The Existence of Predictive Complexity and the Legendre Transformation

Yuri Kalnishkan Volodya Vovk

Technical Report

CLRC-TR-00-04

May 2000

Royal Holloway University of London

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

Abstract

Predictive complexity is a generalisation of Kolmogorov complexity. In this paper we point out some properties of predictive complexity connected with the Legendre (–Young–Fenchel) transformation. Our main result is that mixability is necessary for the existence of conditional predictive complexity (it is known to be sufficient under very mild assumptions).

We formulate a differential criterion of mixability and show that it reduces to a very simple form if we employ the Legendre transformation. The Legendre transformation also turns out to have a probabilistic meaning which allows us to prove that a variant of predictive complexity specifies a unique (up to a parametrisation) mixable game.

Keywords: Mixability, Convexity, Conjugated Function.

Contents

1 Introduction

In this paper, we investigate the problem of existence of predictive complexity. In Section 2, we give the definition of predictive complexity. Unlike [VW98], where the original definition was given, we define predictive complexity for generalised games, which correspond to classes of traditional games.

In Section 3, we discuss necessary and sufficient conditions for the existence of predictive complexity. The most general of known sufficient conditions is mixability; we discuss it in Subsect. 3.1. In Subsect. 3.2, we try to find a necessary condition. We show that mixability is necessary for the existence of complexity for games with signals (Theorem 1). We also make a step towards proving the necessity of mixability for games without signals by showing that convexity is necessary for the existence of complexity.

In Section 4, we try to get an insight into the nature of mixability by applying the Legendre(–Young–Fenchel) transformation. This transformation naturally emerged in [Vovk98] in the investigation of prediction with expert advise and it turns out to be useful for the theory of predictive complexity. In Subsect. 4.1, we define the Legendre transformation and, in Subsect. 4.2, we use it to simplify the differential criterion of mixability. In Subsect. 4.3, we show that it is related to expectations of predictive complexity. As a consequence, we infer that a variant of predictive complexity fully describes the generalised game that specifies it.

2 Definitions

In this paper we restrict ourselves to games with binary outcomes, i.e. we will consider only the *outcome space* $\Omega = \mathbb{B} = \{0, 1\}$. A *game* $\mathfrak{G}$ is a pair (Γ, λ), where Γ stands for a *space of possible predictions*, and $\lambda : \Omega \times \Gamma \to \mathbb{R} \cup \{+\infty\}$ is a *loss function*. We suppose that a definition of computability over $\Omega \times \Gamma$ is given and λ is computable according to this definition.

We also impose some usual restrictions on games:

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 in the second argument.
3. There exists $\gamma \in \Gamma$ such that, for every $\gamma \in \Gamma$ the inequality $\lambda(\omega, \gamma) < +\infty$ holds.
4. For every $\omega \in \Omega$, there is no more then one prediction γ such that $\lambda(\omega, \gamma) = +\infty$ such that

Restrictions 1–3 are rather standard (see e.g. [Vovk98]), restriction 4 was added to get rid of some degenerated cases and to simplify the statements of theorems.

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 [Vovk90]), the set S of all superpredictions is the set of points that lie "north-east" of P.

Clearly, many different games may specify the same set of superpredictions. In many cases, the losses rather then the exact predictions are important and particular games play the role of mere parametrisations. This observation motivates the following definition.

A *generalised game* $\mathfrak{G} = (S)$ consist of a set $S \subseteq (-\infty, +\infty]^2$ which satisfies the following conditions:

- for every point $(x, y) \in S$ and every $a, b \in [0, +\infty]$, we have $(x + a, y + b) \in S$,

- ⋄ the coordinates of points of S are uniformly bounded from below,
- ⋄ the set S is not empty, and
- ⋄ the set S is the closure in the topology of $[-\infty, +\infty]^2$ of the set $S \cap \mathbb{R}^2$,

Lemma 1. *If a game satisfies assumptions 1–4, then its set of superpredictions is a generalised game. Conversely, every generalised game is the set of superpredictions for a game which has the possible predictions space* $[0, 1]$ *and satisfies the assumptions 1–4.*

We need the following generalisation of the concept of a game (see [VW98]). A triple $(\Gamma, \lambda, \Sigma)$ is called a *game with signals*. The set Σ is called the *space of possible signals*. The set Γ and the function λ bear the same names and satisfy the same assumptions as above. An ordinary game may be considered as a game with signals with the space of signals $\Sigma = \{0\}$. A pair (S, Σ), where S is a generalised game and Σ is the set of signals, is a *generalised game with signals*.

