How Many Strings Are Easy to Predict? [⋆]

Yuri Kalnishkan, Vladimir Vovk, and Michael V. Vyugin

Department of Computer Science, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, United Kingdom

Abstract

It is well known in the theory of Kolmogorov complexity that most strings cannot be compressed; more precisely, only exponentially few ($\Theta(2^{n-m})$) binary strings of length n can be compressed by m bits. This paper extends the 'incompressibility' property of Kolmogorov complexity to the 'unpredictability' property of predictive complexity. The 'unpredictability' property states that predictive complexity (defined as the loss suffered by a universal prediction algorithm working infinitely long) of most strings is close to a trivial upper bound (the loss suffered by a trivial minimax constant prediction strategy). We show that only exponentially few strings can be successfully predicted and find the base of the exponent.

1 Introduction

We consider the following on-line prediction problem: given observed outcomes $x_1, x_2, \ldots, x_{n-1}$, the prediction algorithm is required to output a prediction γ_n for the new outcome x_n. Let all outcomes be either 0 or 1 and predictions be real numbers from the interval $[0,1]$. A loss function $\lambda(\omega, \gamma)$ is used to measure the discrepancy between predictions and actual outcomes. The performance of the algorithm is measured by the cumulative loss $\sum_{i=1}^{n} \lambda(x_i, \gamma_i)$. This problem has been extensively studied; see, e.g., [CBFH$^+$97,HKW98,LW94]. The existing literature is mainly concerned with construction of specific prediction algorithms and studying their properties. This paper deals with a more abstract question: how many strings can be successfully predicted?

[⋆] The previous version of this paper was published in proceedings of the 16th Annual Conference on Computational Learning Theory, COLT 2003, vol. 2777 of *Lecture Notes in Artificial Intelligence*, Springer, 2003.

Email addresses: yura@cs.rhul.ac.uk (Yuri Kalnishkan), vovk@cs.rhul.ac.uk (Vladimir Vovk), misha@cs.rhul.ac.uk (Michael V. Vyugin).

The difficulty of predicting a sequence $x_1, x_2, \ldots, x_n$ is formalised by the notion of predictive complexity (introduced in [VW98]). The loss suffered by any prediction algorithm on a sequence is at least the predictive complexity of the sequence and predictive complexity can be approached in the limit if a universal prediction algorithm is allowed to work infinitely long.

This paper shows that most strings have predictive complexity close to the loss of a trivial minimax constant strategy. In other words, we prove that on most strings even the idealised optimal algorithm performs little better than the trivial strategy. Of course, this result is hardly surprising: one would not expect a random string to be predictable. The interesting part is the number of predictable sequences: even though the fraction of such sequences is tiny, we happen to be especially interested in them. Our main result (Theorem 1) says that the idealised optimal algorithm beats the trivial strategy by m on the fraction $\Theta(\beta_0^m)$ of strings of length n, where $\beta_0 \in (0,1)$ is a constant determined by the loss function.

The situation is similar to that with Kolmogorov complexity, which formalises our intuition concerning the shortest description of an object. Results of the theory of predictive complexity are expressed in the same asymptotic fashion as the results about Kolmogorov complexity. In fact, Kolmogorov complexity coincides with predictive complexity for a particular loss function called the logarithmic loss function.

One of the important properties of Kolmogorov complexity is the so called incompressibility property (see [LV97], Sect. 2.2). It states that for most strings Kolmogorov complexity is close to the length of the string (see Appendix A for the exact statements). The intuitive interpretation is that for a random string there is no substantially shorter way to describe it than to list its elements in the most straightforward way. The unpredictability property generalises the incompressibility property and states that for most strings there is no substantially better way to predict their elements than to always use the same trivial minimax strategy. The fraction of strings of length n that have Kolmogorov complexity less than $n - m$ is $\Theta(2^{-m})$; therefore, $\beta_0 = 1/2$ for the logarithmic loss. In general, β_0 is determined by both the local behaviour of the loss function in the neighbourhood of the minimax and its global behaviour.

Sect. 2 defines predictive complexity. Sect. 3 contains the statement of the main theorem followed by a discussion. The main result is proven in Sect. 4.

2 Preliminaries

We consider finite strings of elements from the binary alphabet $\mathbb{B} = \{0, 1\}$ and denote these strings by bold letters. The set of all finite binary strings is denoted by $\mathbb{B}^*$. The length of a string $\boldsymbol{x} \in \mathbb{B}^*$ is denoted by $|\boldsymbol{x}|$. We use similar notation $|A|$ for the number of elements of a set A. The notation $\boldsymbol{x}^{(k)}$ refers to the prefix of $\boldsymbol{x}$ of length k. The notation $\mathbb{N}$ refers to the set of non-negative integers $\{0, 1, 2, \ldots\}$.

We will now define predictive complexity and discuss some of its properties.

2.1 The Definition of Predictive Complexity

An *on-line prediction game* (or simply *game*) $\mathfrak{G}$ is a triple $(\mathbb{B}, [0, 1], \lambda)$, where $\mathbb{B} = \{0, 1\}$ is the *outcome space*[1], $[0, 1]$ is the *prediction space*, and $\lambda : \mathbb{B} \times [0, 1] \to [0, +\infty]$ is a *loss function*. We suppose that λ is computable and continuous.

The following are examples of games: the *square-loss* game with the loss function $\lambda(\omega, \gamma) = (\omega - \gamma)^2$ and the *logarithmic* game with

$$\lambda(\omega, \gamma) = \begin{cases} -\log_2(1 - \gamma) & \text{if} \quad \omega = 0 \ , \\ -\log_2 \gamma & \text{if} \quad \omega = 1 \ . \end{cases}$$

Consider a computable *prediction strategy* $\mathfrak{A} : \mathbb{B}^* \to [0, 1]$; it maps a sequence of outcomes into a prediction. We say that on a finite sequence $x_1 x_2 \ldots x_n \in \mathbb{B}^n$, where $n \in \mathbb{N}$, the strategy $\mathfrak{A}$ suffers loss

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

A function $L : \mathbb{B}^* \to [0, +\infty]$ is a *loss process* if it coincides with the loss of a computable prediction strategy.

An equivalent definition of a loss process can be given as follows. A computable function $L : \mathbb{B}^* \to [0, +\infty]$ is a loss process if $L(\Lambda) = 0$, where Λ is the empty string, and for every $\boldsymbol{x} \in \mathbb{B}^*$ there is $\gamma \in [0, 1]$ such that

$$\begin{cases} L(\boldsymbol{x}0) - L(\boldsymbol{x}) = \lambda(0, \gamma) \ , \\ L(\boldsymbol{x}1) - L(\boldsymbol{x}) = \lambda(1, \gamma) \ . \end{cases} \tag{1}$$

[1] In this paper we restrict ourselves to games with the outcome space $\mathbb{B}$. These games are sometimes called ‘binary’. A more general definition is possible.

