The Weak Aggregating Algorithm and Weak Mixability*

Yuri Kalnishkan and Michael V. Vyugin

Department of Computer Science, Royal Holloway,
University of London, Egham, Surrey, TW20 0EX, UK
{yura, misha}@cs.rhul.ac.uk

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 a sequence of letters from a finite alphabet. 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.

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 case of a discrete outcome space, i.e., $\omega_1, \omega_2, \ldots \in \Omega$ such that $|\Omega| < +\infty$.

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 prediction. 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.

* An early version of this paper was published in November, 2003 as Technical Report CLRC-TR-03-01, Computer Learning Research Centre, Royal Holloway, University of London available at http://www.clrc.rhul.ac.uk/publications/techrep.htm

P. Auer and R. Meir (Eds.): COLT 2005, LNAI 3559, pp. 188–203, 2005.

This problem has been studied intensively; see, e.g., [CBFH$^+$97, HKW98]. Papers [Vov90, Vov98] propose the Aggregating Algorithm that 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 depend on n.

If we can take c equal to 1, the game is called mixable. It is possible to provide a geometrical characterisation of mixable games in terms of the so called sets of superpredictions. 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 can grow when 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, Theor. 1, shows that weak mixability is equivalent to a very simple geometric property of the set of superpredictions, namely, the convexity of its finite part.

If the loss function is bounded, it is possible to predict as well as the best expert up to an additive term of the form $C\sqrt{n}$, provided the finite part of the set of superpredictions is convex. This result follows from a recent paper [HP04]. We shall present our own construction, which is based on ideas from [CBFH$^+$97].

If the game is not bounded, our construction can be applied in a 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 $\langle \Omega, \Gamma, \lambda \rangle$, where Ω is an *outcome space*, Γ is a *prediction space*, and $\lambda : \Omega \times \Gamma \to [0, +\infty]$ is a *loss function*. We assume that Ω is a finite set of cardinality $M < +\infty$; we will refer to elements of Ω as to $\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(M-1)}$. In the simplest binary case $M = 2$ and Ω may be identified with $\mathbb{B} = \{0, 1\}$. We also assume that Γ is a compact topological space and λ is continuous w.r.t. the extended topology of $[-\infty, +\infty]$. Since we treat Ω as a discrete space, the continuity of λ in two arguments is the same as continuity in the second argument. These assumption hold throughout the paper except for Remark 1, where negative losses are discussed.

The *square-loss* game, the *absolute-loss* game, and the *logarithmic* game with the outcome space $\Omega = \mathbb{B}$, prediction space $\Gamma = [0, 1]$, and loss functions $\lambda(\omega, \gamma) = (\omega - \gamma)^2$, $\lambda(\omega, \gamma) = |\omega - \gamma|$, and

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

respectively, are examples of (binary) games. A slightly different example is provided by the *simple prediction game* with $\Omega = \Gamma = \mathbb{B} = \{0,1\}$ and $\lambda(\omega,\gamma) = 0$ if $\omega = \gamma$ and $\lambda(\omega,\gamma) = 1$ otherwise.

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 $\omega_0 \in \Omega$ and $\gamma_0 \in \Gamma$, then there is a sequence $\gamma_n \in \Gamma$ such that $\gamma_n \to \gamma_0$ and $\lambda(\omega,\gamma_n)$ is finite for all $\omega \in \Omega$ and all positive integers n. In order words, any prediction that leads to infinite loss on some outcomes can be approximated by predictions that can only lead to finite loss no matter what outcome occurs. This restriction allows us to exclude some degenerate cases and to simplify the statements of theorems.

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. 6 for more details). All natural loss functions specified by simple analytical expression satisfy these requirements.

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

One can identify a prediction strategy with a function from Ω^* to Γ. Over the first n trials, the strategy $\mathfrak{S}$ suffers the total loss

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

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

2.2 Expert Advice

The problem of prediction with expert advice involve a pool of N experts $\mathcal{E}^{(1)},\mathcal{E}^{(2)},\ldots,\mathcal{E}^{(N)}$, which are working according to the aforementioned 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

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

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 Geometric Interpretation of a Game

Take a game $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ such that $\Omega = \{\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(M-1)}\}$ and $|\Omega| = M$. We say that an M-tuple $(s_0, s_1, \ldots, s_{M-1}) \in (-\infty, +\infty]^M$ is a *superprediction* if there is $\gamma \in \Gamma$ such that the inequalities $\lambda(\omega^{(i)}, \gamma) \le s_i$ hold for every $i = 0, 1, 2, \ldots, M-1$. The set of superpredictions S is an important object characterising the game.

3 Weak Mixability

One may wonder whether the learner can predict as well as the best expert up to an additive constant, i.e., to suffer loss within an additive constant range of the loss of the best expert. It is possible for the so called mixable games; for more details see [Vov90, Vov98]. Examples of mixable games include the square-loss game and the logarithmic game; the simple prediction game and the absolute-loss game are not mixable.

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.

In the worst case, loss grows linearly in the length of the sequence. 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 $\mathcal{E}^{(1)}, \mathcal{E}^{(2)}, \ldots, \mathcal{E}^{(N)}$ $(N = 1, 2, \ldots)$, the inequality

$$\mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{M}}(\omega_1, \omega_2, \ldots, \omega_n) \le \mathrm{Loss}_{\mathcal{E}^{(i)}}(\omega_1, \omega_2, \ldots, \omega_n) + f(n) \tag{4}$$

holds for all $i = 1, 2, \ldots, N$ and every finite sequence $\omega_1, \omega_2, \ldots, \omega_n \in \Omega$, $n = 1, 2, \ldots$.

