Loss Functions, Complexities, and the Legendre Transformation [1]

Yuri Kalnishkan [a], Volodya Vovk [a], and Michael V. Vyugin [a]

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

Abstract

The paper introduces a way of re-constructing a loss function from predictive complexity. We show that a loss function and expectations of the corresponding predictive complexity w.r.t. the Bernoulli distribution are related through the Legendre transformation. It is shown that if two loss functions specify the same complexity then they are equivalent in a strong sense. The expectations are also related to the so called generalized entropy.

1 Introduction

In this paper we consider the on-line prediction framework. A prediction algorithm tries to predict outcomes $\omega_1, \omega_2, \ldots, \omega_n$ that occur one after another. Each time before observing the outcome ω_i the algorithm outputs a prediction γ_i. We assume that the ranges of outcomes and predictions, Ω and Γ, are some sets fixed in advance.

To measure the discrepancy between predictions and outcomes we use a loss function $\lambda(\omega, \gamma)$. The performance of an algorithm on a sequence of outcomes $\omega_1, \omega_2, \ldots, \omega_n$ is measured by the cumulative loss $\sum_{i=1}^{n} \lambda(\omega_i, \gamma_i)$. Consider the problem of formalising the difficulty of predicting elements of a sequence $\omega_1, \omega_2, \ldots, \omega_n$ irrespective of a particular prediction algorithm. This could

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

[1] A previous version of this paper appeared in proceedings of *The Twelfth International Conference on Algorithmic Learning Theory*, volume 2225 of Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, Heidelberg, 2001

have been done if we had at our disposal a certain *universal* prediction algorithm that suffers minimal possible loss. The loss of such an algorithm could have been treated as the intrinsic difficulty of predicting a sequence.

Unfortunately no natural universal algorithm exists in most cases. It is easy to see that for every prediction algorithm there is another algorithm that suffers much smaller loss on some sequences. However the difficulty of predicting can be formalised by the concept of *predictive complexity*. Intuitively, predictive complexity is the loss of a "strategy" that is allowed to work infinitely long but can be approached in the limit. The loss suffered by any actual prediction algorithm on a sequence is at least the predictive complexity of the sequence up to an additive constant. Predictive complexity may be considered as an inherent measure of "learnability" of a string in the same way as Kolmogorov complexity reflects the "simplicity" of a string.

Predictive complexity was introduced in [VW98]. The universal "strategy" is constructed as a mixture of ordinary prediction strategies and thus the theory of predictive complexity is a natural development of prediction with expert advice (cf. [CBFH$^+$97,HKW98,LW94]).

This paper addresses the problem of relations between a loss function and the corresponding predictive complexity (we fix the set of outcomes Ω to be $\{0, 1\}$). Suppose that there exists predictive complexity $\mathcal{K}$ specified by a loss function λ (this is not always the case as some loss functions do not specify complexities at all; however many natural loss functions such as the squared deviation do). Can the same complexity be specified by another loss function or can we recover the loss function from predictive complexity?

We solve this problem considering the "complexity per element" $\frac{1}{n}\mathcal{K}(\zeta)$, where ζ is a string of n elements distributed according to the Bernoulli law and n is large. We show that $\frac{1}{n}\mathcal{K}(\zeta)$ and the loss function λ are related through the Legendre transformation (Appendix B contains a brief introduction to the theory of the Legendre transformation in the one-dimensional case).

We show that if two loss functions specify the same complexity then they are equivalent in a very strong sense, namely, they are mere parameterisations of the same geometrical image. This observation allows us to show that the variants of Kolmogorov complexity, namely, plain, prefix, and monotone, do not correspond to any game and thus are not predictive complexities while another variant of Kolmogorov complexity, the minus logarithm of Levin's a priori semimeasure, is known to be the predictive complexity specified by the logarithmic game.

2 Preliminaries

2.1 Games and Superpredictions

A *game* $\mathfrak{G}$ is a triple $\langle \Omega, \Gamma, \lambda \rangle$, where Ω is called an *outcome space*, Γ stands for a *prediction space*, and $\lambda : \Omega \times \Gamma \to \mathbb{R} \cup \{+\infty\}$ is a *loss function*. We suppose that a definition of computability over Ω and Γ is given and λ is computable according to this definition.

In this paper we are interested in the binary case $\Omega = \mathbb{B} = \{0, 1\}$. We will denote elements of $\mathbb{B}^*$ (i.e., finite strings of elements of $\mathbb{B}$) by bold letters, e.g., $\boldsymbol{x}, \boldsymbol{y}$. The length (i.e. the number of elements) of a string $\boldsymbol{x}$ is denoted by $|\boldsymbol{x}|$. The number of zeros in $\boldsymbol{x}$ is denoted by $\sharp_0 \boldsymbol{x}$ and the number of ones is denoted by $\sharp_1 \boldsymbol{x}$. We denote logarithm to the base 2 by log without a subscript.

