Competing with Markov prediction strategies

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

February 1, 2008

Abstract

Assuming that the loss function is convex in the prediction, we construct a prediction strategy universal for the class of Markov prediction strategies, not necessarily continuous. Allowing randomization, we remove the requirement of convexity.

1 Introduction

This paper belongs to the area of research known as universal prediction of individual sequences (see [2] for a review): the predictor's goal is to compete with a wide benchmark class of prediction strategies. In the previous papers [15] and [14] we constructed prediction strategies competitive with the important classes of Markov and stationary, respectively, continuous prediction strategies. In this paper we consider competing against possibly discontinuous strategies. Our main results assert the existence of prediction strategies competitive with the Markov strategies.

This paper's idea of transition from continuous to general benchmark classes was motivated by Skorokhod's topology for the space D of "càdlàg" functions, most of which are discontinuous. Skorokhod's idea was to allow small deformations not only along the vertical axis but also along the horizontal axis when defining neighborhoods. Skorokhod's topology was metrized by Kolmogorov so that it became a separable space ([1], Appendix III; [11], p. 913), which allows us to apply one of the numerous algorithms for prediction with expert advice (Kalnishkan and Vyugin's Weak Aggregating Algorithm in this paper) to construct a universal algorithm.

In Section 2 we give the main definitions and state our main results, Theorems 1 and 2; their proofs are given in Sections 3 and 4, respectively.

2 Main results

The *game of prediction* between two players, called 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

FOR $n = 1, 2, \ldots$:
  Reality announces $x_n \in \mathbf{X}$.
  Predictor announces $\gamma_n \in \Gamma$.
  Reality announces $y_n \in \mathbf{Y}$.
END FOR.

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 observation 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* $\lambda : \Gamma \times \mathbf{Y} \to \mathbb{R}$.

We will always assume that the signal space $\mathbf{X}$, the prediction space Γ, and the observation space $\mathbf{Y}$ are non-empty sets; $\mathbf{X}$ and Γ will often be equipped with additional structures.

Markov-universal prediction strategies: deterministic case

Predictor's strategies in the prediction protocol will be called *prediction strategies*. Formally such a strategy is a function

$$D : \bigcup_{n=1}^{\infty} (\mathbf{X} \times \mathbf{Y})^{n-1} \times \mathbf{X} \to \Gamma;$$

it maps each history $(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$ to the chosen prediction. In this paper we will be especially interested in *Markov strategies*, which are functions $D : \mathbf{X} \to \Gamma$; intuitively, $D(x_n)$ is the recommended prediction on round n. The restriction to Markov strategies is not a severe one, since the signal x_n can encode as much of the past as we want (cf. [8], footnote 1); in particular, x_n can contain information about the previous observations $y_1, \ldots, y_{n-1}$. In this paper Markov prediction strategies will also be called *prediction rules* (as in [15]; in a more general context, however, it would be risky to omit "Markov" since "prediction rule" is too easy to confuse with "prediction strategy").

For both our theorems we will need the notion of "approximation" to a signal $x \in \mathbf{X}$; intuitively, the "m-approximation" of x is another signal $\phi_m(x)$ which is as close to x as possible but carries only m bits of information. If $\mathbf{X} = [0, 1]$, a reasonable definition of $\phi_m(x)$ would be to take the binary expansion of x but remove all the binary digits starting from the $(m + 1)$th after the binary dot. In general, we will have to equip $\mathbf{X}$ with an "approximation structure"; we will do this following Kolmogorov and Tikhomirov ([12], Section 2, [11], p. 913).

Consider a sequence of mappings $\phi_m : \mathbf{X} \to \mathbf{X}$, $m = 1, 2, \ldots$, such that each ϕ_m is idempotent, in the sense $\phi_m(\phi_m(x)) = \phi_m(x)$ for all $x \in \mathbf{X}$, and $\phi_m(\mathbf{X})$ contains 2^m elements. (Such mappings are coding-theory analogues of

projections in linear algebra and contractions in topology; $\phi_m(x)$ can be thought of as the result of encoding x, sending it over an m-bit channel, and restoring x as well as possible at the receiving end.) It is the sequence $\phi = \{\phi_m \,|\, m = 1, 2, \ldots\}$ that will be referred to as an *approximation structure*.

