The Aggregating Algorithm and Predictive Complexity

Yuri Kalnishkan

Thesis submitted to the University of London
for the degree of Doctor of Philosophy

This is the final version incorporating examiners' comments

Department of Computer Science,
Royal Holloway, University of London,
Egham, Surrey TW20 0EX,
United Kingdom
e-mail: `yura@cs.rhul.ac.uk`

November 11, 2002

Abstract

This thesis is devoted to on-line learning. An on-line learning algorithm receives elements of a sequence one by one and tries to predict every element before it arrives. The performance of such an algorithm is measured by the discrepancies between its predictions and the outcomes. Discrepancies over several trials sum up to total cumulative loss.

The starting point is the Aggregating Algorithm (AA). This algorithm deals with the problem of prediction with expert advice. In this thesis the existing theory of the AA is surveyed and some further properties are established.

The concept of predictive complexity introduced by V. Vovk is a natural development of the theory of prediction with expert advice. Predictive complexity bounds the loss of every algorithm from below. Generally this bound does not correspond to the loss of an algorithm but it is optimal 'in the limit'. Thus it is an intrinsic measure of 'learnability' of a finite sequence. It is similar to Kolmogorov complexity, which is a measure of the descriptive complexity of a string independent of a particular description method. Different approaches to optimality give rise to different definitions of predictive complexity. The problem of determining the strongest complexity existing for a given loss function is addressed in this thesis. Some positive and negative results related to this problem are derived.

The expectations of predictive complexity w.r.t. the Bernoulli distribution provide a powerful tool for the theory of predictive complexity. The relations between these expectations and the loss functions are derived and the expectations are applied to establishing tight linear inequalities between different predictive complexities.

The final result of this thesis is the Unpredictability Property. It is a generalisation of the Incompressibility Property, which holds for Kolmogorov complexity. The Unpredictability Property states that the predictive complexity of most strings is close to a natural upper bound.

Contents

Lists of Statements

Theorems

Corollaries

Lemmas

Propositions

List of Figures

Acknowledgements

First and foremost I would like to thank my supervisors Volodya Vovk and Alex Gammerman. Volodya's ideas and insights laid the foundation for the field this thesis belongs to and have always been instrumental in my work. I have greatly benefited from Alex's enormous expertise in practical issues of Machine Learning.

The next person I would like to thank is my colleague Michael Vyugin. I have been glad to find a co-author and a friend in him. If it were not for our joint work, our discussions, and the encouragement he never failed to give me, this thesis would be lean.

I would like to thank many people for useful and helpful discussions. Perhaps the most valuable comments and suggestions I have received came from Vladimir V'yugin and Alexander Shen (Institute for Information Transmission Problems, Moscow). Meanwhile it was the course of lectures on Kolmogorov complexity delivered by Alexander Shen at the Moscow State University in the year 1997/1998 that introduced me to the subject.

The Department of Computer Science at the Royal Holloway College has been a very friendly research environment and I am grateful to its members. Of the people who are (or at time were) affiliated with the department, I would like to specifically thank Victor Dalmau (now University of California, Santa Cruz), Anna Fukshansky, Leo Gordon, Mark Herbster (now University College, London), Craig Saunders, Steve Schneider, John Shawe-Taylor, Mark Stitson (now Yospace Ltd.), Helen Treharne, and Chris Watkins. Special thanks to Tom Melluish who has done a superb job reading my thesis drafts and correcting my English. Whenever you find a correct article in this text, the credit for it goes to Tom; all incorrect ones are entirely my own fault.

I believe that much of the credit for the excellent atmosphere at the department is shared by the departmental secretaries Gill Adourian and Janet

Hales. Gill and Janet have been more than kind and helpful.

Of the people from other departments of the college I have had very useful discussions with Andrei Soklakov (Mathematics) and Mina Golshan (Physics).

My special thanks go to my friend Andrei Raigorodskii from the Moscow State University. The discussions we had during my trips to Moscow and Andrei's visits to London always added a new perspective to my perception of the subject.

I am grateful to people who showed interest in my research, asked questions, and had discussions with me during the conferences I attended. I would particularly like to thank Olivier Bousquet (Ecole Polytechnique, Centre de Mathématiques Appliquées).

Finally I would like to acknowledge the financial support. I have been funded by the EPSRC grants GR/M14937 'Predictive Complexity: recursion-theoretic variants' (held by Volodya Vovk) and GR/R46670 'Complexity Approximation Principle and Predictive Complexity: Analysis and Applications' (held by Alex Gammerman and Volodya Vovk) and by an Overseas Research Students Awards Scheme grant. I am also grateful to the organisers of COLT99 who partly refunded my travel expenses and to the organisers of ALT99 who supported me through the Mark Gold Award.

Chapter 1

Introduction

In this thesis the general theory of on-line learning algorithms is investigated. These algorithms try to predict elements of a given sequence. An algorithm $\mathfrak{A}$ obtains elements one by one and attempts to predict each element before seeing it. If there is a discrepancy between the prediction and the outcome, we say that the algorithm suffers loss; the loss over several trials adds up to the total loss $\mathrm{Loss}_{\mathfrak{A}}$. The exact description of a learning environment is a triple consisting of a set of possible outcomes, a set of allowed predictions, and a function measuring the loss; a triple of this kind is called a game. The relevant rigorous definitions are formulated and discussed in Chapter 2.

One of the first natural problems emerging within this framework is the problem of prediction with expert advice. Suppose that our learning algorithm is given access to predictions of some 'experts' trying to predict elements of the same sequence. Is it possible to merge experts' predictions in such a way as to achieve the total loss not greatly exceeding that of the best expert? This problem is widely covered in the literature (see, e.g., [LW94, HKW98, CBFH$^+$97]); its settings are discussed in more detail in Chapter 2.

The Aggregating Algorithm (AA) introduced by V. Vovk and discussed in Chapter 3 is a powerful tool for solving the problem of prediction with expert advice. In fact, it is optimal in the following sense: if using any merging technique an algorithm $\mathfrak{A}$ achieves the loss

$$\mathrm{Loss}_{\mathfrak{A}} \le c\,\mathrm{Loss}_{\mathcal{E}_{\mathrm{best}}} + a \ln n \tag{1.1}$$

for all possible sets of experts and all incoming sequences, where $\mathrm{Loss}_{\mathcal{E}_{\mathrm{best}}}$ is the loss of the best expert, n is the number of experts, and c and a are some constants, then the same is achieved by the AA.

All material of Chapters 2 and 3 is not original except for Theorem 1; only the arrangement and the presentation may be treated as new.

The constant c in the bound (1.1) is of great importance. Since the term $a \ln n$ remains constant as new elements of a sequence arrive, the constant c determines the asymptotic behaviour of the cumulative loss $\mathrm{Loss}_{\mathfrak{A}}$. If $c = 1$, the learning algorithm predicts nearly as good as the best expert and games with this property are called mixable. Chapter 4 is concerned with the behaviour of c.

The geometric interpretation of [Vov90, Vov98b] underlies the developments of Chapter 4. This interpretation allows us to derive differential criteria of mixability, which provide us with simpler proofs of mixability for some games (Theorem 8). The criteria can also be applied to the study of games which are not mixable. Although c does not necessarily attain the value 1, we show that for a large class of games $c(\beta) \to 1$ as $\beta \to 1$, where $\beta \in (0,1)$ is a parameter supplied to the AA. In many cases it is possible to determine the rate of convergence. Intuitively it means that by taking the values of β close to 1, we can approximate the 'ideal' situation with $c = 1$ to any degree of precision. However there are situations where $c(\beta)$ does not converge to 1; an example is provided by some unbounded games where $c(\beta)$ is infinite for all values of β.

The concept of predictive complexity and the study of this concept is based on the theory of prediction with expert advice. The introduction of this concept is motivated by the following considerations. It is natural to ask whether it is possible to construct the best computable prediction strategy. The answer to this question is negative unless the game is trivial since every strategy is outperformed by some other strategy on some inputs. However if we extend the class of computable strategies to certain 'superstrategies' (to be more precise, superloss processes) we can often find an optimal element in the class. This optimal element provides a lower bound on the loss of every computable strategy that is tight in some sense; we call it predictive complexity.

There is a similarity between predictive complexity and Kolmogorov complexity. In fact, a variant of Kolmogorov complexity, namely, the negative logarithm of Levin's *a priori* semimeasure, is predictive complexity for the so called logarithmic-loss game. Kolmogorov complexity is an inherent measure of how difficult it is to describe a string: the smaller the complexity the easier it is to describe the string. Similarly, predictive complexity refers to 'predictability' of a string: the smaller the complexity, the more predictable the elements of a string are.

Predictive complexity was introduced in [VW98]. We refer to predictive complexity defined in this paper as 'simple predictive complexity'. If the definition (namely, the approach to optimality) is relaxed, the definitions of weaker complexities emerge. It is shown in [VW98] that predictive complexity exists for mixable games. It remains an open problem to show that mixability is a necessary condition. A more general problem may be formulated in the following way. Given a game, construct the strongest version of predictive complexity available for this game.

This problem is addressed in Chapter 5. Although the complete solution has not been found, some steps towards it have been made. The results of Chapter 4 about the behaviour of $c(\beta)$ allow us to show that many games specify weak complexities. Some negative results are also presented. The construction from [VW98] is included for completeness.

The rest of the thesis consists entirely of original material. Chapter 6 discusses the expectations $\mathcal{K}(\xi_1^{(p)}, \xi_2^{(p)}, \ldots, \xi_n^{(p)})$, where $\mathcal{K}$ is predictive complexity and $\xi_1^{(p)}, \xi_2^{(p)}, \ldots, \xi_n^{(p)}$ are results of n independent Bernoulli trials with the probability of 1 being equal to p. These expectations turn out to have a simple geometrical interpretation and they provide an important tool for investigating predictive complexity. An example of a result that can be shown by using this tool is the uniqueness theorem. It states that if two games specify the same complexity, then their geometric images (sets of superpredictions) coincide. In fact, games with the same set of superpredictions are identical in respect to predictive complexity.

In Chapter 7 the method of expectation is applied to the study of inequalities of the form $\mathcal{K}_1 \geq \mathcal{K}_2$, where $\mathcal{K}_1$ and $\mathcal{K}_2$ are predictive complexities for different games. It turns out that these inequalities have probabilistic and geometric interpretations. A simple criterion for the inequalities to hold is formulated and later applied to the study of relations between two specific complexities, namely, logarithmic-loss complexity $\mathcal{K}^{\log}$ and square-loss complexity $\mathcal{K}^{\mathrm{sq}}$. The intuitive interpretation is that when we compare the complexities of a string $\boldsymbol{x}$ given by different games, we compare the learnability of $\boldsymbol{x}$ in different learning environments.

In Chapter 8 we formulate the Unpredictability Property. There is a natural upper bound on predictive complexity provided by a simple predicting strategy. The Unpredictability Property states that most of the strings have complexity close to this upper bound. This property is the counterpart to the Incompressibility Property for Kolmogorov complexity, which states that most of the strings have complexity close to the maximal possible for their lengths. Strings of this kind can be called random.

The following convention applies throughout this thesis. If a result is new, it is named 'Theorem' or 'Lemma'; this also applies to results which are not essentially new but are given new independent proofs (e.g., Theorem 8). If a result is not original and has been included for completeness sake, it is named 'Proposition'. Propositions bear references to the sources they were taken from.

Some parts of the original contribution of this thesis have been produced by joint work. Michael Vyugin co-authored original results from Chapters 4, 5, 6, and 8. Volodya Vovk co-authored original results from Chapters 6 and 8.

The results from this thesis appear in the following papers published by the author:

- a journal paper:
  - Y. Kalnishkan. General linear relations among different types of predictive complexity. *Theoretical Computer Science*, 271: 181–200, 2002.
- papers in conference proceedings:
  - Y. Kalnishkan. Linear relations between square-loss and Kolmogorov complexity. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 226–232. Association for Computing Machinery, 1999.
  - Y. 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.
  - Y. Kalnishkan, M. Vyugin, and V. Vovk. Losses, complexities, and the Legendre transformation. In *Proceedings of The Twelfth International Conference on Algorithmic Learning Theory, 12th International Conference, ALT2001*, volume 2225 of *Lecture Notes in Artificial Intelligence*, Springer–Verlag, 2001.

  - Y. Kalnishkan and M. Vyugin. Mixability and the existence of weak complexities. In *Learning Theory, 15th Annual Conference on Computational Learning Theory, COLT 2002, Sydney, Australia, July 8-10, 2002, Proceedings*, volume 2375 of *Lecture Notes in Artificial Intelligence*, pages 105–120. Springer, 2002.

- a technical report:

  - Y. Kalnishkan and V. Vovk. The existence of predictive complexity and the Legendre transformation. Technical report, Computer Learning Research Centre Royal Holloway College, 2000.

Some results have not been published yet; this refers to Chapters 4, 5 and 8. On the other hand,

- a paper in conference proceedings:

  - Y. Kalnishkan. Complexity approximation principle and Rissanen's approach to real-valued parameters. In *Proceedings of the 11th European Conference on Machine Learning*, volume 1810 of *Lecture Notes in Artificial Intelligence*, pages 203–210. Springer-Verlag, 2000.

published by the author while doing research in predictive complexity has not been included since it deals with essentially different aspects of the subject.

Chapter 2

The Problem of On-line Prediction

This thesis deals with different variations of the on-line learning framework. In this chapter we discuss the most general on-line learning scheme and give principal definitions. The terminology of 'games' and 'losses' will be used throughout the thesis. One of the main problems about this framework, the problem of prediction with expert advice, is also formulated here.

2.1 The Basic Protocol

Let some events or *outcomes* happen in discrete time. We have a sequence of events

$$\omega_1, \omega_2, \ldots, \omega_n, \ldots \quad ,$$

which occur in succession. For example, ω_i may be a description of the weather on day i after the beginning of an experiment. Our goal is to predict each of the events. Before the event ω_n happens, we (i.e., our prediction method $\mathfrak{A}$) make a *prediction* γ_n. In other terms, $\mathfrak{A}$ works according to the following simple protocol:

Protocol 1.

```
FOR $t = 1, 2, \ldots$
  (1) $\mathfrak{A}$ chooses a prediction $\gamma_t$
  (2) $\mathfrak{A}$ observes the outcome $\omega_t$
END FOR.
```

A prediction method does not have to be oblivious, i.e., it may keep track of previous outcomes as well as some side information[1] or its own internal variables.

In order to develop a mathematical theory of this process, we need to specify several things more precisely. To begin with, let us fix the ranges of possible outcomes and predictions. Let Ω and Γ be the sets of, respectively, all possible outcomes and all possible predictions. We will call them the *outcome space* and the *prediction space*.

Now we need a way of assessing the quality of a prediction or a way to measure the correlation between a prediction and an outcome. Unless we have a measure of this kind, we cannot compare the performance of various prediction methods. Let λ be a *loss function*, i.e., a scalar function of two arguments, one of which is an outcome and another is a prediction. The intuitive meaning of $\lambda(\omega_i, \gamma_i)$ is the discrepancy between ω_i and γ_i or deviation of γ_i from ω_i. Nevertheless in some situations we will allow λ to assume negative values. The infinite value $+\infty$ will also be useful sometimes. Thus we will consider functions λ with different ranges under different circumstances.

The loss over several trials sums up to the total loss. This total or cumulative loss of a method $\mathfrak{A}$ after T trials is the sum

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

where γ_t, $t = 1, 2, \ldots, T$ are predictions output by $\mathfrak{A}$ on trials $1, 2, \ldots, T$. If $\mathfrak{A}$ is deterministic (we consider only deterministic strategies), $\mathrm{Loss}_{\mathfrak{A}}^{\lambda}$ specifies a scalar function on Ω^* (the set of all finite sequences of elements from Ω).

There are a number of points to be made to justify our convention to distinguish between the outcome and prediction spaces. First it is common in learning theory to consider different ‘target’ and ‘hypothesis’ classes (cf. PAC-learning). Secondly by allowing Γ to be larger then Ω, we admit greater flexibility in predictions. If, say, $\omega_i \in \Omega = \{0, 1\}$ represents the presence or absence of rain on day i, it would be natural to allow γ to vary in the interval from 0 to 1 and to represent the probability (in some loose sense of this word) of rain.

[1] ‘Side information’ is not covered in detail in this thesis; it appears only in Chapter 8, where conditional complexity is used. However the problems associated with it are of much interest.

The third argument is the most important and it provides us with a new insight to the process of prediction. We may treat γ as an *action* or a way of behaviour we choose to follow. Now $\lambda(\omega, \gamma)$ becomes the result of an action γ after it faces an 'external' phenomenon ω. In our meteorological example, ω may be the state of the weather on a particular day, while γ may represent the way one dresses in the morning. Inappropriate clothes will result in the person catching a cold; at least you will get wet if you unwisely leave your umbrella at home on a rainy day. A more elaborate example is provided by the stock market. Here ω may represent share prices while γ may be the investment we make in the hope that some particular trends will prevail in the market. Here λ represents the change in our assets.

Despite this 'action' interpretation, borrowed from [VW98], we will preserve the 'prediction' terminology since it is more common in the literature.

2.2 Games

A triple $\langle \Omega, \Gamma, \lambda \rangle$ of elements described in the previous section is called a *game*. In this section we will introduce and discuss several important games.

Arguably, the most natural way to measure the difference between two scalar values a and b is to employ the squared deviation $(a-b)^2$. It is used all the time in mathematical statistics although it is not easy to trace the origin of this convention. The popularity of this measure of difference should, in my opinion, be attributed to the smoothness of the function $y = x^2$, the geometrical interpretation as the squared length, as well as nice properties of the class of square integrable functions.

We will consider the loss function $\lambda(\omega, \gamma) = (\omega - \gamma)^2$ on different domains. The *discrete square-loss game* has binary outcomes ($\Omega = \mathbb{B} = \{0, 1\}$) while predictions are allowed to range through the unit interval ($\Gamma = [0, 1]$). We will mostly be interested in this discrete square-loss game and loosely call it the square-loss game. In the *continuous square-loss game* the outcome is allowed to range over the unit interval as well. This game also appears under the name 'Brier game' in the literature (e.g., [Vov01]).

Similarly, there are two variants of the *A, B-bounded square-loss game*, namely, discrete and continuous. In the former, $\Omega = \{A, B\}$ and $\Gamma = [A, B]$ and in the latter $\Omega = \Gamma = [A, B]$. Naturally we suppose that $A < B$. Unbounded versions of the square-loss game may be considered and are of interest but very little is known about them.

The next most natural measure of discrepancy between a and b is the absolute deviation $|a-b|$. As above, we introduce the *discrete absolute-loss game*

$$\langle\{0,1\},[0,1],|\omega-\gamma|\rangle\ ,$$

the *continuous absolute-loss game*

$$\langle[0,1],[0,1],|\omega-\gamma|\rangle\ ,$$

the *discrete* A,B*-bounded absolute-loss game*

$$\langle\{A,B\},[A,B],|\omega-\gamma|\rangle\ ,$$

and, finally, *continuous* A,B*-bounded absolute-loss game*

$$\langle[A,B],[A,B],|\omega-\gamma|\rangle\ .$$

We now proceed to the logarithmic loss function and the *logarithmic-loss game*, which look less natural but are of fundamental importance. In the logarithmic-loss game, we have $\Omega=\{0,1\}$, $\Gamma=[0,1]$ and

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

(the notation log is a shorthand for the logarithm to the base 2, i.e., $\log_2$). The importance of this game can be justified by the results we will obtain about it. The following comment provides an insight into the nature of this game.

Suppose we have the capital of £1. We split it into two parts, £γ_1 and £$(1-\gamma_1)$, and bet the first part on the outcome 1 and the second on the outcome 0. After the event happens, we are allowed to keep only the part of the capital we bet on the correct outcome. Thereafter we split the remaining capital into proportions γ_2 and $(1-\gamma_2)$ and bet them on the possible outcomes of the second trial and so on. (Note that in this dismal game *you can only lose*; you may treat an act of betting as investing your money into two banks which pay no interest and, moreover one of them eventually goes bankrupt.) The capital we possess after trial T is the product of proportions corresponding to correct guesses. Since our framework requires an additive rather then multiplicative measure of the performance, we take the logarithm of the capital; the pessimistic loss terminology leads us to the negative logarithm.

Another insight into the nature of this definition is given by the Shannon–Fano optimal code.

Note that the case of $\omega = 0$ and $\gamma = 1$ (or $\omega = 1$ and $\gamma = 0$) is possible. In this situation, the loss is $+\infty$ as suggested by $\lim_{x \to 0+}(-\log x)$. We do not exclude this case and we rather prefer to allow λ and Loss^{λ} to assume the value $+\infty$. No ambiguity occurs since we do not allow the value $-\infty$ and thus we will not have to calculate the sum $(+\infty) + (-\infty)$.

A simple generalisation of the logarithmic-loss game is the *β-logarithmic-loss game*, where $\beta \in (0, 1)$, with $\Gamma = [0, 1]$ and

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

Obviously, the usual logarithmic-loss game is 1/2-logarithmic-loss.

The logarithmic-loss game can be generalised to *Cover's game* (introduced in [CO96]; see also [VW98] for discussion). In this game, $\Omega = [0, +\infty)^N \setminus (0, 0, \ldots, 0)$. The outcome $\omega_t = (\omega_t^{(1)}, \omega_t^{(2)}, \ldots, \omega_t^{(N)})$ reflects the change in stock prices between day $t-1$ and day t. Namely, $\omega_t^{(n)}$ is the ratio of the day t closing price of the stock n to the day $(t-1)$ closing price. The elements of $\Gamma \subseteq [0, 1]^N$ are vectors $\gamma = (\gamma^{(1)}, \gamma^{(2)}, \ldots, \gamma^{(N)})$ such that $\sum_{i=1}^N \gamma^{(i)} = 1$. The entries of γ_t are the proportions of the total capital we invest in corresponding stocks at the beginning of day t.

The ratio of our capital at the end of day t to our capital at the end of day $(t-1)$ is the scalar product

$$\gamma_t \cdot \omega_t = \sum_{i=1}^{N} \gamma_t^{(i)} \omega_t^{(i)} \ .$$

Arguing as above, we define the loss function $\lambda(\omega, \gamma) = -\log(\gamma \cdot \omega)$.

The following simple game is also of interest. In this game only exact predictions are of any good. Let $\Omega = \Gamma = \{0, 1\}$ and

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

We call it the *simple prediction game*. It reflects a standard machine learning problem concerning the error count (note that the loss w.r.t. this game coincides with the number of errors). The problem is described e.g. in [LV93], Section 5.4.3 (cf. Exercise 5.3.3 in the edition [LV97]).

In a context similar to ours this game was considered in [LW94].

2.3 Prediction with Expert Advice

One of the central features of the approach of this thesis is that we do not impose any restrictions on possible sequences of outcomes. No assumptions such as i.i.d. or stochasticity are made. We do not restrict ourselves to strings of a particular kind; within our approach *everything may happen.* One may think that no mathematical theory of prediction is under such assumptions. Indeed, every nontrivial guess about the outcome of a trial may be falsified by some outcomes. Every prediction method would fail (i.e., suffer large loss) on some sequence of events; if no sequence is more 'likely' than others, we have no reason to prefer one method to another. Statements of this kind appear in many parts of learning theory and go under names like the 'no-free-lunch theorem'.

In actual practice, when we encounter complete uncertainty, it is natural to assign equal *probability* to every outcome. By doing this, we will end up with some method which is optimal on average. For the absolute-loss, square-loss, and logarithmic-loss games introduced above, this optimal method is to predict $1/2$ every time, but this trivial result is of little theoretical or practical value.

Nevertheless it is possible to formulate and investigate sound problems within our approach. One of the problems is that of prediction with expert advice. We now proceed to formulating its settings.

Suppose that we have a pool of experts. Experts work according to Protocol 1 and all try to predict elements of the same sequence. We admit pools of an arbitrary size. Suppose that the pool is parametrised by θ, which ranges over a set Θ. We will denote experts by $\mathcal{E}_\theta$, where $\theta \in \Theta$, and sometimes identify Θ with the pool. If, say, $\Theta = \{1, 2, \ldots, N\}$, then we have a finite pool $\mathcal{E}_1, \ldots, \mathcal{E}_N$.

Let us make the experts' predictions available to a prediction algorithm $\mathfrak{A}$, the *learner.* The algorithm $\mathfrak{A}$ observes the experts' predictions every time before it makes its own. The following protocol is used:

Protocol 2.

```
FOR $t = 1, 2, \ldots$
    (1) $\mathcal{E}_\theta$ outputs a prediction $\gamma_t^{(\theta)}$ for all $\theta \in \Theta$
    (2) $\mathfrak{A}$ observes $\gamma_t^{(\theta)}$ for all $\theta \in \Theta$
    (3) $\mathfrak{A}$ chooses a prediction $\gamma_t \in \Gamma$
    (4) $\mathfrak{A}$ observes the outcome $\omega_t \in \Omega$
```

```
END FOR.
```

Note that while $\mathfrak{A}$ is an algorithm, we make no assumption regarding the computability of $\mathcal{E}_\theta$. They may be uncomputable or even receive some extra information which is hidden from $\mathfrak{A}$. The 'interior' of experts is unknown to us and we do not know how they come to their conclusions. The learner has no access to the experts' internal variables (if they have any) or other particulars of their work; it only receives their predictions.

The problem of prediction with expert advice is to devise $\mathfrak{A}$ to minimise

$$\mathrm{Loss}^{\lambda}_{\mathfrak{A}}(\omega_1, \omega_2, \ldots, \omega_T) - \inf_{\theta \in \Theta} \mathrm{Loss}^{\lambda}_{\mathcal{E}_\theta}(\omega_1, \omega_2, \ldots, \omega_T) \ , \tag{2.1}$$

where λ is some (fixed) loss function. Since we make no assumptions about the nature of sequence $\omega_1, \omega_2, \ldots, \omega_T$, the difference should be small for all sequences. Different ways of looking at the value of T are possible. The results we will use hold for every T.

We assume the worst-case scenario or, in other words, we consider an antagonistic game[2]. In this game, $\mathfrak{A}$ struggles to decrease the difference (2.1) and its adversary, nature, which produces sequences of ωs and the predictions of experts, tries to maximise it.

There are several remarks to be made about this game. First note that we can treat the experts and the 'side' generating the outcomes ω_t as one player (to develop the meteorological example further, weather bureaus unite with weather makers to trick people into going outdoor without an umbrella on a rainy day). They do not have to be computable or satisfy other assumptions and thus they do not have to be independent. Secondly while the learner has a simple goal of minimising its loss, the goal of experts is more complicated. They minimise the loss of the best expert while trying to deceive the learner (i.e., weather bureau employees have to dress according to their forecasts but they still try to deceive ordinary folk).