There is the following intuition behind the concept of a game. Let $\mathfrak{G} = (\Gamma, \lambda, \Sigma)$ be a game with signals. A *prediction strategy* $\mathfrak{A}$ works according to the protocol below:

```
FOR t = 1, 2, ...
  (1) 𝔄 observes a signal σ_t ∈ Σ
  (2) 𝔄 chooses a prediction γ_t ∈ Γ
  (3) 𝔄 observes the actual outcome ω_t ∈ Ω
  (4) 𝔄 suffers loss λ(ω_t, γ_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. A function $L : \Omega^* \to \mathbb{R} \cup \{+\infty\}$ is called a *loss process* w.r.t. $\mathfrak{G}$ if it coincides with the loss $\mathrm{Loss}_{\mathfrak{A}}$ of some computable strategy (*prediction algorithm*) $\mathfrak{A}$. Note that every loss process is computable.

Different prediction algorithm suffer different losses. If we had found the optimal algorithm, we could have proclaimed its loss to be a measure of complexity of an element of $(\Omega \times \Sigma)^*$. Unfortunately, there is no optimal algorithm in the most of natural cases. This leads to an idea to extend the class of loss processes.

Let $\mathfrak{G} = (S, \Sigma)$ be a generalised game with signals. A function $L : (\Omega \times \Sigma)^* \to \mathbb{R} \cup \{+\infty\}$ is called a *superloss process* w.r.t. $\mathfrak{G}$ (see [VW98]) if the following conditions hold:

- ⋄ $L(\Lambda) = 0$,
- ⋄ for every $n \in \mathbb{N} \cup \{0\}$ and every $\mathbf{x} \in \Omega^n$, $\overline{\sigma} \in \Sigma^n$, and $\sigma \in \Sigma$ the pair

  $$(L(\mathbf{x}0, \overline{\sigma}\sigma) - L(\mathbf{x}, \overline{\sigma}), L(\mathbf{x}1, \overline{\sigma}\sigma) - L(\mathbf{x}, \overline{\sigma}))$$

  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 it is minimal to within an additive constant in the class of all superloss processes. In other words, a superloss process K is universal if, for any other superloss process L, there exists a constant C such that

$$\forall n \in \mathbb{N} \cup \{0\} \forall \mathbf{x} \in \Omega^n \forall \overline{\sigma} \in \Sigma^n : K(\mathbf{x}, \overline{\sigma}) \leq L(\mathbf{x}, \overline{\sigma}) + C \ . \tag{2}$$

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 $\mathcal{K}^{\mathfrak{G}}$.

Complexity with signals is a natural counterpart of conditional Kolmogorov complexity $\mathrm{K}(\mathbf{x}|\overline{\sigma})$. The similarity motivates the notation $\mathcal{K}^{\mathfrak{G}}(\mathbf{x} \leftarrow \overline{\sigma})$ for $\mathcal{K}^{\mathfrak{G}}(\mathbf{x}, \overline{\sigma})$.

3 The Existence of Predictive Complexity

3.1 Mixability as a Sufficient Condition.

The definitions and propositions in this chapter are not new. They reformulate statements which were made before, see e. g. [VW98].

Consider a generalised game $\mathfrak{G} = (S)$. We say that a pair $(\beta, c) \in (0,1) \times [1, +\infty)$ is *allowable* for $\mathfrak{G}$ if, for every finite array of superpredictions

$$(s_0^{(1)}, s_1^{(1)}), (s_0^{(2)}, s_1^{(2)}), \ldots, (s_0^{(n)}, s_1^{(n)}) \in S$$

and weights $w_1, w_2, \ldots, w_n \in [0,1]$ such that $w_1 + w_2 + \ldots + w_n = 1$, there exists $(s_0^*, s_1^*) \in S$ such that the inequalities

$$c \log_\beta \sum_{i=1}^{n} w_i \beta^{s_0^{(i)}} \geq s_0^* \tag{3}$$

$$c \log_\beta \sum_{i=1}^{n} w_i \beta^{s_1^{(i)}} \geq s_1^* \tag{4}$$

hold or, in other words,

$$\left(c \log_\beta \sum_{i=1}^{n} w_i \beta^{s_0^{(i)}}, c \log_\beta \sum_{i=1}^{n} w_i \beta^{s_1^{(i)}} \right) \tag{5}$$

is a superprediction.

Definition 1. A generalised game $\mathfrak{G}$ is mixable if there exists $\beta \in (0,1)$ such that the pair $(\beta, 1)$ is allowable for $\mathfrak{G}$.