The following theorem is the main result of the paper.

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

The merging strategy in the definition of weak mixability is polynomial-time computable modulo the oracle computing λ (see Sect. 6).

The examples of the weakly mixable games are the logarithmic and the square-loss game, which are also mixable, and the absolute-loss game, which is not mixable. The simple prediction game is not weakly mixable.

The rest of the paper contains the proof of the theorem. The 'only if' part follows from Theor. 2 that is proved in Appendix A.

The 'if' splits into two parts, for bounded and for unbounded games. The 'if' part for bounded games follows from [HP04]. In Sect. 4 we shall give an alternative derivation, which gives a slightly better value of the constant C in the additive term $C\sqrt{n}$. The unbounded case is described in Sect. 5.

Remark 1. Let us allow (within this remark) λ to accept negative values; they can be interpreted as 'gain' or 'reward'. If λ accepts the value $-\infty$, the expression for the total loss may include the sum $(-\infty)+(+\infty)$, which is undefined. In order to avoid this ambiguity, it is natural to prohibit λ to take the value $-\infty$. Since λ is assumed to be continuous, this implies that λ is bounded from below, i.e., there is $a > -\infty$ such that $\lambda(\omega, \gamma) \geq a$ for all values of ω and γ.

Consider another game with the loss function $\lambda'(\omega, \gamma) = \lambda(\omega, \gamma) + a$, which is nonnegative. A merging strategy working with nonnegative loss functions can be easily adapted to work with the original game: let the learner just imagine that it is playing the game with λ'. The losses w.r.t. the two games on a string $\omega_1\omega_2\ldots\omega_n$ will differ by an and the upper bounds of the type (4) will be preserved. On the other hand, the sets of superpredictions for the two games will differ by a shift, which preserves convexity. Therefore Theor. 1 remains true for games with loss functions bounded from below.

4 'If' Part for Bounded Games

4.1 Weak Aggregating Algorithm

In this subsection we formulate the Weak Aggregating Algorithm (WAA). Let $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ be a game such that $|\Omega| = M < +\infty$ and let N be the number of experts. Let $\Omega = \{\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(M-1)}\}$.

We describe the WAA using pseudo-code. The WAA accepts N initial normalised weights $q_1, q_2, \ldots, q_N \in [0,1]$ such that $\sum_{i=1}^N q_i = 1$ and a positive number c as parameters. The role of c is similar to that of the learning rate in the theory of prediction with expert advice. Let $\beta_t = e^{-c/\sqrt{t}}$, $t = 1, 2, \ldots$.

(1) $l_1^{(i)} := 0$, $i = 1, 2, \ldots, N$
(2) FOR $t = 1, 2, \ldots$

(3) $w_t^{(i)} := q_i \beta_t^{l_t^{(i)}}$, $i = 1, 2, \ldots, N$

(4) $p_t^{(i)} := \frac{w_t^{(i)}}{\sum_{j=1}^N w_t^{(j)}}$, $i = 1, 2, \ldots, N$

(5) `read experts' predictions` $\gamma_t^{(1)}, \gamma_t^{(2)}, \ldots, \gamma_t^{(N)}$

(6) $g_k := \sum_{j=1}^N \lambda\left(\omega^{(k)}, \gamma_t^{(j)}\right) p_t^{(j)}$, $k = 0, 1, \ldots, M-1$

(7) `output` $\gamma_t \in \Gamma$ `such that` $\lambda(\omega^{(k)}, \gamma_t) \le g_k$ `for all` $k = 0, 1, \ldots, M-1$

(8) `observe` ω_t

(9) $l_{t+1}^{(i)} := l_t^{(i)} + \lambda\left(\omega_t, \gamma_t^{(i)}\right)$, $i = 1, 2, \ldots, N$

(10) `END FOR`

The variable $l_t^{(i)}$ stores the loss of the i-th expert $\mathcal{E}^{(i)}$, i.e., after trial t we have $l_{t+1}^{(i)} = \mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(\omega_1, \omega_2, \ldots, \omega_t)$. The values $w_t^{(i)}$ are weights assigned to experts during the work of the algorithm; they depend on the loss suffered by experts and initial weights q_i. The values $p_t^{(i)}$ are obtained by normalising $w_t^{(i)}$. 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.

This algorithm is applicable if the set of superpredictions S has a convex finite part $S \cap \mathbb{R}^M$. If this is the case, then the point $(g_0, g_1, \ldots, g_{M-1})$ belongs to S and thus γ_t can be found on step (7).

A game $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ is bounded if and only if λ is bounded, i.e., there is $L \in (0, +\infty)$ such that $\lambda(\omega, \gamma) \le L$ for each $\omega \in \Omega$ and $\gamma \in \Gamma$. Examples of bounded games include the square-loss game, the absolute-loss game, and the simple prediction game. The logarithmic game is unbounded.

For bounded games the following lemma holds.