2.4 Regularity Assumptions

In this section we summarise the assumptions we need to make about the games. These assumptions will be used in the treatment of the theory of prediction with expert advice; further tasks we will consider later may require different assumptions.

[2] Which is not one of the games defined in Subsect. 2.2 above!

REG_1 The range of λ is $[0, +\infty]$.

REG_2 The prediction space Γ is a compact topological space.

REG_3 For every $\omega \in \Omega$, the function $\lambda(\omega, \gamma)$ is continuous in the second argument.

REG_4 There exists $\gamma \in \Gamma$ such that, for every $\omega \in \Omega$, the inequality $\lambda(\omega, \gamma) < +\infty$ holds.

REG_5 There is no γ such that, for all $\omega \in \Omega$, the equality $\lambda(\omega, \gamma) = 0$ holds.

REG_6 For every $\gamma_0 \in \Gamma, \omega_0 \in \Omega$ such that $\lambda(\omega_0, \gamma_0) = +\infty$ there is a sequence of $\gamma_n \in \Gamma$, $n = 1, 2, \ldots$, such that $\gamma_n \to \gamma_0$ as $n \to \infty$ and $\lambda(\omega_0, \gamma_n) < +\infty$.

Assumptions REG_1–REG_5 were introduced in [Vov98b, VW98]. Assumption REG_6 essentially means that λ assumes the infinite value only in exceptional situations; these situations may be approximated by final cases.

Clearly, it was the assumption REG_4 that made us prohibit the outcome $(0, 0, \ldots, 0)$ in Cover's game.

In order to speak about the continuity of λ, we need a topology on $[0, +\infty]$. We use the extended topology of $[-\infty, +\infty]$, described in Appendix A.

In fact, continuity follows from computability. If we want the learner to be computable, we need to assume that there is a definition of computability over $\Omega \times \Gamma$ and λ is computable according to some definition. Natural definitions of computability imply continuity; still we preserve REG_3 to highlight the fact that the theory of prediction with expert advice may be considered and still works without mentioning the algorithmic aspects.

The Aggregating Algorithm we are going to discuss involves dealing with measures and integrals in the general case. Therefore we need some assumptions concerning integrability. Whenever we speak about measures over a topological space, we will suppose that they are consistent with the topology, i.e., all Borel sets are measurable. A *probability distribution* is a measure μ such that $\mu(\Gamma) = 1$.

We also assume that there is some fixed σ-algebra $\mathfrak{S}_\Theta$ on Θ and, every time the experts output predictions $\gamma_t^{(\theta)}$, the function $\gamma_t(\theta) = \gamma_t^{(\theta)}$, which maps Θ into Γ, is Borel measurable. Clearly, it is always true for a finite Θ with the discrete topology.

Chapter 3

The Aggregating Algorithm

There are several approaches to the problem of prediction with expert advice and several methods to resolve it (cf. [CBFH$^+$97] and [LW94]). The Aggregating Algorithm is a method introduced in [Vov90]. It is proved to be optimal (see [Vov98b]) in some exact sense and thus it provides a complete solution for a very important and general case.

3.1 The Operation of the Aggregating Algorithm

The Aggregating Algorithm (AA) was proposed in the pioneering paper [Vov90]. We will describe a more general version from [VW98]; a simpler special case will be formulated in parallel. This simple case is very important for the thesis and I put it here for future reference. The rest of this section is organised in the following manner. The text in the left column will refer to the general situation with a game[1] $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ and an arbitrary pool of experts Θ. The right column describes the special case with $\Omega = \{0, 1\}$, $\Gamma = [0, 1]$ and finite $\Theta = \{1, 2, \ldots, N\}$.

The AA accepts a parameter $\beta \in (0, 1)$. Let us fix some β from this range. We start with the definition of the number $c(\beta)$.

[1] The set Γ is supposed to be a topological space; further assumptions will be made later.

A *generalised prediction* is the function $g : \Omega \to [0, +\infty]$ defined by the equality

$$g(\omega) = \log_\beta \int_\Gamma \beta^{\lambda(\omega,\gamma)} P(d\gamma)$$

for every ω, where P is a probability distribution on Γ (consistent with the topology of Γ as described in Sect. 2.4)

A *(simple) generalised prediction* is a pair $g = (g^{(0)}, g^{(1)})$ such that

$$\left\{ \begin{array}{lcl} g^{(0)} & = & \log_\beta \sum_{i=1}^k p_i \beta^{\lambda(0,\gamma_k)} \\ g^{(1)} & = & \log_\beta \sum_{i=1}^k p_i \beta^{\lambda(1,\gamma_k)} \end{array} \right. , \tag{3.1}$$

where k is some positive integer, $p_1, p_2, \ldots, p_k \in [0, 1]$, $\gamma_1, \gamma_2, \ldots, \gamma_k \in \Gamma$, and $p_1 + p_2 + \cdots + p_k = 1$.

Pick a generalised prediction g and consider the number

$$c_\beta(g) = \inf_{\gamma\in\Gamma} \sup_{\omega\in\Omega} \frac{\lambda(\omega,\gamma)}{g(\omega)} .$$

$$c_\beta(g) = \inf_{\gamma\in\Gamma} \max\left(\frac{\lambda(0,\gamma)}{g(\omega)}, \frac{\lambda(1,\gamma)}{g(\omega)} \right) . \tag{3.2}$$

On this particular instance we take the undefined ratio 0/0 to be equal 0. Let $c(\beta)$ be the supremum of $c_\beta(g)$ over all g. If the supremum does not exist, we put $c(\beta) = +\infty$. If it is not clear from the context which game we are referring to, we will write $c(\mathfrak{G}, \beta)$ to specify the game.

Let us show that the definition in the 'finite' case is really a special case of the general definition. We will prove a more general fact.

Theorem 1. *Let β be a number from $(0, 1)$ and $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ be a game satisfying REG$_1$–REG$_5$. If Ω is a compact topological space and $\lambda(\omega, \gamma) : \Omega\times\Gamma \to [0, +\infty]$ is a continuous function (w.r.t. the extended topology) of two arguments, then in the definition of $c(\beta)$ the supremum of $c_g(\beta)$ can be taken over the the generalised predictions induced by finite discrete distributions, i.e., generalised predictions g of the form*

$$g(\omega) = \log_\beta \sum_{i=1}^k p_i \beta^{\lambda(\omega,\gamma_i)} , \tag{3.3}$$

where k is a positive integer, $p_i \in [0, 1]$ $(i = 1, 2, \ldots, k)$ are such that $\sum_{i=1}^k p_i = 1$, and $\gamma_i \in \Gamma$ for every $i = 1, 2, \ldots, k$.

The proofs will be given in Section 3.2.

The idea behind $c(\beta)$ is that every generalised prediction $g(\omega)$ may be replaced by some prediction $\gamma \in \Gamma$ whose loss is only $c(\beta)$ times greater.

Proposition 1 ([VW98]). *Let $G = \langle \Omega, \Gamma, \lambda \rangle$ be a game satisfying REG_1–REG_5 and β be a number from $(0,1)$ such that $c(\beta) < +\infty$. Then for every generalised prediction g there is $\gamma \in \Gamma$ such that, for every $\omega \in \Omega$, the inequality*

$$\lambda(\omega, \gamma) \leq c(\beta) g(\omega) \tag{3.4}$$

holds. In the case of $\Omega = \{0,1\}$ and $g = (g^{(0)}, g^{(1)})$ this reduces to

$$\left\{ \begin{array}{lcl} \lambda(0,\gamma) & \leq & c(\beta) g^{(0)} \\ \lambda(1,\gamma) & \leq & c(\beta) g^{(1)} \end{array} \right. .$$

Let us now proceed to the algorithm itself. Throughout the process of prediction, AA maintains a list of relative weights for experts. After trial t, the weights are

a measure W_t on Θ.	an array of numbers $\left(w_t^{(1)}, w_t^{(2)}, \ldots, w_t^{(N)}\right)$.

The initial weights are to some extent arbitrary and may be regarded as one more parameter of AA. The weights

W_0 are initialised by a probability distribution P_0 on Θ.	$\left(w_0^{(1)}, w_0^{(2)}, \ldots, w_0^{(N)}\right)$ are initialised by an array of numbers $p_0^{(1)}, p_0^{(2)}, \ldots, p_0^{(N)} \in [0,1]$ such that $\sum_{i=1}^N p_0^{(i)} = 1$.

After trial t, AA modifies the weights:

$$W_t(E) = \int_E \beta^{\lambda\left(\omega_t, \gamma_t^{(\theta)}\right)} W_{t-1}(d\theta) \tag{3.5}$$ for every $E \subseteq \Theta$ (this is still a measure; see, say, [Rud74], Proposition 1.25).	$$w_t^{(i)} = w_{t-1}^{(i)} \beta^{\lambda\left(\omega_t, \gamma_t^{(i)}\right)}$$ for $i = 1, 2, \ldots, N$.

On trial t, after observing the experts' predictions, AA normalises the current weights:

let

$$P_{t-1}(A) = W_{t-1}(A)/W_{t-1}(\Theta)$$

for every $A \subseteq \Theta$

let

$$p_{t-1}^{(i)} = w_{t-1}^{(i)} / \sum_{j=1}^{N} w_{t-1}^{(j)} \ ,$$

where $i = 1, 2, \ldots, N$,

and averages the experts' predictions into a generalised prediction

$$g_t(\omega) = \log_\beta \int_\Theta \beta^{\lambda\left(\omega, \gamma_t^{(\theta)}\right)} P_{t-1}(d\theta). \tag{3.6}$$

$$\begin{cases} g_t^{(0)} = \log_\beta \sum_{j=1}^{N} \beta^{\lambda\left(0, \gamma_t^{(j)}\right)} p_{t-1}^{(j)} \\ g_t^{(1)} = \log_\beta \sum_{j=1}^{N} \beta^{\lambda\left(1, \gamma_t^{(j)}\right)} p_{t-1}^{(j)} \ . \end{cases}$$

Now AA finds γ_t such that

$$\lambda(\omega, \gamma_t) \leq c(\beta) g(\omega_t) \tag{3.7}$$

for all $\omega \in \Omega$

$$\begin{cases} \lambda(0, \gamma_t) & \leq & c(\beta) g_t^{(0)} \\ \lambda(1, \gamma_t) & \leq & c(\beta) g_t^{(1)} \end{cases}$$

and outputs it. Such γ_t exists by Proposition 1.

The key property of AA is the following.

Proposition 2 ([Vov90, VW98]). *Let β be a number from $(0,1)$ and $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ be a game satisfying REG$_1$–REG$_5$. For every pool of experts Θ, every positive integer T and every sequence $\omega_1, \omega_2, \ldots, \omega_T \in \Omega$, if a learner $\mathfrak{A}$ uses the AA, then the inequality*

$$\mathrm{Loss}_{\mathfrak{A}}^{\lambda}(\omega_1, \omega_2, \ldots, \omega_T) \leq c(\beta) \log_\beta \int_\Theta \beta^{\mathrm{Loss}_{\mathcal{E}_\theta}^{\lambda}(\omega_1, \omega_2, \ldots, \omega_T)} P_0(d\theta) \ ,$$

where P_0 is the initial distribution, holds.

If the pool is finite, the integral reduces to the sum. If we omit all terms of this sum except for one, we obtain the following corollary.

Corollary 1 ([Vov90]). *If, under the conditions of Proposition 2, the pool* Θ *is finite and* $\Theta = \{1, 2, \ldots, N\}$*, then, for every sequence* $\omega_1, \omega_2, \ldots, \omega_T \in \Omega$*, if a learner* $\mathfrak{A}$ *uses the AA, then the inequality*

$$\mathrm{Loss}^\lambda_{\mathfrak{A}}(\omega_1, \omega_2, \ldots, \omega_T) \leq c(\beta)\,\mathrm{Loss}^\lambda_{\mathcal{E}_i}(\omega_1, \omega_2, \ldots, \omega_T) + \frac{c(\beta)}{\ln(1/\beta)} \ln\left(1/p_0^{(i)}\right) \ ,$$

where $p_0^{(1)}, p_0^{(2)}, \ldots, p_0^{(N)}$ *are the initial weights, holds for all* $i = 1, 2, \ldots, N$*.*

The initial distribution is a parameter that can be altered. If we take the uniform distribution, we will get the next corollary.

Corollary 2 ([Vov90]). *If, under the conditions of Proposition 2, the pool* Θ *is finite and* $\Theta = \{1, 2, \ldots, N\}$*, then, for every sequence* $\omega_1, \omega_2, \ldots, \omega_T \in \Omega$*, if a learner* $\mathfrak{A}$ *uses AA with the uniform initial distribution, then the inequality*

$$\mathrm{Loss}^\lambda_{\mathfrak{A}}(\omega_1, \omega_2, \ldots, \omega_T) \leq c(\beta)\,\mathrm{Loss}^\lambda_{\mathcal{E}_i}(\omega_1, \omega_2, \ldots, \omega_T) + \frac{c(\beta)}{\ln(1/\beta)} \ln N$$

holds.

3.2 Proofs

In this section there are proofs of the statements from the previous section.

Proof of Theorem 1. Consider a generalised prediction g induced by a distribution P over Γ. We will start by showing that $\beta^{g(\omega)}$ may be approximated uniformly in ω to any degree of precision by an expression of the form $\sum_{i=1}^k p_i \beta^{\lambda(\omega,\gamma_i)}$, where k is a positive integer, $p_i \in [0,1]$ $(i = 1, 2, \ldots, k)$ are such that $\sum_{i=1}^k p_i = 1$, and $\gamma_i \in \Gamma$ for every $i = 1, 2, \ldots, k$.

Take some $\varepsilon > 0$. It follows from the compactness of Ω and Γ that $\beta^{\lambda(\omega,\gamma)}$ is uniformly continuous and these sets can be represented as

$$\Omega = \cup_{i=1}^{n_\varepsilon} \Omega_i \tag{3.8}$$

$$\Gamma = \cup_{j=1}^{m_\varepsilon} \Gamma_j \ , \tag{3.9}$$

where, for every $i = 1, 2, \ldots, n_\varepsilon$ and $j = 1, 2, \ldots, m_\varepsilon$, if $\omega_1, \omega_2 \in \Omega_i$ and $\gamma_1, \gamma_2 \in \Gamma_j$, then $\left|\beta^{\lambda(\omega_1,\gamma_1)} - \beta^{\lambda(\omega_2,\gamma_2)}\right| < \varepsilon$.

Without loss of generality we may assume that all Γ_i are disjoint and Borel. Let us pick some $\gamma_j \in \Gamma_j$ and let $p_j = P(\Gamma_i)$, $j = 1, 2, \ldots, m_\varepsilon$. Clearly, for every $\omega \in \Omega_i$, we have

$$\begin{aligned}\left|\int_\Gamma \beta^{\lambda(\omega,\gamma)} P(d\gamma) - \sum_{i=1}^k p_i \beta^{\lambda(\omega_i,\gamma_k)}\right| &\leq \sum_{i=1}^k \left|\int_{\Gamma_j} \beta^{\lambda(\omega,\gamma)} P(d\gamma) - \beta^{\lambda(\omega_i,\gamma_k)} p_i\right| \\ &\leq \sum_{i=1}^k \varepsilon p_i \\ &= \varepsilon \enspace .\end{aligned}$$

Now let $c(\beta)$ be the value obtained from the definition that uses all generalised predictions. By restricting the definition to the generalised predictions of the form (3.3), we can only increase the value. Consider c from the interval $(0, c(\beta))$. Since c is less than $\sup\{c_g(\beta) \mid g \text{ is a generalised prediction}\}$, there is generalised prediction g such that

$$\forall\gamma \in \Gamma \exists\omega \in \Omega : \ \lambda(\omega,\gamma) > cg(\omega) \enspace .$$

The inequality may be rewritten as

$$\beta^{\lambda(\omega,\gamma)/c} < \beta^{g(\omega)} \enspace .$$

The theorem will follow if we replace $\beta^{g(\omega)}$ in this formula with an expression of the form $\sum_{i=1}^k p_i \beta^{\lambda(\omega,\gamma_i)}$. It is sufficient to find some $\varepsilon > 0$ such that

$$\forall\gamma \in \Gamma \exists\omega \in \Omega : \ \beta^{\lambda(\omega,\gamma)/c} + \varepsilon < \beta^{g(\omega)} \enspace .$$

Consider $f(\omega,\gamma) = \beta^{g(\omega)} - \beta^{\lambda(\omega,\gamma)/c}$. It is easy to check that $\beta^{g(\omega)}$ is continuous and thus $f(\omega,\gamma)$ is continuous in two arguments. For every $\gamma \in \Gamma$ there is ω_γ such that $f(\omega_\gamma,\gamma) = \varepsilon_\gamma > 0$. Since the function $f : \Omega \times \Gamma \to \mathbb{R}$ is continuous, for every $\tilde\gamma$, there is an open $\Gamma_{\tilde\gamma} \subseteq \Gamma$ such that $f(\omega_{\tilde\gamma},\gamma) \geq \varepsilon_{\tilde\gamma}/2$ for every $\gamma \in \Gamma_{\tilde\gamma}$. Since Γ is compact, $\{\Gamma_{\tilde\gamma} \mid \tilde\gamma \in \Gamma\}$ contains a finite covering of Γ, i.e., there is a positive integer n and some $\tilde\gamma_1, \tilde\gamma_2, \ldots, \tilde\gamma_n \in \Gamma$ such that $\Gamma = \bigcup_{i=1}^n \Gamma_{\tilde\gamma_i}$. Let

$$\varepsilon = \min_{i=1,2,\ldots,n} \varepsilon_{\tilde\gamma_i}/2 \enspace .$$

□

For completeness sake the following derivation has been included.

Proof of Proposition 2. First note the following inequality for the loss on trial t:

$$\lambda(\omega_t, \gamma_t) \leq c(\beta) \log_\beta \int_\Theta \beta^{\lambda(\omega_t, \gamma_t^{(\theta)})} P_{t-1}(d\theta) \ .$$

The inequality follows immediately from (3.6) and (3.7). Using (3.5), we also obtain

$$W_t(d\theta) = \beta^{\mathrm{Loss}^\lambda_{\mathcal{E}_i}(\omega_1, \omega_2, \ldots, \omega_t)} P_0(d\theta) \ .$$

The rigorous foundation for dealing with $d\theta$ in this fashion is provided by, say, Theorem 1.29 and the subsequent remark in [Rud74].

Now let us prove the proposition by induction.

$$\begin{aligned}
\mathrm{Loss}^\lambda_{\mathfrak{A}}(\omega_1, \omega_2, \ldots, \omega_T) &= \mathrm{Loss}^\lambda_{\mathfrak{A}}(\omega_1, \omega_2, \ldots, \omega_{T-1}) + \lambda(\omega_T, \gamma_T) \\
&\leq c(\beta) \log_\beta \int_\Theta \beta^{\mathrm{Loss}^\lambda_{\mathcal{E}_\theta}(\omega_1, \omega_2, \ldots, \omega_{T-1})} P_0(d\theta) \\
&\quad + c(\beta) \log_\beta \int_\Theta \beta^{\lambda(\omega_T, \gamma_T^{(\theta)})} P_{T-1}(d\theta) \\
&= c(\beta) \log_\beta \int_\Theta W_{T-1}(d\theta) \\
&\quad \times \int_\Theta \frac{\beta^{\lambda(\omega_T, \gamma_T^{(\theta)})}}{W_{T-1}(\Theta)} W_{T-1}(d\theta) \\
&= c(\beta) \log_\beta \int_\Theta \beta^{\lambda(\omega_T, \gamma_T^{(\theta)})} \beta^{\mathrm{Loss}^\lambda_{\mathcal{E}_\theta}(\omega_1, \omega_2, \ldots, \omega_{T-1})} P_0(d\theta) \\
&= c(\beta) \log_\beta \int_\Theta \beta^{\mathrm{Loss}^\lambda_{\mathcal{E}_\theta}(\omega_1, \omega_2, \ldots, \omega_T)} P_0(d\theta) \ .
\end{aligned}$$

□

3.3 Optimality of the Aggregating Algorithm

Proposition 2 together with Corollaries 1 and 2 describe the ability of AA to merge experts' predictions. Had they been the only results about the performance of AA, its significance would have entirely depended on comparison with other techniques of merging, either theoretical (comparison of bounds) or practical (comparison of losses on actual datasets). However AA turns out to be a fundamental development of learning theory because it is optimal in many important situations. In this subsection, the results from [Vov98b] concerning the optimality of AA are outlined.

Consider a game $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ satisfying conditions REG_1–REG_5. Let us formulate an optimality counterpart for Corollary 2.

Informally, optimality of AA means the following. Suppose that there is a merging technique $\mathfrak{A}$ that, for every finite pool Θ of n experts $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$, every positive integer T, and every $\omega_1, \omega_2, \ldots, \omega_T \in \Omega$, achieves loss satisfying the inequalities

$$\mathrm{Loss}_{\mathfrak{A}}^{\lambda}(\omega_1, \omega_2, \ldots, \omega_T) \leq c \, \mathrm{Loss}_{\mathcal{E}_i}^{\lambda}(\omega_1, \omega_2, \ldots, \omega_T) + a \ln n \tag{3.10}$$

for every $i = 1, 2, \ldots, n$ and some nonnegative constants c and a. Then there is $\beta \in (0, 1)$ such that the inequalities

$$\begin{aligned} c(\beta) &\leq c \\ \frac{c(\beta)}{\ln(1/\beta)} &\leq a \end{aligned}$$

hold. In other words, the pairs of (c, a) achieved by the AA are the best and cannot be improved by other merging techniques.

This can be made more precise by utilising the antagonistic game from Subsect. 2.3. Let $\mathcal{G}(c, a)$ be the following antagonistic perfect information game. Initially, one player, called the *Environment*, chooses a positive integer n (the number of experts). Then players act according to Protocol 2, where the Environment generates the experts' predictions $\gamma_t^{(\theta)}$ and outcomes ω_t while the second player, the *Learner*, acts as $\mathfrak{A}$ from the protocol and outputs predictions γ_t. The Environment wins if (3.10) ceases to be true for some $i = 1, 2, \ldots, n$ on some trial $T > 0$. Otherwise the Learner wins. Note that the Learner's victory cannot be established on any trial. It can only win after infinitely many trials have been completed.

Proposition 3 ([Vov98b]). *For every c and a the game $\mathcal{G}(c, a)$ is determined, i.e., either the Environment or the Learner has a winning strategy*[2].

Let $\mathcal{L}$ be the set of all pairs $(c, a) \in [0, +\infty)^2$ such that the Learner has a winning strategy in $\mathcal{G}(c, a)$. Clearly, if $(c, a) \in \mathcal{L}$ then $(c', a') \in \mathcal{L}$ for every

[2] Only deterministic strategies are considered.

$c' \geq c$ and $a' \geq a$. Let the *separation curve* be the boundary of $\mathcal{L}$ relative to the quadrant $(c,a) \subseteq [0,+\infty)^2$ or, in other words, the set

$$\left\{(c,a) \in [0,+\infty)^2 \mid a = \inf\{a' \mid (c,a') \in \overline{\mathcal{L}}\}\right\} \\ \cup \left\{(c,a) \in [0,+\infty)^2 \mid c = \inf\{c' \mid (c',a) \in \overline{\mathcal{L}}\}\right\} \ .$$

Here $\overline{\mathcal{L}}$ refers to the closure of $\mathcal{L}$ w.r.t. the standard topology of R^2 and $\inf \varnothing$ is regarded as nonexistent. On the other hand, consider the curve $\mathcal{AA}$ defined as follows. It consists of all pairs $(c(\beta), c(\beta)/\ln(1/\beta))$, $\beta \in (0,1)$ and the points $(c(0),a(0))$ and $(c(1),a(1))$, where

$$c(0) = \lim_{\beta\to 0+} c(\beta) \ , \qquad c(1) = \lim_{\beta\to 1-} c(\beta) \ ,$$

$$a(0) = \lim_{\beta\to 0+} \frac{c(\beta)}{\ln(1/\beta)} \ , \qquad c(1) = \lim_{\beta\to 1-} \frac{c(\beta)}{\ln(1/\beta)} \ ,$$

provided the coordinates are finite (these limits may be shown to always exist but they can be infinite). Proposition 2 implies that if $c' \geq c(\beta)$ and $a' \geq c(\beta)/\log(1/\beta)$ for some $\beta \in (0,1)$, then the Learner wins in $\mathcal{G}(c',a')$ by using AA.

Now optimality can be formulated.

Proposition 4 ([Vov98b]). *The separation curve coincides with* $\mathcal{AA}$.

Other properties of $\mathcal{L}$ and $c(\beta)$ include

Proposition 5 ([Vov98b]). *(i)* $\mathcal{AA} \subseteq \mathcal{L}$*, i.e.,* $\mathcal{L}$ *is closed w.r.t. the standard topology of* $\mathbb{R}^2$*,*

(ii) $\mathcal{L} \subseteq [1,+\infty) \times [0,+\infty)$*,*

(iii) $c(\beta)$ *is continuous on* $(0,1)$*, and*

(iv) $c(\beta)$ *is non-increasing on* $(0,1)$*.*

Chapter 4

The Multiplicative Constant $c(\beta)$ and the Concept of Mixability

The multiplicative constant $c(\beta)$ is of fundamental importance for AA since it determines its performance when the length of the outcome string tends to infinity. This constant determines how much worse the asymptotic performance of the AA is than the performance of the best expert. It can be said that the AA provides us with $c(\beta)$-competitive algorithms (see e.g. [BDBK^{+}94]), where we treat the loss as the*cost* of a prediction strategy.

Proposition 5, (ii) implies that $c(\beta) \geq 1$. This is a very natural statement; it means that a composition of experts' predictions can not always work better than the best expert.

If Γ is a metric compact, the proof is very simple. Starting from any $\gamma \in \Gamma$, one can construct a sequence $\gamma_1, \gamma_2, \ldots \in \Gamma$ such that $\lambda(\omega, \gamma_i) \geq c(\beta)\lambda(\omega, \gamma_{i+1})$ for every $\omega \in \Omega$ and every positive integer i. There exists a convergent subsequence of $\gamma_1, \gamma_2, \ldots$; let γ_0 be its limit. If $c(\beta) < 1$, then $\lambda(\omega, \gamma_0) = 0$ for every $\omega \in \Omega$. This contradicts REG_5.

Therefore the case $c(\beta) = 1$ is the best that can be achieved. Its importance motivates the following definitions (see [VW98]).