We impose the following restrictions on games in order to exclude degenerate cases:

(1) The set of possible predictions Γ is a compact topological space.
(2) For every $\omega \in \Omega$, the function $\lambda(\omega, \gamma)$ is continuous (w.r.t. the standard topology of $\mathbb{R} \cup \{+\infty\}$) in the second argument.
(3) There exists $\gamma \in \Gamma$ such that, for every $\omega \in \Omega$ the inequality $\lambda(\omega, \gamma) < +\infty$ holds.
(4) If there are $\gamma_0 \in \Gamma, \omega_0 \in \Omega$ such that $\lambda(\omega_0, \gamma_0) = +\infty$, then there is a sequence of $\gamma_n \in \Gamma$, $n = 1, 2, \ldots$, such that $\gamma_n \to \gamma_0$ as $n \to +\infty$ and $\lambda(\omega_0, \gamma_n) < +\infty$.

If a game satisfies these conditions, we call it *regular*.

Conditions 1–3 have been taken from [Vov98]. Condition 2 can in fact be derived from computability of λ because most natural definitions of computability imply continuity. Condition 3 prohibits some degenerated games. Condition 4 essentially means that λ accepts the infinite value only in exceptional situations that can be approximated by finite cases. Appendix A discusses one more aspect of Condition 4.

We say that a pair $(s_0, s_1) \in [-\infty, +\infty]^2$ is a *superprediction* if there exists a prediction $\gamma \in \Gamma$ such that $s_0 \geq \lambda(0, \gamma)$ and $s_1 \geq \lambda(1, \gamma)$. If we let $P = \{(p_0, p_1) \in [-\infty, +\infty]^2 \mid \exists \gamma \in \Gamma : p_0 = \lambda(0, \gamma) \text{ and } p_1 = \lambda(1, \gamma)\}$ (cf. the canonical form of a game in [Vov90]), the set S of all superpredictions is the set of points that lie "north-east" of some point in P.

The set of superpredictions for a regular game is closed. This follows from Conditions 1 and 2. Even a stronger statement is true: the set $S \subseteq [-\infty, +\infty]^2$

is a closure of its finite part $S \cap \mathbb{R}^2$. This follows from Condition 4.

Condition 3 implies that S contains finite points.

Let us describe the intuition behind the concept of a game. Consider a prediction algorithm $\mathfrak{A}$ working according to the following protocol:

```
for $t = 1, 2, \dots$
   (1) $\mathfrak{A}$ chooses a prediction $\gamma_t \in \Gamma$
   (2) $\mathfrak{A}$ observes the actual outcome $\omega_t \in \Omega$
   (3) $\mathfrak{A}$ suffers loss $\lambda(\omega_t, \gamma_t)$
end for
```

Over the first T trials, $\mathfrak{A}$ suffers the total loss

$$\mathrm{Loss}_{\mathfrak{A}}(\omega_1, \omega_2, \dots, \omega_T) = \sum_{t=1}^{T} \lambda(\omega_t, \gamma_t) \ . \tag{1}$$

By definition, put $\mathrm{Loss}_{\mathfrak{A}}(\Lambda) = 0$, where Λ denotes the empty string.

The function $\mathrm{Loss}_{\mathfrak{A}}(\boldsymbol{x})$ can be treated as the predictive complexity of $\boldsymbol{x}$ in the game $\mathfrak{G}$ w.r.t. $\mathfrak{A}$. We will call these functions *loss processes*. Unfortunately, the set of loss processes has no minimal elements except in degenerate cases. The set of loss processes should be extended to the set of superloss processes.

2.2 Superloss Processes and Predictive Complexity

Take a game $\mathfrak{G}$. A function $L : \Omega^* \to \mathbb{R} \cup \{+\infty\}$ is called a *superloss process* w.r.t. $\mathfrak{G}$ (see [VW98]) if the following conditions hold:

- $L(\Lambda) = 0$,
- for every $\boldsymbol{x} \in \Omega^*$, the pair $(L(\boldsymbol{x}0) - L(\boldsymbol{x}), L(\boldsymbol{x}1) - L(\boldsymbol{x}))$ is a superprediction w.r.t. $\mathfrak{G}$, and
- L is semicomputable from above.

