Sub-optimal measures of predictive complexity for absolute loss function

V.V. V'yugin

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

Technical Report

CLRC TR-00-05

April 24, 2002

Department of Computer Science
Egham, Surrey TW20 0EX, England

Abstract

The problem of existence of predictive complexity for the absolute loss game is studied. Predictive complexity is a generalization of Kolmogorov complexity which bounds the ability of any algorithm to predict elements of a sequence of outcomes. For perfectly mixable loss functions (logarithmic and squared difference are among them) predictive complexity is defined like Kolmogorov complexity to within an additive constant. The absolute loss function is not perfectly mixable, and the question on existence of the corresponding predictive complexity which is defined to within an additive constant is open. We prove that in the case of the absolute loss game the predictive complexity can be defined to within an additive term $O(\sqrt{n})$, where n is the length of a sequence of outcomes. We prove also that in some restricted setting this bound cannot be improved.

1 Introduction

A central problem in machine learning (and statistics) is the problem of predicting future events based on past observations. We consider only the simplest case, where events are simple binary outcomes. A prediction algorithm makes its prediction in a form of a real number between 0 and 1. The quality of prediction is measured by a loss function $\lambda(\sigma, p)$, where σ is an outcome and $0 \leq p \leq 1$ is a prediction. Various loss functions $\lambda(\sigma, p)$ are considered in the literature on machine learning and prediction with expert advice (see, for example, [9], [2], [11], [13], [3]). In this paper we concentrate our interest in the *absolute* loss function $\lambda(\sigma, p) = |\sigma - p|$. A popular interpretation of the absolute loss function is that it is the expectation of the learner loss in the simple prediction game, where a biased coin is tossed and outcome 1 is predicted with probability p and outcome 0 is predicted with probability $1 - p$.

In the framework of Dawid [1], Vovk [8], Rissanen [6] and others no assumptions whatsoever are made about the actual sequence of events that is observed. The analysis is done in the worst case over all possible binary outcomes sequences. The typical setting is that we have a set of N experts predicting the same sequence of outcomes. Our goal is to construct an algorithm which performs as well as the best expert no matter what outcome sequence is produced by nature. Specifically, let $L_i(y)$ denote the total loss of an expert $i = 1, 2, \ldots, N$ on a particular sequence y. Then our goal is to minimize the maximum of the difference $L(y) - L_i(y)$ over $i = 1, 2, \ldots, N$, where $L(y)$ is the total loss of our "aggregating" algorithm. A family of predicting algorithms achieving this goal was constructed in Vovk [8], Haussler et al. [2] and others.

In Cesa-Bianchi et al. [3] the upper and lower bounds of the performance of predicting algorithms were obtained in the case of the absolute loss function. An aggregating algorithm P was developed such that for any set of N experts for each expert $i = 1, 2, \ldots, N$ an inequality

$$L_P(y) - L_i(y) \leq c_1 \sqrt{L_i(y) \ln N} + c_2 \ln N \tag{1}$$

holds for all binary sequences y of sufficiently large length n, where $L_P(x)$ is the total loss suffered by P on a sequence x and c_1, c_2 are positive constants. Also, no algorithm P can improve this estimate as $N, n \to \infty$.

Vovk in [11] and [10] proposed in the spirit of Kolmogorov's and Solomonoff's framework an "ideal" setting, where a generalized prediction algorithm is considered that performs as well as any expert from an infinite pool of all possible experts. All experts with computable prediction strategies are in this pool. The correct definition of this algorithm will be given in Sections 2 and 4.1. Under this setting for some "good" loss functions (perfectly mixable [9]), like logarithmic and square, it is possible to prove that like traditional Kolmogorov complexity there exists an optimal measure of predictive complexity $KA(x)$ such that for any other measure of predictive complexity $KA_i(x)$

$$KA(x) \leq KA_i(x) + O(1) \tag{2}$$

holds for each finite sequence x of data. It follows from (2) that in the case of perfectly mixable loss function the *predictive complexity* $KA(x)$ is defined (like Kolmogorov complexity) up to an additive constant.