Lemma 1. *For every $L > 0$, every game $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ such that $|\Omega| < +\infty$ and $\lambda(\omega, \gamma) \le L$ for all $\omega \in \Omega$ and $\gamma \in \Gamma$, and 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 with initial weights $q_1, q_2, \ldots, q_N \in [0, 1]$ such that $\sum_{i=1}^N q_i = 1$ and $c > 0$ achieves loss satisfying*

$$\mathrm{Loss}_{\mathfrak{M}}^{\mathfrak{G}}(\omega_1, \omega_2, \ldots, \omega_n) \le \mathrm{Loss}_{\mathcal{E}^{(i)}}(\omega_1, \omega_2, \ldots, \omega_n) + \left(cL^2 + \frac{1}{c} \ln \frac{1}{q_i}\right) \sqrt{n}$$

for every $i = 1, 2, \ldots, N$ and every finite sequence $\omega_1, \omega_2, \ldots, \omega_n \in \Omega$.

The proof of Lemma 1 is given in Appendix B.

Remark 2. It is easy to see that the result of Lemma 1 will still hold for a countable pool of experts $\mathcal{E}_1, \mathcal{E}_2, \ldots$ We take weights $\sum_{i=1}^{+\infty} q_i = 1$; the sums in lines (4) and (6) from the definition of the WAA become infinite but they clearly converge.

Let us take equal initial weights $q_1 = q_2 = \ldots = q_N = 1/N$ in the WAA. The additive term then reduces to $(cL^2 + (\ln N)/c)\sqrt{n}$. When $c = \sqrt{\ln N}/L$ this expression reaches its minimum. We get the following corollary.

Corollary 1. *Under the conditions of Lemma 1, there is a merging strategy $\mathfrak{M}$ achieving loss satisfying*

$$\mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{M}}(\omega_1, \omega_2, \ldots, \omega_n) \leq \mathrm{Loss}_{\mathcal{E}^{(i)}}(\omega_1, \omega_2, \ldots, \omega_n) + 2L\sqrt{n \ln N} \ .$$

5 'If' Part for Unbounded Games

5.1 Counterexample

The WAA can be applied even in the case of an unbounded game; indeed, the only requirement is that the finite part of the set of superpredictions S is convex. However we cannot guarantee that a reasonable upper bound on the loss of the strategy using it will exist. The same applies to any strategy that uses a linear combination in the same fashion as WAA.

Indeed, consider a game with an unbounded loss function λ. Let ω_0 be such that the function $\lambda(\omega_0, \gamma)$ attains arbitrary large values.

Suppose that there are two experts $\mathcal{E}_1$ and $\mathcal{E}_2$ and on some trial they are ascribed weights $p^{(1)}$ and $p^{(2)}$ such that $p^{(2)} > 0$. Suppose that $\mathcal{E}_1$ outputs $\gamma^{(1)}$ such that $\lambda(\omega_0, \gamma^{(1)}) < +\infty$. The upper estimate on the loss of the learner in the case when the outcome ω_0 occurs is

$$g_0 = p^{(1)}\lambda(\omega_0, \gamma^{(1)}) + p^{(2)}\lambda(\omega_0, \gamma^{(2)}) \ ,$$

where $\gamma^{(2)}$ is the prediction output by $\mathcal{E}_2$. Let us vary $\gamma^{(2)}$. The weights depend on the previous behaviour of the experts and they cannot be changed. If $\lambda(\omega_0, \gamma^{(2)})$ tends to infinity, then g_0 tends to infinity and therefore the difference $g_0 - \lambda(\omega_0, \gamma^{(1)})$ tends to infinity. Thus the learner cannot compete with the first expert.

This example shows that the WAA cannot be straightforwardly generalised to unbounded games. It needs to be altered.

5.2 Approximating Unbounded Games with Bounded

The following lemma allows us to 'cut off' the infinity at a small cost.

Lemma 2. *Let $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ be a game such that $|\Omega| < +\infty$. Then for every $\varepsilon > 0$ there is $L > 0$ with the following property. For every $\gamma \in \Gamma$ there is $\gamma^* \in \Gamma$ such that $\lambda(\omega, \gamma^*) \leq L$ and $\lambda(\omega, \gamma^*) \leq \lambda(\omega, \gamma) + \varepsilon$ for all $\omega \in \Omega$.*

The proof of Lemma 2 is given in Appendix C.

We assume that the game is such that the numbers $L = L_\varepsilon$ can be computed efficiently for every ε and that γ^* can be computed efficiently given $\gamma \in \Gamma$. This is a restriction we impose on games.

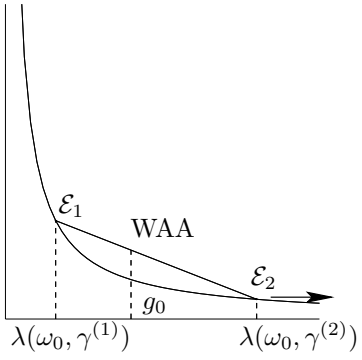

Fig. 1. A counterexample for unbounded games in dimension 2

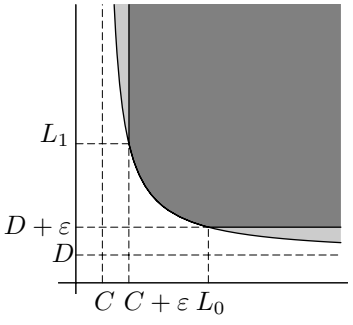

Fig. 2. Computing L_ε in the case of two outcomes

In the case of two outcomes $|\Omega| = 2$ computations are particularly straightforward. See Fig. 2, where