Definition 1. A game $\mathfrak{G}$ is β-*mixable* if $\beta \in (0, 1)$ and $c(\beta) = 1$.

Definition 2. A game $\mathfrak{G}$ is *mixable* if it is β-mixable for some $\beta \in (0, 1)$.

In this chapter we introduce the geometric interpretations of $c(\beta)$ and the concept of mixability after [Vov90] and [Vov98b]. This interpretation allows us to establish many important properties of $c(\beta)$.

We restrict ourselves to the binary case $\Omega = \mathbb{B} = \{0,1\}$ until the final section, where the contrary is explicitly stated. The games with $\Omega = \mathbb{B} = \{0,1\}$ will be referred to as *binary* games.

4.1 Geometric Images of Binary Games

In the binary case $\Omega = \mathbb{B}$, many concepts connected with the AA have a simple geometric interpretation. We are going to use it extensively. A prediction $\gamma \in \Gamma$ generates a pair of numbers $(\lambda(0,\gamma), \lambda(1,\gamma))$ and therefore a point on the extended Euclidean plane $[-\infty,+\infty]^2$. Elements of the set $P = \{(\lambda(0,\gamma), \lambda(1,\gamma)) \mid \gamma \in \Gamma\}$ may be identified with predictions.

Different continuous parametrisations of a given set P do not introduce any change into the way AA operates. All of them are equivalent as far as AA is concerned. In fact, the equivalence can be extended even further. We need the following definition.

Definition 3. A *superprediction* w.r.t. a game $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$ is a pair of numbers $(s_0, s_1) \in [-\infty,+\infty]^2$ such that

$$\begin{aligned} \lambda(0,\gamma) &\leq s_0 \ , \\ \lambda(1,\gamma) &\leq s_1 \end{aligned}$$

for some $\gamma \in \Gamma$.

Within the context of the geometrical interpretation the set S of all superpredictions is the set of all points that lie to the 'north-east' of P. In Fig. 4.1 you can see the sets P and S for the discrete square-loss game.

By analogy with (3.1) we can define a generalised superprediction to be a pair $g = (g^{(0)}, g^{(1)})$ such that

$$\left\{ \begin{array}{lcl} g^{(0)} & = & \log_\beta \sum_{i=1}^k p_i \beta^{s_k^{(0)}} \\ g^{(1)} & = & \log_\beta \sum_{i=1}^k p_i \beta^{s_k^{(1)}} \end{array} \right. ,$$

where k is some positive integer, $p_1, p_2, \ldots, p_k \in [0,1]$ are such that $p_1 + p_2 + \cdots + p_k = 1$, and $s_1 = (s_1^{(0)}, s_1^{(1)}), s_2 = (s_2^{(0)}, s_2^{(1)}), \ldots, s_k = (s_k^{(0)}, s_k^{(1)})$ are superpredictions from S. By analogy with (3.2) for every generalised

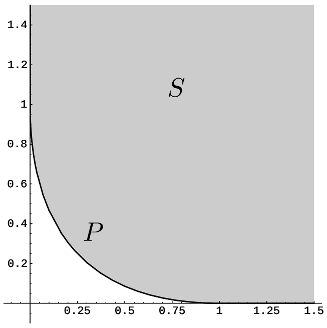

Figure 4.1: The sets of predictions and superpredictions for the discrete square-loss game

superprediction $g = (g^{(0)}, g^{(1)})$ the numbers $\inf_{(s^{(0)},s^{(1)})\in S} \max\left(\frac{s^{(0)}}{g(\omega)}, \frac{s^{(1)}}{g(\omega)}\right)$ can be considered. However it is easy to see that this approach provides us with the same value of $c(\beta)$.

The sets of superpredictions is the terminal point of our generalisation. All games with the same set of superpredictions behave similarly from our point of view. On the other hand, we will see below that games with different sets of superpredictions are substantially different.

The regularity assumptions from Sect. 2.4 reduce to a set of geometrical principles. The following theorem summarises them.

Theorem 2. *Let $S \subseteq (-\infty, +\infty]^2$ be a set such that for every $(x, y) \in S$ and every $u, v \in [0, +\infty]$ we have $(x + u, y + v) \in S$. Then it is a set of superpredictions for some game $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$ satisfying REG_1–REG_5 if and only if the following conditions hold:*

BIN_1 $S \subseteq [0, +\infty]^2$,

BIN_2 $(0, 0) \notin S$,

BIN_3 *S is closed w.r.t. the extended topology of $[-\infty, +\infty]^2$, and*

BIN_4 $S \cap \mathbb{R}^2 \neq \varnothing$.

This game satisfies REG_6 if and only if

BIN_5 *The set S is the closure of its finite part $S \cap \mathbb{R}^2$ w.r.t. the extended topology of $[-\infty, +\infty]^2$.*

We say that a game satisfies a condition BIN_i from the list if its set of superpredictions satisfies it.

An example of S prohibited by Condition BIN_5 is S containing the 'line segment' connecting the points $(a, +\infty)$, $(b, +\infty)$, where $a, b \in \mathbb{R}$, but no points below this segment, i.e., points $(u, v) \in \mathbb{R}^2$ such that $a \leq u \leq b$.

Different games with different loss functions may have the same set of superpredictions S. They are essentially mere parametrisations of S. It would be convenient to have a 'parameter–independent' way of describing S. One0 possible solution to this problem is provided by the following concept.

We will say that $f : I \to \mathbb{R}$, where $I = (a, b) \subseteq \mathbb{R}$ is an open (perhaps infinite) interval, is a *canonical specification* of a game $\mathfrak{G}$ with the set of superpredictions S, if

- f is non-increasing,
- f is semi-continuous from the right,
- f is not constant on every non-void interval $(a', b) \subseteq I$, and
- S is a closure w.r.t. the extended topology of the set $\{(x, y) \in \mathbb{R}^2 \mid \exists \tilde{x} \leq x : f(\tilde{x}) \leq y\}$.

Lemma 1. *Let $\mathfrak{G}$ be a game with the set of superpredictions S which satisfies BIN_1 and BIN_4. Then S satisfies BIN_3 and BIN_5 if and only if either $S = [a, +\infty] \times [b, +\infty]$ for some $a, b \geq 0$ or there is a canonical specification of $\mathfrak{G}$. If under the above conditions there is a canonical specification, it is unique.*

Proof. The 'only if' implication is trivial. Let us prove the 'if' part. Consider $\mathfrak{G}$ with S satisfying BIN_1–BIN_5. Let $\tilde{I} = (0, +\infty)$ and let $\tilde{f}(x) = \inf\{y \in \mathbb{R} \mid (x, y) \in S\}$, where $\inf \varnothing = +\infty$. Let $a = \sup\{x \mid \tilde{f}(x) = +\infty\}$ and $b = \inf\{\tilde{b} \mid \tilde{f}$ is constant on $(b, +\infty)\}$ and consider $I = (a, b)$. It is easy to check that the restriction $\tilde{f}|_I$ is a canonical specification of $\mathfrak{G}$.

Uniqueness follows from the observation that if f is a canonical representation of $\mathfrak{G}$, then $f(x) = \inf\{y \in \mathbb{R} \mid (x, y) \in S\}$. □

4.2 The Geometric Interpretation of $c(\beta)$

We begin with some notation.

Let us introduce several transformations of subsets of the plane. Consider $A \subseteq [-\infty,+\infty]^2$ and real numbers u,v,t. By $A+(u,v)$ denote the shift of A by the vector (u,v), i.e., the set $\{(x+u,y+v)\mid (x,y)\in A\}$ and by tA denote the result of the scaling by t, i.e., the set $\{(tx,ty)\mid (x,y)\in A\}$.

A slightly more complicated transformation is provided by the equation

$$\mathfrak{B}_\beta(x,y) = (\beta^x,\beta^y) \ . \tag{4.1}$$

Clearly, for every $\beta\in(0,1)$, this transformation is a homeomorphism from $(-\infty,+\infty]^2$ onto $[0,+\infty)^2$. It is also a homeomorphism from $[0,+\infty]^2$ onto $[0,1]^2$. The inverse transformation is

$$\mathfrak{B}_\beta^{-1}(x,y) = (\log_\beta x,\log_\beta y) \ . \tag{4.2}$$

Let us define a function on sets $A\subseteq[0,+\infty]^2$. Put $A_\theta = \{t\in[0,+\infty)\mid (t\cos\theta,t\sin\theta)\in A\}\subseteq\mathbb{R}$ and

$$\operatorname{osc}_\theta A = \begin{cases} 1 & \text{if} \quad A_\theta=\varnothing \\ \sup A_\theta/\inf A_\theta & \text{otherwise} \end{cases} \ .$$

This function is a measure of how ‘thick’ the sections of A by the half-line $\{(t\cos\theta,t\sin\theta)\mid t\ge 0\}$ are. Put

$$\operatorname{osc} A = \sup_{\theta\in[0,\pi/2]} \operatorname{osc}_\theta A \ . \tag{4.3}$$

Finally, given $A,B\subseteq[-\infty,+\infty]^2$, the *B-closure* of A, denoted by $\operatorname{cl}_B A$, is the intersection of all shifts of B that contain A, i.e.

$$\operatorname{cl}_B A = \bigcap_{u,v\in\mathbb{R}:B+(u,v)\supseteq A} B+(u,v) \ . \tag{4.4}$$

Now we can give an interpretation of $c(\beta)$ which goes back to [Vov90]. Consider a game $\mathfrak{G}$ with the set of superpredictions S satisfying BIN_1–BIN_4 and $\beta\in(0,1)$. Clearly, generalised superpredictions fill the convex hull of the set $\mathfrak{B}_\beta(S)$. We use the notation $\mathfrak{C}(A)$ from [Egg58] for the convex hull of $A\subseteq\mathbb{R}^2$. It follows from the discussion in Sect. 4.1 above that

$$c(\beta) = \operatorname{osc}\left(\mathfrak{B}_\beta^{-1}\left(\mathfrak{C}\left(\mathfrak{B}_\beta(S)\right)\right)\setminus S\right) \ . \tag{4.5}$$

Since S is closed by BIN_3, the set $\mathfrak{B}_\beta(S)$ is also closed. According to a fact from convex analysis, the convex hull of a closed bounded set $A \subseteq \mathbb{R}^2$ is the intersection of all half-plains that contain A (see, say, [Egg58], Theorem 11). Note that $\mathfrak{B}_\beta$ establishes a correspondence between shifts of the curve $\beta^x + \beta^y = 1$, i.e., the curves $\beta^{x+a} + \beta^{y+b} = 1$, where a, b are some constants, and the straight lines $Ax + By + C = 0$, where $A, B > 0$ and C are some constants. Put $B_\beta = \{(x, y) \subseteq [0, +\infty]^2 \mid \beta^x + \beta^y \leq 1\}$. We have proved the following statement.

Proposition 6 ([Vov90, Vov98b]). *If $\mathfrak{G}$ is a game with the set of superpredictions S satisfying BIN_1–BIN_4, then*

$$c(\beta) = \operatorname{osc}(\operatorname{cl}_{B_\beta} S \setminus S)$$

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

Another important corollary of (4.5) is

Proposition 7 ([Vov90]). *A game $\mathfrak{G}$ with the set of superpredictions S satisfying BIN_1–BIN_4 is β-mixable, where $\beta \in (0, 1)$, if and only if $\mathfrak{B}_\beta(S)$ is convex.*

4.3 Some Simple Properties of $c(\beta)$

In this section we formulate and prove several (mostly trivial) properties of $c(\beta)$ and mixability for future reference.

We start with technical lemmas which describe some geometric properties of $B_\beta = \{(x, y) \subseteq [0, +\infty]^2 \mid \beta^x + \beta^y \leq 1\}$ and its boundary ∂B taken w.r.t. the extended topology of $[-\infty, \infty]^2$.

Lemma 2. *Let $\beta \in (0, 1)$. Then the following statements hold:*

(i) For every two points $(u_1, v_1), (u_2, v_2) \in (-\infty, +\infty]^2$ such that $u_1 < u_2$ and $v_2 < v_1$ there is only one shift of the curve B_β such that ∂B (taken w.r.t. the extended topology) passes through these points.

(ii) The boundaries (w.r.t. the extended topology) of two different shifts of B_β can have no more than 2 common points in $[-\infty, +\infty]$.

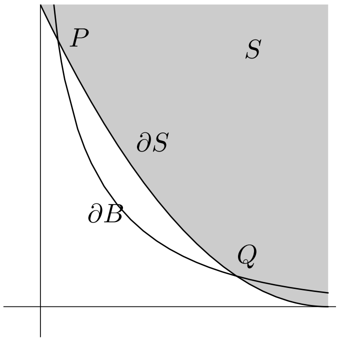

Figure 4.2: The sets from Lemma 3

Proof. Statement (ii) follows from (i). To prove (i), consider the transformation $\mathfrak{B}_\beta$. There is a unique straight line passing through $\mathfrak{B}_\beta(u_1, v_1)$ and $\mathfrak{B}_\beta(u_2, v_2)$. Its equation may be written as $ax + by + c = 0$, where $a, b > 0$ and c are constants. It corresponds to a shift of ∂B_β. □

Lemma 3. *Let $\mathfrak{G}$ be a game with the set of superpredictions S satisfying BIN_1–BIN_5, let $\beta \in (0,1)$. Consider two points $P, Q \in \partial S$. Suppose that there is a shift $B = B_\beta + (a, b)$ such that $P, Q \subseteq \partial B$ and the arc of ∂B between P and Q lies outside S (except for the endpoints). Then the set $\mathrm{cl}_{B_\beta}(S)$ includes the interior of the region bound by arcs of ∂S and ∂B.*

The informal meaning of the lemma is that if a shift of the curve $\beta^x + \beta^y = 1$ passes through two points on the boundary of S and the arc of the curve lies outside S, then the 'lens' bound by the two arcs cannot be cut off by other shifts of the curve (see Fig. 4.2, where the set S is shaded).

Proof. The lemma follows from the definition of $\mathrm{cl}_{B_\beta}(S)$ as the inverse image of the convex hull of $\mathfrak{B}_\beta(S)$. The image of the arc of ∂B_β between P and Q is the line segment connecting $\mathfrak{B}(P)$ and $\mathfrak{B}(Q)$. The lemma follows. □

It is easy to show that $c(\beta)$ remains intact if we reflect S in the straight line $x = y$. Indeed, little changes if we 'swap' the outcomes 0 and 1, i.e., if 0 is renamed 1 and 0 is renamed 1.

Lemma 4. *Let $\mathfrak{G}_1$ be a game with the set of superpredictions S and $\mathfrak{G}_2$ be the game with the set of superpredictions obtained by reflecting S in the bisector of the positive quadrant. Then $\mathfrak{G}_1$ satisfies any of the conditions BIN_1–BIN_5*

if and only if $\mathfrak{G}_2$ *satisfies them and, provided all the conditions are satisfied,* $c(\mathfrak{G}_1, \beta) = c(\mathfrak{G}_2, \beta)$ *for every* $\beta \in (0,1)$.

The following theorem shows that convexity is a necessary condition for mixability. However we will see below that it is not sufficient.

Theorem 3. *Let* $\mathfrak{G}$ *be a game with the set of superpredictions* S *satisfying* BIN_1*–*BIN_4*. If* $S \cap \mathbb{R}^2$ *is not convex, there is* $\delta > 0$ *such that* $c(\beta) > 1 + \delta$ *for all* $\beta \in (0,1)$ *and thus* $\mathfrak{G}$ *is not mixable.*

Proof. There are points $B_0, B_1 \in S$ such that the line segment l connecting B_0 and B_1 is not a subset of S. Without restricting the generality we can assume that $B_0, B_1 \in \partial S$ and the intersection $l \cap S$ consists of B_0 and B_1. Let A be the interior of the set bound by l and the arc of ∂S between B_0 and B_1. Since S is closed, A is not empty.

Lemma 3 implies that $A \subseteq \mathrm{cl}_{B_\beta} S$ for every $\beta \in (0,1)$. Therefore $\mathrm{osc}(\mathrm{cl}_{B_\beta} S \setminus S) \geq \mathrm{osc}\, A$. It remains to note that $\mathrm{osc}\, A > 1$. □

The next theorem allows us to investigate 'parts' of a game separately.

Theorem 4. *Let* $\mathfrak{G}$ *be a game with the set of superpredictions* S *satisfying* BIN_1*–*BIN_5 *and let* $f : (a,b) \to \mathbb{R}$ *be its canonical representation. Let* $c \in (a,b)$ *and let* $f|_{(a,c)}$ *and* $f|_{(c,b)}$ *be the canonical representations of the games* $\mathfrak{G}_1$ *and* $\mathfrak{G}_2$*, respectively. Then* $\mathfrak{G}$ *is mixable if and only if* $\mathfrak{G}_1$ *and* $\mathfrak{G}_2$ *are mixable and* $S \cap \mathbb{R}^2$ *is convex.*

Proof. The 'only if' part is trivial. Let us prove the 'if' part. Let S_1 and S_2 be the sets of superpredictions for $\mathfrak{G}_1$ and $\mathfrak{G}_2$, respectively. Since $\mathfrak{G}_1$ and $\mathfrak{G}_2$ are mixable, there is $\beta \in (0,1)$ such that they are β-mixable and the sets $\mathfrak{B}_\beta(S_1)$ and $\mathfrak{B}_\beta(S_2)$ are convex. Indeed, if $\mathfrak{G}_1$ is β_1-mixable and $\mathfrak{G}_2$ is β_2-mixable, take $\beta = \max(\beta_1, \beta_2)$.

Let $g : (\beta^b, \beta^a) \to \mathbb{R}$ be the function defined by $g(x) = \beta^{f(\log_\beta x)}$, i.e., the graph of g is the image of the graph of f under the transformation $\mathfrak{B}_\beta$. The function g is concave on (β^b, β^c) and on (β^c, β^a). It is concave on the whole interval (β^b, β^a) if and only if the following inequality for one-sided derivatives holds:

$$g'_-(\beta^c) \geq g'_+(\beta^c) \enspace .$$

Since

$$\begin{aligned}\frac{d\beta^{y(t)}}{d\beta^{x(t)}} &= \frac{\ln\beta \cdot \beta^{y(t)} y'(t)}{\ln\beta \cdot \beta^{x(t)} x'(t)} \\ &= \beta^{y(t)-x(t)} \frac{y'(t)}{x'(t)}\end{aligned} \tag{4.6}$$

holds, this inequality follows from $f'_+(c) \geq f'_+(c)$, which in turn follows from the convexity of f. □

Corollary 3. *Let $\mathfrak{G}$ be a game with the set of superpredictions S satisfying BIN_1–BIN_5 and let $f : (a,b) \to \mathbb{R}$ be its canonical representation. Let $c,d \in (a,b)$ are such that $c < d$ and let $f|_{(a,d)}$ and $f|_{(c,b)}$ be the canonical representations of the games $\mathfrak{G}_1$ and $\mathfrak{G}_2$, respectively. Then $\mathfrak{G}$ is mixable if and only if $\mathfrak{G}_1$ and $\mathfrak{G}_2$ are mixable.*

4.4 Differential Criteria of Mixability

If the boundary of S is smooth, some simple criteria of mixability can be derived. The following theorem is a restatement of a theorem from [HKW98].

Theorem 5. *Let $\mathfrak{G}$ be a game with the set of superpredictions S satisfying BIN_1–BIN_5. Suppose that there are twice differentiable functions $x, y : I \to \mathbb{R}$, where $I \subseteq \mathbb{R}$ is an open (perhaps infinite) interval, such that $x' > 0$ and $y' < 0$ on I and S is the closure of the set $\{(u,v) \in \mathbb{R}^2 \mid \exists t \in I : x(t) \leq u, y(t) \leq v\}$ w.r.t. the extended topology of $[-\infty, +\infty]^2$. Then, for every $\beta \in (0,1)$, the game $\mathfrak{G}$ is β-mixable if and only if*

$$\ln\frac{1}{\beta} \leq \frac{y''(t)x'(t) - x''(t)y'(t)}{x'(t)y'(t)(y'(t) - x'(t))} \tag{4.7}$$

holds for every $t \in I$. The game $\mathfrak{G}$ is mixable if and only if the fraction $(y''x' - x''y')/x'y'(y' - x')$ is separated from zero, i.e., there is $\varepsilon > 0$ such that

$$\frac{y''x' - x''y'}{x'y'(y' - x')} \geq \varepsilon \tag{4.8}$$

holds on I.

Proof. Convexity of $\mathfrak{B}_\beta(S)$ is equivalent to concavity of the function with the graph $\{\mathfrak{B}_\beta(x(t), y(t)) \mid t \in I\}$. Since the functions $x(t)$ and $y(t)$ are smooth, this curve is concave if and only if

$$\frac{d^2 \beta^{y(t)}}{d\left(\beta^{x(t)}\right)^2} \leq 0 \tag{4.9}$$

holds on I. Further differentiation of (4.6) yields

$$\begin{aligned}
\frac{d^2 \beta^{y(t)}}{d\left(\beta^{x(t)}\right)^2} &= \frac{1}{\ln\beta \cdot \beta^{x(t)} x'(t)} \Bigg((y'(t) - x'(t)) \ln\beta \cdot \beta^{y(t)-x(t)} \frac{y'(t)}{x'(t)} \\
&\quad + \beta^{y(t)-x(t)} \frac{y''(t)x'(t) - y'(t)x''(t)}{\left(x'(t)\right)^2} \Bigg) \\
&= \frac{\beta^{y(t)-2x(t)}}{\ln\beta \cdot \left(x'(t)\right)^2} \Bigg((y'(t) - x'(t))y'(t) \ln\beta \\
&\quad + \frac{y''(t)x'(t) - y'(t)x''(t)}{x'(t)} \Bigg) .
\end{aligned}$$

Inequality (4.9) reduces to

$$(y'(t) - x'(t))y'(t) \ln\beta \geq -\frac{y''(t)x'(t) - y'(t)x''(t)}{x'(t)} . \tag{4.10}$$

The theorem follows from the assumptions about the signs of the derivatives x' and y'. □

Corollary 4. *Let $\mathfrak{G}$ be a game satisfying BIN_1–BIN_5 with the canonical representation $f : I \to \mathbb{R}$. Suppose that f is twice differentiable on I. Then, for every $\beta \in (0,1)$, the game $\mathfrak{G}$ is β-mixable if and only if*

$$\ln\frac{1}{\beta} \leq \frac{f''(x)}{f'(x)(f'(x) - 1)}$$

holds for every $x \in I$. The game $\mathfrak{G}$ is mixable if and only if the fraction $f''/f'(1 - f')$ is separated from the zero, i.e., there is $\varepsilon > 0$ such that

$$\frac{f''}{f'(f' - 1)} \geq \varepsilon \tag{4.11}$$

holds on I.

Proof. The proof is by taking the parameter $t = x$ in Theorem 5. Note that the derivative of f does not vanish on I. Had it vanished at $x_0 \in I$, the function f would have been constant on $I \cap (x_0, +\infty)$. □

Theorem 4 implies a similar statement for piecewise twice-differentiable functions.

Corollary 5. *Let $\mathfrak{G}$ be a game with the canonical representation $f : I \to \mathbb{R}$. Suppose that f is piecewise twice differentiable on I, i.e., there are numbers $a = x_1 < x_2 < \ldots < x_n = b$, where n is some positive integer, such that f is twice differentiable on every open interval (x_i, x_{i+1}), $i = 1, 2, \ldots, n-1$. Then $\mathfrak{G}$ is β-mixable, where $\beta \in (0,1)$, if and only if*

- *f is convex, and*
- *for every $x \in I$, the inequality*
$$\ln \frac{1}{\beta} \leq \frac{f''(x)}{f'(x)(f'(x)-1)}$$
holds.

We need the following definition to simplify the statements of theorems. Let $\mathfrak{G}$ be a game with the set of superpredictions S satisfying BIN_1–BIN_5 and let $f : (a,b) \to \mathbb{R}$ be the canonical representation of $\mathfrak{G}$. If there is $c < b$ such that the game with the set of superpredictions $S \cap [c, +\infty] \times (-\infty, +\infty]$ is mixable, we say that $\mathfrak{G}$ has a *mixable 0-edge.* If the game obtained by reflecting S in the straight line $x = y$ has a mixable 0-edge, we say that $\mathfrak{G}$ has a *mixable 1-edge.*

The concept of a mixable edge allows us to formulate the following criterion of mixability.

Theorem 6. *Let $\mathfrak{G}$ be a game with the set of superpredictions S satisfying BIN_1–BIN_5 and let $f : (a,b) \to \mathbb{R}$ be its canonical representation. Suppose that f is twice differentiable on (a,b). Then $\mathfrak{G}$ is mixable if and only if $f''(x)$ does not vanish on (a,b) and $\mathfrak{G}$ has mixable 0 and 1-edges.*

Proof. The 'only if' part immediately follows from Theorem 4 and Corollary 4.

Let us prove the 'if' part. There are $c, d \in (a,b)$ such that the games with canonical representations $f|_{(a,c)}$ and $f|_{(d,b)}$ are mixable. Take $c' \in (a,c)$ and

$d' \in (d, b)$. Since the continuous function $f''/f'(f'-1)$ does not vanish on the closed interval $[c', d']$, it is separated from 0 on this interval. Corollary 4 implies that the game with the canonical representation $f|_{(c',d')}$ is mixable. Now we apply Corollary 3. □

In order to make this theorem practical, we need to develop sufficient conditions for the property of having a mixable edge. Particularly simple conditions may be formulated for bounded games.

Let $\mathfrak{G}$ be a game with the set of superpredictions S satisfying BIN_1–BIN_5 and let $f : (a, b) \to \mathbb{R}$ be the canonical representation of $\mathfrak{G}$. If $b < +\infty$, we say that $\mathfrak{G}$ has a *bounded 0-edge.* If the game obtained by reflecting S in the straight line $x = y$ has a bounded 0-edge, we say that $\mathfrak{G}$ has a *bounded 1-edge.* If a game has bounded 0 and 1-edges, it is *bounded.* If an edge is not bounded, it is said to be *unbounded.*

An equivalent definition of a bounded game is a game with the set of superpredictions that can be specified by a bounded loss function.

If $\mathfrak{G}$ has a bounded 0-edge, we may extend f to the segment $(a, b]$. Let $\bar{f} : (a, b] \to \mathbb{R}$ be the function which coincides with f on (a, b) and defined by the equation

$$\bar{f}(b) = \lim_{x \to b-} f(x) \tag{4.12}$$

at b.

