Most sequences are stochastic

V.V. V'yugin

Institute for Information Transmission Problems,
Russian Academy of Sciences,
Bol'shoi Karetnyi per. 19, Moscow GSP-4, 101447, Russia
vld@vyugin.mccme.rssi.ru

and

Computer Learning Research Centre
Royal Holloway, University of London
Egham, Surrey TW20 0EX, England

November 17, 2009

Abstract

The central problem in machine learning (and statistics) is the problem of predicting future events x_{n+1} based on past observations $x_1x_2\ldots x_n$, where $n = 1, 2\ldots$. The main goal is to find a method of prediction that minimizes the total loss suffered on a sequence $x_1x_2\ldots x_{n+1}$ for $n = 1, 2\ldots$. We say that a data sequence is "stochastic" if there exists a simply described prediction algorithm whose performance is close to the best possible one. This optimal performance is defined in terms of Vovk's [8] "predictive complexity", which is a generalization of the notion of Kolmogorov complexity. Predictive complexity gives a limit on the predictive performance of simply described prediction algorithms.

In this paper we argue that data sequences normally occurring in the real world are stochastic; more formally, we prove that Levin's *a priori* semimeasure of non-stochastic sequences is small.

1 Introduction

We present the formal results suggesting some possible explanation of the following phenomenon: for many data sets the performance of the best prediction algorithms (Support Vector Machines, neural nets, boosting in the prediction of hand-written digits) is comparable. Possible explanation is based on a new notion of predictive complexity introduced by Vovk [8], [7] for perfectly mixable loss functions. We also refer readers to V'yugin [12], Sections 2 and 4, where some proofs for predictive complexity are given.

In this paper we consider only perfectly mixable loss functions for which the optimal predictive complexity exists (see below). The problem of existence of "sub-optimal measures" of predictive complexity for absolute loss function, which is not perfectly mixable, is considered in [12].

The predictive complexity of a data sequence gives a lower limit on the predictive performance of simply described prediction algorithms. We suppose that the state-of-the-art algorithms attain predictive complexity.

This is not true for *all* possible data sequences; but we can try to prove this for "normal", in some sense, sequences.

Our problem is closely related to Kolmogorov's theory of stochastic sequences in probabilistic setting (which corresponds to the log-loss game), see Shen [4]. Similar results for this setting were obtained earlier by V'yugin [9].

The plan of this paper: first we give formal definitions of predictive strategy, loss function, predictive complexity; then we discuss the notion of a "normal" sequence (Levin's philosophy) and define the *a priori semimeasure*; state our main results (lower and upper bounds on the *a priori* semimeasure for non-stochastic sequences). In Section 5 we prove main results.

2 Background

Suppose we are given a sequence $x_1, x_2, \ldots, x_i \ldots$ of some data. In this paper we consider only the simplest case, where $x_i \in \{0, 1\}$ (the case, where $x_i \in \{0, \ldots, L-1\}, L \geq 2$, is considered analogously). Our goal is to predict the elements of this data set on-line: we predict x_1, then predict x_2 given x_1, ..., then predict x_i given $x_1, x_2, \ldots x_{i-1}$, etc. At every step i the loss is measured by some function $\lambda(x_i, p_i)$, where the forecast is a real number $p_i \in [0, 1]$ and the actual outcome is x_i. We consider only loss functions computable by algorithms. For example, we consider the squared difference $\lambda(x_i, p_i) = (x_i - p_i)^2$ and the log-loss function $\lambda(x_i, p_i) = -\log p_i$ if $x_i = 1$ and $\lambda(x_i, p_i) = -\log(1 - p_i)$ if $x_i = 0$. Here log means logarithm to the base 2. Other loss functions are considered in Vovk [6].

It is natural to suppose that all predictions are given according to a *prediction strategy* (or *prediction algorithm*) $p_i = S(x_1 x_2 \ldots x_{i-1})$, $i = 2, \ldots$ ($p_1 = S(\Lambda)$, where Λ is the empty sequence). The total loss incurred by Predictor who follows the strategy S over the first n trials $x_1, x_2, \ldots, x_n$ is defined

$$\mathrm{Loss}_S(x_1 x_2 \ldots x_n) = \sum_{i=1}^{n} \lambda(x_i, S(x_1 x_2 \ldots x_{i-1})).$$

The main task is to minimize the total loss suffered on a sequence $x = x_1 x_2 \ldots x_n$ of outcomes. The corresponding game-theoretic interpretation is given in Vovk [7] or in Vovk and Watkins [8].