We will say that a superloss process K is *universal* if for any superloss process L there exists a constant C such that $\forall \boldsymbol{x} \in \Omega^* : K(\boldsymbol{x}) \leq L(\boldsymbol{x}) + C$. The difference between two universal superloss processes w.r.t. $\mathfrak{G}$ is bounded by a constant. If universal superloss processes w.r.t. $\mathfrak{G}$ exist, we may pick one and denote it by $\mathcal{K}^{\mathfrak{G}}$. It follows from the definition that, for every prediction algorithm $\mathfrak{A}$, there is a constant C such that for every $\boldsymbol{x}$ we have $\mathcal{K}^{\mathfrak{G}}(\boldsymbol{x}) \leq \mathrm{Loss}_{\mathfrak{A}}^{\mathfrak{G}}(\boldsymbol{x}) + C$, where $\mathrm{Loss}^{\mathfrak{G}}$ denotes the loss w.r.t. $\mathfrak{G}$. One may call $\mathcal{K}^{\mathfrak{G}}$ *(predictive) complexity* w.r.t. $\mathfrak{G}$.

It is worth mentioning that the regularity conditions are not restrictive from the point of view of predictive complexity. It is shown in Appendix A that a game that does not satisfy Condition 4 will not specify predictive complexity.

2.3 Mixability and the Existence of Predictive Complexity

Mixability was introduced in [Vov98,VW98]. Take a parameter $\beta \in (0,1)$ and consider the homeomorphism $\mathfrak{B}_\beta : (-\infty,+\infty]^2 \to [0,+\infty)^2$ specified by

$$\mathfrak{B}_\beta(x,y) = (\beta^x, \beta^y) \ . \tag{2}$$

A regular 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 $\beta \in (0,1)$.

It can be shown that if a game $\mathfrak{G}$ is β-mixable, $L_1, L_2, \ldots$ is an effective sequence of superloss processes w.r.t. $\mathfrak{G}$ and $p_1, p_2, \ldots \in [0,1]$ is a computable sequence of weights such that $\sum_{i=1}^{+\infty} p_i = 1$, then there is a superloss process L_0 such that

$$L_0(\boldsymbol{x}) \leq L_i(\boldsymbol{x}) + \frac{\ln(1/p_i)}{\ln(1/\beta)} \tag{3}$$

for each $i = 1, 2, \ldots$. This was proved in [VW98] as a part of the proof of the following statement:

Proposition 1 ([VW98]) *If a game $\mathfrak{G}$ with the set of superpredictions S is mixable then there is predictive complexity w.r.t. $\mathfrak{G}$.*

Examples of mixable games are the logarithmic game with $\Gamma = [0,1]$ and

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

and the square-loss game with $\Gamma = [0,1]$ and $\lambda(\omega,\gamma) = (\omega-\gamma)^2$. They specify the logarithmic complexity $\mathcal{K}^{\log}$ and the square-loss complexity $\mathcal{K}^{\mathrm{sq}}$, respectively (see [VW98]). Logarithmic complexity coincides with the negative logarithm of Levin's a priori semimeasure (see [V'y94,LV97] for the definition). The negative logarithm of Levin's a priori semimeasure is a variant of Kolmogorov complexity. Thus we may say that Kolmogorov complexity is a special case of predictive complexity.

3 Convergence to the Entropy

For each game $\langle \Omega, \Gamma, \lambda \rangle$ we define its *generalized entropy* to be the function

$$H(p) = \inf_{\gamma \in \Gamma} ((1-p)\lambda(0,\gamma) + p\lambda(1,\gamma)), \quad p \in [0,1]$$

(cf. [GD02]). In the case of the logarithmic game, the generalized entropy coincides with the regular entropy. The entropy for the square-loss game corresponds to Brier entropy from [GD02].

The following theorem shows the connections between the loss functions and complexities.

Theorem 2 *Let $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ be a mixable game with generalized entropy H and predictive complexity $\mathcal{K}$. Then for every $p \in (0,1)$*

$$\lim_{n \to +\infty} \frac{\mathcal{K}(\xi_1^{(p)} \ldots \xi_n^{(p)})}{n} = H(p) \quad a.s. \ , \tag{4}$$

where $\xi_1^{(p)}, \xi_2^{(p)}, \ldots$ are results of independent Bernoulli trials with the probability of 1 being equal to p.

PROOF. Fix $p \in (0,1)$ and let $\varepsilon > 0$. First we prove that

$$\frac{\mathcal{K}(\xi_1^{(p)} \ldots \xi_n^{(p)})}{n} < H(p) + \varepsilon \tag{5}$$

from some n on. Let γ_0 be a computable prediction such that

$$(1-p)\lambda(0,\gamma_0) + p\lambda(1,\gamma_0) < H(p) + \varepsilon/2 \tag{6}$$

(Condition 2 on p. 3 implies that the set

$$\{\gamma \mid (1-p)\lambda(0,\gamma) + p\lambda(1,\gamma) < H(p) + \varepsilon/2\}$$

is open; therefore, it contains a computable element). By the definition of predictive complexity and the Borel strong law of large numbers, we have with probability one:

$$\begin{align}
\mathcal{K}(\xi_1^{(p)} \ldots \xi_n^{(p)}) &\leq \lambda(0,\gamma_0)\sharp_0(\xi_1^{(p)} \ldots \xi_n^{(p)}) + \lambda(1,\gamma_0)\sharp_1(\xi_1^{(p)} \ldots \xi_n^{(p)}) + O(1) \tag{7}\\
&\leq \lambda(0,\gamma_0)((1-p)n + o(n)) + \lambda(1,\gamma_0)(pn + o(n)) + O(1) \tag{8}\\
&< (H(p) + \varepsilon/2)n + o(n) \ ; \tag{9}
\end{align}$$

this implies (5).

