A Criterion for the Existence of Predictive Complexity for Binary Games

Yuri Kalnishkan Vladimir Vovk Michael V. Vyugin[1]

{yura,vovk,misha}@cs.rhul.ac.uk

Technical Report

CLRC — TR — 04 — 04

March 25, 2004

Revised: May 21, 2004

Computer Learning Research Centre
Royal Holloway University of London
Egham, Surrey TW20 0EX, England

[1]Supported by EPSRC through the grant GR/R46670 "Complexity Approximation Principle and Predictive Complexity: Analysis and Applications"

Abstract

It is well known that there exists a universal (i.e., optimal to within an additive constant if allowed to work infinitely long) algorithm for lossless data compression (Kolmogorov, Levin). The game of lossless compression is an example of an on-line prediction game; for some other on-line prediction games (such as the simple prediction game) a universal algorithm is known not to exist. In this paper we give an analytic characterization of those binary on-line prediction games for which a universal prediction algorithm exists.

Contents

1 Introduction

We consider the following on-line prediction protocol: the prediction algorithm is fed with a sequence of bits $\omega_1, \omega_2, \ldots$, and its goal is to predict each ω_t, $t = 1, 2, \ldots$, given the previous bits $\omega_1, \ldots, \omega_{t-1}$. The loss suffered by the algorithm at each trial is measured by a loss function λ. We are interested in universal prediction algorithms for this problem.

The first universal algorithm was found back in 1965 by Kolmogorov for a batch version of this problem. Kolmogorov found a universal method of compression: for each string one can generate shorter and shorter compressed strings so that optimal compression is achieved in the limit; the original string can be restored from each of these compressed string by a computable function. The length of the shortest compressed string is known as the (plain) *Kolmogorov complexity* of the original string.

Slightly later Levin (in [ZL70]) put Kolmogorov's ideas in an on-line framework constructing a universal algorithm for on-line data compression; the version of Kolmogorov complexity given by Levin's construction is denoted KM (the log of the *a priori* semimeasure). There are also several 'semi-on-line' versions of Kolmogorov complexity (such as prefix and monotone complexity).

Much later (see [VW98]) it was shown that a universal algorithm exists for a wide class of on-line prediction games including, besides some traditional games of prediction, the data compression game and games of playing the market; the loss suffered by the universal algorithm on a string is called the *predictive complexity* of that string. The technical basis for this result is provided by the theory of prediction with expert advice (see [CBFH$^+$97, HKW98, Vov98]).

The first on-line prediction games for which a universal prediction algorithm (equivalently, predictive complexity) does not exist appear in [KVV01, KV02]. They include such important games as the simple prediction game and the absolute-loss game. The main result of this paper is an analytic characterization of those on-line prediction games for which a universal prediction algorithm exists.

2 Preliminaries

2.1 Games and Superpredictions

An *(on-line prediction) game* $\mathfrak{G}$ is a triple $\langle \mathbb{B}, \Gamma, \lambda \rangle$, where $\mathbb{B} = \{0, 1\}$ is the *outcome space*, Γ is the *prediction space*, and $\lambda : \mathbb{B} \times \Gamma \to \mathbb{R} \cup \{+\infty\}$ is the *loss function*.

It is essential to allow λ to take the value $+\infty$; however we do not allow $-\infty$ since the sum $+\infty + (-\infty)$ is undefined. We also prohibit the values of λ to approach $-\infty$: we require that there is some $a \in \mathbb{R}$ such that $\lambda(\omega, \gamma) \geq a$ for all $\omega \in \mathbb{B}$ and $\gamma \in \Gamma$.

Having permitted $+\infty$, we need to ensure that at least some values of λ are finite, otherwise the game is degenerate. We require that there is $\gamma_0 \in \Gamma$ such that both $\lambda(0, \gamma_0)$ and $\lambda(1, \gamma_0)$ are finite numbers. In fact, our main result remains true for the games that do not satisfy these requirements, but they need a special treatment. Appendix B describes the situation in more detail.

We need to impose computability requirements on the games. Assume that with every game there comes an oracle that is able to answer questions about the game. All computations involving the game can include calls to the oracle. Intuitively, the oracle can answer all natural questions about the game. The exact requirements to oracles are formulated in Sect. 5.

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 ones in $\boldsymbol{x}$ and the number of zeroes in $\boldsymbol{x}$ are denoted by $\sharp_1\boldsymbol{x}$ and $\sharp_0\boldsymbol{x}$, respectively.

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 $\omega_t \in \mathbb{B}$
  (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 all loss processes for a given game has no minimal elements except in some 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 $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$. We say that a pair $(s_0, s_1) \in (-\infty, +\infty]^2$ is a superprediction if there is $\gamma \in \Gamma$ such that the inequalities $\lambda(0, \gamma) \leq s_0$ and $\lambda(1, \gamma) \leq s_1$ hold.

The set of superpredictions S is an important object describing the game. Any game can be reduced to the canonical form $\langle \mathbb{B}, S, \lambda_{\mathrm{pr}} \rangle$, where λ_{pr} is the projector: for each $(s_0, s_1) \in S$, we have $\lambda_{\mathrm{pr}}(0, (s_0, s_1)) = s_0$ and $\lambda_{\mathrm{pr}}(1, (s_0, s_1)) = s_1$. We can think of games as of different parameterizations of sets of superpredictions.

The requirements we have imposed on functions λ can be reformulated in terms of S as follows. The set $S \cap \mathbb{R}^2$, the finite part of S, is not empty and there is $a \in \mathbb{R}$ such that $S \subseteq [a, +\infty]^2$.

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

- $L(\Lambda) = 0$,
- for every $\boldsymbol{x} \in \mathbb{B}^*$, 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.

Note that L is semicomputable w.r.t. the oracle of the game, i.e., the algorithm is allowed to make calls to the oracle.

We will say that a superloss process K is *universal* if for any superloss process L there exists a constant C such that

$$K(\boldsymbol{x}) \leq L(\boldsymbol{x}) + C \tag{2}$$

for all $\boldsymbol{x} \in \mathbb{B}^*$. 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 $\mathcal{K}^{\mathfrak{G}}(\boldsymbol{x}) \leq \mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{A}}(\boldsymbol{x}) + C$. We call $\mathcal{K}^{\mathfrak{G}}$ *(predictive) complexity* w.r.t. $\mathfrak{G}$. We will usually omit the superscript and write just $\mathcal{K}$.

2.3 Mixability

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

$$\mathfrak{B}_\beta(x, y) = (\beta^x, \beta^y) \ . \tag{3}$$

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

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

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

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.

Natural examples of games that are not mixable (*non-mixable*) are provided by the absolute loss game with $\Gamma = [0,1]$ and $\lambda(\omega, \gamma) = |\omega - \gamma|$ (the set of superpredictions for the absolute-loss game is shown in Fig. 1) and the simple prediction game with $\Gamma = \mathbb{B}$ and

$$\lambda(\omega, \gamma) = \begin{cases} 0 & \text{if } \omega = \gamma \ , \\ 1 & \text{otherwise} \ . \end{cases} \tag{4}$$

It is easy to establish the results concerning the mixability or non-mixability of the above games with the help of Theorem 2 below.

3 The Main Result and the Idea of the Proof

Theorem 1. *Predictive complexity w.r.t. a game $\mathfrak{G}$ exists if and only if $\mathfrak{G}$ is mixable.*

It is assumed in the theorem that the oracle for the game $\mathfrak{G}$ satisfies the requirements from Sect. 5.

The 'if' part of the theorem is proved in [VW98]. We have included the proof in Appendix F for completeness. The 'only if' part is proved in Appendix D.

In order to illustrate the idea of the proof, we consider a special case. Let us show that there is no predictive complexity for the absolute-loss game. The set of superpredictions for the absolute-loss game is shown in Fig. 1.

The proof is by *reductio ad absurdum*. Suppose that $\mathcal{K}$ is predictive complexity w.r.t. the absolute-loss game.

The set S of superpredictions for the absolute-loss game lies above the straight line $x/2 + y/2 = 1/2$. It is easy to see that

$$\mathbf{E}\mathcal{K}\left(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}\right) \geq n/2 \tag{5}$$

for all positive integers n, 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 equal to $1/2$. Indeed, the mean 'loss' of L on each 'trial' is at least $1/2$ (see Lemma 8).

Consider the prediction strategy that outputs 0 no matter what outcomes actually occur. It suffers loss $L(\boldsymbol{x}) = \sharp_1 \boldsymbol{x}$ for all strings $\boldsymbol{x}$. It follows from the definition

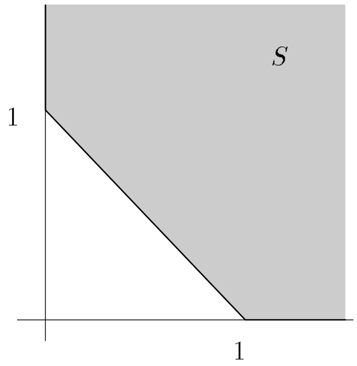

Figure 1 The set of superpredictions S for the absolute-loss game

of predictive complexity that there is a constant C such that $\mathcal{K}(\boldsymbol{x}) \le L(\boldsymbol{x}) + C_1 = \sharp_1\boldsymbol{x} + C_1$ for all strings $\boldsymbol{x}$. Similar considerations imply that $\mathcal{K}(\boldsymbol{x}) \le \sharp_0\boldsymbol{x} + C_0$ for some $C_0 > 0$. Letting $C = \max(C_0, C_1) > 0$ yields

$$\mathcal{K}(\boldsymbol{x}) \le \min(\sharp_0\boldsymbol{x}, \sharp_1\boldsymbol{x}) + C \le \frac{|\boldsymbol{x}|}{2} + C \ . \tag{6}$$

Let $\Xi \subseteq \mathbb{B}^n$ be the set of strings having $n/2 + \sqrt{n}$ or more 0s. We will use the upper bound $\mathcal{K}(\boldsymbol{x}) \le L(\boldsymbol{x}) + C \le n/2 - \sqrt{n} + C$ on $\mathcal{K}$ for $\boldsymbol{x} \in \Xi$ and the upper bound (6) for all other values of $\boldsymbol{x}$. It follows from the DeMoivre-Laplace limit theorem (see, e.g., [Fel68]) that $\Pr(\Xi)$ tends to a positive constant δ as n tends to ∞, where $\Pr(\Xi)$ is the probability of the event $(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}) \in \Xi$. Thus

$$\begin{aligned} \mathbf{E}\mathcal{K}\left(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}\right) &\le \left(\frac{n}{2} - \sqrt{n}\right)\Pr(\Xi) + \frac{n}{2}(1 - \Pr(\Xi)) + C \\ &= \frac{n}{2} - \sqrt{n}(\delta + o(1)) + C \end{aligned}$$

as $n \to +\infty$. This contradicts (5).

In the general case we approximate the boundary of the set of superpredictions S with line segments and perform a similar trick. The approximation takes place near a point where mixability gets violated. In order to find such a point, we need the concept of a canonical specification and a criterion of mixability formulated in Sect. 4.

