Electronic Colloquium on Computational Complexity, Report No. 52 (2001)

Non-linear Inequalities between Predictive and Kolmogorov Complexity

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

April 26, 2001

Abstract

Predictive complexity is a generalization of Kolmogorov complexity which gives a lower bound to ability of any algorithm to predict elements of a sequence of outcomes. A variety of types of loss functions makes it interesting to study relations between corresponding predictive complexities.

Non-linear inequalities between predictive complexity of non-logarithmic type and Kolmogorov complexity (which is close to predictive complexity for logarithmic loss function) are the main subject of consideration in this paper. We prove that asymptotically they differ on sequences of length n in the worst case by a factor equal to $\log n$. These estimates cannot be improved. To obtain these inequalities we present estimates of the cardinality of all sequences of given predictive complexity.

*Computer Learning Research Centre, 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_i based on past observations $x_1x_2 \ldots x_{i-1}$, where $i = 1, 2 \ldots$. The simplest case is when x_i is equal 0 or 1. A prediction algorithm makes its prediction on-line in a form of a real number p_i between 0 and 1. We suppose that the quality of prediction is measured by a specific loss function $\lambda(x_i, p_i)$. The total loss of prediction suffered on a sequence of events $x_1x_2 \ldots x_n$ is measured by the sum of all values $\lambda(x_i, p_i)$, $i = 1, \ldots, n$.

Various loss functions are considered in literature on machine learning and prediction with expert advice (see, for example, [8], [1], [10], [12], [2]). The most important of them are logarithmic loss function and square-loss function. Logarithmic loss function, $\lambda(\sigma, p) = -\log p$ if $\sigma = 1$ and $\lambda(\sigma, p) = -\log(1-p)$ otherwise, is considered in statistics as log-likelihood function, and the corresponding predictive complexity coincides with a variant of Kolmogorov complexity. Square-loss function $\lambda(\sigma, \gamma) = (\sigma - \gamma)^2$ is important to applications, corresponding predictive complexity gives a lower limit to the quality of regression under square loss.

The main goal of prediction is to find a method of prediction which minimizes the total loss suffered on a sequence $x_1x_2 \ldots x_i$ for $i = 1, 2 \ldots$. This "minimal" possible total loss of prediction was formalized by Vovk [10] 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.

A variety of types of loss functions defines the problem of comparative study of corresponding predictive complexities. Kalnishkan [4] provided necessary and sufficient conditions on constant coefficients a_1, a_2 and b_1, b_2, b_3 under which the inequalities

$$a_1K^1(x) + a_2l(x) + c_1 \geq K^2(x) \quad \text{and} \quad \mathrm{b}_1K^1(x) + b_2K^2(x) \leq b_3l(x) + c_2$$

hold for some additive constants c_1, c_2. Here $K^1(x)$ and $K^2(x)$ are predictive complexities of different types, $l(x)$ is the length of a sequence x. Logarithmic KG^{log} and square-loss KG^{sq} complexities can be among K^1 and K^2, in particular inequality $KG^{sq}(x) \leq \frac{1}{4}KG^{log}(x) + c$ holds for some positive constant c. Converse inequalities with constant coefficients between these complexities which can be obtained by Kalnishkan's method have additive members of order $O(l(x))$. To avoid these addends we explore non-linear inequalities. These inequalities hold up to factor equal to logarithm of the length of sequence x. By its definition below $KG^{log}(x)$ coincides with the minus logarithm of the Levin's [11] "a priori" semimeasure (see also [5]) which is close to Kolmogorov complexity $K(x)$ up to addend $O(\log l(x))$. By this reason and by general fundamental importance of Kolmogorov complexity we compare $KG^{sq}(x)$ with $K(x)$.

To obtain these inequalities we estimate the number of all sequences of length n with given upper bound k on predictive complexity (Proposition 2). We deduce from this combinatorial estimation non-linear inequalities between Kolmogorov complexity and predictive complexity of non-logarithmic type (Propositions 3, 4). More advanced estimates for predictive complexity are given in Theorems 3, 4.

Main results of this paper in an asymptotic form are formulated in Theorems 1 and 2.