It remains to prove that

$$\frac{\mathcal{K}(\xi_1^{(p)} \dots \xi_n^{(p)})}{n} > H(p) - \varepsilon$$

from some n on. Consider a superprediction (s_0, s_1). By definition there is $\gamma \in \Gamma$ such that $s_0 \geq \lambda(0, \gamma)$ and $s_1 \geq \lambda(1, \gamma)$ and thus for every $p \in (0, 1)$ we get $(1-p)s_0 + ps_1 \geq H(p)$. Since

$$\left(\mathcal{K}(\xi_1^{(p)} \dots \xi_{n-1}^{(p)} 0) - \mathcal{K}(\xi_1^{(p)} \dots \xi_{n-1}^{(p)}), \mathcal{K}(\xi_1^{(p)} \dots \xi_{n-1}^{(p)} 1) - \mathcal{K}(\xi_1^{(p)} \dots \xi_{n-1}^{(p)})\right)$$

is a superprediction,

$$\mathbf{E}\left(\eta_n \mid \xi_1^{(p)} \dots \xi_{n-1}^{(p)}\right) \geq H(p) \ , \tag{10}$$

where

$$\eta_n = \mathcal{K}(\xi_1^{(p)} \dots \xi_n^{(p)}) - \mathcal{K}(\xi_1^{(p)} \dots \xi_{n-1}^{(p)}) \ .$$

Now we are going to apply the martingale strong law of large numbers. This law can be found in [Shi96] as Theorem VII.5.4. Here we formulate a special case that is sufficient for the purposes of this paper:

Proposition 3 *Let $\xi_1, \xi_2, \dots$ be some sequence of random variables and let $f_1, f_2, \dots$ be a sequence of functions such that f_n is a measurable function of n arguments and*

$$\mathbf{E}(f_n(\xi_1, \dots, \xi_{n-1}, \xi_n) \mid \xi_1, \dots, \xi_{n-1}) = 0 \quad a.s. \tag{11}$$

for all $n = 1, 2, \dots$. If

$$\sum_{n=1}^{+\infty} \frac{\mathbf{E}(f_n^2(\xi_1, \dots, \xi_{n-1}, \xi_n) \mid \xi_1, \dots, \xi_{n-1})}{n^2} < +\infty \quad a.s. \ , \tag{12}$$

then

$$\frac{1}{n} \sum_{i=1}^{n} f_i(\xi_1, \dots, \xi_i) \to 0 \quad as \quad n \to +\infty \quad a.s. \tag{13}$$

In order to show that (12) holds for $f_n(\xi_1^{(p)} \dots \xi_n^{(p)}) = \eta_n - \mathbf{E}\left(\eta_n \mid \xi_1^{(p)} \dots \xi_{n-1}^{(p)}\right)$, $n = 1, 2, \dots$, we need the following lemma.

Lemma 4 *If $\mathcal{K}$ is predictive complexity w.r.t. a mixable game $\mathfrak{G}$, then there is a positive constant c such that*

$$|\mathcal{K}(\boldsymbol{x}b) - \mathcal{K}(\boldsymbol{x})| \leq c \ln n$$

for all $n = 1, 2, \dots$ and $\boldsymbol{x} \in \mathbb{B}^n$ and $b \in \mathbb{B}$.

PROOF of Lemma 4. Take a superprediction $(s,s) \in S \cap \mathbb{R}^2$ and consider the superloss processes L_n, where $n = 1, 2, \ldots$, defined as follows. For every $\boldsymbol{x}$ such that $|\boldsymbol{x}| \le n$, we let $L_n(\boldsymbol{x}) = \mathcal{K}(\boldsymbol{x})$ while for each $\boldsymbol{x}$ of length n and $b \in \mathbb{B}$ we let $L_n(\boldsymbol{x}b) = \mathcal{K}(\boldsymbol{x}) + s$ (the behaviour of L_n on strings longer than $n+1$ is of no importance). Since the game is mixable, we can take the sequence $p_n = 6/(\pi^2 n^2)$, $n = 1, 2, \ldots$ and by using (3) obtain a superloss process L_0 and a constant $a > 0$ such that

$$L_0(\boldsymbol{x}) \le L_n(\boldsymbol{x}) + a \ln n \ .$$

The observation that $\mathcal{K}(\boldsymbol{x}) \le L_0(\boldsymbol{x}) + C$ for some constant C completes that proof. □