Let us fix $\eta > 0$ (*learning rate*) and put $\beta = e^{-\eta} \in (0, 1)$. A loss function $\lambda(x, p)$ is *η-mixable* if for every sequence $p_1, p_2, \ldots$ of predictions and every sequence $r_1, r_2, \ldots$ of nonnegative weights, whose sum do not exceed 1, there

exists a prediction γ such that

$$\lambda(j,\gamma) \le \log_\beta \sum_{i=1}^{\infty} r_i \beta^{\lambda(j,p_i)}$$

holds for all j.

The loss function is *perfectly mixable* if it is η-mixable for some $\eta > 0$. It is known that many popular loss functions such as the log-loss function, square-loss function, Cover's loss function, long-short loss function, Kullback–Leibler loss function, χ^2 loss function, Hellinger loss function etc. (see, e.g., [5], [1], [6] [8]) are perfectly mixable. We can take $0 < \eta \le \ln 2$ in the case of log-loss function and $0 < \eta \le 2$ in the case of square difference [6].

The important construction in this field is the Vovk's aggregating algorithm AA [5], [6]. In the case of perfectly mixable loss functions this algorithm given a finite sequence of predictive strategies $S_1, S_2, \dots S_k$ and weights $r_1, r_2 \dots r_k$, whose sum do not exceed 1, allows us to define their "mixture" – a prediction strategy S such that

$$\mathrm{Loss}_S(x) \le \log_\beta \sum_{i=1}^{k} r_i \beta^{\mathrm{Loss}_{S_i}(x)} \tag{1}$$

for all x, where $\beta = e^{-\eta}$ and the corresponding loss function is η-mixable. The exact construction is given in Section 8.

We fix some universal programming language. Then each computable prediction strategy S is defined by its program, which given a sequence $x = x_1, \dots, x_{i-1}$, some parameter y and integer number k outputs a rational approximation of $S(x)$ with accuracy 2^{-k}. By Kolmogorov complexity $K(S|y)$ of prediction strategy S given parameter y we mean the length of the shortest program having these properties. Unconditional complexity is defined as $K(S) = K(S|\Lambda)$ (for details see [3]).

Now we briefly review the concept of *predictive complexity* from Vovk and Gammerman [7] and Vovk and Watkins [8].

It is natural to consider loss processes corresponding to computable prediction strategies S. In this case, the value $\mathrm{Loss}_S(x)$ can be interpreted as predictive complexity of x. This value, however, depends on S and it is unclear which S to choose. In the most interesting cases a smallest loss functions does not exist – given a computable prediction strategy S, it is easy to construct a computable prediction strategy that greatly outperforms S on at least one outcome sequence. Levin [13], developing ideas of Kolmogorov and Solomonoff, suggested (for a particular loss function) a very natural solution to the problem of nonexistence of a smallest computable loss process. Vovk [7] extended these ideas in a more general setting – for arbitrary loss prosesses.

A non-negative real-valued function g is called *superprediction* if there exists a prediction p such that $g(j) \ge \lambda(j,p)$ for all j.

We will say that a function $KG(x)$ is a *measure of predictive complexity* if the following two conditions hold:

1. $KG(\Lambda) = 0$ and for each x the function $g(j) = KG(xj) - KG(x)$ is a superprediction;

2. KG is *semicomputable from above*, which means that there exists a non-increasing computable sequence of functions KG^t taking rational values such that for every x, $KG(x) = \inf_t KG^t(x)$.

Requirement (1) means that the measure of predictive complexity must be valid: there must exists a prediction strategy that achieves it. (Notice that if $\geq$ is replaced by $=$ in the definition of the superprediction a definition of a loss process will be obtained.) Requirement (2) means that it must be "computable in the limit".

Analogously to item (2) a sequence KG_i is *semicomputable from above*, if there exists a non-increasing by t computable sequence of functions KG_i^t taking rational values such that, for every i and x it holds $KG_i(x) = \inf_t KG_i^t(x)$.

In Vovk and Gammerman [7] and Vovk and Watkins [8] for any η-mixable loss function an *universal* measure of predictive complexity was defined

$$KG(x) = \log_\beta \sum_{i=1}^{\infty} \beta^{KG_i(x)} 2^{-K(i)}, \tag{2}$$

where $KG_i(x)$ is semicomputable from above sequence of *all* measures of predictive complexity, $K(i)$ is the Kolmogorov prefix complexity of the program i enumerating KG_i from above

For the definition and properties of Kolmogorov prefix complexity we refer reader to [3], Section 3. This complexity is based on *prefix-free* code. Any two programs p_1 and p_2 under this way of encoding are incompatible as binary strings. By this reason we have $\sum_{i=1}^{\infty} 2^{-K(i)} \leq 1$.

The index i in KG_i contains all information needed to enumerate it from above, so we call i enumerating program of KG_i. A sequence KG_i is defined in [12], conditions of items (1) and (2) for $KG(x)$ also are verified in that paper.