Theorem 7. *Let $\mathfrak{G}$ be a game with the set of superpredictions S satisfying BIN_1–BIN_5 and let $f : (a, b) \to \mathbb{R}$ be its canonical representation. Let $\mathfrak{G}$ have a bounded 0-edge and let $\bar{f}$ be an extension of f specified by (4.12). If $\bar{f}$ is twice continuously differentiable*[1] *on $(c, b]$, where c is a number from (a, b), and any of the following conditions hold, then $\mathfrak{G}$ has a mixable 0-edge:*

1. *$\bar{f}''(b) > 0$, or*

2. *The derivative $\bar{f}'(b)$ vanishes and there is a positive integer $k > 2$ such that $\bar{f}(x)$ has derivatives up to the order k at b and $f^{(m)}(b) = 0$ for every $m = 1, 2, \ldots, k-1$, but $f^{(k)}(b)$ is of sign $(-1)^k$.*

Proof. If (1) holds, then f'' is separated from 0 in a punctured vicinity of b and so is the fraction $f''/f'(f'-1)$. Thus we can apply Theorem 4.

[1] By the derivative at the point b we mean the left derivative $f'_-(b)$.

By Taylor's theorem, $\bar{f}''(x) = \alpha(x-b)^k + o\left((x-b)^k\right)$ as $x \to b-$, where k is a positive integer and $\alpha > 0$. It is easily shown that $\bar{f}'(x) = \alpha(x-b)^{k+1}/(k+1) + o\left((x-b)^{k+1}\right)$ and therefore

$$\lim_{x \to b-0} \frac{f''(x)}{f'(x)} = -\infty \ .$$

This implies that $f''/f'(1-f')$ is separated from 0 in a punctured vicinity of b. □

The results concerning the mixable 0-edge can be summarised in the following 'procedure' for checking the mixability. Let $f : (a,b) \to \mathbb{R}$ be the canonical representation of $\mathfrak{G}$ with a set of superpredictions S and a bounded 0-edge and let $\bar{f}$ be as above. Suppose that $\bar{f}$ is infinitely differentiable at b. Then:

- if $\bar{f}''(b) > 0$ then $\mathfrak{G}$ has a mixable 0-edge
- if $\bar{f}''(b) < 0$ then $\mathfrak{G}$ does not have a mixable 0-edge and moreover S is not convex
- if $\bar{f}''(b) = 0$ then:
  - if $\bar{f}'(b) < 0$ then $\mathfrak{G}$ does not have a mixable 0-edge
  - if $\bar{f}'(b) = 0$ then:
    - if $\exists k > 2 : \bar{f}'(b) = 0, \bar{f}''(b) = 0, \ldots, \bar{f}^{(k-1)}(b) = 0$ and $\bar{f}^{(k)}(b)$ is of sign $(-1)^k$, then $\mathfrak{G}$ has a mixable 0-edge
    - if $\exists k > 2 : \bar{f}'(b) = 0, \bar{f}''(b) = 0, \ldots, \bar{f}^{(k-1)}(b) = 0$ and $\bar{f}^{(k)}(b) < 0$ is of sign $(-1)^{k+1}$, then $\mathfrak{G}$ does not have a mixable 0-edge and moreover S is not convex.

If all derivatives vanish at b, this procedure comes to no conclusion; we need to analyse the behaviour of the fraction $f''(x)/f'(x)$ as x approaches b. The same applies to the case when there is not enough derivatives at b.

4.5 Mixability of Specific Games

Let us now apply our statements to check whether some specific games are mixable. There are many ways to prove the following theorem (see [Vov90], [Vov98a]); our theory gives a more straightforward method.

Theorem 8. (*i*) *The β-logarithmic-loss game is β'-mixable if and only if $\beta' \in [\beta, 1)$; the logarithmic-loss game is β'-mixable if and only if $\beta' \in [1/2, 1)$.*

(*ii*) *The discrete A, B-bounded square-loss game is β-mixable if and only if $\beta \in [e^{-2/(B-A)^2}, 1)$; the discrete square-loss game is β-mixable if and only if $\beta \in [e^{-2}, 1)$.*

Proof. The proof is by Theorem 5. We take $x(t) = \lambda(0, t)$ and $y(t) = \lambda(1, t)$, where λ is the loss function from the definition, for the β-logarithmic-loss and the discrete square-loss games.

The discrete A, B-bounded square-loss game can be scaled and thus reduced to the game with $\Omega = \Gamma = \{0, 1\}$ and $\lambda(\omega, \gamma) = (B - A)^2(\omega - \gamma)^2$. □

Similarly, Theorem 5 implies that the discrete absolute-loss games are not mixable. We will obtain a more precise result about them later.

We know from Proposition 5, (iv) that $c(\beta)$ does not increase in β in the general case. Theorem 8 allows us to obtain a simple derivation of this property for the binary case.

Corollary 6. *For every $A \subseteq (-\infty, +\infty]^2$, $0 < \beta_1 \le \beta_2 < 1$, and $\theta \in [0, 2\pi)$ we have $\operatorname{osc}_\theta \operatorname{cl}_{B_{\beta_2}} A \le \operatorname{osc}_\theta \operatorname{cl}_{B_{\beta_1}} A$.*

Proof. Indeed, the set B_β is the set of superpredictions for the β-logarithmic-loss game. Since β_1-logarithmic-loss game is β_2-mixable, $\operatorname{cl}_{B_{\beta_2}} B_{\beta_1} = B_{\beta_1}$. Therefore $\operatorname{cl}_{B_{\beta_2}} A \subseteq \operatorname{cl}_{B_{\beta_1}} A$. □

4.6 Non-mixable Games

We now move on to investigating the behaviour of $c(\beta)$ in the general case. In this section we show that in many cases $c(\beta) \to 1$ as $\beta \to 1-$ and investigate the type of convergence. A notable exception is provided by a class of games with $c(\beta) \equiv +\infty$.

4.6.1 Absolute-Loss Games

Let us evaluate $c(\beta)$ for the absolute-loss game.

Theorem 9. *For the discrete A, B-absolute-loss game we have*

$$c(\beta) = \frac{|B - A| \ln 1/\beta}{2 \ln \frac{2}{\beta^{|B-A|}+1}}$$

for every $\beta \in (0, 1)$. Consequently, for the discrete absolute-loss game and every $\beta \in (0, 1)$ we have

$$c(\beta) = \frac{\ln \frac{1}{\beta}}{2 \ln \frac{2}{\beta+1}} \ .$$

We will derive this from a more general statement of independent interest.

Take $a, b > 0$. Let $\mathfrak{G}^{a,b}_{abs}$ be the game with the set of superpredictions $S^{a,b}_{abs} = \{(x, y) \in [0, +\infty]^2 \mid bx + ay > ab\}$. Informally, this is a game with the set of predictions coinciding with the line segment connecting the points $(a, 0)$ and $(0, b)$. The set $S^{a,b}_{abs}$ is shaded darker on Fig. 4.3.

Lemma 5. *For all $a, b > 0$, we have*

$$c(\mathfrak{G}^{a,b}_{abs}, \beta) = \frac{ab \ln \beta}{a \ln \frac{a(1-\beta^{a+b})}{(a+b)(1-\beta^a)} + b \ln \frac{b(1-\beta^{a+b})}{(a+b)(1-\beta^b)}} \ .$$

Proof. Fix $\beta \in (0, 1)$. The curve

$$\beta^x(1 - \beta^b) + \beta^y(1 - \beta^a) = 1 - \beta^{a+b} \tag{4.13}$$

is the shift of $\beta^x + \beta^y = 1$ passing through $(a, 0)$ and $(0, b)$. If follows from Lemma 3 that $D = \mathrm{cl}_{B_\beta}(S^{a,b}_{abs}) \backslash S^{a,b}_{abs}$ is bounded by the line segment connecting the points $(a, 0)$ and $(0, b)$ and the arc of the curve (4.13) between these points.

Let us evaluate $\operatorname{osc} D$. Consider the point $(\tilde{x}, \tilde{y})$ where the tangent to the curve (4.13) is parallel to the chord connecting $(a, 0)$ and $(0, b)$. It follows from the Thales theorem of plane geometry that $\operatorname{osc} D$ is achieved on θ such that the half-line $l_\theta = \{(t \cos \theta, t \sin \theta) \mid t \geq 0\}$ passes through $(\tilde{x}, \tilde{y})$. Figure 4.3 illustrates the proof. The set D is shaded lighter and the set of superpredictions darker.

The derivative of the implicit function $y(x)$ specified by (4.13) is

$$\frac{dy}{dx} = -\beta^{x-y} \frac{1 - \beta^b}{1 - \beta^a}$$

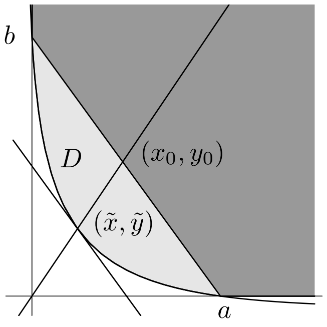

Figure 4.3: The drawing for Lemma 5

and thus $\tilde{x}, \tilde{y}$ may be found from the system

$$\begin{cases} \beta^{\tilde{x}}(1-\beta^b) + \beta^{\tilde{y}}(1-\beta^a) = 1 - \beta^{a+b} \\ -\beta^{\tilde{x}-\tilde{y}} \frac{1-\beta^b}{1-\beta^a} = -\frac{b}{a} \end{cases} .$$

This system can easily be reduced to a system linear in $\beta^{\tilde{x}}$ and $\beta^{\tilde{y}}$. The solution is

$$\tilde{x} = \log_\beta \frac{b(1-\beta^{a+b})}{(a+b)(1-\beta^b)} ,$$
$$\tilde{y} = \log_\beta \frac{a(1-\beta^{a+b})}{(a+b)(1-\beta^a)} .$$

Now let (x_0, y_0) be the intersection of l_θ with the line segment connecting $(a, 0)$ and $(0, b)$. Simple considerations from plane geometry imply that $\operatorname{osc} D = x_0/\tilde{x}$. It is easy to check that

$$x_0 = \frac{ab}{a\frac{\tilde{y}}{\tilde{x}} + b} .$$

The substitutions complete the proof. □

4.6.2 Convergence

One may notice that for the discrete logarithmic-loss game, $c(\beta) \to 1$ as $\beta \to 1$. This observation is a special case of a more general statement.

Theorem 10. *Let $\mathfrak{G}$ be a game with a set of superpredictions S satisfying BIN_1–BIN_5. If $\mathfrak{G}$ is bounded and $S \cap \mathbb{R}^2$ is convex, then $c(\beta) \to 1$ as $\beta \to 1$.*

Proof. In this proof and several subsequent proofs we will use the notation

$$c_\theta(\mathfrak{G}, \beta) = \operatorname{osc}_\theta \left(\mathfrak{B}_\beta^{-1} \left(\mathfrak{C} \left(\mathfrak{B}_\beta(S) \right) \right) \setminus S \right) \quad . \tag{4.14}$$

Clearly, $c(\mathfrak{G}, \beta) = \sup_{\theta \in [0,\pi/2]} c_\theta(\mathfrak{G}, \beta)$. Let us show that $c_\theta(\mathfrak{G}, \beta) \to 1$ uniformly in θ.

It is easy to see that $c_\theta(\beta)$ is continuous in θ if S satisfies BIN_1–BIN_5. On the other hand, for every fixed θ the function $c_\theta(\beta)$ is non-increasing in β (cf. Corollary 6). It follows from Dini's theorem (see, e.g., Theorem 7.13 from [Rud76]) that it is sufficient to show that $c_\theta(\mathfrak{G}, \beta) \to 1$ pointwise. The uniform convergence will follow.

Let $f : (a, b) \to \mathbb{R}$ be a canonical representation of $\mathfrak{G}$. Fix $\theta \in [0, \pi/2]$. If the half-line $l_\theta = \{(t \cos\theta, t \sin\theta) \mid t \geq 0\}$ does not intersect S (this can only happen for $\theta = 0$ and $\theta = \pi/2$), there is nothing to prove. Otherwise let $(x_\theta, y_\theta) \in [0, +\infty]$ be the point where l_θ first meets S, i.e., $(x_\theta, y_\theta) = (t_0 \cos\theta, t_0 \sin\theta)$, where $t_0 = \inf\{t \geq 0 \mid (t\cos\theta, t\sin\theta) \in S\}$.

If $x_\theta \notin (a, b)$, then $c_\theta(\mathfrak{G}, \beta) = 1$. Indeed, consider a point $P = (x, y)$ such that $x \geq b$ and $P \notin S$. It is easy to see that $P \notin \operatorname{cl}_\beta S$ for each $\beta \in (0, 1)$.

Now let $x_\theta \in (a, b)$. The convex set $S \cap \mathbb{R}^2$ has a hyper-plane of support at (x_θ, y_θ). This hyper-plane is a straight line with the equation that may be reduced to the form $qx + ry = s$, where $q, r > 0$. Let $(c, 0)$ and $(0, d)$ be the points where this line intersects the coordinate axes. We have $S \subseteq S_{abs}^{c,d} = \{(x, y) \in [0, +\infty] \mid qx + ry > s\}$, where the latter set is the set of superpredictions for the game $\mathfrak{G}_{abs}^{c,d}$. The inequalities

$$\begin{aligned} 1 &\leq c_\theta(\mathfrak{G}, \beta) \\ &\leq c_\theta(\mathfrak{G}_{abs}^{c,d}, \beta) \\ &\leq c(\mathfrak{G}_{abs}^{c,d}, \beta) \end{aligned}$$

hold. Since $c(\mathfrak{G}_{abs}^{c,d}, \beta) \to 1$ as $\beta \to 1-$, we obtain that $c_\theta(\mathfrak{G}, \beta) \to 1$ as $\beta \to 1-$.

□

In many cases it is possible to determine the type of convergence. Note that the game in the following theorem is not necessarily bounded.

Theorem 11. *Let $\mathfrak{G}$ be a game with the set of superpredictions S satisfying BIN_1–BIN_5. Let $f : (a, b) \to \mathbb{R}$ be its canonical representation. Suppose that $S \cap \mathbb{R}^2$ is convex and the following conditions hold:*

- *at least one of the following is true*
  - *$\mathfrak{G}$ has a mixable 0-edge or*
  - *$\mathfrak{G}$ has a bounded 0-edge and there is $\varepsilon > 0$ such that $f'_-(x) \leq -\varepsilon$ holds on (a, b), and*
- *at least one of the following is true*
  - *$\mathfrak{G}$ has a mixable 1-edge or*
  - *$\mathfrak{G}$ has a bounded 1-edge and there is $T < +\infty$ such that $f'_+(x) \geq -T$ holds on (a, b).*

Then $c(\beta) = 1 + O(\ln(1/\beta))$ as $\beta \to 1-$.

Intuitively the conditions mean the following. The theorem applies if each edge is mixable; it still applies if edges are not mixable, but they are bounded and there is no tangency with vertical and horizontal lines.

Proof. The proof is an elaboration of the proof of Theorem 10. We will show that there is $C \in \mathbb{R}$ such that the inequality

$$c_\theta(\mathfrak{G}, \beta) \leq 1 + C \ln(1/\beta) \tag{4.15}$$

holds for every $\theta \in [0, \pi/2]$.

Suppose that there are $0 < \varepsilon < T < +\infty$ such that $-T \leq f'_-(x) \leq f'_+(x) \leq -\varepsilon$ for every (a, b) (this implies that $\mathfrak{G}$ is bounded). By analogy with the proof of Theorem 10, for every $\theta \in [0, \pi/2]$ consider a half-line l_θ and let (x_θ, y_θ) be the point where it meets S provided it intersects S at all. Let m_{x_θ} be a support line to $S \cap \mathbb{R}^2$ passing through (x_θ, y_θ). For the sake of definitiveness, let m_{x_θ} be the line with the equation $y = (x - x_\theta)(f'_-(x_\theta) + f'_+(x_\theta))/2 + f(x_\theta)$ if $x_\theta \in (a, b)$. All θ such that $x_\theta \notin (a, b)$ can be ignored since $c_\theta(\mathfrak{G}, \beta) = 1$ for them for every β anyway.

Let $\underline{\theta}$ be such that $x_{\underline{\theta}} = b$ and $\overline{\theta}$ be such that $x_{\overline{\theta}} = a$. We should find $C \in \mathbb{R}$ such that (4.15) holds for every $\theta \in (\underline{\theta}, \overline{\theta})$. For every $\theta \in (\underline{\theta}, \overline{\theta})$ let $(c_\theta, 0)$ and $(0, d_\theta)$ be the points where m_{x_θ} cuts the coordinate axis. Since $\mathfrak{G}$ is bounded and slopes of m_{x_θ} are separated from $-\infty$ and 0, there are

$\underline{c}, \underline{d} > 0$ and $\overline{c}, \overline{d} < +\infty$ such that $\underline{c} \le c_\theta \le \overline{c}$ and $\underline{d} \le d_\theta \le \overline{d}$ for every $\theta \in [\underline{\theta}, \overline{\theta}]$.

For every $\theta \in [\underline{\theta}, \overline{\theta}]$, we have in much the same way as in the proof of Theorem 10

$$\begin{aligned} c_\theta(\mathfrak{G}, \beta) &\le c_\theta(\mathfrak{G}^{c_\theta, d_\theta}_{abs}, \beta) & (4.16)\\ &\le (\mathfrak{G}^{c_\theta, d_\theta}_{abs}, \beta)\ . & (4.17)\end{aligned}$$

It remains to show the following. If

$$g(t,c,d) = -\frac{tcd}{c\ln\frac{c(1-e^{-t(c+d)})}{(c+d)(1-e^{-tc})} + d\ln\frac{d(1-e^{-t(c+d)})}{(c+d)(1-e^{-td})}}$$

then $g(t,c,d) = 1 + O(t)$ as $t \to 0+$ and the $O(t)$ term is uniform in $(c,d) \in [\underline{c}, \overline{c}] \times [\underline{d}, \overline{d}]$. Indeed, for every $t > 0$ we have

$$g(t,c,d) = 1 + \frac{c+d}{8}t + t^2 r_2(t,c,d)\ ,$$

where

$$r_2(t,c,d) = \frac{1}{2}\frac{\partial^2 g}{\partial t^2}(\xi, c, d)$$

for some $\xi = \xi(t) \in [0,t]$. Since the derivative $\partial^2 g/\partial t^2$ is uniformly bounded when $(c,d) \in [\underline{c}, \overline{c}] \times [\underline{d}, \overline{d}]$ and $0 \le t \le t_0$, we get the desired result.

The remaining cases can be reduced to the one we have considered. Suppose that $\mathfrak{G}$ has, say, a mixable 0-edge. We will find $b' \in (a,b)$ and $\beta' \in (0,1)$ such that for every θ, if $x_\theta \ge b'$, then $c_\theta(\mathfrak{G}, \beta) = 1$ for every $\beta \in [\beta', 1)$.

It follows from the definition of a mixable 0-edge that there is $b'' \in (a,b)$ such that the game with the canonical representation $f|_{(b'',b)}$ is mixable. Suppose that it is β_1–mixable. Let l be a support hyper-plane to $S \cap \mathbb{R}^2$ passing through $M = (b'', f(b''))$ and let N be the intersection of l with the coordinate axis $x = 0$ (see Fig. 4.4).

Pick $b' \in (b'', b)$. For every $\beta \in [\beta_1, 1)$ there is a shift B of the set B_β such that $S \cap ([b'', +\infty] \times \mathbb{R}) \subseteq B$ and $(b', f(b')) \in \partial B$. For every sufficiently large $\beta \in [\beta_1, 1)$ the segment connecting the points M and N belongs to B.

In order to show this, consider the shift of the curve $\beta^x + \beta^y = 1$ touching a fixed line $y = -sx$, where $s > 0$, at the origin. It is easy to check by direct calculation that the equation of this shift is

$$y(x) = \frac{\ln(1 + t(1-\beta^x))}{\ln\beta}\ .$$

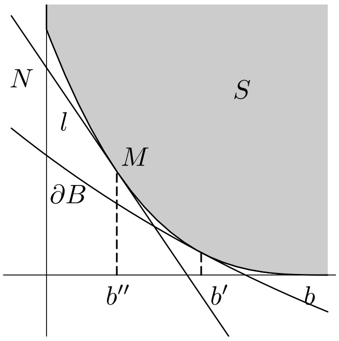

Figure 4.4: The drawing for Theorem 11

For any fixed $s > 0$ and $x \in \mathbb{R}$, the expression $y(x)$ tends to $-sx$ as $\beta \to 1$. Thus for all M' and N' above the line $y = -sx$ the segment $[M', N']$ will lie above the shift for all values of β close to 1.

Pick some $\beta \in (\beta_1, 1)$ such that M and N are inside B. Consider the game with the set of superpredictions

$$S' = (S \cap [b', +\infty] \times \mathbb{R}) \cup (B \cap [0, b'] \times [0, +\infty]) \ \ .$$

It follows from Theorem 4 that this game is mixable, say, β'-mixable. This implies that for every point $P \in \partial(S' \cap \mathbb{R}^2)$ there is a shift B of $B_{\beta'}$ such that $P \in \partial B$ and $S \subseteq S' \subseteq B$. □

4.6.3 Infinite values of $c(\beta)$

If a game does not have bounded edges, $c(\beta)$ does not necessarily behave nicely. The following theorem shows that $c(\beta)$ assumes the value $+\infty$ for a large class of games. Consider a game with the smooth canonical representation $f : (a, +\infty) \to \mathbb{R}$. If the fraction $f''(x)/f'(x)$ is separated from 0 as $x \to +\infty$ then so is the fraction $f''(x)/f'(x)(f'(x) - 1)$ and the game has a mixable 0-edge. The case $f''(x)/f'(x) \to 0$ is completely different.

Theorem 12. *Let $\mathfrak{G}$ be a game with the set of superpredictions S satisfying BIN_1–BIN_5 and an unbounded 0-edge. Let the canonical representation f be twice differentiable on $(a, +\infty)$ for some $a \in \mathbb{R}$. If the following conditions hold*

- $\lim_{x \to +\infty} f(x) = 0$ *and*

- $\lim_{x\to+\infty}\frac{f''(x)}{f'(x)}=0$

then $c(\beta)=+\infty$ *for every* $\beta\in(0,1)$.

Proof. Pick $\beta\in(0,1)$ and consider the image of S under $\mathfrak{B}_\beta$. It is easy to check by evaluating the second derivative (cf. (4.10) and Corollary 4) that the boundary of $\mathfrak{B}_\beta(S)$ in a vicinity of the point $(0,1)$ is the graph of a convex (not concave) function. There exists $u_0>0$ such that the line segment connecting the points $(0,1)$ and $(u_0,\beta^{f(\log_\beta(u_0))})$ lies above the arc of the boundary of $\mathfrak{B}_\beta(S)$ between these points (see Fig. 4.5). The inverse image of the corresponding straight line is a shift $y=b(x)$ of the curve $\beta^x+\beta^y=1$. It is easy to see that $b(x)=\log_\beta(1-\beta^{x+C})$ for some $C\in\mathbb{R}$; we have $0<b(x)<f(x)$ for every $x\in(\log_\beta(u_0),+\infty)$ and $b(x)\to 0$ as $x\to+\infty$. A decomposition yields $b(x)=\left(-\beta^{x+C}+O\left(\beta^{2(x+C)}\right)\right)/\ln\beta$ as $x\to+\infty$.

Now let us pick an arbitrary $\beta_1\in(0,1)$ and show that $c_\theta(\mathfrak{G},\beta_1)\to+\infty$ as $\theta\to 0$ (see 4.14 for the definition of c_θ) . We have proved that there is a function $b_1(x)=\log_\beta(1-\beta^{x+C_1})$, where $C_1\in\mathbb{R}$, and $u_0\in\mathbb{R}$ such that $f(u_0)=b_1(u_0)$ but $0<b_1(x)<f(x)$ for $x>u_0$. It follows from Lemma 3 that the part of subgraph of $y=b_1(x)$ to the left of u_0 belongs to the β_1-closure of S, i.e., $\{(x,y)\in\mathbb{R}^2\mid x\geq u_0 \text{ and } y\geq b_1(x)\}\subseteq \mathrm{cl}_{\beta_1}S$.

Consider some $\beta_2\in(\beta_1,1)$. Clearly, there is $b_2(x)=\log_\beta(1-\beta^{x+C_2})$, where $C_2\in\mathbb{R}$, such that for all sufficiently large values of $x\in\mathbb{R}$ the inequalities

$$0<b_1(x)<b_2(x)<f(x) \tag{4.18}$$

hold. Pick an angle $\theta>0$ and let x_1 and x_2 be such that

$$\frac{b_1(x_1)}{x_1}=\frac{b_2(x_2)}{x_2}=\tan\theta \tag{4.19}$$

(see Fig. 4.6). Suppose that θ is sufficiently small for (4.18) to hold for $x=x_1$ as well as $x=x_2$.

If $x_0\in\mathbb{R}$ is such that $f(x_0)/x_0=\tan\theta$, i.e. $(x_0,f(x_0))$ is the point where the half-line with the gradient θ meets S, then $c_\theta(\mathfrak{G},\beta_1)\geq x_0/x_1\geq x_1/x_2$.

Taking the logarithm of (4.19) yields

$$\begin{aligned}\ln\tan\theta &= (x_1+C_1)\ln\beta_1-\ln\ln(1/\beta_1)+O\left(\beta^{x_1+C_1}\right)-\ln x_1\\ &= x_1\ln\beta_1+o(x_1)\\ &= x_1\ln\beta_1(1+o(1))\end{aligned}$$

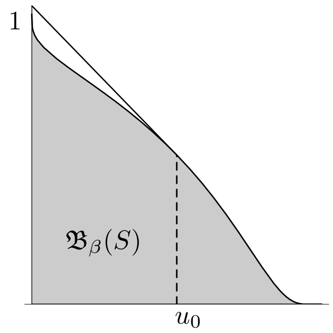

Figure 4.5: The drawing of $\mathfrak{B}_\beta(S)$ for Theorem 12

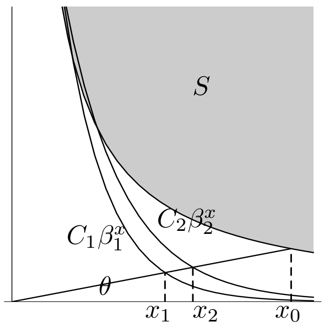

Figure 4.6: The drawing of S, $C_1\beta_1^x$, and $C_1\beta_1^x$ for Theorem 12

as $\theta \to 0$ and, respectively, $x_1 \to +\infty$. Likewise we have

$$\ln \tan \theta = x_2 \ln \beta_2 (1 + o(1))$$

and thus

$$\frac{x_2}{x_1} = \frac{\ln \beta_1}{\ln \beta_2}(1 + o(1))$$

as $\theta \to 0$. Since we can chose β_2 to be as close to 1 as is wished, the theorem follows.