We can now apply Proposition 3:

$$\frac{\mathcal{K}(\xi_1^{(p)} \ldots \xi_n^{(p)})}{n} = \frac{1}{n} \sum_{i=1}^{n} \eta_i \tag{14}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}\left(\eta_i \mid \xi_1^{(p)} \ldots \xi_{i-1}^{(p)})\right) + o(1) \tag{15}$$

$$\ge H(p) + o(1) \tag{16}$$

with probability one ((10 was used to obtain the last line). This completes the proof. □

In the special case of the logarithmic game Theorem 2 is a well-known result in the theory of Kolmogorov complexity (cf. [LV97], Exercise 2.8.3). In combination with Lebesgue's theorem it implies

Corollary 5 *Let $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ be a mixable game with generalized entropy H, $p \in (0,1)$, and $\xi_1^{(p)}, \xi_2^{(p)}, \ldots$ be results of independent Bernoulli trials with the probability of 1 equal to p. Then*

$$\lim_{n \to +\infty} \frac{\mathcal{K}(\xi_1^{(p)} \ldots \xi_n^{(p)})}{n} = H(p)$$

in L_1 and

$$\lim_{n \to +\infty} \frac{\mathbf{E}\mathcal{K}(\xi_1^{(p)} \ldots \xi_n^{(p)})}{n} = H(p) \ .$$

Note that in the proof of Theorem 2 we used mixability only in Lemma 4. One can see from the proof (by taking expectations of (7) and (14) and applying (6) and (10)) that if we do not postulate mixability it is still possible to show the following.

Theorem 6 *Let $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ be a regular game with generalized entropy H, $p \in (0,1)$, and $\xi_1^{(p)}, \xi_2^{(p)}, \ldots$ be results of independent Bernoulli trials with the probability of 1 equal to p. Then*

$$\lim_{n \to +\infty} \frac{\mathbf{E}\mathcal{K}(\xi_1^{(p)} \ldots \xi_n^{(p)})}{n} = H(p) \ .$$

Consider a regular game with the set of superpredictions S. Take the function $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ defined by the formula

$$f(x) = \inf\{y \mid (x,y) \in S\} \tag{17}$$

for each real x (here we let $\inf \varnothing = +\infty$). Since the game satisfies the conditions, the real part of S coincides with the epigraph $\{(x,y) \in \mathbb{R}^2 \mid y \geq f(x)\}$ of f and thus f uniquely determines $S \cap \mathbb{R}^2$, which in turn uniquely determines S. Note that we need Condition 4 to claim that S can be reconstructed from its finite part $S \cap \mathbb{R}^2$.

It turns out that the generalized entropy H can be defined using the Legendre transformation of f. Recall that Appendix B contains an overview of the Legendre transformation. In order to apply the Legendre transformation, we should make sure that f is convex and closed. Convexity is implied by the following lemma while closeness follows from Conditions 1–4.

Lemma 7 *If a regular game $\mathfrak{G}$ specifies predictive complexity, then the intersection of its set of superpredictions S and $\mathbb{R}^2$ is convex.*

The proof is in Appendix C.

Now we can state the expression of H in terms of f.

Proposition 8 *Let $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ be a regular game with generalized entropy H. Let $\mathfrak{G}$ specify complexity $\mathcal{K}$ and let S be the set of superpredictions for $\mathfrak{G}$. Then for every $p \in (0,1)$*

$$H(p) = -pf^* \left(\frac{p-1}{p} \right) \ ,$$

holds, where f^* is the function conjugate to f specified by (17).

PROOF. It suffices to notice that

$$\begin{aligned} H(p) &= \inf_{\gamma\in\Gamma}((1-p)\lambda(0,\gamma)+p\lambda(1,\gamma)) \\ &= \inf_{(x,y)\in S}[(1-p)x+py] \\ &= \inf_{x\in\mathbb{R}}[(1-p)x+pf(x)] \\ &= -p\sup_{x\in\mathbb{R}}\left[\frac{p-1}{p}x-f(x)\right] \\ &= -pf^*\left(\frac{p-1}{p}\right) \quad . \end{aligned}$$

□

Corollary 9 *Let $\mathfrak{G}_1$ and $\mathfrak{G}_2$ be regular games. Suppose they have the sets of superpredictions S_1 and S_2 and specify complexities $\mathcal{K}^1$ and $\mathcal{K}^2$. If there is a function $\delta(n)=o(n)$ as $n\to+\infty$ such that for every $\boldsymbol{x}\in\mathbb{B}^*$ the inequality*

$$|\mathcal{K}^1(\boldsymbol{x})-\mathcal{K}^2(\boldsymbol{x})|\le\delta(|\boldsymbol{x}|) \tag{18}$$

holds, then $S_1=S_2$ and complexities $\mathcal{K}^1$ and $\mathcal{K}^2$ are equal up to a constant.

