Electronic Colloquium on Computational Complexity, Report No. 43 (2001)

ECCC

Predictive complexity and information

Michael V. Vyugin * Vladimir V. V'yugin †

April 26, 2001

Abstract

A new notion of predictive complexity and corresponding amount of information are considered. Predictive complexity is a generalization of Kolmogorov complexity which bounds the ability of any algorithm to predict elements of a sequence of outcomes. We consider predictive complexity for a wide class of bounded loss functions which are generalization of square-loss function. Relations between unconditional $KG(x)$ and conditional $KG(x|y)$ predictive complexities are studied. We define an algorithm which has some "expanding property". It transforms with positive probability sequences of given predictive complexity into sequences of essentially bigger predictive complexity. A concept of amount of predictive information $IG(y:x)$ is studied. We show that this information is non-commutative in a very strong sense and present asymptotic relations between values $IG(y:x)$, $IG(x:y)$, $KG(x)$ and $KG(y)$.

*Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, England, E-mail: misha@cs.rhul.ac.uk

†Institute for Information Transmission Problems, Russian Academy of Sciences, Bol'shoi Karetnyi per. 19, Moscow GSP-4, 101447, Russia, and Computer Learning Research Centre, Royal Holloway, University of London, Egham, Surrey TW20 0EX, England, E-mail: vld@vyugin.mccme.ru

ISSN 1433-8092

1 Introduction

A central problem considered in machine learning (and statistics) is the problem of predicting future event x_n based on past observations $x_1x_2\ldots x_{n-1}$, where $n = 1, 2\ldots$. 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.

The main goal of prediction is to find a method of prediction which minimizes the total loss suffered on a sequence $x = x_1x_2\ldots x_n$ for $n = 1, 2\ldots$. This "minimal" possible total loss of prediction was formalized by Vovk [9] in a notion of predictive complexity. This complexity is a generalization of the notion of Kolmogorov complexity and gives a lower bound to ability of any algorithm to predict elements of a sequence of outcomes.

Predictive complexity $KG(x)$ represents the minimal possible total loss of prediction of a sequence x on-line. It depends on a way in which this loss was suffered. This way is represented by a specific loss function. Various loss functions $\lambda(\sigma, p)$ are considered in literature on machine learning and prediction with expert advice (see, for example, [7], [1], [9], [12], [2]). The most important of them are logarithmic loss function and square-loss function. Logarithmic loss function is close to Kolmogorov complexity [4]. Square-loss function is important to applications, corresponding predictive complexity gives a lower limit to the ability of the method of the least squares.

Overview of our results. In this paper we present results for a class of bounded loss functions which are generalization of the squared difference. We study relations between unconditional and conditional predictive complexities. In particular, a variant of triangle inequality is proved (Proposition 4). We prove that in some cases $KG(x)$ can be essentially bigger than $KG(x|y)$ and $KG(y)$. We define an algorithm Φ which has some "expanding property": this algorithm transforms with positive probability a sequence x of given predictive complexity k into a sequence $\Phi(x)$ with predictive complexity $KG(\Phi(x)) \geq ck \log \frac{n}{k}$, where n is the length of x, c is a constant (Theorem 1). This is impossible for Kolmogorov complexity $K(x)$, since for any computable mapping Φ it holds $K(\Phi(x)) \leq K(x) + O(1)$. We summarize our results in a limit form in Theorems 2 and 3. We also study a concept of amount of predictive information

$$IG(y : x) = KG(x) - KG(x|y).$$

We explore relations between four important values $IG(y : x)$, $IG(x : y)$,

$KG(x)$ and $KG(y)$ in a limit form (Theorems 5, 6, 7). In particular, we prove that $IG(y:x)$ is non-commutative in a very strong sense.

Organisation of the paper. A definition of predictive complexity is given in Section 2. We study the histogram of predictive complexity and relation between Kolmogorov complexity and predictive complexity of square-loss type. In Section 3 we consider a triangle inequality for predictive complexity and show that it cannot be improved. We obtain also relations representing the deviation of the value of $KG(x)$ from $KG(y)$, $KG(x|y)$ and $KG([y,x])$. Most of them are followed from the existence of an algorithm with expanding property. In Section 4 relations for predictive information are considered. Section 6 is rather technical. Here we give detailed proofs of the main theorems, in particular, Subsection 6.4 contains the basic construction of this paper: an algorithm with expanding property.

2 Predictive complexity

We consider only simplest case, where events are simple binary outcomes from $\{0,1\}$. Let us denote $\Xi_n = \{0,1\}^n$ and $\Xi = \cup_{n=0}^{\infty}\Xi_n$. We denote by $l(x)$ the length of a finite sequence $x \in \Xi$, $x^n = x_1 \dots x_n$ is its initial prefix of length n. By Λ we denote the empty sequence from Ξ. It is important to consider the case when some *a priori* information is used in performing predictions. We consider a set of *signals* Δ. For simplicity we will consider the case $\Delta = \Xi$.