The following geometric interpretation is possible. For every $\beta \in (0,1)$, let the homeomorphism $B_\beta : (-\infty, +\infty]^2 \to [0, +\infty)^2 \subseteq \mathbb{R}^2$ be given by $B_\beta(x, y) = (\beta^x, \beta^y)$.

Proposition 1. *A pair $(\beta, 1)$ is allowable for $\mathfrak{G}$ with the set of superpredictions S if and only if $B_\beta(S)$ is convex.*

We may now proceed to the following fundamental statement.

Proposition 2. *If a generalised game (S) is mixable and Σ is a set of signals, then there exists a universal superloss process w.r.t. the generalise game with signals (S, Σ).*

3.2 Necessary Conditions.

It is natural to expect mixability to be a necessary condition for the existence of predictive complexity. We cannot prove it in the general case and here we discuss some partial results.

Theorem 1. *If a generalised game (S) is not mixable, then there exists no predictive complexity w.r.t. the generalised game with signals $(S, [0,1]^2)$.*

To prove this theorem, we need to recall some facts from the theory of prediction with expert advise. Let $\mathfrak{G} = (\Gamma, \lambda)$ be a game. We will consider *experts* which are prediction strategies and prediction algorithms which accept their predictions as signals.

The work of a prediction algorithm and a pool of experts $\mathcal{E}_1, \dots, \mathcal{E}_n$ is described by the following protocol:

FOR $t = 1, 2, \dots$
(1) $\mathcal{E}_1, \dots, \mathcal{E}_n$ output predictions $\gamma_t^{(1)}, \gamma_t^{(2)}, \dots, \gamma_t^{(n)}$
(2) $\mathfrak{A}$ observes $\gamma_t^{(1)}, \gamma_t^{(2)}, \dots, \gamma_t^{(n)}$
(3) $\mathfrak{A}$ chooses a prediction $\gamma_t \in \Gamma$
(4) $\mathfrak{A}$ and $\mathcal{E}_1, \dots, \mathcal{E}_n$ observe the actual outcome $\omega_t \in \Omega$
(5) $\mathfrak{A}$ and $\mathcal{E}_1, \dots, \mathcal{E}_n$ suffer losses $\lambda(\omega_t, \gamma_t)$ and
$\lambda(\omega_t, \gamma_t^{(1)}), \lambda(\omega_t, \gamma_t^{(2)}), \dots, \lambda(\omega_t, \gamma_t^{(n)})$ respectively.
END FOR.

If a pair (β, c) is allowable for $\mathfrak{G}$, then there is an algorithm $\mathfrak{A}$ such that, for any pool of experts $\mathcal{E}_1, \dots, \mathcal{E}_n$, the statement

$$\forall \mathbf{x} \in \Omega^* \forall i \in \{1, 2, \dots, n\} : \mathrm{Loss}_{\mathfrak{A}}(\mathbf{x} \leftarrow (\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_n)) \le c\, \mathrm{Loss}_{\mathcal{E}_i}(\mathbf{x}) + c \frac{\ln n}{\ln 1/\beta} \tag{6}$$

holds (see [Vovk98]). By $\mathrm{Loss}_{\mathfrak{A}}(\mathbf{y} \leftarrow (\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_n))$ we denote the loss of $\mathfrak{A}$ which reads the predictions output by $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_n$. It is possible to say that $\mathfrak{A}$ merges the predictions of $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_n$.

It may be proved that only allowable pairs have this property.

Proposition 3 ([Vovk98]). *Suppose that a pair (β, c) is not allowable for the game $\mathfrak{G}$. Then for every prediction strategy $\mathfrak{A}$ (not necessarily computable) which can read experts' predictions there exist a pool of experts $\mathcal{E}_1, \dots, \mathcal{E}_n$, a string $\mathbf{x} \in \Omega^*$, and a number $1 \le i \le n$ such that*

$$\mathrm{Loss}_{\mathfrak{A}}(\mathbf{x} \leftarrow (\mathcal{E}_1, \dots, \mathcal{E}_n)) > c\, \mathrm{Loss}_{\mathcal{E}_i}(\mathbf{x}) + c \frac{\ln n}{\ln 1/\beta} \quad . \tag{7}$$

We will now adapt this statement to mixable games. Following the argument from [Watk], the number of experts may be reduced to 2.