Unfortunately, for the majority of nontrivial games the class of loss processes does not have a universal (i.e., smallest in some natural sense) element. It can be easily shown using a simple diagonalisation argument that every computable strategy is greatly outperformed by some other computable strategy on some sequences. To overcome this problem we extend the class of loss processes to the class of superloss processes. The original idea goes back to Kolmogorov and Levin, who applied it to what is in our terms the logarithmic game. A *superloss process* is a function $L : \mathbb{B}^* \to [0, +\infty]$ such that

- L is semi-computable from above,
- $L(\Lambda) = 0$, and
- for every $\boldsymbol{x} \in \mathbb{B}^*$ there is $\gamma \in [0, 1]$ such that

$$\begin{cases} L(\boldsymbol{x}0) - L(\boldsymbol{x}) \geq \lambda(0, \gamma) \ , \\ L(\boldsymbol{x}1) - L(\boldsymbol{x}) \geq \lambda(1, \gamma) \ . \end{cases} \tag{2}$$

We say that a superloss process $\mathcal{K}$ is *universal* if for every superloss process L there is a constant C such that $\mathcal{K}(\boldsymbol{x}) \leq L(\boldsymbol{x}) + C$ for all strings $\boldsymbol{x}$. As we will see below, many games, including the logarithmic and the square-loss, have universal superloss processes. A universal process for some game is called *predictive complexity* for that game.

We will need a more general definition of conditional complexity. Let Ξ be an ensemble of constructive objects containing the representations of all finite sequences $\mathbb{B}^*$ and natural numbers $\mathbb{N}$. A function $L : \mathbb{B}^* \times \Xi \to [0, +\infty]$ (we will separate arguments of L by the vertical line | rather than by the comma) is a *conditional superloss process* if

- $L(\Lambda \mid \boldsymbol{y}) = 0$ for all $\boldsymbol{y} \in \mathbb{B}^*$,
- L is semi-computable from above, and
- for every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{B}^*$ there is $\gamma \in [0, 1]$ such that

$$\begin{cases} L(\boldsymbol{x}0 \mid \boldsymbol{y}) - L(\boldsymbol{x} \mid \boldsymbol{y}) \geq \lambda(0, \gamma) \ , \\ L(\boldsymbol{x}1 \mid \boldsymbol{y}) - L(\boldsymbol{x} \mid \boldsymbol{y}) \geq \lambda(1, \gamma) \ . \end{cases} \tag{3}$$

In other terms, it is required that $L(\boldsymbol{x} \mid \boldsymbol{y})$ should be uniformly semicomputable from above and, for each fixed $\boldsymbol{y}$, should be a superloss process. The intuition behind this concept is that the learner may have access to certain additional information.

A conditional superloss process $\mathcal{K}$ is *universal* if for every conditional superloss process L there is a constant C such that the inequality $\mathcal{K}(\boldsymbol{x} \mid \boldsymbol{y}) \leq L(\boldsymbol{x} \mid \boldsymbol{y}) + C$ holds for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{B}^*$. A universal conditional superloss process is called *conditional predictive complexity*. It is easy to see that the existence of conditional predictive complexity $\mathcal{K}(\boldsymbol{x} \mid \boldsymbol{y})$ implies the existence of $\mathcal{K}(\boldsymbol{x})$, and $\mathcal{K}(\boldsymbol{x} \mid \Lambda)$ coincides with $\mathcal{K}(\boldsymbol{x})$ up to an additive constant.

We call a point $(x_0, x_1) \in [0, +\infty]^2$ a *superprediction* w.r.t. $\mathfrak{G}$ if there is a prediction $\gamma \in [0,1]$ such that $x_0 \geq \lambda(0,\gamma)$ and $x_1 \geq \lambda(1,\gamma)$; let S be the set of superpredictions. Geometrically, S is the set of points lying to the north-east of the curve $\{(\lambda(0,\gamma), \lambda(1,\gamma) \mid \gamma \in [0,1]\}$.

We need to define some classes of games in terms of S. We say that $\mathfrak{G}$ is *symmetric* if S is symmetric w.r.t. the straight line $x = y$. For example, if λ is such that $\lambda(0,t) = \lambda(1, 1-t)$, then the game is symmetric.

Mixability is a less trivial property introduced in [VW98]. Take a parameter $\beta \in (0,1)$ and consider the homeomorphism $\mathfrak{B}_\beta : [0,+\infty]^2 \to [0,1]^2$ specified by

$$\mathfrak{B}_\beta(x,y) = (\beta^x, \beta^y) \ . \tag{4}$$

A game $\mathfrak{G}$ with the set of superpredictions S is called *β-mixable* if the set $\mathfrak{B}_\beta(S)$ is convex. A game $\mathfrak{G}$ is *mixable* if it is β-mixable for some $\beta \in (0,1)$.

The mixability property is equivalent to the existence of predictive complexity. It is shown in [VW98] that if a game $\mathfrak{G}$ with the set of superpredictions S is mixable then there is predictive complexity w.r.t. $\mathfrak{G}$. The converse theorem is proven in [KVV04] under certain computability assumptions on games. The same equivalence holds for conditional predictive complexity.

It is easy to see that there is a positive constant C such that for all strings $\boldsymbol{x}$ and $\boldsymbol{y}$ the inequality $\mathcal{K}(\boldsymbol{x} \mid \boldsymbol{y}) \leq \mathcal{K}(\boldsymbol{x}) + C$ holds. On the other hand, by applying the Aggregating Algorithm (see [VW98]) it may be shown that, if the game is mixable, then there is a positive constant c such that $\mathcal{K}(\boldsymbol{x}) \leq \mathcal{K}(\boldsymbol{x} \mid \boldsymbol{y}) + c\mathrm{KP}(\boldsymbol{y})$ for all strings $\boldsymbol{x}$ and $\boldsymbol{y}$, where KP stands for prefix complexity.

The square-loss and the logarithmic games are both mixable and thus they specify conditional and unconditional complexities. We denote them by $\mathcal{K}^{\mathrm{sq}}$ and $\mathcal{K}^{\log}$, respectively. It follows from the definition that the (unconditional) complexity w.r.t. the logarithmic game coincides with KM, the negative logarithm of Levin's a priori semimeasure. Indeed (see [LV97], Sect. 4), $\mathrm{KM} = -\log_2 \mathrm{M}$, where M is an a priori semimeasure defined as follows. An *(enumerable continuous) semimeasure* $\mu : \mathbb{B}^* \to [0,1]$ is a function such that

- μ is semi-computable from below,
- $\mu(\Lambda) \leq 1$, and
- for every $\boldsymbol{x} \in \mathbb{B}^*$ we have

$$\mu(\boldsymbol{x}) \geq \mu(\boldsymbol{x}0) + \mu(\boldsymbol{x}1) \ . \tag{5}$$

A semimeasure M is an *a priori semimeasure* if for every semimeasure μ

there is a constant $C > 0$ such that $CM \geq \mu$. Take $L = -\log_2 \mu$. One can rewrite (5) as $2^{-L(\boldsymbol{x}0)} + 2^{-L(\boldsymbol{x}1)} \leq 2^{-L(\boldsymbol{x})}$, which is equivalent to the existence of $\gamma \in [0,1]$ satisfying (2) with the logarithmic loss function. One can check this by excluding γ from the system. Thus $\mathcal{K}^{\log}$ coincides with KM.