PROOF. For every $p\in(0,1)$ we have

$$\lim_{n\to+\infty}\left|\frac{\mathbf{E}\left[\mathcal{K}^1(\xi_1^{(p)}\dots\xi_n^{(p)})-\mathcal{K}^2(\xi_1^{(p)}\dots\xi_n^{(p)})\right]}{n}\right|\le\frac{\delta(n)}{n}=o(1) \tag{19}$$

as $n\to+\infty$, where $\xi_1^{(p)},\dots,\xi_n^{(p)}$ are as above. This implies that for every $p\in(0,1)$ the equality $\tilde{f}_1(p)=\tilde{f}_2(p)$ holds, where $\tilde{f}_1$ and $\tilde{f}_2$ are defined for the games $\mathfrak{G}_1$ and $\mathfrak{G}_2$ as the limit in (4).

Let f_1 and f_2 be defined for the games $\mathfrak{G}_1$ and $\mathfrak{G}_2$ by (17). Consider the conjugated functions f_1^* and f_2^*. The equalities we have for $\tilde{f}_1$ and $\tilde{f}_2$ together with Theorem 6 and Proposition 8 imply that $f_1^*(t)=f_2^*(t)$ for all $t\in(-\infty,0)$. For every $t>0$ the equality $f_1^*(t)=f_2^*(t)=+\infty$ holds. Since f_1^* and f_2^* are closed, we have $f_1^*(0)=f_2^*(0)$.

It follows from a fundamental property of conjugate functions, namely, $f^{**}=f$ (Proposition 11 from Appendix B), that the functions f_1 and f_2 coincide. □

Corollary 10 *There is no regular game specifying plain Kolmogorov complexity* K*, prefix complexity* KP*, or monotone complexity* Km *as its predictive complexity.*

PROOF. The difference between any of these functions and the negative logarithm of Levin's a priori semimeasure is bounded by a term of logarithmic order of the length of a string. If K is of the complexities K, KP, or Km, then there is a constant $c > 0$ such that the inequality $|K(\boldsymbol{x}) - \mathrm{KM}(\boldsymbol{x})| \leq c \log |\boldsymbol{x}|$ holds (see [V'y94,LV97]).

As we mentioned above, the function KM coincides with $\mathcal{K}^{\log}$, which is complexity w.r.t. the logarithmic game (see [VW98]). If K is predictive complexity w.r.t. a game $\mathfrak{G}$, we can apply Corollary 9. Hence the set of superpredictions for $\mathfrak{G}$ coincides with the set of superpredictions for the logarithmic game and the absolute value of the difference $K(\boldsymbol{x}) - \mathrm{KM}(x)$ is bounded by a constant.

However neither of the differences between these functions and KM can be bounded by a constant (see [V'y94,LV97]). □

4 Acknowledgements

V. Vovk and Y. Kalnishkan have been supported by the EPSRC through the grants GR/M14937 and GR/R46670. Y. Kalnishkan and M. V. Vyugin have also been supported by Overseas Research Students Awards Scheme grants. The authors are grateful to anonymous ALT and TCS reviewers for their insightful and detailed comments.

References

[ATF87] V. M. Alekseev, V. M. Tikhomirov, and S. V. Fomin. *Optimal Control.* Plenum, New York, 1987.

[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.

[GD02] P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. Technical Report 223, Department of Statistical Sciences, University College, London, February 2002.

[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.

[LV97] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications.* Springer, New York, 2nd edition, 1997.

[LW94] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

[Roc70] R. T. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

[RV73] A. W. Roberts and D. E. Varberg. *Convex Functions.* Academic Press, 1973.

[Shi96] A. N. Shiryaev. *Probability.* Springer, New York, 2nd edition, 1996.

[Vov90] V. Vovk. 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.

[Vov98] V. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56:153–173, 1998.

[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.

Appendix A: A Note on Games that Are not Regular

Let us prove that a game that does not satisfy Condition 4 does not specify predictive complexity. Assume the converse. Consider a game that satisfies Conditions 1–3 but does not satisfy Condition 4 and let $\mathcal{K}$ be complexity w.r.t. this game.

Let S be the set of superpredictions for this game. At least one of the following cases is true:

(1) there are numbers a and $\Delta > 0$ such that S contains the point $(a, +\infty)$ but for every finite $(x, y) \in S$ we have $x \geq a + \Delta$, and
(2) there are numbers a and $\Delta > 0$ such that S contains the point $(+\infty, a)$ but for every finite $(x, y) \in S$ we have $y \geq a + \Delta$.

We will consider the first case. The second one can be dealt with in the same fashion.