4 Canonical Specifications

The proof of the main theorem is based on the concept of canonical specification. In this section we introduce canonical specifications and establish a criterion of mixability in terms of canonical specifications. The criterion is of independent interest because it reduces the question of whether a game is mixable to a simple analytical problem.

4.1 Definitions

Let $\mathfrak{G}$ be a game with the set of superpredictions S. Consider the function $\tilde{f} : \mathbb{R} \to (-\infty, +\infty]$ defined by $\tilde{f}(x) = \inf\{y \in \mathbb{R} \mid (x, y) \in S\}$ for each $x \in \mathbb{R}$ (here we assume that $\inf \varnothing = +\infty$).

Let $a = \sup\{x \in \mathbb{R} \mid \tilde{f}(x) = +\infty\}$; since by our assumption S has a non-empty real part, $a < +\infty$. Let $b = \inf\{x \in \mathbb{R} \mid \tilde{f}$ is constant on $[x, +\infty)\}$, where we again assume that $\inf \varnothing = +\infty$. It is possible that $b = +\infty$. Clearly, $b \geq a$. If $b > a$, let f be the restriction of $\tilde{f}$ to (a, b). We will call such $f : (a, b) \to \mathbb{R}$ the *canonical specification* of the game $\mathfrak{G}$.

If $a = b$, then the closure of $S \cap \mathbb{R}^2$ coincides with the set $[a, +\infty] \times [c, +\infty]$ for some $c \in \mathbb{R}$. It is easy to see that a game with the set of superpredictions $S = [a, +\infty] \times [c, +\infty]$ is mixable and if the requirements to the oracle from Sect. 5 hold, then there is predictive complexity w.r.t. this game.

Note that if S is the closure of its final part, then the canonical specification $f : (a, b) \to \mathbb{R}$ uniquely determines S. Indeed, in this case S is the closure w.r.t. the extended topology of the set

For example, the canonical specification for the absolute-loss game defined in Subsect. 2.3 is $f(x) = 1 - x$, $x \in (0, 1)$.

4.2 Derivatives

Let $f : (a, b) \to \mathbb{R}$ be the canonical specification of a game $\mathfrak{G}$ with the set of superpredictions S.

We need to define the following derivatives that make sense even if f is not smooth. First let us define the left and right derivatives:

$$f'_l(x) = \lim_{\Delta x \to 0-} \frac{f(x + \Delta x) - f(x)}{\Delta x} \ , \tag{7}$$

$$f'_r(x) = \lim_{\Delta x \to 0+} \frac{f(x + \Delta x) - f(x)}{\Delta x} \ . \tag{8}$$

If $f'_l(x) = f'_r(x)$ at some point x, we say that the derivative $f'(x) = f'_l(x) = f'_r(x)$ exists.

Suppose the derivative $f'(x)$ exists at $x \in (a, b)$. Let

$$\underline{f}''_l = \liminf_{\Delta x \to 0-} \frac{f(x + \Delta x) - (f(x) + f'(x)\Delta x)}{\frac{1}{2}(\Delta x)^2} \ , \tag{9}$$

$$\underline{f}''_r = \liminf_{\Delta x \to 0+} \frac{f(x + \Delta x) - (f(x) + f'(x)\Delta x)}{\frac{1}{2}(\Delta x)^2} \ . \tag{10}$$

By definition, lower limits always exist though their values may reach $-\infty$. We now let

$$\underline{f}''(x) = \min(\underline{f}''_l(x), \underline{f}''_r(x)) \ . \tag{11}$$

The derivatives $\overline{f}''_l$, $\overline{f}''_r$, and $\overline{f}''$ may be defined in a similar fashion, but they are not necessary for the purposes of this paper.

If $S \cap \mathbb{R}^2$, the real part of S, is convex, then f is convex. Since convex functions always have one-sided derivatives, $f'_l(x)$ and $f'_r(x)$ exist for each $x \in (a, b)$. Moreover they coincide everywhere except on a countable subset of the interval (a, b) (see, e.g., [RV73], Theorem 11.C). The derivatives $\underline{f}''_l$ and $\underline{f}''_r$ are nonnegative for convex functions f.

4.3 An Analytical Characterization of Mixable Games

In this subsection we formulate some results characterizing mixability in terms of canonical specifications.

Theorem 2. *Let $\mathfrak{G}$ be a game with the set of superpredictions S and a canonical specification $f : (a, b) \to \mathbb{R}$. Then $\mathfrak{G}$ is mixable if and only if the following conditions hold: $S \subseteq \overline{S \cap \mathbb{R}^2}$ w.r.t. the extended topology, f is convex, and*

$$\inf_{x \in (a,b) \setminus D} \frac{\underline{f}''(x)}{f'(x)(1 - f'(x))} > 0 \ , \tag{12}$$

where D is the set of points $y \in (a, b)$ such that $f'_l(y) \neq f'_r(y)$.

This theorem follows immediately from the next theorem and Lemmas 2 and 3.

Theorem 3. *Let $\mathfrak{G}$ be a game with the set of superpredictions S and the canonical specification $f : (a, b) \to \mathbb{R}$; let $\beta \in (0, 1)$. Let S be the closure of its final part, $S = \overline{S \cap \mathbb{R}^2}$. Then $\mathfrak{G}$ is β-mixable if and only if the following conditions hold: f is convex and*

$$\underline{f}''(x) \geq -f'(x)(1 - f'(x)) \ln \beta \tag{13}$$

for all $x \in (a, b) \setminus D$, where D is the set of points $y \in (a, b)$ such that $f'_l(y) \neq f'_r(y)$.

The theorem is proved in Appendix E. Note that the computability requirements from Sect. 5 are not used in the proofs of the two theorems from this subsection.

5 Computability

In this section we discuss the requirements the oracle for a game $\mathfrak{G}$ with the set of superpredictions S should satisfy. The general principle is that the oracle should be able to answer all 'reasonable' questions about the game; we do not want, however, to formalize the notion of 'reasonable' and just present a rather long list of questions that the oracle should be able to answer for our proofs to work. All those questions can be easily answered for games specified in a natural analytical form. Thus the requirements are not particularly restrictive as far as applications are concerned.

A *rational interval* is either an interval (p, q), where p and q are rational numbers or $+\infty$ or $-\infty$, one of the intervals $[-\infty, p)$, $(p, +\infty]$, where p is a rational number, or $[-\infty, +\infty]$. A rational interval is open w.r.t. the extended topology of $[-\infty, +\infty]$.

We say that a sequence of nested intervals specifies a number $x \in [-\infty, +\infty]$ if it shrinks to x, i.e., the intersection of all intervals from the sequence is the one-point set $\{x\}$.

In order to prove the existence of complexity for mixable games, we require that S has a constructively open (w.r.t. the oracle) interior. Suppose that the oracle is (simultaneously) getting two sequences of rational intervals shrinking to x and y. If the point (x, y) belongs to the interior of S, the oracle should be able to tell it eventually. This requirement is used in the proof of the 'if' part of the main theorem.

Now suppose that the game has a canonical specification $f : (a, b) \to \mathbb{R}$. The oracle should be able to tell whether

$$\inf_{x \in (a,b)} \frac{\underline{f}''(x)}{f'(x)(1 - f'(x))} = 0 \ , \tag{14}$$

and if so it must produce a sequence of points $x_m \in \mathbb{R}$, $m = 1, 2, \ldots$ (by saying that the oracle produces the sequence we mean that given m and l the oracle produces a rational interval with endpoints $p_{m,l}$ and $q_{m,l}$ so that for each m the intervals with endpoints $p_{m,l}$ and $q_{m,l}$, $l = 1, 2, \ldots$, shrink to x_m) with the following properties: the values $f'(x_m)$ exist for all positive integers m;

$$\lim_{m \to +\infty} \frac{\underline{f}''(x_m)}{f'(x_m)(1 - f'(x_m))} = 0 \ ; \tag{15}$$

x_m converges to $x_0 \in [a,b]$; the sequence x_m is monotone; the values $f(x_m)$, $f'(x_m)$, $\underline{f}''_l(x_m)$, and $\underline{f}''_r(x_m)$ can be produced by the oracle; for each m the oracle can produce a sequence $\Delta x_{m,k}$, $k = 1, 2, \ldots$, such that: $\Delta x_{m,k} \to 0$ as $k \to +\infty$ for all positive integers m; for each positive integer m the values $\Delta x_{m,k}$ are of the same sign; the values $f(x_m + \Delta x_{m,k})$ can be produced by the oracle for all positive integers m and k; for every positive integer m the value of $\underline{f}''(x_m)$ is achieved on the sequence $\Delta x_{m,k}$, $k = 1, 2, \ldots$, i.e.,

$$\underline{f}''(x_m) = \lim_{k \to +\infty} \frac{f(x_m + \Delta x_{m,k}) - (f(x_m) + f'(x_m)\Delta x_{m,k})}{\frac{1}{2}(\Delta x_{m,k})^2} .$$

The above requirements are formulated in the asymmetric fashion. Note that they are in fact of more general nature.

Consider the game $\mathfrak{G}_{\mathrm{m}}$ with the loss function λ_{m} defined as follows: $\lambda_{\mathrm{m}}(0,\gamma) = \lambda(1,\gamma)$ and $\lambda_{\mathrm{m}}(1,\gamma) = \lambda(0,\gamma)$ for every $\gamma \in \Gamma$, where λ is the loss function for $\mathfrak{G}$. It is easy to see that the set S_{m} of superpredictions for $\mathfrak{G}_{\mathrm{m}}$ can be obtained by mirroring S w.r.t. the straight line $x = y$. Let $f_{\mathrm{m}} : (a_{\mathrm{m}}, b_{\mathrm{m}}) \to \mathbb{R}$ be the canonical specification of $\mathfrak{G}_{\mathrm{m}}$.

If the oracle is able to answer the abovementioned questions about $\mathfrak{G}$, then the answers to the same questions about $\mathfrak{G}_{\mathrm{m}}$ can be easily generated. Indeed, the interior of S_{m} can be obtained by mirroring the interior of S. The derivatives of f_{m} can be evaluated using the formulae for the derivatives of the inverse function; the expression $\underline{f}''/(f'(1-f'))$ remains invariant if f is replaced by f^{-1}.

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.

[Fel68] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume I. John Wiley & Sons, Inc, 3rd edition, 1968.

[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.

[Kal02] Y. Kalnishkan. General linear relations among different types of predictive complexity. *Theoretical Computer Science*, 271:181–200, 2002.

[KV02] Y. Kalnishkan and M. V. Vyugin. On the absence of predictive complexity for some games. In *Algorithmic Learning Theory, 13th International Conference, ALT2002, Proceedings*, volume 2533 of *Lecture Notes in Artificial Intelligence*, pages 164–172. Springer, 2002.

[KVV01] Y. Kalnishkan, M. V. Vyugin, and V. Vovk. Losses, complexities and the Legendre transformation. In *Proceedings of the 12th International Conference on Algorithmic Learning Theory, ALT 2001*, volume 2225 of *Lecture Notes in Artificial Intelligence*. Springer, 2001.

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

[Pet75] V. V. Petrov. *Sums of Independent Random Variables*. Springer, 1975.