$$C = \inf_{\gamma \in \Gamma} \lambda(\omega^{(0)}, \gamma) \text{ and } D = \inf_{\gamma \in \Gamma} \lambda(\omega^{(1)}, \gamma);$$

we can take $L_\varepsilon = \max(L_0, L_1)$. If γ is such that the point $(\lambda(\omega^{(0)}, \gamma), \lambda(\omega^{(1)}, \gamma))$ falls into the area to the right of the straight line $x = L_0$, we can take γ^* such that $(\lambda(\omega^{(0)}, \gamma^*), \lambda(\omega^{(1)}, \gamma^*)) = (L_0, D + \varepsilon)$.

5.3 Merging Experts in the Unbounded Case

Consider an unbounded game $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ and N experts $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_N$. Fix some $\varepsilon > 0$. Let L_ε be as above. After obtaining experts' predictions $\gamma_t^{(1)}, \gamma_t^{(2)}, \ldots, \gamma_t^{(N)}$ we can find $\gamma_t^{(1)*}, \gamma_t^{(2)*}, \ldots, \gamma_t^{(N)*}$ as in Lemma 2 and then apply the results for the bounded case to them. By proceeding in this fashion, a strategy $\mathfrak{M}$ suffers loss such that

$$\mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{M}}(\omega_1, \omega_2, \ldots, \omega_n) \le \mathrm{Loss}^{\mathfrak{G}}_{\mathcal{E}^{(i)}}(\omega_1, \omega_2, \ldots, \omega_n) + C_\varepsilon \sqrt{n} + \varepsilon n \tag{5}$$

for all $i = 1, 2, \ldots, N$ and $\omega_1, \omega_2, \ldots, \omega_n \in \Omega$, $n = 1, 2, \ldots$, where $C_\varepsilon = 2L_\varepsilon^2 \sqrt{\ln N}$ (we are applying WAA with equal weights).

This inequality does not allow us to prove Theor. 1. In order to achieve an extra term of the order $o(n)$ we will vary ε.

Take a strictly increasing sequence of integers N_k, $k = 1, 2, \ldots$, and a sequence $\varepsilon_k > 0$, $k = 0, 1, 2, \ldots$. Consider the merging strategy $\mathfrak{M}$ defined as follows. The strategy first takes ε_0 and merges the experts' predictions using the WAA and ε_0 in the fashion described above. 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 applies the WAA and ε_1 until n exceeds N_2 etc (see Fig. 4). Note that each time

n passes through a limit N_i, the current invocation of the WAA terminates and a completely new invocation of the WAA starts working. It does not have to inherit anything from previous invocations.

In Appendix D we show how to choose the sequences ε_k and N_k in such a way as to achieve the desired extra term.

6 Computability Issues

In this section we summarise 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 $\omega \in \Omega$ and $\gamma \in \Gamma$. Secondly, given $x_1, x_2, \ldots, x_n \in [-\infty, +\infty]$, it should be able to find γ (if any) such that $\lambda(\omega^{(i)}, \gamma) \leq x_i$, $i = 1, 2, \ldots, N$. Thirdly, the oracle should be able to compute numbers L_ε and to find γ^* by $\gamma \in \Gamma$ (see Subsect. 5.2).

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.

Acknowledgements

We would like to thank participants of the Kolmogorov seminar on complexity theory at the Moscow State University and Alexander Shen in particular for useful suggestions that allowed us to simplify the WAA. We would also like to thank Volodya Vovk for suggesting an idea that helped us to strengthen an upper bound on the performance of WAA.

We are grateful to anonymous COLT reviewers for their detailed comments. Unfortunately, we could not incorporate all their suggestions into the conference version of the paper due to lack of space.

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.

[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.

[HP04] M. Hutter and J. Poland. Predictions with expert advice by following the perturbed leader for general weights. In *Algorithmic Learning Theory, 15th International Conference, ALT 2004, Proceedings*, volume 3244 of *Lecture Notes in Artificial Intelligence*, pages 279–293. Springer, 2004.

[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.

Appendix A. Proof: 'Only If' Part

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

Theorem 2. *If a game $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$, $|\Omega| = M < +\infty$, has the set of superpredictions S such that its finite part $S \cap \mathbb{R}^M$ is not convex, then there are two strategies $\mathfrak{S}_1$ and $\mathfrak{S}_2$ and a constant $\theta > 0$ such that for any strategy $\mathfrak{S}$ there is a sequence $\omega_n \in \Omega$, $n = 1, 2, \ldots$, such that*

$$\max_{i=1,2} \left(\mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{S}}(\omega_1, \omega_2, \ldots, \omega_n) - \mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{S}_i}(\omega_1, \omega_2, \ldots, \omega_n) \right) \geq \theta n \tag{6}$$

for all positive integers n.

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

Proof. We will use vector notation. If $X = (x_1, \ldots, x_n)$, $Y = (y_1, \ldots, y_n)$ and $\alpha \in \mathbb{R}$, then $X + Y$ and αX are defined in the natural way. By $\langle X, Y \rangle$ we denote the scalar product $\sum_{i=1}^{n} x_i y_i$.