2 Predictive complexity

We consider only simplest case, where events $x_1, x_2, \ldots, x_i \ldots$ are simple binary outcomes from $\{0, 1\}$, nevertheless, our results trivially can be extended to the case of arbitrary finite set of all possible outcomes $\{0, 1, \ldots, L-1\}$, where $L > 1$. It is natural to suppose that all predictions are given according to a *prediction strategy* (or *prediction algorithm*) $p_i = S(x_1, x_2, \ldots x_{i-1})$. We will suppose also that our loss functions are computable. The total loss incurred by Predictor who follows the strategy S over the first n trials is defined

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

The main problem is to find a method of prediction S which minimizes the total loss $L_S(x)$ suffered on a sequence x of outcomes. In machine learning theory several "aggregating algorithms" achieving this goal in the case of finite number of experts were developed [7], [1], [12], [2], [8].

In [10] Vovk suggested a solution of this problem in the case of infinite pool of "computationally efficient" experts. He introduced a notion of *predictive complexity*, which is a generalization of the notion of total loss. A function $KG(x)$ is a *measure of predictive complexity* if the following two conditions hold:

- (i) $KG(\Lambda) = 0$ (where Λ is the empty sequence) and for every x there exists a p such that for each σ $KG(x\sigma) \geq KG(x) + \lambda(\sigma, p)$;

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

We fix some (non-computable) predictive strategy $\Lambda(x) = p$ satisfying condition (i) and call it *universal predictive strategy.* By definition $Loss_\Lambda(x) \leq KG(x)$ for all x.

By Kolmogorov prefix complexity $K(S)$ of prediction strategy S we mean the length of the shortest program (under some universal prefix way of encoding) computing its values with given degree of accuracy (for details see [5], Section 3).

Vovk proved that an "optimal" measure of predictive complexity exists under the same conditions that are sufficient to optimal efficiency of his aggregating algorithm. According to Vovk's [8],[10] framework we fix the *learning rate* $\eta > 0$ 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

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

for all σ. If $c_\eta = 1$ then the corresponding loss function is called η-mixable. By [8] $c_\eta = 1$ for any $0 < \eta \leq 1$ in the case of log-loss function, and for any $0 < \eta \leq 2$ in the case of square difference.

Proposition 1 *[10] 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 any measure of predictive complexity $KG'(x)$ a constant c exists such that $KG(x) \leq KG'(x) + c$ for all x, besides this, $KG(x) \leq Loss_S(x) + (\ln 2/\eta)\ K(S) + c$ for each computable prediction strategy S and each x, where c is a constant.*

Let some η-mixable loss function is given. We fix some $KG(x)$ satisfying conditions of Proposition 1 and call its value the *predictive complexity* of x.

We formulate our results for loss functions from a wide class. We impose the following restrictions on a loss function $\lambda(\sigma, p)$: There exists a computable positive real number b such that

- 1) $\lambda(0,p) \geq b$ or $\lambda(1,p) \geq b$ for each p;
- 2) $\lambda(0,0) = \lambda(1,1) = 0$;

- 3) the loss function $\lambda(\sigma, p)$ is η-mixable for some $\eta > 0$.

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

The inequality $KG(x) \le (1+\gamma)(\ln 2/\eta)K(x) + c$ between complexities $KG(x)$ and $K(x)$ can be obtained from Proposition 1, where $\gamma > 0$ be an arbitrary and c is a positive constant depending on γ. To prove it consider prediction strategy S defined by x such that $S(z) = x_i$ for each z of the length $i - 1$, where $1 \le i \le l(x) - 1$, and $S(z) = 0$, otherwise.

We consider the following additional requirements on loss function. By these restrictions the square-loss function differs from log-loss function.

- 4) $\lambda(0, 1) = \lambda(1, 0) = a$.
- 5) There exists a computable monotonically increasing function $\delta(\epsilon)$ such that $\delta(\epsilon) > 0$ if $\epsilon > 0$ and such that for each $0 \le \epsilon \le 1$ and $0 \le p \le 1$ if $\lambda(0, p) \le a(1-\epsilon)$ and $\lambda(1, p) \le a(1-\epsilon)$ then $\lambda(0, p) \ge a\delta(\epsilon)$ and $\lambda(1, p) \ge a\delta(\epsilon)$.