The function KM differs from plain Kolmogorov complexity K by a term of logarithmic order in the length of the string, i.e., there is $c > 0$ such that for all strings $\boldsymbol{x} \neq \Lambda$ we have $|\mathrm{K}(\boldsymbol{x}) - \mathrm{KM}(\boldsymbol{x})| \leq c \ln |\boldsymbol{x}|$. A proof may be found in [LV97]. The definition of plain Kolmogorov complexity K is given in Appendix A.

3 Main Results and Discussion

Theorem 1 (Unpredictability Property) *Let $\mathfrak{G}$ be a mixable symmetric game specifying conditional predictive complexity $\mathcal{K}$. Suppose that the set S of superpredictions for $\mathfrak{G}$ is such that the boundary ∂S is a twice differentiable curve in a vicinity of the point (B,B), where $B = \inf\{t \in \mathbb{R} \mid (t,t) \in S\}$. Then the inequalities*

$$\sup_{n,m\in\mathbb{N}} \frac{|\{\boldsymbol{x} \in \mathbb{B}^n \mid \mathcal{K}(\boldsymbol{x} \mid m) \leq Bn - m\}|}{2^n \beta_0^m} \leq 1 \tag{6}$$

and

$$\inf_{m\in\mathbb{N}} \varliminf_{n\to\infty} \frac{|\{\boldsymbol{x} \in \mathbb{B}^n \mid \mathcal{K}(\boldsymbol{x} \mid m) \leq Bn - m\}|}{2^n \beta_0^m} > 0 \tag{7}$$

hold, where $\beta_0 \in (0,1)$ is the infimum of all $\beta \in (0,1)$ such that the set $\mathfrak{B}_\beta(S)$ lies below the straight line $x + y = 2\beta^B$.

This theorem assumes not only that the loss function is computable but also that β_0 is computable and that the set

$$\mathfrak{B}_{\beta_0}(S) \cap \{(x,y) \mid x \neq y \text{ and } x + y = 2\beta_0^B\}$$

contains a computable point (if non-empty). For specific loss functions studied in the literature this is always the case.

Let us discuss the theorem informally. Inequality (6) means that for all positive integers n and m the inequality

$$\frac{|\{\boldsymbol{x} \in \mathbb{B}^n \mid \mathcal{K}(\boldsymbol{x} \mid m) \leq Bn - m\}|}{2^n} \leq \beta_0^m$$

holds. This can be interpreted as a statement about the probability provided we assign equal probabilities to all strings of length n. Lemma 2 proves this inequality for any $\beta \in (0,1)$ such that $\mathfrak{B}_\beta(S)$ lies below the straight line $x/2 + y/2 = \beta^B$. The value β_0 simply provides the best bound.

Note that values of β such that $\mathfrak{B}_\beta(S)$ lies below the straight line $x/2 + y/2 = \beta^B$ exist. Indeed, since $\mathfrak{G}$ is mixable, the set $\mathfrak{B}_\beta(S)$ is convex for some $\beta \in (0,1)$. Since the set $\mathfrak{B}_\beta(S)$ is symmetric in the bisector $x = y$, the line $x + y = 2\beta^B$ is a support line for $\mathfrak{B}_\beta(S)$ and convex sets are not cut by their support lines (see, e.g., [Egg58]).

However, in the general case, $\mathfrak{G}$ is not necessarily β_0-mixable. It is possible that for some values of β the game is not β-mixable, i.e., the set $\mathfrak{B}_\beta(S)$ is not convex, but $\mathfrak{B}_\beta(S)$ still lies below the straight line in question (cf. Fig. 1).

Inequality (7) shows that the value β_0 cannot be decreased further. The inequality can be reformulated as follows. There is a constant $\theta > 0$ such that for every positive integer m there is a number $n_0(m)$ such that for all integers $n > n_0(m)$ the inequality

$$\frac{|\{\boldsymbol{x} \in \mathbb{B}^n \mid \mathcal{K}(\boldsymbol{x} \mid m) \leq Bn - m\}|}{2^n} \geq \theta\beta_0^m \tag{8}$$

holds. If we portray a pair (n,m) by a corresponding point in the positive quadrant, (8) holds inside a certain ‘wedge’. Lemmas 4 and 5 give insight into the shape of this wedge. Depending on the set S, there are two cases, which are addressed by the lemmas. In one of the cases we show that $n_0(m)$ can be taken to be equal cm, where c is a constant independent of m and in the other case we take $n_0(m) = cm^3$ (given certain regularity conditions, it is possible to reduce the degree and to take $n_0(m) = cm^2$). The exact shape of the ‘wedge’ is an open problem.

It is easy to see that for the logarithmic game $B = 1$ and $\beta_0 = 1/2$ and thus our theorem states that

$$|\{\boldsymbol{x} \in \mathbb{B}^n \mid \mathrm{KM}(\boldsymbol{x} \mid m) \leq n - m\}| \leq 2^{n-m} \tag{9}$$

for all positive integers n and m. On the other hand, there is $\theta > 0$ such that for every positive integer m for some n on we have

$$|\{\boldsymbol{x} \in \mathbb{B}^n \mid \mathrm{KM}(\boldsymbol{x} \mid m) \leq n - m\}| \geq \theta 2^{n-m} \ . \tag{10}$$

Note that θ is uniform in m while the value $n_0(m)$ such that the inequality holds for all $n > n_0(m)$ can differ for different m.

Appendix A reviews the definition of the plain Kolmogorov complexity K and the incompressibility property. It is remarkable that similar estimates exist for the plain Kolmogorov complexity while their short proof is based on completely different ideas.

4 Proof of the Main Theorem

This section contains the proof of the main theorem. The proof splits into a number of lemmas.

PROOF of Theorem 1 First the existence of values of $\beta \in (0,1)$ such that $\mathfrak{B}_\beta(S)$ lies below the line $x+y=2\beta^B$ is proved in the discussion in Section 3.

Secondly the infimum of such values of β is greater than 0. This is implied by the formula for the second derivative of the function $g_\beta(x)$ whose graph represents $\partial\mathfrak{B}_\beta(S)$ in a vicinity of (β^B,β^B). Indeed, let $x(t)$ and $y(t)$ be smooth functions parameterising the boundary ∂S in a vicinity of (B,B). For definiteness sake, assume that $x(t)$ strictly increases and $y(t)$ strictly decreases. Then $g_\beta(\beta^x)=\beta^y$ and

$$\frac{d^2\beta^{y(t)}}{d\left(\beta^{x(t)}\right)^2} = \frac{\beta^{y(t)-2x(t)}}{\ln\beta\cdot(x'(t))^2}\left((y'(t)-x'(t))y'(t)\ln\beta+\frac{y''(t)x'(t)-y'(t)x''(t)}{x'(t)}\right) .$$

The inequality $g''_\beta(\beta^x)\le 0$ is equivalent to

$$(y'(t)-x'(t))y'(t)\ln\beta \ge -\frac{y''(t)x'(t)-y'(t)x''(t)}{x'(t)} . \tag{11}$$

For every fixed value of t, the left-hand side is a negative value which tends to $-\infty$ as $\beta\to 0$, while the right-hand side is a fixed number. Thus the inequality gets violated for small values of β.