For brevity we will denote finite sequences by bold letters, e.g., $\boldsymbol{x} = \omega_1 \ldots \omega_n \in \Omega^n$. Let $|\boldsymbol{x}|$ be the length of $\boldsymbol{x}$, i.e., the total number of symbols in $\boldsymbol{x}$. We will denote the number of elements equal to $\omega^{(0)}$ in a sequence $\boldsymbol{x}$ by $\sharp_0 \boldsymbol{x}$, the number of elements equal to $\omega^{(1)}$ by $\sharp_1 \boldsymbol{x}$ etc. It is easy to see that $\sum_{i=0}^{M-1} \sharp_i \boldsymbol{x} = |\boldsymbol{x}|$ for every $\boldsymbol{x} \in \Omega^*$. The vector $(\sharp_0 \boldsymbol{x}, \sharp_1 \boldsymbol{x}, \ldots, \sharp_{M-1} \boldsymbol{x})$ will be denoted by $\sharp \boldsymbol{x}$.

There exists a couple of points $B_1 = \left(b_1^{(0)}, b_1^{(1)}, \ldots, b_1^{(M-1)}\right)$ and $B_2 = \left(b_2^{(0)}, b_2^{(1)}, \ldots, b_2^{(M-1)}\right)$ such that $B_1, B_2 \in S \cap \mathbb{R}^M$ but the segment $[B_1, B_2]$ connecting them is not a subset of S. Let $\alpha \in (0, 1)$ be such that $C = \alpha B_1 + (1-\alpha) B_2$ does not belong to S (see Fig. 3). Since λ is continuous and Γ is compact, the set S is closed and thus there is a small vicinity of C that is a subset of $\mathbb{R}^M \setminus S$.

Without restricting the generality one may assume that all coordinates of B_1 and B_2 are strictly positive. Indeed, the points $B_1' = B_1 + t \cdot (1, 1, \ldots, 1)$ and $B_2' = B_2 + t \cdot (1, 1, \ldots, 1)$ belong to S for all positive t. If $t > 0$ is sufficiently

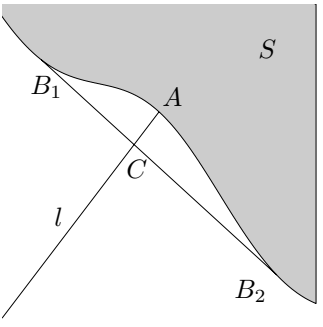

Fig. 3. The drawing for the proof of Theor. 2

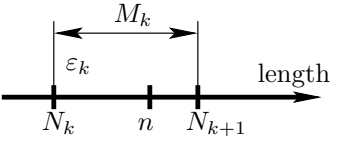

Fig. 4. The sequences of N_k, M_k, and ε_k

small, then $C' = \alpha B_1' + (1-\alpha)B_2'$ still belongs to the vicinity mentioned above and thus C' does not belong to S.

Let us draw a half-line l starting from the origin through C. Let $A = \left(a^{(0)}, a^{(1)}, \ldots, a^{(M-1)}\right)$ be the intersection of l with the boundary ∂S. Such a point really exists. Indeed, $l = \{X \in \mathbb{R}^M \mid \exists t \geq 0 \,:\, X = tC\}$. For sufficiently large t all coordinates of tC are greater than the corresponding coordinates of B_1 and thus $tC \in S$. Now let $t_0 = \inf\{t \geq 0 \mid tC \in S\}$ and $A = t_0 C$. Since $C \notin S$, we get $t_0 > 1$ and thus $A = (1+\delta)C$, where $\delta > 0$.

We now proceed to constructing the strategies $\mathfrak{S}_1$ and $\mathfrak{S}_2$. There are predictions $\gamma_1, \gamma_2 \in \Gamma$ such that $\lambda(\omega^{(i)}, \gamma_1) \leq b_1^{(i)}$ and $\lambda(\omega^{(i)}, \gamma_2) \leq b_2^{(i)}$ for all $i = 0, 1, 2, \ldots, M-1$. Let $\mathfrak{S}_1$ be the oblivious strategy that always predicts γ_1, no matter what outcomes actually occur. Similarly, let $\mathfrak{S}_2$ be the strategy that always predicts γ_2. Without loss of generality it can be assumed that $\mathfrak{S}_1$ and $\mathfrak{S}_2$ are computable. Indeed, the points B_1 and B_2 can be replaced by computable points from their small vicinities. The definitions of $\mathfrak{S}_1$ and $\mathfrak{S}_2$ imply the inequalities

$$\mathrm{Loss}_{\mathfrak{S}_1}(\boldsymbol{x}) \leq \sum_{i=0}^{M-1} \sharp_i \boldsymbol{x} b_1^{(i)} = \langle B_1, \sharp \boldsymbol{x} \rangle \quad \text{and} \quad \mathrm{Loss}_{\mathfrak{S}_2}(\boldsymbol{x}) \leq \sum_{i=0}^{M-1} \sharp_i \boldsymbol{x} b_2^{(i)} = \langle B_2, \sharp \boldsymbol{x} \rangle \tag{7}$$

for all strings $\boldsymbol{x} \in \mathbb{B}^*$.