The condition 5) looks unnatural, but it follows from requirement of strict monotonicity of $\lambda(\sigma, p)$ by p and symmetry condition $\lambda(0, p) = \lambda(1, 1 - p)$.

Without loss of generality we can suppose that $b \ge a\delta(\epsilon)$ for all $0 < \epsilon < 1$. We can also prove that $b \le \frac{1}{2}a$ for any loss function satisfying 1) – 5).

The normalized by $a > 0$ square-loss function $\lambda(\sigma, p) = a(\sigma - p)^2$ satisfies these conditions with $\delta(\epsilon) = \epsilon^2/4$.

3 Summary of results

In this section we summarize main results in an asymptotic form. These results follow from the results of next section.

Let $KG(x)$ be predictive complexity for a loss function satisfying restrictions 1) – 5). Let us define a *worst-case ratio* function

$$f(n) = \sup_{x:l(x)=n} \frac{K(x)}{KG(x)}. \qquad (2)$$

The next theorem follows directly from Theorem 3 (below).

Theorem 1 *The worst-case ratio function $f(n)$ defined by (2) satisfies*

$$\lim_{n\to\infty} \frac{f(n)}{\frac{1}{a}\log n} = 1.$$

Let

$$h_n(t) = \sup_{x\in B_{n,t}} \frac{K(x)}{n}. \tag{3}$$

where

$$B_{n,t} = \{x | l(x) = n, \frac{KG(x)}{n} \le t\}. \tag{4}$$

Define *relative complexities comparing* functions

$$\underline{h}(t) = \lim\inf_{n\to\infty} h_n(t) \tag{5}$$

$$\overline{h}(t) = \lim\sup_{n\to\infty} h_n(t) \tag{6}$$

The following theorem is a direct corollary of Theorem 4 (below).

Theorem 2 *Let $0 < \epsilon < \delta^{-1}(\frac{b}{2a})$. The relative complexities comparing functions $\underline{h}(t)$ and $\overline{h}(t)$ defined by (5) and (6) satisfy*

$$-\frac{t}{a}\log\frac{t}{a} \le \underline{h}(t) \le \overline{h}(t) \le -\frac{t}{a(1-\epsilon)}\log\frac{t}{a(1-\epsilon)} + O(t),$$

when $t \to 0$ (the constant in $O(t)$ depends on ϵ).

4 Non-linear inequalities

In this section we explore some possible connections between Kolmogorov complexity $K(x)$ and predictive complexity $KG(x)$.

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 less than k is bigger than 2^{k-c} and less than 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 less than k. We denote by $\#A$ the cardinality of a finite set A. Let us consider a set

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

In this section we will consider only predictive complexity $KG(x)$ for a loss function satisfying restrictions 1) – 5).

Proposition 2 *Let $0 < \epsilon < 1$ be a rational number. Then there exists a constant c such that for all n and k the following inequalities hold*

$$\sum_{i \leq (k-c)/a} \binom{n}{i} \leq \#A_{n,k} \leq \sum_{i \leq k/b} \binom{k/(a\delta(\epsilon))}{i} \sum_{i \leq k/(a(1-\epsilon))} \binom{n}{i}. \tag{8}$$

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 item 4) of restrictions on loss function there are at least $\sum_{i \leq m} \binom{n}{i}$ 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 upper estimate (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. We assign some labelling to edges $(x, x0)$ and $(x, x1)$ of the binary tree using letters A, B and C, D as follows. For any x consider two cases.

Case 1. There is an edge $(x, x\sigma)$ such that $\lambda(\sigma, \Lambda(x)) \geq a(1-\epsilon)$. In this case we assign C to $(x, x\sigma)$ and D to $(x, x\hat{\sigma})$, where $\hat{\sigma} = 1$ if $\sigma = 0$, and $\hat{\sigma} = 0$ otherwise.

