_1, \ldots\} \text{ and } \{\ldots, y_{-1}, y_0, y_1, \ldots\} \text{ are precompact}\bigr) \\ \Longrightarrow \left(\limsup_{N\to\infty} \left(\frac{1}{N} \sum_{n=1}^{N} \lambda(g_n, y_n) - \frac{1}{N} \sum_{n=1}^{N} \lambda(d_n, y_n) \right) \le 0 \text{ a.s.} \right), \quad (6)$$

where $g_1, g_2, \ldots, d_1, d_2, \ldots$ are independent random variables distributed according to (3)–(4).

Theorem 4 *Let $\mathbf{X}$ and $\mathbf{Y}$ be locally compact metric spaces, Γ be a metric space, and λ be a continuous and large at infinity loss function. There exists a CS universal randomized prediction algorithm.*

3 Proof of Theorem 1

In the rest of the paper we will be using the notation Σ for $(\mathbf{X} \times \mathbf{Y})^\infty \times \mathbf{X}$. By Tikhonov's theorem ([11], Theorem 3.2.4) this is a compact space; it is also metrizable ([11], Theorem 4.2.2). Another standard piece of notation throughout the rest of the paper will be $\sigma_n := (\ldots, x_{n-1}, y_{n-1}, x_n) \in \Sigma$. Remember that λ, as a continuous function on a compact set, is bounded below and above ([11], Theorem 3.10.6).

Let Γ^Σ be the set of all continuous functions from Σ to Γ with the *topology of uniform convergence*, generated by the metric

$$\hat{\rho}(D_1, D_2) := \sup_{\sigma\in\Sigma} \rho\bigl(D_1(\sigma), D_2(\sigma)\bigr),$$

ρ being the metric in Γ (induced by the norm in the containing Banach space). Since the topological space Γ^Σ is separable ([11], Corollary 4.2.18 in combination with Theorem 4.2.8), we can choose a dense sequence $D_1, D_2, \ldots$ in Γ^Σ.

Remark The topology in Γ^Σ is defined via a metric, and this is one the very few places in this paper where we need a specific metric (for brevity we often talk about "metric spaces", but this can always be replaced by "metrizable topological spaces"). Without using the metric, we could say that the topology in Γ^Σ is the compact-open topology ([11], Section 3.4). Since Σ is compact, the compact-open topology on Γ^Σ coincides with the topology of uniform convergence ([11], Theorem 4.2.17). The separability of Γ^Σ now follows from [11], Theorem 3.4.16 in combination with Theorem 4.2.8.

The next step is to apply Kalnishkan and Vyugin's [16] Weak Aggregating Algorithm (WAA) to this sequence. We cannot just refer to [16] and will have to redo their derivation of the WAA's main property since Kalnishkan and Vyugin only consider the case of finitely many "experts" D_k and finite $\mathbf{Y}$. (Although in other respects we will not need their algorithm in full generality and so slightly simplify it.)

Let $q_1, q_2, \ldots$ be a sequence of positive numbers summing to 1, $\sum_{k=1}^{\infty} q_k = 1$. Define

$$l_n^{(k)} := \lambda\left(D_k(\sigma_n), y_n\right), \quad L_N^{(k)} := \sum_{n=1}^{N} l_n^{(k)}$$

to be the instantaneous loss of the kth expert D_k on the nth round and his cumulative loss over the first N rounds. For all $n, k = 1, 2, \ldots$ define

$$w_n^{(k)} := q_k \beta_n^{L_{n-1}^{(k)}}, \quad \beta_n := \exp\left(-\frac{1}{\sqrt{n}}\right)$$

($w_n^{(k)}$ are the weights of the experts to use on round n) and

$$p_n^{(k)} := \frac{w_n^{(k)}}{\sum_{k=1}^{\infty} w_n^{(k)}}$$

(the normalized weights; it is obvious that the denominator is positive and finite). The WAA's prediction on round n is

$$\gamma_n := \sum_{k=1}^{\infty} p_n^{(k)} D_k(\sigma_n) \tag{7}$$

(the series is convergent in the Banach space since the compactness of Γ implies $\sup_{\gamma \in \Gamma} \|\gamma\| < \infty$, and $\gamma_n \in \Gamma$ since

$$\gamma_n - \sum_{k=1}^{K} \frac{p_n^{(k)}}{\sum_{k=1}^{K} p_n^{(k)}} D_k(\sigma_n)$$
$$= \sum_{k=1}^{K} \left(1 - \frac{1}{\sum_{k=1}^{K} p_n^{(k)}}\right) p_n^{(k)} D_k(\sigma_n) + \sum_{k=K+1}^{\infty} p_n^{(k)} D_k(\sigma_n) \to 0 \tag{8}$$

as $K \to \infty$).

Let $l_n := \lambda(\gamma_n, y_n)$ be the WAA's loss on round n and $L_N := \sum_{n=1}^{N} l_n$ be its cumulative loss over the first N rounds.

Lemma 1 ([16], Lemma 9) *The WAA guarantees that, for all N,*

$$L_N \le \sum_{n=1}^{N} \sum_{k=1}^{\infty} p_n^{(k)} l_n^{(k)} - \sum_{n=1}^{N} \log_{\beta_n} \sum_{k=1}^{\infty} p_n^{(k)} \beta_n^{l_n^{(k)}} + \log_{\beta_N} \sum_{k=1}^{\infty} q_k \beta_N^{L_N^{(k)}}. \tag{9}$$

The first two terms on the right-hand side of (9) are sums over the first N rounds of different kinds of mean of the experts' losses (see, e.g., [15], Chapter III, for a general definition of the mean); we will see later that they nearly cancel each other out. If those two terms are ignored, the remaining part of (9) is identical (except that β now depends on n) to the main property of the "Aggregating Algorithm" (see, e.g., [31], Lemma 1). All infinite series in (9) are trivially convergent.