Since the finite part of S is not empty by Condition 3, there is a superloss process L such that $L(\boldsymbol{x})$ is finite for all $\boldsymbol{x} \in \mathbb{B}^*$. Therefore $\mathcal{K}(\boldsymbol{x})$ is finite for all $\boldsymbol{x}$ too. This means that the point $(\mathcal{K}(\boldsymbol{x}0) - \mathcal{K}(\boldsymbol{x}), \mathcal{K}(\boldsymbol{x}1) - \mathcal{K}(\boldsymbol{x}))$ belongs to the finite part of S for all $\boldsymbol{x}$. Thus for all $\boldsymbol{x} \in \mathbb{B}^*$ we have $\mathcal{K}(\boldsymbol{x}0) - \mathcal{K}(\boldsymbol{x}) \geq a + \Delta$. If $\boldsymbol{x}_n$ is the string consisting of n zeroes ($n = 1, 2, \dots$), then $\mathcal{K}(\boldsymbol{x}_n) \geq (a + \Delta)n$.

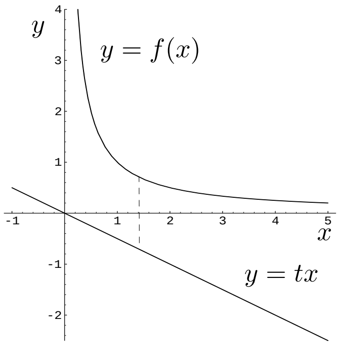

Fig. 1. Evaluation of the Legendre transformation

On the other hand, the function

$$L(\boldsymbol{x}) = \begin{cases} an & \text{if } \boldsymbol{x} = \boldsymbol{x}_n \\ +\infty & \text{otherwise} \end{cases}$$

is a superloss process because $(a, +\infty) \in S$. We have $L(\boldsymbol{x}_n) = an$ and this contradicts the lower bound on $\mathcal{K}(\boldsymbol{x}_n)$ we have just derived.

This remark shows that as far as predictive complexity is concerned, Condition 4 is not restrictive.

Appendix B: Legendre transformation

The *Legendre(–Young–Fenchel) transformation* may be defined for functionals on a locally convex space. However all we need in this paper is just the simplest one-dimensional case. We will follow the treatment of the one-dimensional case in [RV73]; the general theory of this transformation and conjugate functions may be found in [Roc70,ATF87].

Consider a convex function $f : \mathbb{R} \to [-\infty, +\infty]$. The *conjugate* function $f^* : \mathbb{R} \to [-\infty, +\infty]$ is defined by

$$f^*(t) = \sup_{x \in \mathbb{R}} (xt - f(x)) \ . \tag{20}$$

A function $g : \mathbb{R} \to [-\infty, +\infty]$ is called *proper* if $\forall x \in \mathbb{R} : g(x) > -\infty$ and $\exists x \in \mathbb{R} : g(x) < +\infty$. A proper convex function g is *closed* if its epigraph $\{(x, y) \in \mathbb{R}^2 \mid y \geq f(x)\}$ is closed w.r.t. the standard topology of $\mathbb{R}^2$ (cf. Section 7 of [Roc70])

Figure (1) provides an example. In the picture we have

$$f(x) = \begin{cases} \frac{1}{x} & \text{if } x > 0, \\ +\infty & \text{otherwise} \end{cases}$$

and we evaluate $f^*(-1/2)$. The supremum in (20) is achieved at $x = \sqrt{2}$.

Proposition 11 (see [RV73,Roc70]) *If $f : \mathbb{R} \to [-\infty, +\infty]$ is a proper convex function, the following properties hold:*

(i) f^* is convex, proper and closed, and
(ii) if f is closed, $f^{**} = f$.

Appendix C: On a Necessary Condition for the Existence of Predictive Complexity

PROOF (of Lemma 7). Assume the converse. Consider a game $\mathfrak{G}$ with the set of superpredictions S such that $S \cap \mathbb{R}^2$ is not convex but there exists complexity $\mathcal{K}$ w.r.t. $\mathfrak{G}$.

There exist points $B_0, B_1 \in S$ such that the segment $[B_0, B_1]$ is not a subset of S. Without loss of generality we may assume that $B_0 = (b_0, 0), B_1 = (0, b_1)$ (see Fig. 2). Indeed, a game $\mathfrak{G}$ with the set of superpredictions S specify complexity if and only if a game $\mathfrak{G}'$ with the set of superpredictions S' which is a shift of S (i.e., there are $a, b \in \mathbb{R}$ such that $S' = \{(x', y') \in (-\infty, +\infty]^2 \mid \exists (x, y) \in S : x' = x + a, y' = y + b\}$) specifies complexity.

