A Game of Prediction with Expert Advice*

V. Vovk

Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, United Kingdom
E-mail: vovk@dcshbn.ac.uk

Received October 20, 1995; revised October 24, 1996

1. MAIN RESULT

Our learning protocol is as follows. We consider a learner who acts in the following environment. There are a pool of n experts and the nature, which interact with the learner in the following way. At each trial t, t = 1, 2, ...:

1. Each expert i, i = 1, ..., n, makes a prediction γ'i ∈ Γ, where Γ is a fixed prediction space.

2. The learner, who is allowed to see all γ'i, i = 1, ..., n, makes his own prediction γ ∈ Γ.

3. The nature chooses some outcome ω ∈ Ω, where Ω is a fixed outcome space.

4. Each expert i, i = 1, ..., n, incurs loss λ(ω, γ'i) and the learner incurs loss λ(ω, γ), where λ: Ω × Γ → [0, ∞] is a fixed loss function.

We will call the triple (Ω, Γ, λ) our local game. In essence, this is the framework introduced by Littlestone and Warmuth [23] and also studied in, e.g., Cesa-Bianchi et al. [3, 4], Foster [11], Foster and Vohra [12], Freund and Schapire [14], Haussler et al. [15], Littlestone and Long [21], Vovk [31], and Yamanishi [37]. Admitting the possibility of λ(ω, γ) = ∞ is essential for, say, the logarithmic game (see Example 5 below).

One possible strategy for the learner, the Aggregating Algorithm, was proposed in [31]. (That algorithm is described in Appendix A, which is virtually independent of the rest of the paper, and the reader who is mainly interested in the algorithm itself, rather than its properties, might wish to go directly to it.)

If the learner uses the Aggregating Algorithm, then at each time t his cumulative loss is bounded by cL* + a ln n, where c and a are constants that depend only on the local game (Ω, Γ, λ) and L* is the cumulative loss incurred by the best, by time t, expert (see [31]). This motivates considering the following perfect-information game 9(c, a) (the global game) between two players, L (the learner) and E (the environment).

E chooses n ≥ 1 {size of the pool}
FOR i = 1, ..., n
L0(i) := 0 {loss incurred by expert i}
END FOR
L0 := 0 {loss incurred by the learner}
FOR t = 1, 2, ...
FOR i = 1, ..., n
E chooses γ'i ∈ Γ {expert i’s prediction}
END FOR
L chooses γ ∈ Γ {learner’s prediction}
E chooses ω ∈ Ω {outcome}
FOR i = 1, ..., n
L(ω, γ'i) := L(i) - cL(ω, γ'i)
END FOR
L := L - cL(ω, γ)

Player L wins if, for all i and t,

\[ L_t \leq cL^*(t) + a \ln n; \]  \(1\)

otherwise, player E wins.
Let us call the boundary of the set stated otherwise, "strategy" means "deterministic strategy") has a winning strategy in the global game \(G\) (the true outcome is known) by the function

\[
\text{c}(\beta) := \lim_{\beta \to 0} c(\beta), \quad c(1) := \lim_{\beta \to 1} c(\beta),
\]

\[
(a(0) := \lim_{\beta \to 0} a(\beta), \quad a(1) := \lim_{\beta \to 1} a(\beta).
\]

This lemma will be proven in Section 3. Now we can formulate our main result, which will be proven in Sections 4 and 6.

**Theorem 1.** The separation curve is exactly the set

\[
\{(c(\beta), a(\beta)) \mid \beta \in [0, 1]\} \cap [0, \infty]^2.
\]

We conclude this section by discussing how this theorem determines the whole set \(L\). We will see that \(L\) consists of the points on the separation curve and the points "Northeast of" the separation curve. (Lemmas 9, 10, and 11 in Section 3 below might be helpful in visualizing this result.)

The following lemma is a special case of Martin's theorem as presented in [24, Cor. 1].

**Lemma 2.** Each game \(\mathcal{G}(c, a)\) is determined: either \(L\) or \(E\) has a winning strategy.

We say that a point \((c, a)\) is Northeast (resp. Southwest) of a set \(A \subseteq [0, \infty]^2\) if some point \((c', a')\) in \(A\) satisfies \(c' \leq c\) and \(a' \geq a\) (resp. \(c' \geq c\) and \(a' \leq a\)). Suppose the separation curve is nonempty. (It can be empty even when \(\Omega = [0, 1]\) and \(I\) is countable—see Example 6 and Lemma 9 below; in this paper, however, we are not interested in this case. In all examples considered in [15] and [31] the separation curve is nonempty.) It is easy to see that the points \((c, a)\) such that \(\mathcal{G}(c, a) \leftarrow L\) (resp. \(\mathcal{G}(c, a) \leftarrow E\)) are Northeast (resp. Southwest) of the separation curve. Besides, no point outside the separation curve can lie both Northeast and Southwest of the separation curve. The following simple consequence of Assumptions 1 and 2 completes the picture.

**Lemma 3.** \(\mathcal{G}(c, a) \leftarrow L\) when \((c, a)\) belongs to the separation curve.

**Proof.** Notice that L's strategy that beats every oblivious strategy for E will beat every strategy for E (E's strategy is oblivious if it does not depend on the predictions made by the learner); therefore, without loss of generality we can assume that E follows an oblivious strategy. Let \((c_k, a_k)\), \(k = 1, 2, \ldots\) be a sequence of points in \(L\) such that \(c_k \downarrow c\) and \(a_k \downarrow a\) as \(k \to \infty\). For each \(k\) fix a winning strategy \(\mathcal{G}_k\) for L in \(\mathcal{G}(c_k, a_k)\). The learner will win \(\mathcal{G}(c, a)\) acting as follows: if \(\gamma_k\) is the action
suggested by $S^*_k$, L chooses an action that is a limit point (recall Assumption 1) of the sequence $(\gamma_n)_{n=1}^\infty$. Indeed, Assumption 2 implies that, for all $L_t$ (the learner’s cumulative loss by the end of trial $t$) will be a limit point of $L_t[k]$ ($S^*_k$’s cumulative loss by the end of trial $t$ on the actual outcomes); therefore, $L_t \leq c_k L^*_k + a_k \ln n$, $\forall k$, and, consequently, $L_t \leq cL^*_k + a \ln n$. This argument does not work in the case $c = 0$ and $L^*_k = \infty$, but we will soon see (Lemma 5 below) that the separation curve does not intersect the strip $c < 1$.

Assumptions 1 and 2 allow us to strengthen Assumption 4 to:

**Lemma 4.** For some finite set $T \subseteq \Omega$, there exists no $\gamma$ such that, for all $\alpha \in Y$, $\lambda(\alpha, \gamma) = 0$.

**Proof.** Assumption 4 was that

\[ \forall \gamma \in \Gamma \; \exists \alpha \in \Omega: \lambda(\alpha, \gamma) > 0, \tag{3} \]

and we are required to prove that $\Omega$ can be replaced by its finite subset. It suffices to note that (3) means that the sets $\Gamma(\alpha) := \{ \gamma \in \Gamma | \lambda(\alpha, \gamma) > 0 \}$ constitute an open cover of $\Gamma$, which, by Assumption 1, has a finite subcover.

**Lemma 5.** Each game $\mathcal{G}(c, a)$, $c < 1$, is determined in favor of E.

**Proof.** Fix $c < 1$ and $a$. Fix $Y$ whose existence is asserted in Lemma 4, let $\alpha_1, \ldots, \alpha_N$ be an enumeration (without repetition) of the elements of $Y$. Put $d_0 := \infty$, $\Gamma_0 := \{ \gamma \in \Gamma | \lambda(\alpha, \gamma) \leq c, \forall \alpha \in Y \}$. (Here $c$ is a constant large enough for $\Gamma_0$ to be nonempty; the existence of such $C$ follows from Assumption 3), and, for $k = 1, \ldots, N$:

\[
\begin{align*}
d_k &:= \inf \{ \lambda(\alpha_k, \gamma) | \gamma \in \Gamma_{k-1} \}; \\
\Gamma_k &:= \{ \gamma \in \Gamma_{k-1} | \lambda(\alpha_k, \gamma) = d_k \}.
\end{align*}
\]

By induction in $k$ we can prove that each infimum $d_k$ is attained and each $\Gamma_k$ is a nonempty compact set. By the choice of $Y$, not all $d_k$ are zero. Now we can describe E’s strategy that beats L already in the first trial: $n = 1$, the only expert predicts with any $\gamma \in \Gamma_1$, and the way in which $\alpha$ is generated depends on whether the prediction $\delta$ made by L belongs to $\Gamma_0$. If $\delta \in \Gamma_0$, the nature produces outcome $\alpha_1$, where $j := \min \{ k | \lambda(\alpha_k, \delta) > 0 \}$. (By the choice of $Y$, $j < \infty$.) It is easy to see that $\lambda(\alpha_j, \delta) \geq d_j$, so L loses the game. If $\delta \notin \Gamma_0$, the nature produces any $\alpha \in Y$ such that $\lambda(\alpha, \delta) > C$.

**Remark.** Most of all we are interested in the oblivious strategies for the environment (i.e., strategies that ignore the learner’s actions). It is easy to see that Theorem 1 implies the following: Player L has a strategy in $\mathcal{G}(c, a)$ that beats every oblivious strategy for E if and only if $(c, a)$ is Northeast of the curve $(c(\beta), a(\beta))$. (Part “only if” follows from the observation, already used in the proof of Lemma 3, that L’s strategy that beats every oblivious strategy for E will beat every strategy for E.)

Theorem 1 will be proven in Sections 4–6. In Section 4 we describe the learner’s strategy (the Aggregating Algorithm), in Section 6 we describe the environment’s probabilistic strategy, and in Section 5 we state several probability-theoretic results that we need in Section 6.

2. EXAMPLES

In our first several examples we will have $\Omega := \{ 0, 1 \}$. In this case there is a convenient representation for $c(\beta)$ [31, Section 1]. For each set $A \subseteq [-\infty, \infty]^2$ and point $u \in \mathbb{R}^2$ we define the shift $u + A$ of $A$ in direction $u$ to be $\{ u + v | v \in A \}$. The A-closure of $B \subseteq [-\infty, \infty]^2$ is

\[
\text{cl}_A B := \bigcap_{u \in \mathbb{R}^2} B + u 
\]

We write $\text{cl}_A$ for $\text{cl}_A B$ and $\overline{\text{cl}}_A$ for $\text{cl}_A B$, where

\[
A := \{ (x, y) \in [-\infty, \infty]^2 | x \geq 0 \text{ or } y \geq 0 \}, \quad A_y := \{ (x, y) \in [-\infty, \infty]^2 | \beta^x + \beta^y \leq 1 \}.
\]

For any set $B \subseteq [0, \infty]^2$ put

\[
\text{osc}_B := \sup_{x \in [-\infty, \infty]} \text{inf} \{ z | z \in B \},
\]

where $e$ ranges over the vectors in $\mathbb{R}^2$ of length 1, $z$ ranges over $[0, \infty]$, and the conventions for the “extreme cases” are as follows: sup $\emptyset := 0$, inf $\emptyset := \infty$, $\infty/\infty := 1$. It is easy to show [31] that

\[
c(\beta) = \text{osc}(\overline{\text{cl}}_D D) \text{cl}_D D, \quad \forall \beta \in [0, 1], \tag{4}
\]

where $D$ being the graph $\{ (0, 0), (1, 1) | (x, y) \in \Gamma \} \subseteq [0, \infty]^2$ of our local game. (This follows from the fact that the only images of straight lines under the mapping $\log \beta$ that go from the Northwest to the Southeast are shifts of the curve $\beta^{2x} + \beta^{y} = 1$.)