Proof of Lemma 1 The proof is by induction on N. Assuming (9), we obtain

$$\begin{aligned} L_{N+1} = L_N + l_{N+1} &\le L_N + \sum_{k=1}^{\infty} p_{N+1}^{(k)} l_{N+1}^{(k)} \\ &\le \sum_{n=1}^{N+1} \sum_{k=1}^{\infty} p_n^{(k)} l_n^{(k)} - \sum_{n=1}^{N} \log_{\beta_n} \sum_{k=1}^{\infty} p_n^{(k)} \beta_n^{l_n^{(k)}} + \log_{\beta_N} \sum_{k=1}^{\infty} q_k \beta_N^{L_N^{(k)}} \end{aligned}$$

(the first "$\le$" used the "countable convexity" $l_n \le \sum_{k=1}^{\infty} p_n^{(k)} l_n^{(k)}$, which follows from (8) and

$$\lambda \left(\sum_{k=1}^{K} \frac{p_n^{(k)}}{\sum_{k=1}^{K} p_n^{(k)}} D_k(\sigma_n), y_n \right) \le \sum_{k=1}^{K} \frac{p_n^{(k)}}{\sum_{k=1}^{K} p_n^{(k)}} \lambda \left(D_k(\sigma_n), y_n \right)$$

if we let $K \to \infty$). Therefore, it remains to prove

$$\log_{\beta_N} \sum_{k=1}^{\infty} q_k \beta_N^{L_N^{(k)}} \le - \log_{\beta_{N+1}} \sum_{k=1}^{\infty} p_{N+1}^{(k)} \beta_{N+1}^{l_{N+1}^{(k)}} + \log_{\beta_{N+1}} \sum_{k=1}^{\infty} q_k \beta_{N+1}^{L_{N+1}^{(k)}}.$$

By the definition of $p_n^{(k)}$ this can be rewritten as

$$\log_{\beta_N} \sum_{k=1}^{\infty} q_k \beta_N^{L_N^{(k)}} \le - \log_{\beta_{N+1}} \frac{\sum_{k=1}^{\infty} q_k \beta_{N+1}^{L_N^{(k)}} \beta_{N+1}^{l_{N+1}^{(k)}}}{\sum_{k=1}^{\infty} q_k \beta_{N+1}^{L_N^{(k)}}} + \log_{\beta_{N+1}} \sum_{k=1}^{\infty} q_k \beta_{N+1}^{L_{N+1}^{(k)}},$$

which after cancellation becomes

$$\log_{\beta_N} \sum_{k=1}^{\infty} q_k \beta_N^{L_N^{(k)}} \le \log_{\beta_{N+1}} \sum_{k=1}^{\infty} q_k \beta_{N+1}^{L_N^{(k)}}. \tag{10}$$

The last inequality follows from the general result about comparison of different means ([15], Theorem 85), but we can also check it directly (following [16]). Let $\beta_{N+1} = \beta_N^a$, where $0 < a < 1$. Then (10) can be rewritten as

$$\left(\sum_{k=1}^{\infty} q_k \beta_N^{L_N^{(k)}} \right)^a \ge \sum_{k=1}^{\infty} q_k \beta_N^{a L_N^{(k)}},$$

and the last inequality follows from the concavity of the function $t \mapsto t^a$. ∎

Lemma 2 ([16], Lemma 5) *Let L be an upper bound on $|\lambda|$. The WAA guarantees that, for all N and K,*

$$L_N \le L_N^{(K)} + \left(L^2 e^L + \ln\frac{1}{q_K}\right)\sqrt{N}. \tag{11}$$

(There is no term e^L in [16] since it only considers non-negative loss functions.)

Proof From (9), we obtain:

$$\begin{aligned}
L_N &\le \sum_{n=1}^{N}\sum_{k=1}^{\infty} p_n^{(k)} l_n^{(k)} + \sum_{n=1}^{N}\sqrt{n}\ln\sum_{k=1}^{\infty} p_n^{(k)}\exp\left(-\frac{l_n^{(k)}}{\sqrt{n}}\right) + \log_{\beta_N} q_K + L_N^{(K)} \\
&\le \sum_{n=1}^{N}\sum_{k=1}^{\infty} p_n^{(k)} l_n^{(k)} + \sum_{n=1}^{N}\sqrt{n}\left(\sum_{k=1}^{\infty} p_n^{(k)}\left(1 - \frac{l_n^{(k)}}{\sqrt{n}} + \frac{\left(l_n^{(k)}\right)^2}{2n}e^L\right) - 1\right) \\
&\quad + \log_{\beta_N} q_K + L_N^{(K)} \\
&= L_N^{(K)} + \frac{1}{2}\sum_{n=1}^{N}\frac{1}{\sqrt{n}}\sum_{k=1}^{\infty} p_n^{(k)}\left(l_n^{(k)}\right)^2 e^L + \sqrt{N}\ln\frac{1}{q_K} \\
&\le L_N^{(K)} + \frac{L^2 e^L}{2}\sum_{n=1}^{N}\frac{1}{\sqrt{n}} + \sqrt{N}\ln\frac{1}{q_K} \le L_N^{(K)} + \frac{L^2 e^L}{2}\int_0^N \frac{\mathrm{d}t}{\sqrt{t}} + \sqrt{N}\ln\frac{1}{q_K} \\
&\le L_N^{(K)} + L^2 e^L\sqrt{N} + \sqrt{N}\ln\frac{1}{q_K}
\end{aligned}$$