By (2) we obtain that for each measure of predictive complexity KG_i

$$KG(x) \leq KG_i(x) + (\ln 2/\eta)K(i), \tag{3}$$

holds for all x, where ln is the logarithm to the base e.

$KG(x)$ is called the *predictive complexity* of x.

Let S be any computable predictive strategy and p be a program, which given a sequence of outcomes x and a degree of accuracy computes a rational approximation of $S(x)$ with this degree of accuracy. Evidently, there exists a computable function f translating p to some enumerating program of S such that

$$Loss_S(x) = KG_{f(p)}(x). \tag{4}$$

In particular, by (3) for every computable prediction strategy S and for every x

$$KG(x) \leq Loss_S(x) + (\ln 2/\eta)(K(S) + c), \tag{5}$$

where c is a positive constant and $K(S)$ is complexity of S.

3 A priori semimeasure

The concept of predictive complexity is based on thoughtful ideas of Solomonoff and Kolmogorov on the existence of the universal objects in some classes of algorithmically effective objects. Solomonoff proposed ideas of defining the *a priori* probability distribution on the basis of the general theory of algorithms. The main problem he met was that the maximal computable probability distribution does not exist.

Levin [13] gives a precise form of Solomonoff ideas in a concept of a maximal semimeasure semicomputable from below (see also [3], Section 4.5). A real-valued function $P(x)$, where x is a finite binary sequence, is called semimeasure if

- $P(\Lambda) \le 1$;
- $P(x) \ge P(x0) + P(x1)$ for all x;
- the function P is semicomputable from below (see item (2) of the analogous definition of predictive complexity).

Levin proved that there exists a maximal to within a multiplicative positive constant factor semimeasure M semicomputable from below, i.e. for every semimeasure P semicomputable from below a positive constant c exists such that the inequality

$$cM(x) \ge P(x) \tag{6}$$

holds for all x.

It is easy to see that the function $KL(x) = -\log M(x)$ is a variant of the predictive complexity for log-loss function (see also [8]). We will prove that log-loss complexity is the maximal (to within a positive constant factor) among predictive complexities from a wide class.

Proposition 1 *Let $\lambda(\omega, p)$ be any η-mixable loss function satisfying $\lambda(0,0) = \lambda(1,1) = 0$. Then for each $\delta > 0$ a positive constant c exists such that the following hold*

1. $KG(x) \le (\ln 2/\eta)K(x) + c$;
2. $KG(x) \le (1+\delta)(\ln 2/\eta)KL(x) + c$,

where $K(x)$ is the Kolmogorov prefix complexity.

Proof. Let $x = x_1 \dots x_n$. To prove the item (1) consider the prediction strategy S which for every sequence z of length $i-1$, where $i = 1, \dots, l(x)-1$, outputs the i-th element x_i of x using the shortest program p (in the universal programming language) generating x and $S(z) = 0$ for all other z. The length of p is equal to the prefix complexity $K(x)$. Therefore, $K(S) \le K(x) + c$ for some positive constant c. By definition $\mathrm{Loss}_S(x) = 0$. The item (1) follows from (5).

The item (2) follows from the inequalities between $K(x)$ and $KL(x)$ [3], Section 4. □

At the rest of this section we consider some notions needed to give an interpretation of Proposition 3 below. Levin [2] considered combinations of probabilistic and deterministic processes as the most general class of processes of generating data. With each probabilistic process some computable probability distribution can be assigned. Each deterministic process is realized by means of an algorithm. Algorithmic processes transform sequences generated by probabilistic processes into new sequences. More precise, a probabilistic computer F is a Turing machine supplied with an additional input tape. In the process of computation this machine reads a sequence ω on this tape and produces a sequence $\omega' = F(\omega)$. We suppose that there is a computable probability distribution μ in the set of all possible ω. So we can calculate the probability

$$P(x) = \mu\{\omega | x \subseteq F(\omega)\}$$

of that the result $F(\omega)$ of the computation begins with a finite sequence x. Strictly speaking, $P(x)$ is not a probability distribution, since $F(\omega)$ may be finite for an infinite ω. It is easy to see that $P(x)$ is a semimeasure semicomputable from below. The converse result is proved in [13]: for every semimeasure $P(x)$ semicomputable from below a probabilistic computer $F(\omega)$ exists such that

$$P(x) = \mu\{\omega | x \subseteq F(\omega)\},$$

for all x, where $\mu(x) = 2^{-l(x)}$ is the uniform probability distribution in the set of all binary sequences.

Therefore, by (6) $M(x)$ is an universal upper bound of the probability of generating sequences x by probabilistic computers.