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

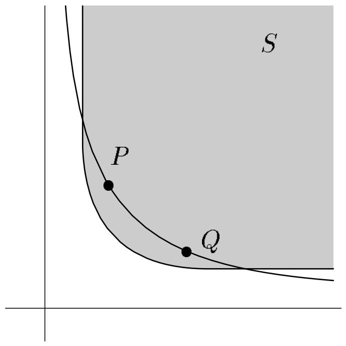

Figure 2 An equivalent definition of mixability

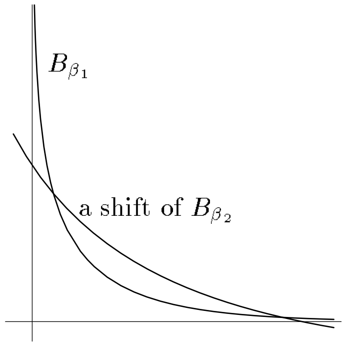

Figure 3 An arc of B_{β_2} connecting points of B_{β_1}

[Val64] F. A. Valentine. *Convex Sets.* McGraw-Hill Book Company, 1964.

[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.

[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: Closure, Mixability and the Existence of Complexity

This section contains some simple lemmas about topological properties of the set of superpredictions S.

Let us start with a technical lemma offering a restatement of the concept of mixability.

Lemma 1. *A game $\mathfrak{G}$ with the set of superpredictions S is β-mixable if and only if for each two points $P = (x_1, y_1)$ and $Q = (x_2, y_2)$ from S such that neither the pair of inequalities $x_1 \geq x_2$ and $y_1 \geq y_2$ is true nor the pair of inequalities $x_1 \leq x_2$ and $y_1 \leq y_2$ is true, there is a shift of the curve $B_\beta = \{(x, y) \in (-\infty, +\infty]^2 \mid \beta^x + \beta^y = 1\}$ such that P and Q belong to the shift and the arc between P and Q is a subset of S.*

The statement of the lemma is illustrated in Fig. 2.

Proof. The lemma is implied by the following observation. The inverse image of a straight line that can be described by the equation $ax + by = 1$ with $a, b \geq 0$ is a shift of B_β. □

Lemma 2. *Let $\mathfrak{G}_1$ be a game with the set of superpredictions S_1 and $\mathfrak{G}_2$ be a game with the set of superpredictions S_2 such that S_1 is the closure of S_2 w.r.t. the extended topology of $[-\infty, +\infty]^2$. Then $\mathfrak{G}_1$ is mixable if and only if $\mathfrak{G}_2$ is mixable and there is predictive complexity w.r.t. $\mathfrak{G}_1$ if and only if there is predictive complexity w.r.t. $\mathfrak{G}_2$.*

Proof. Let L_1 be a superloss process w.r.t. $\mathfrak{G}_1$. If L_2 is defined by $L_2(\boldsymbol{x}) = L_1(\boldsymbol{x}) + (1 - 1/|\boldsymbol{x}|)$ for all $\boldsymbol{x} \in \mathbb{B}^*$, then L_2 is a superloss process w.r.t. $\mathfrak{G}_2$ and $|L_1(\boldsymbol{x}) - L_2(\boldsymbol{x})| \leq 1$ for all $\boldsymbol{x} \in \mathbb{B}^*$. This implies equivalence for predictive complexity.

If $\mathfrak{G}_2$ is β-mixable, then $\mathfrak{G}_1$ is β-mixable, because the closure of the convex set $\mathfrak{B}_\beta(S_2)$ is convex (see, e.g., [Val64], Theorem 1.11).

In order to show the equivalence for mixability, we need Lemma 1. Suppose that $\mathfrak{G}_1$ is β_1-mixable and $\beta_2 \in (\beta_1, 1)$. One can check by direct calculation that each two points on the curve B_{β_1} can be connected by an arc of the curve B_{β_2} and the arc will lie above B_{β_1} (see Fig. 3). Thus $\mathfrak{G}_2$ is β_2-mixable. □

Note that β-mixability of $\mathfrak{G}_1$ above does not imply β-mixability of $\mathfrak{G}_2$ in the general case. Indeed, let $\mathfrak{G}_1$ be the game with the set of superpredictions $S_1 = \{(x, y) \in (-\infty, +\infty]^2 \mid \beta^x + \beta^y \leq 1\}$. Let us remove a finite arc from the boundary of S_1 and denote the remaining set by S_2. Clearly, $\mathfrak{B}_\beta(S_2)$ is not convex.

However the lemma implies that it is sufficient to consider the games with closed sets of superpredictions.

Lemma 3. *Let $\mathfrak{G}$ be a game with the set of superpredictions S such that the real part $S \cap \mathbb{R}^2$ is not empty. If $S \not\subseteq \overline{S \cap \mathbb{R}^2}$, i.e., S is not a subset of the closure of its real part, then $\mathfrak{G}$ is not mixable and there is no predictive complexity w.r.t. $\mathfrak{G}$.*

Proof. Let $S \not\subseteq \overline{S \cap \mathbb{R}^2}$. At least one of the following cases is true:

1. there is a number $a \in \mathbb{R}$ and $\Delta \in (0, +\infty)$ 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 is a number a and $\Delta > 0$ such that S contains the point $(+\infty, a)$ but for every finite $(x, y) \in S$ we have $y \geq a + \Delta$.

It can be assumed thar a is computable. We will consider the first case. The second one can be dealt with in the same fashion.

Let $\mathcal{K}$ be predictive complexity w.r.t. $\mathfrak{G}$. Since the finite part of S is not empty, 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, then $\mathcal{K}(\boldsymbol{x}_n) \geq (a + \Delta)n$, $n = 1, 2, \ldots$.

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 $\mathcal{K}(\boldsymbol{x}_n)$ derived above.

The image $\mathfrak{B}_\beta(S)$ is not convex due to the presence of the line segment connecting $\mathfrak{B}_\beta(a, +\infty)$ and $\mathfrak{B}_\beta(a + \Delta, +\infty)$. □

This section shows that it is sufficient to consider the games with the set of superpredictions S such that S is the closure of its real part $S \cap \mathbb{R}^2$.

Appendix B: Degenerate Cases

We have excluded games with sets of superpredictions S such that $S \cap \mathbb{R}^2 = \varnothing$ from consideration. However Theorem 1 remains true for these games after certain adjustments. In order to consider games of this kind, we need to impose different requirements on the oracle.

Consider a game $\mathfrak{G}$ with the set of superpredictions $S = I_1 \times \{+\infty\} \cup \{+\infty\} \times I_2$, where I_1, I_2 are intervals of the form $(a, +\infty]$ or $[a, +\infty]$. The interior of S w.r.t. the extended topology is empty, so it is constructively open. Let us require that the interiors of the intervals I_1 and I_2 are constructively open w.r.t. the oracle associated with $\mathfrak{G}$, i.e., if a sequence of rational intervals shrinks to x in the interior of I_1 or I_2, then this can be finally determined by the oracle. Under this requirement the following theorem holds.

Theorem 4. *Let $\mathfrak{G}$ be a game with the set of superpredictions S. If the real part of S is empty, i.e., $S \cap \mathbb{R}^2 = \varnothing$, then the following conditions are equivalent:*

- *there is $a \in (-\infty, +\infty]$ such that S is one of the sets $[a, +\infty] \times \{+\infty\}$, $(a, +\infty] \times \{+\infty\}$, $\{+\infty\} \times [a, +\infty]$, or $\{+\infty\} \times (a, +\infty]$*
- *$\mathfrak{G}$ is mixable*
- *there is predictive complexity w.r.t. $\mathfrak{G}$*

Proof. Clearly, it is sufficient to consider the case of closed S (cf. Lemma 2). Let $S = [a, +\infty] \times \{+\infty\} \cup \{+\infty\} \times [b, +\infty]$, where $a, b \in \mathbb{R}$. Consider the superloss process L_1 such that $L_1(\boldsymbol{x}) = a|\boldsymbol{x}|$ if $\boldsymbol{x}$ consists only of 0s and $L_1(\boldsymbol{x}) = +\infty$ otherwise and the superloss process L_2 such that $L_2(\boldsymbol{x}) = b|\boldsymbol{x}|$ if $\boldsymbol{x}$ consists only of 1s and $L_2(\boldsymbol{x}) = +\infty$ otherwise. If $\mathcal{K}$ is predictive complexity w.r.t. $\mathfrak{G}$, then $\mathcal{K} \le L_1 + C$ and $\mathcal{K} \le L_2 + C$ for some constant C. Thus the values $\mathcal{K}(\boldsymbol{x})$ should be finite for all $\boldsymbol{x}$ that consist only of 0s or only of 1s. However the pair $(\mathcal{K}(0) - \mathcal{K}(\Lambda), \mathcal{K}(1) - \mathcal{K}(\Lambda))$ is a superprediction and thus at least one of its components is infinite.

On the other hand, the image $\mathfrak{B}_\beta(S)$ consists of two orthogonal line segments and thus cannot be convex for any $\beta \in (0, 1)$.

If, e.g., $S = [a, +\infty] \times \{+\infty\}$, then $B_\beta(S)$ is a line segment and thus the game is mixable. The function $\mathcal{K}$ such that $\mathcal{K}(\boldsymbol{x}) = a|\boldsymbol{x}|$ if $\boldsymbol{x}$ consists only of 0s and $\mathcal{K}(\boldsymbol{x}) = +\infty$ otherwise is predictive complexity. □

Appendix C: Convexity, Mixability and the Existence of Complexity

This section contains results concerning the convexity of the real part of the set of superpredictions.

Lemma 4. *Let $\mathfrak{G}$ be a game with the set of superpredictions S. If $\mathfrak{G}$ is mixable, then $S \cap \mathbb{R}^2$ is convex.*

Proof. The set $\{(x, y) \in \mathbb{R}^2 \mid \beta^x + \beta^y \le 1\}$ is convex. The lemma follows from Lemma 1. □

We need a technical lemma to simplify the calculations.

Lemma 5. *Let $\mathfrak{G}_1$ be a game with the set of superpredictions S_1 and $\mathfrak{G}_2$ be a game with the set of superpredictions S_2. If the set S_2 is the shift of S_1 along a vector $(u, v) \in \mathbb{R}^2$, then $\mathfrak{G}_1$ is mixable if and only if $\mathfrak{G}_2$ is mixable. Moreover if the numbers u and v are computable, then $\mathfrak{G}_1$ specifies predictive complexity if and only if $\mathfrak{G}_2$ specifies predictive complexity.*

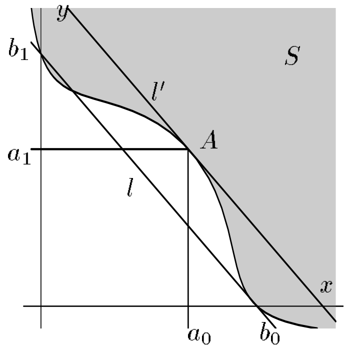

Figure 4 A set of superpredictions S that is not convex

Proof. The equivalence for mixability follows from the restatement of the concept of mixability given in Lemma 1.