(in the second "$\le$" we used the inequalities $e^t \le 1 + t + \frac{t^2}{2}e^{|t|}$ and $\ln t \le t - 1$). ∎

Now it is easy to prove Theorem 1. Let γ_n be the predictions output by the WAA. Consider any continuous stationary prediction strategy D. Since every continuous function on a metric compact is uniformly continuous ([11], Theorem 4.3.32), for any $\epsilon > 0$ we can find $\delta > 0$ such that $|\lambda(\gamma_1, y) - \lambda(\gamma_2, y)| < \epsilon$ whenever $\rho(\gamma_1, \gamma_2) < \delta$. We can further find K such that $\hat{\rho}(D_K, D) < \delta$, and (11) then gives, for all biinfinite $\ldots, x_{-1}, y_{-1}, x_0, y_0, x_1, y_1, \ldots$,

$$\begin{aligned}
&\limsup_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N}\lambda(\gamma_n, y_n) - \frac{1}{N}\sum_{n=1}^{N}\lambda(D(\sigma_n), y_n)\right) \\
&\qquad \le \limsup_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N}\lambda(\gamma_n, y_n) - \frac{1}{N}\sum_{n=1}^{N}\lambda(D_K(\sigma_n), y_n)\right) + \epsilon \\
&\qquad\qquad \le \limsup_{N\to\infty}\left(L^2 e^L + \ln\frac{1}{q_K}\right)\frac{1}{\sqrt{N}} + \epsilon = \epsilon;
\end{aligned}$$

since ϵ can be arbitrarily small the WAA is CS universal.

4 Proof of Theorem 2

Let us first recall some useful facts about the probability measures on a metric compact Ω (we will be following [32]). The Banach space of all continuous real-valued functions on Ω with the usual pointwise addition and scalar action and the sup norm will be denoted $C(\Omega)$. By one of the Riesz representation theorems ([10], 7.4.1; see also 7.1.1), the mapping $\mu \mapsto I_\mu$, where $I_\mu(f) := \int_\Omega f \, \mathrm{d}\mu$, is a linear isometry between the set of all finite Borel signed measures μ on Ω with the total variation norm and the dual space $C'(\Omega)$ to $C(\Omega)$ with the standard dual norm ([27], Chapter 4). We will identify the finite Borel signed measures μ on Ω with the corresponding $I_\mu \in C'(\Omega)$. This makes the set $\mathcal{P}(\Omega)$ of probability measures on Ω a convex closed subset of $C'(\Omega)$.

We will be interested, however, in a different topology on $C'(\Omega)$, the weakest topology for which all evaluation functionals $\mu \in C'(\Omega) \mapsto \mu(f)$, $f \in C(\Omega)$, are continuous. This topology is known as the *weak* topology* ([27], 3.14), and the topology inherited by $\mathcal{P}(\Omega)$ is known as the *topology of weak convergence* ([3], Appendix III). The point mass δ_ω, $\omega \in \Omega$, is defined to be the probability measure concentrated at ω, $\delta_\omega(\{\omega\}) = 1$. The simple example of a sequence of point masses δ_{ω_n} such that $\omega_n \to \omega$ as $n \to \infty$ and $\omega_n \neq \omega$ for all n shows that the topology of weak convergence is different from the dual norm topology: $\delta_{\omega_n} \to \delta_\omega$ holds in one but does not hold in the other.

It is not difficult to check that $\mathcal{P}(\Omega)$ remains a closed subset of $C'(\Omega)$ in the weak* topology ([6], III.2.7, Proposition 7). By the Banach–Alaoglu theorem ([27], 3.15) $\mathcal{P}(\Omega)$ is compact in the topology of weak convergence (this is a special case of Prokhorov's theorem, [3], Appendix III, Theorem 6). In the rest of this paper, $\mathcal{P}(\Omega)$ (and all other spaces of probability measures) are always equipped with the topology of weak convergence.

Since Ω is a metric compact, $\mathcal{P}(\Omega)$ is also metrizable (by the well-known Prokhorov metric: [3], Appendix III, Theorem 6).

Define

$$\lambda(\gamma, y) := \int_\Gamma \lambda(g, y) \gamma(\mathrm{d}g), \tag{12}$$

where γ is a probability measure on Γ. This is the loss function in a new game of prediction with the prediction space $\mathcal{P}(\Gamma)$; it is convex in γ.

Let us check that the loss function (12) is continuous. If $\gamma_n \to \gamma$ and $y_n \to y$ for some $(\gamma, y) \in \mathcal{P}(\Gamma) \times \mathbf{Y}$,

$$|\lambda(\gamma_n, y_n) - \lambda(\gamma, y)| \le |\lambda(\gamma_n, y_n) - \lambda(\gamma_n, y)| + |\lambda(\gamma_n, y) - \lambda(\gamma, y)| \to 0$$

(the first addend tends to zero because of the uniform continuity of $\lambda : \Gamma \times \mathbf{Y} \to \mathbb{R}$ and the second addend by the definition of the topology of weak convergence).