Case 2. Case 1 does not hold, i.e. $\lambda(\sigma, \Lambda(x)) \leq a(1-\epsilon)$ for all σ. In this case we assign the letter A to $(x, x0)$ and letter B to $(x, x1)$ if $\lambda(0, \Lambda(x)) \geq b$ and assign these letters vise versa, otherwise.

Evidently, two different sequences of length n have different labellings.

If some edge $(x, x\sigma)$ labeled by C then $\lambda(\sigma, \Lambda(x)) \geq a(1-\epsilon)$ and, hence, for any path x of length n having more than $\frac{k}{a(1-\epsilon)}$ letters C it holds $KG(x) \geq Loss_\Lambda(x) \geq k$.

By definition if some edge $(x, x\sigma)$ labeled by A or by B then $\lambda(\sigma, \Lambda(x)) \leq a(1-\epsilon)$ for all σ. Then by item 5) of the requirement on loss function we have $\lambda(\sigma, \Lambda(x)) \geq a\delta(\epsilon)$ for all σ. Hence, for any path x of the length n having more than $k/(a\delta(\epsilon))$ letters A or B it holds $KG(x) \geq Loss_\Lambda(x) \geq k$.

Hence, any sequence x of length n, on which $KG(x) \leq k$, can have no more than $k/(a\delta(\epsilon))$ letters A or B and no more than $\frac{k}{a(1-\epsilon)}$ letters C, the rest part of x are letters D. It also has no more than $\frac{k}{b}$ letters A.

By means of this labelings, every sequence $x \in A_{n,k}$ can be recovered from the following pair (α, β) of sequences. The first element of this pair is the sequence α of all letters A and B assigned to edges on x in the original order. This sequence contains no more than $\frac{k}{b}$ letters A. It is also can not be longer than $k/(a\delta(\epsilon))$. The second element of the pair is the sequence β of all letters C and D assigned to edges on x in the original order. This sequence contains no more than $\frac{k}{a(1-\epsilon)}$ letters C. Given these two sequences (α, β), the whole sequence x can be recovered as follows. Let $x^{i-1} = x_1 \dots x_{i-1}$, where $1 \leq i \leq n$, be already recovered by some initial fragments α^{s-1} and β^{q-1} of sequences α and β. We can place x^{i-1} in the binary tree supplied by new labellings and so define letters assigned to edges $(x^{i-1}, x^{i-1}0)$ and $(x^{i-1}, x^{i-1}1)$. Comparing these letters with α_s and β_q we can define which sequence must be used in recovering of the next member of x. The corresponding letter α_s or β_q of this sequence determines the member x_i of the sequence x.

Note, that the labelling and, hence, our method of recovering are incomputable. It gives us only a possibility to estimate the number of elements of the set $A_{n,k}$. The method of recovering shows that to do this, it is enough to estimate the number of all such pairs (α, β). It can be estimated as follows:

$$\#A_{n,k} \leq \sum_{i \leq k/b} \binom{k/(a\delta(\epsilon))}{i} \sum_{i \leq k/(a(1-\epsilon))} \binom{n}{i}.$$

□

Note, that upper estimate (8) is valid when $k \leq \min\{na\delta(\epsilon), na(1-\epsilon)\}$ (this means that k must be much smaller than n for small ϵ). A less strong but more simple upper estimate

$$\#A_{n,k} \leq \sum_{i \leq k/b} \binom{n}{i}$$

can be obtained using an analogous labeling only by letters A and B.

Proposition 3 *Let* $0 < \gamma < 1$ *and* $0 < \epsilon < \delta^{-1}(\frac{b}{2a})$.

- *(i) If in addition* $\epsilon \leq \frac{1}{2}$ *then a positive constant* c *exists such that for all* x

$$K(x) \leq \frac{KG(x)}{a(1-\epsilon)} \left(\log l(x) - (1-\gamma) \log \frac{KG(x)}{a(1-\epsilon)} \right) - \quad (9)$$

$$2\log\left(\frac{a\delta(\epsilon)}{b}\right)\frac{KG(x)}{b}+c. \tag{10}$$

- *(ii) For all sufficiently large n for all x of length n if $KG(x)\leq\frac{n}{2}a(1-\epsilon)$ then*

$$\frac{K(x)}{n}\leq H\left(\frac{KG(x)}{na(1-\epsilon)}\right)-2\log\left(\frac{a\delta(\epsilon)}{b}\right)\frac{KG(x)}{bn}+\frac{7\log n}{n}, \tag{11}$$

