The Weak Aggregating Algorithm and Weak Mixability

Yuri Kalnishkan Michael V. Vyugin[1]

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

Technical Report

CLRC — TR — 03 — 01

November 2003

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

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

Abstract

This paper resolves the problem of predicting as well as the best expert up to an additive term $o(n)$, where n is the length of the sequence. For the bounded games the paper introduces the Weak Aggregating Algorithm that allows us to obtain additive terms of the form $C\sqrt{n}$. A modification of the Weak Aggregating Algorithm that covers unbounded games is also described.

Contents

1 Introduction **3**

2 Preliminaries **4**
- 2.1 On-line Prediction 4
- 2.2 Expert Advice 5
- 2.3 Expert Advice in Mixable Games 5
  - 2.3.1 Geometric Interpretation of a Game 5
  - 2.3.2 The Result for Mixable Games 5

3 Weak Mixability **6**

4 Proof: "Only If" part **7**

5 Proof: "If" Part for Bounded Games **9**
- 5.1 The Value $c(\beta)$ 9
  - 5.1.1 Closure 9
  - 5.1.2 Oscillation 10
- 5.2 Convergence of $c(\beta)$ 10
  - 5.2.1 A Special Case 10
  - 5.2.2 The Case of $(0, 0)$-Games 12
  - 5.2.3 The Case of Arbitrary (C, D)-Games 13
- 5.3 Weak Aggregating Algorithm 14

6 Proof: "If" Part for Unbounded Games **16**
- 6.1 Approximating Unbounded Games with Bounded 16
- 6.2 Weak Aggregation for Unbounded Games 16

7 Computability Issues **18**

1 Introduction

This paper deals with the problem of prediction with expert advice. We consider the on-line prediction protocol, where outcomes $\omega_1, \omega_2, \ldots$ occur in succession while a prediction strategy tries to predict them. Before seeing an event ω_t the prediction strategy produces a prediction γ_t. We are interested in the binary case, where $\omega_1, \omega_2, \ldots \in \mathbb{B} = \{0, 1\}$.

We use a loss function $\lambda(\omega, \gamma)$ to measure the discrepancies between predictions and outcomes. A loss function and a prediction space (a set of possible predictions) Γ specify the game, i.e., a particular prediction environment. The performance of a learner $\mathfrak{S}$ w.r.t. a game is measured by the cumulative loss suffered on the sequence of outcomes $\omega_1, \omega_2, \ldots, \omega_n$

$$\mathrm{Loss}_{\mathfrak{S}}(\omega_1, \omega_2, \ldots, \omega_n) = \sum_{t=1}^{n} \lambda(\omega_t, \gamma_t) \ . \tag{1}$$

In the problem of prediction with expert advice we have N prediction strategies $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_N$ that try to predict elements of the same sequence. Their predictions become available to the merging prediction strategy $\mathfrak{M}$ every time before $\mathfrak{M}$ outputs its own. The goal of $\mathfrak{M}$ is to predict nearly as well as the best expert, i.e., to suffer loss that is little bigger than the smallest of the experts' losses.

This problem has been studied intensively; see, e.g., [CBFH$^+$97, HKW98]. In [Vov90, Vov98] the Aggregating Algorithm was proposed. It allows $\mathfrak{M}$ to achieve loss satisfying the inequality

$$\mathrm{Loss}_{\mathfrak{M}}(\omega_1, \omega_2, \ldots, \omega_n) \le c\, \mathrm{Loss}_{\mathcal{E}_i}(\omega_1, \omega_2, \ldots, \omega_n) + a \ln N \tag{2}$$

for all $i = 1 \ldots, N$ and all possible sequences of outcomes $\omega_1, \omega_2, \ldots, \omega_n$, $n = 1, 2, \ldots$, where the constants c and a are optimal and are specified by the game. Note that neither c nor a depends on n.

If we can take c equal to 1, the game is called mixable. It is possible to provide a geometrical characterization of mixable games in terms of the so called sets of superpredictions. The problem of prediction with expert advice and the Aggregating Algorithm are discussed in more detail in Sect. 2. The Aggregating Algorithm fully resolves the problem of predicting as well as the best expert up to an additive constant.

There are interesting games that are not mixable, e.g., the absolute loss game introduced in Sect. 2. The Aggregating Algorithm still works for some of such games, but it only allows us to achieve values of c greater than 1.

In this paper we take a different approach to non-mixable games. We fix $c = 1$ but consider $a(n)$ that grows as the length n of the sequence increases. We study the problem of predicting as well as the best expert up to $o(n)$ as $n \to +\infty$, where n is the length of the sequence. Sect. 3 introduces the corresponding concept of weak mixability. The main result of this paper, Theorem 1, shows that weak mixability is equivalent to a very simple geometric property of the set of superpredictions, namely, the convexity of its real part.

In Sect. 5 we introduce the Weak Aggregating Algorithm, which works for bounded loss functions. It is obtained by modifying the Aggregating Algorithm and it allows us to predict as well as the best expert up to an additive term of the form $C\sqrt{n}$, provided the real part of the set of superpredictions is convex.

If the game is not bounded, the Weak Aggregating Algorithm can be applied in a slightly different form to predict as well as the best expert up to $o(n)$.

2 Preliminaries

2.1 On-line Prediction

A *game* $\mathfrak{G}$ is a triple $\langle \mathbb{B}, \Gamma, \lambda \rangle$, where $\mathbb{B} = \{0, 1\}$ is an *outcome space*[1], Γ is a *prediction space*, and $\lambda : \mathbb{B} \times \Gamma \to [0, +\infty]$ is a *loss function*. We assume that Γ is a compact topological space and λ is continuous w.r.t. the extended topology of $[-\infty, +\infty]$. The *square-loss* game with $\Gamma = [0, 1]$ and $\lambda(\omega, \gamma) = (\omega - \gamma)^2$, the *absolute-loss* game with the same prediction space and $\lambda(\omega, \gamma) = |\omega - \gamma|$, and the *logarithmic* game with the same prediction space and

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

are examples of games. A slightly different example is provided by the *simple prediction game* with $\Gamma = \mathbb{B} = \{0, 1\}$ and

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

It is essential to allow λ to accept the value $+\infty$; this assumption is necessary in order to take into account the logarithmic game as well as other unbounded games. However we impose the following restriction: if $\lambda(\omega_0, \gamma_0) = +\infty$ for some $\omega_0 \in \mathbb{B}$ and $\gamma_0 \in \Gamma$, then there is a sequence $\gamma_n \in \Gamma$ such that $\gamma_n \to \gamma_0$ and $\lambda(\omega_0, \gamma_n) \to +\infty$ as $n \to +\infty$. This restriction allows us to exclude some degenerate cases and to simplify the statements of theorems.

A prediction strategy $\mathfrak{S}$ works according to the following protocol:

(1) FOR $t = 1, 2, \ldots$
(2) $\quad\mathfrak{S}$ chooses a prediction $\gamma_t \in \Gamma$
(3) $\quad\mathfrak{S}$ observes the actual outcome $\omega_t \in \Omega$
(4) END FOR

We say that over the first T trials, $\mathfrak{S}$ suffers the total loss

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

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

A prediction strategy is allowed to remember all previous outcomes. It can be oblivious and ignore the prehistory, but it does not have to. One can identify a prediction strategy with a function from $\mathbb{B}^*$ to Γ. Suppose that λ can be computed by an oracle. We assume that the oracle is capable of more than just outputting the values of λ, e.g., it can solve some simple inequalities involving λ. See Sect. 7 for more details.