It is natural to suppose that all predictions are given according to a *prediction strategy* (or *prediction algorithm*) S. When performing prediction p_i the strategy S uses two input sequences, a sequence $x^{i-1} = x_1, x_2, \dots x_{i-1}$ of previous outcomes, and a sequence $y^i = y_1, y_2, \dots, y_i$ of signals, i.e $p_i = S(x^{i-1}, y^i)$ and

$$Loss_S(x^n|y^n) = \sum_{i=1}^{n} \lambda(x_i, S(x^{i-1}, y^i)).$$

The value $y^n = 0^n$, $n = 0, 1, \dots$, corresponds to the case when *a priori* information does not used (here by 0^n we denote the sequence of n zeros).

The value $Loss_S(x^n|y^n)$ can be interpreted as predictive complexity of x^n given y^n. This value, however, depends on S and it is unclear which S to choose. Levin [11], 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 [8]

extended these ideas in a more general setting for a class of mixable loss functions.

Let us fix $\eta > 0$ (the *learning rate*) and put $\beta = e^{-\eta} \in (0,1)$. Let c_η be the infinum of all c such that for each simple probability distribution $P(\gamma)$ on $[0,1]$ (i.e. having a finite domain) there exists a prediction $\hat{\gamma}$ such that

$$\lambda(\sigma, \hat{\gamma}) \le c \log_\beta \sum_\gamma \beta^{\lambda(\sigma,\gamma)} P(\gamma) \tag{1}$$

for all σ. If $c_\eta = 1$ then the corresponding loss function (game) is called η-mixable. We can take $0 < \eta \le 1$ in the case of log-loss function, where $\lambda(1,p) = -\log p$ and $\lambda(0,p) = -\log(1-p)$, and $0 < \eta \le 2$ in the case of square difference $\lambda(\sigma,\gamma) = (\sigma - \gamma)^2$ (see [7]).

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

- (i) $KG(\Lambda|\Lambda) = 0$ and for every x, y of equal length and each extension γ of y there exists a prediction p depending on x and $y\gamma$ such that inequality
$$KG(x\sigma|y\gamma) \ge KG(x|y) + \lambda(\sigma, p) \tag{2}$$
holds for each σ.

- (ii) KG is *semicomputable from above*, which means that there exists a computable sequence of simple functions KG^t taking rational values and such that for every x and y it holds $KG(x|y) = \inf_t KG^t(x|y)$.

By a simple function we mean a function which takes rational values or $+\infty$ and equals $+\infty$ for almost all $x \in \Xi$.

Requirement (i) means that the measure of predictive complexity must be valid: there exists a prediction strategy that achieves it. (Notice that if $\ge$ is replaced by $=$ in (2), the definition of a total loss function will be obtained.) Requirement (ii) means that $KG(x|y)$ must be "computable in the limit".

The main advantage of such definition is that a semicomputable from above sequence $KG_i(x|y)$ of all measures of predictive complexity can be constructed. More precisely, there exists a computable from i, t, x, y sequence $KG_i^t(x|y)$ of simple functions such that

- (iii) $KG_i^{t+1}(x|y) \le KG_i^t(x|y)$ for all i, t, x;

- (iv) $KG_i(x|y) = \inf_t KG_i^t(x|y)$ for all i, x;

- (v) for each measure of predictive complexity $KG(x|y)$ there exists an i such that $KG(x|y) = KG_i(x|y)$ for all x and y.

We will use Kolmogorov prefix complexity which is a modification of the plain Kolmogorov complexity (see e.g. Li and Vitanyi [5], Section 3).

Proposition 1 *Let a loss function $\lambda(\omega, p)$ be computable and η-mixable for some $\eta > 0$. Then there exists a measure of predictive complexity $KG(x)$ such that for each measure of predictive complexity $KG_i(x|y)$*

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

holds for all x, y, where $K(i)$ is the Kolmogorov prefix complexity of the program i enumerating KG_i from above.

The proof this proposition is based on Vovk's aggregating algorithm [9] (see Section 6.1).

Let some η-mixable loss function $\lambda(\sigma, p)$ be given. Put

$$b = \inf_p \sup_\sigma \lambda(\sigma, p). \tag{4}$$

We suppose that $b > 0$. We suppose also that the loss function is computable, and hence, it is continuous by p in the interval $[0, 1]$. Then the infimum in (4) is attained for some computable real number $\hat{p}$. For log-loss function $b = 1$, $\hat{p} = 1$ and $b = \frac{1}{4}$, $\hat{p} = \frac{1}{4}$ in the case of squared difference.

We impose a very natural condition

$$\lambda(0, 0) = \lambda(1, 1) = 0, \tag{5}$$

which holds in both cases of log-loss and square-loss functions. We also consider additional requirements on loss function by which square-loss function differs from log-loss function:

$$\lambda(0, 1) = \lambda(1, 0) = a. \tag{6}$$

Now, when restrictions on a loss function were specified let us fix some $KG(x|y)$ satisfying conditions of Proposition 1 and call its value the conditional *predictive complexity* of x given y. In the case when y is trivial, i.e. consists only from zeros we consider (unconditional) predictive complexity $KG(x)$ of a sequence x.