Now let us consider an arbitrary strategy $\mathfrak{S}$ and construct a sequence $\boldsymbol{x}_n = \omega_1 \omega_2 \ldots \omega_n$ satisfying the requirements of the theorem. The sequence is constructed by induction. Let $\boldsymbol{x}_0 = \Lambda$. Suppose that $\boldsymbol{x}_n$ has been constructed. 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. There is some $\omega^{(i_0)} \in \Omega$ such that $\lambda(\omega^{(i_0)}, \gamma) \geq a^{(i_0)}$. Indeed, if this is not true and the inequalities $\lambda(\omega^{(i)}, \gamma) < a^{(i)}$ hold for all $i = 1, 2, \ldots, M-1$, then there is a vicinity of A that is a subset of S. This contradicts the definition of A. We let $\boldsymbol{x}_{n+1} = \boldsymbol{x}_n \omega_{i_0}$. The construction implies

$$\mathrm{Loss}_{\mathfrak{S}}(\boldsymbol{x}_n) \geq \sum_{i=0}^{M-1} \sharp_i \boldsymbol{x}_n a^{(i)} = \langle A, \sharp \boldsymbol{x}_n \rangle \ . \tag{8}$$

Let $\varepsilon = \min_{j=1,2;\ i=0,1,2,\ldots,M-1} b_j^{(i)} > 0$. We get $\langle B_j, \boldsymbol{x} \rangle = \sum_{i=0}^{M-1} b_j^{(i)} \sharp_i \boldsymbol{x} \geq \varepsilon |\boldsymbol{x}|$ for all strings $\boldsymbol{x} \in \mathbb{B}^*$ and $j = 1, 2$. Since $A = (1+\delta)(\alpha B_1 + (1-\alpha) B_2)$ we get

$$\begin{aligned} \langle A, \sharp \boldsymbol{x} \rangle &= (1+\delta)(\alpha \langle B_1, \sharp \boldsymbol{x} \rangle + (1-\alpha) \langle B_2, \sharp \boldsymbol{x} \rangle) \\ &\geq \alpha \langle B_1, \sharp \boldsymbol{x} \rangle + (1-\alpha) \langle B_2, \sharp \boldsymbol{x} \rangle + \delta \varepsilon |\boldsymbol{x}| \end{aligned}$$

for all strings $\boldsymbol{x}$. Let $\theta = \delta\varepsilon$; note that ε and δ do not depend on $\mathfrak{S}$. By combining this inequality with (7) and (8) we obtain the inequality

$$\mathrm{Loss}_{\mathfrak{S}}(\boldsymbol{x}_n) \geq \alpha \, \mathrm{Loss}_{\mathfrak{S}_1}(\boldsymbol{x}_n) + (1-\alpha) \, \mathrm{Loss}_{\mathfrak{S}_2}(\boldsymbol{x}_n) + \theta n$$

for all positive integers n.

It is easy to see that

$$\begin{aligned} \mathrm{Loss}_{\mathfrak{S}}(\boldsymbol{x}_n) - \mathrm{Loss}_{\mathfrak{S}_1}(\boldsymbol{x}_n) &\geq (1-\alpha)(\mathrm{Loss}_{\mathfrak{S}_2}(\boldsymbol{x}) - \mathrm{Loss}_{\mathfrak{S}_1}(\boldsymbol{x})) + \theta n \ , \\ \mathrm{Loss}_{\mathfrak{S}}(\boldsymbol{x}_n) - \mathrm{Loss}_{\mathfrak{S}_2}(\boldsymbol{x}_n) &\geq \alpha(\mathrm{Loss}_{\mathfrak{S}_1}(\boldsymbol{x}) - \mathrm{Loss}_{\mathfrak{S}_2}(\boldsymbol{x})) + \theta n \ . \end{aligned}$$

If $\mathrm{Loss}_{\mathfrak{S}_2}(\boldsymbol{x}) \geq \mathrm{Loss}_{\mathfrak{S}_1}(\boldsymbol{x})$ the former difference is greater than or equal to θn, otherwise the latter difference is greater than or equal to θn. By combining these facts we obtain (6). □

Appendix B. Proof of Lemma 1

In this appendix we prove Lemma 1. We start with the following lemma.

Lemma 3. *Let $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ be a game such that $|\Omega| < +\infty$ and let N be the number of experts. Let the finite part of the set of superpredictions $S \cap \mathbb{R}^M$ be convex. If $\mathfrak{M}$ is a merging strategy following the WAA, then for every $t = 1, 2, \ldots$ we get*

$$\beta_t^{\mathrm{Loss}_{\mathfrak{M}}^{\mathfrak{G}}(\omega_1,\ldots,\omega_t)} \geq \beta_t^{\sum_{j=1}^{t} \alpha(j)} \sum_{i=1}^{N} q_i \beta_t^{\mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(\omega_1,\ldots,\omega_t)} \ , \tag{9}$$

where

$$\alpha(j) = \log_{\beta_j} \frac{\beta_j^{\sum_{i=1}^{N} \lambda(\omega_j, \gamma_j^{(i)}) p_j^{(i)}}}{\sum_{i=1}^{N} \beta_j^{\lambda(\omega_j, \gamma_j^{(i)})} p_j^{(i)}} \tag{10}$$

for $j = 1, 2, \ldots, t$, in the notation introduced above.