Let A be any set of binary sequences of length n. We define

$$M(A) = \sum_{x \in A} M(x).$$

Let some property Π_n defines for any n a set A_n of binary sequences of length n. Then for every probability distribution μ and algorithmic process F there exists a positive constant c (depending only on μ and F) such that

$$\mu\{\omega | F(\omega) \in A_n\} \leq cM(A_n)$$

for all n. According to our framework, if $M(A_n) \longrightarrow 0$ as n tends to infinity, then for all sufficiently large n the sequences of the length n having property Π_n form a scarce part of all objects generating in combinations of deterministic and probabilistic processes.

4 Stochastic and non-stochastic sequences

Let α and γ be some nonnegative numbers. A sequence x is called (α, γ)-*stochastic* if there exists a prediction strategy S such that $K(S) \leq \alpha$ and

$$\mathrm{Loss}_S(x) - KG(x) \leq \gamma.$$

The following Proposition 2 shows that for a wide class of loss functions "non-stochastic" sequences exist. Consider some conditions for a loss function sufficient to this proposition holds:

1. $\lambda(0,0) = \lambda(1,1) = 0$;
2. there exists a positive real number b such that $\lambda(0,p) \geq b$ or $\lambda(1,p) \geq b$ for each p;
3. the loss function $\lambda(\omega,p)$ is η-mixable for some $\eta > 0$.

The log-loss function and the squared difference satisfy these condition with $b = 1$, and $b = \frac{1}{4}$, accordingly.

Proposition 2 *For any loss function satisfying the conditions of items (1)–(3) above a positive constant c exists such that for every n there exists a binary sequence x of the length n satisfying*

1. $\mathrm{Loss}_P(x) - KG(x) \geq bn - (2\ln 2/\eta)\alpha - (\ln 2/\eta)(\log n + 2\log\log n) - c$ *for each prediction strategy P such that $K(P) \leq \alpha$;*
2. $M(x) \geq 2^{-\alpha - \log n - 2\log\log n - c}$.

As follows from the proof of Proposition 2 (Section 6.1 below) $K(x) \leq \alpha + \log n + 2\log\log n + c$, i.e. Kolmogorov complexity of x (from Proposition 2) can be sufficiently small when the total loss of each simple predictive strategy on x is sufficiently large. This holds since x is defined (in Section 6.1) by diagonal method in terms of prediction strategies of small complexity.

5 Main result

The predictive complexity determines asymptotically the minimal possible loss of forecasting. It includes also arbitrarily complex prediction strategies. Here we impose the restriction $K(S) \leq \alpha$, where α reflecting degree of computational resources allowed. We show that even in the case when α is small with respect to the length n of binary sequences (for instance $\alpha = O(\log n)$) this can help us to reach almost the minimal possible total loss incurred over most elements of the data set.

In this section we estimate how large can be the set of all non-stochastic sequences of the length n. We prove that the a priori semimeasure of this set is asymptotically decreases as n increases.

Let $D^n_{\alpha,\gamma}$ be the set of all binary sequences of the length n which are not (α,γ)-stochastic. For every $x \in D^n_{\alpha,\gamma}$ we have

$$\mathrm{Loss}_S(x) - KG(x) > \gamma$$

for each prediction strategy S such that $K(S) \leq \alpha$.

Proposition 3 • *For any perfectly mixable loss function satisfying condition of item (1) from Section 4 a positive constant c exists such that for every n, α and $1 \le \gamma \le n$ the estimate*

$$M(D^n_{\alpha,\gamma}) = \sum_{x \in D^n_{\alpha,\gamma}} M(x) \le 2^{-\alpha + 2\log n + 2\log\log n - \log\gamma + c} \tag{7}$$

holds;

- *If the predictive complexity is based on the log-loss function the corresponding estimate is the following*

$$M(D^n_{\alpha,\gamma}) = \sum_{x \in D^n_{\alpha,\gamma}} M(x) \le 2^{-\alpha + \log n + 2\log\log n - \log\gamma + c}, \tag{8}$$

This proposition is a generalization of Theorem 3 [9] for wide class of loss functions.

Inequality (7) or (8) can be interpreted as an upper bound on the probability of generating non-stochastic sequences by a probabilistic computer.

For every m by Proposition 3 the probability of generating (α, γ)-stochastic sequences is $\ge 1 - 2^{-m}$ when $\alpha + \log\gamma > 2\log n + 2\log\log n + c + m$. The last condition wittingly holds for each $1 \le \gamma \le n$ if

$$\alpha \ge 2\log n + 2\log\log n + c + m. \tag{9}$$

Proposition 3 shows that given an upper bound α satisfying (9) on the complexity of prediction strategies most sequences x are stochastic with respect to a simple prediction strategy Q with $K(Q) \le \alpha$, i.e. $\mathrm{Loss}_Q(x)$ is close to its minimal value $KG(x)$.