A very natural problem arises: to estimate the cardinality of all sequences of predictive complexity less than k. A trivial property of Kolmogorov complexity and predictive complexity for log-loss function is that the cardinality of all binary sequences x of complexity $\leq k$ is $\geq 2^{k-c}$ and $\leq 2^k$ for some positive constant c. In the case of predictive complexity of non-logarithmic type the cardinality of the set of all sequences of bounded complexity is infinite. We can estimate the number of sequences of length n having predictive complexity $\leq k$. We denote by $\#A$ the cardinality of a finite set A. Let us consider a set

$$A_{n,k} = \{y | l(y) = n, KG(y) \leq k\}. \tag{7}$$

Proposition 2 *There exists a constant c such that for all n and k*

$$\sum_{i \leq (k-c)/a} \binom{n}{i} \leq \#A_{n,k} \leq \sum_{i \leq k/b} \binom{n}{i}. \tag{8}$$

Proof. Let a sequence x of length n has no more than m ones. Consider prediction strategy $S(z) = 0$ for all z. Then by (5) and (6) there are at least $\sum\limits_{i \leq m} \binom{n}{i}$ of x such that $KG(x) \leq Loss_S(x) + c \leq am + c \leq k$, where c is a constant. Then $m \leq (k - c)/a$ and we obtain the left-hand side of the inequality (8).

To prove the right-hand side of the inequality (8) consider the *universal* prediction strategy $\Lambda(x) = p$, where $p = p(x)$ is the prediction from the item (i) of definition of the measure of predictive complexity. By definition $Loss_\Lambda(x) \leq KG(x)$ for each x. By (4) for any x we have $\lambda(0, \Lambda(x)) \geq b$ or $\lambda(1, \Lambda(x)) \geq b$. By this property we assign new labelling to edges of the binary tree using letters A and B. We assign A to $(x, x0)$ and B to $(x, x1)$ if $\lambda(0, \Lambda(x)) \geq b$, and assign B to $(x, x0)$ and A to $(x, x1)$ otherwise. Evidently, two different sequences of length n have different labellings. For each edge $(x, x\sigma)$ labeled by A it holds $\lambda(\sigma, \Lambda(x)) \geq b$ and, hence, for any sequence x of length n having more than m As it holds $KG(x) \geq Loss_\Lambda(x) \geq bm$. Therefore, $\#A_{n,k} \leq \sum\limits_{i \leq k/b} \binom{n}{i}$. $\square$

A more strong upper bound of the cardinality of the set $A_{n,k}$ is given in [10]. Let $K(x)$ be the Kolmogorov (prefix) complexity of x.

Proposition 3 *A positive constant c exists such that for all n and $k \leq \frac{bn}{2}$ for all x of the length n such that $KG(x) \leq k$ the inequality*

$$K(x) \leq (\frac{k}{b} + 2) \log n - \frac{k}{b} \log \frac{k}{4b} + c \tag{9}$$

holds.

Sketch of the proof. Let us consider the recursively enumerable set $A_{n,k}$ defined by (7) above. We can specify any $x \in A_{n,k}$ by n, k and the ordinal number of x in the natural enumeration of $A_{n,k}$, i.e. $K(x) \leq \log \#A_{n,k} + 2\log n + 2\log k + c$ for some constant c. The needed upper bound now follows from (8). For details see Section 6.3. □

3 Triangle inequality

The following Proposition 4 is an analogue of the corresponding result for the prefix Kolmogorov complexity $K(x) \leq K(x|y) + K(y) + O(1)$ [5].

Proposition 4 *Positive constants c_1 and c_2 exist such that for all x and y of length n*

$$KG(x) \leq KG(x|y) + (\ln 2/\eta)K(y) + c_1 \leq \tag{10}$$

$$KG(x|y) + c_2 KG(y) \log\left(\frac{n}{KG(y)}\right). \tag{11}$$

Inequality (11) holds if $KG(y) \leq bn/2$.

Proof. This proposition is a direct corollary of Proposition 1. For any finite sequence y define $\bar{y} = y0^\infty$. The measure of predictive complexity $S(x) = KG(x|\bar{y}^{l(x)})$ can be enumerated from below by a program depending from y. Then by (3) for any x such that $l(x) = l(y)$

$$KG(x) \leq KG(x|y) + (\ln 2/\eta)K(y) + c_1,$$

where c_1 is a positive constant. Applying the upper bound on $K(y)$ from Proposition 3 we obtain the needed result. □

The following theorem shows that inequality (11) cannot be improved. We construct a computable mapping having some "expanding property": it transforms sequences of given predictive complexity into sequences of essentially bigger predictive complexity.

Let $C_{n,k}$ be a set of sequences y of the length n having k changes from 0 to 1 or from 1 to 0 (occurences of combinations 01 and 10 in y); it is also convenient to consider a case $y_1 = 1$ (i.e. a case when y starts from 1) as a change.

Let us consider a computable predictive strategy S such that $S(\Lambda) = 0$ and $S(x1) = 1$, $S(x0) = 0$ for all x. By (6) we have $Loss_S(y) \le a(k+1)$ for each $y \in C_{n,k}$. Therefore, $KG(y) \le ak + O(1)$ for each $y \in C_{n,k}$. Let $\lceil r \rceil$ be the least integer number s such that $s \ge r$.

Theorem 1 *For any n and $k \le \frac{n}{2}$ a computable mapping Φ from $C_{n,k}$ to a subtree $\{x | l(x) = n, 0^{n-m} \subseteq x\}$, where $m = \lceil \log 4^k \binom{n}{k} \rceil$, exists such that for a portion $\ge 1/2$ of all $y \in C_{n,k}$ the output $x = \Phi(y)$ satisfies*

$$KG(x) \ge \frac{b}{20} \log \binom{n}{k}, \tag{12}$$

$$KG(x|y) = O(\log n). \tag{13}$$

We have also $KG(y) \le ak + O(1)$ for each $y \in C_{n,k}$. In the case $k = k(n) = o(n)$ the factor $b/20$ in (12) can be replaced on $b/2$.