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

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. After that we make some transformations of the upper estimate (8) of Proposition 2 and replace k on $KG(x)$. For details see Section 6.1. □

Proposition 4 *Let $0<\gamma<1$, $0<\epsilon<\delta^{-1}(\frac{b}{2a})$. Then a positive constant c exists such that for each sufficiently large n and each $k\leq\frac{1}{2}na(1-\epsilon)$ a binary sequence x of length n exists such that*

$$k(1-\gamma)(1-\epsilon)\leq KG(x)\leq k+c, \tag{12}$$

$$K(x)\geq\log\binom{n}{k/a}-1\geq nH\left(\frac{KG(x)}{an}\right)-2\log n \tag{13}$$

and also

$$K(x)\geq\frac{KG(x)}{a}\left(\log n-\log\frac{KG(x)}{a}\right)-2. \tag{14}$$

Sketch of the proof. We will find x satisfying the condition of this proposition in the set $A_{n,k}$ defined by (7). We must estimate minimal k' such that $\#A_{n,k}\geq\binom{n}{(k-c)/a}>2\#A_{n,k'}$. We show in Section 6.2 that this inequality holds for all sufficiently large n if $k'=(k-c)(1-\gamma)(1-\epsilon)$, where c is a constant from lower estimate (8). By incompressibility property of Kolmogorov complexity and lower estimate (8) an $x\in A_{n,k}-A_{n,k'}$ exists such that $K(x)\geq\log\binom{n}{(k-c)/a}-2$. After that, using appropriate estimates of binomial coefficients and replacing k on $k-c$ we obtain inequalities (12), (13) and (14). For details see Section 6.2. □

The next corollary from propositions 3 and 4 gives precise relations between normalized Kolmogorov and predictive complexities. This result is too technical and it is reformulated in the Section 3 in a more convenient form.

Corollary 1 *Let $0 < \epsilon < \delta^{-1}(\frac{b}{2a})$. Then for all sequences x of sufficiently large length if $KG(x) \leq \frac{1}{2}na(1-\epsilon)$ then*

$$\frac{K(x)}{l(x)} \leq H\left(\frac{KG(x)}{a(1-\epsilon)l(x)}\right) - 2\log\left(\frac{a\delta(\epsilon)}{b}\right)\frac{KG(x)}{bl(x)} + \frac{7\log l(x)}{l(x)}$$

and for each sufficiently large n there is some x of length n such that

$$\frac{K(x)}{l(x)} \geq H\left(\frac{KG(x)}{al(x)}\right) - \frac{2\log l(x)}{l(x)}$$

Proof. This corollary follows from (11) and (13). □

Theorem 3 *Let $0 < \epsilon < \min\{\frac{1}{2}, \delta^{-1}(\frac{b}{2a})\}$. Then there exists a constant c such that for all n*

$$\frac{1}{a}\log n - c \leq f(n) \leq \frac{1}{a(1-\epsilon)}\log n - \frac{2}{b}\log\delta(\epsilon) + c$$

where $f(n)$ is the worst-case ratio function defined by (2).

Proof. The right-hand inequality follows directly from (9). The left-hand inequality can be derived from (12) and (14) of Proposition 4. It is enough to let $k = n^{\epsilon}$. Taking $\epsilon \to 0$ we obtain the needed inequality. □

Theorem 4 *Let $0 < \epsilon < \delta^{-1}(\frac{b}{2a})$. Then for each real number $t \leq \frac{1}{2}a(1-\epsilon)$*

$$H\left(\frac{t}{a}\right) \leq \underline{h}(t) \leq \overline{h}(t) \leq H\left(\frac{t}{a(1-\epsilon)}\right) - \frac{2}{b}\log\frac{a\delta(\epsilon)}{b}t. \qquad (15)$$

where $\underline{h}(t)$ and $\overline{h}(t)$ are relative complexities comparing functions defined by (5) and (6).