If $\mathbf{X}$ is a totally bounded (say, compact) metric space, there is an approximation structure ϕ such that

$$\lim_{m\to\infty} \rho\left(x, \phi_m(x)\right) = 0 \tag{1}$$

uniformly in $x \in \mathbf{X}$. (We often let ρ stand for the metric in various metric spaces, always clear from the context.) In fact, the *mth Kolmogorov diameter*

$$\mathcal{K}_m(\mathbf{X}) := \frac{1}{2} \inf_{\phi} \sup_{x\in\mathbf{X}} \operatorname{diam}\left(\phi_m^{-1}(\phi_m(x))\right)$$

of $\mathbf{X}$ is essentially the inverse function to the ϵ-entropy $\mathcal{H}_\epsilon(\mathbf{X})$. See [9] for precise values and estimates of $\mathcal{K}_m(\mathbf{X})$ for numerous totally bounded metric spaces $\mathbf{X}$.

A prediction strategy is *Markov-universal* for a loss function λ and an approximation structure ϕ if it guarantees that for any prediction rule D and any $m = 1, 2, \ldots$ there exists a number $N_{D,m}$ such that for any $N \geq N_{D,m}$ and any sequence $x_1, y_1, x_2, y_2, \ldots$ of Reality's moves its responses γ_n satisfy

$$\frac{1}{N}\sum_{n=1}^{N} \lambda(\gamma_n, y_n) \leq \frac{1}{N}\sum_{n=1}^{N} \lambda\Big(D(\phi_m(x_n)), y_n\Big) + 2^{-m}.$$

Theorem 1 *Suppose $\mathbf{X}$ is equipped with an approximation structure ϕ, Γ is a closed convex subset of a separable Banach space, and the loss function $\lambda(\gamma, y)$ is bounded, convex in the variable $\gamma \in \Gamma$, and uniformly continuous in $\gamma \in \Gamma$ uniformly in $y \in \mathbf{Y}$. There exists a Markov-universal for λ and ϕ prediction strategy.*

A Markov-universal prediction strategy will be constructed in the next section. Theorem 1 says that, under its conditions,

$$\limsup_{N\to\infty} \left(\frac{1}{N}\sum_{n=1}^{N} \lambda(\gamma_n, y_n) - \frac{1}{N}\sum_{n=1}^{N} \lambda\Big(D(\phi_m(x_n)), y_n\Big)\right) \leq 0 \tag{2}$$

uniformly in $x_1, y_1, x_2, y_2, \ldots$ for all $m = 1, 2, \ldots$ and all $D : \mathbf{X} \to \Gamma$.

If $\mathbf{X}$ is a compact metric space and (1) holds uniformly in $x \in \mathbf{X}$, (2) implies

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

for all continuous prediction rules D; this is close to Theorem 1 in [15]. The advance of this paper as compared to [15] is that our main results do not assume that D is continuous.

Markov-universal prediction strategies: randomized case

When the loss function $\lambda(\gamma, y)$ is not required to be convex in γ, the conclusion of Theorem 1 may become false ([6], Theorem 2). The situation changes if we consider randomized prediction strategies.

A *randomized prediction strategy* is a function

$$D : \bigcup_{n=1}^{\infty} (\mathbf{X} \times \mathbf{Y})^{n-1} \times \mathbf{X} \to \mathcal{P}(\Gamma)$$

mapping the past to the probability measures on the prediction space. In other words, this is a strategy for Predictor in the extended game of prediction with the prediction space $\mathcal{P}(\Gamma)$. A *Markov randomized prediction strategy*, or *randomized prediction rule* for brevity, is a function $D : \mathbf{X} \to \mathcal{P}(\Gamma)$.