□

A simple example of a game satisfying this theorem is provided by the game with the canonical representation $f(x) = 1/x$, $x \in (0, +\infty)$. Indeed, we have $f'(x) = -1/x^2$ and $f''(x) = 2/x^3$; thus

$$\frac{f''(x)}{f'(x)} = \frac{2}{x} \to 0$$

as $x \to +\infty$. The theorem implies that $c(\beta) \equiv +\infty$ for this game.

The first example of a game with $c(\beta) \equiv +\infty$ was constructed in [Vov98b], Example 6, but the construction was rather artificial and the set of superpredictions of the resulting game was not convex.

4.7 Continuous Games

This section of the chapter stands out because we are considering non-binary games here. We will not investigate them deeply; we just reproduce a theorem from [HKW98] which allows us to reduce the problem of finding $c(\beta)$ for continuous games introduced above to the question concerning corresponding discrete games.

Proposition 8 ([HKW98]). *Take a game* $\mathfrak{G} = \langle [0,1], [0,1], \lambda \rangle$. *Consider* $\mathfrak{G}_{bin} = \langle \mathbb{B}, [0,1], \lambda_{bin} \rangle$, *where* λ_{bin} *is the restriction* $\lambda|_{\mathbb{B}\times[0,1]}$. *Let* $\beta \in (0,1)$ *and* $c(\beta) = c$ *for* $\mathfrak{G}_{bin}$. *Consider a function*

$$g(\omega, \gamma_1, \gamma_2) = \left(\lambda(\omega, \gamma_1) - \frac{\lambda(\omega, \gamma_2)}{c} \right) \ln \beta \ .$$

If

$$\frac{\partial^2 g(\omega, \gamma_1, \gamma_2)}{\partial \omega^2} + \left(\frac{\partial g(\omega, \gamma_1, \gamma_2)}{\partial \omega} \right)^2 \le 0 \qquad (4.20)$$

holds for all $\omega, \gamma_1, \gamma_2 \in [0,1]$, *then* $c(\beta) = c$ *for* $\mathfrak{G}$.

Proof. Take a positive integer k, a set of weights $p_1, p_2, \ldots, p_k \in [0,1]$ such that $\sum_{i=1}^k p_i = 1$, and k predictions $\gamma_1, \gamma_2, \ldots, \gamma_k \in [0,1]$. It follows from the definition of $c(\beta)$ that there is $\gamma \in [0,1]$ such that

$$\begin{cases} \lambda(0,\gamma) & \le c \log_\beta \sum_{i=1}^k p_i \beta^{\lambda(0,\gamma_k)} \\ \lambda(1,\gamma) & \le c \log_\beta \sum_{i=1}^k p_i \beta^{\lambda(1,\gamma_k)} \ . \end{cases} \qquad (4.21)$$

Now consider the inequality

$$\lambda(\omega, \gamma) \le c \log_\beta \sum_{i=1}^k p_i \beta^{\lambda(\omega, \gamma_k)} \ . \qquad (4.22)$$

It is equivalent to

$$\beta^{-\frac{\lambda(\omega,\gamma)}{c}} \sum_{i=1}^k p_i \beta^{\lambda(\omega,\gamma_k)} \le 1$$

and therefore to

$$\sum_{i=1}^k p_i e^{g(\omega, \gamma_i, \gamma)} \le 1 \ . \qquad (4.23)$$

Consider the left-hand side as a function of ω. Equation 4.20 implies that it is convex in ω and (4.21) means that (4.23) holds for $\omega = 0, 1$. This proves that (4.23 holds for every $\omega \in [0, 1]$ and thus (4.22) holds for every $\omega \in [0, 1]$. □

Note the constructive nature of the proof (cf. [Vov98a]). Suppose we have a formula or an algorithm which constructs γ given experts' predictions $\gamma_1, \gamma_2, \ldots, \gamma_k$ and weights $p_1, p_2, \ldots, p_k$ and ensures that (4.21) is true for a discrete game (a *substitution rule* from [Vov98a]). The proof of the proposition implies that the same rule may be used for the corresponding continuous game provided the conditions of the proposition are satisfied.

Chapter 5

Predictive Complexity: Definitions and Existence

We have seen that the Aggregating Algorithm is a powerful tool for merging different pools of prediction strategies. The question arises of whether it is possible to merge *all* strategies. Indeed, there are countably many algorithms working according to Protocol 1. If we could merge them all we would end up with a universal prediction method. Similar ideas (in a different context) were proposed by Solomonoff back in sixties; see [LV97] for an overview of Solomonoff's research.

Unfortunately, the straightforward approach fails for most games. It follows from simple diagonalization-style considerations, every predicting algorithm is significantly outperformed by some other algorithm on some strings. Indeed, consider, say, the discrete square-loss game and a prediction algorithm $\mathfrak{A}$. Let us construct an infinite sequence $\boldsymbol{x} = (x_1, x_2, \ldots) \in \mathbb{B}^\infty$ and a strategy $\mathfrak{A}_0$ performing better then $\mathfrak{A}$ on finite prefixes of $\boldsymbol{x}$. On the first trial $\mathfrak{A}$ suffers loss of at least $1/4$ on at least one of the outcomes 0 or 1. We can effectively find such an outcome (if both 0 and 1 qualify, let us take 0 for definiteness sake) and choose x_1 being equal to this outcome. The same procedure can be repeated over consecutive trials and thus $\boldsymbol{x}$ is constructed by induction. The loss suffered by $\mathfrak{A}$ on prefixes of $\boldsymbol{x}$ satisfies the inequality

$$\mathrm{Loss}^{\mathrm{sq}}_{\mathfrak{A}}(x_1, x_2, \ldots, x_n) \geq \frac{n}{4}$$

for every positive integer n. On the other hand, since the sequence $\boldsymbol{x}$ is computable, we may consider $\mathfrak{A}_0$ which predicts x_k on trial k[1]. Thus

$$\mathrm{Loss}^{\mathrm{sq}}_{\mathfrak{A}_0}(x_1, x_2, \ldots, x_n) = 0 \ .$$

This problem can be overcome at a certain cost. We widen the set of all computable strategies to the class of 'superstrategies'. For many important games, this larger class exhibits nicer properties and contains an element universal in some natural sense. In this chapter we define these superstrategies, introduce several notions of universality and investigate under which conditions universal superstrategies exist.

We restrict ourselves to binary games although many definitions and constructions can be easily extended to the non-binary case.

5.1 Weaker Regularity Assumptions

It is natural to develop some parts of the theory of predictive complexity under assumptions that are weaker than those for the theory of the AA. Sometimes the conditions REG_1 and REG_5 or their equivalents BIN_1 and BIN_2 can be relaxed. While it makes no sense to consider negative losses in the theory of prediction with expert advice, they may be acceptable in the theory of predictive complexity.

Consider the following condition:

BIN_1' There are $a, b \in \mathbb{R}$ such that $S \subseteq [a, +\infty] \times [b, +\infty]$.

This requirement is equivalent to allowing a loss function λ to assume values from $[r, +\infty]$, where r is some finite number (perhaps negative). In some situations we will consider BIN_1' instead of BIN_1 and BIN_2.

5.2 Loss and Superloss Processes

Let $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$ be a game with the set of superpredictions S satisfying BIN_1' and BIN_3–BIN_4. We will speak about finite sequences of outcomes, i.e., finite strings of binary digits, and denote them by bold lowercase letters, e.g., $\boldsymbol{x}$, $\boldsymbol{y}$, or $\boldsymbol{z}$.

[1] The algorithm $\mathfrak{A}_0$ is oblivious; the outcomes do not affect its predictions.

A function $L : \mathbb{B}^* \to (-\infty, +\infty]$ is a *loss process* if there is an algorithm $\mathfrak{A}$ working according to Protocol 1 such that $L = \mathrm{Loss}^{\lambda}_{\mathfrak{A}}$. In other words, the set of all loss processes coincides with the set of losses of algorithms. Note that a loss process is computable.

The discussion at the beginning of this chapter explains the need to extend this class. A function $L : \mathbb{B}^* \to (-\infty, +\infty]$ is a *superloss process* if

- $L(\Lambda) = 0$, where Λ is the empty string,
- L is semi-computable from above, and
- for every $\boldsymbol{x} \in \mathbb{B}^*$ there is $\gamma \in \Gamma$ such that

$$\begin{cases} L(\boldsymbol{x}0) - L(\boldsymbol{x}) \geq \lambda(\gamma, 0) \ , \\ L(\boldsymbol{x}1) - L(\boldsymbol{x}) \geq \lambda(\gamma, 1) \ . \end{cases} \tag{5.1}$$

The last condition means that the pair $(L(\boldsymbol{x}0) - L(\boldsymbol{x}), L(\boldsymbol{x}1) - L(\boldsymbol{x}))$ is a superprediction.

Note that every loss process is a superloss process. Indeed, if L is a loss process, it satisfies (5.1); moreover the inequalities may be replaced by equalities. One may use the prediction output by $\mathfrak{A}$ on input $\boldsymbol{x}$ as γ for (5.1).

One of the key properties of superloss processes is the following.

Proposition 9 ([VW98]). *Let $\mathfrak{G} = \langle \mathbb{B}, [0,1], \lambda \rangle$ be a game with the set of superpredictions S satisfying BIN_1' and BIN_3–BIN_4. Then the set of all superloss processes for $\mathfrak{G}$ is enumerable, i.e., there is a computable sequence $L_1, L_2, \ldots$ of superloss processes w.r.t. $\mathfrak{G}$ which contains all of them.*

Proof. We give the proof from [VV01].

The computability of $\lambda(\omega, \gamma)$ implies that there is a computable sequence of functions $\lambda^t : [0,1] \cap \mathbb{Q} \to \mathbb{Q} \cup \{+\infty\}$ $(t = 1, 2, \ldots)$ decreasing in t (i.e., $\lambda^{t+1}(\omega, \gamma) \leq \lambda^t(\omega, \gamma)$ for every positive integer t, each $\omega \in \mathbb{B}$, and each $\gamma \in [0,1] \cap \mathbb{Q}$) and converges to λ (i.e., $\lim_{t \to +\infty} \lambda^t(\omega, \gamma) = \lambda(\omega, \gamma)$ for every $\omega \in \mathbb{B}$ and $\gamma \in [0,1] \cap \mathbb{Q}$).

It follows from the existence of the universal enumerable set that there is an enumerable set $W \subseteq \mathbb{N} \times \mathbb{B}^* \times (\mathbb{Q} \cup \{+\infty\})$ such that the set of its sections $W_i = \{(\boldsymbol{x}, q) \mid (i, \boldsymbol{x}, q) \in W\}$, $i = 1, 2, \ldots$, coincides with the set of all enumerable subsets of $\mathbb{B}^* \times (\mathbb{Q} \cup \{+\infty\})$. Let W^t be a finite subset of W enumerable in t steps and $W_i^t = \{(\boldsymbol{x}, q) \mid \exists q' \leq q : (i, \boldsymbol{x}, q') \in W^t\}$.

Let us define a computable set of functions $L_i^t : \mathbb{B}^* \to (\mathbb{Q} \cup \{+\infty\})$ (i, t are positive integers) decreasing in t ($L_i^{t+1}(\boldsymbol{x}) \leq L_i^t(\boldsymbol{x})$ for every positive integers i, t and every $\boldsymbol{x} \in \mathbb{B}^*$). Suppose that $L_i^1, L_i^2, \ldots, L_i^{t-1}$ has already been constructed. Consider all functions L such that

- the graph of L is a subset of W_i^t,
- for every $\boldsymbol{x} \in \mathbb{B}^*$ there is rational $\gamma \in [0, 1]$ such that

$$\begin{cases} L(\boldsymbol{x}0) - L(\boldsymbol{x}) \geq \lambda^t(\gamma, 0) \ , \\ L(\boldsymbol{x}1) - L(\boldsymbol{x}) \geq \lambda^t(\gamma, 1) \ , \end{cases}$$

  and

- $L^t(\boldsymbol{x}) \leq L_i^{t-1}(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathbb{B}^*$.

Let L_i^t be a minimal one of these functions, i.e., let L_i^t be such that there is no L satisfying the conditions and $\boldsymbol{x} \in \mathbb{B}^*$ such that $L(\boldsymbol{x}) < L_i^t(\boldsymbol{x})$ (if there are many such L we just take the first in an enumeration).

Now it is possible to define a computable sequence of functions $L_i : \mathbb{B}^* \to (\mathbb{Q} \cup \{+\infty\})$ ($i = 1, 2, \ldots$). Put $L_i(\boldsymbol{x}) = \inf_{t \in \mathbb{N}} L_i^t(\boldsymbol{x})$. Every L_i is a superloss process. On the other hand, if L is a measure of predictive complexity, its subgraph coincides with W_i for some i. One can check that L coincides with L_i. □

5.3 Simple Predictive Complexity

In this section we introduce the most important form of predictive complexity. We will refer to it as to *simple predictive complexity* or just *predictive complexity.*

5.3.1 Definition

Definition 4 ([VW98]). A superloss process $\mathcal{K}$ w.r.t. a game $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ is called *(simple) predictive complexity* w.r.t. $\mathfrak{G}$ if for every other superloss process L w.r.t. $\mathfrak{G}$ there is a constant C such that the inequality

$$\mathcal{K}(\boldsymbol{x}) \leq L(\boldsymbol{x}) + C \qquad (5.2)$$

holds for every $\boldsymbol{x} \in \Omega^*$.

We will use the notation $\mathcal{K}^{\mathfrak{G}}$ for complexity w.r.t. a game $\mathfrak{G}$ if it is not clear from the context which game we are referring to.

Note that if there is a superloss process satisfying Definition 4, it is not unique. However the absolute value of the difference between two such superloss processes is uniformly bounded by a constant.

A similar situation occurs in the case of Kolmogorov complexity. There are many universal functions and complexities any two of them define differ by a constant. There are two approaches that can be taken. The first one is to fix a particular universal function and to define complexity using that particular universal function. Another one is to thing of Kolmogorov complexity as defined up to an additive constant.

The first approach involves specifying a particular universal function and usually raises the issue of the complexity of a universal function. However within that approach we can actually speak of the complexity of an individual object. The theory built using the second approach is necessarily asymptotic. The value of complexity of a particular sequence $\boldsymbol{x}$ makes little sense and only the relations among the values of complexity for infinitely many strings may be the subject of a non-trivial mathematical investigation. This is the price we pay for generality and universality.

In this thesis we are following the second approach. We say that predictive complexity is *a* universal superloss process. We do not bother constructing a particular universal superloss process; instead, we are formulating our results in the asymptotic fashion.

Since every loss process is a superloss process, (5.2) hold for loss processes as well. In other words, for every prediction strategy $\mathfrak{A}$ there is a constant C such that for every $\boldsymbol{x} \in \Omega^*$ the inequality

$$\mathcal{K}^{\mathfrak{G}}(\boldsymbol{x}) \leq \mathrm{Loss}^{\lambda}_{\mathfrak{A}}(\boldsymbol{x}) + C \tag{5.3}$$

holds. Thus $\mathcal{K}^{G}$ bounds the loss w.r.t. $\mathfrak{G}$ of every prediction strategy from below aand this bound is tight in a certain sense. We may say that predictive complexity is an intrinsic measure of predictability of a string in the same way as Kolmogorov complexity is an intrinsic measure of complexity of a string independent of a particular description method.

5.3.2 Existence

The problem of the existence of predictive complexity has not been fully resolved yet. We will formulate some sufficient and necessary conditions below and in subsequent sections, but our results do not form a complete solution.

The theory of prediction with expert advice provides a sufficient condition for the existence of predictive complexity.

Proposition 10 ([VW98]). *If a game $\mathfrak{G}$ with the set of superpredictions S satisfying BIN_1–BIN_4 is mixable, then there is simple predictive complexity w.r.t. $\mathfrak{G}$.*

Proof. Suppose that $\mathfrak{G}$ is β-mixable, where $\beta \in (0,1)$. By Proposition 9, there is an enumerable sequence $L_1, L_2, \ldots$ of all superloss processes w.r.t. $\mathfrak{G}$. Pick a positive sequence $q_1, q_2, \ldots \geq 0$ such that $\sum_{i=1}^{+\infty} q_i = 1$ (e.g., $q_i = 1/2^i$). Let us define the function $\mathcal{K} : \mathbb{B}^* \to \mathbb{R}$ by the following formula:

$$\mathcal{K}(\boldsymbol{x}) = \log_\beta \sum_{i=1}^{+\infty} q_i \beta^{L_i(\boldsymbol{x})} \quad .$$

First we will show that $\mathcal{K}$ is a superloss process. Indeed,

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

where $\omega = 0, 1$. The point $\mathfrak{B}_\beta(\mathcal{K}(\boldsymbol{x}0) - \mathcal{K}(\boldsymbol{x}), \mathcal{K}(\boldsymbol{x}1) - \mathcal{K}(\boldsymbol{x}))$ belongs to $\mathfrak{B}_\beta(S)$ since a convergent convex mixture of a countable number of points from a convex set in $\mathbb{R}^2$ belongs to the set.

Secondly since $\log_\beta t$ is a decreasing function of t, for every positive integer i we have

$$\begin{aligned}\mathcal{K}(\boldsymbol{x}) &\leq \log_\beta q_i \beta^{L_i(\boldsymbol{x})} \\ &= L_i(\boldsymbol{x}) + \log_\beta q_i \quad .\end{aligned}$$

□

5.3.3 Some Specific Complexities

It follows from Proposition 10 and Theorem 8 that there is simple predictive complexity for some games we have introduced.

The discrete square-loss game is mixable and thus simple predictive complexity w.r.t. this game exists. We will denote it by $\mathcal{K}^{\mathrm{sq}}$.

Another mixable game we have introduced is the logarithmic-loss game. It defines logarithmic-loss complexity $\mathcal{K}^{\log}$. It is remarkable that this function has been known for mathematicians for some time. It is easy to see (just by comparing the definitions) that it coincides with a variant of Kolmogorov complexity, namely, the negative logarithm of Levin's *a priori* semimeasure or KM, which is described in [V'y94, LV97].

Let us perform the derivation. The definition of Levin's *a priori* semimeasure is as follows. A function $Q : \mathbb{B}^* \to [0, 1]$ is an *(enumerable continuous) semimeasure* if

- $Q(\Lambda) \le 1$, where Λ is the empty string,
- for every $\boldsymbol{x} \in \mathbb{B}^*$, we have $Q(x0) + Q(x1) \le Q(x)$ and
- Q is enumerable from below.

It is shown in [V'y94, LV97] that there is a semimeasure $\mathbf{M}$ such that for any other semimeasure Q there is a constant $c > 0$ such that $cQ(\boldsymbol{x}) \le \mathbf{M}(\boldsymbol{x})$ holds for all $\boldsymbol{x} \in \mathbb{B}^*$. By definition, put $\mathrm{KM} = -\log_2 \mathbf{M}$.

Now for each semimeasure Q consider the function $L_Q = -\log Q$. The class of all L_Q, as Q ranges over the class of all semimeasures, is the class of all functions $L : \mathbb{B}^* \to [0, +\infty]$ such that

- $L(\Lambda) \ge 0$, where Λ is the empty string,
- for every $\boldsymbol{x} \in \mathbb{B}^*$, we have $2^{-L(x0)} + 2^{-L(x1)} \le 2^{-L(x)}$ and
- Q is enumerable from above.

Clearly, KM has the following property. For every L there is a constant C such that $\mathrm{KM}(\boldsymbol{x}) \le L(\boldsymbol{x}) + C$.

Now consider the definition of a superloss process w.r.t. the logarithmic-loss game. A superloss process L is semicomputable from above and for every $\boldsymbol{x} \in \mathbb{B}^*$ there is $\gamma \in [0, 1]$ such that

$$\begin{cases} L(\boldsymbol{x}0) - L(\boldsymbol{x}) \ge -\log(1-\gamma) \ , \\ L(\boldsymbol{x}1) - L(\boldsymbol{x}) \ge -\log\gamma \ . \end{cases}$$

Excluding γ from the system provides us with with the condition $2^{-L(x0)} + 2^{-L(x1)} \leq 2^{-L(x)}$. Thus KM equals $\mathcal{K}^{\log}$ up to an additive constant.

Note that KM differs from other variants of Kolmogorov complexity, namely, plain complexity K, prefix complexity KP, and monotone complexity Km, by terms bounded by $C \log |\boldsymbol{x}|$, where C is a constant.

5.4 Weaker Predictive Complexities

Definition 4 can be weakened in many ways. The reasons for considering weaker complexities are as follows. First they help to clarify the problem of the existence of predictive complexity. Proposition 10 shows that mixability is a sufficient condition for the existence of predictive complexity. There are reasons to believe that it is also necessary (e.g., the role of mixability in prediction with expert advice and the optimality of the Aggregating Algorithm) though the problem remains open. Weaker versions of predictive complexity provide an approach to solving this problem. Secondly weaker complexities are of independent interest. When we cannot construct simple complexity, we sometimes can still use weaker versions.

Consider a game $\mathfrak{G}$ and superloss processes w.r.t. $\mathfrak{G}$. We will weaken Definition 4 by introducing extra non-constant terms. The easiest way to do it is to consider terms depending on the length of strings $\boldsymbol{x}$.

A superloss process $\mathcal{K}$ is predictive complexity up to $f(n)$, where $f : \mathbb{N} \to [0, +\infty)$, if for every other superloss process L there is a constant C such that the inequality $\mathcal{K}(\boldsymbol{x}) \leq L(\boldsymbol{x}) + Cf(|\boldsymbol{x}|)$ holds for every $\boldsymbol{x} \in \mathbb{B}^*$.

Assumption BIN_4 implies that there is a process L such that $L(\boldsymbol{x}) = O(n)$ as $n \to +\infty$ and thus the definition becomes most interesting in the case $f(n) = o(n)$ as $n \to +\infty$.

This consideration motivates the definition of predictive complexity up to $o(n)$. A superloss process $\mathcal{K}$ satisfies this definition if for every L we have $\mathcal{K}(\boldsymbol{x}) \leq L(\boldsymbol{x}) + o(|\boldsymbol{x}|)$. The term $o(|\boldsymbol{x}|)$ does not have to be uniform in L.

Another approach is to consider functions of $\mathcal{K}$ as extra terms. Definition of complexities up to $f(\mathcal{K})$ or $o(\mathcal{K})$ may be given in much the same way as the definitions up to $f(n)$ or $o(n)$. However we will not discuss these variants of complexity since few facts are known about them.

An interesting variant of complexity is complexity up to $M \cdot \mathcal{K}$. We say that a superloss process $\mathcal{K}$ is complexity w.r.t. $\mathfrak{G}$ up to $M \cdot \mathcal{K}$, where M is a positive number, if for every superloss process L there is a constant C such that the inequality

$$\mathcal{K}(\boldsymbol{x}) \leq L(\boldsymbol{x}) + M\mathcal{K}(\boldsymbol{x}) + C \tag{5.4}$$

holds for every $\boldsymbol{x} \in \mathbb{B}^*$. This definition makes sense only if $M \in (0,1)$. If this is the case, (5.4) can be rewritten as

$$\mathcal{K}(\boldsymbol{x}) \leq \frac{1}{1-M}L(\boldsymbol{x}) + \frac{C}{1-M} \tag{5.5}$$

or, since C is arbitrary, as

$$\mathcal{K}(\boldsymbol{x}) \leq \frac{1}{1-M}L(\boldsymbol{x}) + C \ . \tag{5.6}$$

One can easily see that the coefficient $1/(1-M)$ is greater than or equal to 1; it tends to 1 as M tends to 0 and tends to $+\infty$ as M tends to 1.

We can define complexity up to the multiplicative constant m ($m \geq 1$) by the requirement that for every superloss process L there is a constant C such that $\mathcal{K}(\boldsymbol{x}) \leq mL(\boldsymbol{x}) + C$ for every $\boldsymbol{x} \in \mathbb{B}^*$. Clearly, $\mathcal{K}$ is complexity up to $M \cdot \mathcal{K}$ if and only if it is complexity up to the multiplicative constant $1/(1-M)$.

5.5 Complexities and Shifts

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 $S_2 = S_1 + (u,v)$, where u, v are some real constants. Any $L : \mathbb{B}^* \to (0,+\infty]$ is a superloss process w.r.t. $\mathfrak{G}$ if and only if $L(\boldsymbol{x}) + u\sharp_0\boldsymbol{x} + v\sharp_1\boldsymbol{x}$ is a superloss process w.r.t. $\mathfrak{G}_2$. The notation $\sharp_0\boldsymbol{x}$ stands for the number of 0s in a string $\boldsymbol{x} \in \mathbb{B}^*$; the notation $\sharp_1\boldsymbol{x}$ stands for the number of 1s, respectively. Therefore there is simple predictive complexity w.r.t. $\mathfrak{G}_1$ if and only if there is simple predictive complexity w.r.t. $\mathfrak{G}_2$ and $\mathcal{K}^{\mathfrak{G}_1}$ is simple predictive complexity w.r.t. $\mathfrak{G}_1$ if and only if $\mathcal{K}^{\mathfrak{G}_1}(\boldsymbol{x}) + u\sharp_0\boldsymbol{x} + v\sharp_1\boldsymbol{x}$ is simple predictive complexity w.r.t. $\mathfrak{G}_2$. The same applies to complexities w.r.t. $f(n)$ and $o(n)$.

5.6 On the Existence of Weak Complexity

Let us start with a simple elaboration of Proposition 10.

Theorem 13. *Let $\mathfrak{G}$ be a game with the set of superpredictions S satisfying BIN_1–BIN_4. If $c = c(\beta) < +\infty$ for some β, then there is predictive complexity up to the multiplicative constant c w.r.t. $\mathfrak{G}$.*

Proof. Consider $\mathcal{K}$ defined by the formula

$$\mathcal{K}(\boldsymbol{x}) = c \log_\beta \sum_{i=1}^{+\infty} q_i \beta^{L_i(\boldsymbol{x})} \enspace ,$$

where L_i is an enumeration of all superloss processes and $\sum_{i=1}^{+\infty} q_i = 1$.