Lemma 2. *Suppose that a game $\mathfrak{G}$ is not mixable. Then for every prediction strategy $\mathfrak{A}$ (not necessarily computable) which can read experts predictions and every constant $C > 0$ there exist experts $\mathcal{E}_1, \mathcal{E}_2$, a string $\mathbf{x} \in \Omega^*$, and a number $1 \le i \le 2$ such that*

$$\mathrm{Loss}_{\mathfrak{A}}(\mathbf{x} \leftarrow (\mathcal{E}_1, \mathcal{E}_2)) > \mathrm{Loss}_{\mathcal{E}_i}(\mathbf{x}) + C \quad . \tag{8}$$

Proof. Assume the converse, i.e. suppose that there exist a strategy $\mathfrak{A}$ and a constant $C > 0$ such that, for every pair of experts $\mathcal{E}_1, \mathcal{E}_2$, every string $\mathbf{x} \in \Omega^*$, and every $1 \le i \le 2$, the inequality

$$\mathrm{Loss}_{\mathfrak{A}}(\mathbf{x} \leftarrow (\mathcal{E}_1, \mathcal{E}_2)) \le \mathrm{Loss}_{\mathcal{E}_i}(\mathbf{x}) + C \tag{9}$$

holds. Consider the strategy $\mathfrak{A}'$ which works with pools of experts of arbitrary size in the following way. Confronting a pool $\mathcal{E}_1, \ldots, \mathcal{E}_n$, the strategy $\mathfrak{A}'$ extends it to a pool of size m, where $m = 2^{\lceil \log_2 n \rceil}$ (say, by introducing fake experts $\mathcal{E}_{n+1}, \ldots, \mathcal{E}_m$ which all behave like $\mathcal{E}_n$). The strategy $\mathfrak{A}'$ runs several copies of $\mathfrak{A}$ in parallel. Every copy relates to one of $\log_2 m$ levels. There are $m/2^i$ copies of $\mathfrak{A}$ on the level i, where $i = 1, 2, \ldots, \log_2 m$. The first copy of the first level accepts the predictions output by experts $\mathcal{E}_1, \mathcal{E}_2$, the second copy reads $\mathcal{E}_3, \mathcal{E}_4$, and so on. The first copy of level i reads the predictions output by first and second copies of level $(i-1)$ as experts predictions, the second accepts third and fourth of level $(i-1)$, etc. The predictions made by the only copy on the level $\log_2 m$ are output as predictions of $\mathfrak{A}'$.

Using induction, one may check that, for any $\mathcal{E}_1, \ldots, \mathcal{E}_n$, every $\mathbf{x} \in \Omega^*$, and every $i = 1, 2, \ldots, n$, the inequality

$$\mathrm{Loss}_{\mathfrak{A}'}(\mathbf{x} \leftarrow (\mathcal{E}_1, \ldots, \mathcal{E}_n)) \leq \mathrm{Loss}_{\mathcal{E}_i}(\mathbf{x}) + C\lceil \log_2 n \rceil \tag{10}$$

holds. Clearly, for every natural number n, we have $C\lceil \log_2 n \rceil \leq 2C \log_2 n$. If we let $\beta = \exp(-1/(2C))$, we come to a contradiction with Proposition 3. $\square$

Proof of Theorem 1. Let $\mathfrak{G} = \{[0,1], \lambda, [0,1]^2\}$ be a game with the set of signals $[0,1]^2$ and the set of superpredictions S. Let us assume that there exists complexity $\mathcal{K}$ w.r.t. $\mathfrak{G}$.

Consider an element of $(\Omega \times [0,1]^2)^\infty$, i.e. three infinite strings $\mathbf{x}$, $\overline{\sigma}^{(1)}$ and $\overline{\sigma}^{(2)}$. We will denote their n-th elements by $\mathbf{x}_n$, $\sigma_n^{(1)}$, and $\sigma_n^{(2)}$, and n-element prefixes by $\overline{\mathbf{x}}_n$, $\overline{\sigma}_n^{(1)}$, and $\overline{\sigma}_n^{(2)}$, respectively. It follows from the definition of predictive complexity, that there is a constant $C > 0$ such that, for any n, the inequalities

$$\mathcal{K}(\overline{\mathbf{x}}_n \leftarrow (\overline{\sigma}_n^{(1)}, \overline{\sigma}_n^{(2)})) \leq \sum_{i=1}^{n} \lambda(\mathbf{x}_n, \sigma_n^{(1)}) + C \tag{11}$$

$$\mathcal{K}(\overline{\mathbf{x}}_n \leftarrow (\overline{\sigma}_n^{(1)}, \overline{\sigma}_n^{(2)})) \leq \sum_{i=1}^{n} \lambda(\mathbf{x}_n, \sigma_n^{(2)}) + C \tag{12}$$

hold.

Now we will construct three particular strings by induction. Let $\mathbf{x}_0 = \sigma_0^{(1)} = \sigma_0^{(2)} = \Lambda$. Suppose that the prefixes $\mathbf{x}_n$, $\sigma_n^{(1)}$, and $\sigma_n^{(2)}$ have been constructed before step t and $\mathcal{K}(\overline{\mathbf{x}}_n \leftarrow (\overline{\sigma}_n^{(1)}, \overline{\sigma}_n^{(2)})) = K_{t-1}$. Below is the description of the step t.