If a function $\mathfrak{S} : \mathbb{B}^* \to \Gamma$ is computable by an algorithm that has access to such an oracle, we call the strategy computable; similarly, if the function is polynomial-time computable modulo the work of the oracle, we call the strategy polynomial-time computable. For all natural loss functions specified by simple analytical expressions, e.g., the loss functions specified above, this definition leads to the same natural concepts of a computable and polynomial-time computable strategy.

[1] In this paper we restrict ourselves to binary games. A more general definition is possible.

2.2 Expert Advice

In the problem of prediction with expert advice we have a pool of N experts $\mathcal{E}^{(1)}, \mathcal{E}^{(2)}, \ldots, \mathcal{E}^{(N)}$, which are working according to the foregoing protocol. On trial t they output predictions $\gamma_t^{(1)}, \gamma_t^{(2)}, \ldots, \gamma_t^{(N)}$. A *merging strategy* $\mathfrak{M}$ is allowed to observe the experts' prediction before outputting its own, i.e., it works according to the following protocol:

(1) `FOR` $t = 1, 2, \ldots$
(2) $\quad\mathcal{E}^{(1)}, \mathcal{E}^{(2)}, \ldots, \mathcal{E}^{(N)}$ `output predictions` $\gamma_t^{(1)}, \gamma_t^{(2)}, \ldots, \gamma_t^{(N)} \in \Gamma$
(3) $\quad\mathfrak{M}$ `chooses a prediction` $\gamma_t \in \Gamma$
(4) $\quad\mathfrak{M}$ `observes the actual outcome` $\omega_t \in \Omega$
(5) `END FOR`

The goal of the merging strategy is to suffer loss that is not much worse than the loss of the best expert. By the best expert after trial t we mean the expert that has suffered the smallest loss over t trials.

One may think of a merging strategy as of a function

$$\mathfrak{M} : \bigcup_{N=0}^{+\infty} \bigcup_{t=1}^{+\infty} \left(\mathbb{B}^{t-1} \times \left(\Gamma^N\right)^t\right) \to \Gamma \ . \tag{3}$$

Here N is the number of experts and t is the number of a trial; the information available to $\mathfrak{M}$ before making a prediction on trial t consists of $t-1$ previous outcomes and t arrays each consisting of N experts' predictions.

When we speak about computability, we assume that the algorithm computing $\mathfrak{M}$ receives experts' predictions as inputs. The experts do not have to be computable in any sense. The learner has no access to their internal "mechanics"; the only thing it knows about them is their predictions.

2.3 Expert Advice in Mixable Games

We introduce mixability as a geometric property of games and then show its relevance to the problem of prediction with expert advice.

2.3.1 Geometric Interpretation of a Game

Take a game $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$. We say that a pair $[0, s_1) \in (-\infty, +\infty]^2$ is a *superprediction* if there is $\gamma \in \Gamma$ such that the inequalities $\lambda(0, \gamma) \leq s_0$ and $\lambda(1, \gamma) \leq s_1$ hold. The set of superpredictions S is an important geometrical object characterizing the game.

Let $\beta \in (0, 1)$. Consider the transformation $\mathfrak{B}_\beta$ specified by $\mathfrak{B}_\beta(x, y) = (\beta^x, \beta^y)$. This transformation is a homeomorphism of $[0, +\infty]^2$ onto $[0, 1]^2$ (we consider the extended topology on $[-\infty, +\infty]$ and the induced topology on the Cartesian product $[-\infty, +\infty]^2$).

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

2.3.2 The Result for Mixable Games

The Aggregating Algorithm (AA) was first introduced in [Vov90]. It provides a very general solution to the problem of prediction with expert advice. The AA is polynomial-time computable modulo the oracle computing λ (see Sect. 7).

The following proposition connects the geometrical definition of mixability with the problem of prediction with expert advice. The AA accepts a parameter $\beta \in (0,1)$.

Proposition 1 ([Vov90, Vov98]). *For every game $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$*

(i) if $\mathfrak{G}$ is β-mixable for some $\beta \in (0,1)$, then for every finite set of experts $\mathcal{E}^{(1)}, \mathcal{E}^{(2)}, \dots, \mathcal{E}^{(N)}$ ($N = 1, 2, \dots$) the merging strategy $\mathfrak{M}$ following the AA with the parameter β achieves loss satisfying

$$\mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{M}}(\omega_1, \omega_2, \dots, \omega_n) \leq \mathrm{Loss}_{\mathcal{E}^{(i)}}(\omega_1, \omega_2, \dots, \omega_n) + \frac{\ln N}{\ln(1/\beta)}$$

for every $i = 1, 2, \dots, N$ and every finite sequence $\omega_1, \omega_2, \dots, \omega_n \in \mathbb{B}$, $n = 1, 2, \dots$;

(ii) if there is a merging strategy $\mathfrak{M}$ and a positive constant a such that for every pair of experts $\mathcal{E}^{(1)}$ and $\mathcal{E}^{(2)}$ the strategy $\mathfrak{M}$ achieves

$$\mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{M}}(\omega_1, \omega_2, \dots, \omega_n) \leq \mathrm{Loss}_{\mathcal{E}^{(i)}}(\omega_1, \omega_2, \dots, \omega_n) + a$$

for all $i = 1, 2$ and every finite sequence $\omega_1, \omega_2, \dots, \omega_n \in \mathbb{B}$, $n = 1, 2, \dots$, then $\mathfrak{G}$ is mixable.

The proposition implies that an equivalent definition of mixability can be given in terms of prediction with expert advice. The AA fully resolves the problem of predicting as well as the best expert up to an additive constant. The class of games that allow predicting as well as the best expert up to an additive constant is exactly the class of mixable games.

It can be shown that the square-loss and the logarithmic games are mixable. However the absolute-loss game is not.

3 Weak Mixability

For non-mixable games it is not possible to predict as well as the best expert up to an additive constant. Let us relax this requirement and ask whether it is possible to predict as well as the best expert up to a larger term.

On average loss of a strategy grows linearly in the length of the sequence[2]. Therefore all terms of slower growth can be considered small as compared to loss. This motivates the following definition.

A game $\mathfrak{G}$ is *weakly mixable* if there is a function $f : \mathbb{N} \to \mathbb{R}$ such that $f(n) = o(n)$ as $n \to +\infty$ and a merging strategy $\mathfrak{M}$ such that for every finite set of experts $\mathcal{E}^{(1)}, \mathcal{E}^{(2)}, \dots, \mathcal{E}^{(N)}$ ($N = 1, 2, \dots$) the inequality

$$\mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{M}}(\omega_1, \omega_2, \dots, \omega_n) \leq \mathrm{Loss}_{\mathcal{E}^{(i)}}(\omega_1, \omega_2, \dots, \omega_n) + f(n)$$

holds for all $i = 1, 2, \dots, N$ and every finite sequence $\omega_1, \omega_2, \dots, \omega_n \in \mathbb{B}$, $n = 1, 2, \dots$.

The following theorem is the main result of the paper.

Theorem 1. *A game $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$ with the set of superpredictions S is weakly mixable if and only if the real part of S, the set $S \cap \mathbb{R}^2$, is convex.*

The merging strategy in the definition of weak mixability can be chosen to be polynomial-time computable modulo the oracle computing λ (see Sect. 7).

The rest of the paper contains the proof of the theorem.

[2] This statement can be made precise by considering sequences $\xi_1, \xi_2, \dots, \xi_n$, where ξ_i are results of independent Bernoulli trials. We do not discuss it in more detail in this paper though

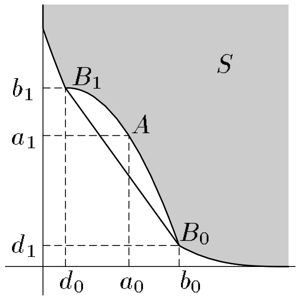

Figure 1 A set of superpredictions that is not convex

4 Proof: “Only If” part

Here we will derive a statement that is slightly stronger than that required by Theorem 1.