By Proposition 2 non-learnable objects can also exist, but by Proposition 3 and Levin's philosophy we shall meet them very rarely; non-predictable fluctuations of prices on financial markets form a scarce part in the stream of all financial data.

Now we formulate the main result which is a corollary from Propositions 2 and 3.

Theorem 1 *For any loss function satisfying conditions of items (1)–(3) of Section 4 a positive constant c exists such that*

$$2^{-\alpha - \log n - 2\log\log n - c} \le M(D^n_{\alpha,\gamma}) \le 2^{-\alpha + 2\log n + 2\log\log n - \log\gamma + c} \tag{10}$$

for each n, α and $0 \le \gamma \le bn - (2\ln 2/\eta)\alpha - (\ln 2/\eta)(\log n + 2\log\log n) - c$.

An open problem arises – can we eliminate the logarithmic terms in the inequalities (10)?

6 Proofs

6.1 Proof of Proposition 2

Let the corresponding degree of accuracy sufficient for estimation below is given. For any α let $p_1, p_2, \ldots, p_k$ be all programs of length $\leq \alpha$ which given this degree of accuracy terminate for all $z, l(z) \leq n$. For any $j = 1, \ldots k$ let $P_j(z)$ be an output of p_j on z.

We have $k < 2^{\alpha+1}$. By means of AA (Section 8) we can define an averaging prediction strategy P_α such that

$$\mathrm{Loss}_{P_\alpha}(x) \leq \log_\beta \sum_{i=1}^{k} k^{-1} \beta^{\mathrm{Loss}_{P_i}(x)}. \tag{11}$$

Let p be a program among $p_1, p_2, \ldots, p_k$ with the maximal terminating time. By means of the program p we recover k and using AA we compute all values $P_\alpha(z)$, $l(z) \leq n$, with corresponding degree of accuracy.

After that we can define a sequence $x = x_1 x_2 \ldots x_n$ as follows We compute rational approximations of $\lambda(1, P_\alpha(x_1 \ldots x_{s-1}))$ and $\lambda(0, P_\alpha(x_1 \ldots x_{s-1}))$ from below until at least one of

$$\lambda(1, P_\alpha(x_1 \ldots x_{s-1})) > b - 2^{-(s+1)} \tag{12}$$

or

$$\lambda(0, P_\alpha(x_1 \ldots x_{s-1})) > b - 2^{-(s+1)} \tag{13}$$

will hold (it is supposed that $x_1 \ldots x_{s-1} = \Lambda$ for $s = 1$). By item (2) of restrictions on loss function the inequality (12) or inequality (13) always will hold. If (12) was computed the first define $x_s = 1$, and define $x_s = 0$, otherwise. As follows from the definition $\mathrm{Loss}_{P_\alpha}(x) \geq bn - 1$. By (11)

$$\mathrm{Loss}_{P_i}(x) \geq bn - (\ln 2/\eta)\alpha - 1$$

for all $i \leq k$ (i.e. for all P such that $K(P) \leq \alpha$).

By definition $K(x|n) \leq \alpha + c$, for some positive constant c. The ordinary inequalities between conditional and unconditional prefix complexities (see [3], Section 3) imply

$$K(x) \leq K(x|n) + \log n + 2 \log\log n + c \tag{14}$$

for some positive constant c. Adding item (1) of Proposition 1 we obtain

$$KG(x) \leq (\ln 2/\eta)(\alpha + \log n + 2 \log\log n) + c$$

for some positive constant c. Hence, for all P such that $K(P) \leq \alpha$, we have

$$\mathrm{Loss}_P(x) - KG(x) \geq bn - (2 \ln 2/\eta)\alpha - (\ln 2/\eta)(\log n + 2 \log\log n) - c$$

for some positive constant c.

Since $KL(x) \leq K(x) + c$ holds for some c (see [3], Section 4.5), by (14) we have

$$M(x) \geq 2^{-\alpha - \log n - 2 \log\log n - c}.$$

$\square$.

6.2 Proof of Proposition 3

Simple intuitive explanation of the idea of the proof: the α-simple prediction strategy will be an approximation to the universal "prediction semistrategy".

Since the universal measure of predictive complexity is semicomputable from above, the function $Q(x) = \beta^{KG(x)}$ is lower semicomputable. This means that there exists a non-decreasing computable sequence of functions $k_i(x)$ taking rational values such that $Q(x) = \sup_i k_i(x)$.