The absolute loss function is not perfectly mixable. We prove that in the absolute loss case the term $O(1)$ in (2) can be replaced on a term $O(\sqrt{l(x)})$ where $l(x)$ is the length of x, namely, a *sub-optimal* measure of predictive complexity $KA(x)$ exists such that for any other measure of predictive complexity $KA_i(x)$ the inequality

$$KA(x) \leq KA_i(x) + O(\sqrt{l(x)}) \tag{3}$$

holds for all x. It is clear that for any two sub-optimal measures of predictive complexity $KA(x)$ and $KA'(x)$

$$KA(x) = KA'(x) + O(\sqrt{l(x)}).$$

We fix some sub-optimal (or universal) measure $KA(x)$ satisfying (3) and call it *predictive complexity for the absolute loss game*. By (4) the predictive complexity is defined to within an $O(\sqrt{l(x)})$ term.

A natural question arises whether (2) holds for absolute loss case, i.e. whether predictive complexity $KA(x)$ can be defined such that (2) holds for this $KA(x)$. It would be ideal to prove that the predictive complexity for absolute loss function cannot be defined with better accuracy than $O(\sqrt{l(x)})$: for any measure of predictive complexity $KA(x)$ there exists another measure of predictive complexity $KA'(x)$ such that

$$\limsup_{n\to\infty} \sup_{x:l(x)=n} \frac{KA(x) - KA'(x)}{\sqrt{n}} > 0.$$

This question still remains open. This paper studies this problem in a more restricted setting. A stronger version of (2) is

$$KA(x) \leq KA_i(x) + O(K(i)), \tag{4}$$

where $K(i)$ is Kolmogorov prefix complexity of a program i enumerating predictive complexity $KA_i(x)$ from above.

Inequality (4) in many ways is more useful than (2). We study to what degree (4) can be extended to the absolute loss case. We define a measure of predictive complexity $KA(x)$ such that

$$KA(x) \leq KA_i(x) + O(\sqrt{l(x)})K(i) \tag{5}$$

for each i and x. We prove that inequality (5) cannot be improved in the following sense. The total loss $L_S(x)$ suffered by any computable prediction strategy S on a sequence x can be represented as $L_S(x) = KA_i(x)$ for some i. Then by (5) the inequality

$$KA(x) \leq L_S(x) + O(\sqrt{n})K(S), \tag{6}$$

holds for each computable prediction strategy S for each n and for each sequence x of the length n, where $K(S)$ is the Kolmogorov prefix complexity of the prediction strategy S. We extend the result of Cesa-Bianchi et al. [3] (Section 3.2, Theorem 8) to an arbitrary linearly bounded measure of predictive complexity $KA(x)$. We prove that if some nondecreasing function f and constants c_1, c_2, c_3 satisfy the inequality

$$KA(x) \leq L_S(x) + f(c_1 n)(c_2 + c_3 K(S))$$

for each computable prediction strategy S for each n and for each sequence x of the length n then

$$\liminf_{n\to\infty} \frac{f(n)}{\sqrt{n/\log^2 n}} = \infty.$$

2 Measures of predictive complexity

Let a sequence $x_1, x_2, \ldots, x_i \ldots$ of some data is given, where $x_i \in \{0,1\}$. 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+1} given $x_1, x_2, \ldots x_i$, etc. At every step i the loss is measured by some nonnegative 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. Several loss functions were considered in Vovk [9], Haussler et al. [2] (log-loss, Hellinger, etc.), Vovk and Watkins [11] (financial theory), Yamanishi [13] (logistic, etc).

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

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

The main task is to minimize the total loss suffered on a sequence $x = x_1 x_2 \ldots x_n$ of outcomes. A set of all possible outcomes σ, a set of all predictions p, and a loss function $\lambda(\sigma, p)$ are called the *game*. The corresponding game-theoretic interpretation was introduced by Littlestone and Warmuth [5] (for details we refer readers to Vovk [11]).

Let us fix $\eta > 0$ (the *learning rate*) and put $\beta = e^{-\eta} \in (0,1)$. Let c_η be an infinum of all c such that for each sequence of weights $p_1, p_2, \dots$ with a sum ≤ 1 there exists a prediction $\hat{\gamma}$ such that

$$\lambda(j, \hat{\gamma}) \leq c \log_\beta \sum_i p_i \beta^{\lambda(j,\gamma)} \tag{7}$$

for $j = 0, 1$. By Vovk [8], [9] (Section 2), $c_\eta = 1$ for an appropriate η, in the case of logarithmic loss function $\lambda(j,p)$, where $\lambda(j,p) = -\log p$ if $j = 1$, and $\lambda(j,p) = -\log(1-p)$ if $j = 0$, and in the case of square loss function $\lambda(j,\gamma) = (j-\gamma)^2$. By $\log p$ we mean the logarithm of p by the base 2. For an absolute loss function $\lambda(j,p) = |j-p|$ it was proven in Haussler et al. [2] (Section 4.2) that $c_\eta = (\ln \frac{1}{\beta})/(2 \ln \frac{2}{1+\beta})$.