There is a (possibly uncomputable) prediction strategy $\mathfrak{S}$ which reads signals and suffers loss

$$\mathrm{Loss}_{\mathfrak{S}}(\mathbf{y} \leftarrow (\xi, \eta)) \leq \mathcal{K}(\overline{\mathbf{x}}_n \mathbf{y} \leftarrow (\sigma_n^{(1)}\xi, \sigma_n^{(2)}\eta)) \tag{13}$$

for every m and every $\mathbf{y} \in \Omega^m$, $\xi, \eta \in [0,1]^m$. By Lemma 2 there exist experts $\mathcal{E}_1$ and $\mathcal{E}_2$ and a string $\mathbf{y} \in [0,1]^*$ such that at least one of the inequalities

$$\mathrm{Loss}_{\mathfrak{S}}(\mathbf{y} \leftarrow (\mathcal{E}_1, \mathcal{E}_2)) \geq \mathrm{Loss}_{\mathcal{E}_1}(\mathbf{y}) + (t - K_{t-1}) \tag{14}$$

$$\mathrm{Loss}_{\mathfrak{S}}(\mathbf{y} \leftarrow (\mathcal{E}_1, \mathcal{E}_2)) \geq \mathrm{Loss}_{\mathcal{E}_2}(\mathbf{y}) + (t - K_{t-1}) \tag{15}$$

holds. We extend $\mathbf{x}_n$ with $\mathbf{y}$, $\overline{\sigma}_n^{(1)}$ with the predictions of $\mathcal{E}_1$, and $\overline{\sigma}_n^{(2)}$ with the predictions of $\mathcal{E}_2$.

The contradiction with (11), (12) completes the proof. $\square$

This theorem does not provide us with a general result on mixability and the existence of predictive complexity. We can, however, prove the following lemma. We use the notation $\sharp_0\mathbf{x}$ and $\sharp_1\mathbf{x}$ for the numbers of zeroes and ones in a binary string $\mathbf{x}$.

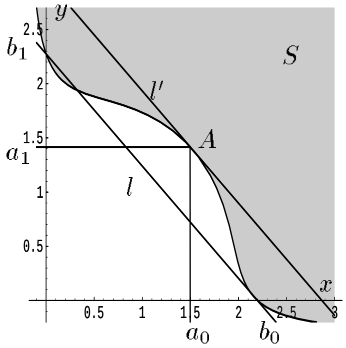

Figure 1 The set S is coloured grey.

Lemma 3. *Let $\mathfrak{G} = (S)$ be a generalised game with the set of superpredictions S such that $S \cap \mathbb{R}^2$ is not convex. Then predictive complexity w.r.t. $\mathfrak{G}$ does not exist.*

Proof. Assume the converse. Consider a generalised game $\mathfrak{G} = (S)$ with the set of superpredictions S such that $S \cap \mathbb{R}^2$ is not convex but complexity $\mathcal{K}$ w.r.t. $\mathfrak{G}$ exists. There exists a point $A = (a_0, a_1) \in \partial S$ and a straight line l with the equation $\alpha_0 x + \alpha_1 y = \rho$ such that $\alpha_0, \alpha_1 > 0$, the straight line l has common points with S in the both half-planes $\{x, y\} | \mid x > a_0\}$ and $\{x, y\} | \mid x < a_0\}$, but $\alpha_0 a_0 + \alpha_1 a_1 > \rho$ i.e. A lies to the north-east of l.

This might be proved in the following way. There are points $B, C \in S \cap \mathbb{R}^2$ such that the line segment $[B, C]$ has a common point D with $\mathbb{R}^2 \setminus S$. Since $\mathbb{R}^2 \setminus S$ is open, D must be surrounded by a vicinity of points from $S \cap \mathbb{R}^2$. We may now draw l through B and C and let A be the intersection of a straight line passing through D and different from l with ∂S.

Without loss of generality we may assume that $a_0, a_1 > 0$ and there exist points $(b_0, 0), (0, b_1) \in l \cap (S \cap \mathbb{R}^2)$ (see Fig. 1). Indeed, a generalised game $\mathfrak{G} = (S)$ specify complexity if and only if the generalised game $\mathfrak{G}' = (S')$ specify complexity, where S' is a shift of S (i.e. there are $a, b \in \mathbb{R}$ such that $S' = \{(x', y') \in (-\infty, +\infty]^2 \mid \exists (x, y) \in S : x' = x + a, y' = y + b\}$).

Since $b_0 \sharp_0 \mathbf{x}$ and $b_1 \sharp_1 \mathbf{x}$ are superloss processes, there is $C > 0$ such that, for every $\mathbf{x} \in \mathbb{R}^2$, the inequalities

$$\mathcal{K}(\mathbf{x}) \le b_0 \sharp_0 \mathbf{x} + C \tag{16}$$

$$\mathcal{K}(\mathbf{x}) \le b_1 \sharp_1 \mathbf{x} + C \tag{17}$$

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