We will say that a randomized prediction strategy outputting γ_n is *Markov-universal* for a loss function λ and an approximation structure ϕ if, for any randomized prediction rule D and any $m = 1, 2, \ldots$, there exists $N_{D,m}$ such that, for any sequence $x_1, y_1, x_2, y_2, \ldots$ of Reality's moves,

$$\sup_{N \geq N_{D,m}} \left(\frac{1}{N} \sum_{n=1}^{N} \lambda(g_n, y_n) - \frac{1}{N} \sum_{n=1}^{N} \lambda(d_n, y_n) \right) \leq 2^{-m} \tag{3}$$

with probability at least $1 - 2^{-m}$, where $g_1, g_2, \ldots, d_1, d_2, \ldots$ are independent random variables distributed as

$$g_n \sim \gamma_n, \;\; d_n \sim D(\phi_m(x_n)), \quad n = 1, 2, \ldots . \tag{4}$$

Intuitively, the word "probability" after (3) refers only to the prediction strategies' internal randomization; it is not assumed that Reality behaves stochastically. We will use this definition only in the case where the loss function λ is continuous in the prediction, and so (3) will indeed be an event having a probability.

Theorem 2 *Suppose the signal space $\mathbf{X}$ is equipped with an approximation structure ϕ, Γ is a separable topological space, and the loss function λ is bounded and such that the set of functions $\{\lambda(\cdot, y) \,|\, y \in \mathbf{Y}\}$ is equicontinuous. There exists a randomized prediction strategy that is Markov-universal for λ and ϕ.*

A Markov-universal prediction strategy is constructed in Section 4. The randomized version of (2), immediately following from Theorem 2, is

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

for all $m = 1, 2, \ldots$ and all $D : \mathbf{X} \to \Gamma$, where $g_1, g_2, \ldots, d_1, d_2, \ldots$ are independent and distributed as (4).

3 Proof of Theorem 1

Let us fix a dense countable subset Γ^* of Γ. We will say that a function $D : \mathbf{X} \to \Gamma$ is *m-elementary* if $D(\mathbf{X}) \subseteq \Gamma^*$ and $D(x)$ depends on x only via $\phi_m(x)$; a function is *elementary* if it is m-elementary for some m. There are countably many elementary functions; let us enumerate them as $D_1, D_2, \ldots$. We will refer to these functions as *experts*. We will apply a special case of Kalnishkan and Vyugin's [6] Weak Aggregating Algorithm (WAA) to the sequence of experts (as in [14]).

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

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

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

$$\gamma_n := \sum_{k=1}^{\infty} p_n^{(k)} D_k(x_n). \tag{5}$$

To make this series convergent, we may take $q_k := 2^{-k}$ and reorder D_k so that $\sup_x \|D_k(x)\| \le k$ for all k. In this case we will automatically have $\gamma_n \in \Gamma$ since

$$\begin{aligned} &\gamma_n - \sum_{k=1}^{K} \frac{p_n^{(k)}}{\sum_{k=1}^{K} p_n^{(k)}} D_k(x_n) \\ &\qquad = \sum_{k=1}^{K} \left(1 - \frac{1}{\sum_{k=1}^{K} p_n^{(k)}}\right) p_n^{(k)} D_k(x_n) + \sum_{k=K+1}^{\infty} p_n^{(k)} D_k(x_n) \to 0 \end{aligned} \tag{6}$$

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 ([6], Lemma 9) *The WAA guarantees that, for all $N = 1, 2, \ldots$,*

$$L_N \le \sum_{n=1}^{N} \sum_{k=1}^{\infty} p_n^{(k)} l_n^{(k)} - \sum_{n=1}^{N} \log_{\beta_n} \sum_{k=1}^{\infty} p_n^{(k)} \beta_n^{l_n^{(k)}} + \log_{\beta_N} \sum_{k=1}^{\infty} q_k \beta_N^{L_N^{(k)}}. \tag{7}$$

The first two terms on the right-hand side of (7) are sums over the first N rounds of different kinds of mean of the experts' losses (see, e.g., [5], 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 (7) is identical (except that β now depends on n) to the main property of the "Aggregating Algorithm" (see, e.g., [13], Lemma 1). All infinite series in (7) are trivially convergent.