If $c_\eta = 1$ for some η then the corresponding loss function (game) is called perfectly mixable.

The absolute loss game is not perfectly mixable, since $c_\eta > 1$ for each $\eta > 0$.

The Vovk's aggregating algorithm AA [8], [9] given a finite sequence of predictive strategies $S_1, S_2, \dots S_k$ and weights $r_1, r_2 \dots r_k$ with a sum ≤ 1 allows us to define their "mixture", i.e. a prediction strategy S such that

$$L_S(x) \leq c_\eta \log_\beta \sum_{i=1}^{k} r_i \beta^{L_{S_i}(x)} \tag{8}$$

for all x.

Let us consider the total loss function corresponding to a computable prediction strategy S. In this case, the value $L_S(x)$ can be interpreted as a predictive complexity of x. This value, however, depends on S and it is unclear which S to choose. Levin [12], developing ideas of Kolmogorov and Solomonoff, suggested (for the logarithmic loss function) a very natural solution to the problem of existence of a smallest measure of predictive complexity. Vovk [10] extended these ideas in a more general setting for a wide class of loss functions.

We suppose that our loss function $\lambda(\sigma, p)$ is computable by an algorithm. This means that if $\lambda(\sigma,p) < \infty$ then given an arbitrary degree of accuracy $\epsilon > 0$ we can compute a rational approximation of $\lambda(\sigma,p)$ with the accuracy ϵ using some rational approximation of the real number p. More precise, there are two computable sequences of simple functions $\lambda^t(\sigma,p)$ nonincreasing by t, and $\lambda_t(\sigma,p)$ nondecreasing by t, such that $\lambda(\sigma,p) = \inf_t \lambda^t(\sigma,p)$ and $\lambda(\sigma,p) = \sup_t \lambda_t(\sigma,p)$. By a simple function we mean a function whose domain is a union of intervals with rational endpoints. This function is constant on each of them and takes rational values or $+\infty$. Simple functions are constructive objects and can be enumerated by positive integer numbers. By this definition any computable loss function $\lambda(\sigma,p)$ is continuous by p.

Let Ξ be a set of all finite binary sequences. By a simple function on Ξ we mean a function which takes nonnegative rational values or $+\infty$ and equals $+\infty$ for almost all $x \in \Xi$.

A function $KA(x)$ is a *measure of predictive complexity* if the following two conditions hold:

- (i) $KA(\Lambda) = 0$ and for each x there exists an p such that

$$KA(xj) \geq KA(x) + \lambda(j, p) \tag{9}$$

  for each $j = 0, 1$.

- (ii) $KA(x)$ is *semicomputable from above*, which means that there exists a computable sequence of simple functions $KA^t(x)$ nonincreasing by t, and such that $KA(x) = \inf_t KA^t(x)$ for each x.

Requirement (i) 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 (9), the definition of a total loss function will be obtained.) Requirement (ii) means that it must be "computable in the limit".

The main advantage of this definition is that a semicomputable from above sequence $KA_i(x)$ of all measures of predictive complexity can be constructed. Semicomputability of the sequence $KA_i(x)$ means that there exists a computable from i, t, x sequence of simple functions $KA_i^t(x)$ such that

- (iii) $KA_i^{t+1}(x) \leq KA_i^t(x)$ for all i, t, x;
- (iv) $KA_i(x) = \inf_t KA_i^t(x)$ for all i, x;
- (v) for each measure of predictive complexity $KA(x)$ there exists an i such that $KA = KA_i$.

A sequence $KA_i(x)$ satisfying (i)-(v) will be constructed in Section 4.1.

We suppose that some universal programming language is fixed. The index i in KA_i contains all information needed to enumerate it from above, so we call i an enumerating program of KA_i.

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 the value of $S(x)$ with this degree of accuracy. By $K(S)$ we denote the length of the shortest program p computing S. Evidently, there exists a computable function f translating p to some enumerating program of S, namely

$$L_S(x) = KA_{f(p)}(x). \tag{10}$$

The following theorem is based on Vovk's construction [11].