Unfortunately, Theorem 1 cannot be applied to the new game of prediction directly: the theorem assumes that Γ is a subset of a Banach space, whereas the dual to an infinite-dimensional Banach space is never even metrizable in the weak* topology ([27], 3.16). The proof of Theorem 1, however, still works for the new game.

It is clear that the mixture (7) is a probability measure. The result of the previous section is still true, and the randomized prediction strategy (7) produces $\gamma_n \in \mathcal{P}(\Gamma)$ that are guaranteed to satisfy

$$\limsup_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N}\lambda(\gamma_n,y_n)-\frac{1}{N}\sum_{n=1}^{N}\lambda(D(\sigma_n),y_n)\right)\le 0, \tag{13}$$

for any continuous stationary randomized prediction strategy D. The loss function is bounded in absolute value by a constant L, and so the law of the iterated logarithm (see, e.g., [28], (5.8)) implies that

$$\limsup_{N\to\infty}\frac{\left|\sum_{n=1}^{N}\bigl(\lambda(g_n,y_n)-\lambda(\gamma_n,y_n)\bigr)\right|}{\sqrt{2L^2N\ln\ln N}}\le 1, \tag{14}$$

$$\limsup_{N\to\infty}\frac{\left|\sum_{n=1}^{N}\bigl(\lambda(d_n,y_n)-\lambda(D(\sigma_n),y_n)\bigr)\right|}{\sqrt{2L^2N\ln\ln N}}\le 1 \tag{15}$$

with probability one. Combining the last two inequalities with (13) gives

$$\limsup_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N}\lambda(g_n,y_n)-\frac{1}{N}\sum_{n=1}^{N}\lambda(d_n,y_n)\right)\le 0 \text{ a.s.}$$

Therefore, the WAA (applied to $D_1, D_2, \ldots$) is a universal continuous randomized prediction strategy.

5 Proof of Theorem 3

In view of Theorem 1, we only need to get rid of the assumption of compactness of $\mathbf{X}$, Γ, and $\mathbf{Y}$.

Game of removal

The proofs of Theorems 3 and 4 will be based on the following game (an abstract version of the "doubling trick", [7]) played in a topological space X:

GAME OF REMOVAL $G(X)$

FOR $n = 1, 2, \ldots$:
  Remover announces compact $K_n \subseteq X$.
  Evader announces $p_n \notin K_n$.
END FOR.

Winner: Evader if the set $\{p_1, p_2, \ldots\}$ is precompact; Remover otherwise.

Intuitively, the goal of Evader is to avoid being removed to the infinity. Without loss of generality we will assume that Remover always announces a non-decreasing sequence of compact sets: $K_1 \subseteq K_2 \subseteq \cdots$.

Lemma 3 (Gruenhage) *Remover has a winning strategy in $G(X)$ if X is a locally compact and paracompact space.*

Proof We will follow the proof of Theorem 4.1 in [12] (the easy direction). If X is locally compact and σ-compact, there exists a non-decreasing sequence $K_1 \subseteq K_2 \subseteq \cdots$ of compact sets covering X, and each K_n can be extended to compact K_n^* so that $\operatorname{Int} K_n^* \supseteq K_n$ ([11], Theorem 3.3.2). Remover will obviously win $G(X)$ choosing $K_1^*, K_2^*, \ldots$ as his moves.

If X is the sum of locally compact σ-compact spaces X_s, $s \in S$, Remover plays, for each $s \in S$, the strategy described in the previous paragraph on the subsequence of Evader's moves belonging to X_s. If Evader chooses $p_n \in X_s$ for infinitely many X_s, those X_s will form an open cover of the closure of $\{p_1, p_2, \ldots\}$ without a finite subcover. If x_n are chosen from only finitely many X_s, there will be infinitely many x_n chosen from some X_s, and the result of the previous paragraph can be applied. It remains to remember that each locally compact paracompact can be represented as the sum of locally compact σ-compact subsets ([11], Theorem 5.1.27). ∎

Large at infinity loss functions

We will need the following useful property of large at infinity loss functions.

Lemma 4 *Let λ be a loss function that is large at infinity. For each compact set $B \subseteq \mathbf{Y}$ and each constant M there exists a compact set $C \subseteq \Gamma$ such that*

$$\forall \gamma \notin C, y \in B : \quad \lambda(\gamma, y) > M. \tag{16}$$

Proof For each point $y^* \in B$ fix a neighborhood $O_{y^*} \ni y^*$ and a compact set $C(y^*) \subseteq \Gamma$ such that $\lambda(\Gamma \setminus C(y^*), O_{y^*}) \subseteq (M, \infty)$. Since the sets O_{y^*} form an open cover of B, we can find this cover's finite subcover $\{O_{y_1^*}, \ldots, O_{y_n^*}\}$. It is clear that

$$C := \bigcup_{j=1,\ldots,n} C\left(O_{y_j^*}\right)$$

satisfies (16). ∎

In fact, the only property of large at infinity loss functions that we will be using is that in the conclusion of Lemma 4. In particular, it implies the following lemma.