It is a superloss process since

$$\begin{aligned}\mathcal{K}(\boldsymbol{x}\omega) - \mathcal{K}(\boldsymbol{x}) &= c \left(\log_\beta \sum_{i=1}^{\infty} q_i \beta^{L_i(\boldsymbol{x}\omega)} - \log_\beta \sum_{i=1}^{\infty} q_i \beta^{L_i(\boldsymbol{x})} \right) \\ &= c \log_\beta \sum_{i=1}^{\infty} \frac{\beta^{L_i(\boldsymbol{x})}}{\frac{1}{q_i} \sum_{j=1}^{\infty} q_j \beta^{L_i(\boldsymbol{x})}} \beta^{L(\boldsymbol{x}\omega) - L(\boldsymbol{x})} \enspace ,\end{aligned}$$

for the both values $\omega = 0, 1$, and thus $(\mathcal{K}(\boldsymbol{x}0) - \mathcal{K}(\boldsymbol{x}), \mathcal{K}(\boldsymbol{x}1) - \mathcal{K}(\boldsymbol{x}))$ belongs to S.

For every positive integer i we have

$$\begin{aligned}\mathcal{K}(\boldsymbol{x}) &\leq c \log_\beta q_i \beta^{L_i(\boldsymbol{x})} \\ &= cL_i(\boldsymbol{x}) + \log_\beta q_i \enspace .\end{aligned}$$

□

If the set of superpredictions is not convex, we cannot construct an essentially stronger complexity. This is implied by the following negative result.

Theorem 14. *Let $\mathfrak{G}$ be a game with the set of superpredictions S satisfying BIN_1–BIN_4. If $S \cap \mathbb{R}^2$ is not convex, then there is no predictive complexity up to $o(n)$ 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 not convex but there exists complexity $\mathcal{K}$ w.r.t. $\mathfrak{G}$.

There exist points $B_0, B_1 \in S$ such that the segment connecting B_0 and B_1 is not a subset of S. Without loss of generality (cf. Sect. 5.5) we may assume that $B_0 = (b_0, 0)$ and $B_1 = (0, b_1)$ (see Fig. 5.1).

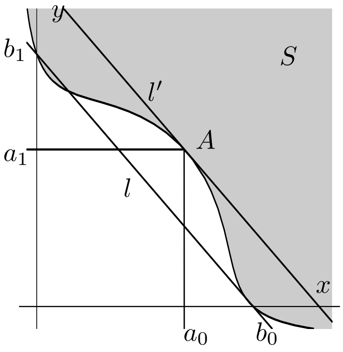

Figure 5.1: The drawing for Theorem 14. The set S is shaded

Let us denote the line containing this segment by l and let us assume that it has the equation $\alpha_0 x + \alpha_1 y = \rho$, where $\alpha_0, \alpha_1, \rho > 0$. There exists a point $A = (a_0, a_1) \in \partial S$, where $a_0, a_1 > 0$, that lies above the straight line l, i.e., $\alpha_0 a_0 + \alpha_1 a_1 = \rho + \delta > \rho$.

Since $b_0 \sharp_0 \boldsymbol{x}$ and $b_1 \sharp_1 \boldsymbol{x}$ are superloss processes, the inequalities

$$\mathcal{K}(\boldsymbol{x}) \leq b_0 \sharp_0 \boldsymbol{x} + f(|\boldsymbol{x}|) \tag{5.7}$$

$$\mathcal{K}(\boldsymbol{x}) \leq b_1 \sharp_1 \boldsymbol{x} + f(|\boldsymbol{x}|) \tag{5.8}$$

hold for every $\boldsymbol{x} \in \mathbb{B}^*$, where $f(n) = o(n)$ as $n \to +\infty$. At the same time, there is a sequence of strings $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ such that for any $n \in \mathbb{N}$ we have $|\boldsymbol{x}_n| = n$ and

$$\mathcal{K}(\boldsymbol{x}_n) \geq a_0 \sharp_0 \boldsymbol{x}_n + a_1 \sharp_1 \boldsymbol{x}_n \ . \tag{5.9}$$

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

$$\mathcal{K}(\boldsymbol{x}_n 0) - \mathcal{K}(\boldsymbol{x}_n) \geq a_0 \tag{5.10}$$

$$\mathcal{K}(\boldsymbol{x}_n 1) - \mathcal{K}(\boldsymbol{x}_n) \geq a_1 \tag{5.11}$$

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

Combining (5.7), (5.8) and (5.9), we get

$$a_0 \sharp_0 \boldsymbol{x}_n + a_1 \sharp_1 \boldsymbol{x}_n \leq b_0 \sharp_0 \boldsymbol{x}_n + f(n) \tag{5.12}$$

$$a_0 \sharp_0 \boldsymbol{x}_n + a_1 \sharp_1 \boldsymbol{x}_n \leq b_1 \sharp_1 \boldsymbol{x}_n + f(n) \tag{5.13}$$

as $n \to +\infty$. If we multiply (5.12) by α_0/a_1, (5.13) by α_1/a_0 and then add them together, we obtain

$$\frac{\delta}{a_1}\sharp_0 \boldsymbol{x}_n + \frac{\delta}{a_0}\sharp_1 \boldsymbol{x}_n \leq f(n)$$

as $n \to +\infty$. This is a contradiction since at least one of the values $\sharp_0 \boldsymbol{x}_n$ and $\sharp_1 \boldsymbol{x}_n$ is greater than or equal to $n/2$ for infinitely many $n \in \mathbb{N}$. □

An example of a game with the non-convex set of superpredictions is provided by the simple prediction game. It has the set of superpredictions $S = \{(x, y) \in [0, +\infty]^2 \mid x \geq 1 \text{ or } y \geq 1\}$.

It is easy to see that, if $c(\beta) \to 1$ as $\beta \to 1$, then for every $\varepsilon > 0$ there is complexity up to the multiplicative constant $1 + \varepsilon$. However the convergence of $c(\beta)$ to 1 implies the existence of stronger types of complexity. The following theorem generalises a result from [V'y02]; the idea of varying the values of β in the proof goes back to [CBFH$^+$97].

Theorem 15. *Let $\mathfrak{G}$ be a game with the set of superpredictions satisfying BIN$_1$–BIN$_4$. If*

- *$\mathfrak{G}$ is bounded and*
- *$c(\beta) \to 1$ for $\mathfrak{G}$ as $\beta \to 1$*

then there exists predictive complexity w.r.t. $\mathfrak{G}$ up to $f(n)$, where

$$\begin{aligned} f(n) &= \sqrt{n} + \sum_{k=1}^{n} \left(1 - \frac{1}{c(e^{-1/\sqrt{k}})}\right) && (5.14) \\ &= o(n) \text{ as } n \to +\infty \ . && (5.15) \end{aligned}$$

Proof. The proof is similar to a proof from [V'y02].

We may assume that $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$ and there is $l > 0$ such that $\lambda(\omega, \gamma) \leq l$ for every $\omega \in \Omega$ and $\gamma \in \Gamma$.

Put $\beta_n = e^{-1/\sqrt{n}}$, $n = 1, 2, \ldots$. Let L_i, $i = 1, 2, \ldots$, be an effective enumeration of all superloss processes w.r.t. $\mathfrak{G}$. Let the sequence L_i^* be defined by

$$L_i^*(\boldsymbol{x}) = L_i(\boldsymbol{x}) + l \sum_{k=1}^{|\boldsymbol{x}|-1} \left(1 - \frac{1}{c(\beta_n)}\right) \ ,$$

for every $\boldsymbol{x} \in \mathbb{B}^*$. Pick a sequence $p_i > 0$, $i = 1, 2, \ldots$, such that $\sum_{i=1}^{+\infty} p_i = 1$ and consider the function defined by

$$\mathcal{K}(\boldsymbol{x}) = \log_{\beta_n} \sum_{i=1}^{+\infty} p_i \beta_n^{L_i^*(\boldsymbol{x})}$$

for every $\boldsymbol{x}$ of length n, $n = 1, 2, \ldots$.

Let us check that $\mathcal{K}$ is a superloss process w.r.t. $\mathfrak{G}$.

Fix some arbitrary $n \in \mathbb{N}$ and $\boldsymbol{x}$ of length n. We have

$$\beta_n^{\mathcal{K}(\boldsymbol{x})} = \sum_{i=1}^{+\infty} p_i \beta_n^{L_i^*(\boldsymbol{x})} \tag{5.16}$$

and, for each $\omega = 0, 1$,

$$\beta_{n+1}^{\mathcal{K}(\boldsymbol{x}\omega)} = \sum_{i=1}^{+\infty} p_i \beta_{n+1}^{L_i^*(\boldsymbol{x})} \quad . \tag{5.17}$$

We cannot manipulate with these formulae because they include different bases β_n and β_{n+1}. This obstacle can be overcome with the following trick. Since the function $y = x^a$ is convex on $[0, +\infty)$ for every $a \geq 1$, the inequality $\left(\sum_i p_i t_i\right)^a \leq \sum_i p_i t_i^a$ holds for every $t_1, t_2, \ldots \geq 0$ and every sequence $p_1, p_2, \ldots \geq 0$ which sums up to 1. This implies

$$\beta_n^{\mathcal{K}(\boldsymbol{x}\omega)} = \left(\beta_{n+1}^{\mathcal{K}(\boldsymbol{x}\omega)}\right)^{\log_{\beta_{n+1}} \beta_n} \tag{5.18}$$

$$\leq \sum_{i=1}^{+\infty} p_i \left(\beta_{n+1}^{L_i^*(\boldsymbol{x})}\right)^{\log_{\beta_{n+1}} \beta_n} \tag{5.19}$$

$$= \sum_{i=1}^{+\infty} p_i \beta_n^{L_i^*(\boldsymbol{x}\omega)} \tag{5.20}$$

for each $\omega = 0, 1$. Combining (5.16) and (5.20), we obtain

$$\beta_n^{\mathcal{K}(\boldsymbol{x}\omega) - \mathcal{K}(\boldsymbol{x})} \leq \sum_{i=1}^{+\infty} q_i \beta_n^{L_i^*(\boldsymbol{x}\omega) - L_i^*(\boldsymbol{x})} \tag{5.21}$$

for each $\omega = 0, 1$, where

$$q_i = \frac{p_i \beta_n^{L_i^*(\boldsymbol{x})}}{\sum_{k=1}^{+\infty} p_k \beta_n^{L_k^*(\boldsymbol{x})}} \quad .$$

Since all L_i are superloss processes, these exists a sequence of superpredictions $(s_1^{(0)}, s_1^{(1)}), (s_2^{(0)}, s_2^{(1)}), \ldots \in S$ such that for every $i = 1, 2, \ldots$ and each $\omega = 0, 1$ we have

$$L_i^*(\boldsymbol{x}\omega) - L_i^*(\boldsymbol{x}) = L_i(\boldsymbol{x}\omega) - L_i(\boldsymbol{x}) + l\left(1 - \frac{1}{c(\beta_n)}\right) \tag{5.22}$$

$$\geq s_i^{(\omega)} + l\left(1 - \frac{1}{c(\beta_n)}\right) \quad . \tag{5.23}$$

It can be assumed that $s_i^{(\omega)} \leq l$ for all i and ω.

If we combine (5.21) with (5.23) and take into account the definition of $c(\beta)$, we obtain

$$\beta_n^{\mathcal{K}(\boldsymbol{x}\omega) - \mathcal{K}(\boldsymbol{x})} \leq \beta_n^{\left(1 - \frac{1}{c(\beta_n)}\right)} \sum_{i=1}^{+\infty} q_i \beta_n^{s_i^{(\omega)}} \tag{5.24}$$

$$\leq \beta_n^{\left(1 - \frac{1}{c(\beta_n)}\right) + \frac{s^{(\omega)}}{c(\beta_n)}} \tag{5.25}$$

$$= \beta_n^{s^{(\omega)} + (l - s^{(\omega)})\left(1 - \frac{1}{c(\beta_n)}\right)} \tag{5.26}$$

$$\leq \beta^{s^{(\omega)}} \tag{5.27}$$

for each $\omega = 0, 1$, where $(s^{(0)}, s^{(1)})$ is a superprediction. This implies the inequalities

$$\mathcal{K}(\boldsymbol{x}0) - \mathcal{K}(\boldsymbol{x}) \geq s^{(0)} \quad ,$$
$$\mathcal{K}(\boldsymbol{x}1) - \mathcal{K}(\boldsymbol{x}) \geq s^{(1)} \quad .$$

Thus $\mathcal{K}$ is a superloss process.

It follows from our definition of $\mathcal{K}$ that for every $\boldsymbol{x}$ of length n and every $i = 1, 2, \ldots$ we get

$$\mathcal{K}(\boldsymbol{x}) \leq L_i^*(\boldsymbol{x}) + \log_{\beta_n} p_i \tag{5.28}$$

$$= L_i(\boldsymbol{x}) + \sum_{k=1}^{n-1}\left(1 - \frac{1}{c(\beta_k)}\right) + \sqrt{n} \ln \frac{1}{p_i} \quad . \tag{5.29}$$

A simple lemma from calculus completes the proof.

Lemma 6. *Let $a_n \geq 0$, $n = 1, 2, \ldots$ be such that $a_n = o(1)$ as $n \to +\infty$. Then $\sum_{k=1}^{n} a_k = o(n)$ as $n \to +\infty$.*

□

Now Theorems 10 and 11 can be used to establish the existence of predictive complexity.

Corollary 7. *Let $\mathfrak{G}$ be a game with a set of superpredictions S satisfying BIN_1–BIN_5. If $\mathfrak{G}$ is bounded and $S \cap \mathbb{R}^2$ is convex, there is $f : \mathbb{N} \to \mathbb{R}$ such that $f(n) = o(n)$ as $n \to +\infty$ and there is complexity up to $f(n)$ w.r.t. $\mathfrak{G}$.*

Corollary 8. *Let $\mathfrak{G}$ be a bounded game with the set of superpredictions S satisfying BIN_1–BIN_5. Let $f : (a, b) \to \mathbb{R}$ be its canonical representation. Suppose that $S \cap \mathbb{R}^2$ is convex and the following conditions hold:*

- *$\mathfrak{G}$ has a mixable 0-edge or there is $\varepsilon > 0$ such that $f'_-(x) \leq -\varepsilon$ holds on (a, b), and*
- *$\mathfrak{G}$ has a mixable 1-edge or there is $T < +\infty$ such that $f'_+(x) \geq -T$ holds on (a, b).*

Then there is complexity w.r.t. $\mathfrak{G}$ up to $\sqrt{n}$.

Note that we had to require $\mathfrak{G}$ to be bounded; this was not necessary in Theorem 11, which still holds for some unbounded functions. However this construction does not work for them.

Chapter 6

Expectations of Predictive Complexity

While discussing the problem of prediction with expert advice and the properties of the Aggregating Algorithm, we emphasised the absence of any law generating the outcomes. Indeed, the independence of any restriction of this kind is a very important feature of the AA.

On the other hand, it is interesting to study how our constructions behave when outcomes are generated by a certain mechanism. Suppose that we have constructed predictive complexity $\mathcal{K}$ for some game. We may treat its argument as a random variable with a particular distribution. We will consider the Bernoulli distribution. Let $\xi_1, \xi_2, \dots, \xi_n$ be independent variables and let each of them assume the value 1 with probability p and the value 0 with probability $1-p$; we will refer to this scheme as to Bernoulli trials. The random variable $\mathcal{K}(\xi_1, \xi_2, \dots, \xi_n)$ turns out to be a useful tool for studying predictive complexities.

In this chapter we evaluate the expectations $\mathbf{E}\mathcal{K}(\xi_1, \xi_2, \dots, \xi_n)$ and show that the expectations and the loss function are mutually related; the relation is established by the Legendre transformation. This result allows us to prove a uniqueness theorem. It states that if two games specify the same predictive complexity, they are equivalent in a very strong sense, namely, they have the same set of superpredictions.

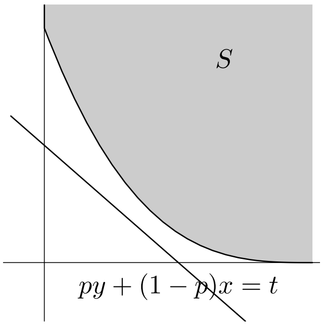

Figure 6.1: The drawing of S above the straight line $py + (1-p)x = t$ for Lemma 7

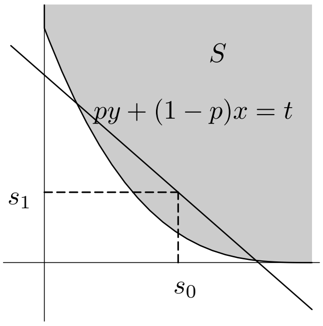

Figure 6.2: The drawing of S intersecting the straight line $py + (1-p)x = t$ for Lemma 8

6.1 The Main Lemmas on Expectations

In this section we evaluate the expectations of $\mathcal{K}$ when the argument is distributed as results of Bernoulli trials. We will evaluate the expectations by revealing their geometric meaning.

The following lemma provides a lower estimate which holds for the expectation of every superloss process w.r.t. $\mathfrak{G}$. This does not require predictive complexity w.r.t. $\mathfrak{G}$ to exist in any sense.

Lemma 7. *Let L be a superloss process and S be the set of superpredictions w.r.t. a game $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$. If $p \in (0,1)$ and $t \in \mathbb{R}$ are such that*

$$py + (1-p)x \geq t \tag{6.1}$$

holds for every $(x,y) \in S$, then

$$\mathbf{E}L(\xi_1^{(p)}, \xi_2^{(p)} \ldots \xi_n^{(p)}) \geq tn \ , \tag{6.2}$$

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

Geometrically (6.1) means that S lies 'north-east' of the straight line $py + (1-p)x = t$ (see Fig. 6.1).

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

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

where $\xi^{(p)}$ is a result of one Bernoulli trial with the probability of 1 being equal to p. It now follows that

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

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

The next lemma provides an upper estimate on the expectation of predictive complexity.

Lemma 8. *Let $\mathfrak{G} = \langle\mathbb{B},\Gamma,\lambda\rangle$ be a game with the set of superpredictions S satisfying BIN_1' and BIN_3–BIN_5, let a pair of computable numbers $(s_0,s_1)\in\mathbb{R}^2$ be a superprediction w.r.t. $\mathfrak{G}$, and*

$$ps_1 + (1-p)s_0 = t \ , \tag{6.3}$$

where $p\in(0,1)$. Then

- *if $\mathcal{K}$ is simple predictive complexity w.r.t. $\mathfrak{G}$, then there is a constant C such that*
$$\mathbf{E}\mathcal{K}(\xi_1^{(p)}\dots\xi_n^{(p)}) \leq tn + C$$
holds for every $n\in\mathbb{N}$,

- *if $\mathcal{K}$ is predictive complexity w.r.t. $\mathfrak{G}$ up to $f(n)$, then there is a constant C such that*
$$\mathbf{E}\mathcal{K}(\xi_1^{(p)}\dots\xi_n^{(p)}) \leq tn + Cf(n)$$
holds for every $n\in\mathbb{N}$, and

- *if $\mathcal{K}$ is predictive complexity w.r.t. $\mathfrak{G}$ up to $o(n)$, then there is $f:\mathbb{N}\to\mathbb{R}$ such that $f(n)=o(n)$ as $n\to+\infty$ and*
$$\mathbf{E}\mathcal{K}(\xi_1^{(p)}\dots\xi_n^{(p)}) \leq tn + f(n) \ ,$$
holds for every $n\in\mathbb{N}$,

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

Equation (6.3) means that the straight line $py + (1-p)x = t$ intersects S (see Fig. 6.2).

Proof. The proof is by considering the superloss process $L(\boldsymbol{x}) = s_0\sharp_0\boldsymbol{x} + s_1\sharp_1\boldsymbol{x}$. □

6.2 Expectations and the Legendre Transformation

The lemmas from the previous section suggest that the shape of the set of superpredictions S determines expectations and vice versa. The following theorem makes the relation explicit. The Legendre transformation and the concept of the conjugate function emerge naturally; Appendix B provides a brief introduction to the theory of the Legendre transformation.

Theorem 16. *Let* $\mathfrak{G}$ *be a game with the set of superpredictions* S *satisfying* BIN_1' *and* BIN_3*–*BIN_5*. If* $\mathcal{K}$ *is complexity w.r.t.* $\mathfrak{G}$ *up to* $o(n)$*, then for every* $p \in (0,1)$

(i) there exists a finite limit

$$\tilde{f}(p) = \lim_{n\to\infty} \frac{\mathbf{E}\mathcal{K}(\xi_1^{(p)} \ldots \xi_n^{(p)})}{n} \quad , \tag{6.4}$$

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

(ii) the equality

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

holds, where f^* *is the function conjugate to* f *specified by* $f(x) = \inf\{y \mid (x,y) \in S\}$[1] *for every* $x \in \mathbb{R}$*, and*

[1] We assume that $\inf \varnothing = +\infty$.

(iii) if $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$, then the equality

$$\tilde{f}(p) = \inf_{\gamma \in \Gamma} ((1-p)\lambda(0,\gamma) + p\lambda(1,\gamma)) \tag{6.5}$$

holds.

Proof. It follows from Lemmas 7 and 8 that for every $p \in (0,1)$ we have

$$\alpha(p)n \leq \mathbf{E}\mathcal{K}(\xi_1^{(p)} \dots \xi_n^{(p)}) \leq \alpha(p)n + o(n) \ , \tag{6.6}$$

as $n \to +\infty$, where

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

□

Depending on the ‘quality’ of complexity, the $o(n)$ term in (6.6) can be specified more precisely. If $\mathcal{K}$ is simple predictive complexity, this term may be replaced by a constant; if $\mathcal{K}$ is predictive complexity up to $g(n)$ such that $g(n) = o(n)$ as $n \to +\infty$, then the term may be replaced by $Cg(n)$, where C is some constant.

The function $\tilde{f}(p)$ defined by (6.5) is called *generalised entropy* in the literature. In the case of the logarithmic-loss game it coincides with the regular entropy. We will use this concept in our study of predictive complexity.

6.3 The Uniqueness Theorem

Reversibility of the Legendre transformation allows us to prove the following theorem.

Theorem 17 (Uniqueness Theorem). *Let $\mathfrak{G}_1$ and $\mathfrak{G}_2$ be two games with sets of superpredictions S_1 and S_2 satisfying BIN_1' and BIN_3–BIN_5 and let $\mathcal{K}_1$ and $\mathcal{K}_2$ be the complexities w.r.t $\mathfrak{G}_1$ and $\mathfrak{G}_2$, respectively, up to $o(n)$. If* *there is a function $\delta(n) = o(n)$ as $n \to \infty$ such that for every $\boldsymbol{x} \in \mathbb{B}^*$ the inequality*

$$|\mathcal{K}_1(\boldsymbol{x}) - \mathcal{K}_2(\boldsymbol{x})| \le \delta(|\boldsymbol{x}|)$$

holds, then $S_1 = S_2$.

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

$$\left| \frac{\mathbf{E}\Big[\mathcal{K}_1(\xi_1^{(p)} \dots \xi_n^{(p)}) - \mathcal{K}_2(\xi_1^{(p)} \dots \xi_n^{(p)})\Big]}{n} \right| \le \frac{\delta(n)}{n} = o(1)$$

as $n \to \infty$, where $\xi_1^{(p)}, \dots, \xi_n^{(p)}$ are results of n independent Bernoulli trials with the probability of 1 being equal to p. This implies that for every $p \in (0,1)$ the equality $\tilde{f}_1(p) = \tilde{f}_2(p)$ holds, where $\tilde{f}_1$ and $\tilde{f}_2$ are defined for the games $\mathfrak{G}_1$ and $\mathfrak{G}_2$ by (6.4). Thus $f_1^*(t) = f_2^*(t)$ for all $t \in (-\infty, 0)$, where f_1 and f_2 are defined in the same way as f in Theorem 16, (*ii*). We have $f_1^*(0) = f_2^*(0)$ since f_1^* and f_2^* are convex and closed (see Appendix B). For every $t > 0$ the equality $f_1^*(t) = f_2^*(t) = +\infty$ holds. It follows from a fundamental property of conjugate functions, namely, $f^{**} = f$, that the functions f_1 and f_2 coincide, where f_1 and f_2 are defined in the same way as f in (*ii*) of Theorem 16. This implies that $S_1 = S_2$. □

Theorem 17 has a remarkable corollary.

Corollary 9. *There is no game specifying plain Kolmogorov complexity* K*, prefix complexity* KP*, or monotone complexity* Km *as its predictive complexity.*

Proof. The difference between any of this functions and the negative logarithm of Levin's *a priori* semimeasure is bounded by a term of logarithmic order of the length of a string. If one of the functions had been predictive complexity for a game, this game would have been equivalent to the logarithmic-loss game, which has

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

$\gamma \in [0,1]$, and complexity would have coincided with logarithmic complexity $\mathcal{K}^{\log}$. But $\mathcal{K}^{\log}$ coincides with the negative logarithm of Levin's *a priori* semimeasure (see [VW98]). However neither of the differences between these functions and KM is bounded by a constant (see [V'y94, LV97]). □

Chapter 7

Linear Inequalities

The technique we developed in the previous chapter, the method of expectations, has another application. It can be used to analyse linear inequalities between complexities. Consider the following problem. Given two predictive complexities, $\mathcal{K}_1$ and $\mathcal{K}_2$ (w.r.t. $\mathfrak{G}_1$ and $\mathfrak{G}_2$), does the inequality $\mathcal{K}_1 \geq \mathcal{K}_2$ hold, at least up to some extra terms?

In this section we will resolve this problem. The inequality $\mathcal{K}_1 \geq \mathcal{K}_2$ turns out to hold if and only if the inequality $\mathbf{E}\mathcal{K}_1 \geq \mathbf{E}\mathcal{K}_2$ holds for expectations taken w.r.t. every Bernoulli distribution on strings and this is equivalent to the geometric fact $S_1 \subseteq S_2$, where S_1 and S_2 are sets of superpredictions for the respective games. The only extra additive terms which are worth considering are those of the order $|\boldsymbol{x}|$ and adding a term of this kind is equivalent to a shift of one of the sets of superpredictions.

7.1 General Inequalities

The following theorem establishes the triple equivalence between the case of a straightforward inequality, the inequality between expectations, and the geometrical interpretation.