Theorem 1 *Let $KA_i(x)$ be a semicomputable from above sequence of all measures of predictive complexity, $\lambda(\omega, \gamma)$ be a loss function and $\beta = e^{-\eta}$, where η is a learning rate. Then there exists a measure of predictive complexity $KA(x)$ such that*

$$KA(x) \leq c_\eta KA_i(x) + c_\eta(\ln 2/\eta)K(i), \tag{11}$$

holds for all i and x, where $K(i)$ is the Kolmogorov prefix complexity of a number i (see e.g. Li and Vitanyi [4], Section 3).

In particular, for each computable prediction strategy S and for each x

$$KA(x) \leq c_\eta L_S(x) + c_\eta(\ln 2/\eta)(K(S) + c), \tag{12}$$

where c is a constant.

The proof of this theorem is given in Section 4.1.

For perfectly mixable loss functions we have $c_\eta = 1$ for an appropriate η, and so, inequality (11) is analogous to that holds for optimal Kolmogorov complexity.

3 Main results

In Theorems 2 and 3 below we consider only the absolute loss function $\lambda(\omega, \gamma) = |\omega - \gamma|$. In the previous theorem the learning rate η was a constant. In the case of an absolute loss function we have in (11) the constant factor c_η which is bigger than 1 for each $\eta > 0$. In this section we show that AA with a variable learning rate allows us to construct predictive complexity with more optimal (in several cases) bound of its relative performance. Recall, that $l(x)$ denotes the length of a finite binary sequence x.

Theorem 2 *Let $KA_i(x)$ be a semicomputable from above sequence of all measures of predictive complexity. Then there exists a sub-optimal measure of predictive complexity $KA(x)$ such that for each i and each x*

$$KA(x) \leq KA_i(x) + (\sqrt{l(x)} \ln 2) K(i). \tag{13}$$

In particular, for each computable prediction strategy S and for each x

$$KA(x) \leq L_S(x) + (\sqrt{l(x)} \ln 2)(K(S) + c), \tag{14}$$

where c is a constant.

The proof of this theorem is given in Section 4.2.

Corollary 1 *For any two sub-optimal measures of predictive complexity for absolute loss function $KA(x)$ and $KA'(x)$*

$$KA(x) = KA'(x) + O(\sqrt{l(x)}). \tag{15}$$

We fix some sub-optimal measure of predictive complexity $KA(x)$ satisfying (13), (14) and call it *the predictive complexity for absolute loss function.* Relation (15) means that predictive complexity for the absolute loss function is defined to within an $O(\sqrt{l(x)})$ term.

In the following Theorem 3 we show that an additional square root in the estimates (13) and (14) cannot be essentially decreased.

We call a measure of predictive complexity $KA(x)$ linearly bounded if for some positive constant c_0 the inequality

$$KA(x) \leq c_0 l(x) \tag{16}$$

holds for all x. The sub-optimal measures of predictive complexity defined by (20) and in Theorem 2 are linearly bounded, since we can take in (14) S equal to a trivial predictive strategy which always predicts $\frac{1}{2}$.

The following theorem shows that we cannot construct a sub-optimal measure of predictive complexity which is better than that in Theorem 2.

Theorem 3 *Let $KA(x)$ be a linearly bounded measure of predictive complexity for absolute loss function and let f be a nondecreasing function such that for some constants c_1, c_2 and c_3 the inequality*

$$KA(x) \leq L_S(x) + f(c_1 n)(c_2 + c_3 K(S)) \tag{17}$$

holds for each computable prediction strategy S for each n and for each sequence x of the length n. Then

$$\liminf_{n\to\infty} \frac{f(n)}{\sqrt{n/\log^2 n}} = \infty. \tag{18}$$

The proof of this theorem is given in Section 4.3.

Whether can we replace $\sqrt{l(x)}$ in (14) by $\sqrt{L_S(x)}$ is an open problem.

4 Proofs

4.1 Proof of Theorem 1

A sequence $KA_i(x)$ of all measures of predictive complexity can be defined using standard methods of the theory of algorithms as follows. We will consider the recursively enumerable (r.e.) sets as consisting of pairs (x, r), where x is a finite binary sequence and r is a nonnegative rational number (all such pairs can be effectively encoded using all natural numbers). Let W be a universal r.e. set such that for each r.e. set A (consisting of pairs (x, r) as mentioned above) there exists a natural number i such that $A = W_i$, where W_i is a projection of W defined by i, i.e. $W_i = \{(x, r) | (i, x, r) \in W\}$. The existence of this set is the central result of the theory of algorithms (see Rogers [7]).

By computability of $\lambda(\sigma, p)$ a computable sequence of simple functions $\lambda^t(\sigma, p)$ exists such that $\lambda^{t+1}(\sigma, p) \leq \lambda^t(\sigma, p)$ for all t, σ, p and $\lambda(\sigma, p) = \inf_t \lambda^t(\sigma, p)$.