Lemma 5 *Under the conditions of Theorem 3, for each compact set $B \subseteq \mathbf{Y}$ there exists a compact convex set $C = C(B) \subseteq \Gamma$ such that for each continuous stationary prediction strategy $D : \Sigma \to \Gamma$ there exists a continuous stationary prediction strategy $D' : \Sigma \to C$ that dominates D in the sense*

$$\forall \sigma \in \Sigma, y \in B : \quad \lambda(D'(\sigma), y) \le \lambda(D(\sigma), y). \tag{17}$$

Proof Without loss of generality B is assumed non-empty. Fix any $\gamma_0 \in \Gamma$. Let

$$M_1 := \sup_{y \in B} \lambda(\gamma_0, y),$$

let $C_1 \subseteq \Gamma$ be a compact set such that

$$\forall \gamma \notin C_1, y \in B : \quad \lambda(\gamma, y) > M_1 + 1,$$

let

$$M_2 := \sup_{(\gamma, y) \in C_1 \times B} \lambda(\gamma, y),$$

and let $C_2 \subseteq \Gamma$ be a compact set such that

$$\forall \gamma \notin C_2, y \in B : \quad \lambda(\gamma, y) > M_2 + 1.$$

It is obvious that $M_1 \le M_2$ and $\gamma_0 \in C_1 \subseteq C_2$. We can and will assume C_2 convex (see [27], Theorem 3.20(c)).

Let us now check that C_1 lies inside the interior of C_2. Indeed, for any fixed $y \in B$ and $\gamma \in C_1$, we have $\lambda(\gamma, y) \le M_2$; since $\lambda(\gamma', y) > M_2 + 1$ for all $\gamma' \notin C_2$, some neighborhood of γ will lie completely in C_2.

Let $D : \Sigma \to \Gamma$ be a continuous stationary prediction strategy. We will show that (17) holds for some continuous stationary prediction strategy D' taking values in the compact convex set $C(B) := C_2$. Namely, we define

$$D'(\sigma) := \begin{cases} D(\sigma) & \text{if } D(\sigma) \in C_1 \\ \frac{\rho(D(\sigma), \Gamma \setminus C_2)}{\rho(D(\sigma), C_1) + \rho(D(\sigma), \Gamma \setminus C_2)} D(\sigma) + \frac{\rho(D(\sigma), C_1)}{\rho(D(\sigma), C_1) + \rho(D(\sigma), \Gamma \setminus C_2)} \gamma_0 & \text{if } D(\sigma) \in C_2 \setminus C_1 \\ \gamma_0 & \text{if } D(\sigma) \in \Gamma \setminus C_2 \end{cases}$$

where ρ is the metric on Γ; the denominator $\rho(D(\sigma), C_1) + \rho(D(\sigma), \Gamma \setminus C_2)$ is positive since already $\rho(D(\sigma), C_1)$ is positive. Since C_2 is convex, we can see that D' indeed takes values in C_2. The only points x at which the continuity of D' is not obvious are those for which $D(\sigma)$ lies on the boundary of C_1: in this case one has to use the fact that C_1 is covered by the interior of C_2.

It remains to check (17); the only non-trivial case is $D(\sigma) \in C_2 \setminus C_1$. By the convexity of $\lambda(\gamma, y)$ in γ, the inequality in (17) will follow from

$$\frac{\rho(D(\sigma), \Gamma \setminus C_2)}{\rho(D(\sigma), C_1) + \rho(D(\sigma), \Gamma \setminus C_2)} \lambda(D(\sigma), y) + \frac{\rho(D(\sigma), C_1)}{\rho(D(\sigma), C_1) + \rho(D(\sigma), \Gamma \setminus C_2)} \lambda(\gamma_0, y) \le \lambda(D(\sigma), y),$$

i.e.,

$$\lambda(\gamma_0, y) \le \lambda(D(\sigma), y).$$

Since the left-hand side of the last inequality is at most M_1 and its right-hand side exceeds $M_1 + 1$, it holds true. ∎

Remark If the loss function is allowed to depend on the infinite past, the σs in Lemma 5 will have to be restricted to a compact set $A \subseteq \Sigma$ and the compact set C will depend not only on B but also on A (see Lemma 18 of [32]).

The proof

For each compact $B \subseteq \mathbf{Y}$ fix a compact convex $C(B) \subseteq \Gamma$ as in Lemma 5. Predictor's strategy ensuring (5) is constructed from Remover's winning strategy in $G(\mathbf{X} \times \mathbf{Y})$ (see Lemma 3; metric spaces are paracompact by the Stone theorem, [11], Theorem 5.1.3) and from Predictor's strategies $\mathcal{S}(A, B)$ outputting predictions

$$\gamma_n \in C(B) \tag{18}$$

and ensuring the consequent of (5) for all continuous

$$D : (A \times B)^\infty \times A \to C(B) \tag{19}$$

under the assumption that $(x_n, y_n) \in A \times B$ for given compact $A \subseteq \mathbf{X}$ and $B \subseteq \mathbf{Y}$ (the existence of such $\mathcal{S}(A, B)$ is asserted in Theorem 1). Remover's moves are assumed to be of the form $A \times B$ for compact $A \subseteq \mathbf{X}$ and $B \subseteq \mathbf{Y}$. Predictor is simultaneously playing the game of removal $G(\mathbf{X} \times \mathbf{Y})$ as Evader.

At the beginning of the game of prediction Predictor asks Remover to make his first move $A_1 \times B_1$ in the game of removal; without loss of generality we assume that $A_1 \times B_1$ contains all (x_n, y_n), $n \le 0$ (there is nothing to prove if $\{(x_n, y_n) \mid n \le 0\}$ is not precompact). Predictor then plays the game of prediction using the strategy $\mathcal{S}(A_1, B_1)$ until Reality chooses $(x_n, y_n) \notin A_1 \times B_1$ (forever if Reality never chooses such (x_n, y_n)). As soon as such (x_n, y_n) is chosen, Predictor announces (x_n, y_n) in the game of removal and notes Remover's response (A_2, B_2). He then continues playing the game of prediction using the strategy $\mathcal{S}(A_2, B_2)$ until Reality chooses $(x_n, y_n) \notin A_2 \times B_2$, etc.