The equivalence for predictive complexity is implied by the following observation. A function $L_1 : \mathbb{B}^* \to (-\infty, +\infty]$ is a superloss process w.r.t. $\mathfrak{G}_1$ if and only if $L_2(\boldsymbol{x}) = L_1(\boldsymbol{x}) + u\sharp_0\boldsymbol{x} + v\sharp_1\boldsymbol{x}$ is a superloss process w.r.t. $\mathfrak{G}_2$. $\square$

The next lemma is valid only for closed sets $S \cap \mathbb{R}^2$.

Lemma 6. *[KVV01] Let $\mathfrak{G}$ be a game with the set of superpredictions S. If $S \cap \mathbb{R}^2$ is closed w.r.t. the standard topology of $\mathbb{R}^2$, but is not convex, then there is no predictive complexity w.r.t. $\mathfrak{G}$.*

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

There exist computable points $B_0, B_1 \in S \cap \mathbb{R}^2$ such that the segment $[B_0, B_1]$ is not a subset of S. Without loss of generality (see Lemma 5) we may assume that $B_0 = (b_0, 0), B_1 = (0, b_1)$ (see Fig. 4).

Let us denote the straight line passing through B_0 and B_1 by l; it has the equation $\alpha_0 x + \alpha_1 y = \rho$, where $\alpha_0, \alpha_1, \rho > 0$. There exists a point $A = (a_0, a_1)$ with $a_0, a_1 > 0$ on the arc of boundary of S between B_0 and B_1 lying above l (see Fig. 4). 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 $\delta > 0$.

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

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

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

hold. At the same time, there is a sequence of strings $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots$ such that for any positive integer n we have $|\boldsymbol{x}_n| = n$ and

$$\mathcal{K}(\boldsymbol{x}_n) \ge a_0\sharp_0\boldsymbol{x_n} + a_1\sharp_1\boldsymbol{x_n} \ . \tag{18}$$

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 $\{(x, y) \mid x \ge a_0\}$ or $\{(x, y) \mid y \ge a_1\}$, i.e., at least one of the inequalities

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

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

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

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

$$a_0\sharp_0\boldsymbol{x}_n + a_1\sharp_1\boldsymbol{x}_n \le b_0\sharp_0\boldsymbol{x}_n + C \ , \tag{21}$$
$$a_0\sharp_0\boldsymbol{x}_n + a_1\sharp_1\boldsymbol{x}_n \le b_1\sharp_1\boldsymbol{x}_n + C \tag{22}$$

for every positive integer 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

$$b_0 = \frac{\alpha_0 a_0 + \alpha_1 a_1}{\alpha_0} - \frac{\delta}{\alpha_0} \ , \tag{23}$$
$$b_1 = \frac{\alpha_0 a_0 + \alpha_1 a_1}{\alpha_1} - \frac{\delta}{\alpha_1} \ . \tag{24}$$

If we multiply (21) by α_0/a_1, (22) by α_1/a_0, add the equations together, and substitute the expressions (23) and (24) for b_0 and b_1, we obtain

$$\frac{\delta}{a_1}\sharp_0\boldsymbol{x}_n + \frac{\delta}{a_0}\sharp_1\boldsymbol{x}_n \le C_1 \ , \tag{25}$$

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. □

Appendix D: Proof of the Main Result

In this section we prove the 'only if' part of Theorem 1, which is the main result of this paper. The proof is divided into several stages.

Introductory stage.

Assume the converse. Let $\mathfrak{G}$ be a non-mixable game with the set of superpredictions S and $\mathcal{K}$ be predictive complexity w.r.t. $\mathfrak{G}$.

Taking into account lemmas from Appendix A, we may assume that the set S is the closure of its real part $S \cap \mathbb{R}^2$. Lemma 5 implies that we can assume that $S \subseteq [0, +\infty]^2$. Finally, it follows from Lemma 6 that we may assume that the set $S \cap \mathbb{R}^2$ is convex.

The case $S = [a, +\infty] \times [c, +\infty]$ is trivial (see Subsect. 4.1). Let $f : (a, b) \to \mathbb{R}$ be the canonical specification of $\mathfrak{G}$. The function f is convex. It follows from Theorem 2 that

$$\inf_{x \in (a,b) \setminus D} \frac{\underline{f}''(x_m)}{f'(x_m)(1 - f'(x_m))} = 0 \ , \tag{26}$$

where $D \subseteq (a, b)$ is the set of points where $f'(x)$ does not exist. The requirements from Sect. 5 state that the oracle can produce a sequence $x_n \in (a, b)$, $n = 1, 2, \ldots$, such that the derivative $f'(x_n)$ exists for each n and

$$\lim_{n \to +\infty} \frac{\underline{f}''(x_n)}{f'(x_n)(1 - f'(x_n))} = 0 \ . \tag{27}$$

It can be assumed that this sequence is monotone and converges to $x_0 \in [a, b]$ (it is possible that $x_0 = b = +\infty$).

For each $m = 1, 2, \ldots$ there is a sequence $\Delta x_{m,k}$, $k = 1, 2, \ldots$, such that $\Delta x_{m,k} \to 0$ as $k \to +\infty$ and

$$f(x_m + \Delta x_{m,k}) = f(x_m) + f'(x_m)\Delta x_{m,k} + \underline{f}''(x_m)\frac{\Delta x_{m,k}^2}{2} + \nu_m(\Delta x_{m,k})\Delta x_{m,k}^2 , \tag{28}$$

where for each $m = 1, 2, \dots$ we have $\nu_m(t) \to 0$ as $t \to 0$.

For each positive integer m let

$$p_m = 1/(1 - f'(x_m)) \ . \tag{29}$$

Since $f'(x_m) < 0$, we have $p_m \in (0, 1)$. It is easy to check that the line $(1 - p_m)x + p_m y = (1 - p_m)x_m + p_m f(x_m)$ is a tangent line to S at $(x_m, f(x_m))$. A simple calculation yields

$$f'(x_m) = -\frac{1 - p_m}{p_m} \ . \tag{30}$$

Single point processes.

We need to construct some superloss processes. To make the descriptions more intuitive, we will use the term 'strategy' in the following general sence. Take a superloss process L. We say that L is the loss of a strategy $\mathfrak{A}$ that outputs the prediction $(L(\boldsymbol{x}0) - L(\boldsymbol{x}), L(\boldsymbol{x}1) - L(\boldsymbol{x}))$ on input $\boldsymbol{x}$ for all sequences $\boldsymbol{x} \in \mathbb{B}^*$.

Let $\mathfrak{A}_m$ be the strategy that always predicts $(x_m, f(x_m))$, $m = 1, 2, \dots$. We denote the loss of $\mathfrak{A}_m$ on a string $\boldsymbol{x} \in \mathbb{B}^*$ by $L_m(\boldsymbol{x})$. Now we are going to construct sequences of positive numbers $\varepsilon_{m,k}$, $m, k = 1, 2, \dots$ such that for every m we have $\varepsilon_{m,k} \to 0$ as $k \to 0$ and strategies $\mathfrak{A}_m^{\varepsilon_{m,k}}$.

The requirements from Sect. 5 imply that we can assume that all $\Delta x_{m,k}$ are of the same sign. There are two possibilities for the sign. If they are all negative, we let $\varepsilon_{m,k} = -\Delta x_{m,k}$ and if they are all positive, we let $\varepsilon_{m,k} = \Delta x_{m,k}$ for all positive integers m and k. These two possibilities are illustrated by Figures 5 and 6, where $E_m = (1 - p_m)x_m + p_m f(x_m)$ so that $(1 - p_m)x + p_m y = E_m$ is the tangent line to S at $(x_m, f(x_m))$.

Let $\xi_1^{(p)}, \xi_2^{(p)}, \dots, \xi_n^{(p)}$ be the outcomes of Bernoulli trials with the probability of success equal to p and let

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2} dx \ . \tag{31}$$

We will now show that there are functions $\alpha_m(t)$ such that for every positive integer m we have $\alpha_m(t) \to 0$ as $t \to 0$ and the inequality

$$\begin{aligned} L_m(\xi_1^{(p_m)}, \xi_2^{(p_m)}, \dots, \xi_n^{(p_m)}) - L_m^{\varepsilon_{m,k}}(\xi_1^{(p_m)}, \xi_2^{(p_m)}, \dots, \xi_n^{(p_m)}) \geq \\ -\underline{f}''(x_m)\frac{\varepsilon_{m,k}^2}{2} np_m + \varepsilon_{m,k}\sqrt{n\frac{1 - p_m}{p_m}} - \underline{f}''(x_m)\frac{\varepsilon_{m,k}^2}{2}\sqrt{np_m(1 - p_m)} \\ + \alpha_m(\varepsilon_{m,k})\varepsilon_{m,k}^2\left(\sqrt{np_m(1 - p_m)} + np_m\right) \ , \end{aligned} \tag{32}$$

holds with probability at least $\Phi(-1) - \frac{1}{\sqrt{np_m(1-p_m)}}$. Let us consider the two possibilities mentioned above separately.

The case of negative increments.

First suppose that $\Delta x_{m,k}$ are all negative and $\varepsilon_{m,k} = -\Delta x_{m,k}$. Let $\mathfrak{A}_m^{\varepsilon}$ be the strategy that always outputs the prediction $(x_m - \varepsilon, f(x_m - \varepsilon))$, $\varepsilon \in (0, x_m)$. We denote the loss of $\mathfrak{A}_m^{\varepsilon}$ on $\boldsymbol{x} \in \mathbb{B}^*$ by $L_m^{\varepsilon}(\boldsymbol{x})$. See Fig. 5 for an illustration of this case.

Take a positive integer m, a string $\boldsymbol{x}$ of length n and $\varepsilon \in (0, x_m - a)$. It is easy to see that if $\boldsymbol{x}$ contains $l^{(0)}$ zeroes or more, then $L_m^{\varepsilon}(\boldsymbol{x}) - L_m(\boldsymbol{x}) \geq \varepsilon l^{(0)} - (f(x_m -$

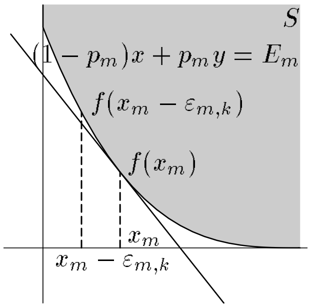

Figure 5 The case $\varepsilon_{m,k} = -\Delta x_{m,k}$

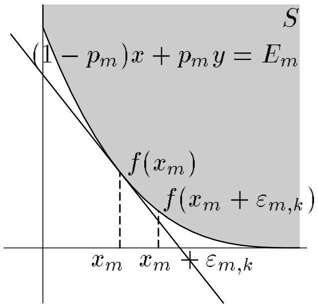

Figure 6 The case $\varepsilon_{m,k} = \Delta x_{m,k}$

$\varepsilon) - f(x_m))(n - l^{(0)})$. Let $l^{(0)} = l^{(0)}_{n,m} = n(1-p_m) + \sqrt{np_m(1-p_m)}$. Taking into account (28), for each $\boldsymbol{x}$ having $l^{(0)}_{n,m}$ zeroes or more, we get

$$\begin{aligned} L_m(\boldsymbol{x}) - L_m^{\varepsilon_{m,k}}(\boldsymbol{x}) \geq\ & \varepsilon_{m,k}\left(n(1-p_m) + \sqrt{np_m(1-p_m)}\right) \\ & - \left(-f'(x_m)\varepsilon_{m,k} + \underline{f}''(x_m)\frac{\varepsilon^2_{m,k}}{2} + \nu_m(-\varepsilon_{m,k})\varepsilon^2_{m,k}\right) \qquad (33) \\ & \times \left(np_m - \sqrt{np_m(1-p_m)}\right) \\ =\ & -\underline{f}''(x_m)\frac{\varepsilon^2_{m,k}}{2}np_m + \varepsilon_{m,k}\sqrt{n\frac{1-p_m}{p_m}} \\ & + \underline{f}''(x_m)\frac{\varepsilon^2_{m,k}}{2}\sqrt{np_m(1-p_m)} \qquad (34) \\ & + \nu_m(-\varepsilon_{m,k})\varepsilon^2_{m,k}\left(\sqrt{np_m(1-p_m)} - np_m\right) \ . \end{aligned}$$

In order to estimate the probability of getting at least $l^{(0)}_{n,m}$ zeroes among the outcomes $\xi^{(p)}_1, \xi^{(p)}_2, \ldots, \xi^{(p)}_n$ of Bernoulli trials with the probability of success equal to p, we will rely on the following proposition, which is a special case of the Berry-Essen inequality.