In the examples given below we will use the following simple observation: there exists a shift of the curve $\beta^x + \beta^y = 1$ containing points $(x_1, y_1)$ and $(x_2, y_2)$ if and only if $(y_1 - y_2)(x_2 - x_1) > 0$; if this condition is satisfied, there is only one such shift, namely

\[
(\beta^{x_1} - \beta^{y_1})(\beta^x + (\beta^{x_2} - \beta^{y_2}) \beta^y = \beta^{x_2} + \gamma - \beta^{y_1} + \gamma, \tag{5}
\]

(This can be checked by direct substitution of $(x, y) := (x_1, y_1)$ and $(x, y) := (x_2, y_2)$ into (5).)
Example 1 (Simple Prediction Game; Littlestone and Warmuth [23]). Here $\Omega = \Gamma = \{0, 1\}$ and

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

It is easy to check (using (4) and the observation above) that

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

Example 2 (J. C. Thompson; Dawid [8]). Again $\Omega = \Gamma = \{0, 1\}$. In this example (and also in Example 7 below) we will say “action” instead of “prediction.” The learner is a fruit farmer who can take protective measures (action $\gamma = 1$) to guard her plants against frost (outcome $\omega = 1$) at a cost of $a > 0$; if she does not protect ($\gamma = 0$), she faces a loss of $b > a$ if there is frost ($\omega = 1$), and 0 if not ($\omega = 0$). Therefore, the loss function is

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

Here $c(\beta)$ is the solution to

$$(\beta^a - \beta^b) \beta^{\psi(\beta)} + (1 - \beta^a) \beta^{\psi(\beta)} = \beta^a - \beta^{a+b}.$$}

For $a = 1$ and $b = 2$, we can solve this equation and obtain

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

Example 3 (Absolute Loss Game). Here $\Omega = \{0, 1\}$, $\Gamma = \{0, 1\}$, and $\lambda(\omega, \gamma) = |\omega - \gamma|$. As can be easily seen from the result of Example 1, now

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

As proven in Haussler et al. [15, Sect. 4.2], $c(\beta)$ is also given by this formula when $\Omega = \{0, 1\}$.

Example 4 (Brier Game). Here $\Omega = \{0, 1\}$, $\Gamma = \{0, 1\}$, and $\lambda(\omega, \gamma) = (\omega - \gamma)^2$. In this game, $c(\beta) = 1$ for $\beta \geq e^{-2}[31, 15]$. As shown by Haussler et al. [15, Example 4.4], this is also true when $\Omega = \{0, 1\}$.

Example 5 (Logarithmic Game). Here $\Omega = \{0, 1\}$, $\Gamma = \{0, 1\}$.

$$\lambda(\omega, \gamma) = \omega \ln \frac{e^\omega}{\gamma} + (1 - \omega) \ln \frac{1 - e^\omega}{1 - \gamma}.$$ 

Now $c(\beta) = 1$ for $\beta \geq e^{-1}$ (DeSantis et al. [9]); Haussler et al. [15, Example 4.3] prove that this is true for $\Omega = \{0, 1\}$ as well.

Example 6. This example is rather artificial; it demonstrates that it is possible that $c(\beta) = \infty$, for some $\beta \in ]0, 1[$. Let $\varepsilon_1, \varepsilon_2, ...$ be a decreasing sequence of positive numbers such that $\varepsilon_k \to 0$ very fast; e.g., it suffices to take

$$\varepsilon_1 := \frac{1}{2}, \quad \varepsilon_k + 1 := \varepsilon_k^k, \quad k \geq 1.$$ 

Fix arbitrary $\beta \in ]0, 1[$ and put

$$\Omega := \{0, 1\}, \quad \Gamma := \{1, 2, ..., \infty\}, \quad \lambda(0, \gamma) := \varepsilon_\gamma, \quad \lambda(1, \gamma) := \max(\log_\beta \varepsilon_\gamma, 0), \quad \forall \gamma$$

(with $\varepsilon_\infty$ interpreted as 0). Checking Assumptions 1–4 is straightforward.

Now we prove that $c(\beta) = \infty$. Let $K$ be an arbitrarily large number; we will prove that $c(\beta) \geq K$. As can be easily seen from representation (4), it suffices to prove that, for large $k$, the point $(1/K)(\lambda(0, k), \lambda(1, k + 1))$ of the $(x, y)$-plane lies Northeast of the shift of the curve $\beta^x + \beta^y = 1$ that goes through the points $(\lambda(0, k), \lambda(1, k))$ and $(\lambda(0, k + 1), \lambda(1, k + 1))$ corresponding to the predictions $k$ and $k + 1$, respectively. Since this shift is

$$(\beta^{\lambda(1, k)} - \beta^{\lambda(1, k + 1)}) \beta^{\lambda(0, k + 1)} + (\beta^{\lambda(0, k + 1)} - \beta^{\lambda(0, k)}) \beta^{\lambda(1, k + 1)} = \beta^{\lambda(0, k + 1) + \lambda(1, k)} - \beta^{\lambda(0, k) + \lambda(1, k + 1)},$$

(cf. (5)), we are required to prove that

$$(\beta^{\lambda(1, k)} - \beta^{\lambda(1, k + 1)}) \beta^{\lambda(0, k)} + (\beta^{\lambda(0, k + 1)} - \beta^{\lambda(0, k)}) \beta^{\lambda(1, k + 1)} < \beta^{\lambda(0, k + 1) + \lambda(1, k)} - \beta^{\lambda(0, k) + \lambda(1, k + 1)},$$

i.e.,

$$(\varepsilon_k - \varepsilon_{k + 1}) \beta^{\lambda(K)\varepsilon_k} + (\beta^{\varepsilon_k + 1} - \beta^{\varepsilon_k}) \varepsilon_k^{\lambda(K)\varepsilon_k + 1} < \beta^{\varepsilon_k + 1} \varepsilon_k - \beta^{\varepsilon_k + 1} \varepsilon_k.$$ 

Since, as $\delta \to 0$,

$$\beta^\delta = e^{-\delta \ln(1/\beta)} = 1 - \delta \ln \frac{1}{\beta} + o(\delta),$$
we can rewrite this inequality as
\[
(e_k - e_{k+1}) \left( 1 - \frac{e_k}{K} \ln \frac{1}{\beta} + o(e_k) \right)
+ (e_k - e_{k+1} + o(e_k)) \left( \ln \frac{1}{\beta} \right) e_{k+1}^{\frac{1}{K}}
< \left( 1 - e_{k+1} \ln \frac{1}{\beta} + o(e_{k+1}) \right) e_k
- \left( 1 - e_k \ln \frac{1}{\beta} + o(e_k) \right) e_{k+1},
\]
which simplifies to
\[-(e_k - e_{k+1}) \frac{e_k}{K} + (e_k - e_{k+1}) e_{k+1}^{\frac{1}{K}} + o(e_k^2) < 0.
\]

It remains to notice that, as \(k \to \infty\),
\[-(e_k - e_{k+1}) \frac{e_k}{K} + (e_k - e_{k+1}) e_{k+1}^{\frac{1}{K}} + o(e_k^2) = 0. \]

\[\text{EXAMPLE 7 (Freund and Schapire [14])}. \text{ The learner is a gambler; he has } K \text{ friends who are very successful in horse-race betting. Frustrated by his own persistent losses, he decides he will wager a fixed sum of money in every race but that he will apportion his money among his friends based on how well they are doing. His goal is to allocate each race’s wager in such a way that his total winnings will be reasonably close to what he would have won had he bet everything with the luckiest of his friends.}

\text{Freund and Schapire formalize this game as follows (their terminology is different from ours). The outcome space } \Omega \text{ is } \{0, 1\}^k; \text{ an outcome } \omega = \omega_1 \cdots \omega_K \text{ of trial } t \text{ means that friend } k \text{’s loss would be } \omega_k \in \{0, 1\} \text{ had the gambler given him all his money (emararked for trial } t). \text{ The action space } \Gamma \text{ is the set of all vectors } \gamma = \gamma_1 \cdots \gamma_K \text{ such that } \gamma_1 + \cdots + \gamma_K = 1; \text{ action } \gamma = \gamma_1 \cdots \gamma_K \text{ means that the gambler gives fraction } \gamma_k \text{ of his money to friend } k, k = 1, \ldots, K. \text{ The loss function, mixture loss, is given by the dot product}
\[
\lambda(\omega, \gamma) := \gamma \cdot \omega = \sum_{k=1}^{K} \gamma_k \omega_k.
\]

\text{Freund and Schapire are interested in the experts who are actually the same persons as the friends: expert } k, k = 1, \ldots, K, \text{ recommends giving all money at every trial to friend } k. \text{ Under this assumption, they prove [14, Theorem 2] that the gambler can ensure that, for all } t \text{ and } k,
\[
L_t \leq \frac{(\ln(1/\beta)) L_t(k) + K}{1 - \beta}, \quad (6)
\]
in the notation of (1). On the other hand, our Theorem 1 implies that the gambler can ensure that, for all } t \text{ and } k,
\[
L_t \leq c(\beta) L_t(k) + \frac{c(\beta)}{\ln(1/\beta)} \ln K. \quad (7)
\]

\text{Lemmas 6 and 7 below show that}
\[
c(\beta) < \frac{\ln(1/\beta)}{1 - \beta}, \quad \forall \beta \in ]0, 1[;
\]
therefore, (7) is stronger than (6) (Freund and Schapire use the Weighted Majority Algorithm deriving (6)). But Lemma 7 shows that, when } K \text{ is large, } (\ln(1/\beta))/(1 - \beta) \text{ is close to } c(\beta).