Theorem 2. *If a game $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$ has the set of superpredictions S such that its real part $S \cap \mathbb{R}^2$ is not convex, then there are strategies $\mathfrak{S}_0$ and $\mathfrak{S}_1$ such that for any strategy $\mathfrak{S}$ there is a sequence $\omega_n \in \mathbb{B}$, $n = 1, 2, \ldots$, such that*

$$\max_{i=1,2} \left(\mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{S}}(\omega_1, \omega_2, \ldots, \omega_n) - \mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{S}_i}(\omega_1, \omega_2, \ldots, \omega_n) \right) = \Omega(n) \ . \tag{4}$$

If the loss function is computable, the strategies can be chosen to be computable.

Proof. There are points $B_0, B_1 \in S \cap \mathbb{R}^2$ such that the line segment connecting them is not a subset of S. It follows from the definition of S that the segment is neither vertical nor horizontal. Since λ is continuous, S is closed and thus there is a point on the segment such that a non-empty open vicinity of that point does not intersect S.

Let $B_0 = (b_0, d_1)$ and $B_1 = (d_0, b_1)$ (see Fig. 1). For the sake of being definite, assume that $b_0 > d_0$. There is a point $A = (a_0, a_1) \in S$, where $a_0 \in (d_0, b_0)$ and $a_1 \in (d_1, b_1)$, such that A lies above the segment but the intersection $([0, a_0] \times [0, a_1]) \cap S$ consists of only one point A. The fact that A lies above the segment can be expressed as follows. Let $\alpha_0 x + \alpha_1 y = \rho$, where $\alpha_0, \alpha_1, \rho > 0$, be an equation of the straight line passing through B_0 and B_1, i.e.,

$$\alpha_0 b_0 + \alpha_1 d_1 = \rho \tag{5}$$

and

$$\alpha_0 d_0 + \alpha_1 b_1 = \rho \ . \tag{6}$$

Then

$$\alpha_0 a_0 + \alpha_1 a_1 = \rho + \delta \ , \tag{7}$$

where $\delta > 0$.

We now proceed to constructing the strategies $\mathfrak{S}_0$ and $\mathfrak{S}_1$. There are predictions $\gamma_0, \gamma_1 \in \Gamma$ such that

$$\lambda(0, \gamma_0) \le b_0 \qquad \lambda(0, \gamma_1) \le d_0 \tag{8}$$

$$\lambda(1, \gamma_0) \le d_1 \qquad \lambda(1, \gamma_1) \le b_1 \ . \tag{9}$$

Let $\mathfrak{S}_0$ be the strategy that always predicts γ_0, no matter what outcomes actually occur. Similarly, let $\mathfrak{S}_1$ be the strategy that always predicts γ_1.

We will denote finite sequences of bits by bold letters, e.g., $\boldsymbol{x} = \omega_1 \ldots \omega_n \in \mathbb{B}^n$. Let $|\boldsymbol{x}|$ be the length of $\boldsymbol{x}$, i.e., the total number of bits in $\boldsymbol{x}$. We will denote the number of zeros in a sequence $\boldsymbol{x}$ by $\sharp_0 \boldsymbol{x}$ and the number of ones by $\sharp_1 \boldsymbol{x}$. Obviously, $\sharp_0 \boldsymbol{x} + \sharp_1 \boldsymbol{x} = |\boldsymbol{x}|$ for every $\boldsymbol{x} \in \mathbb{B}^*$.

The definition of $\mathfrak{S}_0$ and $\mathfrak{S}_1$ imply that

$$\mathrm{Loss}_{\mathfrak{S}_0}(\boldsymbol{x}) \le b_0 \sharp_0 \boldsymbol{x} + d_1 \sharp_1 \boldsymbol{x} \tag{10}$$

and

$$\mathrm{Loss}_{\mathfrak{S}_1}(\boldsymbol{x}) \le d_0 \sharp_0 \boldsymbol{x} + b_1 \sharp_1 \boldsymbol{x} \ . \tag{11}$$

Now consider an arbitrary strategy $\mathfrak{S}$. We will construct a sequence $\boldsymbol{x}_n \in \mathbb{B}^n$, $n = 1, 2, \ldots$, such that

$$\max_{i=0,1} \left(\mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{S}}(\boldsymbol{x}_n) - \mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{S}_i}(\boldsymbol{x}_n) \right) = \Omega(n) \ . \tag{12}$$

The sequence is constructed by induction. Suppose that $\boldsymbol{x}_n$ has been constructed (we take $\boldsymbol{x}_0 = \Lambda$, the empty string). Let γ be the prediction output by $\mathfrak{S}$ on the $(n+1)$-th trial, provided the previous outcomes were elements constituting the strings $\boldsymbol{x}_n$ in the correct order. Since $S \subseteq ([a_0, +\infty] \times [0, +\infty]) \cup ([0, +\infty] \times [a_1, +\infty])$, the point $(\lambda(0,\gamma), \lambda(1,\gamma))$ belongs to at least one of these quadrants. In other terms, at least one of the inequalities

$$\begin{aligned} \lambda(0,\gamma) &\ge a_0, \\ \lambda(1,\gamma) &\ge a_1 \end{aligned}$$

holds. If the former is true, we let $\boldsymbol{x}_{n+1} = \boldsymbol{x}_n 0$; otherwise let $\boldsymbol{x}_{n+1} = \boldsymbol{x}_n 1$. The construction implies

$$\mathrm{Loss}_{\mathfrak{S}}(\boldsymbol{x}_n) \ge a_0 \sharp_0 \boldsymbol{x}_n + a_1 \sharp_1 \boldsymbol{x}_n \ . \tag{13}$$

Suppose that (12) does not hold, i.e.,

$$\max_{i=0,1} \left(\mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{S}}(\boldsymbol{x}_n) - \mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{S}_i}(\boldsymbol{x}_n) \right) = o(n)$$

as $n \to +\infty$. Combining this with (10), (11), and (13) yields

$$a_0 \sharp_0 \boldsymbol{x}_n + a_1 \sharp_1 \boldsymbol{x}_n \le b_0 \sharp_0 \boldsymbol{x}_n + d_1 \sharp_1 \boldsymbol{x}_n + o(n), \tag{14}$$

$$a_0 \sharp_0 \boldsymbol{x}_n + a_1 \sharp_1 \boldsymbol{x}_n \le d_0 \sharp_0 \boldsymbol{x}_n + b_1 \sharp_1 \boldsymbol{x}_n + o(n) \ . \tag{15}$$

Let $\tilde{a}_0, \tilde{a}_1, \tilde{b}_0, \tilde{b}_1 > 0$ be such that

$$\begin{aligned} a_0 &= d_0 + \tilde{a}_0 & \qquad b_0 &= d_0 + \tilde{b}_0 \\ a_1 &= d_1 + \tilde{a}_1 & \qquad b_1 &= d_1 + \tilde{b}_1 \ . \end{aligned}$$

The substitution into the system (14), (15) reduces it to

$$\tilde{a}_0 \sharp_0 \boldsymbol{x}_n + \tilde{a}_1 \sharp_1 \boldsymbol{x}_n \le \tilde{b}_0 \sharp_0 \boldsymbol{x}_n + o(n), \tag{16}$$

$$\tilde{a}_0 \sharp_0 \boldsymbol{x}_n + \tilde{a}_1 \sharp_1 \boldsymbol{x}_n \le \tilde{b}_1 \sharp_1 \boldsymbol{x}_n + o(n) \ . \tag{17}$$