Proof. This theorem follows directly from Corollary 1. □

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 3

We will use the following estimates of the binomial coefficients from [3], Section 6.1.

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k \tag{16}$$

and estimates

$$\sum_{i \leq m} \binom{n}{i} \leq (m+1)\binom{n}{m}, \tag{17}$$

$$\log \binom{n}{s} \leq nH(\frac{s}{n}) \tag{18}$$

for any $m \leq \frac{n}{2}$ and $s \leq n$. We use also inequality

$$\frac{H(p)}{p} \leq -2 \log p \tag{19}$$

for all $0 < p < \frac{1}{2}$.

Let $k \leq \frac{n}{2} a(1-\epsilon)$. We have also $\frac{k}{b} \leq \frac{1}{2} \frac{k}{a\delta(\epsilon)}$ for all $\epsilon < \delta^{-1}(\frac{b}{2a})$.

To prove inequality (9) let us consider the recursively enumerable set $A_{n,k}$ defined by (7). We can specify any $x \in A_{n,k}$ by n, k and the ordinal number of x in the natural enumeration of $A_{n,k}$. Using an appropriate encoding of all triples of positive integer numbers by upper estimate (8) of Proposition 2 and using (16), (17), (18), (19) we obtain for all $x \in A_{n,k}$

$$K(x) \leq \log \# A_{n,k} + 2\log n + 2\log k + c \leq \tag{20}$$

$$\log \frac{k}{b} \binom{k/(a\delta(\epsilon))}{k/b} + \log \frac{k}{a(1-\epsilon)} \binom{n}{k/(a(1-\epsilon))} + \tag{21}$$

$$2\log n + 2\log k + c \leq \tag{22}$$

$$\log \frac{k}{b} + \frac{k}{a\delta(\epsilon)} H\left(\frac{a\delta(\epsilon)}{b}\right) + \log \frac{k}{a(1-\epsilon)} + \log \left(\frac{en}{k/(a(1-\epsilon))}\right)^{k/(a(1-\epsilon))} + \tag{23}$$

$$2\log n + 2\log k + c' = \tag{24}$$

$$\log \frac{k}{b} + \frac{k}{a\delta(\epsilon)} H\left(\frac{a\delta(\epsilon)}{b}\right) + \log \frac{k}{a(1-\epsilon)} + \tag{25}$$

$$\frac{k}{a(1-\epsilon)}\left(\log n + \log e - \log\frac{k}{a(1-\epsilon)}\right) + 2\log n + 2\log k + c' \le \quad (26)$$

$$\left(\frac{k}{a(1-\epsilon)} + 2\right)\left(\log n - (1-\gamma)\log\frac{k}{a(1-\epsilon)}\right) - \frac{2k}{b}\log\frac{a\delta(\epsilon)}{b} + c'', \quad (27)$$

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

Put $k = KG(x) + 2a(1-\epsilon)$. Then by inequalities (20)–(27), we obtain in the case $KG(x) \le \frac{1}{2}na(1-\epsilon) + 2a(1-\epsilon) = a(1-\epsilon)(\frac{n}{2}+2)$ the following inequality

$$K(x) \le \frac{KG(x)}{a(1-\epsilon)}\left(\log n - (1-\gamma)\log\frac{KG(x)}{a(1-\epsilon)}\right) - \frac{2KG(x)}{b}\log\frac{a\delta(\epsilon)}{b} + c \quad (28)$$

for some positive constant c.

Consider two strategies $S_1(z) = 0$ and $S_2(z) = 1$ for all z. Then for each x of length n it holds $Loss_{S_1}(x) \le \frac{a}{2}n$ or $Loss_{S_2}(x) \le \frac{a}{2}n$. Therefore the inequality $KG(x) \le \frac{1}{2}an + c$ holds for some positive constant c. If $KG(x) \ge \frac{n}{2}a(1-\epsilon)$ we have for all n and all x of length n

$$K(x) \le n + 2\log n + c_1 \le \frac{n}{2}\log n - \frac{n}{2}(1-\gamma)\log n + c_2 \le \quad (29)$$

$$\frac{KG(x)}{a(1-\epsilon)}\left(\log n - (1-\gamma)\log\frac{KG(x)}{a(1-\epsilon)} - (1-\gamma)\log(2(1-\epsilon))\right) + c_3, \quad (30)$$

where c_1, c_2, c_3 are positive constants. Inequality (9), (10) follows from (29), (30) when $\epsilon \le \frac{1}{2}$. Item (i) is proved.