\text{LEMMA 6. For Freund and Schapire’s game,}
\[
c(\beta) = \left( \ln \frac{1}{\beta} \right) \left( K \ln \frac{K}{K + \beta - 1} \right).
\]

\text{Proof. We will only prove inequality } \leq \gamma; \text{ the last step of this proof will show that in fact } \text{ holds. We are required to prove that, for any simple probability distribution } P \text{ in } \Gamma, \text{ there exists an action } \delta \in \Gamma \text{ such that, for all } \omega \in \Omega,
\[
\delta \cdot \omega \leq \frac{\ln(1/\beta)}{K \ln(K + \beta - 1))} \log_\gamma \sum_{\gamma'} \beta^{\gamma' \cdot \omega P(\gamma')},
\]
i.e.,
\[
\delta \cdot \omega \leq - \frac{1}{K \ln(K + \beta - 1))} \ln \sum_{\gamma'} \beta^{\gamma' \cdot \omega P(\gamma')}. \quad (8)
\]

\text{First we show that it suffices to prove (8) for } P \text{ concentrated on the extreme points of the simplex } \Gamma. \text{ This simplex has } K \text{ extreme points, viz., } \gamma^k = \gamma_1^k \cdots \gamma_K^k, k = 1, \ldots, K, \text{ where}
\[
\gamma^k_j = \begin{cases} 1, & \text{if } j = k, \\ 0, & \text{otherwise.} \end{cases}
\]

\text{To prove (8), we represent each } \gamma \in \text{dom } P \text{ as a weighted average of the extreme points of } \Gamma,
\[
\gamma = \sum_{k=1}^{K} q_{\gamma,k} \gamma^k, \quad \gamma \in \text{dom } P,
\]
\[
q_{\gamma,k} \in [0, 1], \text{ being the coefficients of the expansions, } \sum_{k=1}^{K} q_{\gamma,k} = 1. \text{ The convexity of the function } \xi \in \mathbb{R} \mapsto \beta^\xi \text{ implies}
\[
\sum_{\gamma} \beta^{\gamma' \cdot \omega P(\gamma)} = \sum_{\gamma} \beta^{\sum_{k=1}^{K} q_{\gamma,k} \gamma^k \cdot \omega P(\gamma)} = \sum_{k} \beta^{\sum_{\gamma} q_{\gamma,k} \gamma^k \cdot \omega P(\gamma)} \leq \sum_{k} \sum_{\gamma} \beta^{\gamma' \cdot \omega q_{\gamma,k} P(\gamma)} = \sum_{k} \beta^{\gamma' \cdot \omega q_{\gamma,k}}
\]
\text{in the notation of (1). On the other hand, our Theorem 1 implies that the gambler can ensure that, for all } t \text{ and } k,
\[
L_t \leq c(\beta) L_t(k) + \frac{c(\beta)}{\ln(1/\beta)} \ln K. \quad (7)
\]
where \( q_k = \sum_{i} q_{i,k} P^i(\gamma) \). Therefore, it suffices to prove that for any weights \( q_1, \ldots, q_2 \in [0, 1] \) for the friends (\( \sum_k q_k = 1 \)), there exists an action \( \delta \in I \) such that, for all \( \omega \in \Omega \),

\[
\delta \cdot \omega \leq -\frac{1}{K \ln(K/(K + \beta - 1))} \ln \sum_{k=1}^{K} \beta^\omega q_k.
\]

Let us prove that the function

\[
\omega \in \mathbb{R}^K \mapsto \ln \sum_{k=1}^{K} \beta^\omega q_k
\]

is convex. Fix any \( \omega, a \in \mathbb{R}^K \). We must only prove that

\[
\zeta \in \mathbb{R} \mapsto \ln \sum_{k} (\beta^{\omega_k} + \zeta a_k) q_k
\]

is convex. Writing this as

\[
\zeta \mapsto \ln \sum_{k} \beta^{\omega_k} e^{\zeta a_k},
\]

we can see that it is sufficient to prove that

\[
\zeta \mapsto \ln \sum_{k} \beta^{\omega_k} e^{\zeta a_k}
\]

is convex; without loss of generality we assume \( \sum_k p_k = 1 \). The convexity of (10) is proven in Lemma 14 below.

Since the right-hand side of (9) is concave, it suffices to prove (9) only for \( \omega \) that are extreme points of the cube \([0, 1]^K\). Letting \( I \) run over the subsets of \( \{1, \ldots, K\} \), we transform (9) to

\[
\sum_{k \in I} \delta_k \leq -\frac{1}{K \ln(K/(K + \beta - 1))} \ln \left( \sum_{k \in I} q_k + \sum_{k \notin I} q_k \right),
\]

i.e.,

\[
\sum_{k \in I} \delta_k \leq -\frac{1}{K \ln(K/(K + \beta - 1))} \ln \left( 1 + (\beta - 1) \sum_{k \in I} q_k \right).
\]

For \( I = \emptyset \), (11) is obviously true, so we assume \( I \neq \emptyset \). Let us show that it suffices to establish (11) only in the case of one-element \( I \). To do so, it suffices to prove that (11) holds for \( I = I_1 \cup I_2 \) as soon as (11) holds for \( I = I_1 \) and \( I = I_2 \), where \( I_1 \) and \( I_2 \) are disjoint nonempty subsets of \( \{1, \ldots, K\} \). The last assertion follows from

\[
-\ln \left( 1 + (\beta - 1) \sum_{k \in I_1} q_k \right) - \ln \left( 1 + (\beta - 1) \sum_{k \in I_2} q_k \right)
\]

\[
\leq -\ln \left( 1 + (\beta - 1) \sum_{k \in I_1 \cup I_2} q_k \right).
\]

which is essentially a special case of

\[
(1 + x_1)(1 + x_2) \geq 1 + (x_1 + x_2),
\]

where \( x_1, x_2 \leq 0 \).

Substituting \( I := \{k\} \), we get

\[
\delta_k \leq -\frac{1}{K \ln(K/(K + \beta - 1))} \ln(1 + (\beta - 1) q_k).
\]

The existence of such \( \delta \) will follow from

\[
-\frac{1}{K \ln(K/(K + \beta - 1))} \sum_{k=1}^{K} \ln(1 + (\beta - 1) q_k) \geq 1,
\]

i.e.,

\[
\sum_{k=1}^{K} \ln(1 + (\beta - 1) q_k) \leq -K \ln \frac{K}{K + \beta - 1};
\]

\[
\prod_{k=1}^{K} (1 + (\beta - 1) q_k) \leq \left( \frac{K + \beta - 1}{K} \right)^K.
\]

It remains to note that

\[
\sum_{k=1}^{K} (1 + (\beta - 1) q_k) = K + \beta - 1
\]

is constant, and so the product attains its maximum when all \( q_k \) are equal, \( q_k = 1/K, k = 1, \ldots, K \).

**Lemma 7.** For \( \beta \in ]0, 1[ \),

\[
K \ln \frac{K}{K + \beta - 1} > 1 - \beta
\]

and, as \( K \to \infty \),

\[
K \ln \frac{K}{K + \beta - 1} = 1 - \beta + O(1/K).
\]

**Proof.** By Taylor’s formula, for some \( \theta \in ]0, 1[ \),

\[
K \ln \frac{K}{K + \beta - 1} = -K \ln \frac{K + \beta - 1}{K} = -K \ln \left( 1 + \frac{\beta - 1}{K} \right)
\]

\[
= -K \left( \frac{\beta - 1}{K} \frac{1}{2} \left( \frac{\beta - 1}{K} \right)^2 \right) \left( 1 + \theta \frac{\beta - 1}{K} \right)^2
\]

\[
= 1 - \beta + \frac{1}{2} (1 - \beta)^2 \frac{1}{K} \left( 1 + \theta \frac{\beta - 1}{K} \right)^2.
\]
3. FUNCTIONS $c(\beta)$ AND $a(\beta)$

In this section we study the curve $\{c(\beta), a(\beta)\mid \beta \in [0, 1]\}$, which will be shown to coincide with the separation curve inside $[0, \infty[^2$.

**Lemma 8.** $c(\beta) \geq 1, \forall \beta$.

**Proof.** If $c(\beta) < 1$ for some $\beta$, then
\[ \forall \gamma \in \Gamma \exists \delta \in \Gamma \forall \omega \in \Omega: \delta(\omega, \delta) < c(\omega, \gamma), \] (12)
where $c$ is a constant between $c(\beta)$ and 1. By Assumption 3, there is $\gamma_1 \in \Gamma$ such that $\delta(\omega, \gamma_1) < \infty$, for all $\omega$. By (12), we can find $\gamma_2, \gamma_3, \ldots$ such that
\[ \delta(\omega, \gamma_k + 1) < c(\omega, \gamma_k). \forall \omega, k = 1, 2, \ldots. \]
Let $\gamma$ be a limit point of the sequence $\gamma_1, \gamma_2, \ldots$. Then, for each $\omega$, $\delta(\omega, \gamma)$ is a limit point of the sequence $\delta(\omega, \gamma_k)$ and, therefore, $\delta(\omega, \gamma) = 0$. The existence of such $\gamma$ would contradict Assumption 4.

This lemma shows that we always have $c(\beta) > 0$ (and hence $a(\beta) > 0$), $\beta$ ranging over $[0, 1]$ (through it is possible that $c(\beta) = \infty$ and $a(\beta) = \infty$: see Example 6 above).

We use the words “increase” and “decrease” in a wide sense: say, an increasing function may be constant on some pieces of its domain.

**Lemma 9.** As $\beta \in [0, 1]$, increases, $c(\beta)$ decreases and $a(\beta)$ increases.

**Proof.** The case of $a(\beta)$ is simple:
\[
 a(\beta) = \frac{1}{\ln(1/\beta)} \inf \{ c \mid \forall P \exists \delta \forall \omega: \delta(\omega, \delta) \leq c \log \beta \sum_{\gamma} P(\gamma_{\omega}) \}
\leq \frac{c}{\ln(1/\beta)} \inf \{ \beta \mid \forall P \exists \delta \forall \omega: \delta(\omega, \delta) \}
\leq \frac{c}{\ln(1/\beta)} \inf \{ \beta \mid \forall P \exists \delta \forall \omega: \delta(\omega, \delta) \}
\leq \frac{c}{\ln(1/\beta)} \inf \{ \beta \mid \forall P \exists \delta \forall \omega: \delta(\omega, \delta) \}
\]
introducing the notation $b := c/\ln(1/\beta)$:
\[
 a(\beta) = \inf \{ b \mid \forall P \exists \delta \forall \omega: \delta(\omega, \delta) \}
\leq b \left( -\ln \sum_{\gamma} P(\gamma_{\omega}) \right). \] (13)
Therefore, it is sufficient to prove that, as $\beta$ increases, $\beta^{\log_{\beta}(\gamma)}$ increases. The last assertion is obvious.

Analogously, in the case of $c(\beta)$ it is sufficient to prove, for each $P$ and $\omega$, that, as $\beta$ increases, $\log_{\beta} \sum_{\gamma} P(\gamma_{\omega})$ also increases. Fix $\alpha$ and $\beta$ such that $0 < \alpha < \beta < 1$. We are required to prove
\[
 \log_{\beta} \alpha E^{\xi} \leq \log_{\beta} E^{\beta^{\xi}}, \] (14)
where $\xi = \xi_{\alpha,p}$ is an extended random variable (i.e., a random variable that is allowed to take value $\infty$) that takes each value $\delta(\omega, \gamma)$ with probability $P(\gamma)$. Let $p > 1$ be such that $\alpha = \beta^p$. We can rewrite (14) as
\[
 \log_{\beta} E^{\beta^{\xi}} \leq \log_{\beta} E^{\beta^{\xi}}; \quad \frac{1}{p} \log_{\beta} E^{\beta^{\xi}} \geq \log_{\beta} E^{\beta^{\xi}}.
\]
We continue putting $\eta := \beta^{\xi}$:
\[
 \frac{1}{p} \log_{\beta} E^{\eta} \geq \log_{\beta} E^{\eta}; \quad (E^{\eta})^{1/p} \geq E^{\eta}.
\]
The last inequality follows from the monotonicity of the $L^p$ norms (see, e.g., Williams [36, Section 6.7]).