Let W^t be a finite subset of W enumerated in t steps. Define

$$W_i^t = \{(x, r) | \exists r'((i, x, r') \in W^t, r \geq r')\} \cup (\Xi \times \{+\infty\}).$$

It is easy to define a computable sequence of simple functions $KA_i^t(x)$ such that $KA_i^0(x) = \infty$ and $KA_i^{t+1}(x) \leq KA_i^t(x)$ for all x. Besides, $KA_i^t(x)$ is a minimal (under $\leq$) simple function whose graph is a subset of W_i^t and such that for each x a rational p exists for which

$$KA_i^t(x\sigma) - KA_i^t(x) \geq \lambda^t(\sigma, p) \tag{19}$$

holds for each $\sigma = 0, 1$. Define $KA_i(x) = \inf_t KA_i^t(x)$ for each i and x. It follows from (19) and continuity of $\lambda(\sigma, p)$ by p that for any i the function $KA_i(x)$ is a measure of predictive complexity.

Let a function $KA(x)$ satisfies the conditions (i), (ii) of the definition of a measure of predictive complexity and $W_i = \{(x, r) | r > KA(x)\}$, where r is a rational number. It is easy to verify that $KA(x) = KA_i(x)$ for all x.

Let r_i be a semicomputable from below sequence of real numbers such that $\sum\limits_{i=1}^{\infty} r_i \le 1$. For instance, we can take $r_i = 2^{-K(i)}$, where $K(i)$ is the Kolmogorov prefix complexity of i.

Analogously to Vovk and Gammerman [10] and Vovk and Watkins [11] define a function $KA(x)$ as follows

$$KA(x) = c_\eta \log_\beta \sum_{i=1}^{\infty} \beta^{KA_i(x)} r_i, \tag{20}$$

We prove that $KA(x)$ is a measure of predictive complexity. By definition $KA(x)$ is semicomputable from above, i.e (ii) holds. We must verify (i). Indeed, by (20) for every x and $j = 0, 1$

$$KA(xj) - KA(x) = c_\eta \log_\beta \sum_{i=1}^{\infty} q_i \beta^{KA_i(xj) - KA_i(x)} \ge \tag{21}$$

$$c_\eta \log_\beta \sum_{i=1}^{\infty} q_i \beta^{\lambda(j,\gamma_i)} \ge \lambda(j, \gamma), \tag{22}$$

where

$$q_i = \frac{r_i \beta^{KA_i(x)}}{\sum\limits_{s=1}^{\infty} r_s \beta^{KA_s(x)}}.$$

Here for any i a prediction γ_i satisfying

$$KA_i(xj) - KA_i(x) \ge \lambda(j, \gamma_i)$$

exists since each element of the sequence $KA_i(x)$ satisfies the condition (i) of the measure of predictive complexity. A prediction γ satisfying (22) exists by definition of the constant c_η from Section 2. For further details see [11], Section 7.6.

4.2 Proof of Theorem 2

Let us define

$$KA_i^*(x) = KA_i(x) + \sum_{k=1}^{l(x)-1} \frac{1}{2\sqrt{k}}, \tag{23}$$

where $KA_i(x)$ is a semicomputable from above sequence of all measures of predictive complexity satisfying (i)-(v) of Section 2.

Let $\beta_n = e^{-\frac{1}{\sqrt{n}}}$. For any binary sequence x of length n define

$$KA(x) = \log_{\beta_n} \sum_{i=1}^{\infty} p_i \beta_n^{KA_i^*(x)}, \tag{24}$$

where $p_i = 2^{-K(i)}$.

By definition the function $KA(x)$ is semicomputable from above. Let us check the condition (i) of the measure of predictive complexity. We have for each x of length n

$$\beta_n^{KA(x)} = \sum_{i=1}^{\infty} p_i \beta_n^{KA_i^*(x)} \tag{25}$$

and for each $j = 0, 1$

$$\beta_{n+1}^{KA(xj)} = \sum_{i=1}^{\infty} p_i \beta_{n+1}^{KA_i^*(xj)}. \tag{26}$$

Let $\epsilon = \log_{\beta_{n+1}} \beta_n - 1$; then $\beta_n = \beta_{n+1}^{1+\epsilon}$.