The proof is much simpler when $KG(x) = KL(x) = -\log M(x)$ is the Levin's predictive complexity for log-loss function. In this case $\beta = \frac{1}{2}$ and $Q(x) = M(x)$. By definition of M we have

$$\sum_{l(x)=n} \beta^{KG(x)}(x) = \sum_{l(x)=n} M(x) \le 1. \tag{15}$$

Let p be a finite binary sequence representing the rational approximation of the real number $\sum_{l(x)=n} M(x)$ from below with accuracy $2^{-\alpha}$. Then using p and n we can effectively find integer numbers t and k such that the following conditions hold

1. $\sum_{l(x)=n} \beta^{KG^{t,k}(x)} > \sum_{l(x)=n} \beta^{KG(x)} - 2^{-\alpha}$, where $KG^{t,k}(x) = \log_\beta \sum_{i=1}^{k} \beta^{KG_i^t(x)} 2^{-i}$, and $KG_i^t(x)$ is some rational approximation from above of $KG_i(x)$ computed in t steps.
2. for each x of length $\le n$ and for each $i \le k$ the difference $KG_i^t(xj) - KG_i^t(x)$ is a superprediction.

The mixture of superpredictions is also a superprediction (we check this in Section 8). Then by item (2) the difference $KG^{t,k}(xj) - KG^{t,k}(x)$ is a superprediction for each x of the length $\le n$, and by the definition of the superprediction and by AA there exists a prediction strategy Q such that

$$\mathrm{Loss}_Q(x) \le KG^{t,k}(x) \tag{16}$$

for all such x. Since this construction is algorithmically effective,

$$K(Q|n) \le \alpha + c, \tag{17}$$

where c is a positive constant.

Temporarily in the definition of $D^n_{\alpha,\gamma}$ we will consider prediction strategies S which is α-simple conditional with respect to n, i.e. such that $K(S|n) \le \alpha$.

By definition for every $x \in D^n_{\alpha,\gamma}$ we have

$$\mathrm{Loss}_S(x) - KG(x) > \gamma$$

for each prediction strategy S such that $K(S|n) \le \alpha$.

Therefore, by equation (16) and by item (1) above for every $x \in D^n_{\alpha-c,\gamma}$ we obtain

$$\beta^\gamma \sum_{x\in D^n_{\alpha-c,\gamma}} \beta^{KG(x)} > \sum_{x\in D^n_{\alpha-c,\gamma}} \beta^{\mathrm{Loss}_Q(x)} \geq$$
$$\sum_{x\in D^n_{\alpha-c,\gamma}} \beta^{KG^{k,t}(x)} > \sum_{x\in D^n_{\alpha-c,\gamma}} \beta^{KG(x)} - 2^{-\alpha},$$

where c is such that (17) holds. This implies

$$(1-\beta^\gamma) \sum_{x\in D^n_{\alpha-c,\gamma}} \beta^{KG(x)} \leq 2^{-\alpha}. \tag{18}$$

In the case of log-loss function $\beta = \frac{1}{2}$ and $\beta^{KG(x)} = M(x)$, and then by (18) we have

$$M(D^n_{\alpha-c,\gamma}) \leq 2^{-\alpha+1}$$

for every $\gamma \geq 1$.

For other types of predictive complexities the sum $\sum_{l(x)=n} \beta^{KG(x)}$ can exceed 1. For instance, in the case of square-loss function this sum is of order of exponent from n. Let us consider the general case. We replace the inequality (15) on

$$\sum_{l(x)=n} \beta_n^{KG(x)} M(x) \leq 1,$$

where $\beta_n = e^{-\frac{1}{n}}$. Let p be a finite binary sequence representing the rational approximation of the real number $\sum_{l(x)=n} \beta_n^{KG(x)} M(x)$ from below with accuracy $2^{-\alpha}$. After that, as above using p and n we effectively find t and k such that the conditions of items (1') and (2) hold, where item

1'. $\sum_{l(x)=n} \beta_n^{KG^{t,k}(x)} M^t(x) > \sum_{l(x)=n} \beta_n^{KG(x)} M(x) - 2^{-\alpha}$.

is used instead of item (1) above. Here

$$KG^{t,k}(x) = \log_\beta \sum_{i=1}^{k} \beta^{KG_i^t(x)} 2^{-i},$$

where $\beta = e^{-\eta}$ and η is such that our loss function is η-mixable, $KG_i^t(x)$ is some rational approximation from above of $KG_i(x)$ and $M^t(x)$ is some rational approximation from below of $M(x)$ computed in t steps.