This lemma implies Lemma 1. In addition, it implies that we have only two possibilities for each local game: either $c(\beta)$ and $a(\beta)$ are finite for all $\beta \in [0, 1]$, (1) (in view of Theorem 1), this means that the separation curve is empty; or else $c(\beta)$ and $a(\beta)$ are finite for all $\beta \in [0, 1]$. (1). If, say, $c(\beta)$ is finite for some $\beta$, then $a(\beta)$ is finite for this $\beta$; $a(\alpha)$ and, hence, $c(\alpha)$ are finite for $\alpha < \beta$; $c(\alpha)$ and, hence, $a(\alpha)$ are finite for $\alpha > \beta$.) In particular, we have $c(\beta) = a(\beta) = \infty$, $\forall \beta \in [0, 1]$, for the local game of Example 6.

**Lemma 10.** The functions $c(\beta)$ and $a(\beta)$ are continuous.

**Proof.** It suffices to consider only the case of $a(\beta)$ (since $c(\beta) = a(\beta) \ln(1/\beta)$). Fix any $\beta \in [0, 1]$. Since $a(\beta)$ increases (see Lemma 9), the values
\[
 a(\beta-) := \lim_{\beta \to \alpha^+} a(\alpha), \quad a(\beta+) := \lim_{\beta \to \alpha^+} a(\alpha)
\]
exist and $a(\beta-) \leq a(\beta) \leq a(\beta+)$. Since $c(\beta)$ decreases, we have
\[
 a(\beta-) = \lim_{\beta \to \alpha^+} \left( c(\alpha) \frac{1}{\ln(1/\beta)} \right) = c(\beta-) \frac{1}{\ln(1/\beta)} \geq a(\beta),
\]
\[
 a(\beta+) = \lim_{\beta \to \alpha^+} \left( c(\alpha) \frac{1}{\ln(1/\beta)} \right) = c(\beta+) \frac{1}{\ln(1/\beta)} \leq a(\beta).
\]
Therefore, $a(\beta-) = a(\beta) = a(\beta+)$, which means that $a(\alpha)$ is continuous at $\beta$. 

---

File: DISTL2 155607. By CV, Date: 03/06/98, Time: 11:52 LOPM. V&R Page 01:01
Codes: 5084 Signs: 2998. Length: 56 pic 0 pts, 236 mm
LEMMA 11. Suppose $c(\beta) < \infty$ ($\beta \in ]0, 1[$). The following three sets form a partition of the quadrant $[0, \infty]^2$ of the $(c, \alpha)$-plane:

- the points on the curve $\{(c, \beta), a(\beta)\} | \beta \in [0, 1] \}$;
- the points Northeast of and outside this curve;
- the points Southwest of and outside this curve.

In other words, these three sets are pairwise disjoint and their union coincides with the whole quadrant.

Proof. It suffices to note that, since

$$a(0) = \lim_{\beta \to 0} \frac{c(\beta)}{\ln(1/\beta)} = 0, \quad a(1) = \lim_{\beta \to 1} \frac{c(\beta)}{\ln(1/\beta)} = \infty,$$

the points $(c(0), a(0))$ and $(c(1), a(1))$ lie on the border of the square $[0, 0)(\infty, 0)(\infty, \infty)(0, \infty)$ of the extended $(c, \alpha)$-plane.

4. STRATEGY FOR THE LEARNER

In this section we will prove one half of Theorem 1: each game $\mathcal{G}(c, \beta, a(\beta))$ with $(c, \beta, a(\beta)) \in ]0, \infty]^2$ is determined in favor of $L$. The proof will follow from a simple analysis of the Aggregating Algorithm; this section is, however, independent of Appendix A (where we describe the Aggregating Algorithm in a “pure” form, disentangled from its analysis).

We begin with a simple lemma.

LEMMA 12. For each $\beta \in ]0, 1[$, the infimum in the definition of $c(\beta)$ is attained.

Proof. We are required to prove

$$\forall P \exists \delta \forall \omega: \lambda(\alpha, \delta) \leq c(\beta) \log \beta \sum_{\gamma} \beta^{t(\alpha, \gamma)}P(\gamma). \quad (15)$$

Fix any $P$; let $c_1c_2...$ be a decreasing sequence such that $c_k \to c(\beta)$ as $k \to \infty$. By the definition of $c(\beta)$, for each $k$ there exists $\delta_k$ such that

$$\forall \omega: \lambda(\alpha, \delta_k) \leq c_k \log \beta \sum_{\gamma} \beta^{t(\alpha, \gamma)}P(\gamma).$$

Let $\delta$ be a limit point (whose existence follows from Assumption 1) of the sequence $\delta_1, \delta_2...$. Then, for each $\omega$, $\lambda(\alpha, \delta)$ is a limit point of the sequence $\lambda(\alpha, \delta_k)$ (by Assumption 2) and, therefore,

$$\lambda(\alpha, \delta) \leq c(\beta) \log \beta \sum_{\gamma} \beta^{t(\alpha, \gamma)}P(\gamma)$$

(recall that, by Lemma 8, $c(\beta) > 0$).

First we make the learner’s task easier: he is allowed to make predictions that are functions of the outcome $\omega$, $\mathcal{G} \to [0, \infty]$ and incurs loss $g(\omega)$, where $\omega$ is the outcome chosen by the nature; we will only require that $g$ can be represented as mixture $(2)$. Now the learner can act simply by computing the weighted average of the experts’ suggestions: at trial $t+1$, $t = 0, 1, ...$, his prediction $\gamma_{t+1}$ is defined by the equality

$$\beta_{t+1} = \frac{\sum_{i=1}^{n} \beta^{t(\alpha, \gamma_{t+1}^{(i)})} \beta_{t+1}^{(i)}}{\sum_{i=1}^{n} \beta^{t(\alpha, \gamma_{t+1}^{(i)})}}, \quad \forall \omega. \quad (16)$$

When $L$ uses this strategy, we have

$$\beta_{t+1} = \frac{1}{n} \sum_{i=1}^{n} \beta^{t(\alpha, \gamma_{t+1}^{(i)})}, \quad \forall \omega. \quad (17)$$

Indeed, for $t=0$ this equality is obvious, and an easy induction in $t$ gives

$$\beta_{t+1} = \beta_{t} + \frac{1}{n} \sum_{i=1}^{n} \beta^{t(\alpha, \gamma_{t+1}^{(i)})}. \quad (18)$$

By Lemma 12, however, he can make predictions $\gamma_{t+1}$ with

$$\lambda(\alpha, \gamma_{t+1}) \leq c(\beta) \gamma_{t+1}(\alpha), \quad \forall \omega,$$

in this case we will have, instead of $(18)$,

$$L_t \leq c(\beta) \log \beta \sum_{\gamma} \beta^{t(\alpha, \gamma)}P(\gamma). \quad (19)$$

So far we have assumed that the denominator of the ratio in $(16)$ is positive; in the case where it is zero, the learner can choose $\gamma_{t+1}$ arbitrarily.

It remains to consider the case $\beta \in [0, 1]$. If $\beta = 1$, we have $a(\beta) = \infty$ (by Lemma 8). Therefore, we assume $\beta = 0$.

LEMMA 13.

$$c(0) = \inf \{ c | \forall \delta \exists \omega: \lambda(\alpha, \delta) \leq c \min_{\gamma \in D} \lambda(\alpha, \gamma) \}, \quad (19)$$

$D$ ranging over the finite subsets of $\Gamma$; this infimum is attained.
Proof. Let \( e^* \) stand for the right-hand side of (19). For all \( \beta \in [0, 1] \), we have \( c(\beta) \leq e^* \), which implies, by Lemma 10 (the continuity of \( c(\beta) \)), \( c(0) \leq e^* \).

Let us prove \( e^* < c(0) \). We assume \( c(0) < \infty \). For each \( \beta \in [0, 1] \), we have

\[
\forall \beta \in [0, 1] \quad e(\alpha, \beta) \leq c(0) \log \left( \frac{1}{|D|} \sum_{\gamma \in D} \beta^{\alpha(\gamma)} \right)
\]

(see Lemmas 12 and 9); taking \( P \) to be the uniform probability distribution in a finite set \( D \subseteq \Gamma \), we get

\[
\forall \beta \in [0, 1] \quad e(\alpha, \beta) \leq c(0) \log \left( \frac{1}{|D|} \sum_{\gamma \in D} \beta^{\alpha(\gamma)} \right)
\]

| \( D \), or \( \# D \), stands for the number of elements in a set \( D \). Let \( \beta_1, \beta_2, \ldots \) be a decreasing sequence of numbers in \( [0, 1] \) such that \( \text{inf}_{\beta_k} \beta_k = 0 \). For each \( \beta \) let \( \delta_{\beta} \) be a limit point of the sequence \( \delta(\beta_k), k = 1, 2, \ldots \). Then, for each \( \alpha, e(\alpha, \delta_{\beta}) \) is a limit point of \( \delta(\alpha, \delta_{\beta}(\beta_k)) \), and we get

\[
\forall \beta \in [0, 1] \quad \lim_{\beta \to 0} \log \left( \frac{1}{|D|} \sum_{\gamma \in D} \beta^{\alpha(\gamma)} \right) = c(0) \min_{\gamma \in D} \delta(\alpha, \gamma).
\]

Recalling the definition of \( e^* \), we obtain \( e^* < c(0) \). \( \square \)

We are only required to consider the case \( c(0) < \infty \). Now L’s goal is to ensure \( L_i \leq c(0) \) \( L_i(i) \) (we have \( a(0) = 0 \) here). We replace (16) by

\[
g_{t+1}(\alpha) = \min_k \delta(\alpha, \gamma_{t+1}(i)), \quad \forall \alpha.
\]

Instead of (18), we have now \( L_i \leq L_i(i) \), for all \( i, t \), and Lemma 13 ensures that L can attain his goal.

5. LARGE DEVIATIONS

In this section we introduce notions and state assertions that will be used to analyze the probabilistic strategy for the environment presented in the next section.

First we recall some simple results of convex analysis. For each function \( \phi : \mathbb{R} \to [-\infty, \infty] \) its Young-Fenchel transform \( \phi^*: \mathbb{R} \to [-\infty, \infty] \) is defined by the equality

\[
\phi^*(x) := \sup_{\zeta} (\zeta x - \phi(\zeta)).
\]

Function \( \phi^* \) is convex and closed. If \( \phi \) is convex and closed and does not take value \( -\infty \), the Young-Fenchel transform \( \phi^{**} \) of \( \phi^* \) coincides with \( \phi \) (this is part of the Fenchel-Moreau theorem; see, e.g., [1, Subsection 2.6.3]).

Now we can move on to the main topic of this section, the theory of large deviations. The material of this section is well known (cf., e.g., Borovkov [2, Section 8.8]).

In this paper we usually consider simple random variables \( \xi \), which take only finitely many values. The distribution of such \( \xi \) is completely determined by the probabilities \( \text{prob}\{\xi = y\}, y \in \mathbb{R} \) (these probabilities are different from 0 for only finitely many \( y \)). We will identify simple random variables and the corresponding probability distributions in \( \mathbb{R} \).