By the concavity of a function $y = x^{1+\epsilon}$, where $\epsilon > 0$, we obtain an inequality

$$\left(\sum p_i a_i\right)^{1+\epsilon} \leq \sum p_i a_i^{1+\epsilon}.$$

Using this inequality we obtain

$$\beta_n^{KA(xj)} = \beta_{n+1}^{(1+\epsilon)KA(xj)} \leq \sum_{i=1}^{\infty} p_i \beta_{n+1}^{(1+\epsilon)KA_i^*(xj)} = \sum_{i=1}^{\infty} p_i \beta_n^{KA_i^*(xj)}. \tag{27}$$

Dividing (27) on (25) we obtain

$$\beta_n^{KA(xj)-KA(x)} \leq \sum_{i=1}^{\infty} q_i \beta_n^{KA_i^*(xj)-KA_i^*(x)}, \tag{28}$$

where

$$q_i = \frac{p_i \beta_n^{KA_i^*(x)}}{\sum\limits_{k=1}^{\infty} p_k \beta_n^{KA_k^*(x)}}$$

are weights summing to 1.

By definition for each $i = 1, 2, \ldots$ a prediction $\hat{\gamma}_i$ exists such that for $j = 0, 1$

$$KA_i^*(xj) - KA_i^*(x) = KA_i(xj) - KA_i(x) + \frac{1}{2\sqrt{n}} \geq \lambda(j, \hat{\gamma}_i) + \frac{1}{2\sqrt{n}}. \tag{29}$$

By definition of $\beta_n = e^{-\frac{1}{\sqrt{n}}}$ and by [2] (Section 4.2) or by [9] (Section 2) for any n we can put $c_\eta = (\ln \frac{1}{\beta_n})/(2 \ln \frac{2}{1+\beta_n})$ in the absolute loss case. By (28), (29) and by definition (7) of c_η a prediction $\hat{\gamma}$ exists such that for $j = 0, 1$

$$\beta_n^{KA(xj)-KA(x)} \leq \sum_{i=1}^{\infty} q_i \beta_n^{\lambda(j,\hat{\gamma}_i)} \beta_n^{\frac{1}{2\sqrt{n}}} \leq$$
$$\beta_n^{c_\eta^{-1}\lambda(j,\hat{\gamma})} \beta_n^{\frac{1}{2\sqrt{n}}} \leq \beta_n^{(1-\frac{1}{2\sqrt{n}})\lambda(j,\hat{\gamma})} \beta_n^{\frac{1}{2\sqrt{n}}} =$$
$$\beta_n^{\lambda(j,\hat{\gamma})+\frac{1}{2\sqrt{n}}(1-\lambda(j,\hat{\gamma}))} \leq \beta_n^{\lambda(j,\hat{\gamma})}.$$

Then we have $KA(xj) - KA(x) \geq \lambda(j, \hat{\gamma})$, i.e. (i) is true. Hence, the function $KA(x)$ defined by (24) is a measure of predictive complexity. By (24) for each i and x of length n we have $\beta_n^{KA(x)} \geq p_i \beta_n^{KA_i^*(x)}$. Then we obtain

$$KA(x) \leq KA_i^*(x) + \log_{\beta_n} p_i = KA_i(x) + \frac{1}{2}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k}} + \sqrt{n}K(i)\ln 2) \leq \tag{30}$$

$$KA_i(x) + \sqrt{n}(1 + K(i)\ln 2). \tag{31}$$

We can ignore the term 1 in (31) since $K(i)$ is defined up to an additive constant. The inequality (14) follows from (10), since it holds $K(f(p)) \le K(p) + c$ for some constant c.

4.3 Proof of Theorem 3

Any finite binary sequence $\alpha = \alpha_1\alpha_2\ldots\alpha_n$ of length n defines a *static* prediction strategy $S(x_1 \ldots x_{i-1}) = \alpha_i$ for $1 \le i \le n$. The total loss suffered by this prediction strategy over $x = x_1x_2\ldots x_n$ is equal to

$$L_\alpha(x) = L_S(x) = \sum_{i=1}^{n} |x_i - \alpha_i|.$$

We will consider the mathematical expectation of an arbitrary function g with respect to the uniform measure $L(x) = 2^{-l(x)}$

$$E_n(g(x)) = \sum_{l(x)=n} g(x)L(x).$$

Let Ξ_n be a set of all binary sequences of the length n. For any finite set E by $|E|$ we denote the cardinality of E. For any set $E \subseteq \Xi_n$ of finite binary sequences of the length n (static prediction strategies) we define

$$R_n(E) = E_n(\min_{\alpha\in E} L_\alpha(x))$$

and

$$R_{N,n} = \min_{|E|=N, E\subseteq\Xi_n} R_n(E).$$

Our proof is based on the following probabilistic result from Cesa-Bianchi et al. [3].