Let $\tilde{\rho} = \rho - (\alpha_0 d_0 + \alpha_1 d_1)$. It is easy to check that $\tilde{\rho} \geq 0$. The substitution into (5), (6), and (7) yields

$$\alpha_0 \tilde{b}_0 = \tilde{\rho},$$
$$\alpha_1 \tilde{b}_1 = \tilde{\rho},$$
$$\alpha_0 \tilde{a}_0 + \alpha_1 \tilde{a}_1 = \tilde{\rho} + \delta \ .$$

Multiplying (16) by $\alpha_0/\tilde{a}_1$, multiplying (17) by $\alpha_1/\tilde{a}_0$, and then adding them together yield

$$\frac{\delta}{\tilde{a}_1} \sharp_0 \boldsymbol{x}_n + \frac{\delta}{\tilde{a}_0} \sharp_1 \boldsymbol{x}_n \leq o(n) \tag{18}$$

This is not possible since at least one of the values $\sharp_0 \boldsymbol{x}_n$ or $\sharp_1 \boldsymbol{x}_n$ accepts a value greater than or equal to $n/2$ for infinitely many positive integers n. □

5 Proof: "If" Part for Bounded Games

A game $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$ is *bounded* if and only if λ is bounded, i.e., there is $L \in (0, +\infty)$ such that $\lambda(\omega, \gamma) \leq L$ for each $\omega = 0, 1$ and $\gamma \in \Gamma$.

Remark 1. Let S be the set of superpredictions w.r.t. $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$. If $\mathfrak{G}$ is bounded, then the boundary ∂S consists of two half-lines and, perhaps, a bounded curve. Conversely, if S, the set of superpredictions for a game $\mathfrak{G}$, has this property, then there is a bounded game $\mathfrak{G}'$ that has the same set of superpredictions S. For example, it can be the game $\mathfrak{G}' = \langle \mathbb{B}, \Gamma, \lambda_L \rangle$, where $\lambda'(\omega, \gamma) = \max(\lambda(\omega, \gamma), L)$ and L is sufficiently large.

We will construct the Weak Aggregating Algorithm (WAA) that works for bounded games and has the following property.

Theorem 3. *If a game $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$ with the set of superpredictions S is bounded and the set $S \cap \mathbb{R}^2$, the real part of S, is convex, then there is a constant $C > 0$ such that for every finite set of experts $\mathcal{E}^{(1)}, \mathcal{E}^{(2)}, \ldots, \mathcal{E}^{(N)}$ ($N = 1, 2, \ldots$) the merging strategy $\mathfrak{M}$ following the WAA achieves loss satisfying*

$$\mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{M}}(\omega_1, \omega_2, \ldots, \omega_n) \leq \mathrm{Loss}_{\mathcal{E}^{(i)}}(\omega_1, \omega_2, \ldots, \omega_n) + (C \ln N)\sqrt{n}$$

for every $i = 1, 2, \ldots, N$ and every finite sequence $\omega_1, \omega_2, \ldots, \omega_n \in \mathbb{B}$.

5.1 The Value $c(\beta)$

This subsection discusses the value $c(\beta)$. It appears naturally in the theory of the AA and we need it to introduce the WAA.

5.1.1 Closure

Let $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$ be a game with the set of superpredictions S. Take $\beta \in (0, 1)$ and consider the convex hull of the set $\mathfrak{B}_\beta(S)$. We denote it by $\mathfrak{C}(\mathfrak{B}_\beta(S))$. Consider the inverse image $\mathfrak{B}_\beta^{-1}(\mathfrak{C}(\mathfrak{B}_\beta(S)))$.

There is a different way to describe the resulting set. For every set $M \subseteq [-\infty, +\infty]^2$ and a pair $a, b \in \mathbb{R}$, by $M + (a, b)$ denote the shift of the set M by the vector (a, b), i.e., the set $\{(x + a, y + b) \mid (x, y) \in M\}$. If $L \subseteq [-\infty, +\infty]^2$, let the *$M$-closure of L* be the intersection of all shifts of M that contain L, i.e.,

$$\mathrm{cl}_M(L) = \bigcap_{a, b \in \mathbb{R} \,:\, M + (a, b) \supseteq L} M + (a, b) \ .$$

Let $B_\beta = \{(x,y) \in [0,+\infty]^2 \mid \beta^x + \beta^y \geq 1\}$ be the "supergraph" of the curve $\beta^x + \beta^y = 1$. It is easy to see that

$$\mathfrak{B}_\beta^{-1}(\mathfrak{C}(\mathfrak{B}_\beta(S))) = \mathrm{cl}_{B_\beta}(S) \ ,$$

i.e., the set we are interested in is the B_β-closure of the set of superpredictions S.

Remark 2. An intersection of convex sets is convex. It is easy to check that the set $B_\beta \cap \mathbb{R}^2$ is convex. This implies that the set $\mathrm{cl}_{B_\beta}(S) \cap \mathbb{R}^2$ is always convex.

If $\mathfrak{G}$ is β-mixable, then the set $\mathfrak{B}_\beta(S)$ is convex. Hence $\mathfrak{C}(\mathfrak{B}_\beta(S)) = \mathfrak{B}_\beta(S)$ and $\mathrm{cl}_{B_\beta}(S) = S$. We have shown that for every mixable game the real part of the set of superpredictions is convex.

5.1.2 Oscillation

First, we define oscillations for one-dimensional sets. Let $M \subseteq [0,+\infty)$. If $M = \varnothing$, let $\mathrm{osc}\, M = 1$. If $M = \{0\}$, let $\mathrm{osc}\, M = 1$. Otherwise let $\mathrm{osc}\, M = \sup M / \inf M$, where by definition $a/0 = +\infty$ for all $a \in (0,+\infty]$.

Now let $M \subseteq [0,+\infty)^2$. Take $\theta \in [0,\pi/2]$. Put $\mathrm{osc}_\theta\, M$ to be the oscillation of the section of M by the line that passes through the origin and has the slope θ, i.e.,

$$\mathrm{osc}_\theta\, M = \mathrm{osc}\{t \in [0,+\infty) \mid (t\cos\theta, t\sin\theta) \in M\} \ .$$

The oscillation of M is the supremum of all oscillations, i.e.,

$$\mathrm{osc}\, M = \sup_{\theta \in [0,\pi/2]} \mathrm{osc}_\theta\, M$$

(cf. the definition from [Vov90]).

Let $\beta \in (0,1)$ and $\theta \in [0,\pi/2]$. By $c(\beta,\theta,\mathfrak{G})$ we denote $\mathrm{osc}_\theta \left(\mathrm{cl}_{B_\beta}(S) \setminus S\right)$ and by $c(\beta,\Gamma)$ we denote $\mathrm{osc}\left(\mathrm{cl}_{B_\beta}(S) \setminus S\right)$, where B_β is as above. Obviously, $c(\beta,\mathfrak{G}) = \sup_{\theta\in[0,\pi/2]} c(\beta,\theta,\mathfrak{G})$. If it is clear from the context what game we are referring to, the argument $\mathfrak{G}$ will be omitted.

5.2 Convergence of $c(\beta)$

It is easy to see that oscillations are always greater than or equal to 1. In this subsection we show that for every finite game $\mathfrak{G}$ the value $c(\beta)$ converges to 1 as $\beta \to 1$ and determine the rate of convergence.

Theorem 4. *For every bounded game $\mathfrak{G}$ there are positive constants c and ε such that $c(\beta) \leq 1 + c\ln(1/\beta)$ for all $\beta \in [1-\varepsilon, 1)$.*

The theorem essentially states that $c(\beta) = 1 + O(\ln(1/\beta))$ as $\beta \to 1-$. Since $c(\beta)$ is completely specified by the set of superpredictions S, the theorem remains true for the class of games such that the boundary ∂S consists of two half-lines and, perhaps, a bounded curve (see Remark 1).

5.2.1 A Special Case

We start by evaluating $c(\beta)$ for some particular games.