Sketch of the proof. We construct a mapping Φ which compresses the set $C_{n,k}$ into the set $\{x | l(x) = n, 0^{n-m} \subseteq x\}$, where $m = \lceil \log 4^k \binom{n}{k} \rceil$, such that the density of of the image of Φ in this set is sufficiently large, namely, 4^{-k}. We prove also a variant of "incompressibility lemma", from which follows that the large portion of this image consists of elements x of predictive complexity $KG(x)$ satisfying (12) (see Section 6.4 for a detailed construction). □

In the following theorems we present results of Proposition 4 and Theorem 1 in an asymptotic form. For any functions $\alpha(n)$ and $\beta(n)$ the expression $\alpha(n) = \Theta(\beta(n))$ means that there exist constants c_1 and c_2 such that $c_1\beta(n) \le \alpha(n) \le c_2\beta(n)$ holds for all n. The expression $\alpha(n) = \Omega(\beta(n))$ means that there exists a constant c_1 such that $\alpha(n) \ge c_1\beta(n)$ for all n.

Theorem 2 *Let a function $k(n)$ be unbounded, $\nu(n) = \Omega(\log n)$, and $k(n) = O(n)$, $\nu(n) = O(n)$. Then*

$$\sup_{x : \exists y (KG(y) \le k(n), KG(x|y) \le \nu(n))} KG(x) = \Theta\left(\nu(n) + k(n) \log\left(\frac{n}{k(n)}\right)\right) \tag{14}$$

Proof. The part $\le$ follows from Proposition 4. In case $\nu(n) \ge k(n) \log\left(\frac{n}{k(n)}\right)$ the part $\ge$ of (14) is evident. Otherwise, the statement follows from Theorem 1. □

For any $x = x_1 \dots x_n$ and $y = y_1 \dots y_n$ we consider "a pair" $[y, x] = y_1 x_1 \dots y_n x_n$. The following theorem defines some limits of that predictive complexity of a pair can be less than complexity of their elements.

Theorem 3 *Let* $k(n) = \Omega(\log n)$, $\nu(n) = \Omega(k(n))$, $k(n) = O(n)$ *and* $\nu(n) = O(n)$. *Then*

$$\sup_{x:\exists y(KG(y)\le k(n), KG([y,x])\le \nu(n))} KG(x) = \Theta\left(\nu(n) + k(n)\log\left(\frac{n}{k(n)}\right)\right). \quad (15)$$

Proof. In case $\nu(n) \ge k(n)\log\left(\frac{n}{k(n)}\right)$ the part $\ge$ of (15) is evident. Otherwise, we use Theorem 1 and its proof. Let us consider the mapping Φ from the proof of Theorem 1 (see Section 6.4) and and the sequences y and $x = \Phi(y)$ satisfying the conditions of this proposition. Define a computable prediction strategy S

$$S(y_1x_1\ldots y_{i-1}x_{i-1}y_i) = \Phi(y^i)_i = x_i,$$
$$S(y_1x_1\ldots y_ix_i) = 0.$$

Then by (6) we have $Loss_S([y,x]) \le ak(n)$ and, therefore,

$$KG([y,x]) \le ak(n) + (\ln 2/\eta)K(S) \le ak(n) + (2\ln 2/\eta)\log n \le c_1\nu(n) \quad (16)$$

for some constant c_1. By Theorem 1 $KG(y) \le c_2\log n \le c_3k(n)$ for some constants c_2, c_3. The part $\ge$ of (15) follows from these inequalities after normalizing of $k(n)$ and $\nu(n)$.

The $\le$ part of (15) follows from Proposition 4 and an obvious inequality

$$KG(x|y) \le KG([y,x]) + c,$$

where c is a positive constant. □

4 Predictive information

The *amount of predictive of information* in a sequence y about a sequence x of the same length in the process of on-line prediction was defined by Vovk [9]

$$IG(y:x) = KG(x) - KG(x|y). \quad (17)$$

In this section we explore relations between four important values $IG(y:x)$, $IG(x:y)$, $KG(x)$ and $KG(y)$ in a limit form. These results are mainly based on the construction of Theorem 1.

Predictive information is non-commutative in a very strong sense. Define

$$g_1(n) = \sup_{l(x)=l(y)=n} \frac{IG(x:y)}{IG(y:x)}. \quad (18)$$

Theorem 4 *It holds*

$$g_1(n) = \Theta(n). \tag{19}$$

Sketch of the proof. Consider the random string x (eg sequence of coin tosses). Another string y will be equal to x without its first sign. These strings provide needed example. For the exact proof see Section 6.2. □

Let us define

$$g_2(n) = \sup_{l(x)=l(y)=n} \frac{IG(y:x)}{KG(y)}. \tag{20}$$

The main results of Section 3 can be summarized in the following theorem.

Theorem 5 *It holds* $g_2(n) = \Theta(\log n)$.