Inequality (6) follows from Lemma 2 below.

In order to prove (7), we need to consider two cases. Either there is $\Delta\in(0,\beta_0^B]$ such that the inverse image $\mathfrak{B}_{\beta_0}^{-1}(\beta_0^B-\Delta,\beta_0^B+\Delta)$ is a superprediction, or the line $x+y=2\beta_0^B$ and the curve $\partial\mathfrak{B}_{\beta_0}(S)$ have a contact of the second order [2] at (β_0^B,β_0^B) (see Figs. 1 and 2).

Indeed, suppose that g_{β_0} has a strictly negative second derivative at the point β_0^B. By continuity, for some small $\beta<\beta_0$ the second derivative of g_β at β^B will remain negative and thus there are $\varepsilon,\delta>0$ such for every $\beta\in[\beta_0-\varepsilon,\beta_0]$ the

[2] We say that curves $y=f_1(x)$ and $y=f_2(x)$ have a contact of the n-th order at a point (x_0,y_0) if $y_0=f_1(x_0)=f_2(x_0)$, $f_1'(x_0)=f_2'(x_0)$,..., $f_1^{(n)}(x_0)=f_2^{(n)}(x_0)$. This simple definition if sufficient for our purposes; of course it can be refined and made coordinate-independent.

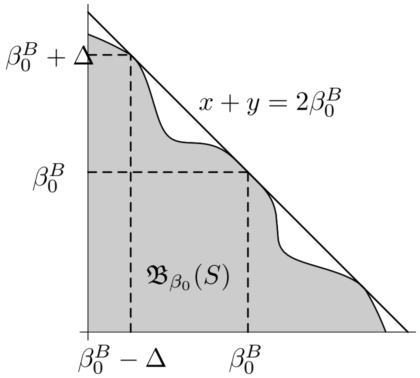

Fig. 1. The case of positive Δ; the set $\mathfrak{B}_{\beta_0}(S)$ is shaded

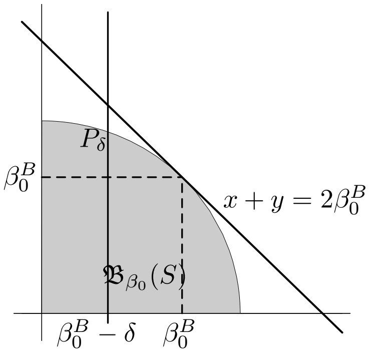

Fig. 2. The case of the contact of the second order; the set $\mathfrak{B}_{\beta_0}(S)$ is shaded

part of $\mathfrak{B}_\beta(S)$ in the stripe $[\beta^B-\delta, \beta^B+\delta]\times\mathbb{R}$ lies below the line $x+y=2\beta^B$. Since β_0 is the infimum, there exists Δ in question.

The cases are considered in Lemmas 4 and 5. $\square$

4.1 The Upper Bound on Probability

In this subsection we derive the upper bound on the number of strings of low complexity.

Lemma 2 *Let $\mathfrak{G}$ be a symmetric game with the set of superpredictions S; let $B=\inf\{t\in\mathbb{R}\mid (t,t)\in S\}$ and let L be a conditional superloss process w.r.t. $\mathfrak{G}$. Let $\beta\in(0,1)$ be such that $\mathfrak{B}_\beta(S)$ lies below the straight line $x+y=2\beta^B$. Then for all positive integers n and m we have*

$$\frac{\left|\{\boldsymbol{x}\in\mathbb{B}^n\mid \exists k\in\{1,2,\ldots,n\}: L(\boldsymbol{x}^{(k)}\mid m)\leq Bk-m\}\right|}{2^n}\leq\beta^m\quad . \tag{12}$$

PROOF. Consider the function $M_m(\boldsymbol{x})=\beta^{L(\boldsymbol{x}|m)-B|\boldsymbol{x}|}$. If we show that for every fixed m it is a supermartingale w.r.t. the Bernoulli distribution with the probability of success equal to 1/2 and apply a variant of Doob's inequality, the bound will follow. The definitions of a martingale and a supermartingale and the required inequality may be found in Appendix B. We must check that $\frac{1}{2}M_m(\boldsymbol{x}0)+\frac{1}{2}M_m(\boldsymbol{x}1)\leq M(\boldsymbol{x})$.

Lemma 3 *Under the conditions of Lemma 2, for every $\boldsymbol{x}\in\mathbb{B}^*$, the inequality*

$$\frac{1}{2}\beta^{L(\boldsymbol{x}0|m)-B(|\boldsymbol{x}0|)}+\frac{1}{2}\beta^{L(\boldsymbol{x}1|m)-B(|\boldsymbol{x}1|)}\leq\beta^{L(\boldsymbol{x}|m)-B|\boldsymbol{x}|}$$

holds for every positive integer m.

PROOF of Lemma 3 It follows from the definition of predictive complexity that the pair $(L(\boldsymbol{x}0 \mid m) - L(\boldsymbol{x} \mid m), L(\boldsymbol{x}1 \mid m) - L(\boldsymbol{x} \mid m))$ is a superprediction, i.e., belongs to S. The conditions of Lemma 2 imply that for every $(x, y) \in \mathfrak{B}_\beta(S)$, the inequality $x/2 + y/2 \leq \beta^B$ holds. The lemma follows. □

It follows from this lemma that $M_m(\boldsymbol{x})$ is a supermartingale. We can now apply Prop. 9. □

4.2 Tightness of the Bound

In this subsection we show that the bound from the previous subsection is tight.

Lemma 4 *Let $\mathfrak{G}$ be a symmetric game with the set of superpredictions S; let $B = \inf\{t \in \mathbb{R} \mid (t, t) \in S\}$ and let $\mathfrak{G}$ specify conditional complexity $\mathcal{K}$. Let $\beta_0 \in (0, 1)$ and $\Delta \in (0, \beta_0^B]$ be such that the point $\mathfrak{B}_{\beta_0}^{-1}((\beta_0^B - \Delta, \beta_0^B + \Delta))$ is a superprediction. Then there are positive constants c and θ such that for every positive integer m and positive integer $n \geq cm$ the inequality*

$$\theta\beta_0^m \leq \frac{|\{\boldsymbol{x} \in \mathbb{B}^n \mid \mathcal{K}(\boldsymbol{x} \mid m) \leq Bn - m\}|}{2^n}$$

holds.

The conditions of the lemma are pictured in Fig. 1.