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

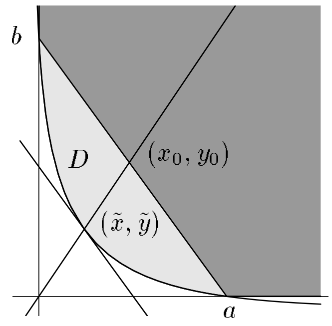

Figure 2 The drawing for Lemma 1

Lemma 1. *For all $a, b > 0$, we have*

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

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

$$\beta^x(1-\beta^b) + \beta^y(1-\beta^a) = 1 - \beta^{a+b} \tag{19}$$

is the shift of $\beta^x + \beta^y = 1$ passing through $(a, 0)$ and $(0, b)$. It is easy to see that$D = \left(\mathrm{cl}_{B_\beta} S^{a,b}_{\text{abs}}\right) \setminus S^{a,b}_{\text{abs}}$ is bounded by the line segment connecting the points $(a, 0)$ and $(0, b)$ and the arc of curve (19) between these points.

Let us evaluate $\operatorname{osc} D$. Consider the point $(\tilde{x}, \tilde{y})$ where the tangent to curve (19) is parallel to the chord connecting $(a, 0)$ and $(0, b)$. It follows from the Thales theorem of plane geometry that $\operatorname{osc}_\theta D$ achieves its maximum on θ_0 such that the half-line $l_{\theta_0} = \{(t \cos\theta_0, t \sin\theta_0) \mid t \geq 0\}$ passes through $(\tilde{x}, \tilde{y})$.

Figure 2 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 (19) is

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

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_0} with the line segment connecting $(a, 0)$ and $(0, b)$. Simple considerations from plane geometry imply that $\operatorname{osc}_{\theta_0} D = x_0/\tilde{x}$. It is easy to check that

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

The substitutions complete the proof of the lemma. □

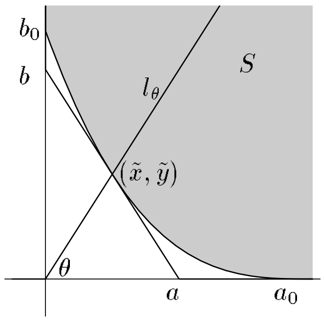

Figure 3 The drawing for the proof of convergence for the case of a $(0,0)$-game

5.2.2 The Case of $(0,0)$-Games

Let $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$ be a game with the set of superpredictions S. We say that $\mathfrak{G}$ is a (C,D)-game, where $C, D \geq 0$, if $C = \inf\{x \mid \exists y \in \mathbb{R} : (x,y) \in S\}$ and $D = \inf\{y \mid \exists x \in \mathbb{R} : (x,y) \in S\}$. Clearly, S is a subset of the quadrant $[C,+\infty]\times[D,+\infty]$ and this quadrant is the intersection of all quadrants that contain S.

Let us prove Theorem 4 for the case of a bounded $(0,0)$-game $\mathfrak{G}$ with the set of superpredictions S.

Since the game is bounded, the boundary of S consists of two half-lines and perhaps a bounded curve. The case $S = [0,+\infty]^2$ is trivial; therefore we assume that the half lines are $[a_0,+\infty] \times \{0\}$ and $\{0\} \times [b_0,+\infty]$, where $0 < a_0, b_0 < +\infty$.

Take $\theta \in (0, \pi/2)$. Let $(\tilde{x}, \tilde{y})$ be the point where the half-line with the slope θ first meets S, i.e., $(\tilde{x}, \tilde{y}) = l_\theta \cap \partial S$, where $l_\theta = \{(t\cos\theta, t\sin\theta) \mid t \geq 0\}$. Since $S \cap \mathbb{R}^2$ is convex, it has a tangent at $(\tilde{x}, \tilde{y})$ (see, e.g., [Egg58], Th. 8). One can assume that this tangent is neither horizontal nor vertical. Let $(a,0)$ and $(0,b)$ be the points where the tangent cuts the coordinate axes. Clearly, $0 < a \leq a_0$ and $0 < b \leq b_0$.

Consider the game $\mathfrak{G}^{a,b}_{\text{abs}}$ with the set of superpredictions $S^{a,b}_{\text{abs}}$. It is easy to see that for every $\beta \in (0,1)$

$$\begin{aligned} c(\beta, \theta, \mathfrak{G}) &\leq c(\beta, \theta, \mathfrak{G}^{a,b}_{\text{abs}}) \\ &\leq c(\beta, \mathfrak{G}^{a,b}_{\text{abs}}) \ . \end{aligned}$$

Thus

$$\begin{aligned} c(\beta, \mathfrak{G}) &= \sup_{\theta \in (0,\pi/2)} c(\beta, \theta, \mathfrak{G}) \\ &\leq \sup_{a \in (0,a_0], b \in (0,b_0]} c(\beta, \mathfrak{G}^{a,b}_{\text{abs}}) \ . \end{aligned}$$

It remains to consider the behavior of the function

$$g(a,b,t) = \frac{-abt}{a \ln \frac{a(1-e^{-t(a+b)})}{(a+b)(1-e^{-ta})} + b \ln \frac{b(1-e^{-t(a+b)})}{(a+b)(1-e^{-tb})}}$$

for small positive t and positive bounded a and b.

It is easy to check that

$$g(a,b,t) = 1 + \frac{a+b}{8} t + O\left(t^2\right)$$

as $t \to 0$. The O term is uniform over $a \in [0,a_0]$, $b \in [0,b_0]$, provided $a_0, b_0 < +\infty$.

This proves Theorem 4 for the case of bounded $(0,0)$-games.

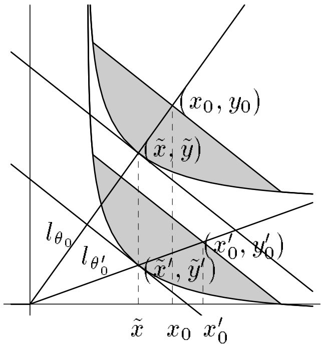

Figure 4 The shift in the proof of Lemma 2

5.2.3 The Case of Arbitrary (C, D)-Games

It remains to consider an arbitrary bounded (C, D)-game $\mathfrak{G}$ $(C, D \geq 0)$. Let S be the set of superpredictions for $\mathfrak{G}$. The case $S = [C, +\infty] \times [D, +\infty]$ is trivial; therefore we assume that the half lines in the boundary of S are $[a_0, +\infty] \times \{D\}$ and $\{C\} \times [b_0, +\infty]$, where $C < a_0 < +\infty$ and $D < b_0 < +\infty$.

Arguing as above, take $\theta \in (0, \pi/2)$ and let $(\tilde{x}, \tilde{y}) = \{(t \cos\theta, t \sin\theta) \mid t \geq 0\} \cap \partial S$. If $\tilde{x} = C$ or $\tilde{y} = D$, i.e., the half-line intersects ∂S at one of two half-lines, then $c(\beta, \theta, \mathfrak{G}) = 1$ for all $\beta \in (0, 1)$.

Assume this is not the case. There is a tangent to S at $(\tilde{x}, \tilde{y})$ that is neither vertical nor horizontal. Let (a, D) and (C, b) be the points where the tangent intersects the half-lines. We have $C \leq a \leq a_0$ and $D \leq b \leq b_0$. Consider the game $\mathfrak{G}_\theta$ with the set of superpredictions $S_\theta = \{(x, y) \mid x \in [C, +\infty], y \in [D, +\infty]$, and $(b_0 - D)x + (a_0 - C)y \geq a_0 b_0 - CD\}$, i.e., the S_θ is bounded by half-lines $[a, +\infty] \times \{D\}$ and $\{C\} \times [b, +\infty]$ and by the line segment connecting (a, D) with (C, b). Clearly, $c(\beta, \theta, \mathfrak{G}) \leq c(\beta, \mathfrak{G}_\theta)$. We could proceed by calculating $c(\beta, \mathfrak{G}_\theta)$. Instead, we will prove the following lemma.