Proposition 1. *Let* $p \in (0,1)$. *If* $S_n = \xi^{(p)}_1 + \xi^{(p)}_2 + \ldots + \xi^{(p)}_n$ *and*

$$F^{(p)}_n(x) = \Pr\left\{\frac{S_n - np}{\sqrt{np(1-p)}} \leq x\right\} \ , \tag{35}$$

then

$$\sup_{-\infty\leq x\leq+\infty} |F^{(p)}_n(x) - \Phi(x)| \leq \frac{p^2+(1-p)^2}{\sqrt{np(1-p)}} \tag{36}$$

for all $n = 1, 2, \ldots$.

A more general statement of the Berry-Essen Inequality and a proof may be found in [Pet75], Sect. V.2.

It is easy to see that

$$\Pr\{S_n \leq np - \sqrt{np(1-p)}\} = \Pr\left\{\frac{S_n - np}{\sqrt{np(1-p)}} \leq -1\right\} \tag{37}$$

$$= F^{(p)}_n(-1) \tag{38}$$

and $p^2 + (1-p)^2 \leq 1$ for all $p \in (0,1)$. Thus the probability to get $l^{(0)}_{n,m}$ or more zeroes among $\xi^{(p)}_1, \xi^{(p)}_2, \ldots, \xi^{(p)}_n$ is at least $\Phi(-1) - \frac{1}{\sqrt{np(1-p)}}$.

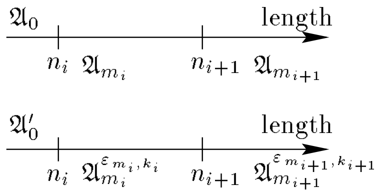

Figure 7 The operation of the strategies $\mathfrak{A}_0$ and $\mathfrak{A}'_0$

The case of positive increments.

Now suppose that all $\Delta x_{m,k}$ are positive. Let $\varepsilon_{m,k} = \Delta x_{m,k}$ and let $\mathfrak{A}_m^\varepsilon$ be the strategy that always predicts $(x_m + \varepsilon, f(x_m + \varepsilon))$, $\varepsilon \in (x_m, b)$. As above, we denote the loss of $\mathfrak{A}_m^\varepsilon$ on $\boldsymbol{x} \in \mathbb{B}^*$ by L_m^ε. See Fig. 6 for an illustration of this case.

If $\boldsymbol{x}$ of length n has $l^{(1)}$ ones or more, then $L_m^\varepsilon(\boldsymbol{x}) - L_m(\boldsymbol{x}) \geq (f(x_m) - f(x_m + \varepsilon))l^{(1)} - \varepsilon(n - l^{(1)})$. Let $l^{(1)} = l_{n,m}^{(1)} = np_m + \sqrt{np_m(1-p_m)}$. Taking into account (28), for each $\boldsymbol{x}$ having $l_{n,m}^{(1)}$ ones or more, we get

$$
\begin{aligned}
L_m(\boldsymbol{x}) - L_m^{\varepsilon_{m,k}}(\boldsymbol{x}) \geq &- \left(f'(x_m)\varepsilon_{m,k} + f''(x_m)\frac{\varepsilon_{m,k}^2}{2} + \nu_m(\varepsilon_{m,k})\varepsilon_{m,k}^2 \right) \\
&\times \left(np_m + \sqrt{np_m(1-p_m)} \right) \\
&- \varepsilon_{m,k} \left(n(1-p_m) - \sqrt{np_m(1-p_m)} \right) \qquad (39)\\
= &-\underline{f}''(x_m)\frac{\varepsilon_{m,k}^2}{2} np_m + \varepsilon_{m,k}\sqrt{n\frac{1-p_m}{p_m}} \\
&- \underline{f}''(x_m)\frac{\varepsilon_{m,k}^2}{2}\sqrt{np_m(1-p_m)} \qquad (40)\\
&+ \nu_m(\varepsilon_{m,k})\varepsilon_{m,k}^2 \left(\sqrt{np_m(1-p_m)} + np_m \right) \enspace .
\end{aligned}
$$

Arguing as above, one can show that the probability to get $l_{n,m}^{(1)}$ or more ones among $\xi_1^{(p)}, \xi_2^{(p)}, \ldots, \xi_n^{(p)}$ is at least $\Phi(-1) - \frac{1}{\sqrt{np(1-p)}}$.

Approaching the contradiction.

It follows from Lemma 8 that

$$
\begin{aligned}
\mathbf{E}\mathcal{K}\left(\xi_1^{(p_m)}, \xi_2^{(p_m)}, \ldots, \xi_n^{(p_m)}\right) &\geq \mathbf{E}L_m\left(\xi_1^{(p_m)}, \xi_2^{(p_m)}, \ldots, \xi_n^{(p_m)}\right) \qquad (41)\\
&= ((1-p_m)x_m + p_m f(x_m))\, n \enspace , \qquad (42)
\end{aligned}
$$

where $\xi_1^{(p_m)}, \xi_2^{(p_m)}, \ldots, \xi_n^{(p_m)}$ are results of n independent Bernoulli trials with the probability of success equal to p_m. We are going to obtain a contradiction by beating this estimate through the use of (32). In order to do this, we will construct two strategies, $\mathfrak{A}_0$ and $\mathfrak{A}'_0$ that work as follows. There is a strictly increasing sequence of positive integers $n_1 < n_2 < \ldots$ and for each $\boldsymbol{x}$ of length n such that $n_{i-1} < n \leq n_i$ the strategy $\mathfrak{A}_0$ simulates $\mathfrak{A}_{m_i}$ while $\mathfrak{A}'_0$ simulates $\mathfrak{A}_{m_i}^{\varepsilon_{m_i,k_i}}$ for some m_i and k_i. Figure 7 presents a diagram of how the strategies work. We denote the loss of $\mathfrak{A}_0$ by L_0 and the loss of $\mathfrak{A}'_0$ by L'_0.

There is a number of possibilities for the location of the limit point $x_0 \in [a, b]$. One of the following statements is true:

1. $x_0 = a$,

2. $x_0 \in (a, b)$, or

3. $x_0 = b$ (note that b may be equal to $+\infty$).

The first case can be reduced to the others by mirroring S w.r.t. the straight line $x = y$ (recall the remark about mirroring at the end of Sect. 5). In the second case $\underline{f}''(x_n) \to 0$ since $f'(x)$ is separated from 0 in a vicinity of x_0. In the third case there are two possibilities, either $f'(x_n)$ is separated from zero and, respectively, $\underline{f}''(x_n) \to 0$, or $f'(x_n) \to 0$ and thus $\underline{f}''(x_n)/f'(x_n) \to 0$ as $n \to +\infty$. The first possibility is not essentially different from the second case above, while the other one requires a separate treatment.

We obtain two essentially different situations to consider:

1. $x_m \to x_0 \in (a, b]$, there is $c > 0$ such that $f'(x_m) \leq -c < 0$ for all $m = 1, 2, \ldots$, and $\underline{f}''(x_m) \to 0$ as $m \to +\infty$

2. $x_m \to b$, $f'(x_m) \to 0$, and $\underline{f}''(x_m)/f'(x_m) \to 0$ as $m \to +\infty$.

In the first case, we can assume that all predictions made by $\mathfrak{A}_m$ and $\mathfrak{A}_m^{\varepsilon_{m,k}}$, where m and k are positive integers, belong to some finite square $[0, M]^2 \subseteq \mathbb{R}^2$, where $0 < M < +\infty$. On the other hand, it follows from the definition (29) that p_m converges to $p_0 \in (0, 1)$ as $m \to +\infty$.

In the second case we cannot guarantee that all predictions are taken from some bounded area. It is possible that $x_0 = +\infty$. We will rely on monotonicity of x_m. As far as the probabilities are concerned, (29) implies that $p_m \to 1$ as $m \to +\infty$.

Let us consider the cases one by one.

The case $f'(x_m) \not\to 0$.

Suppose that $\mathfrak{A}_0$ and $\mathfrak{A}'_0$ have already been constructed for all x of length less than or equal to n_{i-1}. Consider the following set of inequalities:

$$1 \leq \underline{f}''(x_m)\varepsilon_{m,k}^2 n \leq 2 \ , \tag{43}$$

$$n_{i-1} < n \ , \tag{44}$$

$$M n_{i-1} \leq \frac{1}{4\sqrt{\underline{f}''(x_m)}} \frac{\Phi(-1)}{2} \sqrt{\frac{1 - p_m}{p_m}} \ , \tag{45}$$

$$|\alpha_m(\varepsilon_{m,k})| \leq \underline{f}''(x_m) \ . \tag{46}$$

They are consistent, i.e., there are positive integers n, m, and k that satisfy all these inequalities. For example, first let m to be sufficiently large to satisfy (45) and then tune k and n.