Lemma 5 *Let $\mathfrak{G}$ be a symmetric game with the set of superpredictions S, let $B = \inf\{t \in \mathbb{R} \mid (t, t) \in S\}$, and let the boundary ∂S be represented by a twice differentiable curve in a vicinity of the point (B, B); let $\mathfrak{G}$ specify conditional predictive complexity $\mathcal{K}$. Let $\beta_0 \in (0, 1)$ be such that the curves $x/2 + y/2 = \beta_0^B$ and $\mathfrak{B}(\partial S)$ have a contact of the second order at the point (β_0^B, β_0^B). Then there are positive constants c and θ such that for every positive integer m and positive integer $n \geq cm^3$ the inequality*

$$\theta\beta_0^m \leq \frac{|\{\boldsymbol{x} \in \mathbb{B}^n \mid \mathcal{K}(\boldsymbol{x} \mid m) \leq Bn - m\}|}{2^n} \tag{13}$$

holds.

The statements of the lemmas and some details of the proofs are illustrated by Figs. 1 and 2.

Let us prove the lemmas.

PROOF of Lemma 4 Let $(\log_{\beta_0}(\beta_0^B - \Delta), \log_{\beta_0}(\beta_0^B + \Delta))$ be a superprediction. For every positive integer m we will construct a superloss process L_m that achieves

$$p(n,m) = \frac{|\{\boldsymbol{x} \in \mathbb{B}^n \mid L_m(\boldsymbol{x}) \leq Bn - m\}|}{2^n} \geq \frac{1}{4}\beta_0^m \tag{14}$$

for every $n \geq c_1 m + c_2$, where c_1 and c_2 are some constants independent of m or n.

In order to construct these superloss processes, we need the metaphor of a 'superstrategy'. Within this proof the word 'superstrategy' is taken to mean a prediction algorithm that on every trial outputs a superprediction and suffers corresponding loss. The total loss of a superstrategy is a superloss process.

Let $\mathfrak{A}$ be the superstrategy that always outputs the same superprediction $(\log_{\beta_0}(\beta_0^B - \Delta), \log_{\beta_0}(\beta_0^B + \Delta))$ and let $L(\boldsymbol{x})$ be the loss of this superstrategy. The superstrategy $\mathfrak{A}_m$ works as follows. It imitates $\mathfrak{A}$ as long as the inequality $L(\boldsymbol{x}) > B|\boldsymbol{x}| - m$ holds (note that the inequality is true for $\boldsymbol{x} = \Lambda$ because $m > 0$). After the inequality gets violated, the superstrategy switches to the superprediction (B, B). Let $L_m(\boldsymbol{x})$ be the loss of $\mathfrak{A}_m$. Put $A = B - \log_{\beta_0}(\beta_0^B + \Delta) > 0$ so that $(B|\boldsymbol{x}| - m) - L_m(\boldsymbol{x})$ does not exceed A. In other terms, $L(\boldsymbol{x})$ cannot jump over the threshold $B|\boldsymbol{x}| - m$ by more than A.

Let $M(\boldsymbol{x}) = \beta_0^{L(\boldsymbol{x}) - B|\boldsymbol{x}|}$ and $M_m(\boldsymbol{x}) = \beta_0^{L_m(\boldsymbol{x}) - B|\boldsymbol{x}|}$. These processes are martingales w.r.t. the Bernoulli distribution with the probability of success being equal to 1/2. The identity $(M(\boldsymbol{x}0) + M(\boldsymbol{x}1))/2 = M(\boldsymbol{x})$ implies that

$$\mathbf{E}M(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}) = \mathbf{E}M_m(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}) = 1$$

for every positive integer m, where $\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}$ are results of n independent Bernoulli trials with the probability of success being equal to 1/2. Since $\beta_0^{-A} = 1 + \Delta\beta_0^{-B} \leq 2$ we get $M_m(\boldsymbol{x}) \leq \beta_0^{-m-A} \leq 2\beta_0^{-m}$ for every $\boldsymbol{x} \in \mathbb{B}^m$.

Fix a positive integer m as in the statement of the theorem. Pick $\boldsymbol{x}$ of length n and consider the 'trajectories'

$$\langle 1, M(\boldsymbol{x}^{(1)}), M(\boldsymbol{x}^{(2)}), \ldots, M(\boldsymbol{x}^{(n)}) \rangle$$

and

$$\langle 1, M_m(\boldsymbol{x}^{(1)}), M_m(\boldsymbol{x}^{(2)}), \ldots, M_m(\boldsymbol{x}^{(n)}) \rangle \ .$$

Consider $\varepsilon > 0$ such that $\varepsilon < 1 \leq \beta_0^{-m}$. There are three mutually exclusive options:

(1) $M(\boldsymbol{x}^{(k)}) \geq \beta_0^{-m}$ for some $k \leq n$ and thus $\beta_0^{-m} \leq M_m(\boldsymbol{x}) \leq 2\beta_0^{-m}$.
(2) $M(\boldsymbol{x}^{(k)}) < \beta_0^{-m}$ for all $k \leq n$ and $M_m(\boldsymbol{x}) = M(\boldsymbol{x}) \leq \varepsilon$.
(3) $M(\boldsymbol{x}^{(k)}) < \beta_0^{-m}$ for all $k \leq n$ and $\beta_0^{-m} > M_m(\boldsymbol{x}) = M(\boldsymbol{x}) > \varepsilon$.

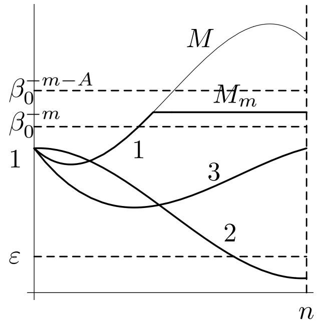

Fig. 3. Three options for trajectories from the proof of Lemma 4

These three options are shown in Fig. 3, where the values of $M(\boldsymbol{x}^{(k)})$ and $M_m(\boldsymbol{x}^{(k)})$ are plotted against values of k.

The expectation of $M_n(\boldsymbol{x})$ over all $\boldsymbol{x}$ of length n splits into the sum of three terms corresponding to the three classes of trajectories

$$1 = \mathbf{E}M_m(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}) = \Sigma_1 + \Sigma_2 + \Sigma_3 \ , \tag{15}$$

where $\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}$ are as above. The following bounds hold:

$$\begin{aligned}
&\Sigma_1 \leq 2\beta_0^{-m} \Pr\{M_m(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}) \geq \beta_0^{-m})\},\\
&\Sigma_2 \leq \varepsilon,\\
&\Sigma_3 \leq \beta_0^{-m} \Pr\{\varepsilon < M_m(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}) < \beta_0^{-m}\}\\
&\quad \leq \beta_0^{-m} \Pr\{M(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}) > \varepsilon\} \ .
\end{aligned}$$

The event $\{M_m(\boldsymbol{x}) \geq \beta_0^{-m})\}$ coincides with the event $\{L_m(\boldsymbol{x}) \leq Bn - m\}$ and thus $\Pr\{M_m(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}) \geq \beta_0^{-m})\} = p(n, m)$. If we denote the value

$$\begin{gathered}
\Pr\{M(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}) > \varepsilon\} = \\
\Pr\{L(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}) - Bn < \log_{\beta_0} \varepsilon\}
\end{gathered} \tag{16}$$

by $\alpha_\varepsilon(n)$, we obtain the inequality

$$1 \leq 2\beta_0^{-m} p(n, m) + \varepsilon + \beta_0^{-m} \alpha_\varepsilon(n) \tag{17}$$

and thus

$$p(n, m) \geq \frac{\beta_0^m}{2} - \frac{\varepsilon \beta_0^m}{2} - \frac{\alpha_\varepsilon(n)}{2} \ . \tag{18}$$

We will construct an upper bound on $\alpha_\varepsilon(n)$. The case $\Delta = \beta_0^B$ (i.e., the point appears on the line $x + y = 2\beta_0^B$ at the intersection with a coordinate axis) is trivial: the probability that $L(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)})$ is finite equals $1/2^n$ and this upper bound is sufficient for our purposes.