Lemma 2. *Let $\mathfrak{G}$ be a game with the set of superpredictions $S = S^{a,b}_{\mathrm{abs}} + (C, D)$, where $a, b, C, D \geq 0$. Then $c(\beta, \mathfrak{G}) \leq c(\beta, \mathfrak{G}^{a,b}_{\mathrm{abs}})$.*

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

Arguing as in the proof of Lemma 1, the set $(\mathrm{cl}_{B_\beta} S) \setminus S$ is bounded by the line segment connecting the points $(C + a, D)$ and $(C, D + b)$ and the arc of the curve that is a shift of $\beta^x + \beta^y = 1$ connecting the same points (on Fig. 4 this set is grey).

The value $c(\beta, \theta, \mathfrak{G})$ achieves its maximum on θ_0 such that the half-line $l_{\theta_0} = \{(t \cos\theta_0, t \sin\theta_0) \mid t \geq 0\}$ passes through $(\tilde{x}, \tilde{y})$ on he arc, where the point $(\tilde{x}, \tilde{y})$ is such that there is a tangent to the arc passing through $(\tilde{x}, \tilde{y})$ and parallel to the chord connecting $(C + a, D)$ and $(C, D + b)$. If (x_0, y_0) is the intersection of l_{θ_0} with the chord, then $c(\beta, \mathfrak{G}) = x_0/\tilde{x}$.

Let $\mathfrak{G}'$ be a game with the set of superpredictions $S' = S - (0, D)$. The set $\mathrm{cl}_{B_\beta}(S') \setminus S'$ is a shift of the set $\mathrm{cl}_{B_\beta}(S) \setminus S$. Let θ'_0, $l_{\theta'_0}$, $(\tilde{x}', \tilde{y}')$, and (x'_0, y'_0) be defined for for $\mathfrak{G}'$ in the same fashion as the corresponding objects for $\mathfrak{G}$. We have $c(\beta, \mathfrak{G}') = x'_0/\tilde{x}'$ (see Fig. 4, where the set $\mathrm{cl}_{B_\beta}(S') \setminus S'$ is grey too).

It is easy to see that $\tilde{x} = \tilde{x}'$. Since $\theta'_0 \leq \theta_0$, we get $x'_0 \geq x_0$. This implies $x_0/\tilde{x} \leq x'_0/\tilde{x}'$ and thus $c(\beta, \mathfrak{G}) \leq c(\beta, \mathfrak{G}')$.

It remains to show that $c(\beta, \mathfrak{G}') \leq c(\beta, \mathfrak{G}^{a,b}_{\mathrm{abs}})$. This can be done by a similar argument applied to a shift along the x-axis and the equality $c(\beta, \mathfrak{G}') = y'_0/\tilde{y}'$. $\square$

5.3 Weak Aggregating Algorithm

We can now formulate the Weak Aggregating Algorithm (WAA). Let $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$ be a game and let N be the number of experts. Take a sequence $\beta_t = e^{-1/\sqrt{t}}$, $t = 1, 2, \ldots$.

We describe the WAA using pseudo-code. The WAA accepts N parameters $q_1, q_2, \ldots, q_N$, where $q_i \geq 0$, $i = 1, 2, \ldots, N$. Assume that at least one of q_i is strictly positive.

(1) $l_1^{(i)} := 0$, $i = 1, 2, \ldots, N$
(2) `FOR` $t = 1, 2, \ldots$
(3) $\quad w_t^{(i)} := q_i \beta_t^{l_t^{(i)}}$, $i = 1, 2, \ldots, N$
(4) $\quad p_t^{(i)} := \frac{w_t^{(i)}}{\sum_{j=1}^N w_t^{(i)}}$, $i = 1, 2, \ldots, N$
(5) $\quad$ `read experts' predictions` $\gamma_t^{(1)}, \gamma_t^{(2)}, \ldots, \gamma_t^{(N)}$
(6) $\quad g_0 := \log_{\beta_t} \sum_{j=1}^N \beta_t^{\lambda\left(0, \gamma_t^{(j)}\right)} p_t^{(j)}$ `and` $g_1 := \log_{\beta_t} \sum_{j=1}^N \beta_t^{\lambda\left(1, \gamma_t^{(j)}\right)} p_t^{(j)}$
(7) $\quad$ `output` $\gamma_t \in \Gamma$ `such that` $\lambda(0, \gamma_t) \leq c(\beta_t) g_0$ `and` $\lambda(1, \gamma_t) \leq c(\beta_t) g_1$
(8) $\quad$ `observe` ω_t
(9) $\quad l_{t+1}^{(i)} := l_t^{(i)} + \lambda\left(\omega_t, \gamma_t^{(i)}\right)$, $i = 1, 2, \ldots, N$
(10) `END FOR`

The variable l_i stores the loss of the i-th expert $\mathcal{E}^{(i)}$, i.e., after trial t we have $l_t^{(i)} = \mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(\omega_1, \omega_2, \ldots, \omega_t)$.

Note that it is sufficient to have only one set of variables $p^{(i)}$, $i = 1, 2, \ldots, N$, one set of variables $w^{(i)}$, $i = 1, 2, \ldots, N$, and one set of variables $l^{(i)}$, $i = 1, 2, \ldots, N$ to save memory. The subscript t has been added in order to simplify referring to these variables in the proofs below.

Since the game is bounded, there is $L \in (0, +\infty)$ such that $\lambda(\omega, \gamma) \leq L$ for all $\omega = 0, 1$, $\gamma \in \Gamma$.

In order to obtain a bound on the performance of WAA, we need the following lemma.

Lemma 3. *If $\mathfrak{G}$ is bounded, $q_i' = q_i / \sum_{j=1}^N q_j$ for all $i = 1, 2, \ldots, N$, and $\mathfrak{M}$ is a merging strategy following the WAA (with (7')), then for every $t = 1, 2, \ldots$ we get*

$$\beta_t^{\mathrm{Loss}_{\mathfrak{M}}^{\mathfrak{G}}(\omega_1, \ldots, \omega_t)} \geq \beta_t^{L \sum_{k=1}^t \left(1 - \frac{1}{c(\beta_k)}\right)} \sum_{i=1}^N q_i' \beta_t^{\mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(\omega_1, \ldots, \omega_t)} \tag{20}$$

in the notation introduced above.

Proof. The proof is by induction on t. Let us assume that (20) holds and then derive the corresponding inequality for the step $t + 1$.