Proof. The right-hand inequality follows directly from (11). The left-hand inequality can be derived from Theorem 1. Its enough to let $k = \sqrt{n}$. □

Let us define also

$$g_3(n) = \sup_{l(x)=l(y)=n} \frac{IG(y:x)}{KG(x)}. \tag{21}$$

Theorem 6 *It holds*

$$\lim_{n\to\infty} g_3(n) = 1. \tag{22}$$

Proof. The part (22) follows from Theorem 1, where we let $k = n^{1/2}$. □

Theorem 7 *Let $k(n) \leq bn$ for all n. Then a constant c exists such that*

$$\sup_{(x,y):l(x)=l(y)=n, KG(y)\leq k(n)} IG(y:x) = \Theta\left(k(n)\log\left(\frac{n}{k(n)}\right)\right) \tag{23}$$

$$\sup_{(x,y):l(x)=l(y)=n, KG(x)\leq k(n)} IG(y:x) = k(n) + O(1) \tag{24}$$

$$\sup_{(x,y):l(x)=l(y)=n, IG(x:y)\leq c} IG(y:x) = \Theta(n) \tag{25}$$

Proof. The right-hand part of (23) follows from Proposition 4. The left-hand part of (23) follows from Theorem 1. To prove (24) put $x = y$ and note that $KG(x|x) = O(1)$. Relation (25) follows from the proof of Theorem 4. □

5 Acknowledgements

Authors are grateful to Volodya Vovk for useful discussions and for his suggestions about formulating of main results of the paper.

6 Appendix

6.1 Proof of Proposition 1

A sequence $KG_i(x|y)$ of all measures of predictive complexity can be defined using standard methods of the theory of algorithms.

Let r_i be a semicomputable from below sequence of real numbers such that the series $\sum_{i=1}^{\infty} r_i$ is convergent and its sum does not exceed 1. We can take $r_i = 2^{-K(i)}$. Analogously to [8] and [9] a measure of predictive complexity $KG(x|y)$ can be defined

$$KG(x|y) = \log_\beta \sum_{i=1}^{\infty} \beta^{KG_i(x|y)} r_i, \tag{26}$$

where $\beta = e^{-\eta}$. By definition $KG(x|y)$ is semicomputable from above, i.e (ii) holds. We must verify (i). Indeed, by (26) for every x, y of equal length and $\sigma, \beta \in \{0,1\}$

$$KG(x\sigma|y\beta) - KG(x|y) = \log_\beta \sum_{i=1}^{\infty} q_i \beta^{KG_i(x\sigma|y\beta) - KG_i(x|y)} \geq \tag{27}$$

$$\log_\beta \sum_{i=1}^{\infty} q_i \beta^{\lambda(\sigma,\gamma_i)} \geq \lambda(\sigma,\gamma), \tag{28}$$

where

$$q_i = \frac{r_i \beta^{KG_i(x|y)}}{\sum_{s=1}^{\infty} r_s \beta^{KG_s(x|y)}}.$$

Here for any i a prediction $\gamma_i = \gamma(x, y\beta)$ satisfying

$$KG_i(x\sigma|y\beta) - KG_i(x|y) \geq \lambda(\sigma,\gamma_i)$$

exists since each element of the sequence $KG_i(x|y)$ satisfies the condition (i) of the measure of predictive complexity. A prediction γ satisfying (28) exists by definition of the constant c_η in Section 2. For further details see [9], Section 7.6.

6.2 Proof of Theorem 4

For any sequence $x = x_1 \dots x_n$ define its left shift $Tx = x_2 \dots x_n 0$. The proof of the theorem is based on the following simple lemma.

Lemma 1 *It holds*

$$KG(x|Tx) = O(1), \tag{29}$$

$$KG(Tx) = KG(Tx|x) + O(1) \tag{30}$$

The proof of this lemma is given in Section 6.5.

To prove the right-hand side of (19) take $y = Tx$. Then by Lemma 1 a positive constant c exists such that

$$IG(Tx : x) = KG(x) - KG(x|Tx) \geq KG(x) - c$$
$$IG(x : Tx) = KG(Tx) - KG(Tx|x) \leq c$$

for all x. Using the definition (4) of b and the diagonal argument it easy to construct a sequence x of length n such that $KG(x) \geq bn$. From this the right-hand side of (19) follows. The left-hand side of (19) follows from the inequality

$$KG(x) \leq bl(x) + c$$

for all x, where c is a positive constant. This inequality can be easily proved using a computable prediction strategy which always predicts $\hat{p}$, where $\hat{p}$ minimizes the condition (4). $\Box$.

6.3 Proof of Proposition 3

To prove inequality (9) let us consider the recursively enumerable set (7). We can specify any $x \in A_{n,k}$ by n, k and the ordinal number of x in the natural enumeration of $A_{n,k}$. Let $k \leq bn/2$. Then using an appropriate encoding of all triples of positive integer numbers we obtain for all x of the length n

$$K(x) \leq \log \#A_{n,k} + 2\log n + 2\log k + c \leq \tag{31}$$

$$\log \frac{k}{b}\binom{n}{k/b} + 2\log n + 2\log k + c \leq \tag{32}$$

$$\log k \left(\frac{en}{k/b}\right)^{k/b} + 2\log n + 2\log k + c' \leq \tag{33}$$

$$\left(\frac{k}{b}+2\right)\log n-\left(1-\frac{2}{\log(k/b)}\right)\frac{k}{b}\log\frac{k}{b}+c''= \tag{34}$$

$$\left(\frac{k}{b}+2\right)\log n-\frac{k}{b}\log\frac{k}{4b}+c''. \tag{35}$$

where c, c' and c'' are positive constants.