Now take n_i, m_i, and k_i equal to some n, m, and k satisfying the inequalities. Since the behaviour of $\mathfrak{A}_0$ differs from that of $\mathfrak{A}_{m_i}$ and the behaviour of $\mathfrak{A}'_0$ differs from that of $\mathfrak{A}_{m_i}^{\varepsilon_{m_i,k_i}}$ only on strings of length less than or equal to n_{i-1}, it is easy to see that for all $\boldsymbol{x}$ such that $n_{i-1} < |\boldsymbol{x}| \leq n_i$ we have

$$|L_0(\boldsymbol{x}) - L_{m_i}(\boldsymbol{x})| \leq 2M n_{i-1} \leq \frac{1}{2\sqrt{\underline{f}''(x_m)}} \frac{\Phi(-1)}{2} \sqrt{\frac{1 - p_{m_i}}{p_{m_i}}} \ , \tag{47}$$

$$|L'_0(\boldsymbol{x}) - L_{m_i}^{\varepsilon_{m_i,k_i}}(\boldsymbol{x})| \leq 2M n_{i-1} \leq \frac{1}{2\sqrt{\underline{f}''(x_m)}} \frac{\Phi(-1)}{2} \sqrt{\frac{1 - p_{m_i}}{p_{m_i}}} \ . \tag{48}$$

On the other hand, it follows from (32) that

$$L_{m_i}\left(\xi_1^{(p_{m_i})},\xi_2^{(p_{m_i})},\ldots,\xi_{n_i}^{(p_{m_i})}\right) - L_{m_i}^{\varepsilon_{m_i},k_i}\left(\xi_1^{(p_{m_i})},\xi_2^{(p_{m_i})},\ldots,\xi_{n_i}^{(p_{m_i})}\right) \geq \frac{1}{\sqrt{\underline{f}''(x_{m_i})}}\sqrt{\frac{1-p_{m_i}}{p_{m_i}}} - 3\left(p_{m_i}+\sqrt{\frac{p_{m_i}(1-p_{m_i})}{n_i}}\right) \quad (49)$$

with probability at least $\Phi(-1)-\frac{1}{\sqrt{n_i p_{m_i}(1-p_{m_i})}}$. If i is sufficiently large, (49) holds with probability at least $\Phi(-1)/2$.

It follows from the definition of predictive complexity (2), that there is a constant C such that

$$\mathcal{K}(\boldsymbol{x}) \leq \min(L_0(\boldsymbol{x}), L_0'(\boldsymbol{x})) + C \quad (50)$$

holds for all strings $\boldsymbol{x}\in\mathbb{B}^*$. Combining this with (47) and (48), we get

$$\mathcal{K}(\boldsymbol{x}) \leq \min(L_{m_i}(\boldsymbol{x}), L_{m_i}^{\varepsilon_{m_i},k_i}(\boldsymbol{x})) + \frac{1}{2\sqrt{\underline{f}''(x_m)}}\frac{\Phi(-1)}{2}\sqrt{\frac{1-p_{m_i}}{p_{m_i}}} + C \ . \quad (51)$$

for all $\boldsymbol{x}$ such that $n_{i-1}<|\boldsymbol{x}|\leq n_i$; note that the constant C does not depend on i. Consider the upper bound on expectations $\mathbf{E}\mathcal{K}\left(\xi_1^{p_{m_i}},\xi_2^{p_{m_i}},\ldots,\xi_{n_i}^{p_{m_i}}\right)$ given by (51). Taking into account lower estimate (41) and inequality (49), we get

$$\left(\frac{1}{2\sqrt{\underline{f}''(x_{m_i})}}\sqrt{\frac{1-p_{m_i}}{p_{m_i}}} - 3\left(p_{m_i}+\sqrt{\frac{p_{m_i}(1-p_{m_i})}{n_i}}\right)\right)\frac{\Phi(-1)}{2} \leq C \quad (52)$$

for all sufficiently large i. Since $f''(x_m)\to 0+$ and $p_m\to p_0$, where $p_0\in(0,1)$, as $m\to\infty$, the left-hand side tends to $+\infty$ as $i\to+\infty$. This is a contradiction.

The case $f'(x_m)\to 0$.

In the second case there are two essential differences. First instead of $f''(x_m)\to 0$ we are going to rely on the convergence

$$\frac{\underline{f}''(x_m)p_m}{1-p_m} = -\frac{\underline{f}''(x_m)}{f'(x_m)} \to 0+ \quad (53)$$

as $m\to+\infty$. Secondly since x_m may converge to $+\infty$, we cannot rely on the bounds similar to (47) and (48); we will need a finer estimate instead.

Let $\theta = x_1+|x_1-f(x_1)|$. Arguing as above, suppose that $\mathfrak{A}_0$ and $\mathfrak{A}_0'$ have already been constructed for all x of length less than or equal to n_{i-1}. Consider the following set of inequalities:

$$1 \leq \underline{f}''(x_m)\varepsilon_{m,k}^2 n \leq 2 \ , \quad (54)$$

$$n_{i-1} < n \ , \quad (55)$$

$$\frac{1}{\sqrt{np_m(1-p_m)}} \leq \frac{\Phi(-1)}{2} \ , \quad (56)$$

$$2n_{i-1}(\theta+1) \leq \frac{\Phi(-1)}{4}\sqrt{\frac{1-p_m}{\underline{f}''(x_m)p_m}} \ , \quad (57)$$

$$\varepsilon_{m,k} \leq 1 \ , \quad (58)$$

$$|\alpha_m(\varepsilon_{m,k})| \le \underline{f}''(x_m) \ . \tag{59}$$

They are consistent, i.e., there are positive integers n, m, and k that satisfy these inequalities. For example, first let m to be sufficiently large to satisfy (56) and (57) and then tune k and n.

Now take n_i, m_i, and k_i equal to some n, m, and k satisfying the inequalities (recall that the way $\mathfrak{A}_0$ and $\mathfrak{A}_0'$ operate is illustrated by Fig. 7). It follows from (32) that

$$L_{m_i}\left(\xi_1^{(p_{m_i})},\xi_2^{(p_{m_i})},\ldots,\xi_{n_i}^{(p_{m_i})}\right) - L_{m_i}^{\varepsilon_{m_i},k_i}\left(\xi_1^{(p_{m_i})},\xi_2^{(p_{m_i})},\ldots,\xi_{n_i}^{(p_{m_i})}\right) \ge \sqrt{\frac{1-p_{m_i}}{\underline{f}''(x_{m_i})p_{m_i}}} - 3\left(p_{m_i} + \sqrt{\frac{p_{m_i}(1-p_{m_i})}{n_i}}\right) \tag{60}$$

with probability at least $\Phi(-1)/2$. In fact, (60) is the same as (49), we just group some factors differently. Let $\Xi_{n_i} \subseteq \mathbb{B}^{n_i}$ be the set of all strings $\boldsymbol{x}$ of length n_i such that (60) holds on $\boldsymbol{x}$, i.e., (60) holds if and only if $\xi_1^{(p_{m_i})}\xi_2^{(p_{m_i})}\ldots\xi_{n_i}^{(p_{m_i})} \in \Xi_{n_i}$,where $i = 1,2,\ldots$.

Instead of (50), we are going to use the following elaboration. We use L_0' to estimate the value $\mathcal{K}(\boldsymbol{x})$ on $\boldsymbol{x} \in \Xi_{n_i}$ and L_0 on $\boldsymbol{x} \in \mathbb{B}^{n_i} \setminus \Xi_{n_i}$. There is a constant C such that

$$\mathbf{E}\mathcal{K}\left(\xi_1^{(p_{m_i})},\xi_2^{(p_{m_i})},\ldots,\xi_{n_i}^{(p_{m_i})}\right) \le \sum_{\boldsymbol{x}\in\Xi_{n_i}} L_0'(\boldsymbol{x})\Pr{}_{p_{m_i}}(\boldsymbol{x}) + \sum_{\boldsymbol{x}\in\mathbb{B}^{n_i}\setminus\Xi_{n_i}} L_0(\boldsymbol{x})\Pr{}_{p_{m_i}}(\boldsymbol{x}) + C \tag{61}$$

for all $i = 1,2,\ldots$, where $\Pr_p(\boldsymbol{x})$ is the probability of the event $\xi_1^{(p)}\xi_2^{(p)}\ldots\xi_n^{(p)} = \boldsymbol{x}$.

In order to obtain an estimate in terms of L_{m_i} and $L_{m_i}^{\varepsilon_{m_i},k_i}$, consider the equality

$$\sum_{\boldsymbol{x}\in\Xi_{n_i}} L_0'(\boldsymbol{x})\Pr{}_{p_{m_i}}(\boldsymbol{x}) + \sum_{\boldsymbol{x}\in\mathbb{B}^{n_i}\setminus\Xi_{n_i}} L_0(\boldsymbol{x})\Pr{}_{p_{m_i}}(\boldsymbol{x}) = \sum{}'(L_0'(\tilde{\boldsymbol{x}}\bar{\boldsymbol{x}}) - L_0'(\tilde{\boldsymbol{x}}))\Pr{}_{p_{m_i}}(\tilde{\boldsymbol{x}}\bar{\boldsymbol{x}}) + \sum{}''(L_0(\tilde{\boldsymbol{x}}\bar{\boldsymbol{x}}) - L_0(\tilde{\boldsymbol{x}}))\Pr{}_{p_{m_i}}(\tilde{\boldsymbol{x}}\bar{\boldsymbol{x}}) + \sum{}' L_0'(\tilde{\boldsymbol{x}})\Pr{}_{p_{m_i}}(\tilde{\boldsymbol{x}}\bar{\boldsymbol{x}}) + \sum{}'' L_0(\tilde{\boldsymbol{x}})\Pr{}_{p_{m_i}}(\tilde{\boldsymbol{x}}\bar{\boldsymbol{x}}) \ , \tag{62}$$

where the sum $\sum'$ is taken over all $\tilde{\boldsymbol{x}}$ of length n_{i-1} and $\bar{\boldsymbol{x}}$ of length $n_i - n_{i-1}$ such that $\tilde{\boldsymbol{x}}\bar{\boldsymbol{x}} \in \Xi_{n_i}$ while the sum $\sum''$ is taken over all $\tilde{\boldsymbol{x}}$ of length n_{i-1} and $\bar{\boldsymbol{x}}$ of length $n_i - n_{i-1}$ such that $\tilde{\boldsymbol{x}}\bar{\boldsymbol{x}} \in \mathbb{B}^{n_i} \setminus \Xi_{n_i}$. The formula will hold if we replace L_0 by L_{m_i} and L_0' by $L_{m_i}^{\varepsilon_{m_i},k_i}$. Since by construction we have

$$L_0(\tilde{\boldsymbol{x}}\bar{\boldsymbol{x}}) - L_0(\tilde{\boldsymbol{x}}) = L_{m_i}(\tilde{\boldsymbol{x}}\bar{\boldsymbol{x}}) - L_{m_i}(\tilde{\boldsymbol{x}}) \ , \tag{63}$$

$$L_0'(\tilde{\boldsymbol{x}}\bar{\boldsymbol{x}}) - L_0'(\tilde{\boldsymbol{x}}) = L_{m_i}^{\varepsilon_{m_i},k_i}(\tilde{\boldsymbol{x}}\bar{\boldsymbol{x}}) - L_{m_i}^{\varepsilon_{m_i},k_i}(\tilde{\boldsymbol{x}}) \tag{64}$$

for all $\tilde{\boldsymbol{x}}$ of length n_{i-1} and $\bar{\boldsymbol{x}}$ of length $n_i - n_{i-1}$, we get

$$\left|\left(\sum_{\boldsymbol{x}\in\Xi_{n_i}} L_0'(\boldsymbol{x})\Pr{}_{p_{m_i}}(\boldsymbol{x}) + \sum_{\boldsymbol{x}\in\mathbb{B}^{n_i}\setminus\Xi_{n_i}} L_0(\boldsymbol{x})\Pr{}_{p_{m_i}}(\boldsymbol{x})\right) - \left(\sum_{\boldsymbol{x}\in\Xi_{n_i}} L_{m_i}^{\varepsilon_{m_i},k_i}(\boldsymbol{x})\Pr{}_{p_{m_i}}(\boldsymbol{x}) + \sum_{\boldsymbol{x}\in\mathbb{B}^{n_i}\setminus\Xi_{n_i}} L_{m_i}(\boldsymbol{x})\Pr{}_{p_{m_i}}(\boldsymbol{x})\right)\right| \le$$