The function x^α, where $0 < \alpha < 1$, is increasing in x, $x \geq 0$. If is also concave in x, $x \geq 0$. For every set of weights $p_i \in [0, 1]$, $i = 1, \ldots, n$ such that $\sum_{i=1}^n p_i = 1$ and every array of $x_i \geq 0$, $i = 1, \ldots, n$, we get

$$\left(\sum_{i=1}^n p_i x_i\right)^\alpha \geq \sum_{i=1}^n p_i x_i^\alpha \ .$$

Therefore (20) implies

$$\beta_{t+1}^{\mathrm{Loss}_{\mathfrak{M}}^{\mathfrak{G}}(\omega_1,\dots,\omega_t)} = \left(\beta_t^{\mathrm{Loss}_{\mathfrak{M}}^{\mathfrak{G}}(\omega_1,\dots,\omega_t)}\right)^{\log_{\beta_t}\beta_{t+1}} \tag{21}$$

$$\geq \left(\sum_{i=1}^N q_i'\beta_t^{\mathrm{Loss}_{\mathcal{E}(i)}^{\mathfrak{G}}(\omega_1,\dots,\omega_t)+L\sum_{k=1}^t\left(1-\frac{1}{c(\beta_k)}\right)}\right)^{\log_{\beta_t}\beta_{t+1}} \tag{22}$$

$$\geq \sum_{i=1}^N q_i'\beta_{t+1}^{\mathrm{Loss}_{\mathcal{E}(i)}^{\mathfrak{G}}(\omega_1,\dots,\omega_t)+L\sum_{k=1}^t\left(1-\frac{1}{c(\beta_k)}\right)} \tag{23}$$

As we saw above, it can be assumed that $\lambda(\omega,\gamma_t)\leq L$ for $\omega=0,1$ and $t=1,2,\dots$. Since $c(\beta)\geq 1$, we get

$$\left(L-\lambda\left(\omega_{t+1},\gamma_{t+1}\right)\right)\left(1-\frac{1}{c(\beta_{t+1})}\right)\geq 0 \tag{24}$$

for every $i=1,2,\dots,N$. Taking into account steps (6) and (7) in the algorithm, we get

$$\beta_{t+1}^{\lambda(\omega_{t+1},\gamma_{t+1})} \geq \beta_{t+1}^{\lambda(\omega_{t+1},\gamma_{t+1})+(L-\lambda(\omega_{t+1},\gamma_{t+1}))\left(1-\frac{1}{c(\beta_{t+1})}\right)} \tag{25}$$

$$= \beta_{t+1}^{L\left(1-\frac{1}{c(\beta_{t+1})}\right)+\frac{\lambda(\omega_{t+1},\gamma_{t+1})}{c(\beta_{t+1})}} \tag{26}$$

$$\geq \beta_{t+1}^{L\left(1-\frac{1}{c(\beta_{t+1})}\right)}\sum_{i=1}^N p_{t+1}^{(i)}\beta_{t+1}^{\lambda\left(\omega_{t+1},\gamma_{t+1}^{(i)}\right)} \quad . \tag{27}$$

We have

$$p_t^{(i)} = \frac{q_i'\beta_t^{\mathrm{Loss}_{\mathcal{E}(i)}^{\mathfrak{G}}(\omega_1,\dots,\omega_t)}}{\sum_{j=1}^N q_j'\beta_t^{\mathrm{Loss}_{\mathcal{E}(j)}^{\mathfrak{G}}(\omega_1,\dots,\omega_t)}} \quad , \tag{28}$$

$i=1,2,\dots,N$, $t=1,2,\dots$. Multiplying (23) by (27) and substituting the expressions for $p_{t+1}^{(i)}$ completes the proof of the lemma. □

To obtain an upper bound on $\mathrm{Loss}_{\mathfrak{M}}$ in terms of $\mathrm{Loss}_{\mathcal{E}(i)}$ we drop all terms except for the i-th in the sum (20) and then take the logarithm to the base β_t. For each $i=1,2,\dots,N$, if $q_i\neq 0$, we get

$$\mathrm{Loss}_{\mathfrak{M}}^{\mathfrak{G}}(\omega_1,\dots,\omega_t)\leq \mathrm{Loss}_{\mathcal{E}(i)}^{\mathfrak{G}}(\omega_1,\dots,\omega_t)+\sqrt{t}\ln\frac{1}{q_i'}+L\sum_{k=1}^t\left(1-\frac{1}{c(\beta_k)}\right) \quad . \tag{29}$$

Theorem 4 implies that there is a constant $c>0$ such that for all sufficiently large k the inequality

$$c(\beta_k)\leq 1+c\ln(1/\beta_t) \tag{30}$$

$$=1+\frac{c}{\sqrt{t}} \tag{31}$$

holds for all sufficiently large k. Thus there is $C>0$ such that

$$\sum_{k=1}^t\left(1-\frac{1}{c(\beta_k)}\right)\leq C+c\sum_{k=1}^t\frac{1}{c+\sqrt{t}} \tag{32}$$

$$\leq C+c\int_{x=0}^t\frac{1}{\sqrt{x}}dx \tag{33}$$

$$=C+2c\sqrt{t} \tag{34}$$

This proves Theorem 3.

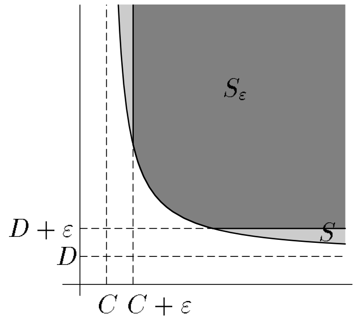

Figure 5 A bounded approximation of an unbounded game

6 Proof: "If" Part for Unbounded Games

6.1 Approximating Unbounded Games with Bounded

Suppose that $\mathfrak{G}$ is not bounded. If $\mathfrak{G}$ is one of the games described in Remark 1, then the WAA can be modified in such a way as to achieve the result of Theorem 3. Indeed, we can take a sufficiently large constant $L \in (0, +\infty)$ and require that on step (7) in the pseudo-code only predictions γ_t such that $\lambda(\omega, \gamma_t) \leq L$ for both $\omega = 0, 1$ are output.

The rest of this section is devoted to "essentially" unbounded games.

Let $\mathfrak{G} = \langle \mathbb{B}, \Gamma, \lambda \rangle$ be a (C, D)-game $(C, D \geq 0)$ that is not bounded and let $\varepsilon > 0$ be a positive constant. We will construct a bounded ε-approximation of $\mathfrak{G}$.

Let S be the set of superpredictions for $\mathfrak{G}$. Consider a game $\mathfrak{G}_\varepsilon = \langle \mathbb{B}, \Gamma_\varepsilon, \lambda_\varepsilon \rangle$ with the set of superpredictions $S_\varepsilon = S \cap ([C+\varepsilon, +\infty] \times [D+\varepsilon, +\infty])$ (see Fig. 5).

Since $S_\varepsilon \subseteq S$, for every prediction $\gamma' \in \Gamma_\varepsilon$ there is a prediction $\gamma \in \Gamma$ such that $\lambda(\omega, \gamma) \leq \lambda_\varepsilon(\omega, \gamma')$ for each $\omega = 0, 1$. Every strategy w.r.t. $\mathfrak{G}_\varepsilon$ can be easily reconstructed into a strategy w.r.t. $\mathfrak{G}$ suffering loss that is the same or smaller.

On the other hand, for every prediction $\gamma \in \Gamma$ there is a prediction $\gamma' \in \Gamma_\varepsilon$ such that $\lambda_\varepsilon(\omega, \gamma') \leq \lambda(\omega, \gamma) + \varepsilon$ for each $\omega = 0, 1$.

Suppose that the real part of S, the set $S \cap \mathbb{R}^2$, is convex. This implies that for every $\varepsilon > 0$ the same holds for S_ε, i.e., the set $S_\varepsilon \cap \mathbb{R}^2$ is convex. Theorem 3 can be applied to $\mathfrak{G}_\varepsilon$. Let us pick a constant C defined by the statement of the theorem and denote $C \ln 2$ by C_ε.

Consider two experts $\mathcal{E}^{(1)}$ and $\mathcal{E}^{(2)}$ that output predictions from Γ. The predictions can be replaced by predictions from Γ_ε so that the above inequalities are satisfied. Then there is a merging strategy following the WAA that merges the experts' predictions w.r.t. the game $\mathfrak{G}_\varepsilon$. Its predictions can again be converted into predictions from Γ. Thus there is a strategy $\mathfrak{M}$ such that

$$\mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{M}}(\omega_1, \omega_2, \ldots, \omega_n) \leq \mathrm{Loss}^{\mathfrak{G}}_{\mathcal{E}^{(i)}}(\omega_1, \omega_2, \ldots, \omega_n) + C_\varepsilon \sqrt{n} + \varepsilon n \qquad (35)$$

for all $i = 1, 2$ and $\omega_1, \omega_2, \ldots, \omega_n \in \mathbb{B}$, $n = 1, 2, \ldots$.