6.4 Proof of Theorem 1

For any n and any sequence z of the length $\le n$ by a subtree $\Xi_n(z)$ we mean a set

$$\Xi_n(z)=\{y|l(y)=n, z\subseteq y\}.$$

The height of the subtree is the number $m=n-l(z)$.

The main technical result of this paper is a construction of a mapping Φ from $C_{n,k}$ into a subtree $\Xi_n(0^l)$ of height $m=\lceil\log 4^k\binom{n}{k}\rceil$, where $l=n-m$.

Let $y*y'$ denotes the maximal joint prefix of y and y', i.e. a sequence z of maximal length such that $z\subseteq y$ and $z\subseteq y'$. The main requirement to Φ is that

$$l(\Phi(y)*\Phi(y'))\ge l(y*y') \tag{36}$$

for all $y,y'\in C_{n,k}$.

We construct the mapping Φ as a result of a recursive procedure COMP(n,k,σ), where n and k be positive integer numbers and $\sigma=0$ or 1.

Procedure COMP(n,k,σ).

For any n the procedure COMP$(n,0,0)$ returns the identical mapping Φ on the set $\{0^n\}$, the procedures COMP$(n,1,1)$ and COMP$(n,0,1)$ return the identical mapping Φ on the set $\{1^n\}$.

Let $C^\sigma_{n,k}$ be the set of all sequences from $C_{n,k}$ starting from σ ($\sigma=0,1$).

If $k>0$ the procedure returns a mapping Φ from $C^\sigma_{n,k}$ into a subtree $\Xi_n(\sigma^l)$ for some l. Without loss of generality we suppose that $\sigma=0$.

In the following we will construct this mapping Φ by series of subsequent reassignments of the values of Φ. We start with the mapping Φ identical on $C^0_{n,k}$. We will do also a series of subsequent transformations of subtrees of the initial tree Ξ_n.

Consider the *base* which is a sequence of n zeros 0^n (in the case $\sigma=1$ the base is a sequence of n ones 1^n). There are $n-1$ subtrees $\Xi_n(0^i1)$ on the base, $i=1,\dots,n-1$. Each sequence $x\in C^0_{n,k}$ of length n belongs to one of such trees $\Xi_n(0^i1)$. We call them *basic subtrees*.

For each $i = 1, \ldots, n-1$ we apply the procedure COMP$(n-i, k-1, 1)$. For each i the procedure returns a subtree $\Xi_{n-i}(1^{k_i})$ and a mapping Φ_i from $C^1_{n-i,k-1}$ into this subtree.

Induction hypothesis: for each i the set $\Phi_i(C^1_{n-i,k-1}) \cap \Xi_{n-i}(1^{k_i})$ occupies at least $4^{-(k-1)}$ portion of all sequences of basic subtree $\Xi_{n-i}(1^{k_i})$. In other words,

$$\#\Phi_i(C^1_{n-i,k-1} \cap \Xi_{n-i}(1^{k_i})) \geq 4^{-(k-1)} \#\Xi_{n-i}(1^{k_i}).$$

There is a natural one-to-one correspondence between sequences $0^i 1x \in C^0_{n,k}$ and sequences $x \in C^1_{n-i,k-1}$. Then we can redefine Φ on $C^0_{n,k}$ such that for each i and for each $1x \in C^1_{n-i,k-1}$ it holds $\Phi(0^i 1x) = 0^i \Phi_i(1x)$.

Denote $A_i = \Xi_n(0^i 1^{k_i})$ and call it the compressed basic subtree. By the induction hypothesis we obtain that $\Phi(C^0_{n,k}) \cap (\cup_i A_i)$ occupies at least $4^{-(k-1)}$ portion of all sequences from $\cup_i A_i$.

At the current step of induction we will change this assignment of Φ according instructions given below.

Each initial basic subtree $\Xi_n(0^i 1)$, $1 \leq i \leq n-1$ contains (before applying the procedure COMP) a subtree $T = \Xi_n(0^i 11)$ isomorphic to its upper neighbour $\Xi_n(0^{i+1}1)$. This subtree T will be transformed by the procedure COMP as a part of $\Xi_n(0^i 1)$ on the previous inductive step into a subtree T' of corresponding resulting basic subtree $A_i = \Xi_n(0^i 1^{k_i})$. The height of A_i is equal to $h_i = n - i - k_i$.

For each i if $h_i < h_{i+1}$ then replace A_{i+1} on the corresponding subtree T' of A_i. Doing these replacements we simultaneously changing corresponding assignments of the values of Φ. This process provides us that $h_i \geq h_{i+1}$ for all i.

Now we will transform all compressed basic subtrees A_i into a resulting compressed subtree $\Xi_n(0^l)$ and will change the corresponding assignments of Φ such that the density of the image of Φ in the resulting subtree $\Xi_n(0^l)$ will be at least 4^{-k}. The reconstruction consists of subsequent application of the following two operations on subtrees of a binary tree.