Let us consider the item (ii). In the case $KG(x) \le \frac{n}{2}a(1-\epsilon)$ inequality (11) can be obtained by applying inequality (18) to the second binomial coefficient of (21) as follows.

$$K(x) \le \frac{k}{a\delta(\epsilon)}H\left(\frac{a\delta(\epsilon)}{b}\right) + nH\left(\frac{k}{na(1-\epsilon)}\right) + 6\log n + c, \quad (31)$$

where c is a positive constant.

Putting $k = KG(x)$ in (31) and dividing on n we obtain for any $\epsilon < \delta^{-1}(\frac{b}{2a})$ for all sufficiently large n

$$\frac{K(x)}{n} \le H\left(\frac{KG(x)}{na(1-\epsilon)}\right) + \frac{KG(x)}{na\delta(\epsilon)}H\left(\frac{a\delta(\epsilon)}{b}\right) + \frac{7\log n}{n} \le$$
$$H\left(\frac{KG(x)}{na(1-\epsilon)}\right) - 2\log\left(\frac{a\delta(\epsilon)}{b}\right)\frac{KG(x)}{bn} + \frac{7\log n}{n}.$$

6.2 Proof of Proposition 4

We will find x satisfying the condition of this proposition in the set $A_{n,k}$ defined by (7). We must find some k' such that $\#A_{n,k} > 2\#A_{n,k'}$.

By the upper and lower estimates (8) of Proposition 2 it is sufficient that k' be satisfy

$$\binom{n}{(k-c)/a} > 2 \sum_{i \le k'/(a(1-\epsilon))} \binom{n}{i} \sum_{i \le k'/b} \binom{k'/(a\delta(\epsilon))}{i}, \tag{32}$$

where c is a constant from the lower estimate (8).

We will find k' satisfying $k' \le \frac{n}{2}a(1-\epsilon)$. By (16) inequality (32) follows from

$$\left(\frac{na}{k-c}\right)^{\frac{k-c}{a}} \ge \frac{4k'}{b}\left(\frac{eb}{a\delta(\epsilon)}\right)^{\frac{k'}{b}}\left(\frac{ena(1-\epsilon)}{k'}\right)^{\frac{k'}{a(1-\epsilon)}}\frac{k'}{a(1-\epsilon)}. \tag{33}$$

Inequality (33) holds for all sufficiently large n if $k' = (k-c)(1-\gamma)(1-\epsilon)$. Then for each sufficiently large n we have $\#A_{n,k} > 2\#A_{n,k'}$ and

$$(k-c)(1-\gamma)(1-\epsilon) \le KG(x) \le k \tag{34}$$

for all $x \in A_{n,k} - A_{n,k'}$. We have also $k' \le \frac{n}{2}a(1-\epsilon)$ if $k \le \frac{1}{2}na(1-\epsilon) + c$.

By incompressibility property of Kolmogorov complexity we have that an $x \in A_{n,k} - A_{n,k'}$ exists such that

$$K(x) \ge \log\binom{n}{(k-c)/a} - 2 \ge nH\left(\frac{k-c}{an}\right) - 2\log n. \tag{35}$$

Here we used the last inequality on the page 66 of [5]. We obtain also by (16)

$$K(x) \ge \log\binom{n}{(k-c)/a} - 2 \ge \frac{k-c}{a}\log n - \frac{k-c}{a}\log\frac{k-c}{a} - 2 = \tag{36}$$

$$\frac{k-c}{a}\left(\log n - \log\frac{k-c}{a}\right) - 2. \tag{37}$$

Now replacing in the proof of the proposition k on $k+c$ and putting $k = KG(x)$ we obtain from (35) and (37) inequalities (13) and (14). Inequality (12) follows from (34).

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

[7] 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.

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

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

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

[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'