If $\Delta < \beta_0^B$ we will use the Chernoff bound in Hoeffding's form (see [Hoe63], Theor. 1). We need the following simple corollary. If $X_1, X_2, \ldots, X_n$ are independent Bernoulli trials with the probability of success equal to $p \in (0,1)$, $S = X_1 + X_2 + \cdots + X_n$, and $t \geq 0$, then

$$\Pr\left\{\frac{S}{n} - p \leq -t\right\} \leq e^{-2nt^2} \quad . \tag{19}$$

Let us 'straighten' the function $L(\boldsymbol{x}) - B|\boldsymbol{x}|$ in order to apply the Chernoff bound. The value

$$\begin{aligned} d &= \frac{(L(\boldsymbol{x}0) - B|\boldsymbol{x}0|) + (L(\boldsymbol{x}1) - B|\boldsymbol{x}1|)}{2} - (L(\boldsymbol{x}) - B|\boldsymbol{x}|) \\ &= \frac{1}{2}(\log_{\beta_0}(\beta_0^B - \Delta) + \log_{\beta_0}(\beta_0^B + \Delta) - 2B) \\ &= \frac{\ln\left(1 - \left(\frac{\Delta}{\beta_0^B}\right)^2\right)}{2\ln\beta_0} \\ &> 0 \end{aligned}$$

is independent of $\boldsymbol{x} \in \mathbb{B}^*$ so that $\mathbf{E}(L(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}) - Bn) = nd$; let $r = (\log_{\beta_0}(\beta_0^B - \Delta) - B) - d = d - (\log_{\beta_0}(\beta_0^B + \Delta) - B) > 0$. The function

$$S(\boldsymbol{x}) = \frac{L(\boldsymbol{x}) - B|\boldsymbol{x}| - d|\boldsymbol{x}|}{2r} + \frac{|\boldsymbol{x}|}{2}$$

can be treated as the sum of outcomes of independent Bernoulli trials with the probability of success equal to 1/2 and thus satisfies (19). By substituting the definition of $S(\boldsymbol{x})$ and $p = 1/2$ into (19) we get the inequality

$$\Pr\{L(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}) - Bn \leq dn - 2nrt\} \leq e^{-2nt^2} \quad , \tag{20}$$

which holds for all $t \geq 0$, where $\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}$ are results of independent Bernoulli trials with the probability of success equal to 1/2.

Let us now choose t and apply (20) in order to construct an upper bound on $\alpha_\varepsilon(n)$ defined by (16). Let $\log_{\beta_0}\varepsilon = dn - 2nrt$. If $n \geq \frac{1}{d}\log_{\beta_0}\varepsilon$, then $t(\varepsilon) = \frac{d}{2r} - \frac{1}{2nr}\log_{\beta_0}\varepsilon \geq 0$ and (20) can be applied. If, moreover, $n \geq \frac{2}{d}\log_{\beta_0}\varepsilon$ then $0 < d/(4r) \leq t \leq d/(2r)$ and $\alpha_\varepsilon(n) \leq e^{-2nt^2} \leq e^{-\frac{n}{8}\left(\frac{d}{r}\right)^2}$.

Fix $\varepsilon = 1/4$. We can achieve $\alpha_{1/4}(n) \leq \beta_0^m/4$ by taking

$$n \geq \max(\frac{2}{d}\log_{\beta_0}\frac{1}{4}, 8\left(\frac{r}{d}\right)^2(m\ln(1/\beta_0) + \ln 4)) \quad . \tag{21}$$

The substitution to (18) yields $p(n,m) \geq \beta_0^m/4$. We can take $c_1 = 8\left(\frac{r}{d}\right)^2 \ln\frac{1}{\beta_0}$ and $c_2 = \max\left(\frac{2}{d}\frac{\ln 4}{\ln(1/\beta_0)}, 8\left(\frac{r}{d}\right)^2\ln 4\right)$. $\square$

We are now moving on to the other lemma.

PROOF of Lemma 5

The proof is based upon the proof of Lemma 4.

For every small $\delta > 0$, consider the point $(\beta_0^B - \delta, \beta_0^B + \delta)$. It is not the image of a superprediction, but we still can define the function $L_m^{(\delta)}$ treating the point $(\log_{\beta_0}(\beta_0^B - \delta), \log_{\beta_0}(\beta_0^B + \delta))$ in the same fashion as we treated $(\log_{\beta_0}(\beta_0^B - \Delta), \log_{\beta_0}(\beta_0^B + \Delta))$ in the definition of L_m above. The processes $L_m^{(\delta)}$ are no longer superloss processes w.r.t. $\mathfrak{G}$, but for every fixed δ they still satisfy (14). We will use the notation $d(\delta)$, $r(\delta)$, $c_1(\delta)$, and $c_2(\delta)$ for numbers defined in the same fashion as d, r, c_1, and c_2 in the construction above.

We will use these processes to construct a superloss process $\tilde{L}_m$ and a constant $C > 0$ such that

$$p(n,m) = \frac{\left|\{\boldsymbol{x} \in \mathbb{B}^n \mid \tilde{L}_m(\boldsymbol{x}) \leq Bn - m + C\}\right|}{2^n} \geq \frac{1}{4}\beta_0^m \tag{22}$$

for every $n \geq \tilde{c}_1 m^3 + \tilde{c}_2$, where $\tilde{c}_1$ and $\tilde{c}_2$ are some constants independent of m, n.

Now take points on $\partial\mathfrak{B}_{\beta_0}(S)$ approximating $(\beta_0^B - \delta, \beta_0^B + \delta)$. For definiteness sake, for every small $\delta > 0$ let P_δ be the intersection of $x = \beta_0^B - \delta$ and the boundary $\partial\mathfrak{B}_{\beta_0}(S)$ (see Fig. 2). The distance between P_δ and $(\beta_0^B - \delta, \beta_0^B + \delta)$ is $O(\delta^3)$ as δ approaches 0. Let $\tilde{L}_m^{(\delta)}$ be the superloss process which uses the components of the superprediction $\mathfrak{B}_{\beta_0}^{-1}(P_\delta)$ exactly where $L_m^{(\delta)}$ uses numbers $\log_{\beta_0}(\beta_0^B - \delta)$ and $\log_{\beta_0}(\beta_0^B + \delta)$. Since

$$\begin{aligned}\log_{\beta_0}(\beta_0^B \pm \delta + O(\delta^3)) - \log_{\beta_0}(\beta_0^B \pm \delta) &= \log_{\beta_0}\left(1 + \frac{O(\delta^3)}{\beta_0^B \pm \delta}\right) \\ &= O(\delta^3)\end{aligned}$$

as $\delta \to 0$, there is $t > 0$ such that $|\tilde{L}_m^{(\delta)}(\boldsymbol{x}) - L_m^{(\delta)}(\boldsymbol{x})| \leq t|\boldsymbol{x}|\delta^3$ for every small positive δ and every string $\boldsymbol{x}$.