1) *Joining to the nearest upper neighbour.* Suppose that there are at least two basic subtrees of type $\Xi_n(0^i 1^{3+s_i})$ and $\Xi_n(0^{i+1}1^{2+s_i})$ of the same height h, where $s_i \geq 0$. We transform each assignment of type $\Phi(y) = 0^i 1^{3+s_i} u$ to $\Phi(y) = 0^{i+1}1^{1+s_i}0u$, i.e. move subtree $\Xi_n(0^i 1^{3+s_i})$ to $\Xi_n(0^{i+1}1^{1+s_i}0)$. We will apply this transformation only to the first (left) pair of the basic subtrees of height h. Then the height of the resulting subtree will be $h+1$ and will not exceed the height of the previous basic subtree.

2) *Moving up to the nearest upper free vertex on the base.* Suppose that there is a basic subtree $\Xi_n(0^i1^2)$ and there is no basic subtree of type $\Xi_n(0^{i+1}1)$. In this case we change each assignment $\Phi(y) = 0^i1^2u$ to $\Phi(y) = 0^{i+1}1u$, i.e. we move basic subtree $\Xi_n(0^i1^2u)$ to $\Xi_n(0^{i+1}1u)$.

By definition, after these transformations, the image of Φ still occupies at least $4^{-(k-1)}$ portion of all sequences from each transformed basic subtree.

Given initial compressed basic subtrees we transform them by applying these two operations as many times as possible. Suppose that among final basic subtrees there is a basic subtree of type $\Xi(0^i1^{j+3})$, where $j \geq 0$. Then there are only two possibilities. The first one is that there is no basic subtree of type $\Xi_n(0^{i+1}1^s)$, where $s \geq 1$, and then the operation 2) must be applied. The second one is that there is a subtree $\Xi_n(0^{i+1}1^{2+s_i})$ (and there is no subtree $\Xi_n(0^{i+1}10)$). In this case the operation 1) must be applied. No basic subtree of type $\Xi_n(0^{i+1}1)$ exists, since otherwise the property $h_i \geq h_{i+1}$ for heights of basic subtrees will be failed. The contradiction obtained shows that there is no basic subtree of type $\Xi(0^i1^{j+3})$ after all transformations.

As a result we obtain the base 0^n and a sequence of the new basic subtrees on it of the type $\Xi_n(0^i1^j)$, where $i \geq 1$ and $j = 1$ or $j = 2$. We obtained also a new assignment to values of Φ.

Consider the first basic subtree B of maximal height. Since, the operations 1) and 2) cannot be applied to B, it is of the form $\Xi_n(0^l1)$ or $\Xi_n(0^l11)$, where $l \geq 1$. Then the portion of sequences of B among all sequences of $\Xi_n(0^l)$ is at least $\frac{1}{4}$. Therefore, the portion of sequences in $\Xi_n(0^l)$ which are of a form $\Phi(y)$, where $y \in C^0_{n,k}$, can decrease no more than in 4 times, and so it is $\geq \frac{1}{4}4^{-(k-1)} = 4^{-k}$.

We declare the mapping Φ and the subtree $\Xi_n(0^l)$ as the results returning by procedure COMP$(n, k, 0)$. We also proved that the induction hypothesis holds for these results.

End of the procedure COMP.

Let Φ and $\Xi_n(0^l)$ be outputs of COMP$(n, k, 0)$. Since the image $\Phi(C_{n,k})$ contains $\binom{n}{k}$ sequences in the binary tree $\Xi_n(0^l)$ by density condition this tree have $\leq 4^k\binom{n}{k}$ sequences.

The following lemma will be used to prove the existence of elements of big predictive complexity in a set of given cardinality.

Lemma 2 *For any n let $m < n$ and z be a sequence of the length $n - m$. Then for any set $W \subseteq \{x|l(x) = n, z \subseteq x\}$ for at least $1/2$ portion of all*

$x \in W$ it holds

$$KG(x) \geq bl_{max},$$

where l_{max} is the maximal integer number l such that $l \leq \frac{m}{2}$ and

$$H\left(\frac{l}{m}\right) \leq \frac{\log \#W}{m} - \frac{\log m}{m} - \frac{1}{m}, \tag{37}$$

where $H(p) = -p\log p - (1-p)\log(1-p)$ is the Shannon entropy.

Proof. Let $\Lambda(x)$ be the universal predictive strategy as in the proof of Proposition 2 such that $Loss_\Lambda(x) \leq KG(x)$ for all x. By definition of b (see (4)) for any x we have $\lambda(0, \Lambda(x)) \geq b$ or $\lambda(1, \Lambda(x)) \geq b$. Using this property we assign the labelling to edges using letters A and B as in the proof of Proposition 2. Then for any sequence x of the length m having more than k As it holds $Loss_\Lambda(x) \geq bk$.