Let \( \xi \) be a simple random variable. For each \( \xi \in \mathbb{R} \) we put

\[
\psi(\xi) := \sum_y e^{i\gamma} \text{prob}\{\xi = y\}
\]

(20)

(this is the moment generating function) and define a new simple random variable \( \eta \) by

\[
\text{prob}\{\eta = y\} = \frac{1}{\psi(\xi)} e^{i\gamma} \text{prob}\{\xi = y\}, \quad \forall y \in \mathbb{R}
\]

(21)

(sometimes \( \eta \) is called Cramér’s transform of \( \xi \), but we will use this term in a different sense). Note that

\[
\mathbb{E}(\eta) = \sum_y y \frac{1}{\psi(\xi)} e^{i\gamma} \text{prob}\{\xi = y\} = \frac{\psi'(\xi)}{\psi(\xi)}
\]

(22)

and

\[
\text{var} \eta = \sum_y y^2 \frac{1}{\psi(\xi)} e^{i\gamma} \text{prob}\{\xi = y\} - \mathbb{E}^2\eta
\]

\[
= \frac{\psi'(\xi)}{\psi(\xi)} - \left( \frac{\psi'(\xi)}{\psi(\xi)} \right)^2 = (\ln \psi)^*(\xi).
\]

(23)

Since the variance is always nonnegative, we obtain

LEMMA 14. The function \( \ln \psi \) is convex.

Put

\[
S_N := \sum_{k=1}^N \xi_k, \quad Z_N := \sum_{k=1}^N \eta_k,
\]

where \( \xi_k \) (resp. \( \eta_k \)) are independent random variables distributed as \( \xi \) (resp. \( \eta \)).

LEMMA 15. For all \( N \) and \( z \),

\[
\text{prob}\{Z_N = z\} = \frac{1}{\psi(\xi)} e^{i\gamma} \text{prob}\{S_N = z\}.
\]
\textbf{Proof.} We find
\[
\Pr\{Z_N = z\} = \sum \Pr\{\eta_1 = y_1, \ldots, \eta_N = y_N\} = \sum \frac{1}{\psi(z)} e^{\psi(z)} \Pr\{\xi = y_1\} \ldots \\
\times \frac{1}{\psi(z)} e^{\psi(z)} \Pr\{\xi_N = y_N\} = \frac{1}{\psi(z)} e^{\psi(z)} \Pr\{S_N = z\},
\]
all sums being over the \(N\)-tuples \(y_1, \ldots, y_N\) such that \(y_1 + \cdots + y_N = z\).

We are interested in the probability \(\Pr\{S_N \leq \alpha N\}\), \(\alpha\) being some constant. Lemma 17 below gives a lower estimate for this probability in terms of the Young–Fenchel transform
\[
A(x) := \sup(\alpha \xi - \ln \psi(\xi))
\]
of the convex function \(\ln \psi\); we will say that \(A = A_\xi\) is Cramér’s transform of the random variable \(\xi\). First we state some basic properties of Cramér’s transform; they are proved in Appendix B.

\textbf{Lemma 16.} For each simple random variable \(\xi\), \(A = A_\xi\) satisfies
\[
A(x) < \infty \iff x \in \{\min \xi, \max \xi\};
\]
\[
\min \xi, \max \xi \Rightarrow A(x) = -\ln \Pr\{\xi = x\};
\]
\(A\) is continuous on \(\min \xi, \max \xi\). If \(\var{\xi} > 0\), \(A\) is a smooth (i.e., infinitely differentiable) function on \(\min \xi, \max \xi\) and
\[
A'(E_\xi) = A'(E_\xi) = 0, \quad A''(E_\xi) = \frac{1}{\var{\xi}}.
\]
If \(\var{\xi} = 0\),
\[
A(x) = \begin{cases} 0, & \text{if} \quad x = E_\xi, \\ \infty, & \text{otherwise}. \end{cases}
\]

The following two lemmas are also proved in Appendix B; the idea of their proof is to express the probability \(\Pr\{S_N \leq \alpha N\}\) through probabilities \(\Pr\{Z_N = z\}\) (see Lemma 15) and approximate the latter by a Gaussian distribution.

\textbf{Lemma 17.} Let \(\xi\) be a simple random variable and \(\alpha \in \mathbb{R}\). There is a constant \(C = C(\xi, \alpha)\) such that, for all \(N \geq 0\),
\[
\Pr\{S_N \leq \alpha N\} \geq \frac{1}{C \sqrt{N+1}} \exp(-NA_\xi(\alpha))
\]
(29)
where \(S_N = \xi_1 + \cdots + \xi_N\) and \(\xi_1, \xi_2, \ldots\) are independent random variables distributed as \(\xi\).

We will also consider extended simple random variables \(\xi\), which are allowed to take value \(\infty\) (but not \(-\infty\)). The weight \(w(\xi)\) of such \(\xi\) is defined to be \(\Pr\{\xi < \infty\}\); we will require that \(w(\xi) > 0\). The moment generating function \(\psi\) of \(\xi\) is defined by (20) (which we will also write as \(\psi(\xi) := E e^{\xi}\), \(y\) ranging over \(\mathbb{R}\), and Cramér’s transform \(A = A_\xi\) of \(\xi\) is defined by (25).

\textbf{Lemma 18.} Lemma 17 continues to hold when \(\xi\) is allowed to be an extended simple random variable.

\section{6. Strategy for the Environment}

The aim of this section is to prove the remaining half of Theorem 1. Fix a point \((c, a) \in [0, \infty]^2\) Southwest of and outside the curve \((c(\beta), a(\beta))\); we are required to prove \(1(c, a) = E\). (If \(c(\beta) = \infty\) for \(\beta \in ]0, 1]\), \((c, a)\) is an arbitrary point of \([0, \infty]^2\); recall that either \(c(\beta) = \infty\), \(\forall \beta \in ]0, 1]\), or \(c(\beta) < \infty\), \(\forall \beta \in ]0, 1]\). We will present E’s probabilistic strategy in the global game \(1(c, a)\) which will enable us to prove \(1(c, a) = E\). Fix \(\beta \in ]0, 1[\) such that \(c < c(\beta)\) and \(a < a(\beta)\). (So \(\beta\) can be taken arbitrarily if \(c(\beta) = \infty\), \(\beta \in ]0, 1]\). Since \(c(\beta) = a(\beta)\), \(\ln(1/\beta) > a(\ln 1/\beta)\), we can, increasing \(c\) if necessary, also ensure that
\[
c > a \ln \frac{1}{\beta}.
\]

Fix constant \(\epsilon \in ]c, a(\beta)[\). By the definition of \(c(\beta)\), there is a simple probability distribution \(P\) in \(\Gamma\) such that
\[
\forall \delta \in \Gamma \quad \exists \alpha \in \lambda(\alpha, \delta) > \epsilon \log_\beta \sum_{\gamma} b^{2\alpha(\gamma, \gamma)} P(\gamma).
\]
(31)
For each \(\omega \in \Omega\) define \(\Gamma(\omega)\) to be the set of all \(\delta \in \Gamma\) that satisfy the inequality of (31); (31) asserts that the sets \(\Gamma(\omega)\) compose an open cover of \(\Gamma\). By Assumption 1, there exists a finite subcover \(\{\Gamma(\omega)\} \{\alpha \in T\}\) of this cover. Therefore, we can rewrite (31) as
\[
\forall \delta \in \Gamma \quad \exists \alpha \in T, \lambda(\omega, \delta) > \epsilon \log_\beta \sum_{\gamma} b^{2\alpha(\gamma, \gamma)} P(\gamma).
\]
(32)
We say that \( \omega \in \Gamma \) is **trivial** if \( \lambda(\omega, \gamma) = 0 \), for all \( \gamma \in \text{dom} \, P \).
Let \( \Gamma^* \) be the set of trivial \( \omega \in \Gamma \). Let us fix any function \( v: \Gamma \to \mathbb{R} \) such that, for each \( \delta \in \Gamma \):

- \( \lambda(v(\delta), \delta) > c \log \sum_{\gamma} \beta^h(\delta, \gamma) P(\gamma); \)
- if there exists a nontrivial \( \omega \in \Gamma \) for which the inequality of (32) holds, then \( v(\delta) \) is nontrivial;
- if \( v(\delta) \) is trivial, then

\[
\forall \delta \in \Gamma^*, \quad v(\delta) = \arg \max_{\omega \in \Gamma^*} \lambda(\omega, \delta).
\]

For each \( \omega \in \Gamma \), define \( \xi_{\omega} \) to be the extended simple random variable that takes each value \( \lambda(\omega, \gamma), \gamma \in \text{dom} \, P \), with probability \( P(\gamma) \). (We assume, without loss of generality, that \( \Gamma \) contains no \( \omega \) such that \( \lambda(\omega, \gamma) = \infty \), \( \forall \gamma \in \text{dom} \, P \); therefore, \( w(\xi_{\omega}) > 0 \), for all \( \omega \in \Gamma \).) Let \( A_{\omega} \) be Cramer’s transform of \( \xi_{\omega} \).

**Lemma 19.** There exist constants \( z_{\omega} (\omega \in \Gamma) \) and \( \epsilon > 0 \) such that, for all \( \omega \in \Gamma \),

\[
\inf_{\delta \in \arg \max_{\omega \in \Gamma^*} \lambda(\omega, \delta)} \lambda(\omega, \delta) > c z_{\omega} + a A_{\omega}(z_{\omega}) + \epsilon, \quad (33)
\]

\[
0 \leq A_{\omega}(z_{\omega}) < \infty. \quad (34)
\]

**Proof.** Of course, we can drop \( \epsilon > 0 \) in (33). First we assume that \( \omega \) is nontrivial. In this case, (33) (without \( \epsilon > 0 \)) can be deduced with the following chain of inequalities:

\[
\inf_{\delta \in \arg \max_{\omega \in \Gamma^*} \lambda(\omega, \delta)} \lambda(\omega, \delta) > c \log \sum_{\gamma} \beta^h(\omega, \gamma) P(\gamma)
\]

\[
= \frac{c}{\ln(1/\beta)} \ln \sum_{\gamma} \beta^h(\omega, \gamma) P(\gamma)
\]

\[
= \frac{c}{\ln(1/\beta)} \ln E^{\xi_{\omega}}
\]

\[
= \frac{c}{\ln(1/\beta)} \inf \left( \frac{1}{z_{\omega}} + a A_{\omega}(z_{\omega}) \right)
\]

\[
\geq c z_{\omega} + a A_{\omega}(z_{\omega}). \quad (39)
\]

Inequality (35) follows from the definition of \( v \) (notice that we have used the non-triviality of \( \omega \) here), and equality (36) follows from the definition of \( \xi_{\omega} \).

Let us prove equality (37). Put \( \phi(\xi) := \ln E^{\xi_{\omega}} \) (therefore, \( A_{\omega} = \phi_{\ast, \omega} \), recall that the notation \( E \) implies summing only over the finite values of \( \xi_{\omega} \)). By the Fenchel–Moreau theorem we can transform the infimum in (37) as follows:

\[
\inf_{\zeta} \left( \frac{1}{\beta} + A_{\omega}(\zeta) \right) = -\sup_{\zeta} \left( -\frac{1}{\beta} - \phi_{\ast, \omega}(\zeta) \right)
\]

\[
= -\phi_{\ast, \omega}(\frac{1}{\beta}) = -\phi_{\ast, \omega}(\ln \beta)
\]

\[
= -\ln E \exp(\xi_{\omega} \ln \beta) = -\ln E^{\xi_{\omega}}
\]

(see the proof of Lemma 18).