In the proof of Lemma 1 we will use the following property of "countable convexity" of λ:

$$l_n \le \sum_{k=1}^{\infty} p_n^{(k)} l_n^{(k)}. \tag{8}$$

This property follows from (6) and

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

if we let $K \to \infty$.

Proof of Lemma 1 The proof is by induction on N. For $N = 1$, (7) follows from the countable convexity (8) and $p_1^{(k)} = q_k$. Assuming (7), 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 "$\le$" again used the countable convexity (8)). 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

$$\log_{\beta_N} \sum_{k=1}^{\infty} q_k \beta_N^{L_N^{(k)}} \le \log_{\beta_{N+1}} \sum_{k=1}^{\infty} q_k \beta_{N+1}^{L_N^{(k)}}. \tag{9}$$

The last inequality follows from the general result about comparison of different means ([5], Theorem 85), but we can also check it directly (following [6]). Let $\beta_{N+1} = \beta_N^a$, where $0 < a < 1$. Then (9) can be rewritten as

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

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

Lemma 2 ([6], Lemma 5) *Let L be an upper bound on $|\lambda|$. The WAA guarantees that, for all N and K,*

$$L_N \leq L_N^{(K)} + \left(L^2 e^L + \ln \frac{1}{q_K}\right)\sqrt{N}. \tag{10}$$

Proof From (7), we obtain:

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

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

Remark There is no term e^L in [6] since that paper only considers non-negative loss functions. (Notice that even without assuming non-negativity this term is very crude and can be easily improved.)

Now it is easy to prove Theorem 1. The definition of Markov-universality can be restated as follows: a prediction strategy outputting γ_n is Markov-universal if and only if for any prediction rule D, any $m = 1, 2, \ldots$, and any $\epsilon > 0$ there exists $N_{D,m,\epsilon}$ such that, for any $N \geq N_{D,m,\epsilon}$ and any $x_1, y_1, x_2, y_2, \ldots$,

$$\frac{1}{N}\sum_{n=1}^{N} \lambda(\gamma_n, y_n) \leq \frac{1}{N}\sum_{n=1}^{N} \lambda\Big(D(\phi_m(x_n)), y_n\Big) + \epsilon. \tag{11}$$

Let γ_n be output by the WAA and let us consider any prediction rule D, any $m \in \{1, 2, \ldots\}$, and any $\epsilon > 0$. Choose $\delta > 0$ such that $|\lambda(\gamma, y) - \lambda(\gamma', y)| < \epsilon/2$ whenever $\rho(\gamma, \gamma') < \delta$ and choose an m-elementary expert D_K such that, for all $x \in \phi_m(\mathbf{X})$, $\rho(D(x), D_K(x)) < \delta$.

From (10) we obtain

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\lambda(\gamma_n, y_n) - \frac{1}{N}\sum_{n=1}^{N}\lambda\Big(D(\phi_m(x_n)), y_n\Big) \\
&\qquad \le \frac{1}{N}\sum_{n=1}^{N}\lambda(\gamma_n, y_n) - \frac{1}{N}\sum_{n=1}^{N}\lambda\Big(D_K(\phi_m(x_n)), y_n\Big) + \frac{\epsilon}{2} \\
&\qquad = \frac{1}{N}\sum_{n=1}^{N}\lambda(\gamma_n, y_n) - \frac{1}{N}\sum_{n=1}^{N}\lambda\Big(D_K(x_n), y_n\Big) + \frac{\epsilon}{2} \\
&\qquad \le \left(L^2 e^L + \ln\frac{1}{q_K}\right)\frac{1}{\sqrt{N}} + \frac{\epsilon}{2}; \quad (12)
\end{aligned}$$

now (11) is obvious.