There exists a prediction strategy Q such that $\mathrm{Loss}_Q(x) \leq KG^{t,k}(x)$ for all x of length $\leq n$ and $K(Q|n) \leq \alpha + c$, where c is a positive constant. Then by (1') we obtain

$$\sum_{l(x)=n} \beta_n^{\mathrm{Loss}_Q(x)} M(x) > \sum_{l(x)=n} \beta_n^{KG(x)} M(x) - 2^{-\alpha}. \tag{19}$$

By definition for every $x \in D^n_{\alpha-c,\gamma}$ we have

$$\mathrm{Loss}_Q(x) - KG(x) > \gamma.$$

Therefore, by (19) we obtain

$$\beta_n^\gamma \sum_{x \in D^n_{\alpha-c,\gamma}} \beta_n^{KG(x)} M(x) > \sum_{x \in D^n_{\alpha-c,\gamma}} \beta_n^{\mathrm{Loss}_Q(x)} M(x) > \sum_{x \in D^n_{\alpha-c,\gamma}} \beta_n^{KG(x)} M(x) - 2^{-\alpha}.$$

This implies

$$(1-\beta_n^\gamma) \sum_{x \in D^n_{\alpha-c,\gamma}} \beta_n^{KG(x)} M(x) \le 2^{-\alpha}. \tag{20}$$

By item (2) of Proposition 1 for every $\delta > 0$ a positive constant $c > 0$ exists such that $KG(x) \le (1+\delta)(\ln 2/\eta) KL(x) + c$. We have also $KL(x) \le n + c$ for all x of length n, where c is a positive constant ([3], Section 4.5). Hence $KG(x) \le cn$ for some $c > 0$, where n is the length of x.

Since $\beta_n = e^{-\frac{1}{n}}$ we have

$$\beta_n^{KG(x)} = e^{-\frac{1}{n} KG(x)} \ge e^{-c},$$

and

$$1 - \beta_n^\gamma \ge \frac{\gamma}{2n}$$

for $0 < \gamma \le n$. Therefore, by (20) we obtain the estimate

$$\sum_{x \in D^n_{\alpha-c,\gamma}} M(x) \le 2^{-\alpha} \frac{2n}{\gamma} e^c = 2^{-\alpha + \log n - \log \gamma + c \log e + 1}. \tag{21}$$

To eliminate the condition n in $K(Q|n)$ we consider the following estimate of prefix Kolmogorov complexity [3], Section 3.

$$K(Q) \le K(Q|n) + K(n) + c' \le \alpha + \log n + 2 \log\log n + c$$

for some positive constants c' and c. Replacing α in (21) on $\alpha - \log n - 2\log\log n - c$ and returning to the previous (unconditional) definition of $D^n_{\alpha,\gamma}$ we obtain the needed estimate

$$\sum_{x \in D^n_{\alpha,\gamma}} M(x) \le 2^{-\alpha + 2\log n + 2\log\log n - \log\gamma + c} \tag{22}$$

for some positive constant c. □

7 Acknowledgements

The motivation of this paper is due to Volodya Vovk and Alex Gammerman. Author is deeply grateful to them for valuable discussions.

8 Appendix

In this section we present the Vovk's aggregating algorithm AA and prove that the mixture of superpredictions is also a superprediction.

Let a loss function $\lambda(\omega, p)$ be η-mixable and $\beta = e^{-\eta}$. Let also a finite sequence $P_1, P_2, \ldots, P_k$ of computable prediction strategies and a sequence $r(1), \ldots, r(k)$ of nonnegative real numbers which sum do not exceed 1 are given.

A computable prediction strategy P will be defined such that for each binary sequence $y_1 \ldots y_n$ the inequality

$$\mathrm{Loss}_P(y_1 \ldots y_m) \le \log_\beta \sum_{i=1}^{k} r(i) \beta^{\mathrm{Loss}_{P_i}(y_1 \ldots y_m)} \tag{23}$$

holds.

Put initial weghts $r_0(i) = r(i), i = 1, \ldots, k$. After each trial $y_j, j = 1, \ldots, n$, the weights are updated as follows

$$r_j(i) = \beta^{\lambda(y_j, P_i(y_1 \ldots y_{j-1}))} r_{j-1}(i).$$

As follows from this definition

$$r_j(i) = \beta^{\mathrm{Loss}_{P_i}(y_1 \ldots y_j)} r_0(i).$$

After each trial y_{j-1} a superprediction g_j is defined

$$g_j(\omega) = \log_\beta \sum_{i=1}^{k} \beta^{\lambda(\omega, P_i(y_1 \ldots y_{j-1}))} r^*_{j-1}(i),$$

where

$$r^*_{j-1}(i) = \frac{r_{j-1}(i)}{\sum_{s=1}^{k} r_{j-1}(s)}$$

are the normalized weights.

Since the loss function is η-mixable for each $j = 1, \ldots, n$ a real number p_j exists (can be effectively computed with arbitrary degree of accuracy) such that

$$\lambda(\omega, p_j) \le g_j(\omega)$$

for all ω. Define $P(y_1 \ldots y_{j-1}) = p_j$.

By mathematical induction on $t = 1, \ldots n$ we shall prove that

$$\sum_{j=1}^{t} g_j(y_j) = \log_\beta \sum_{i=1}^{k} \beta^{\mathrm{Loss}_{P_i}(y_1 \ldots y_t)} r(i). \tag{24}$$

The inequality (23) follows from (24) when $t = n$.