Now we consider the case where \( \omega \) is trivial. We take \( z_{\omega} \) to be 0; therefore, \( A_{\omega}(z_{\omega}) = 0 \) (by Lemma 16). We are required to prove

\[
\inf_{\delta \in \arg \max_{\omega \in \Gamma^*} \lambda(\omega, \delta)} \lambda(\omega, \delta) > 0. \quad (40)
\]

Let \( \Gamma^* \) be the set of those \( \delta \in \Gamma \) for which \( v(\delta) \) is trivial. Since \( \Gamma^* \) (being a closed subset of \( \Gamma \)) is compact and the function \( \delta \mapsto \max_{\omega \in \Gamma^*} \lambda(\omega, \delta) \) is continuous and positive on \( \Gamma^* \), we have

\[
\inf_{\delta \in \Gamma^*} \max_{\omega \in \Gamma^*} \lambda(\omega, \delta) > 0,
\]

i.e., \( \inf_{\delta \in \Gamma^*} \lambda(\omega, \delta) > 0 \), which implies (40).

Fix such \( z_{\omega} \) and \( \epsilon \).

Now we can describe the probabilistic strategy for \( \varepsilon \) in \( \mathcal{G}(\varepsilon, a) \):

- the number \( n \) of experts is (large and) chosen as specified below;
- the outcome chosen by \( \varepsilon \) always coincides with \( \delta \); \( \delta \) being the prediction made by \( \lambda \);
- each expert predicts the value \( \gamma \in \text{dom} \, P \) with constant probability \( P(\gamma) \).

Let us look at how this simple strategy helps us prove that \( \lambda \) does not have a winning strategy in \( \mathcal{G}(\varepsilon, a) \). Assume, on the contrary, that \( \lambda \) has a winning strategy in this game, and let this winning strategy play against \( \varepsilon \)’s probabilistic strategy just described.
Let $\gamma_{1\cdots T}$ be the random sequence of L's predictions and $\omega_0,\omega_2,\ldots$ be the random sequence of outcomes during this play. For each $T \geq 1$ and $\omega \in \mathcal{Y}$, let $m_d(T) \in \{0,1\}$ be the fraction $\frac{\sum \{ T | \omega_0 = \omega \}}{T}$ of $\omega$'s among the first $T$ outcomes $\omega_0 \cdots \omega_T$. Define stopping time $\tau$ by

$$\tau := \min \left\{ T | T \geq \frac{\ln n}{\sum_{\omega \in \mathcal{Y}} m_\omega(T) A_\omega(z_\omega) + \varepsilon} \right\}. \quad (41)$$

Note that

$$\ln n \max_{\omega \in \mathcal{Y}} A_\omega(z_\omega) \leq \tau \leq \ln n \varepsilon \quad \text{.} \quad (42)$$

Let $T$ be a number for which the probability of $\tau = T$ is the largest; this largest probability is at least $1/(C_1 \ln n)$ ($C_1, C_2, \ldots$ stand for positive constants).

We say that a sequence $\omega_1 \cdots \omega_T \in \mathcal{Y}^T$ is suitable if

$$(\omega_1 = \kappa_1, \ldots, \omega_T = \kappa_T) \Rightarrow \tau = T. \quad \text{Fix a suitable } \omega_1 \cdots \omega_T \text{ with the largest probability of the event } \{ \omega_1 = \kappa_1, \ldots, \omega_T = \kappa_T \}; \text{ since the probability that } \omega_1 \cdots \omega_T \text{ will be suitable is at least } 1/(C_1 \ln n), \text{ this largest probability is at least } (1/(C_1 \ln n)) |\mathcal{Y}|^{-T}. \text{ We say that the random sequence } \gamma_{1\cdots T} \text{ of L's predictions agrees with } \kappa_1 \cdots \kappa_T \text{ if } \gamma(T) = \kappa_i, i = 1, \ldots, T. \text{ It is obvious that L has a strategy in } \mathcal{G}(C, \alpha) \text{ (a simple modification of his winning strategy) such that his predictions } \gamma_{1\cdots T} \text{ always agree with the sequence } \kappa_1 \cdots \kappa_T \text{ and, with probability at least } (1/(C_1 \ln n)) |\mathcal{Y}|^{-T}, \text{ L}_T \leq cL_T(i) + a \ln n, \quad \forall i. \quad (43)$$

We will arrive at a contradiction proving that our probabilistic strategy for E fails condition (43) with probability greater than $1 - \frac{1}{1/(C_1 \ln n)) |\mathcal{Y}|^{-T}$ when playing against any strategy for L that ensures agreement with $\kappa_1 \cdots \kappa_T$.

For each $\omega \in \mathcal{Y}$, let $m_\omega \in \{0,1\}$ be the fraction $\frac{\sum \{ t \in \{1,\ldots, T\} | \omega_t = \omega \}}{T}$ of $\omega$'s in the sequence $\omega_1 \cdots \omega_T$.

By the definition of E's strategy, the first $T$ outcomes will be $\omega_1 = \kappa_1, i = 1, \ldots, T$. So the sequence $\omega_1 \cdots \omega_T$ contains $m_\omega$ of $\omega$'s and, therefore, L's cumulative loss during the first $T$ trials is at least

$$R := \sum_{\omega \in \mathcal{Y}} m_\omega T \inf_{\delta \in \mathcal{M}(\omega, \delta)} \lambda(\omega, \delta). \quad (44)$$

Let $\mathcal{A}$ be the event that, for at least one expert $i$, the cumulative loss

$$\sum_{t \in \{1,\ldots, T\} | \omega_t = \omega} \lambda(\omega, \gamma(i))$$

on the $\omega$'s of $\omega_1 \cdots \omega_T$, for all $\omega \in \mathcal{Y}$, is at most $Tm_\omega z_\omega$ (as it were, the specific per $\omega$ loss is at most $z_\omega$). On event $\mathcal{A}$, the cumulative loss of the best expert (during the first $T$ trials) is at most

$$\rho := T \sum_{\omega \in \mathcal{Y}} m_\omega z_\omega. \quad (45)$$

Let us show that we can choose the number $n$ of experts so that the probability of $\mathcal{A}$ failing is less than $(1/(C_1 \ln n)) |\mathcal{Y}|^{-T}$. Lemma 18 shows that, for each expert and each $\omega \in \mathcal{Y}$, the loss incurred by him on the $\omega$'s of the sequence $\omega_1 \cdots \omega_T$ is at most $Tm_\omega z_\omega$ with probability at least

$$\frac{1}{C_2 \sqrt{Tm_\omega + 1}} \exp(-Tm_\omega A_\omega(z_\omega)) \geq \frac{1}{T} \exp(-Tm_\omega A_\omega(z_\omega)) \quad \text{.}$$

(since $n$ is large, $T$ is large as well: see (42)). Since for each expert these $|\mathcal{Y}|$ events are independent, the probability of their conjunction is at least

$$\frac{1}{T^n} \exp\left(-T \sum_{\omega \in \mathcal{Y}} m_\omega A_\omega(z_\omega)\right) \quad \text{.}$$

The probability that $\mathcal{A}$ will fail is at most

$$\left(1 - \frac{1}{T^n} \exp\left(-T \sum_{\omega \in \mathcal{Y}} m_\omega A_\omega(z_\omega)\right)\right)^n \quad \text{.}$$

since $\omega_1 \cdots \omega_T$ is suitable, the natural logarithm of this expression is

$$n \ln \left(1 - \frac{1}{T^n} \exp\left(-T \sum_{\omega \in \mathcal{Y}} m_\omega A_\omega(z_\omega)\right)\right) \leq -n \frac{T}{T^n} \exp(-T \ln n - C_3)$$

$$< -\frac{n}{T^n} \exp(\varepsilon T - \ln n - C_3)$$

$$= \frac{\exp(\varepsilon T)}{C_4 T^n} \leq -\ln(C_1 \ln n) - T \ln |\mathcal{Y}|,$$

all inequalities holding for $n$ (equivalently, $T$) sufficiently large; the second $< \ldots$ follows from (41). Therefore, we can take $n$ so large that the probability of $\mathcal{A}$ failing is less than $(1/(C_1 \ln n)) |\mathcal{Y}|^{-T}$. 

---


Codes: 4928 Signs: 2961. Length: 56 pic 0 pts, 236 mm
It remains to prove that \( R > cp + c \ln n \), which reduces (see (44), (45), and (41)) to proving
\[
\sum_{\omega \in \mathcal{Y}} m_{\omega} \inf_{\delta \in \mathcal{D}(\omega)} \lambda(\omega, \delta) > c \sum_{\omega \in \mathcal{Y}} m_{\omega} z_{\omega} + a \left( \sum_{\omega \in \mathcal{Y}} m_{\omega} A_{\omega}(z_{\omega}) + \epsilon \right).
\]
It suffices to multiply (33) by \( m_{\omega} \) and sum over \( \omega \in \mathcal{Y} \).

7. GAMES WITH INFINITE \( c \) OR \( a \)

So far we have considered only the global games \( \mathcal{G}(c, a) \) with \( (c, a) \in [0, \infty)^4 \). Intuitively, \( c = \infty \) means that \( E \) is required to ensure that some expert predicts perfectly and \( a = \infty \) means that there is only one expert. In the case where \( c = \infty \) or \( a = \infty \) (or both \( c = \infty \) and \( a = \infty \)), the picture is as follows. Player \( L \) wins \( \mathcal{G}(c, \infty) \) if and only if \( c \geq 1 \). All games \( \mathcal{G}(\infty, a), a \geq 0(0) \), are determined in favor of \( L \); the winner is not determined by the separation curve alone (who the winner is depends not only on the curve \((c, a)\)) but also on other aspects of the local game \((\Omega, \Gamma, \lambda)\). This section is devoted to the proof of these facts.

First we prove the existence of \( L \)’s relevant strategies. If \( c = 1 \) and \( a = \infty \), \( L \)’s goal \( L_i \leq L_i(i) + \infty \) in \( n \) is easy to attain: if \( n > 1 \), this condition is vacuous, and if \( n = 1 \), it suffices for \( L \) to repeat the predictions made by the only expert.