Now, to estimate from below the maximal total loss of Λ on sequences from W we must estimate from below the maximal number of As occurring in sequences from W. This estimate l satisfies inequality

$$\sum_{i=1}^{l} \binom{m}{i} < \frac{1}{2}\#W. \tag{38}$$

The inequality (38) follows from $l\binom{m}{l} < \frac{1}{2}\#W$. Since the elementary inequality

$$\binom{m}{l} \leq 2^{mH(\frac{l}{m})}$$

holds (see [3], Section 6.1), the inequality (38) also follows from

$$l2^{mH(\frac{l}{m})} \leq \frac{1}{2}\#W. \tag{39}$$

This inequality follows from

$$H\left(\frac{l}{m}\right) \leq \frac{\#W}{m} - \frac{\log m}{m} - \frac{1}{m}. \tag{40}$$

Hence, we have $KG(x) \geq bl$ for at least $1/2$ portion of all $x \in W$, where l is the maximal number satisfying (40). □

To apply Lemma 2 get

$$m = \lceil \log 4^k \binom{n}{k} \rceil,$$
$$W = \Phi(C_{n,k}).$$

We have $\#W = \binom{n}{k}$. We must find maximal l such that the inequality

$$H\left(\frac{l}{m}\right) \leq \frac{\log \binom{n}{k}}{2k + \log \binom{n}{k}} - \frac{\log m}{m} - \frac{1}{m} \tag{41}$$

holds. It holds $\binom{n}{k} \geq 2^k$ if $k \leq \frac{n}{2}$. Then since $\log \binom{n}{k} \geq k$, the first term of (41) is bigger than $\frac{1}{3}$. Hence, for m sufficiently large it is sufficient to find l such that

$$H\left(\frac{l}{m}\right) \leq \frac{3}{10}. \tag{42}$$

It is easy to verify by table values of Shannon entropy that the inequality (42) wittingly holds if

$$\frac{l}{m} \leq \frac{1}{20}$$

and so, we can take estimate

$$l_{max} = \frac{m}{20}.$$

By Lemma 2 an x of the length n exists such that

$$KG(x) \geq \frac{b}{20}\left(2k + \log \binom{n}{k}\right) \geq \frac{b}{20} \log \binom{n}{k}.$$

Let $\Phi(y) = x$. By the prefix property (36) of mapping Φ each prefix x^i of length i can be recovered from a prefix y^i of length i. Hence, a prediction strategy S exists which computes the i-th member of x given $y^i = y_1 \ldots y_i$. By definition we have $Loss_S(x|y) = 0$. This predictive strategy S is trivially defined by the mapping Φ given n and k as parameters. Hence, by (3) we have $KG(x|y) \leq (2 \ln 2/\eta) \log n$ for all sufficiently large n. The proof of the proposition is completed.

6.5 Proof of Lemma 1

To prove (29) consider a computable prediction strategy S such that for any sequences x and y of length $n \geq 2$ it holds $S(x|y) = y_{n-1}$ (for $n = 1$ define $S(x|y) = 0$). Then $Loss_S(x|Tx) \leq b$ for each x, and by definition $KG(x|Tx) \leq b + c$ hold for all x, where c is a positive constant c.

Let us prove (30). $KG(Tx|x) \leq KG(Tx) + O(1)$ by definition. To prove the converse inequality define a function $S(u)$ using an idea of Vovk's [7] aggregating algorithm (see Section 6.1).

$$S(u) = \log_\beta(2^{-1}\beta^{KG(u|0u)} + 2^{-1}\beta^{KG(u|1u)}). \tag{43}$$

We show that $S(u)$ is a measure of predictive complexity. Indeed, for any σ

$$S(u\sigma) - S(u) = \log_\beta \sum_{i=0}^{1} 2^{-1}\beta^{KG(u\sigma|iu)} - \log_\beta \sum_{i=0}^{1} 2^{-1}\beta^{KG(u|iu)} =$$

$$\log_\beta \sum_{i=0}^{1} q_i \beta^{KG(u\sigma|iu)-KG(u|iu)} \geq \log_\beta \sum_{i=0}^{1} 2^{-1}\beta^{\lambda(\sigma,\hat{p}(iu))} \geq \lambda(\sigma, \hat{p}(u)),$$

where

$$q_i = \frac{2^{-1}\beta^{KG(u|iu)}}{\sum_{j=0}^{1} 2^{-1}\beta^{KG(u|ju)}},$$

$i = 0, 1$, and predictions $\hat{p}(iu)$ and $\hat{p}(u)$ exist by η-mixability property of the loss function $\lambda(\sigma, p)$. By definition (43) of $S(u)$ we have for $i = 0, 1$

$$S(u) \leq KG(u|iu) + (\ln 2/\eta)$$

for all u. Hence, by definition

$$KG(Tx) \leq KG(Tx|x) + c$$

for all x, where c is a positive constant.

References

[1] 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.

[2] 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

[3] Cormen, H., Leiserson, E., Rivest, R (1990) *Introduction to Algorithms.* New York: McGraw Hill.

[4] Kalnishkan, Y. (1999) General linear relations among different types of predictive complexity. In. *Proc. 10th international Conference on Algorithmic Learning Theory–ALT '99, v. 1720 of Lecture Notes in Artificial Intelligence*, pp. 323–334, Springer–Verlag.

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

[6] 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.

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

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

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

[10] Vyugin, M., V'yugin, V. (2001) Non-linear inequalities between predictive and Kolmogorov complexity (submitted for publication).

[11] 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.

[12] 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.

ECCC ISSN 1433-8092
http://www.eccc.uni-trier.de/eccc
ftp://ftp.eccc.uni-trier.de/pub/eccc
ftpmail@ftp.eccc.uni-trier.de, subject 'help eccc'