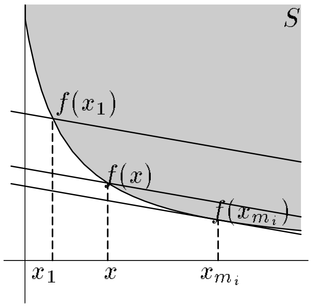

Figure 8 The estimate on expectations

$$\begin{aligned}
&\mathbf{E}\max\left(L_0\left(\xi_1^{p_{m_i}},\xi_2^{p_{m_i}},\dots,\xi_{n_{i-1}}^{p_{m_i}}\right),L_0'\left(\xi_1^{p_{m_i}},\xi_2^{p_{m_i}},\dots,\xi_{n_{i-1}}^{p_{m_i}}\right)\right)\\
&+\mathbf{E}\max\left(L_{m_i}\left(\xi_1^{p_{m_i}},\xi_2^{p_{m_i}},\dots,\xi_{n_{i-1}}^{p_{m_i}}\right),L_{m_i}^{\varepsilon_{m_i,k_i}}\left(\xi_1^{p_{m_i}},\xi_2^{p_{m_i}},\dots,\xi_{n_{i-1}}^{p_{m_i}}\right)\right)\le\\
&\mathbf{E}L_0\left(\xi_1^{p_{m_i}},\xi_2^{p_{m_i}},\dots,\xi_{n_{i-1}}^{p_{m_i}}\right)+\mathbf{E}L_{m_i}\left(\xi_1^{p_{m_i}},\xi_2^{p_{m_i}},\dots,\xi_{n_{i-1}}^{p_{m_i}}\right)+2n_{i-1}\ . \quad (65)
\end{aligned}$$

The last inequality follows from (58) and the following observation: for all $\boldsymbol{x}\in\mathbb{B}^*$ and positive integers m and k, we have $|L_m(\boldsymbol{x})-L_m^{\varepsilon_{m,k}}(\boldsymbol{x})|\le\varepsilon_{m,k}|\boldsymbol{x}|$.

We will now show that

$$\mathbf{E}L_0\left(\xi_1^{p_{m_i}},\xi_2^{p_{m_i}},\dots,\xi_{n_{i-1}}^{p_{m_i}}\right)\le\theta n_{i-1}, \quad (66)$$

$$\mathbf{E}L_{m_i}\left(\xi_1^{p_{m_i}},\xi_2^{p_{m_i}},\dots,\xi_{n_{i-1}}^{p_{m_i}}\right)\le\theta n_{i-1}\ . \quad (67)$$

Indeed, while the length is less than n_{i-1}, the strategies $\mathfrak{A}_0$ and $\mathfrak{A}_{m_i}$ output predictions $(x,f(x))$, where $x\in[x_1,x_{m_i}]$. Since S is convex, all such points $(x,f(x))$ lie above the tangent line $(1-p_{m_i})x+p_{m_i}y=(1-p_{m_i})x_{m_i}+p_{m_i}f(x_{m_i})$ but below the parallel line $(1-p_{m_i})x+p_{m_i}y=(1-p_{m_i})x_1+p_{m_i}f(x_1)$ (see Fig. 8). Hence

$$\begin{aligned}
(1-p_{m_i})x+p_{m_i}f(x)&\le(1-p_{m_i})x_1+p_{m_i}f(x_1) \quad (68)\\
&\le\theta\ . \quad (69)
\end{aligned}$$

This provides us with the desired inequalities.

Taking into account (57), we get

$$\begin{aligned}
&\mathbf{E}\mathcal{K}\left(\xi_1^{(p_{m_i})},\xi_2^{(p_{m_i})},\dots,\xi_{n_i}^{(p_{m_i})}\right)\le\\
&\qquad\sum_{\boldsymbol{x}\in\Xi_{n_i}}L_{m_i}^{\varepsilon_{m_i,k_i}}(\boldsymbol{x})\Pr{}_{p_{m_i}}(\boldsymbol{x})+\sum_{\boldsymbol{x}\in\mathbb{B}^{n_i}\setminus\Xi_{n_i}}L_{m_i}(\boldsymbol{x})\Pr{}_{p_{m_i}}(\boldsymbol{x})+\\
&\qquad\qquad\frac{\Phi(-1)}{4}\sqrt{\frac{1-p_m}{\underline{f}''(x_m)p_m}}+C\ . \quad (70)
\end{aligned}$$

Comparison with (41) and (60) yields

$$\left(\frac{1}{2}\sqrt{\frac{1-p_{m_i}}{\sqrt{\underline{f}''(x_{m_i})}p_{m_i}}}-3\left(p_{m_i}+\sqrt{\frac{p_{m_i}(1-p_{m_i})}{n_i}}\right)\right)\frac{\Phi(-1)}{2}\le C\ . \quad (71)$$

The left-hand side tends to infinity as i tends to $+\infty$ and this is a contradiction.

Appendix E: A Derivation of the Analytical Characterization

Proof of Theorem 3. Let S be the set of superpredictions for S. By definition $\mathfrak{G}$ is β-mixable if and only if the set $\mathfrak{B}_\beta(S)$ (see (3) for the definition of $\mathfrak{B}_\beta$) is convex.

Taking into account Lemma 4, assume that f is convex.

The set $\mathfrak{B}_\beta(S)$ is the closure of its non-empty interior and thus it is convex if and only if through each point of the boundary $\partial\mathfrak{B}_\beta(S)$ there passes a support hyperplane (line). In fact, it is sufficient to have *weak support line* at each boundary point; a line l is called a weak support line for the set M at a point P if there is a vicinity N of the point P such that $M \cap N$ is not cut by l (see, e.g., [Val64], Theorem 4.10).

Let g be the function defined by $g(u) = \beta^{f(\log_\beta u)}$, $u \in (\log_\beta a, \log_\beta b)$. The graph of g is the image of the graph of f under $\mathfrak{B}_\beta$.

Take a point $x_0 \in D$. Since f is convex, $f'_l(x_0) < f'_r(x_0)$. It is easy to see that $g'_l(\beta^{x_0}) > g'_r(\beta^{x_0})$. Indeed, differentiation yields

$$\begin{aligned} \frac{d\beta^y}{d\beta^x} &= \frac{\ln\beta \cdot \beta^y y'}{\ln\beta \cdot \beta^x x'} \\ &= \beta^{y-x}\frac{y'}{x'} \ . \end{aligned} \tag{72}$$

This implies that $\mathfrak{B}_\beta(S)$ has a weak support line at $(\beta^{x_0}, g(\beta^{x_0}))$.

It remains to consider $x_0 \in (a,b) \setminus D$. The inverse image of a line $rx + sy = t$ where $r, s, t > 0$ w.r.t. the transformation $\mathfrak{B}_\beta$ is a shift of the curve $\beta^x + \beta^y = 1$. The set $\mathfrak{B}_\beta(S)$ has a weak support line at $\mathfrak{B}_\beta((x_0, f(x_0)))$ if and only if there is a shift $y = B(x)$ of the curve $\beta^x + \beta^y = 1$ passing through $(x_0, f(x_0))$ and a vicinity V of the point $(x_0, f(x_0))$ such that $S \cap V$ lies above the shift.

The tangent lines to such a shift of $\beta^x + \beta^y = 1$ at to S at $(x_0, f(x_0))$ should coincide. We need to take the second derivatives into account.

The equation $\beta^x + \beta^y = 1$ can be rewritten as $y = \log_\beta(1 - \beta^x)$. Differentiation yields

$$y'_x = \frac{1}{1 - \beta^{-x}} \ , \tag{73}$$

$$y''_{xx} = -\frac{\beta^{-x}\ln\beta}{(1 - \beta^{-x})^2} \ . \tag{74}$$

It is easy to check that $y''_{xx} = y'_x(1 - y'_x)\ln\beta$. Thus if $y = B(x)$ is the shift of $\beta^x + \beta^y = 1$ that has the same derivative as f at x_0, then $B''(x_0) = f'(x_0)(1 - f(x_0))\ln\beta$.

The above implies that if $\mathfrak{B}_\beta(S)$ has a weak support line at $(\beta^{x_0}, \beta^{f(x_0)})$, then $B''(x_0) \geq \underline{f}''(x_0)$. This proves the 'if' part of the theorem.

On the other hand, if $B''(x_0) > \underline{f}''(x_0)$, then $\mathfrak{B}_\beta(S)$ has a weak support line at $(\beta^{x_0}, \beta^{f(x_0)})$. This implies that under the conditions of the theorem $\mathfrak{G}$ is $\tilde{\beta}$-mixable for each $\tilde{\beta}$ such that $\beta < \tilde{\beta} < 1$. The 'only if' part of the theorem follows by continuity. Indeed, if the set $\mathfrak{B}_{\beta'}(S)$ is convex for all $\beta' > \beta$, then the set $\mathfrak{B}_\beta(S)$ is convex too. □

Appendix F: On the Existence of Complexity for Mixable Games

In this section we prove the 'if' part of Theorem 1. The proof essentially repeats the proof from [VW98], but we include it for completeness.

Let $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$ be a game with the set of superpredictions S. The lemmas from Appendix A imply that it is sufficient to consider the case when S is the closure of its finite part, i.e., $S = \overline{S \cap \mathbb{R}^2}$.

By definition, a superloss process is semicomputable from above. There are different equivalent definitions of semicomputability; let us assume the following, which is convenient for this proof. A function $L : \mathbb{B}^* \to [-\infty, +\infty]$ is semicomputable from above if there is a recursively enumerable set $W \subseteq \mathbb{B}^* \times (\mathbb{Q} \cup \{+\infty\})$ ($\mathbb{Q}$ denotes the set of rational numbers) such that

$$L(\boldsymbol{x}) = \inf\{t \mid (\boldsymbol{x}, t) \in W\} \tag{75}$$

for all $\boldsymbol{x} \in \mathbb{B}^*$. Here we assume that $\inf \varnothing = +\infty$. Note that since for all superloss processes $L(\Lambda) = 0$ by definition, we can restrict the condition (75) to non-empty $\boldsymbol{x}$.

For this proof we need the following definition. A superloss process L is *strict* if for each string $\boldsymbol{x}$ the superprediction $(L(\boldsymbol{x}0) - L(\boldsymbol{x}), L(\boldsymbol{x}1) - L(\boldsymbol{x}))$ belongs to the interior of S. In the definition of predictive complexity $\mathcal{K}$ it is sufficient to require that $\mathcal{K}$ is smaller than every strict superloss process, since for every superloss process L the function $\tilde{L}(\boldsymbol{x}) = L(\boldsymbol{x}) + (1 - 1/|\boldsymbol{x}|)$ is a strict superloss process and $|L(\boldsymbol{x}) - \tilde{L}(\boldsymbol{x})| \leq 1$.