When $t = 1$ we have

$$g_1(y_1) = \log_\beta \sum_{i=1}^{k} \beta^{\lambda(y_1, P_i(\Lambda))} r(i).$$

When $t > 1$ we have

$$\log_\beta \sum_{i=1}^{k} \beta^{\mathrm{Loss}_{P_i}(y_1 \ldots y_t)} r(i) - \log_\beta \sum_{i=1}^{k} \beta^{\mathrm{Loss}_{P_i}(y_1 \ldots y_{t-1})} r(i) =$$

$$\log_\beta \frac{\sum_{i=1}^{k} \beta^{\mathrm{Loss}_{P_i}(y_1 \cdots y_{t-1}) + \lambda(y_t, P_i(y_1 \cdots y_{t-1}))} r(i)}{\sum_{i=1}^{k} \beta^{\mathrm{Loss}_{P_i}(y_1 \cdots y_{t-1})} r(i)} =$$

$$\log_\beta \sum_{i=1}^{k} \beta^{\lambda(y_t, P_i(y_1 \ldots y_{t-1}))} r^*_{t-1}(i) = g_t(y_t).$$

Let $g_i(x)$, $i = 1, \ldots, n$, be a sequence of superpredictions and $r(i)$, $i = 1, \ldots, n$, be a sequence of nonnegative weights with sum ≤ 1. We prove that their mixture

$$g(x) = \log_\beta \sum_{i=1}^{n} r(i) \beta^{g_i(x)}$$

is also a superprediction.

By definition for each $1 \leq i \leq n$ there exists an p_i such that $g_i(xj) - g_i(x) \geq \lambda(j, p_i)$ for all j. Then for all j

$$g(xj) - g(x) = \log_\beta \sum_{i=1}^{n} r(i) \beta^{g_i(xj)} - \log_\beta \sum_{i=1}^{n} r(i) \beta^{g_i(x)} \geq$$

$$\log_\beta \sum_{i=1}^{n} q(i) \beta^{g_i(xj) - g_i(x)} \geq \log_\beta \sum_{i=1}^{n} q(i) \beta^{\lambda(j, p_i)} \geq \lambda(j, p),$$

where

$$q(i) = \frac{r(i) \beta^{g_i(x)}}{\sum_{s=1}^{n} r(s) \beta^{g_s(x)}}$$

and a prediction p exists by definition of η-mixable function.

References

[1] Haussler, D., Kivinen, J. and Warmuth, M.K. (1994) Tight worst-case loss bounds for predicting with expert advice. Technical Report UCSC-CRL-94-36, University of California at Santa Cruz, revised December 1994. Short version in P. Vitányi, editor, *Computational Learning Theory*, Lecture Notes in Computer Science, volume 904, pages 69–83, Springer, Berlin, 1995.

[2] Levin, L.A. (1984) Randomness conservation inequalities; Information and independence in mathematical theories. *Inform. Comp.*, **61**, 15–37.

[3] Li, M. and Vitányi, P. (1997) *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York, 2nd edition.

[4] Shen, A. (1983) The concept of (α, β)-stochasticity in the Kolmogorov sense and its properties. *Sov. Math. Dokl.*, 28, 295–299.

[5] Vovk, V. (1990) Aggregating strategies. In M. Fulk and J. Case, editors, *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, pages 371–383, San Mateo, CA, 1990. Morgan Kaufmann.

[6] Vovk, V. (1998) A game of prediction with expert advice. *J. Comput. Syst. Sci.*, 56:153–173.

[7] Vovk, V. and Gammerman, A. (1999) Complexity estimation principle, *The Computer Journal*, *42:4*, 318–322.

[8] Vovk, V., Watkins, C.J.H.C. (1998) Universal portfolio selection, *Proceedings of the 11th Annual Conference on Computational Learning Theory*, 12–23.

[9] V'yugin, V.V. (1985) Nonstochastic objects, *Problems Inform. Transmission* **21**, 77–88.

[10] V'yugin, V.V. (1987) On the defect of randomness of a finite object with respect to measures with given complexity bounds, *SIAM Theory Probab. Appl.* **32**, 508–512.

[11] V'yugin, V.V. (1998) Non-stochastic infinite and finite sequences, *Theoretical Computer Science.* **207**, 363–382.

[12] V'yugin, V.V. (2000) Sub-optimal measures of predictive complexity for absolute loss function, *Information and Computation* (submitted for publication 5/19/2000).

[13] Zvonkin, A.K. and Levin, L.A. (1970) The complexity of finite objects and the algorithmic concepts of information and randomness, *Russ. Math. Surv.* **25**, 83–124.