6.2 Weak Aggregation for Unbounded Games

Consider a game $\mathfrak{G}$ with the set of superpredictions S such that $S \cap \mathbb{R}^2$ is convex and let $\mathfrak{G}_\varepsilon$, S_ε and C_ε be as above, $\varepsilon > 0$.

Take a strictly increasing sequence of integers N_j, $j = 1, 2, \ldots$, and a sequence $\varepsilon_j > 0$, $j = 0, 1, 2, \ldots$. Consider the merging strategy $\mathfrak{M}$ defined as follows. The strategy first takes ε_0 and the approximation $\mathfrak{G}_{\varepsilon_0}$. It then merges the experts' predictions using the WAA applied w.r.t. $\mathfrak{G}_{\varepsilon_0}$. This continues while n, the length of the sequence of outcomes, is less than or equal to N_1. Then the strategy switches to ε_1 and the approximation $\mathfrak{G}_{\varepsilon_1}$ and applies the WAA w.r.t. $\mathfrak{G}_{\varepsilon_1}$ until n exceeds N_2 etc. Note that each time n passes through a bound N_i, the current invocation of the WAA terminates and a completely new invocation of the WAA starts working. It does not inherit anything from previous invocations.

Take $M_0 = N_1$ and $M_j = N_{j+1} - N_j$, $j = 1, 2 \ldots$. Let a positive integer n be such that $N_k < n \leq N_{k+1}$. Applying (35) yields

$$\mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{M}}(\omega_1, \omega_2, \ldots, \omega_n) \leq \mathrm{Loss}_{\mathcal{E}^{(i)}}(\omega_1, \omega_2, \ldots, \omega_n) + \alpha(n)$$

for all $i = 1, 2$ and

$$\alpha(n) = \sum_{j=0}^{k-1} M_j \varepsilon_j + \sum_{j=0}^{k-1} C_{\varepsilon_j} \sqrt{M_j} + \varepsilon_k (n - N_k) + C_{\varepsilon_k} \sqrt{n - N_k} \ . \tag{36}$$

We will show that the sequences N_i and ε_i can be chosen in such a way so as to ensure that $\alpha(n) = o(n)$ as $n \to +\infty$.

We will formulate conditions sufficient for the terms in (36) to be of $o(n)$ order of magnitude. First, note that

(1) $\lim_{j\to+\infty} \varepsilon_j = 0$

is sufficient to ensure that $\varepsilon_k + (n - N_k) = o(n)$ as $n \to +\infty$. Secondly, if, moreover,

(2) M_j is non-decreasing in j

and

(3) ε_j is non-increasing,

then $\sum_{j=0}^{k-1} M_j \varepsilon_j = o(n)$. Indeed, let m be a positive integer such that $m \leq k$. Condition (2) implies that $M_m \leq n/(k - m)$. Indeed, if $M_m > n/(k-m)$, then the same holds for all M_j, $j \geq m$ and thus the sum $\sum_{j=m}^{k-1} M_j > n$. We get

$$\frac{1}{n} \sum_{j=0}^{k-1} M_j \varepsilon_j = \frac{1}{n} \sum_{j=0}^{m} M_j \varepsilon_j + \frac{1}{n} \sum_{j=m+1}^{k-1} M_j \varepsilon_j \tag{37}$$

$$\leq \frac{(m+1) M_m \varepsilon_0}{n} + \frac{\varepsilon_{m+1}}{n} \sum_{j=m+1}^{k-1} M_j \tag{38}$$

$$\leq \frac{(m+1)\varepsilon_0}{k - m} + \varepsilon_{m+1} \ . \tag{39}$$

If we let $m = \sqrt{k}$, both the terms tend to 0 as k tends to $+\infty$, i.e., as $n \to +\infty$.

Thirdly, similar considerations imply that if

(4) $\lim_{j\to+\infty} M_j = +\infty$

and

(5) $C_{\varepsilon_j} \leq \sqrt[8]{M_j}$, $j = 0, 1, 2, \ldots$,

then $\sum_{j=0}^{k} C_{\varepsilon_j} \sqrt{M_j} \leq \sum_{j=0}^{k} M_j / M_j^{3/8} = o(n)$.

It remains to consider the last term in (36). There are two cases, either $n-N_k \le M_k^{3/4}$ or $n-N_k > M_k^{3/4}$. If the former case we get

$$\frac{1}{n} C_{\varepsilon_k} \sqrt{n-N_k} \le \frac{M_k^{1/8}\sqrt{n-N_k}}{N_k} \tag{40}$$

$$\le \frac{M_k^{1/8} M_k^{3/8}}{M_{k-1}} \tag{41}$$

$$= \frac{\sqrt{M_k}}{M_{k-1}}\ , \tag{42}$$

while in the latter case we get

$$\frac{1}{n} C_{\varepsilon_k} \sqrt{n-N_k} \le \frac{M_k^{1/8}\sqrt{M_k}}{M_k^{3/4}} \tag{43}$$

$$= \frac{1}{M_k^{1/8}} \tag{44}$$

$$\to 0 \tag{45}$$

as $k \to +\infty$. To ensure the convergence in the former case it is sufficient to have

(6) $M_{j-1} \le M_j^{3/4}$, $j = 1, 2, \ldots$.

Let us show that the conditions (1)–(6) are consistent, i.e., construct the sequences ε_j and M_j. Let $M_0 = \max(2, \lceil C_{\varepsilon_0}^8 \rceil)$ and $M_{j+1} = \lceil M_j^{4/3} \rceil$, $j = 0, 1, 2, \ldots$. The sequence ε_j is constructed as follows. Suppose that all ε_j have been constructed for $j \le k$. If $C_{\varepsilon_k/2} \le M_k^{1/8}$, we let $\varepsilon_{k+1} = \varepsilon_k/2$; otherwise we let $\varepsilon_{k+1} = \varepsilon_k$. Since $M_k \to +\infty$ and C_ε is finite for every $\varepsilon > 0$, we will be able to divide ε_k by 2 eventually and thus ensure that $\varepsilon_j \to 0$ as $j \to +\infty$.

7 Computability Issues

In this section we discuss the properties that an oracle computing λ should satisfy. The general principle is that the oracle should be capable of answering all "reasonable" questions that can be easily answered for a loss function specified by a simple analytical expression. Thus these requirements are not particularly restrictive.

First, the oracle should be able to evaluate the values $\lambda(\omega, \gamma)$, where $\omega = 0, 1$ and $\gamma \in \Gamma$. Secondly, it should be able to evaluate the values $c(\beta)$, $\beta \in (0, 1)$. Finally, given $x, y \in [-\infty, +\infty]$, it should be able to find γ (if any) such that $\lambda(0, \gamma) \le x$ and $\lambda(1, \gamma) \le y$.

When we say that the oracle is supplied with a number $x \in [-\infty, +\infty]$, we assume that it is given a sequence of rational intervals I_i that shrinks to x, i.e., $x = \cap_{i=1}^{+\infty} I_i$. A rational interval is one of the intervals $[-\infty, p)$, (p, q), or $(q, +\infty]$, where p and q are rational.

If we say that the oracle outputs $x \in [-\infty, +\infty]$, we mean that it outputs a sequence of rational intervals that shrinks to x. We assume that elements $\gamma \in \Gamma$ can be approximated and dealt with in a similar fashion.

References

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

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

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

[Vov90] V. Vovk. Aggregating strategies. In M. Fulk and J. Case, editors, *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, pages 371–383, San Mateo, CA, 1990. Morgan Kaufmann.

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