Lemma 1 *It holds*

$$\liminf_{N\to\infty}\liminf_{n\to\infty} \frac{\frac{n}{2} - R_{N,n}}{\sqrt{\frac{n}{2}\ln N}} \ge 1. \tag{32}$$

For the proof see Corollary 7 and the proof of Lemma 6 from [3].

We will use also the following trivial

Lemma 2 *Let $\xi(y)$ be any function on Ξ_n and $E_n(\xi) \ge \gamma$. Then for any $\epsilon > 0$*

$$L\{y | l(y) = n, \xi(y) > (1-\epsilon)\gamma\} \ge \frac{\epsilon\gamma}{\max_{l(y)=n}\xi(y)}.$$

Proof. Indeed, this inequality follows from

$$\gamma \le E_n(\xi) = \sum_{\{y|l(y)=n,\xi(y)\le(1-\epsilon)\gamma\}} \xi(y)L(y) + \sum_{\{y|l(y)=n,\xi(y)>(1-\epsilon)\gamma\}} \xi(y)L(y) \le$$
$$(1-\epsilon)\gamma + \max_{l(y)=n}\xi(y)L\{y|l(y)=n, \xi(y) > (1-\epsilon)\gamma\}.$$

□

For any n define $\beta_n = e^{-\frac{1}{n}}$ and $\alpha_n = 2\log n$. Let p_n be a finite binary sequence representing the rational approximation of the real number $\sum\limits_{l(x)=n} \beta_n^{KA(x)} L(x)$ from below with accuracy $2^{-\alpha_n}$ (this sum is ≤ 1). Then using p_n and n we can effectively find an integer number t such that the following conditions hold

- ◇ (1) $\sum\limits_{l(x)=n} \beta_n^{KA^t(x)} L(x) > \sum\limits_{l(x)=n} \beta_n^{KA(x)} L(x) - 2^{-\alpha_n}$, where $KA^t(x)$ is some approximation from above of $KA(x)$ computed in t steps.
- ◇ (2) for each x of the length $\le n$ there exists a real number $\hat{\gamma}$ such that for each $j = 0, 1$
  $KA^t(xj) - KA^t(x) \ge \lambda(j, \hat{\gamma})$;
- ◇ (3) $KA^t(x) \le 2c_1 l(x)$ for all $x, l(x) \le n$, where c_1 is from (16).

By $E(\lambda(j,\gamma)) = (1-\gamma)\frac{1}{2} + \gamma\frac{1}{2} = \frac{1}{2}$ and (2) we obtain $E_n(KA^t(y)) \ge \frac{n}{2}$.

Let

$$D_{n,t} = \{x | l(x) = n, KA^t(x) - KA(x) > 1\}.$$

We have $\beta_n^{KA^t(x)} < \beta_n^{KA(x)} \beta_n$ for each $x \in D_{n,t}$. Then

$$\beta_n \sum_{x \in D_{n,t}} \beta_n^{KA(x)} L(x) \ge \sum_{x \in D_{n,t}} \beta_n^{KA(x)} L(x) - 2^{-\alpha_n}.$$

Therefore, we obtain

$$(1-\beta_n) \sum_{x \in D_{n,t}} \beta_n^{KA(x)} L(x) \le 2^{-\alpha_n}.$$

We have also $1 - \beta_n \ge 1/2n$ and $\beta_n^{KA(x)} \ge e^{-c_0}$, where c_0 is that from (16). Then we obtain $\frac{1}{2n} e^{-c_0} \sum\limits_{x \in D_{n,t}} L(x) \le 2^{-\alpha_n}$ or

$$L(D_{n,t}) \le 2^{-\alpha_n + \log n + 1 + c_0 \log e} \le \frac{c}{n}$$

for some positive constant c. Hence, we have

$$KA^t(x) \le KA(x) + 1 \tag{33}$$

for all x of the length n with an exception of a portion c/n of such x.