The superloss processes $\tilde{L}_m$ are constructed as follows. For every m we will choose $\delta(m) > 0$ and a positive integer $n_0(m)$. The process $\tilde{L}_m$ imitates $\tilde{L}_m^{(\delta(m))}$ as long as the length of the string is less than $n_0(m)$. The remaining 'tail' is provided by the trivial strategy predicting (B, B). As soon in the length $n_0(m)$ is reached, the superstrategy switches to (B, B). The problem is to choose $\delta(m)$ and $n_0(m)$ such that $\tilde{L}_m$ will still be close enough to $L_m^{(\delta(m))}(\boldsymbol{x})$ for strings of length up to $n_0(m)$ and the inequality (22) will be achieved.

It is easy to check that

$$d(\delta) = \frac{\ln\left(1 - \left(\frac{\delta}{\beta_0^B}\right)^2\right)}{2 \ln \beta_0} \qquad\qquad \sim \bar{d}\delta^2$$

and

$$r(\delta) = (\log_{\beta_0}(\beta_0^B - \delta) - B) - d(\delta) \ \sim \bar{r}\delta$$

as $\delta \to 0$, where $\bar{d}$ and $\bar{r}$ are some positive numbers. Thus we obtain the inequalities $c_1(\delta) \leq \bar{c}_1/\delta^2$ and $c_2(\delta) \leq \bar{c}_2/\delta^2$ for all sufficiently small δ and some positive constants $\bar{c}_1$ and $\bar{c}_2$.

Consider three inequalities

$$\begin{aligned} n_0(m) &\geq m\frac{\bar{c}}{\delta^2(m)} \ , \qquad &(23)\\ 1 &\geq tn_0(m)\delta^3(m) \ , &(24)\\ \eta &\geq \delta(m) \ . &(25) \end{aligned}$$

The first one, with $\bar{c} = \bar{c}_1 + \bar{c}_2$, implies that $n_0(m) \geq (c_1(\delta(m)) + c_2(\delta(m)))m \geq c_1(\delta(m))m + c_2(\delta(m))$; the second ensures that the difference $\tilde{L}_m^{(\delta(m))}(\boldsymbol{x}) - L_m^{(\delta(m))}(\boldsymbol{x})$ is bounded by a constant in the absolute value for strings $\boldsymbol{x}$ of length $|\boldsymbol{x}| = n_0(m)$; and the last one, where $\eta > 0$ is a small constant, guarantees that $\delta(m)$ is sufficiently small. Inequalities (23) and (24) imply

$$\sqrt{\frac{m\bar{c}}{n_0(m)}} \leq \delta(m) \leq \sqrt[3]{\frac{1}{tn_0(m)}} \ .$$

These two inequalities are consistent if and only if

$$\frac{m^3\bar{c}^3}{n_0^3(m)} \leq \frac{1}{t^2 n_0^2(m)} \ ,$$

i.e., $n_0(m) \geq m^3\bar{c}^3t^2$. Similarly, (23) and (25) imply

$$\sqrt{\frac{m\bar{c}}{n_0(m)}} \leq \delta(m) \leq \eta$$

and these inequalities are consistent if and only if $n_0(m) \geq m\bar{c}/\eta^2$. Let

$$\begin{aligned} n_0(m) &= \left\lceil \max\left\{ m^3\bar{c}^3t^2, \frac{m\bar{c}}{\eta^2} \right\} \right\rceil \\ \delta(m) &= \sqrt{\frac{m\bar{c}}{n_0(m)}} \ . \end{aligned}$$

The lemma follows. □

Remark 6 *If the boundary ∂S can be represented by a thrice differentiable curve in a vicinity of the point (B, B), the construction from the proof of Lemma 5 can be strengthened to show that (13) holds for all $n \geq cm^2$ for some c.*

Indeed, since the set $\mathfrak{B}_{\beta_0}(S)$ is symmetric in the straight line $x = y$, the contact between the boundary and the line $x + y = 2\beta^B$ at (β^B, β^B) has the third order. This observation implies that $|\tilde{L}_m^{(\delta(m))}(\boldsymbol{x}) - L_m^{(\delta(m))}(\boldsymbol{x})| \leq t|\boldsymbol{x}|\delta^4(m)$. Inequality (24) may thus be replaced by $1 \geq tn_0(m)\delta^4(m)$.

Acknowledgements

The authors were supported by the EPSRC through the grant GR/R46670 'Complexity Approximation Principle and Predictive Complexity: Analysis and Applications'.

The authors would like to thank anonymous COLT and *Information and Computation* reviewers for their detailed comments.

References

[CBFH$^+$97] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

[Egg58] H. G. Eggleston. *Convexity.* Cambridge University Press, Cambridge, 1958.

[HKW98] D. Haussler, J. Kivinen, and M. K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925, 1998.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

[KT75] S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes.* Academic Press, London, 1975.

[KVV04] Y. Kalnishkan, V. Vovk, and M. V. Vyugin. A criterion for the existence of predictive complexity for binary games. In *Algorithmic Learning Theory, 15th International Conference, ALT 2004, Proceedings*, volume 3244 of *Lecture Notes in Artificial Intelligence*, pages 249–263. Springer, 2004.

[LV97] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications.* Springer, New York, 1997.

[LW94] N. Littlestone and M. K. Warmuth. The Weighted Majority Algorithm. *Information and Computation*, 108:212–261, 1994.

[VW98] V. Vovk and C. J. H. C. Watkins. Universal portfolio selection. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 12–23, 1998.

[V'y94] V. V. V'yugin. Algorithmic entropy (complexity) of finite objects and its applications to defining randomness and amount of information. *Selecta Mathematica formerly Sovietica*, 13:357–389, 1994.

[Wil91] D. Williams. *Probability with Martingales.* Cambridge University Press, Cambridge, 1991.

[ZL70] A. K. Zvonkin and L. A. Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Math. Surveys*, 25:83–124, 1970.

Appendix A: Kolmogorov Complexity

In this appendix we briefly survey the definition of Kolmogorov complexity and the incompressibility property.