$$\mathcal{K}(\mathbf{x}_n) \ge a_0 \sharp_0 \mathbf{x} + a_1 \sharp_0 \mathbf{x} \ . \tag{18}$$

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

$$\mathcal{K}(\mathbf{x_n} 0) - \mathcal{K}(\mathbf{x}_n) \ge a_0 \tag{19}$$

$$\mathcal{K}(\mathbf{x_n} 1) - \mathcal{K}(\mathbf{x}_n) \ge a_1 \tag{20}$$

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

Combining (16), (17) and (18) we get

$$a_0 \sharp_0 \mathbf{x}_n + a_1 \sharp_0 \mathbf{x}_n \le b_0 \sharp_0 \mathbf{x}_n + C \tag{21}$$

$$a_0 \sharp_0 \mathbf{x}_n + a_1 \sharp_1 \mathbf{x}_n \le b_1 \sharp_1 \mathbf{x}_n + C \tag{22}$$

for every $n \in \mathbb{N}$. Since $(b_0, 0)$ and $(0, b_1)$ lie below l', where l' is parallel to l and passes through A, we have

$$b_0 = \frac{\alpha_0 a_0 + \alpha_1 a_1}{\alpha_0} - \delta_0 \tag{23}$$

$$b_1 = \frac{\alpha_0 a_0 + \alpha_1 a_1}{\alpha_1} - \delta_1 \ , \tag{24}$$

where $\delta_0, \delta_1 > 0$. Multiplying (21) by α_0/a_1, (22) by α_1/a_0 and then adding them together we obtain

$$\frac{\alpha_0 \delta_1}{a_1} \sharp_0 \mathbf{x}_n + \frac{\alpha_1 \delta_0}{a_0} \sharp_1 \mathbf{x}_n \leq C_1 \ , \tag{25}$$

where $C_1 > 0$ is a constant. This leads to a contradiction since $\alpha_0 \delta_1 / a_1 > 0$, $\alpha_1 \delta_0 / a_0 > 0$, and at least one of the values $\sharp_0 \mathbf{x}_n$, $\sharp_1 \mathbf{x}_n$ is unbounded. □

Remark 1. In the definition of a generalised game, we require that $S = \overline{S \cap \mathbb{R}^2}$ in the topology of $[-\infty, +\infty]^2$. This restriction is natural since a generalised game which does not satisfy it do not specify a predictive complexity. Indeed, consider a generalised game (S) such that for every $(x, y) \in S \cap \mathbb{R}^2$ we have $x > a_2$, but the segment $[[a_1, +\infty], [a_2, +\infty]]$ is a subset of S, where $a_1 < a_2$. Suppose that $\mathcal{K}$ is complexity w.r.t. (S). Let us denote by $\mathbf{x}_n$ the string of n 0s. Clearly, there is $C > 0$ such that $\mathcal{K}(\mathbf{x}_n) \leq n a_1 + C$ for every n. On the other hand, for every n, the value $\mathcal{K}(\mathbf{x}_n 1)$ is finite and thus the superprediction $(\mathcal{K}(\mathbf{x}_{n+1}) - \mathcal{K}(\mathbf{x}_n), \mathcal{K}(\mathbf{x}_n 1) - \mathcal{K}(\mathbf{x}_n))$ is finite. It means that, for any natural n, we have $\mathcal{K}(\mathbf{x}_n) \geq n a_2$. This is a contradiction.

4 The Legendre Transformation

4.1 Definition and Some Properties.

The *Legendre(–Young–Fenchel) transformation* may be defined for the functionals on locally convex spaces. In this paper we need only the most simple one-dimensional case. The general theory of this transformation and conjugated functions may be found in [Rock70, ATF87]. In [RV73] the one-dimensional case is covered.

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

$$f^*(y) = \sup_{x \in \mathbb{R}} (xy - f(x)) \ . \tag{26}$$

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 g is *closed* if its *epigraph* $\{(x, y) \in \mathbb{R}^2 \mid y \geq g(x)\}$ is closed in the standard topology of $\mathbb{R}^2$.

The *subdifferential* of g is the set $\partial g = \{(x, y) \in \mathbb{R}^2 \mid g(x) \in \mathbb{R}, y \in \partial g(x)\}$, where $\partial g(x) = \{(x, y) \in \mathbb{R}^2 \mid \forall t \in \mathbb{R} : ty + g(x) \leq g(t)\}$ is the subdifferential at the point x.

Proposition 4 (see [RV73, Rock70]). *If $f : \mathbb{R} \to [-\infty, +\infty]$ is a proper convex function, the following properties hold:*

(i) f^* is convex, proper and closed;

(ii) ∂f and ∂f^* are symmetric w.r.t. the straight line $x = y$;

(iii) if f is closed, $f^{**} = f$.

If f is differentiable, (ii) reduces to $f' = (f^*)^{-1}$, where g^{-1} denotes the inverse function.