Let \( c = \infty \); \( L \)’s goal is to ensure \( L_i \leq \infty L_i(i) + a(0) \) in \( n \). In other words, \( L \) must ensure \( L_i \leq a(0) \) in \( n \) as long as \( \min L_i = 0 \). For each \( \beta \in [0, 1] \), define \( \gamma_{\beta i}^{i+1}: \Omega \to [0, \infty) \) by
\[
\beta \gamma_{\beta i}^{i+1}(\omega) = \frac{1}{|E_i|} \sum_{\omega \in E_i} \beta^{\alpha(\omega, \gamma_{\beta i}^{i+1}(\omega))},
\]
where \( E_i := \{ i | L_i(i) = 0 \} \) is the set of perfect experts by time \( t \) (cf. (16)), and choose \( \gamma_{\beta i}^{i+1} \in \Gamma \) such that
\[
\lambda(\omega, \gamma_{\beta i}^{i+1}) \leq c(\beta) \gamma_{\beta i}^{i+1}(\omega), \quad \forall \omega.
\]
Let \( \gamma_{\beta i}^{i+1} \) be a limit point of a sequence \( \gamma_{\beta k}^{i+1} \) with \( \beta_k \downarrow 0 \). We find
\[
\lambda(\omega, \gamma_{\beta i}^{i+1}) \leq \lim_{\beta \to 0} \beta(\beta) \log \left( \frac{1}{|E_i|} \sum_{\omega \in E_i} \beta^{\alpha(\omega, \gamma_{\beta i}^{i+1}(\omega))} \right)
= \lim_{\beta \to 0} \left( -a(\beta) \ln \left( \frac{1}{|E_i|} \sum_{\omega \in E_i} \beta^{\alpha(\omega, \gamma_{\beta i}^{i+1}(\omega))} \right) \right)
= -a(0) \ln \frac{|E_i|}{|E_i|}.
\]
Provided \( F_i \neq \emptyset \), where \( F_i := \{ i \in E_i | \lambda(\omega, \gamma_{\beta i}^{i+1}(i)) = 0 \} \). Taken as \( \omega \) the outcome chosen by the nature, we obtain
\[
\lambda(\omega, \gamma_{\beta i}^{i+1}) \leq a(0) \ln \frac{|E_i|}{|E_i|},
\]
therefore, \( L \)’s total loss \( L_i \) at each trial \( t \) such that \( \min L_i(i) = 0 \) (equivalently, such that \( E_i \neq \emptyset \)) is at most
\[
a(0) \ln \frac{|E_i|}{|E_i|} + \ldots + a(0) \ln \frac{|E_{i+1}|}{|E_i|} = a(0) \ln \frac{|E_n|}{|E_i|}
\leq a(0) \ln(n).
\]

It remains to construct the relevant strategies for player \( E \). First suppose \( c < 1, a = \infty \). Player \( E \)’s goal is to ensure, for some \( t \) and \( i, L_t > cL(i) + \infty \) in \( n \). He is forced to choose \( n = 1 \). After that it suffices for him to apply the strategy of Lemma 5.

Now we consider the remaining case \( c = \infty, 0 \leq a < a(0) \). Player \( E \)’s goal is to ensure \( L_t > \infty L_t(i) + a \) in \( n \) for some \( t \) and \( i \). We will give two examples with identical separation curves and \( a(0) > 0 \): in the first, \( E \) will be able to do so for any \( a < a(0) \); in the second, he will be able to do so for no \( a < a(0) \). Consider the simple prediction case (see Example 1 in Section 2). Here \( a(0) = 1/\ln 2 \) and \( c(0) = \infty \). A winning strategy for \( E \) is as follows: he takes \( n = 2^k \) with any positive integer \( k \); the outcome is always opposite to \( L \)’s prediction; at trial \( t, t = 1, \ldots, k \), expert \( i \in \{ 1, \ldots, 2^k \} \) predicts with the \( r \)th bit in the binary expansion of \( i - 1 \) (we assume that the length of this binary expansion is exactly \( k \); if necessary, we add zeroes on the left). After the \( k \)th trial we will have \( L_t = k, L_t(i) = 0 \) for some \( i \), and \( a \) in \( n < a(0) \) in \( n = \log_2 n = k \).

Now consider the local game \((\Omega, \Gamma, \lambda)\) with \( \Omega := \{ 0, 1 \} \times \{ 1, 2, \ldots, k \}, \Gamma := \{ 0, 1 \} \), and
\[
\lambda(j, i, y) = \begin{cases} \epsilon_i, & \text{if } j = y, \\ 1, & \text{otherwise,} \end{cases}
\]
where \( \epsilon_i \) is a decreasing sequence of numbers in \( ]0, 1[ \) such that \( \epsilon_i \to 0 (i \to \infty) \). All \( c(\beta) \) and \( a(\beta) \) are here exactly the same as in the case of the simple prediction game. But now, for all \( t \) and \( i, L_t(i) > 0 \); therefore, \( E \) cannot win the game.

It remains an open problem to give an explicit formula for the value
\[
\inf \{ a | \mathcal{G}(\infty, a) - L \} \in [0, a(0)].
\]

8. CONNECTIONS WITH LITERATURE

The Aggregating Algorithm was proposed in [31] as a common generalization of the Bayesian merging scheme (Dawid [7, Section 4]; DeSantis et al. [9]) and the Weighted
Majority Algorithm (Vovk [32, Theorem 5] and Littlestone and Warmuth [23]; I am using the name coined by Littlestone and Warmuth). Earlier, algorithms with similar properties were proposed by Foster [11] (for the case of the Brier loss function; see Example 4 above) and Foster and Vohra [12] (for a wide class of loss functions). The environment’s strategy described in Section 6 resulted from a series of small steps: cf. Theorem 6 of [32], Theorem 2 of [31], and [33]; the key idea is from Cohen [5]. (In Section 6 we did not use the powerful techniques of Cesa-Bianchi et al. [4] and Haussler et al. [15]; these techniques may be useful in strengthening our result.) The idea of considering the value $\beta = 0$ is due to Littlestone and Long [21].

The main contribution of this paper is the proof that the Aggregating Algorithm and bound (1) with $c = c(\beta), a = a(\beta)$ are in some sense optimal. If we understand “optimal” in a stronger sense, however, the bound and even the algorithm in some situations can be improved. Cesa-Bianchi et al. [4] give a simple example concerning the bound. Let us consider the game whose only difference from $\vartheta(c, a)$ is that the number of experts is not chosen by the adversary but is given $a priori$ as $n := 3$; the outcome and prediction spaces are $\{0, 1\}$ and the loss function is $|a - \gamma|$ (i.e., the simple prediction game is being played; see Example 1). The Aggregating Algorithm (which, in this case, coincides with the Weighted Majority Algorithm) ensures, for $\beta < 1/2$,

$$L_t \leq 2L_t(i) + 1, \quad \forall t, i.$$ 

This coincides with (1) when $c = 2$ and $a = 1/\ln 3$, and the point $(2, 1/\ln 3)$ lies Southwest of and outside the separation curve for the simple prediction game (since for this game $c(\beta) > 2, \forall \beta \in [0, 1]$). Therefore, bound (1) can be improved when $c$ and $a$ are allowed to depend on $n$.

The next step has been made by Littlestone and Long [21], who give an example where not only the bound but also the algorithm itself can be improved. Their example violates one more assumption of our Theorem 1: the loss function $\lambda$ as well is now allowed to depend on the number of experts.

**Example 8** (Littlestone and Long; modified). There are $n > 2$ experts, two possible outcomes, 0 and 1, and two possible predictions, $\gamma_0$ and $\gamma_1$. The loss function is

$$\lambda(0, \gamma_0) = \lambda(1, \gamma_1) = 0,$$

$$\lambda(0, \gamma_1) = 1, \quad \lambda(1, \gamma_0) = (n - 1) \ln n.$$ 

**Lemma 20.** Under the conditions of Example 8, the following holds. If $L_t$ at each trial $t$ predicts $\gamma_1$ unless all “best experts” $i \in \arg \min L_{i, t} - i(j)$ unanimously predict $\gamma_0$, then (1) will hold with $c = (n - 1)(\ln n + 1)$ and $a = (n - 1)\ln n$. On the other hand, the Aggregating Algorithm loses $\vartheta(c, a)$ when $a < n - 1$ (in particular, when $c = (n - 1)(\ln n + 1)$ and $a = (n - 1)\ln n$).

**Proof.** First we prove the assertion about the simple strategy that always predicts $\gamma_1$ unless all best experts $i \in \arg \min L_{i, t} - i(j)$ unanimously predict $\gamma_0$. It is sufficient to prove that, for all $t$,

$$L_t \leq (n - 1)\ln n + 1 - |\arg \min L_{i, t}(j)|$$

(46) (cf. (1)), $L^*_t = \min_j L_{i, t}(j)$ being the loss incurred by the argument $L_{i, t}(j)$ best experts. To see that (46) is indeed true, notice that it is true for $t = 0$ and that, for $t > 0$, it follows from

$$L_{t - 1} \leq (n - 1)\ln n + 1 + n - |\arg \min L_{i, t - 1}(j)|.$$ 

It remains to prove the assertion about the Aggregating Algorithm. Let, at trial 1, all experts but one predict $\gamma_0$ and the outcome be 1. We are only required to prove that the Aggregating Algorithm will predict $\gamma_0$, whatever $\beta \in [0, 1]$. The $(n - 1)$-1 mixture of $\gamma_0, \gamma_1$ (see (16); in Section 2, we represent prediction $\gamma \in \Gamma$ by the point $(\lambda(0, \gamma), \lambda(1, \gamma))$ of the $(x, y)$-plane) is

$$\left(\log \left(\frac{n - 1}{n}; \frac{1}{n}\right) - \log \left(\frac{n - 1}{n}; \frac{1}{n}\right)\right):$$

therefore, $L$ predicting $\gamma_0$ is equivalent to

$$\log \left(\frac{n - 1}{n}; \frac{1}{n}\right) > (n - 1)\ln n$$

which in turn is equivalent to

$$\ln \left(\frac{n - 1}{n}; \frac{1}{n}\right) < (n - 1)\ln n \ln \left(\frac{n - 1}{n}; \frac{1}{n}\right).$$

(47)

So we are only required to prove (47).

For $\beta = 0$ inequality (47) becomes

$$\ln n > (n - 1)\ln n \ln \left(\frac{n - 1}{n}; \frac{1}{n}\right),$$

i.e.,

$$\frac{1}{n - 1} > \ln \left(\frac{n - 1}{n}; \frac{1}{n}\right).$$
The last inequality is a special case of the inequality $t > \ln(1 + t)$, where $t > 0$.

For $\beta = 1$, (47) turns into the equality $0 = 0$. Let us rewrite (47) in the equivalent form

$$n - \frac{1}{n} \beta^{(n-1) \ln n} + \frac{1}{n} < \left( \frac{n - 1 + \beta}{n} \right)^{(n-1) \ln n}.$$  

Since this inequality holds for $\beta = 0$ and the corresponding non-strict inequality holds for $\beta = 1$, it suffices to prove the existence of a point $\alpha \in [0, 1]$ such that

$$\beta < \alpha \Rightarrow \frac{d}{d\beta} \left[ \frac{n - 1}{n} \beta^{(n-1) \ln n} + \frac{1}{n} \right] \left( \frac{n - 1 + \beta}{n} \right)^{(n-1) \ln n},$$  

$$\beta > \alpha \Rightarrow \frac{d}{d\beta} \left[ \frac{n - 1}{n} \beta^{(n-1) \ln n} + \frac{1}{n} \right] \left( \frac{n - 1 + \beta}{n} \right)^{(n-1) \ln n}.$$  

Since the function

$$f(\beta) := \frac{d}{d\beta} \left[ \frac{n - 1}{n} \beta^{(n-1) \ln n} + \frac{1}{n} \right] \left( \frac{n - 1 + \beta}{n} \right)^{(n-1) \ln n}$$  

$$= \frac{n - 1}{n} (n - 1 \ln n \beta^{(n-1) \ln n - 1} + \left( \frac{n - 1 + \beta}{n} \right)^{(n-1) \ln n - 1} - \frac{1}{n})$$  

$$= (n - 1) \left( \frac{\beta \ln n}{n - 1 + \beta} \right)^{(n-1) \ln n - 1}$$  

is monotonic and satisfies $f(0) = 0$ and $f(1) > 1$, (48) indeed holds.