The proof is in two steps. First we show that there exist an effective enumeration $L_1, L_2, \ldots$ such that for each positive integer i the function L_i is a superloss process and each strict superloss process appears in this enumeration. The effectiveness means that given a positive integer i we can construct an algorithm that enumerates the set $W_i \subseteq \mathbb{B} \times (\mathbb{Q} \cup \{+\infty\})$ such that $W = W_i$ defines $L = L_i$ by (75). Secondly we show that the processes from the enumeration can be mixed to construct the predictive complexity.

Let us construct the enumeration. There is a universal recursively enumerable set $U \subseteq \mathbb{B}^* \times (\mathbb{Q} \cup \{+\infty\}) \times \mathbb{N}$, where $\mathbb{N}$ is the set of positive integers, i.e., U is a recursively enumerable set such that for every recursively enumerable $W \subseteq \mathbb{B}^* \times (\mathbb{Q} \cup \{+\infty\})$ there is $i \in \mathbb{N}$ such that $W = W_i = \{(\boldsymbol{x}, q) \mid (\boldsymbol{x}, q, i) \in U\}$. Each set W_i defines a function L_i by (75). We will now describe an effective procedure that reconstructs an enumeration of W into an enumeration of $W' \subseteq W$ with the following property: W' defines a superloss process and if W defines a strict superloss process, then W' defines the same strict superloss process.

We run the algorithm enumerating W and output elements of W'. Let $\widetilde{W'}$ be the part of W' output so far (note that $\widetilde{W'}$ is finite). Suppose that a pair $(\boldsymbol{x}, t_{\boldsymbol{x}}) \in W$, $\boldsymbol{x} \neq \Lambda$, arrives. We will call it a *request*. A request is processed as follows. If $\boldsymbol{x} = \Lambda$, the request is immediately output. If $t_{\boldsymbol{x}} \geq \min\{t \mid (\boldsymbol{x}, t) \in \widetilde{W'}\}$ (here $\min \varnothing = +\infty$), the request is discarded. Otherwise let ω be the last bit of $\boldsymbol{x}$, let $\boldsymbol{y}$ be the prefix of $\boldsymbol{x}$ just one bit shorter than $\boldsymbol{x}$, and let $\boldsymbol{z}$ be $\boldsymbol{y}$ appended with the bit different from ω. Let $t_{\boldsymbol{y}} = \min\{t \mid (\boldsymbol{y}, t) \in \widetilde{W'}\}$ if $\boldsymbol{y} \neq \Lambda$ and 0 otherwise and let $t_{\boldsymbol{z}} = \min\{t \mid (\boldsymbol{z}, t) \in \widetilde{W'}\}$. Consider the pair $P_{\boldsymbol{x}}$ that equals $(t_{\boldsymbol{x}} - t_{\boldsymbol{y}}, t_{\boldsymbol{z}} - t_{\boldsymbol{y}})$ if $\omega = 0$ and $(t_{\boldsymbol{z}} - t_{\boldsymbol{y}}, t_{\boldsymbol{x}} - t_{\boldsymbol{y}})$ otherwise.

Let us now check whether $P_{\boldsymbol{x}}$ belongs to the interior of S. Since the interior of S is constructively open (see Sect. 5), we can check it effectively in the limit. We start the algorithm that checks it. Since it may work infinitely long in the case when the pair is not in the interior, we run the algorithm 'in the background', i.e., we perform other activities while the algorithm works. At each moment we have a number of requests processed in this way. When we get a positive answer, i.e., we see that the pair $P_{\boldsymbol{x}}$ belongs to the interior of S, we output $(\boldsymbol{x}, t_{\boldsymbol{x}})$ and restart processing all the requests taking the amendment to $\widetilde{W'}$ into account. From time to time we pause processing the requests to get a new pair from the enumeration of W.

It is easy to see that if W defines a strict superloss process, then W' defines the same process. On the other hand, since by assumption S is closed, W' always defines a superloss process. Thus we have constructed the enumeration we wanted.

It follows from the definition of mixability that $\mathfrak{G}$ is β-mixable with some $\beta \in (0,1)$. Let

$$\mathcal{K} = \log_\beta \sum_{i=1}^{\infty} \frac{1}{2^i} \beta^{L_i} \quad . \tag{76}$$

Let us show that $\mathcal{K}$ is a superloss process. For each $\omega = 0, 1$ we have

$$\begin{aligned} \mathcal{K}(\boldsymbol{x}\omega) - \mathcal{K}(\boldsymbol{x}) &= \log_\beta \sum_{i=1}^{\infty} \frac{1}{2^i} \beta^{L_i(\boldsymbol{x}\omega)} - \log_\beta \sum_{i=1}^{\infty} \frac{1}{2^i} \beta^{L_i(\boldsymbol{x})} \\ &= \log_\beta \sum_{i=1}^{\infty} \frac{\beta^{L_i(\boldsymbol{x})}}{2^i \sum_{j=1}^{\infty} \frac{1}{2^j} \beta^{L_i(\boldsymbol{x})}} \beta^{L(\boldsymbol{x}\omega) - L(\boldsymbol{x})} \quad . \end{aligned}$$

The point $\mathfrak{B}_\beta(\mathcal{K}(\boldsymbol{x}1) - \mathcal{K}(\boldsymbol{x}), \mathcal{K}(\boldsymbol{x}0) - \mathcal{K}(\boldsymbol{x}))$ is thus a convergent convex mixture of the points $\mathfrak{B}_\beta(L_i(\boldsymbol{x}0) - L_i(\boldsymbol{x}), L_i(\boldsymbol{x}1) - L_i(\boldsymbol{x}))$, $i = 1, 2, \ldots$. Let us show that $\mathfrak{B}_\beta(\mathcal{K}(\boldsymbol{x}1) - \mathcal{K}(\boldsymbol{x}), \mathcal{K}(\boldsymbol{x}0) - \mathcal{K}(\boldsymbol{x})) \in \mathfrak{B}_\beta(S)$.

Lemma 7. *Let M be a convex closed set in a topological vector space and let $\mathbf{s}_1, \mathbf{s}_2, \ldots$ belong to M. If the sum $\sum_{i=1}^{+\infty} p_i \mathbf{s}_i$, where all p_i are positive and satisfy the equality $\sum_{i=1}^{+\infty} p_i = 1$, converges to $\mathbf{s}$, then $\mathbf{s} \in M$.*

Proof of the lemma. We have

$$\mathbf{s} = \lim_{n\to+\infty} \sum_{i=1}^{n} p_i \mathbf{s}_i \tag{77}$$

$$= \lim_{n\to+\infty} \left(\sum_{j=1}^{n} p_j \right) \left(\sum_{i=1}^{n} \frac{p_i}{\sum_{j=1}^{n} p_j} \mathbf{s}_i \right) \tag{78}$$

$$= \lim_{n\to+\infty} \sum_{i=1}^{n} \frac{p_i}{\sum_{j=1}^{n} p_j} \mathbf{s}_i \tag{79}$$

The third equality follows from second because $\lim_{n\to+\infty} \left(\sum_{j=1}^{n} p_j \right) = 1$. Since

$$\sum_{i=1}^{n} \frac{p_i}{\sum_{j=1}^{n} p_j} = 1 \quad , \tag{80}$$

the sum

$$\sum_{i=1}^{n} \frac{p_i}{\sum_{j=1}^{n} p_j} \mathbf{s}_i \tag{81}$$

is a finite convex mixture of elements from M and thus belongs to M (see, e.g., [Val64], Sect. I.C). The limit of this sums belongs to M because M is closed. $\square$

Thus $(\mathcal{K}(\boldsymbol{x}1) - \mathcal{K}(\boldsymbol{x}), \mathcal{K}(\boldsymbol{x}0) - \mathcal{K}(\boldsymbol{x}))$ is a superprediction and $\mathcal{K}$ is a superloss process.

Clearly, for each i we have $\mathcal{K} \le L_i + i \frac{\ln 2}{\ln(1/\beta)}$.

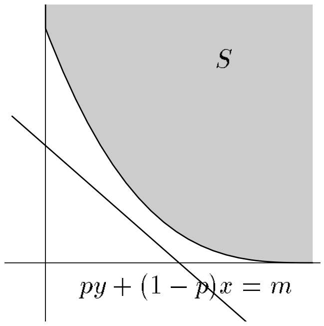

Figure 9 The drawing of S above the straight line $py + (1-p)x = m$ for Lemma 8

Appendix G: A Lower Bound on Expectations

This section contains a lemma (it appeared as Lemma 3 in [Kal02]) providing a lower bound for the expectations of predictive complexity.

Lemma 8. *Let $\mathfrak{G}$ be a game with the set of superpredictions S and $p \in (0,1)$. If for every $(u,v) \in S$ the inequality $(1-p)u + pv \geq m$ holds, then for each superloss process L w.r.t. $\mathfrak{G}$ we get*

$$\mathbf{E}L(\xi_1^{(p)} \ldots \xi_n^{(p)}) \geq mn \ , \tag{82}$$

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

Geometrically the conditions of the lemma mean that S lies to the 'north-east' of the line $px + (1-p)y = m$, see Fig. 9.

Proof. Consider a string $\boldsymbol{x}$. The point $(L(\boldsymbol{x}0) - L(\boldsymbol{x}), L(\boldsymbol{x}1) - L(\boldsymbol{x})) = (s_0, s_1)$ is a superprediction. We have

$$\mathbf{E}(L(\boldsymbol{x}\xi^{(p)}) - L(\boldsymbol{x})) = (1-p)s_0 + ps_1 \geq m \ , \tag{83}$$

where $\xi^{(p)}$ is a result of one Bernoulli trial with the probability of 1 being equal to p. On the other hand

$$\begin{aligned}
\mathbf{E}L(\xi_1^{(p)} \ldots \xi_n^{(p)}\xi_{n+1}^{(p)}) &= \sum_{\omega_1,\ldots,\omega_n,\omega_{n+1} \in \mathbb{B}} \Pr\{(\omega_1 \ldots \omega_n\omega_{n+1})\} L(\omega_1 \ldots \omega_n\omega_{n+1}) \\
&= \sum_{\omega_1,\ldots,\omega_n \in \mathbb{B}} \Pr\{(\omega_1 \ldots \omega_n)\} \mathbf{E}L(\omega_1 \ldots \omega_n \xi^{(p)}) \\
&\geq m + \sum_{\omega_1,\ldots,\omega_n \in \mathbb{B}} \Pr\{(\omega_1 \ldots \omega_n)\} L(\omega_1 \ldots \omega_n) \\
&= m + \mathbf{E}L(\xi_1^{(p)} \ldots \xi_n^{(p)}) \ ,
\end{aligned} \tag{84}$$

where $\Pr\{(\omega_1 \ldots \omega_k)\}$ is for the probability of the event $\{(\xi_1^{(p)} \ldots \xi_k^{(p)}) = (\omega_1 \ldots \omega_k)\}$. The lemma follows. $\square$