Sometimes we will apply this transformation to functions defined on (possibly infinite) intervals $I \subseteq \mathbb{R}$. In such cases we will define the function to be $+\infty$ outside I.

4.2 Differential Criteria.

The following theorem was originally proved in [HKW94]. We will give a short derivation from the general theory of mixability.

Theorem 2. *Consider a game $\mathfrak{G} = ([0,1], \lambda)$. Suppose that the functions $\lambda(0,t) = x(t)$ and $\lambda(1,t) = y(t)$ are monotone, twice differentiable in every point of $(0,1)$, and their derivatives do not vanish on $(0,1)$. Then $\mathfrak{G}$ is mixable if and only if there is $\varepsilon > 0$ such that*

$$\frac{y''x' - x''y'}{x'y'(y' - x')} \geq \varepsilon \tag{27}$$

holds on $(0,1)$. If $\mathfrak{G}$ is mixable, the smallest β such that the pair $(\beta, 1)$ is allowable is given by the formula

$$\beta_0 = \sup_{t\in(0,1)} \exp\left(-\frac{y''(t)x'(t) - x''(t)y'(t)}{x'(t)y'(t)(y'(t) - x'(t))}\right) . \tag{28}$$

Proof. Since x, y are monotone, the boundary $\partial(S \cap \mathbb{R}^2)$ of the set of real-valued superpredictions consists of the curve $\{(x(t), y(t)) \mid t \in (0,1)\}$ and possibly one or two half-lines. It follows from Proposition 1, that the pair $(\beta, 1)$ is allowable if and only if the curve $\{(\beta^{x(t)}, \beta^{y(t)}) \mid t \in (0,1)\}$ in concave. Having calculated the derivative $d^2(\beta^y)/d(\beta^x)^2$, we come to the condition

$$\ln\beta \geq -\frac{y''(t)x'(t) - x''(t)y'(t)}{x'(t)y'(t)(y'(t) - x'(t))} . \tag{29}$$

□

The criterion we obtained has an advantage of being symmetric in x, y. At this point we will loose this symmetry to obtain different interesting results.

Let $I = [a,b] \subseteq (-\infty, +\infty]$ be a (possibly infinite) interval. We will say that a continuous function $f : I \setminus \{+\infty\} \to (-\infty, +\infty]$ *canonically specifies* a generalised game (S) if:

1. f is not constant on every non-void interval $(a', b') \subseteq I$,

2. $S \cap \mathbb{R}^2$ coincides with the epigraph of F, where

$$F(x) = \begin{cases} f(x), & x \in I \setminus \{+\infty\} \\ f(b), & x > b. \end{cases} \tag{30}$$

Note that not every game can be canonically specified. However, if $S \cap \mathbb{R}^2$ is convex, then (S) is canonically specified by a unique f, which is convex. By Lemma 3, for every game (S) such that there exists complexity w.r.t. (S), the set $S \cap \mathbb{R}^2$ is convex and thus (S) is canonically specified by some f.

Theorem 3. *Let $\mathfrak{G} = (S)$ be a generalised game canonically specified by f with $I = [a, b]$. Suppose that f is twice differentiable on (a, b); then $\mathfrak{G}$ is mixable if and only if*

$$\exists \varepsilon > 0 \forall t \in (a, b) : \frac{f''(t)}{f'(t)(f'(t) - 1)} \geq \varepsilon \ . \tag{31}$$

If $\mathfrak{G}$ is mixable, the smallest β such that the pair $(\beta, 1)$ is allowable is given by the formula

$$\beta_0 = \sup_{t \in (0,1)} \exp\left(-\frac{f''(t)}{f'(t)(f'(t) - 1)}\right) \ . \tag{32}$$

Proof. Similar to Theorem 2. □

Corollary 1. *Under the conditions of Theorem 3, $\mathfrak{G}$ is mixable if and only if there exists $c > 0$ such that, for every t from the range of f', the inequality*

$$(f^*(t))'' \geq \frac{c}{t(t-1)} \tag{33}$$

holds.

Proof. The proof follows from the observation

$$f''(x) = \frac{1}{(f^*)''(f'(x))} \ . \tag{34}$$

□

4.3 The Legendre Transformation and Expectations of Complexity.

We have seen that the Legendre transformation helps to simplify a differential criterion of mixability. We will now show that is has a meaning in terms of expectations of predictive complexity.