There exists a point $A = (a_0, a_1)$ with $a_0, a_1 > 0$ on the boundary of S and above the straight line passing through B_0 and B_1. Let us denote this line by l and let us assume that it has the equation $\alpha_0 x + \alpha_1 y = \rho$, where $\alpha_0, \alpha_1, \rho > 0$. Let l' be the line parallel to l and passing through A. The equation of l' can be written as $\alpha_0 x + \alpha_1 y = \rho + \delta$, where $\delta > 0$.

Let us denote the numbers of 1s and 0s in a string $\boldsymbol{x}$ by $\sharp_1 \boldsymbol{x}$ and $\sharp_0 \boldsymbol{x}$, respectively. Since the functions $b_0 \sharp_0 \boldsymbol{x}$ and $b_1 \sharp_1 \boldsymbol{x}$ are superloss processes, there is $C > 0$ such that, for every $\boldsymbol{x} \in \mathbb{B}^*$, the inequalities

$$\mathcal{K}(\boldsymbol{x}) \leq b_0 \sharp_0 \boldsymbol{x} + C \tag{21}$$
$$\mathcal{K}(\boldsymbol{x}) \leq b_1 \sharp_1 \boldsymbol{x} + C \tag{22}$$

holds. At the same time, there is a sequence of strings $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ such that for any $n \in \mathbb{N}$ we have $|\boldsymbol{x}_n| = n$ and

$$\mathcal{K}(\boldsymbol{x}_n) \geq a_0 \sharp_0 \boldsymbol{x_n} + a_1 \sharp_1 \boldsymbol{x_n} \ . \tag{23}$$

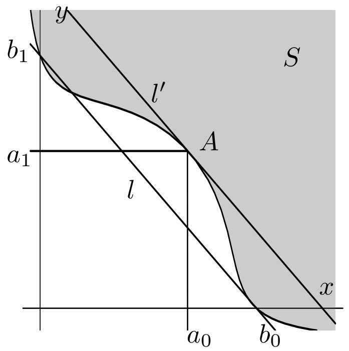

Fig. 2. A set of superpredictions S that is not convex

The construction of $\boldsymbol{x}_n$ is by induction. Let $\boldsymbol{x}_0 = \Lambda$. Suppose we have constructed $\boldsymbol{x}_n$. The point $(\mathcal{K}(\boldsymbol{x}_n 0) - \mathcal{K}(\boldsymbol{x}_n), \mathcal{K}(\boldsymbol{x}_n 1) - \mathcal{K}(\boldsymbol{x}_n))$ should lie in at least one of the half-planes $\{(x, y) \mid x \geq a_0\}$ or $\{(x, y) \mid y \geq a_1\}$ i.e., at least one of the inequalities

$$\mathcal{K}(\boldsymbol{x}_n 0) - \mathcal{K}(\boldsymbol{x}_n) \geq a_0 \tag{24}$$
$$\mathcal{K}(\boldsymbol{x}_n 1) - \mathcal{K}(\boldsymbol{x}_n) \geq a_1 \tag{25}$$

hold. We define $\boldsymbol{x}_{n+1}$ to be either $\boldsymbol{x}_n 0$ or $\boldsymbol{x}_n 1$ accordingly.

Combining (21), (22) and (23) we get

$$a_0 \sharp_0 \boldsymbol{x}_n + a_1 \sharp_1 \boldsymbol{x}_n \leq b_0 \sharp_0 \boldsymbol{x}_n + C \tag{26}$$
$$a_0 \sharp_0 \boldsymbol{x}_n + a_1 \sharp_1 \boldsymbol{x}_n \leq b_1 \sharp_1 \boldsymbol{x}_n + C \tag{27}$$

for every $n \in \mathbb{N}$. Since $(b_0, 0), (0, b_1) \in l$, we get $\alpha_0 b_0 = \alpha_1 b_1 = \rho$, while $A \in l'$ implies that $\alpha_0 a_0 + \alpha_1 a_1 = \rho + \delta$. Therefore

$$b_0 = \frac{\alpha_0 a_0 + \alpha_1 a_1}{\alpha_0} - \frac{\delta}{\alpha_0} \tag{28}$$
$$b_1 = \frac{\alpha_0 a_0 + \alpha_1 a_1}{\alpha_1} - \frac{\delta}{\alpha_1} \quad . \tag{29}$$

If we multiply (26) by α_0/a_1, (27) by α_1/a_0, add the equations together, and substitute the expressions (28) and (29) for b_0 and b_1, we obtain

$$\frac{\delta}{a_1} \sharp_0 \boldsymbol{x}_n + \frac{\delta}{a_0} \sharp_1 \boldsymbol{x}_n \leq C_1 \quad , \tag{30}$$

where $C_1 > 0$ is a constant. This is a contradiction since $\delta/a_1 > 0$, $\delta/a_0 > 0$, and at least one of the values $\sharp_0 \boldsymbol{x}_n$, $\sharp_1 \boldsymbol{x}_n$ is unbounded. □