Proof (of Lemma 3). The proof is by induction on t. Let us assume that (9) 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 (9) implies

$$\beta_{t+1}^{\mathrm{Loss}_{\mathfrak{M}}^{\mathfrak{G}}(\omega_1,\ldots,\omega_t)} = \left(\beta_t^{\mathrm{Loss}_{\mathfrak{M}}^{\mathfrak{G}}(\omega_1,\ldots,\omega_t)}\right)^{\log_{\beta_t}\beta_{t+1}} \tag{11}$$

$$\geq \left(\beta_t^{\sum_{j=1}^t \alpha(j)} \sum_{i=1}^N q_i \beta_t^{\mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(\omega_1,\ldots,\omega_t)}\right)^{\log_{\beta_t}\beta_{t+1}} \tag{12}$$

$$\geq \beta_{t+1}^{\sum_{j=1}^t \alpha(j)} \sum_{i=1}^N q_i \beta_{t+1}^{\mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(\omega_1,\ldots,\omega_t)} \tag{13}$$

Step (7) of the algorithm implies that $\lambda(\omega_{t+1}, \gamma_{t+1}) \leq \sum_{i=1}^N \lambda\left(\omega_{t+1}, \gamma_{t+1}^{(i)}\right) p_{t+1}^{(i)}$. By exponentiating this inequality we get

$$\beta_{t+1}^{\lambda(\omega_{t+1},\gamma_{t+1})} \geq \beta_{t+1}^{\sum_{i=1}^N \lambda\left(\omega_{t+1},\gamma_{t+1}^{(i)}\right) p_{t+1}^{(i)}} \tag{14}$$

$$= \frac{\beta_{t+1}^{\sum_{i=1}^N \lambda\left(\omega_{t+1},\gamma_{t+1}^{(i)}\right) p_{t+1}^{(i)}}}{\sum_{i=1}^N \beta_{t+1}^{\lambda\left(\omega_{t+1},\gamma_{t+1}^{(i)}\right)} p_{t+1}^{(i)}} \sum_{i=1}^N \beta_{t+1}^{\lambda\left(\omega_{t+1},\gamma_{t+1}^{(i)}\right)} p_{t+1}^{(i)} \tag{15}$$

$$= \beta_{t+1}^{\alpha(t+1)} \sum_{i=1}^N \beta_{t+1}^{\lambda\left(\omega_{t+1},\gamma_{t+1}^{(i)}\right)} p_{t+1}^{(i)} \quad . \tag{16}$$

Multiplying (13) by (16) and substituting

$$p_{t+1}^{(i)} = \frac{w_{t+1}}{\sum_{j=1}^N w_{t+1}^{(j)}} = \frac{q_i \beta_{t+1}^{\mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(\omega_1,\ldots,\omega_t)}}{\sum_{j=1}^N q_j \beta_{t+1}^{\mathrm{Loss}_{\mathcal{E}^{(j)}}^{\mathfrak{G}}(\omega_1,\ldots,\omega_t)}}$$

completes the proof on the lemma. □

By taking the logarithm of (9) we get

$$\mathrm{Loss}_{\mathfrak{M}}^{\mathfrak{G}}(\omega_1, \ldots, \omega_t) \leq \sum_{j=1}^t \alpha(j) + \log_{\beta_t} \sum_{i=1}^N q_i \beta_t^{\mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(\omega_1,\ldots,\omega_t)}$$

$$\leq \sum_{j=1}^t \alpha(j) + \log_{\beta_t} q_i + \mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(\omega_1, \ldots, \omega_t)$$

for every $i = 1, 2, \ldots, N$. We have $\log_{\beta_t} q_i = -\frac{\sqrt{t}}{c} \ln q_i$. It remains to estimate the first term.

Recall that L is an upper bound on λ. By applying the inequality $\ln x \le x-1$ we get

$$\begin{aligned}\alpha(t) &= \sum_{i=1}^{N} \lambda(\omega_t, \gamma_t^{(i)}) p_t^{(i)} + \frac{\sqrt{t}}{c} \ln \sum_{i=1}^{N} \beta_t^{\lambda(\omega_t, \gamma_t^{(i)})} p_t^{(i)} \\ &\le \sum_{i=1}^{N} \lambda(\omega_t, \gamma_t^{(i)}) p_t^{(i)} + \frac{\sqrt{t}}{c} \left(\sum_{i=1}^{N} \beta_t^{\lambda(\omega_t, \gamma_t^{(i)})} p_t^{(i)} - 1 \right)\end{aligned}$$

By using Taylor's series with Lagrange's remainder term we obtain

$$\beta_t^{\lambda(\omega_t, \gamma_t^{(i)})} = e^{-c\lambda(\omega_t, \gamma_t^{(i)})/\sqrt{t}} = 1 - \frac{c\lambda(\omega_t, \gamma_t^{(i)})}{\sqrt{t}} + \frac{1}{2} \left(\frac{c\lambda(\omega_t, \gamma_t^{(i)})}{\sqrt{t}} \right)^2 e^{\xi} \ ,$$

where $\xi \in [-c\lambda(\omega_t, \gamma_t^{(i)})/\sqrt{t}, 0]$ and thus

$$\beta_t^{\lambda(\omega_t, \gamma_t^{(i)})} \le 1 - \frac{c\lambda(\omega_t, \gamma_t^{(i)})}{\sqrt{t}} + \frac{c^2 L^2}{2t} \ .$$

Therefore $\alpha(t) \le cL^2/2\sqrt{t}$ and summing yields

$$\sum_{j=1}^{t} \alpha(j) \le \sum_{j=1}^{t} \frac{cL^2}{2\sqrt{t}} \le \frac{cL^2}{2} \int_0^t \frac{dx}{\sqrt{x}} = cL^2\sqrt{t} \ .$$

This completes the proof.