4 Proof of Theorem 2

A convenient pseudo-metric on Γ can be defined by

$$\rho(g, g') := \sup\left\{\lambda(g, y) - \lambda(g', y) \,|\, y \in \mathbf{Y}\right\}, \quad g, g' \in \Gamma$$

(cf. [3], Corollary 11.3.4). Let us redefine Γ as the quotient space obtained from the original Γ by identifying g and g' for which $\rho(g, g') = 0$ ([4], Section 2.4); in other words, we will not distinguish predictions that always lead to identical losses. Now ρ becomes a metric on Γ. Let Γ^* be a countable dense subset of the original topological space Γ (which is separable as a subset of a separable Banach space); the condition of equicontinuity implies that Γ^* (formally defined as the set of equivalence classes containing elements of the original Γ^*) remains a dense subset in Γ equipped with the metric ρ.

We define the norm of a function $f : \Gamma \to \mathbb{R}$ as

$$\|f\|_{\mathrm{BL}} := \sup_{g, g' \in \Gamma : g \neq g'} \frac{|f(g) - f(g')|}{\rho(g, g')} + \sup_{g \in \Gamma} |f(g)|;$$

this norm is finite for bounded Lipschitz functions (which form a Banach space under this norm: see [3], Section 11.2). Notice that

$$\|\lambda\|_{\mathrm{BL}} := \sup_{y \in \mathbf{Y}} \|\lambda(\cdot, y)\|_{\mathrm{BL}} < \infty. \quad (13)$$

Next define

$$\lambda(\gamma, y) := \int_{\Gamma} \lambda(g, y)\gamma(\mathrm{d}g), \quad (14)$$

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 linear and, therefore, convex in γ. (In general, the role of randomization in this paper is to make the loss function convex in the prediction.)

As a metric on $\mathcal{P}(\Gamma)$ we will take the Fortet–Mourier metric ([3], Section 11.3) defined as

$$\beta(\gamma, \gamma') := \sup_{f:\|f\|_{\mathrm{BL}} \le 1} \left| \int_\Gamma f \, \mathrm{d}(\gamma - \gamma') \right| .$$

The topology on $\mathcal{P}(\Gamma)$ induced by this metric is called the *topology of weak convergence* ([1]; weak convergence is called simply "convergence" in [3]; for the proof of equivalence of several natural definitions of the topology of weak convergence, see [3], Theorem 11.3.3).

Let us check that the loss function (14) is also bounded Lipschitz, in the sense of (13): if $\gamma, \gamma' \in \mathcal{P}(\gamma)$ and $y \in \mathbf{Y}$,

$$|\lambda(\gamma, y) - \lambda(\gamma', y)| = \left| \int_\Gamma \lambda(g, y)(\gamma - \gamma')(\mathrm{d}g) \right| \le \|\lambda\|_{\mathrm{BL}} \, \beta(\gamma, \gamma').$$

It is easy to see that the space $\mathcal{P}(\Gamma)$ with metric β is separable: e.g., the set of probability measures concentrated on finite subsets of Γ^* and taking rational values is dense in $\mathcal{P}(\Gamma)$ (cf. [1], Appendix III). Let us enumerate the elements of a dense countable set in $\mathcal{P}(\Gamma)$ as $D_1, D_2, \ldots$; as in the previous section, we will use the WAA to merge all *experts* D_k.

The convergence of the mixture (5) to a probability measure on Γ is now obvious. The countable convexity (8) now holds with equality,

$$\lambda \left(\sum_{k=1}^\infty p_n^{(k)} D_k(x_n), y_n \right) = \sum_{k=1}^\infty p_n^{(k)} \lambda \left(D_k(x_n), y_n \right),$$

and follows from the general fact that

$$\int f \, \mathrm{d} \sum_{k=1}^\infty p_k P_k = \sum_{k=1}^\infty p_k \int f \, \mathrm{d} P_k$$

for bounded Borel $f : \Gamma \to \mathbb{R}$, positive $p_1, p_2, \ldots$ summing to 1, and $P_1, P_2, \ldots \in \mathcal{P}(\Gamma)$ (this is obviously true for simple f and follows for arbitrary integrable f from the definition of Lebesgue integral: see, e.g., [3], Section 4.1).