Let us check that this strategy for Predictor will always ensure (5). If Reality chooses (x_n, y_n) outside Predictor's current $A_k \times B_k$ finitely often, the consequent of (5) will be satisfied for all continuous stationary $D : \Sigma \to C(B_K)$ (B_K being the second component of Remover's last move (A_K, B_K)) and so, by Lemma 5, for all continuous stationary $D : \Sigma \to \Gamma$. If Reality chooses (x_n, y_n) outside Predictor's current $A_k \times B_k$ infinitely often, the set of (x_n, y_n), $n = 1, 2, \ldots$, will not be precompact, and so the antecedent of (5) will be violated.

6 Proof of Theorem 4

When γ ranges over $\mathcal{P}(C)$ (identified with the subset of $\mathcal{P}(\Gamma)$ consisting of the measures concentrated on C) for a compact $C \subseteq \Gamma$, the loss function (12), as we have seen, is continuous. The following analogue of Lemma 5 will be useful.

Lemma 6 *Under the conditions of Theorem 4, for each compact set $B \subseteq \mathbf{Y}$ there exists a compact convex set $C = C(B) \subseteq \Gamma$ such that for each continuous stationary randomized prediction strategy $D : \Sigma \to \mathcal{P}(\Gamma)$ there exists a continuous stationary randomized prediction strategy $D' : \Sigma \to \mathcal{P}(C)$ such that (17) holds (D' dominates D "on average").*

(In fact, this lemma is not needed for the proof of Theorem 4 as we stated it, but it will imply that γ_n dominate $D(\sigma_n)$ on average, for any continuous stationary randomized prediction strategy D: see (20).)

Proof Define γ_0, M_1, C_1, M_2, and C_2 as in the proof of Lemma 5. Fix a continuous function $f_1 : \Gamma \to [0,1]$ such that $f_1 = 1$ on C_1 and $f_1 = 0$ on $\Gamma \setminus C_2$ (such an f_1 exists by the Tietze–Uryson theorem, [11], Theorem 2.1.8). Set $f_2 := 1 - f_1$. Let $D : \Sigma \to \mathcal{P}(\Gamma)$ be a continuous stationary randomized prediction strategy. For each $\sigma \in \Sigma$, split $D(\sigma)$ into two measures on Γ absolutely continuous with respect to $D(\sigma)$: $D_1(\sigma)$ with Radon–Nikodym density f_1 and $D_2(\sigma)$ with Radon–Nikodym density f_2; set

$$D'(\sigma) := D_1(\sigma) + |D_2(\sigma)| \, \delta_{\gamma_0}$$

(letting $|P| := P(\Gamma)$ for P a measure on Γ). It is clear that the stationary randomized prediction strategy D' is continuous (in the topology of weak convergence, as usual), takes values in $\mathcal{P}(C_2)$, and

$$\begin{aligned}
\lambda(D'(\sigma), y) &= \int_\Gamma \lambda(\gamma, y) f_1(\gamma) D(\sigma)(\mathrm{d}\gamma) + \lambda(\gamma_0, y) \int_\Gamma f_2(\gamma) D(\sigma)(\mathrm{d}\gamma) \\
&\le \int_\Gamma \lambda(\gamma, y) f_1(\gamma) D(\sigma)(\mathrm{d}\gamma) + \int_\Gamma M_1 f_2(\gamma) D(\sigma)(\mathrm{d}\gamma) \\
&\le \int_\Gamma \lambda(\gamma, y) f_1(\gamma) D(\sigma)(\mathrm{d}\gamma) + \int_\Gamma \lambda(\gamma, y) f_2(\gamma) D(\sigma)(\mathrm{d}\gamma) = \lambda(D(\sigma), y)
\end{aligned}$$

for all $(\sigma, y) \in \Sigma \times B$. So we can take $C(B) := C_2$. ∎

Fix one of the mappings $B \mapsto C(B)$ whose existence is asserted by the lemma.

We will prove that the prediction strategy of the previous section with (18) replaced by $\gamma_n \in \mathcal{P}(C(B))$ and (19) replaced by

$$D : (A \times B)^\infty \times A \to \mathcal{P}(C(B))$$

is CS universal. Let $D : \Sigma \to \mathcal{P}(\Gamma)$ be a continuous stationary randomized prediction strategy, i.e., a continuous stationary prediction strategy in the new game of prediction with loss function (12). Let (A_K, B_K) be Remover's last move (if Remover makes infinitely many moves, the antecedent of (6) is false, and there is nothing to prove), and let $D' : \Sigma \to \mathcal{P}(C(B_K))$ be a continuous stationary randomized prediction strategy satisfying (17) with $B := B_K$. From some n on our randomized prediction algorithm produces $\gamma_n \in \mathcal{P}(\Gamma)$ concentrated on $C(B_K)$, and they will satisfy

$$\limsup_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N}\lambda(\gamma_n,y_n)-\frac{1}{N}\sum_{n=1}^{N}\lambda(D(\sigma_n),y_n)\right)$$
$$\le\limsup_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N}\lambda(\gamma_n,y_n)-\frac{1}{N}\sum_{n=1}^{N}\lambda(D'(\sigma_n),y_n)\right)\le 0. \quad (20)$$

This is an interesting property but slightly different from what Theorem 4 asserts.