By Lemma 1 (32) there exists an N_0 such that for each $N \ge N_0$ there exists an n_N such that for each $n \ge n_N$ there is a set $E_{n,N}$ of N static prediction strategies of the length n such that

$$\frac{\frac{n}{2} - E_n(\min_{\alpha \in E_{n,N}} L_\alpha(y))}{\sqrt{\frac{n}{2} \ln N}} > \frac{1}{2}.$$

The inequality $E_n(KA^t(y)) \geq \frac{n}{2}$ implies

$$E_n(KA^t(y) - \min_{\alpha \in E_{n,N}} L_\alpha(y)) \geq \frac{n}{2} - E_n(\min_{\alpha \in E_{n,N}} L_\alpha(y)) > \frac{1}{2}\sqrt{\frac{n}{2}\ln N}.$$

By Lemma 2 for $\epsilon = \frac{1}{2}$, $\xi(y) = KA^t(y) - \min_{\alpha \in E_{n,N}} L_\alpha(y)$ and $\gamma = \frac{1}{2}\sqrt{\frac{n}{2}\ln N}$ we have

$$L\{y|l(y) = n, KA^t(y) - \min_{\alpha \in E_{n,N}} L_\alpha(y) > \frac{1}{4}\sqrt{\frac{n}{2}\ln N}\} \geq \tag{34}$$

$$\frac{\sqrt{\frac{n}{2}\ln N}}{8c_1 n} = \frac{c\sqrt{\ln N}}{\sqrt{n}}, \tag{35}$$

where c is a constant.

A finite set $E_{n,N}$ of cardinality N satisfying (34) can be found effectively by n, N and p_n (using exhaustive search).

Since we have for some constant $c > 0$

$$L\{y|l(y) = n, KA^t(y) - KA(y) > 1\} < \frac{c}{n}$$

for all sufficiently large n, for each such n there exists an y of the length n such that

$$KA^t(y) - \min_{\alpha \in E_{n,N}} L_\alpha(y) > \sqrt{\frac{n}{32}\ln N}, \tag{36}$$

$$KA^t(y) \leq KA(y) + 1. \tag{37}$$

For each y satisfying (36) there exists an $\alpha \in E_{n,N}$ such that

$$KA^t(y) - L_\alpha(y) > \sqrt{\frac{n}{32}\ln N}. \tag{38}$$

Given $E_{n,N}$ we can specify any $\alpha \in E_{n,N}$ using $\log N + c$ bit, where c is a constant, and so,

$$K(\alpha) \leq 2\log N + 3\log n + c, \tag{39}$$

where c is a constant.

Combining (37), (38) and (39) with the condition of Theorem 3 we obtain

$$L_\alpha(y) + \sqrt{\frac{n}{32}\ln N} \leq KA^t(y) \leq KA(y) + 1 \leq$$
$$L_\alpha(y) + f(c_1 n)(c_2 + c_3 K(\alpha)) + 1.$$

Then

$$\sqrt{\frac{n}{32}\ln N} \leq f(c_1 n)(c_2 + 3c_3\log n + 2c_3\log N + c_4) + 1,$$

where c_4 is a constant. Hence, a constant c exist such that for each $N \geq N_0$ for all sufficiently large n it holds

$$\frac{f(n)}{\sqrt{n/\log^2 n}} \geq c\sqrt{\ln N}.$$

In other words,

$$\liminf_{n\to\infty} \frac{f(n)}{\sqrt{n/\log^2 n}} = \infty.$$

5 Acknowledgements

The problem of existing of predictive complexity for the absolute loss game was posed by Volodya Vovk. Author is grateful to him for useful discussions. Author also thanks an anonimous referee whose suggestions clarify the place of this work in the theory of predictive complexity.

References

[1] Dawid, A.P. (1984) Statistical theory: The prequential approach (with discussion). *J. of the Royal Statist. Soc., Series A*, **147**, 278–292.

[2] Haussler, D., Kivinen, J., 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.

[3] Cesa-Bianchi, N., Freund, Y., Helmbold, D.P., Haussler, D., Schapire, R.E., Warmuth, M.K. (1997) How to use expert advice. *Journal of the ACM*, **44**, 427–485

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

[5] Littlestone, N., Warmuth, M.K., (1994) The weighted majority algorithm, *Inform. and Comput.*, **108**, 212–261.

[6] Rissanen, J. (1986) Statistical complexity and modeling. *Annals. of Statist.*, **14**, 1080–1100.

[7] Rogers, H. (1967) Theory of recursive functions and effective computability, New York: McGraw Hill.

[8] 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.

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

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

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

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

[13] Yamanishi, K. (1995) Randomized approximate aggregating strategies and their applications to prediction and discrimination, in *Proceedings, 8th Annual ACM Conference on Computational Learning Theory*, 83–90, Assoc. Comput. Mach., New York.