Therefore, it is easy to check that the chain (12) still works (with $\mathcal{P}(\Gamma)$ equipped with metric β) and we can rephrase the previous section's result as follows. For any randomized prediction rule D, any $m = 1, 2, \ldots$, and any $\epsilon > 0$ there exists $N_{D,m,\epsilon}$ such that, for any $N \ge N_{D,m,\epsilon}$ and any $x_1, y_1, x_2, y_2, \ldots$, the WAA's predictions $\gamma_n \in \mathcal{P}(\Gamma)$ are guaranteed to satisfy

$$\frac{1}{N} \sum_{n=1}^N \lambda(\gamma_n, y_n) \le \frac{1}{N} \sum_{n=1}^N \lambda\Big(D(\phi_m(x_n)), y_n \Big) + \frac{\epsilon}{2} \tag{15}$$

(cf. (11)).

The loss function is bounded in absolute value by a constant L, and so the law of the iterated logarithm (in Kolmogorov's finitary form, [7], the end of the introductory section; the condition that the cumulative variance tends to infinity is easy to get rid of: see, e.g., [10], (5.8)) implies that for any $\delta > 0$ there exists N_δ such that the conjunction of

$$\sup_{N \geq N_\delta} \left| \sum_{n=1}^{N} \big(\lambda(g_n, y_n) - \lambda(\gamma_n, y_n) \big) \right| \leq \sqrt{2.01 L^2 N \ln \ln N}$$

and

$$\sup_{N \geq N_\delta} \left| \sum_{n=1}^{N} \big(\lambda(d_n, y_n) - \lambda(D(x_n), y_n) \big) \right| \leq \sqrt{2.01 L^2 N \ln \ln N}$$

holds with probability at least $1 - \delta$. Combining the last two inequalities with (15) we can see that for any randomized prediction rule D, any $m = 1, 2, \ldots$, any $\epsilon > 0$, and any $\delta > 0$ there exists $N_{D,m,\epsilon,\delta}$ such that, for any $x_1, y_1, x_2, y_2, \ldots$, the WAA's responses $\gamma_n \in \mathcal{P}(\Gamma)$ to $x_1, y_1, x_2, y_2, \ldots$ are guaranteed to satisfy

$$\sup_{N \geq N_{D,m,\epsilon,\delta}} \left(\frac{1}{N} \sum_{n=1}^{N} \lambda(g_n, y_n) - \frac{1}{N} \sum_{n=1}^{N} \lambda(d_n, y_n) \right) \leq \epsilon$$

with probability at least $1 - \delta$. This is equivalent to the WAA (applied to $D_1, D_2, \ldots$) being a Markov-universal randomized prediction strategy.

5 Conclusion

An interesting theoretical problem is to state more explicit versions of Theorems 1 and 2: for example, to give an explicit expression for $N_{D,m}$.

The field of lossy compression is now well developed, and it would be interesting to apply our prediction algorithms (perhaps with the Weak Aggregating Algorithm replaced by an algorithm based on, say, gradient descent [2] or defensive forecasting [15]) to the approximation structures induced by popular lossy compression algorithms.

Acknowledgments

This work was partially supported by MRC (grant S505/65).

References

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

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

[3] Richard M. Dudley. *Real Analysis and Probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, England, revised edition, 2002.

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

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

[6] 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.

[7] Andrei N. Kolmogorov. Über das Gesetz des iterierten Logarithmus. *Mathematische Annalen*, 101:126–135, 1929.

[8] Andrei N. Kolmogorov. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Mathematische Annalen*, 104:415–458, 1931.

[9] 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.

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

[11] Albert N. Shiryaev. Kolmogorov: life and creative activities. *Annals of Probability*, 17:866–944, 1989.

[12] Vladimir M. Tikhomirov. ϵ-entropy and ϵ-capacity (in Russian). In Yury V. Prokhorov and Albert N. Shiryaev, editors, *Kolmogorov. Teoriya Informatsii i Teoriya Algoritmov*, pages 262–269. Nauka, Moscow, 1987.

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

[14] Vladimir Vovk. Competing with stationary prediction strategies. Technical Report `arXiv:cs.LG/0607067`, `arXiv.org` e-Print archive, July 2006.

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