Let (S) be a generalised game which specify complexity $\mathcal{K}$. We may consider expectations $\mathbf{E}\mathcal{K}(\xi_1^{(p)} \ldots \xi_n^{(p)})$, where $\xi_1^{(p)}, \ldots, \xi_n^{(p)}$ are results of n independent Bernoulli trials with the probability of 1 being equal to p. For any $p \in [0, 1]$, let us define a (finite or infinite) value

$$\tilde{f}(p) = \lim_{n \to \infty} \frac{\mathbf{E}\mathcal{K}(\xi_1^{(p)} \ldots \xi_n^{(p)})}{n} \ . \tag{35}$$

provided the limit exists.

Theorem 4. *Let (S) be a generalised game canonically specified by f and let $\mathcal{K}$ be complexity w.r.t. (S). Then, for every $p \in [0, 1]$, the number $\tilde{f}(p)$ exists. We have*

$$\tilde{f}(p) = -pf^*\left(\frac{p-1}{p}\right) \ , \tag{36}$$

for any $p \in (0, 1)$, $\tilde{f}(0) = \min\{x \mid \exists y : (x, y) \in S\}$, and $\tilde{f}(1) = \min\{y \mid \exists x : (x, y) \in S\}$.

Proposition 5 ([Kaln99]). *Let a game (S) specify complexity $\mathcal{K}$. Let $p \in [0,1]$ and let the numbers $\rho_1 \leq \rho_2$ be such that*

$$\forall (x,y) \in S \cap \mathbb{R}^2 : (1-p)x + py \geq \rho_1 \ , \tag{37}$$

but

$$\exists (x_0,y_0) \in S \cap \mathbb{R}^2 : (1-p)x_0 + py_0 \leq \rho_1 \ . \tag{38}$$

Then there is $C > 0$ such that, for every $n \in \mathbb{N}$, we have

$$\rho_1 n \leq \mathbf{E}\mathcal{K}(\xi_1^{(p)} \ldots \xi_n^{(p)}) \leq \rho_2 n + C \ , \tag{39}$$

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

Proof of Theorem 4. It follows from Proposition 5 that there is $C > 0$ such that, for every $p \in (0,1)$ and every $n \in \mathbb{N}$, we have

$$\alpha(p) n \leq \mathbf{E}\mathcal{K}(\xi_1^{(p)} \ldots \xi_n^{(p)}) \leq \alpha(p) n + C \ , \tag{40}$$

where

$$\alpha(p) = \inf_{(x,y)\in S\cap\mathbb{R}^2} (1-p)x + py \tag{41}$$

$$= \inf_{x \in I\setminus\{+\infty\}} (1-p)x + pf(x) \tag{42}$$

$$= -p \sup_{x \in I\setminus\{+\infty\}} \frac{p-1}{p}x - f(x) \tag{43}$$

$$= -pf^*\left(\frac{p-1}{p}\right) \tag{44}$$

(we suppose that $f : I \setminus \{+\infty\} \to (-\infty, +\infty]$). ☐

Corollary 2. *If two games (S_1), (S_2) specify the same complexity, then $S_1 = S_2$.*

Proof. The corollary may be derived directly from Proposition 5 but we will give a proof which relies on Theorem 4. If f canonically specifies (S) and $\mathcal{K}$ is the complexity w.r.t. (S), then $\mathcal{K}$ uniquely determines $\tilde{f}(p)$ for every $p \in [0,1]$ and thus $f^*(t)$ for every negative t. The value $f^*(0)$ is specified by semicontinuity. Clearly, $f^*(t) = +\infty$ for every positive t. On the other hand, f is convex and thus $f^{**} = f$. ☐

Acknowledgments

This work was supported by EPSRC through the grant GR/M14937 ("Predictive complexity: recursion-theoretic variants"); the first author was also supported by an ORS Awards Scheme grant.

References

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

[HKW94] D. Haussler, J. Kivinen, and M. K. Warmuth. Tight worst-case loss bounds for predicting with expert advise. Technical Report UCSC-CRL-94-36, University of California at Santa Cruz, revised December 1994.

[Kaln99] Yuri Kalnishkan. General linear relations among different types of predictive complexity. In *Proc. 10th International Conference on Algorithmic Learning Theory - ALT '99*, volume 1720 of *Lecture Notes in Artificial Intelligence*, pages 323–334. Springer-Verlag, 1999.

[Rock70] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

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

[Vovk90] 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.

[Vovk98] 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.

[Watk] C. J. H. C. Watkins. Personal communications.