A *programming language* P is a partial computable function from $\mathbb{B}^* \times \mathbb{B}^*$ to $\mathbb{B}^*$. Informally, an argument $\boldsymbol{p}$ of $P(\boldsymbol{p}, \boldsymbol{y})$ is a program, $\boldsymbol{y}$ is an input and the value $P(\boldsymbol{p}, \boldsymbol{y})$ is the result of executing of program p on input y; this result may be undefined. Given a programming language P, one may define complexity of $\boldsymbol{x}$ given $\boldsymbol{y}$ w.r.t. P by the equation

$$\mathrm{K}_P(\boldsymbol{x} \mid \boldsymbol{y}) = \min\{n \mid \exists p \in \mathbb{B}^n \ : P(\boldsymbol{p}, \boldsymbol{y}) = \boldsymbol{x}\} \ , \tag{26}$$

where $\min(\varnothing) = +\infty$ by definition.

A fundamental theorem of Kolmogorov's (see any of [ZL70,V'y94,LV97]) states that there is a universal programming language U, i.e., a language such that for every P there is a constant $C > 0$ such that for every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{B}^*$ we have

$$\mathrm{K}_U(\boldsymbol{x} \mid \boldsymbol{y}) \leq \mathrm{K}_P(\boldsymbol{x} \mid \boldsymbol{y}) + C.$$

Clearly, the difference between the two complexities specified by universal programming languages is bounded by a constant. We may pick one universal programming language U and define *conditional (plain) Kolmogorov complexity* $\mathrm{K} = \mathrm{K}_U$.

Unconditional Kolmogorov complexity can be defined by $\mathrm{K}(\boldsymbol{x}) = \mathrm{K}(\boldsymbol{x} \mid \Lambda)$, where Λ is the empty string. We have $\mathrm{K}(\boldsymbol{x} \mid \boldsymbol{y}) \leq \mathrm{K}(\boldsymbol{x}) + C$ for some constant C.

Proposition 7 (Incompressibility Property)

(i) There is a constant C such that for every $\boldsymbol{x} \in \mathbb{B}^*$ the inequality

$$\mathrm{K}(\boldsymbol{x}) \leq |\boldsymbol{x}| + C \tag{27}$$

holds.

(ii) For every positive integer n and every $m < n$ we have

$$|\{\boldsymbol{x} \mid |\boldsymbol{x}| = n \text{ and } \mathrm{K}(\boldsymbol{x}) \leq n - m\}| \leq 2^{n-m+1} \ . \tag{28}$$

This statement can be found in any of the sources [ZL70,V'y94,LV97]. For completeness, we give a short proof.

PROOF. The proof of (i) is by considering the programming language which performs the identity mapping. The statement (ii) follows from the observation that there can be no more then 2^s strings of complexity s since each of them is generated by its own program of length s. □

The bound in (ii) is tight since the following holds:

Proposition 8 *There is a positive constant θ such that for all positive integers n and $m < n$ we have*

$$|\{\boldsymbol{x} \mid |\boldsymbol{x}| = n \text{ and } \mathrm{K}(\boldsymbol{x} \mid m) \leq n - m\}| \geq \theta 2^{n-m+1} \ . \tag{29}$$

PROOF Sketch Consider the function $P_m(\boldsymbol{y}) := 0^m\boldsymbol{y}$ which transforms any string $\boldsymbol{y}$ of length $n - m$ into the string $P_m(\boldsymbol{y})$ of length n, and $K(P_m(\boldsymbol{y}) \mid m) \leq n - m + C$ for some constant C independent of $\boldsymbol{y}$, n and m. The number of $\boldsymbol{y}$'s of length $n - m$ is 2^{n-m}. To obtain the statement of theorem, replace m by $m + C$ and take $\theta = 2^{-C-1}$. We omit some technical details needed because we should change m in the condition for $m + C$. □

Appendix B: Martingales

Here we briefly review a general definition of a *(super)martingale*, adapt it to our special case, and formulate an inequality necessary for the derivation of the incompressibility property.

We are going to use (more or less) the terminology and notation from [Wil91]. Throughout this appendix Ω refers to a *sample space*; its elements $\omega \in \Omega$ are *sample points*. A *filtered space* is a quadruple $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, \Pr)$ where $\mathcal{F}$ is a σ-algebra on Ω, σ-algebras $\mathcal{F}_n$, $n = 0, 1, 2, \ldots$, are sub-σ-algebras of $\mathcal{F}$ such that

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F} \ , \tag{30}$$

and $\Pr$ is a probability measure on $(\Omega, \mathcal{F})$. A sequence of random variables $X_0, X_1, X_2, \ldots$ on Ω is a *martingale* w.r.t. $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, \Pr)$ if for every $n = 0, 1, 2, \ldots$ the variable X_n is measurable w.r.t. $\mathcal{F}_n$, and for every $n \geq 1$ we have

- $\mathbf{E}_{\Pr}(|X_n|) < +\infty$, and
- $\mathbf{E}_{\Pr}(X_n \mid \mathcal{F}_{n-1}) = X_{n-1}$.

In the definition of a supermartingale the last condition should be replaced by $\mathbf{E}_{\Pr}(X_n \mid \mathcal{F}_{n-1}) \leq X_{n-1}$. In the above expressions $\mathbf{E}_{\Pr}$ stands for the expectation taken w.r.t. the probability Pr.

Non-negative martingales satisfy Doob's inequality (see, e.g., [Wil91]); we need a version of this inequality for supermartingales. The following statement may be found, e.g., in [KT75] (Lemma 5.2):

Proposition 9 *If non-negative random variables $Z_0, Z_1, Z_2, \ldots$ form a supermartingale w.r.t.* $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, \Pr)$*, then for every $c > 0$ and positive integer n we have*

$$\Pr\left\{\max_{k=0,1,2,\ldots} Z_k \geq c\right\} \leq \frac{\mathbf{E}Z_0}{c} \ .$$

Consider the case of the Bernoulli distribution with the probability of 1 equal to p. The sample space is the set of all infinite binary strings $\mathbb{B}^\infty$. The σ-algebra $\mathcal{F}$ is generated by all cylinders $\Gamma_{\boldsymbol{x}}$, $\boldsymbol{x} \in \mathbb{B}^*$, where

$$\Gamma_{\boldsymbol{x}} = \{\boldsymbol{x}\tau \mid \tau \in \mathbb{B}^\infty\} \ .$$

For every $n = 0, 1, 2, \ldots$, the σ-algebra $\mathcal{F}_n$ is generated by the cylinders $\Gamma_{\boldsymbol{x}}$ such that $|\boldsymbol{x}| = n$. A function measurable w.r.t. $\mathcal{F}_n$ may be identified with a function defined on $\mathbb{B}^n$. Thus a sequence of random variables $X_0, X_1, X_2, \ldots$ such that X_n is measurable w.r.t. $\mathcal{F}_n$, $n = 0, 1, 2, \ldots$, may be identified with a function $L : \mathbb{B}^* \to \mathbb{R}$. In order to be a martingale, it should satisfy the condition $pL(\boldsymbol{x}1) + (1-p)L(\boldsymbol{x}0) = L(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathbb{B}^*$. If for every $\boldsymbol{x} \in \mathbb{B}^*$ we have $pL(\boldsymbol{x}1) + (1-p)L(\boldsymbol{x}0) \leq L(\boldsymbol{x})$, it is a supermartingale.