Appendix C. Proof of Lemma 2

Let $|\Omega| = M$ and $\Omega = \{\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(M-1)}\}$.

For every $L > 0$ let $\Gamma_L = \{\gamma \in \Gamma \mid \lambda(\omega, \gamma) \le L \text{ for all } \omega \in \Omega\}$ and let $P_L = \big\{ \big(\lambda(\omega^{(0)}, \gamma), \lambda(\omega^{(1)}, \gamma), \ldots, \lambda(\omega^{(M-1)}, \gamma)\big) \mid \gamma \in \Gamma_L \big\}$. In other terms, $P_L = P \cap [0, L]^M$, where $P = \{\big(\lambda(\omega^{(0)}, \gamma), \lambda(\omega^{(1)}, \gamma), \ldots, \lambda(\omega^{(M-1)}, \gamma)\big) \mid \gamma \in \Gamma\}$ is the set of all 'predictions'. For every $\varepsilon > 0$ let $U_{L,\varepsilon}$ be the ε-vicinity of the set P_L, i.e., the union of all open balls of radius ε having points of P_L as their centres. Finally, let $S_{L,\varepsilon} = \{X \in [-\infty, +\infty]^M \mid X \ge Y \text{ for some } Y \in U_{L,\varepsilon}\}$.

Now fix $\varepsilon > 0$. We have $S \subseteq \bigcup_{L>0} S_{L,\varepsilon}$. Indeed, consider a point $X = \big(\lambda(\omega^{(0)}, \gamma), \lambda(\omega^{(1)}, \gamma), \ldots, \lambda(\omega^{(M-1)}, \gamma)\big)$ for some $\gamma \in \Gamma$. If all coordinates of X are finite, $X \in P_L$ for some sufficiently large L. If some of the coordinates are infinite, γ can still be approximated by predictions that can only lead to finite loss and thus X belongs to some $S_{L,\varepsilon}$.

The covering $\bigcup_{L>0} S_{L,\varepsilon}$ has a finite subcovering. Indeed, let us take some $\beta \in (0, 1)$ and apply the transformation $\mathfrak{B}_\beta$ specified by

$$\mathfrak{B}_\beta(x_0, x_1, \ldots, x_{M-1}) = (\beta^{x_0}, \beta^{x_1}, \ldots, \beta^{x_{M-1}}) \ .$$

The set $\mathfrak{B}_\beta(S)$ is a compact set and all sets $\mathfrak{B}_\beta(S_{L,\varepsilon})$ are open if considered as subsets of the space $[0,+\infty)^M$ with the standard Euclidean topology.

Therefore there is $L > 0$ such that $S \subseteq S_{L,\varepsilon}$. The lemma follows.

Appendix D. Choosing the Sequences

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 $N_k < n \leq N_{k+1}$ (see Fig. 4). Applying (5) 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 $i = 1, 2, \ldots, N$, where N is the number of experts and

$$\alpha(n) = \sum_{j=0}^{k} M_j \varepsilon_j + \sum_{j=0}^{k} C_{\varepsilon_j} \sqrt{M_j} + \varepsilon_k (n - N_k) + C_{\varepsilon_k} \sqrt{n - N_k} \ ; \tag{17}$$

note that the former two terms correspond to the previous invocations of WAA and the later two correspond to the current invocation.

We will formulate conditions sufficient for the terms in (17) 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 $\sum_{j=0}^{k} M_j \varepsilon_j = o(n)$. Indeed, let m be a positive integer such that $m \leq k$. Condition (2) implies that $M_m \leq n/(k-m+1)$. Indeed, if $M_m > n/(k-m+1)$, then the same holds for all M_j, $j \geq m$ and thus $\sum_{j=m}^{k} M_j > n$. We get

$$\begin{aligned} \frac{1}{n} \sum_{j=0}^{k} M_j \varepsilon_j &= \frac{1}{n} \sum_{j=0}^{m} M_j \varepsilon_j + \frac{1}{n} \sum_{j=m+1}^{k} M_j \varepsilon_j \\ &\leq \frac{(m+1) M_m \varepsilon_0}{n} + \frac{\varepsilon_{m+1}}{n} \sum_{j=m+1}^{k} M_j \leq \frac{(m+1)\varepsilon_0}{k-m+1} + \varepsilon_{m+1} \ . \end{aligned}$$

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, moreover,

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

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

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

It remains to consider the last term in (17). 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

$$\frac{1}{n} C_{\varepsilon_k}\sqrt{n - N_k} \le \frac{M_k^{1/8}\sqrt{n - N_k}}{N_k} \le \frac{M_k^{1/8} M_k^{3/8}}{M_{k-1}} = \frac{\sqrt{M_k}}{M_{k-1}} \ ,$$

while in the latter case we get

$$\frac{1}{n} C_{\varepsilon_k}\sqrt{n - N_k} \le \frac{M_k^{1/8}\sqrt{M_k}}{M_k^{3/4}} = \frac{1}{M_k^{1/8}} \to 0$$

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

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

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 $j \to +\infty$.