According to the proof of Lemma 6, we can, and we will, assume that $D'(\sigma_n)$ generates outcomes d'_n in two steps: first d_n is generated from $D(\sigma_n)$, and then it is replaced by γ_0 with probability $f_2(\sigma_n)$. The loss function is bounded in absolute value on the compact set $C(B_K)\times B_K$ by a constant L. From the law of the iterated logarithm (see (14) and (15)) applied to the losses of γ_n and d'_n we now obtain, instead of (20),

$$\limsup_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N}\lambda(g_n,y_n)-\frac{1}{N}\sum_{n=1}^{N}\lambda(d_n,y_n)\right)$$
$$\le\limsup_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N}\lambda(g_n,y_n)-\frac{1}{N}\sum_{n=1}^{N}\lambda(d'_n,y_n)\right)$$
$$=\limsup_{N\to\infty}\left(\frac{1}{N}\sum_{n=1}^{N}\lambda(\gamma_n,y_n)-\frac{1}{N}\sum_{n=1}^{N}\lambda(D'(\sigma_n),y_n)\right)\le 0 \text{ a.s.};$$

it remains to compare this with (6).

7 Stationarity and continuity

As we said earlier, the assumption of stationarity is very natural for prediction strategies: it just means that the arbitrary origin of time is not taken into account (in the spirit of the invariance principle in statistics; see, e.g., [21], Section 6.1). Stationary strategies can detect and make use of all kinds of trends and one-off phenomena; e.g., they can perform well when the rate of environment change is constantly increasing (as in our own environment). There need not be stationarity in the environment.

Interestingly, our prediction algorithms are continuous (or can be made continuous) but not stationary. First we discuss the continuity of the prediction algorithms constructed in the proofs of our four theorems.

Theorem 1 It is easy to check that the WAA is continuous; by the Weierstrass M-test, (7) converges uniformly and so its sum is continuous.

Theorem 2 To check that γ_n is a continuous function of σ_n in the topology of weak convergence, we only need to check that $\int f\,\mathrm{d}\gamma_n$ is a continuous function of σ_n for each $f\in C(\Sigma)$. This again follows from the Weierstrass M-test.

Theorem 3 As described, Predictor's strategy is not continuous since his behavior changes suddenly when Reality outputs (x_n, y_n) outside his current $A_k \times B_k$, but it is clear that it can be "smoothed around the edges" to ensure continuity.

Theorem 4 The situation is analogous to Theorem 3.

For concreteness, we will discuss stationarity only in the case of Theorem 1. We know that the WAA is a prediction strategy that is continuous as a function of the type $\Sigma \times \{1, 2, \ldots\} \to \Gamma$. It is not stationary (i.e., we cannot get rid of the $\{1, 2, \ldots\}$) because it has to keep track of the experts' losses since the beginning of the game of prediction. Stationary strategies can depend on time only in a limited way: e.g., in terms of our own environment, they can depend on the time of day or the season. But the WAA's dependence is much heavier: it has to know precisely the time that has elapsed since the beginning.

Let us now check that there are no universal continuous stationary prediction strategies under conditions of Theorem 1. Suppose Γ is such that there exists $f : \Gamma \to \Gamma$ without fixed points (i.e., $f(\gamma) \neq \gamma$ for all $\gamma \in \Gamma$; we can take, e.g., a circle as Γ). If D were a universal continuous stationary strategy, we could define another continuous stationary strategy $D'(\sigma) := f(D(\sigma))$ and make Reality collude with D' (i.e., output y_n leading to a significantly smaller loss for D'; this can be done for an appropriate choice of λ, and in fact can be done for all usual λ).

Stationary Reality

A standard problem in probability theory is where Reality is governed by a stationary probability measure; of course, only stationary prediction strategies are considered. In this subsection we will list several references for this problem, considering, for simplicity, only the case where the signals x_n are absent (formally, we assume that $\mathbf{X}$ is a one-element set and omit the x_n, which now do not carry any information, from our notation).

The problem of prediction has been studied extensively for both strictly stationary sequences of observations and wide sense stationary sequences (the definitions and a general discussion of "strict sense" and "wide sense" concepts can be found in [9], Chapter 2, Sections 8 and 3). We will first assume that $\ldots, y_{-1}, y_0, y_1, \ldots$ form a wide sense stationary sequence of random variables and then a strictly stationary sequence.

The natural mode of prediction for wide sense stationary sequences is linear prediction. The problem of linear prediction (not necessarily one-step-ahead, as in this paper) of wide sense stationary sequences was posed and solved by Kolmogorov [17, 18, 19]; later but independently this was done by Wiener [33].

Kolmogorov and Wiener assumed the probability distribution of the observations known. There are many efficient ways to estimate the spectral density of this probability distribution (in terms of which the optimal linear predictor is expressed); see, e.g., [2], Chapter 9, for a review. (An early idea of spectral estimation was proposed by Einstein in 1914: see [24], p. 363.)

The problem of existence of universal prediction strategies for strictly stationary and ergodic sequences of observations was posed by Cover [8], and such strategies were found by Ornstein [26] for finite $\mathbf{Y}$ and Algoet [1] for $\mathbf{Y}$ a Polish space. Papers [14, 13, 25] construct such strategies using techniques very similar to those of this paper.