Theorem 18 (Theorem on Linear Inequalities). *Let $\mathfrak{G}_1$ and $\mathfrak{G}_2$ be games with sets of superpredictions S_1 and S_2 satisfying BIN_1' and BIN_3–BIN_5. Let $\mathcal{K}_1$ and $\mathcal{K}_2$ be complexities w.r.t. $\mathfrak{G}_1$ and $\mathfrak{G}_2$, respectively, up to $f(n)$ such that $f(n) = o(n)$ as $n \to +\infty$. Then the following statements are equivalent:*

(i) there is a constant $C \in \mathbb{R}$ *such that for every* $\boldsymbol{x} \in \mathbb{B}^*$ *the inequality* $\mathcal{K}_1(\boldsymbol{x}) + Cf(|\boldsymbol{x}|) \geq \mathcal{K}_2(\boldsymbol{x})$ *holds,*

(ii) $S_1 \subseteq S_2$*,*

(iii) for every $p \in (0,1)$ *there is a constant* $C_p \in \mathbb{R}$ *such that for every* $n \in \mathbb{N}$ *the inequality* $\mathbf{E}\mathcal{K}_1(\xi_1^{(p)} \dots \xi_n^{(p)}) + C_p f(n) \geq \mathbf{E}\mathcal{K}_2(\xi_1^{(p)} \dots \xi_n^{(p)})$ *holds, where* $\xi_1^{(p)}, \dots, \xi_n^{(p)}$ *are results of n independent Bernoulli trials with the probability of 1 being equal to p, and*

(iv) for every $p \in (0,1)$ *the inequality* $\tilde{h}_1(p) \geq \tilde{h}_2(p)$ *holds, where* h_1 *and* h_2 *are generalised entropies for the games* $\mathfrak{G}_1$ *and* $\mathfrak{G}_2$*.*

If, however statements (i)–(iii) do not hold, there is $p_0 \in (0,1)$ *such that*

$$\mathbf{E}\mathcal{K}_1(\xi_1^{(p_0)} \dots \xi_n^{(p_0)}) - \mathbf{E}\mathcal{K}_2(\xi_1^{(p_0)} \dots \xi_n^{(p_0)}) = \Omega(n) \qquad (7.1)$$

as $n \to +\infty$*.*

Proof. Statements (*iii*) and (*iv*) are equivalent by Theorem 16.

The implication (*i*) ⇒ (*iii*) is trivial.

Let us prove that (*ii*) ⇒ (*i*). Let L be a superloss process w.r.t. $\mathfrak{G}_1$. It follows from the definition, that, for every $\boldsymbol{x} \in \mathbb{B}^*$, we have $(L(\boldsymbol{x}0) - L(\boldsymbol{x}), L(\boldsymbol{x}1) - L(\boldsymbol{x})) \in S_1 \subseteq S_2$. It implies that L is a superloss process w.r.t. $\mathfrak{G}_2$. Take $L = \mathcal{K}_1$. Now (*i*) follows from the definition of predictive complexity.

It remains to prove that (*iii*) ⇒ (*ii*) and that the last claim from the statement of the theorem is true. Let us assume that condition (*ii*) is violated, i.e., there exists a pair $(s_0, s_1) \in S_1 \setminus S_2$. It follows from BIN_5 that s_0 and s_1 can be assumed to belong to $\mathbb{R}^2$. We will find $p_0 \in (0,1)$ such that (7.1) holds.

Lemma 9. *Let* $S \subseteq \mathbb{R}^2$ *be a set satisfying the following conditions:*

- *S is closed,*
- *S is convex,*
- *for every* $(x,y) \in S$ *and every* $s,t > 0$ *we have* $(x+s, y+t) \in S$*, and*
- *there are* $x_0, y_0 \in \mathbb{R}$ *such that* $S \subseteq [x_0, +\infty) \times [y_0, +\infty)$*.*

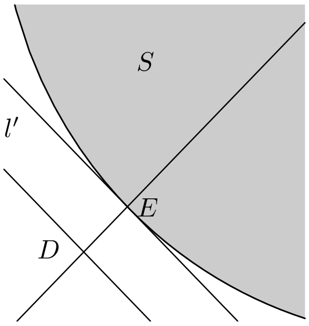

Figure 7.1: The drawing for Lemma 9. The set S is shaded

If $D \in \mathbb{R}^2$ is a point such that $D \notin S$, then there is a straight line l with the equation $ax + by = c$, where $a, b > 0$ and c are some constants, that passes through D but such that S lies above l and separated from l, i.e., there is $\delta > 0$ such that for every $(x, y) \in S$ the inequality $ax + by \geq c + \delta$ holds.

Proof. Let $D = (u_0, v_0)$. The proof can be derived from the Separation Theorem for convex sets (see, say, [Egg58]) but we will give a self-contained proof. Since S is closed in the standard topology of $\mathbb{R}^2$, there exists a point $E \in S$ which is closest to D. The convexity of S implies that all the points of S lie on one side of the straight line l' which is perpendicular to DE and passes through E, and D lies on the other side (see Fig. 7.1). Let l'' be parallel to l and pass through D and let its equation be $a''x + b''y = c''$. It follows from the properties of S that $a'', b'' \geq 0$ (or this can be achieved by multiplying the both sides by -1). If neither of a'' and b'' equals 0, i.e., l'' is neither vertical nor horizontal, we may take l to coincide with l''. If not, l can be obtained by slightly turning l'' around D; this is possible since $S \subseteq [x_0, +\infty) \times [y_0, +\infty)$. □

This lemma can be applied to $S_2 \cap \mathbb{R}^2$ and the point (s_0, s_1). After the appropriate normalisation, the equation of the resulting line reduces to $p_0 y + (1 - p_0)x = t$ for some $p_0 \in (0, 1)$ and $t \in \mathbb{R}$. We have $p_0 s_1 + (1 - p_0)s_0 = t$, but there is $\delta > 0$ such that for every $x, y \in S_2$ the inequality $p_0 y + (1 - p_0)x \geq t + \delta$ holds. It remains to apply Lemmas 7 and 8. □

In Sect. 5.5 we discussed the following fact. Let S is the set of superpredictions for a game $\mathfrak{G}$ and let $\mathfrak{G}'$ be a game with the set of superpredictions $S + (a, a)$, i.e., the set obtained by a shift of S along the straight line $x = y$.

Then there is a correspondence between superloss processes w.r.t. $\mathfrak{G}$ and superloss processes w.r.t. $\mathfrak{G}'$, namely, $L(\boldsymbol{x})$ is a superloss process w.r.t. $\mathfrak{G}$ if and only if $L(\boldsymbol{x}) + a|\boldsymbol{x}|$ is a superloss process w.r.t. $\mathfrak{G}'$. A similar statement applies to scaling. If $\mathfrak{G}'$ has the set of superpredictions aS, then $L(\boldsymbol{x})$ is a superloss process w.r.t. $\mathfrak{G}$ if and only if $aL(\boldsymbol{x})$ is a superloss process w.r.t. $\mathfrak{G}'$. These observations imply the following corollary.

Corollary 10. *Under the conditions of Theorem 18 the following statements are equivalent, where $a, b \in \mathbb{R}$ and $a \geq 0$:*

(i) there is $C \in \mathbb{R}$ such that for all $\boldsymbol{x} \in \mathbb{B}^*$ the inequality $a\mathcal{K}_1(\boldsymbol{x}) + b|\boldsymbol{x}| + Cf(|\boldsymbol{x}|) \geq \mathcal{K}_2(\boldsymbol{x})$ holds,

(ii) $aS_1 + (b, b) \subseteq S_2$, and

(iii) for every $p \in (0, 1)$ there exists $C_p \in \mathbb{R}$ such that for all $n \in \mathbb{N}$ the inequality $a\mathbf{E}\mathcal{K}_1(\xi_1^{(p)} \ldots \xi_n^{(p)}) + bn + C_p f(n) \geq \mathbf{E}\mathcal{K}_2(\xi_1^{(p)} \ldots \xi_n^{(p)})$ holds, where $\xi_1^{(p)}, \ldots, \xi_n^{(p)}$ are results of n independent Bernoulli trials with the probability of 1 being equal to p, and

(iv) for every $p \in (0, 1)$ the inequality $a\tilde{h}_1(p) + b \geq \tilde{h}_2(p)$ holds, where h_1 and h_2 are generalised entropies for the games $\mathfrak{G}_1$ and $\mathfrak{G}_2$.

One may wonder whether the extra term $b|\boldsymbol{x}|$ can be replaced by a smaller term. The next corollary clarifies the situation.

Corollary 11. *Suppose that under the conditions of Theorem 18 the following statement holds:*

For every $p \in (0, 1)$ there exists a function $\alpha_p : \mathbb{N} \to \mathbb{R}$ such that $\alpha_p(n) = o(n)$ as $n \to +\infty$ and for every $n \in \mathbb{N}$ the inequality

$$a\mathbf{E}\mathcal{K}_1(\xi_1^{(p)} \ldots \xi_n^{(p)}) + bn + \alpha_p(n) \geq \mathbf{E}\mathcal{K}_2(\xi_1^{(p)} \ldots \xi_n^{(p)})$$

holds, where $a, b \in \mathbb{R}$, $a \geq 0$, and $\xi_1^{(p)}, \ldots, \xi_n^{(p)}$ are results of n independent Bernoulli trials with the probability of 1 being equal to p.

Then there is $C \in \mathbb{R}$ such that the inequality

$$a\mathcal{K}_1(\boldsymbol{x}) + b|\boldsymbol{x}| + Cf(|\boldsymbol{x}|) \geq \mathcal{K}_2(\boldsymbol{x})$$

holds for every $\boldsymbol{x} \in \mathbb{B}^*$.

Corollary 12. *If under the conditions of Theorem 18 there exists a function $g : \mathbb{N} \to \mathbb{R}$ such that $g(n) = o(n)$ as $n \to +\infty$ and, for every $\boldsymbol{x} \in \mathbb{B}^*$, the inequality*

$$a\mathcal{K}_1(\boldsymbol{x}) + b|\boldsymbol{x}| + g(|\boldsymbol{x}|) \geq \mathcal{K}_2(\boldsymbol{x}) \ ,$$

where $a, b \in \mathbb{R}$ and $a \geq 0$, holds, then there is $C \in \mathbb{R}$ such that the inequality

$$a\mathcal{K}_1(\boldsymbol{x}) + b|\boldsymbol{x}| + Cf(|\boldsymbol{x}|) \geq \mathcal{K}_2(\boldsymbol{x})$$

holds.

Clearly, this corollary trivialises when $g = O(f)$, but it may be useful in other cases.

There is a class of linear inequalities which is not covered by Theorem 18. We do not have an equally general result for them. The following theorem deals with symmetric games. A game $\mathfrak{G}$ with the set of superpredictions S is called *symmetric* if the set S is symmetric w.r.t. the straight line $x = y$.

Theorem 19. *Let $\mathfrak{G}_1$ and $\mathfrak{G}_2$ be two symmetric games with sets of superpredictions S_1 and S_2 satisfying BIN_1' and BIN_3–BIN_5 and let $\mathcal{K}_1$ and $\mathcal{K}_2$ be complexities w.r.t $\mathfrak{G}_1$ and $\mathfrak{G}_2$, respectively, up to $f(n)$ such that $f(n) = o(n)$ as $n \to +\infty$. Let*

$$r_1 = \inf\{t \in \mathbb{R} \mid (t,t) \in S_1\}$$
$$r_2 = \inf\{t \in \mathbb{R} \mid (t,t) \in S_2\} \ .$$

Then for every $a_1, a_2 \geq 0$ the following statements are equivalent:

(i) there is a constant $C \in \mathbb{R}$ such that for every $\boldsymbol{x} \in \mathbb{B}^*$ the inequality $a_1\mathcal{K}_1(\boldsymbol{x}) + a_2\mathcal{K}_2(\boldsymbol{x}) \leq b|\boldsymbol{x}| + Cf(|\boldsymbol{x}|)$ holds,

(ii) $a_1 r_1 + a_2 r_2 \leq b$,

(iii) for every $p \in (0,1)$ there is a constant $C \in \mathbb{R}$ such that for every $n \in \mathbb{N}$ the inequality $a_1\mathbf{E}\mathcal{K}_1(\xi_1^{(1/2)} \ldots \xi_n^{(1/2)}) + a_2\mathbf{E}\mathcal{K}_2(\xi_1^{(1/2)} \ldots \xi_n^{(1/2)}) \leq bn + Cf(n)$ holds, where $\xi_1^{(1/2)}, \ldots, \xi_n^{(1/2)}$ are results of n independent Bernoulli trials with the probability of 1 being equal to $1/2$, and

(iv) the inequality $a_1\tilde{h}_1(1/2) + a_2\tilde{h}_2(1/2) \leq b$ holds, where h_1 and h_2 are generalised entropies for the games $\mathfrak{G}_1$ and $\mathfrak{G}_2$.

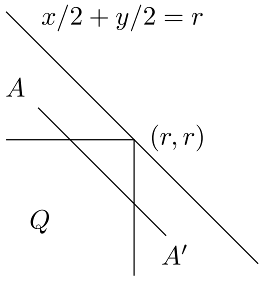

Figure 7.2: The quadrant Q and the segment $[A, A']$ from the proof of Lemma 10.

If, however statements (i)–(iii) do not hold, then

$$a_1\mathbf{E}\mathcal{K}_1(\xi_1^{(p_0)} \dots \xi_n^{(p_0)}) + a_2\mathbf{E}\mathcal{K}_2(\xi_1^{(p_0)} \dots \xi_n^{(p_0)}) - bn = \Omega(n)$$

as $n \to +\infty$.

Proof. The proof is similar to that of Theorem 18 but a little simpler. We need the following lemma.

Lemma 10. *Let $S \subseteq \mathbb{R}^2$ be a set satisfying the following conditions:*

- *S is closed,*
- *S is convex,*
- *for every $(x, y) \in S$ and every $s, t > 0$ we have $(x + s, y + t) \in S$, and*
- *S is symmetric w.r.t. the straight line $x = y$.*

If $r = \inf\{t \in \mathbb{R} \mid (t, t) \in S\}$ then S lies above its tangent $x/2 + y/2 = r$, i.e., for every $(x, y) \in S$ the inequality $x/2 + y/2 \geq r$ holds.

Proof. The set S does not intersect the set $Q = (-\infty, r)^2$. If there is $A = (x_0, y_0) \in S$ which lies below the straight line $x/2 + y/2 = r$ then $A' = (y_0, x_0) \in S$ lies below this line too. Since S is convex, the segment connecting A and A' must lie inside S but it necessarily intersects Q (see Fig. 7.2).

□

The lemma implies that there is $C > 0$ such that the following inequalities hold:

$$r_1 n \le \mathbf{E}\mathcal{K}_1(\xi_1^{(1/2)}\ldots\xi_n^{(1/2)}) \le r_1 n + Cf(n) \ ,$$
$$r_2 n \le \mathbf{E}\mathcal{K}_2(\xi_1^{(1/2)}\ldots\xi_n^{(1/2)}) \le r_2 n + Cf(n) \ .$$

Now we can prove the theorem. The implication $(i) \Rightarrow (iii)$ is trivial. The implication $(ii) \Rightarrow (i)$ can be proved by observing that for all $\boldsymbol{x} \in \mathbb{B}^*$ the inequalities $Kcal_1(\boldsymbol{x}) \le r_1|\boldsymbol{x}| + Cf(|\boldsymbol{x}|)$ and $Kcal_2(\boldsymbol{x}) \le r_2|\boldsymbol{x}| + Cf(|\boldsymbol{x}|)$ hold for some constant $C > 0$. The implication $(iii) \Rightarrow (ii)$ is proved in the same way as in Theorem 18.

□

7.2 Linear Inequalities between Square-Loss and Logarithmic-Loss Complexities

In this section we apply our general results to study inequalities between two specific complexities, $\mathcal{K}^{\mathrm{sq}}$ and $\mathcal{K}^{\log}$.

7.2.1 Expectations

Our proofs rely upon the probabilistic criterion from Corollary 10. We need the entropies

$$h_{sq}(p) = \min_{0\le\gamma\le1}(p(1-\gamma)^2 + (1-p)\gamma^2) \tag{7.2}$$
$$= p(1-p) \tag{7.3}$$

and

$$h_{log}(p) = \min_{0\le\gamma\le1}(-p\log\gamma - (1-p)\log(1-\gamma)) \tag{7.4}$$
$$= -p\log p - (1-p)\log(1-p) \ . \tag{7.5}$$

Corollary 10 and Theorem 19 imply the following lemma. It is one of the possible ways to reduce the study of linear relations between $\mathcal{K}^{\log}$ and $\mathcal{K}^{\mathrm{sq}}$ to a problem of calculus.

Lemma 11. *Consider real $a, a_1, a_2, \ge 0$ and real b. There is $C \in \mathbb{R}$ such that for all $\boldsymbol{x} \in \mathbb{B}^*$ the inequality*

$$a\mathcal{K}^{\mathrm{sq}}(\boldsymbol{x}) + b|\boldsymbol{x}| + C \ge \mathcal{K}^{\log}(\boldsymbol{x})$$

holds if and only if $ah_{sq}(p) + b \geq h_{log}(p)$ *holds for every* $p \in [0,1]$. *There is* $C \in \mathbb{R}$ *such that for all* $\boldsymbol{x} \in \mathbb{B}^*$ *the inequality*

$$a\mathcal{K}^{\log}(\boldsymbol{x}) + b|\boldsymbol{x}| + C \geq \mathcal{K}^{\mathrm{sq}}(\boldsymbol{x})$$

holds if and only if $ah_{log}(p) + b \geq h_{sq}(p)$ *holds for every* $p \in [0,1]$. *There is* $C \in \mathbb{R}$ *such that for all* $\boldsymbol{x} \in \mathbb{B}^*$ *the inequality*

$$a_1\mathcal{K}^{\mathrm{sq}}(\boldsymbol{x}) + a_2\mathcal{K}^{\log}(\boldsymbol{x}) \leq b|\boldsymbol{x}| + C$$

holds if and only if $a_1 h_{sq}(1/2) + a_2 h_{log}(1/2) \leq b$ *holds.*

7.2.2 Case $a\mathcal{K}^{\mathrm{sq}}(\boldsymbol{x}) + b|\boldsymbol{x}| + C \geq \mathcal{K}^{\log}(\boldsymbol{x})$

To describe the boundary of the set

$$M = \{(a,b) \mid a \geq 0 \text{ and } \exists C > 0 \, \forall x \in \mathbb{B}^* : a\mathcal{K}^{\mathrm{sq}}(x) + b|x| + C \geq \mathcal{K}^{\log}(x)\} \ , \tag{7.6}$$

we introduce $\varphi(a) = \inf\{b \mid (a,b) \in M\}$, where $a \geq 0$. By Proposition 5, the points $(a, \varphi(a))$ belong to M. Let

$$f(a,p) = h_{log}(p) - ah_{sq}(p) = -p\log p - (1-p)\log(1-p) - ap(1-p) \ ,$$

where $a \geq 0$ and $p \in [0,1]$. Clearly,

$$\varphi(a) = \max_{p\in[0,1]} f(a,p) \ . \tag{7.7}$$

Theorem 20. *For every* $a \in [0, 2/\ln 2]$, *we have*

$$\varphi(a) = 1 - \frac{a}{4} \ .$$

Proof. Let us fix any $a \in [0, 2/\ln 2]$ and calculate $\max_{p\in[0,1]} f(a,p)$. We have

$$\frac{\partial f(a,p)}{\partial p} = a(2p-1) + \log(1-p) - \log p \ , \tag{7.8}$$

$$\frac{\partial^2 f(a,p)}{\partial p^2} = 2a - \frac{1}{p(1-p)\ln 2} \ . \tag{7.9}$$

Since $\max_{p\in[0,1]} p(1-p) = 1/4$, the function $f(a,p)$ is concave in the second argument for every $a \in [0, 2/\ln 2]$. On the other hand, the derivative $\partial f(a,p)/\partial p$ vanishes at $p = 1/2$. Hence the maximum in (7.7) is attained at the point $p = 1/2$. The substitution of $p = 1/2$ into the definition of f completes the proof. □

The behaviour of φ on the interval $(2/\ln 2, +\infty)$ is more complicated because the maximum is no longer attained at $p = 1/2$. We do not know any explicit formula for φ on this interval. The following lemmas describe some properties of φ.

Lemma 12.

$$\varphi(a) \sim \frac{2^{-a}}{\ln 2} \quad \text{as} \quad a \to +\infty \ .$$

Proof. Consider the equation

$$\frac{\partial f(a,p)}{\partial p} = 0 \ . \tag{7.10}$$

It is equivalent to the equation $p = r(a,p)$, where

$$r(a,p) = \frac{1}{1 + 2^{a(1-2p)}} \ .$$

For every $a > 0$, the function $r(a,p)$ is concave in the second argument in the interval $[1/2, 1]$ and convex in the second argument in the interval $[0, 1/2]$ because

$$\frac{\partial^2 r(a,p)}{\partial p^2} = \frac{a^2 2^{a(1+2p)}(2^a - 2^{2ap}) 4 \ln^2 2}{(2^a + 4^{ap})^3} \ .$$

On the other hand, for every $a > 0$, we have $r(a,0) > 0$, $r(a,1/2) = 1/2$, and $r(a,1) < 1$. For any $a > 2/\ln 2$

$$\left.\frac{\partial r(a,p)}{\partial p}\right|_{p=\frac{1}{2}} = \frac{1}{2} a \ln 2 > 1 \ .$$

Thus, for every fixed $a > 2/\ln 2$, Eq. (7.10) has 3 roots (see Fig. 7.3). If we denote the smallest one by $\xi(a)$, the roots are $\xi(a)$, $1/2$ and $1 - \xi(a)$. Since

$$\left.\frac{\partial^2 f(a,p)}{\partial p^2}\right|_{p=\frac{1}{2}} > 0$$

and $\lim_{p\to 0+} \partial f/\partial p = -\lim_{p \to 1-0} \partial f/\partial p = +\infty$, the point $p = 1/2$ is the point of a local minimum of f and both $p = \xi(a)$ and $p = 1 - \xi(a)$ are points where the maximum from (7.7) is attained.

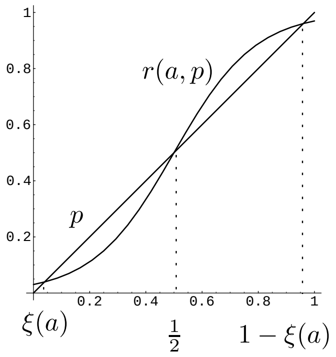

Figure 7.3: The function $r(a,p)$ with $a=4$

The function $r(a,p)$ is strictly increasing in p for every $a>0$. Obviously, for every $p\in[0,1/2)$, this function is strictly decreasing in a and $\lim_{a\to+\infty} r(a,p)=0$. These observations imply that $\xi(a)$ is strictly decreasing and $\lim_{a\to+\infty}\xi(a)=0$.

One can easily see the function $\xi(a)$ maps the half-line $(2/\ln 2,+\infty)$ onto the interval $(0,1/2)$. One may consider the inverse function $a(\xi)$ which maps the interval $(0,1/2)$ onto $(2/\ln 2,+\infty)$. Equation (7.10) implies that

$$\begin{aligned} a(\xi) &= \frac{1}{1-2\xi}\log\left(\frac{1}{\xi}-1\right) && (7.11)\\ &= -\log\xi + o(1) \text{ as } \xi\to 0+\ \ . && (7.12)\end{aligned}$$

Let us now substitute ξ for p and $a(\xi)$ for a in the definition of $f(a,p)$. One may check by direct calculation that

$$\varphi(a(\xi)) = f(a(\xi),\xi) = \frac{\xi}{\ln 2} + o(\xi) \qquad (7.13)$$

as $\xi\to 0$. Equation (7.12) implies that $\xi = 2^{-a}+o(2^{-a})$. Substituting this into (7.13) completes the proof. □

Lemma 13. *For every $a\ge 0$, we have*

$$\varphi(a)\ge\log(1+2^{-a})\ .$$

Proof. For every $a\ge 0$ and $p\in[0,1]$, the estimate

$$f(a,p)\ge -ap-p\log p-(1-p)\log(1-p)=h(a,p) \qquad (7.14)$$

holds. The function h is concave in the second argument and its derivative $\partial h(a,p)/\partial p$ vanishes at $p = 1/(1+2^a)$. Substituting this value of p into (7.14) completes the proof. □

7.2.3 Case $a\mathcal{K}^{\log}(\boldsymbol{x}) + b|\boldsymbol{x}| + C \geq \mathcal{K}^{\mathrm{sq}}(\boldsymbol{x})$

This case is simpler.

Theorem 21. *If $a \geq 0$, the following conditions are equivalent:*

- *There is $C \in \mathbb{R}$ such that for all $\boldsymbol{x} \in \mathbb{B}^*$ the inequality $a\mathcal{K}^{\log}(\boldsymbol{x}) + b|\boldsymbol{x}| + C \geq \mathcal{K}^{\mathrm{sq}}(\boldsymbol{x})$ holds.*
- *$b \geq \max(\frac{1}{4} - a, 0)$.*

Proof. Let

$$s(a,p) = h_{sq}(p) - ah_{log}(p) = p(1-p) - a\left(-p\log p - (1-p)\log(1-p)\right) \;,$$

where $a \geq 0$ and $p \in [0,1]$. Clearly the derivative

$$\frac{\partial s(a,p)}{\partial p} = 1 - 2p - a\left(\log(1-p) - \log p\right)$$

always vanishes at $p = 1/2$. The function $u(p) = \log(1-p) - \log p$ is convex on $[0,1/2]$ and concave on $[1/2,1]$. One can easily check that for every fixed $a \in [0,1/4]$ the function $s(a,p)$ has a local maximum at $p = 1/2$ and local minimums at some points from $(0,1/2)$ and $(1/2,0)$. The value $s(a,1/2) = 1/4 - a$ is the maximal on $[0,1]$.

If $a \geq 1/4$, we have $\sup_{p\in[0,1]} s(a,p) = 0$. Indeed, $s(a,p)$ decreases in the first argument and, for every $a \geq 0$, we have $s(a,0) = 0$.

The theorem follows.

□

7.2.4 Case $a_1\mathcal{K}^{\mathrm{sq}}(\boldsymbol{x}) + a_2\mathcal{K}^{\log}(\boldsymbol{x}) \leq b|\boldsymbol{x}| + C$

This case is trivial. The next theorem follows immediately from Lemma 11.