8 Conclusion

An interesting direction of further research is to obtain non-asymptotic versions of our results. If the benchmark class of continuous stationary prediction strategies is compact, loss bounds can be given in terms of ϵ-entropy [20]. In general, one can give loss bounds in terms of a nested family of compact sets whose union is dense in the set of continuous stationary prediction strategies (in analogy with Vapnik and Chervonenkis's principle of structural risk minimization [29]).

Acknowledgments

I am grateful to Yura Kalnishkan and Ilia Nouretdinov for useful comments. The construction of CS universal prediction strategies is based on Alex Smola's and Gábor Lugosi's suggestions. This work was partially supported by MRC (grant S505/65).

References

[1] Paul H. Algoet. Universal schemes for prediction, gambling and portfolio selection. *Annals of Probability*, 20:901–941, 1992. Corrections: 23:474–478, 1995.

[2] T. W. Anderson. *The Statistical Analysis of Time Series*. Wiley, New York, 1971. Wiley Classics Library edition: 1994.

[3] Patrick Billingsley. *Convergence of Probability Measures*. Wiley, New York, 1968.

[4] Lenore Blum, Felipe Cucker, Michael Shub, and Steve Smale. *Complexity and Real Computation*. Springer, New York, 1998.

[5] Lenore Blum, Michael Shub, and Steve Smale. On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines. *Bulletin of the American Mathematical Society*, 21:1–46, 1989.

[6] Nicolas Bourbaki. *Eléments de mathématique, Livre VI, Intégration, Chapitres 1 à 4*. Hermann, Paris, first edition, 1952.

[7] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006.

[8] Tom M. Cover. Open problems in information theory. In *Moscow Information Theory Workshop*, New York, 1975. IEEE Press.

[9] Joseph L. Doob. *Stochastic Processes*. Wiley, New York, 1953.

[10] Richard M. Dudley. *Real Analysis and Probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, England, 2002. Originally published in 1989.

[11] Ryszard Engelking. *General Topology*, volume 6 of *Sigma Series in Pure Mathematics*. Heldermann, Berlin, second edition, 1989.

[12] Gary Gruenhage. The story of a topological game. *Rocky Mountain Journal of Mathematics*, 2006. To appear.

[13] László Györfi and Gábor Lugosi. Strategies for sequential prediction of stationary time series. In Moshe Dror, Pierre L'Ecuyer, and Ferenc Szidarovszky, editors, *Modeling Uncertainty: An Examination of its Theory, Methods, and Applications*. Kluwer, 2001.

[14] László Györfi, Gábor Lugosi, and G. Morvai. A simple randomized algorithm for consistent sequential prediction of ergodic time series. *IEEE Transactions on Information Theory*, 45:2642–2650, 1999.

[15] G. H. Hardy, John E. Littlewood, and George Pólya. *Inequalities*. Cambridge University Press, Cambridge, second edition, 1952.

[16] Yuri Kalnishkan and Michael V. Vyugin. The Weak Aggregating Algorithm and weak mixability. In Peter Auer and Ron Meir, editors, *Proceedings of the Eighteenth Annual Conference on Learning Theory*, volume 3559 of *Lecture Notes in Computer Science*, pages 188–203, Berlin, 2005. Springer. The journal version is being prepared for the Special Issue of *Journal of Machine Learning Research* devoted to COLT'2005; all references are to the journal version.

[17] Andrei N. Kolmogorov. Sur l'interpolation et extrapolation des suites stationnaires. *Comptes rendus de Séances de l'Academie des Sciences*, 208:2043–2045, 1939.

[18] Andrei N. Kolmogorov. Interpolation and extrapolation of stationary random sequences (in Russian). *Izvestiya AN SSSR. Mathematics series*, 5:3–14, 1941.

[19] Andrei N. Kolmogorov. Stationary sequences in Hilbert space (in Russian). *Byulleten' MGU. Mathematics*, 2(6):1–40, 1941.

[20] Andrei N. Kolmogorov and Vladimir M. Tikhomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces (in Russian). *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

[21] E. L. Lehmann. *Testing Statistical Hypotheses*. Springer, New York, second edition, 1986.

[22] Per Martin-Löf. *Notes on Constructive Mathematics*. Almqvist & Wiksell, Stockholm, 1970.

[23] Som A. Naimpally and Brian D. Warrack. *Proximity Spaces*, volume 59 of *Cambridge Tracts in Mathematics and Mathematical Physics*. Cambridge University Press, London, 1970.

[24] H. Joseph Newton. A conversation with Emanuel Parzen. *Statistical Science*, 17:357–378, 2002.

[25] Andrew B. Nobel. On optimal sequential prediction for general processes. *IEEE Transactions on Information Theory*, 49:83–98, 2003.

[26] D. S. Ornstein. Guessing the next output of a stationary process. *Israel Journal of Mathematics*, 30:292–296, 1978.

[27] Walter Rudin. *Functional Analysis*. McGraw-Hill, Boston, second edition, 1991.

[28] Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* Wiley, New York, 2001.

[29] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[30] Vladimir Vovk. Aggregating strategies. In Mark Fulk and John Case, editors, *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–383, San Mateo, CA, 1990. Morgan Kaufmann.

[31] Vladimir Vovk. Competitive on-line statistics. *International Statistical Review*, 69:213–248, 2001.

[32] Vladimir Vovk. Predictions as statements and decisions. Technical Report `arXiv:cs.LG/0606093`, `arXiv.org` e-Print archive, June 2006.

[33] Norbert Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*. Technology Press of the Massachusetts Institute of Technology, Cambridge, MA, 1949. Reprinted from a secret 1942 publication.