Theorem 22. *For every real $a_1, a_2 > 0$ and every real b,the following conditions are equivalent:*

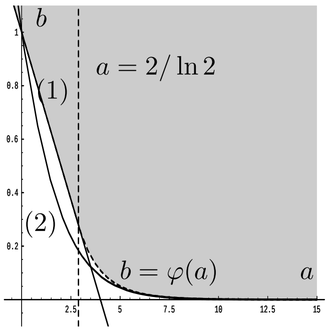

Figure 7.4: $a\mathcal{K}^{\mathrm{sq}}(\boldsymbol{x}) + b|\boldsymbol{x}| + C \geq \mathcal{K}^{\log}$

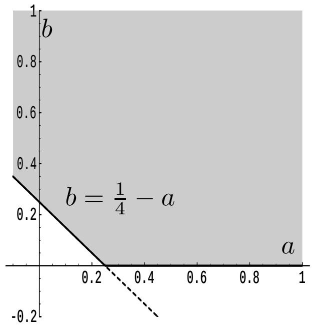

Figure 7.5: $a\mathcal{K}^{\log}(\boldsymbol{x}) + b|\boldsymbol{x}| + C \geq \mathcal{K}^{\mathrm{sq}}$

- *There is* $C \in \mathbb{R}$ *such that for all* $\boldsymbol{x} \in \mathbb{B}^*$ *the inequality* $a_1\mathcal{K}^{\mathrm{sq}}(\boldsymbol{x}) + a_2\mathcal{K}^{\log}(\boldsymbol{x}) \leq b|\boldsymbol{x}| + C$ *holds.*

- $a_1/4 + a_2 \leq b$*.*

7.2.5 Conclusion

Here we summarise the results of this section. In Fig. 7.4 the set of all pairs (a, b) such that there is $C \in \mathbb{R}$ such that for all $\boldsymbol{x} \in \mathbb{B}^*$ the inequality $a\mathcal{K}^{\mathrm{sq}}(\boldsymbol{x}) + b|\boldsymbol{x}| + C \geq \mathcal{K}^{\log}(\boldsymbol{x})$ holds is coloured grey. The curve $b = 1 - a/4$ is denoted by (1) and the curve $b = \log(1 + 2^{-a})$ is denoted by (2). The curve $b = \varphi(a)$ was plotted by means of a simple numerical evaluation (cf. Lemma 11). In Fig. 7.5 the set of all the pairs (a, b) such that there is $C \in \mathbb{R}$ such that for all $\boldsymbol{x} \in \mathbb{B}^*$ the inequality $a\mathcal{K}^{\log}(\boldsymbol{x}) + b|\boldsymbol{x}| + C \geq \mathcal{K}^{\mathrm{sq}}(\boldsymbol{x})$ holds is shaded.

Chapter 8

Incompressibility and Unpredictability

Kolmogorov complexity is the length of the shortest description of a finite string. The length of the shortest description is less than or equal to the length of the string itself. If there is a short description of a string, the string has certain regularity; otherwise it may be called random. The Incompressibility Property for Kolmogorov complexity states that most of the strings are random in this sense, i.e., their Kolmogorov complexity is close to the length.

Proposition 11 (Incompressibility Property).

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

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

holds.

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

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

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

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

In this chapter the Incompressibility Property is extended to the case of predictive complexity. We are going to apply the theory of martingales and Doob's inequality. Appendix C contains a brief survey on martingales. We also prove that the bound we derive is tight.

Our results are restricted to symmetric games defined in Chapter 7. Recall that a game $\mathfrak{G}$ with the set of superpredictions S is called *symmetric* if S is symmetric w.r.t. the straight line $x = y$.

8.1 Conditional Complexity

In order to formulate some results concerning unpredictability, we need the concept of conditional predictive complexity. The definition is an elaboration of the unconditional definition from Chapter 5. Two approaches, the batch and the on-line, are possible. We will only formulate the batch definition, leaving the on-line case beyond the scope of this thesis.

Let $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$ be a game with the set of superpredictions S satisfying BIN_1' and BIN_3–BIN_4. A function $L : \mathbb{B}^* \times \mathbb{B}^* \to (-\infty, +\infty]$ (we will separate arguments of L by the vertical line $\mid$ rather then by the comma) is a *conditional superloss process* if

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

$$\begin{cases} L(\boldsymbol{x} \mid \boldsymbol{y}) + \lambda(\gamma, 0) \leq L(\boldsymbol{x}0 \mid \boldsymbol{y}) \ , \\ L(\boldsymbol{x} \mid \boldsymbol{y}) + \lambda(\gamma, 1) \leq L(\boldsymbol{x}1 \mid \boldsymbol{y}) \ . \end{cases} \tag{8.1}$$

In other words, L is semicomputable from above as a function of two arguments and for every fixed $\boldsymbol{y} \in \mathbb{B}^*$ the function $L(\cdot \mid \boldsymbol{y})$ is a superloss process.

A conditional superloss process $\mathcal{K}$ is *conditional (simple) prediction complexity* if for every conditional superloss process L there is a constant such that the inequality $\mathcal{K}(\boldsymbol{x} \mid \boldsymbol{y}) \leq L(\boldsymbol{x} \mid \boldsymbol{y}) + C$ holds for every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{B}^*$.

It is easy to see (cf. Proposition 9) that there is an effective enumeration of all conditional superloss processes $L_1, L_2, \ldots$. The following equivalent of Proposition 10 holds:

Proposition 12. *If a game $\mathfrak{G}$ with the set of superpredictions S satisfying BIN_1–BIN_4 is mixable, there is conditional simple predictive complexity w.r.t. $\mathfrak{G}$.*

The following inequalities show relations between conditional and unconditional complexities. Since the unconditional complexity $\mathcal{K}$ can be treated as a conditional superloss process, there is a constant C such that for all $\boldsymbol{y} \in \mathbb{B}^*$ the inequality

$$\mathcal{K}(\boldsymbol{x} \mid \boldsymbol{y}) \leq \mathcal{K}(\boldsymbol{x}) + C \tag{8.2}$$

holds. On the other hand, if the game is β-mixable, then there is C such that for all $\boldsymbol{y} \in \mathbb{B}^*$ the inequality

$$\mathcal{K}(\boldsymbol{x}) \leq \mathcal{K}(\boldsymbol{x} \mid \boldsymbol{y}) + \frac{\mathrm{KP}(\boldsymbol{y})}{\ln(1/\beta)} + C \tag{8.3}$$

holds, where KP stands for prefix complexity.

We will be considering complexity $\mathcal{K}(\boldsymbol{x} \mid m)$, where m is a positive integer. We assume that there is a natural computable mapping of $\mathbb{N}$ into $\mathbb{B}$ and identify a subset of $\mathbb{B}$ with $\mathbb{N}$.

8.2 The Unpredictability Property

The following theorem shows that most of the strings $\boldsymbol{x}$ have complexity close to $B|\boldsymbol{x}|$. The notation $\boldsymbol{x}^{(k)}$ stands for the prefix of length k (i.e., first k bits) of a binary string $\boldsymbol{x}$.

Theorem 23. *Let $\mathfrak{G}$ be a symmetrical game with the set of superpredictions S satisfying BIN_1' and BIN_3–BIN_5; let $B = \inf\{t \mid (t,t) \in S\}$. Suppose that $\mathfrak{G}$ specifies conditional complexity $\mathcal{K}$. Then*

(i) there is $C > 0$ such that for every string $\boldsymbol{x}$ the bound

$$\mathcal{K}(\boldsymbol{x}) \leq B|\boldsymbol{x}| + C \tag{8.4}$$

holds, and

(ii) if $\beta \in (0,1)$ is such that the set $\mathfrak{B}_\beta(S)$ lies below the straight line $x + y = 2\beta^B$, then for every positive integer n and every real non-negative m we have

$$\frac{\left|\{\boldsymbol{x} \in \mathbb{B}^n \mid \exists k \in \{1,2,\ldots,n\} : \mathcal{K}(\boldsymbol{x}^{(k)} \mid m) \leq Bk - m\}\right|}{2^n} \leq \beta^m \ . \tag{8.5}$$

Proof. Part (i) is trivial. The function $L(\boldsymbol{x}) = B|x|$ is a superloss process w.r.t. $\mathfrak{G}$ and thus there is $C > 0$ such that $\mathcal{K}(\boldsymbol{x}) \leq L(\boldsymbol{x}) + C$.

We are now moving on to (ii). Let β be such that $\mathfrak{B}_\beta(S)$ lies below the straight line $x + y = 2\beta^B$. Consider the function $\beta^{\mathcal{K}(\boldsymbol{x}|m)-B|\boldsymbol{x}|}$. If we show that for every fixed m it is a supermartingale and apply Proposition 14 in the case of the Bernoulli distribution with the probability of 1 equal to 1/2, the bound will follow.

Lemma 14. *Under the conditions of Theorem 23, for every $\boldsymbol{x} \in \mathbb{B}$, the inequality*

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

holds for every positive integer m.

Proof of Lemma 14. By definition of predictive complexity, the pair $(\mathcal{K}(\boldsymbol{x}1 \mid m) - \mathcal{K}(\boldsymbol{x} \mid m), \mathcal{K}(\boldsymbol{x}0 \mid m) - \mathcal{K}(\boldsymbol{x} \mid m))$ is a superprediction, i.e., belongs to S. The conditions of Theorem 23 imply that for every $(x, y) \in \mathfrak{B}_\beta(S)$, the inequality $x/2 + y/2 \leq \beta^B$ holds. The lemma follows. □

It follows from this lemma that $\beta^{\mathcal{K}(\boldsymbol{x}|m)-B|\boldsymbol{x}|}$ is a supermartingale (see Appendix C). □

Note that (ii) holds for every conditional superloss process, not just $\mathcal{K}$.

8.3 Tightness of the Bound

In this section we prove that the value of β in bound (8.5) cannot be decreased.

Theorem 24. *Let $\mathfrak{G}$ be a symmetrical game with the set of superpredictions S satisfying BIN_1' and BIN_3–BIN_5; let $B = \inf\{t \mid (t,t) \in S\}$. Let $\mathfrak{G}$ specify conditional complexity $\mathcal{K}$. Let $\beta \in (0,1)$ be such that the set $\mathfrak{B}_\beta(S)$ does not lie below the straight line $x + y = 2\beta^B$. Then there are positive constants c and θ such that for every non-negative computable number m and positive integer $n \geq cm$ the inequality*

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

holds.

Note that if there are values of β satisfying the conditions of Theorem 23 then their infimum is greater than 0. Indeed, consider Inequalities (4.7) and (4.8). Clearly, the boundary of $\mathfrak{B}_\beta(S)$ is the graph of a convex function in a small vicinity of (B, B) for sufficiently small β.

Proof. For every computable m, we will construct a superloss process L_m that achieves

$$p(n, m) = \frac{|\{\boldsymbol{x} \in \mathbb{B}^n \mid L_m(\boldsymbol{x}) \leq Bn - m\}|}{2^n} \geq \frac{1}{4}\beta^m \tag{8.6}$$

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

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

There is $\Delta \in (0, \beta^B]$ such that $(\log_\beta(\beta^B - \Delta), \log_\beta(\beta^B + \Delta))$ is a superprediction. Let $\mathfrak{A}$ be the superstrategy that always outputs $(\log_\beta(\beta^B - \Delta), \log_\beta(\beta^B + \Delta))$ and let $L(\boldsymbol{x})$ be the loss of this superstrategy. The superstrategy $\mathfrak{A}_m$ works as follows. It imitates $\mathfrak{A}$ as long as the inequality $L(\boldsymbol{x}) > B|\boldsymbol{x}| - m$ holds. After the inequality gets violated, the superstrategy switches to the superprediction (B, B). Let $L_m(\boldsymbol{x})$ be the loss of $\mathfrak{A}_m$. Put $A = B - \log_\beta(\beta^B + \Delta) > 0$ so that $(B|\boldsymbol{x}| - m) - L_m(\boldsymbol{x})$ does not exceed A.

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

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

for every positive computable m, where $\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}$ are results of n independent Bernoulli trials with the probability of success being equal to 1/2. Note that $M_m(\boldsymbol{x}) \leq \beta^{-m-A} \leq 2\beta^{-m}$ for every $\boldsymbol{x} \in \mathbb{B}^m$.

Fix a positive computable m. Pick $\boldsymbol{x}$ of length n and consider the 'trajectories'

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

and

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

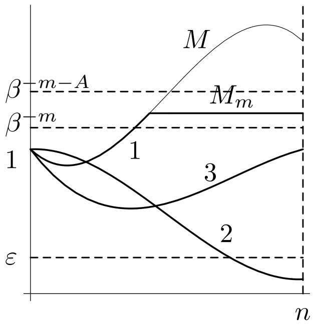

Figure 8.1: Three options for trajectories from the proof of Theorem 24

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

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

These three options are shown in Fig. 8.1, where the values of $M(\boldsymbol{x}^{(k)})$ and $M_m(\boldsymbol{x}^{(k)})$ are plotted against those of k.

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

$$1 = \mathbf{E} M_m(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}) = \Sigma_1 + \Sigma_2 + \Sigma_3 \ , \tag{8.7}$$

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

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

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

$$\Pr\{M(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}) > \varepsilon\} =$$
$$\Pr\{L(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}) - Bn < \log_\beta \varepsilon\}$$

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

$$1 \le 2\beta^{-m} p(n,m) + \varepsilon + \beta^{-m}\alpha_\varepsilon(n) \tag{8.8}$$

and thus

$$p(n,m) \ge \frac{\beta^m}{2} - \varepsilon\beta^m - \alpha_\varepsilon(n) \ . \tag{8.9}$$

We will construct an upper bound for $\alpha_\varepsilon(n)$ by means of the Chernoff bound. A derivation of the Chernoff bound may be found in [Gal68]; we need the following simple form of the bound. If $X_1, X_2, \ldots, X_n$ are independent Bernoulli trials with the probability of success equal to $p \in (0,1)$, $S = X_1 + X_2 + \cdots + X_n$, and $\gamma \in [0,1]$, then

$$\Pr\{S < (1-\gamma)pn\} \le e^{-np\gamma^2/2} \ . \tag{8.10}$$

The function $L(\boldsymbol{x}) - B|\boldsymbol{x}|$ may be treated as a biased random walk. The value

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

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

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

can be treated as the sum of outcomes of independent Bernoulli trials and thus satisfies (8.10). The Chernoff bound implies that for every $\gamma \in [0,1]$ we have

$$\Pr\{L(\xi_1^{(1/2)}, \xi_2^{(1/2)}, \ldots, \xi_n^{(1/2)}) - Bn \le -\gamma rn + dn\} \le e^{-n\gamma^2/4} \ .$$

Therefore if n and ε are such that the value $\gamma = \frac{d}{r} - \frac{1}{rn}\log_\beta \varepsilon$ falls within the segment $[0,1]$, then $\alpha_\varepsilon(n) \leq e^{-n\gamma^2/4}$. It is easy to check that $d/r \leq 1$. Hence for each $\varepsilon \in (0,1)$ the inequalities $d/(2r) < \gamma < d/r \leq 1$ hold for every $n \geq n_0 = \frac{2}{d}\log_\beta \varepsilon$.

Fix $\varepsilon = 1/8$. For every $n > \max(n_0, 16\left(\frac{r}{d}\right)^2(m\ln(1/\beta) + \ln 8))$ we have $-n\gamma^2/4 \leq m\ln(1/\beta) + \ln 8$ and thus $\alpha_{1/8}(n) \leq e^{-n\gamma^2/4} \leq \beta^m/8$. The substitution to (8.9) yields $p(n,m) \geq \beta^m/4$.

□

8.4 An Alternative Derivation

It is remarkable that a weaker form of the upper bound from Theorem 23 can be derived directly from the Incompressibility Property for Kolmogorov complexity. In this section we give this derivation.

Let $\mathfrak{G}$, S, β and B be as in Theorem 23. We will now prove independently that there is $C_K > 0$ such that for every positive integer n and $m \leq n$ there are at least $2^n - 2^{m-n}$ strings $\boldsymbol{x}$ of length n with the complexity

$$\mathcal{K}(\boldsymbol{x}) \geq nB - \frac{\ln 2}{\eta}m - C_K\frac{\log n}{\eta} \ .$$

Proof. Since the image of S under the transformation

$$(x', y') = \mathfrak{B}(x, y)$$

is convex and symmetric w.r.t. the straight line $x' = y'$, the straight line

$$\frac{x}{2} + \frac{y}{2} = \beta^B \ , \tag{8.11}$$

passing through the point $x' = y' = \beta^B$, must be a support line for the image. The set S thus lies 'north-east' to

$$\Pi = \{(x,y) \mid \frac{\beta^x}{2} + \frac{\beta^y}{2} = \beta^B\} \ , \tag{8.12}$$

which is the inverse image of the line (8.11), i.e., for every $(x,y) \in S$ there are $(\tilde{x}, \tilde{y}) \in \Pi$ such that $x \geq \tilde{x}$ and $y \geq \tilde{y}$.

Consider the *β-logarithmic* game with the loss function

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

or

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

Obviously, this game specifies complexity $\frac{\mathcal{K}^{\log}}{\log(1/\beta)}$ and has the set of predictions

$$P_\beta=\{(x,y)\mid \beta^x+\beta^y=1\} .$$

Since the equation defining Π in (8.12) may be rewritten as

$$\beta^{x-B-\frac{1}{\log\beta}}+\beta^{y-B-\frac{1}{\log\beta}}=1 ,$$

the curve Π is a shift of P_β, or $P_\beta=\Pi+(-B-\frac{1}{\log\beta},-B-\frac{1}{\log\beta})$. This implies that the set $S+(-B-\frac{1}{\log\beta},-B-\frac{1}{\log\beta})$ is contained within the set of superpredictions of the β-logarithmic-loss game. It now follows from Theorem 18 that there is a constant $C_1\geq 0$ such that for every string $\boldsymbol{x}$ the inequality

$$\mathcal{K}(\boldsymbol{x})-\left(B+\frac{1}{\log\beta}\right)|\boldsymbol{x}|\geq\frac{\mathcal{K}^{\log}(\boldsymbol{x})}{\log\frac{1}{\beta}}-C_1 \tag{8.13}$$

holds.

Let us apply the standard Incompressibility Property for the plain Kolmogorov complexity now. Since logarithmic-loss and plain Kolmogorov complexity K are related by the equation

$$|\mathcal{K}^{\log}(\boldsymbol{x})-\mathrm{K}(\boldsymbol{x})|\leq C_2\log|\boldsymbol{x}| ,$$

where C_2 does not depend on $\boldsymbol{x}$, for every n and $m\leq n$ there are at least 2^n-2^{n-m} strings of length n with the logarithmic-loss complexity

$$\mathcal{K}^{\log}(\boldsymbol{x})\geq n-m+1-C_2\log n .$$

The substitution to (8.13) yields

$$\begin{aligned}\mathcal{K}(\boldsymbol{x})&\geq\frac{n-m+1-C_2\log n}{\log\frac{1}{\beta}}+\left(B+\frac{1}{\log\beta}\right)n-C_1\\ &\geq nB-\frac{m}{\log\frac{1}{\beta}}-C_K\frac{\log n}{\log\frac{1}{\beta}} .\end{aligned}$$

□

Notation

This chapter summarises some notation used throughout the thesis.

AA	Aggregating Algorithm
$\mathfrak{B}_\beta$	the transformation $\mathfrak{B}_\beta(x,y) = (\beta^x, \beta^y)$, page 43.
$\mathfrak{C}(A)$	the convex hull of a set $A \subseteq \mathbb{R}^2$
$c(\beta)$	the multiplicative constant emerging in the Aggregating Algorithm, page 29
$h(p)$	generalised entropy, page 83
K	plain Kolmogorov complexity (cf. C in [LV97])
KP	prefix Kolmogorov complexity (cf. K in [LV97])
$\mathcal{K}^{\log}$	logarithmic-loss complexity, page 69
$\mathcal{K}^{\mathrm{sq}}$	square-loss complexity, page 69
log	logarithm to the base 2, i.e., $\log_2$
$\mathbb{N}$	the set of positive integers $\{1, 2, 3, \ldots\}$
$\lvert\boldsymbol{x}\rvert$	the length of a finite string $\boldsymbol{x}$
$\sharp_0\boldsymbol{x}$	the number of zeroes in a finite string $\boldsymbol{x} \in \mathbb{B}^*$, page 71
$\sharp_1\boldsymbol{x}$	the number of ones in a finite string $\boldsymbol{x} \in \mathbb{B}^*$, page 71

Bibliography[1]

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

[BDBK$^+$94] S. Ben-David, A. Borodin, R. Karp, G. Tardos, and A. Widgerson. On the power of randomisation in on-line algorithms. *Algorithmica*, 11:2–14, 1994.

[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.

[CO96] T. M. Cover and E. Ordentlich. Universal portfolios with side information. *IEEE Transactions on Information Theory*, 42(2), March 1996.

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

[Gal68] R. G. Gallager. *Information Theory and Reliable Communication.* John Wiley and Sons, INC, 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.

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

[LV93] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications.* Springer, New York, 1993.

[1] The author's papers on the subject of the thesis are enumerated in Chapter 1, Introduction.

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

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

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

[Rud74] W. Rudin. *Real and Complex Analysis.* TaTa McGrow–Hill Publishing Co., second edition, 1974.

[Rud76] W. Rudin. *Principles of Mathematical Analysis.* McGraw–Hill Book Company, third edition, 1976.

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

[Vov90] V. Vovk. Aggregating strategies. In M. Fulk and J. Case, editors, *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, pages 371–383, San Mateo, CA, 1990. Morgan Kaufmann.

[Vov98a] V. Vovk. Competitive on-line linear regression. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

[Vov98b] V. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56:153–173, 1998.

[Vov01] V. Vovk. Probability theory for the Brier game. *Theoretical Computer Science*, 261:57–79, 2001. Preliminary version in M. Li and A. Maruoka, editors, *Algorithmic Learning Theory*, vol. 1316 of *Lecture Notes in Computer Science*, pages 323–338.

[VV01] M. V. Vyugin and V. V. V'yugin. Non-linear inequalities between predictive and Kolmogorov complexities. In *Proc. 12th International Conference on Algorithmic Learning Theory — ALT '01*, 2001. To be published.

[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.

[V'y02] V. V. V'yugin. Sub-optimal measures of predictive complexity for absolute loss function. *Information and Computation*, 175:146–157, 2002.

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

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

Appendix A

The Infinity and the Extended Topology

It is often convenient in the theory of prediction with expert advice and the theory of predictive complexity to consider the extended real line $[-\infty, +\infty]$. The statements of many theorems become simpler if we treat $-\infty$ and $+\infty$ as numbers similar to those from $\mathbb{R} = (-\infty, +\infty)$. In fact, in most cases we need only $+\infty$. The logarithmic-loss game provides an example of how introducing $+\infty$ can simplify definitions and statements since it is essential that the loss function for this game assumes the value $+\infty$.

We use the following (more or less standard) conventions for performing arithmetic operations with $+\infty$ (cf. [Rud74], 1.21). For any $a \in \mathbb{R}$, the inequality $a < +\infty$ and the equality $a + (+\infty) = +\infty$ hold. If $a \in \mathbb{R}$ and $a > 0$, then $a \cdot (+\infty) = +\infty$ and

$$a^{+\infty} = \lim_{x \to +\infty} a^x = \begin{cases} 0 & \text{if} \quad a \in (0,1) \\ 1 & \text{if} \quad a = 1 \\ +\infty & \text{if} \quad a > 1 \ . \end{cases}$$

A slightly less obvious convention is to let $0 \cdot (+\infty) = 0$.

We also need some topological properties of $[-\infty, +\infty]$. The *extended topology* will be employed. By definition, an open subset of $[\infty, +\infty]$ is any of the sets U, $U \cup (a_1, +\infty]$, $[-\infty, a_2) \cup U$, or $[-\infty, a_2) \cup U \cup (a_1, +\infty]$, where U is an open subset of $\mathbb{R}$ and $a_1, a_2 \in \mathbb{R}$. In other words, the extended topology is generated by the base consisting of all open subsets of $\mathbb{R}$ and sets of the form $(a_1, +\infty]$ and $[-\infty, a_2)$, where $a_1, a_2 \in \mathbb{R}$. The extended topology of the

Cartesian product $[-\infty, +\infty]^2$ is introduced as the product of the extended topologies.

The continuity w.r.t. the extended topology is a very natural property. A function $f : M \to [-\infty, +\infty]$ is continuous at a point $x_0 \in M$ if and only if $\lim_{x \to x_0} f(x) = f(x_0)$ no matter whether $f(x_0)$ is finite or infinite.

Appendix B

The Legendre Transformation

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

Consider a function $f : \mathbb{R} \to [-\infty, +\infty]$. It is called *convex* if its epigraph $\{(x, y) \in \mathbb{R}^2 \mid y \geq f(x)\}$ is convex. The *conjugate* function $f^* : \mathbb{R} \to [-\infty, +\infty]$ to a convex function f is defined by

$$f^*(t) = \sup_{x \in \mathbb{R}} (xt - f(x)) \ . \tag{B.1}$$

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 for each real α the *level set* $L_\alpha = \{x \in \mathbb{R} \mid g(x) \leq \alpha\}$ is closed w.r.t. the standard topology of $\mathbb{R}$.

Figure B.1 provides an example. In the picture we have

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

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

Proposition 13 (see [RV73, Roc70]). *If $f : \mathbb{R} \to [-\infty, +\infty]$ is a proper convex function, the following properties hold:*

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

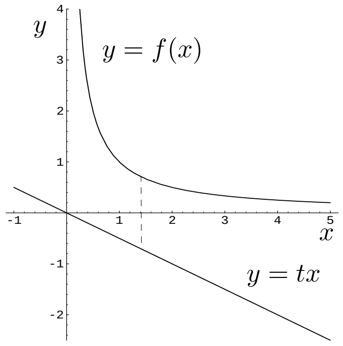

Figure B.1: Evaluation of the Legendre transformation

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

Conjugate functions have a number of interesting properties, e.g., we have $xy \leq f(x) + f^*(y)$ and $(f^*)' = (f')^{-1}$ where applicable; the latter of these equalities is a special case of $\partial(f^*) = (\partial f)^{-1}$, where ∂ refers to the *subdifferential.* The equality $xy = f(x) + f^*(y)$ holds if and only if $y \in \partial f(x)$. These properties may be employed to derive properties of expectations of predictive complexity though this investigation falls beyond the scope of this thesis.

Appendix C

Martingales

Let us start with a general definition of a *(super)martingale* from probability theory. Later we will adapt it to our special case.

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

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots \subseteq \mathcal{F} \ ,$$

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

- $\mathbf{E}_{\mathrm{Pr}}(|X_n|) < +\infty$, and

- $\mathbf{E}_{\mathrm{Pr}}(X_n \mid \mathcal{F}_{n-1}) = X_{n-1}$.

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

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

Proposition 14. *If non-negative random variables $Z_0, Z_1, Z_2, \ldots$ form a supermartingale w.r.t. $(\Omega, \mathcal{F}, \{\mathcal{F}\}_n, \Pr)$, then for every $c > 0$ and positive integer n we have*

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

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

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

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