These examples evoke several questions: What is the strictest sense in which bound (1) with $c = c(\beta)$, $a = a(\beta)$ is optimal? What is the strictest sense in which the Aggregating Algorithm is optimal? Does the learner have a better strategy?

When it is known that the cumulative loss of the best expert is 0, Littlestone and Long [21] propose to let $\beta \rightarrow 0$ in the Aggregating Algorithm. Cesa-Bianchi et al. [3] consider an opposite situation where we want to take into account the possibility that the cumulative loss of the best expert may be much larger than $\ln n$. In this case we would like to let $\beta \rightarrow 1$, but this would lead to $a(\beta) \rightarrow \infty$ and make bound (1) (with $c = c(\beta)$, $a = a(\beta)$) vacuous. In essence, the idea of Cesa-Bianchi et al. [3] is that the linear combination $cL_i + a(\beta) \ln n$ of (1) should be replaced by a function like

$$c(1) L_i + b(\beta) \ln n + a(1) \ln n$$  

(in Cesa-Bianchi et al. [3] $c(1) = 1$; in this case the idea of using (49) is especially appealing). It would be interesting to study the separation curve in the $(b, a)$-plane for the global game determined by (49) (or by some more suitable expression, such as the slightly different expression in [3, Theorem 12]).

The main result of this paper is closely connected with Theorem 3.1 of Haussler et al. [15]. In that theorem the authors find, for the global games $G(c, a)$ corresponding to a wide class of local games, the intersection of the separation curve with the straight line $c = 1$ (when non-empty, this is perhaps the most interesting part of the separation curve). In Section 4 of [15] Haussler et al. consider the continuous-valued outcomes.

Some papers (Littlestone et al. [22], Kivinen and Warmuth [18]; Section 5 of Littlestone [20] can also be regarded this way) set a different task for the learner: his performance must be almost as good as the performance of the best linear combination of experts (the prediction space must be a linear space here). In some sense, our task (approximating the best expert) and the task of approximating the best linear combination of experts reduce to each other:

- a single expert can always be represented as a degenerate linear combination;
- we can always replace the old pool of experts by a new pool that consists of the relevant linear combinations of experts.

Even so, these reductions are not perfect: e.g., the second reduction will lead to a continuous pool of experts, which will make maintaining the weights for the experts much more difficult; on the other hand, we can hope that, by applying the Aggregating Algorithm (which was shown to be in some sense optimal) to the enlarged pool of experts, we will be able to obtain sharper bounds on the performance of the learner.

A fascinating direction of research is “tracking the best expert” (Littlestone and Warmuth [23], Herbster and Warmuth [17]). The nice results (such as Theorem 5.7 of [17]) obtained in that direction, however, correspond to only one side of Theorem 1; namely, they assert the existence of a good strategy for the learner.

Much of the work in the area of on-line prediction, including this paper, has been profoundly influenced (sometimes indirectly) by Ray Solomonoff’s thinking: for accounts of Solomonoff’s research, see Li and Vitányi [19] and Solomonoff [28].
APPENDIX A: AGGREGATING ALGORITHM

In this appendix we describe an algorithm (the Aggregating Algorithm, or AA) that the learner can use to make predictions based on the predictions made by a pool $\Theta$ of experts. Fix $\beta \in [0, 1]$; the parameter $\beta$ determines how fast AA learns (sometimes this is represented in the form $\beta = e^{-\gamma}$, where $\gamma > 0$ is called the learning rate).

In the bulk of the paper we assume that the pool is finite, $\Theta = \{1, \ldots, n\}$. Under this assumption, AA is optimal in the sense that it gives a winning strategy for $L$ in the game $\mathcal{G}(c, (a, b))$ described in Section 1. The first description of AA was given in [31, Section 1]; Haussler et al. [15] noted that the proof that AA gives a winning strategy in $\mathcal{G}(c, (a, b))$ makes no use of the assumption made in [31] that $\Omega = \{0, 1\}$.

In this Appendix we will not assume that $\Theta$ is finite; dropping this assumption will not make the algorithm more complicated if we ignore, as we are going to do here, the exact statement of the regularity conditions needed for the existence of various integrals. The observation that AA works for infinite sets of experts was made by Freund [13].

Let $\mu$ be a fixed measure on the pool $\Theta$ and $\pi_0$ be some probability density with respect to $\mu$ (this means that $\pi_0 \geq 0$ and $\int \pi_0 \, dq = 1$; in what follows, we will drop "with respect to $\mu$"). The prior density $\pi_0$ specifies the initial weights assigned to the experts. In the finite case $\Theta = \{1, \ldots, n\}$, it is natural to take $\mu([i]) = 1, \ldots, n$ (the counting measure); to construct L's winning strategy in $\mathcal{G}(c, (a, b))$ (in the proof of Theorem 1) we only need to consider equal weights for the experts: $\pi_0(i) = 1/n, i = 1, \ldots, n$.

In addition to choosing $\beta, \mu$, and $\pi_0$, we also need to specify a "substitution function" in order to be able to apply AA. A pseudoprediction is defined to be any function of the type $\Omega \to [0, \infty]$ and a substitution function is a function $\Sigma$ that maps every pseudoprediction $g: \Omega \to [0, \infty]$ into a prediction $\Sigma(g) \in \Gamma$. A "real prediction" is defined to be any function of the type $\Omega \to [0, 1]$ and a substitution function is a function $\Sigma$ that maps every pseudoprediction $g: \Omega \to [0, 1]$ into a prediction $\Sigma(g) \in \Gamma$. A "real prediction" is identified with the pseudoprediction $g$ defined by $g(\omega) := \Sigma(g)$. We say that a substitution function $\Sigma$ is a minimax substitution function if, for every pseudoprediction $g$,

$$\Sigma(g) \in \arg \min_{\gamma \in \Gamma} \sup_{\omega \in \Omega} \frac{\lambda(\omega, \gamma)}{g(\omega)},$$

where $0/0$ is set to $0$.

**Lemma 1.** Under Assumptions 1–4 (see Section 1), a minimax substitution function exists.

**Proof.** This proof is similar to the proof of Lemma 12. Let $g$ be a pseudoprediction; put

$$c(g) := \inf_{\gamma \in \Gamma} \sup_{\omega \in \Omega} \frac{\lambda(\omega, \gamma)}{g(\omega)},$$

(with the same convention $0/0 := 0$). The case $c(g) = \infty$ is trivial, so we assume that $c(g)$ is finite. Let $c_1, c_2, \ldots$ be a decreasing sequence such that $c_k \to c(g)$ as $k \to \infty$. By the definition of $c(g)$, for each $k$ there exists $\delta_k \in \Gamma$ such that

$$\forall \omega: \frac{\lambda(\omega, \delta_k)}{g(\omega)} \leq c_k.$$

Let $\delta$ be a limit point (whose existence follows from Assumption 1) of the sequence $\delta_1, \delta_2, \ldots$. Then, for each $\omega$, $\lambda(\omega, \delta)$ is a limit point of the sequence $\lambda(\omega, \delta_k)$ (by Assumption 2) and, therefore,

$$\frac{\lambda(\omega, \delta)}{g(\omega)} \leq c(g).$$

This means that we can put $\Sigma(g) := \delta$.

Fix a minimax substitution function $\Sigma$. Now we have all we need to describe how AA works. At every trial $\ell = 1, 2, \ldots$ the learner updates the experts' weights as

$$\pi_t(i) := \beta^{\pi_{t-1}(i)} \pi_{t-1}(i), \quad i \in \Theta,$$

where $\pi_t$ is the prior density on the experts. (So the weight of an expert $i$ whose prediction $\gamma(i)$ leads to a large loss $\lambda(\omega, \gamma(i))$ gets slashed.) The prediction made by AA at trial $\ell$ is obtained from the weighted average of the experts' predictions by applying the minimax substitution function

$$\gamma_i := \Sigma(\pi_t),$$

where the pseudoprediction $g_i$ is defined by

$$g_i(\omega) := \log_{\beta} \int_{\Theta} \beta^{\pi_{t-1}(i)} \pi_{t-1}(i) \mu(\omega) \, d\omega,$$

and $\pi^*_t$ are the normalized weights,

$$\frac{\pi^*_t(i)}{\sum_{i \in \Theta} \pi^*_t(i) \mu(\omega)} \mu(\omega),$$

(assuming that the denominator is positive; if it is 0, $(\pi_0\mu)$-almost all experts have suffered infinite loss and, therefore, AA is allowed to make any prediction). This completes the description of the algorithm.

**Remark.** When implementing AA for various specific loss functions, it is important to have an easily computable minimax substitution function $\Sigma$. It is not difficult to see that easily computable substitution functions exist in all examples considered in Section 2 and in other natural examples considered in literature. (In the case of a finite
pool $\theta$ the pseudoprediction that $\Sigma$ is fed with is represented by the corresponding simple probability distribution $P$ in $\theta$; cf. (2).)

Having described AA for a possibly infinite pool of experts, we can add a few more examples to the examples given in Section 2.

**Example 9 (Cover and Ordentlich [6]).** The learner is investing in a market of $K$ stocks. The behavior of the market at trial $t$ is described by a non-negative price relative vector $\omega_t \in [0, \infty[^K$. The $k$th entry $\omega_{t,k}$ of the $t$th price relative vector $\omega_t$ denotes the ratio of closing to opening price of the $k$th stock for the $t$th trading day. An investment at time $t$ in this market is specified by a portfolio vector $\gamma_t \in [0, \infty[^K$ with non-negative entries $\gamma_{t,k}$ summing to 1: $\gamma_{t,1} + \cdots + \gamma_{t,K} = 1$. The entries of $\gamma_t$ are the proportions of current wealth invested in each stock at time $t$. This is a special case of our learning protocol with the outcome and prediction spaces

$$\Omega = [0, \infty[^K, \Gamma = \{ \gamma = \gamma_1 \cdots \gamma_K \in [0, \infty[^K \mid \gamma_1 + \cdots + \gamma_K = 1 \},$$

respectively. An investment using portfolio $\gamma$ increases the investor’s wealth by a factor of $\gamma \cdot \omega = \sum_{k=1}^{K} \gamma_k \omega_k$ if the market performs according to the price relative vector $\omega = \omega_1, \cdots, \omega_K$. The loss function is defined to be the minus logarithm of this increase:

$$\ell(\omega, \gamma) := -\ln(\gamma \cdot \omega). \quad (50)$$

Let us consider the pool of experts $\Theta = \Gamma$; expert $\gamma$’s prediction is always $\gamma$. Notice that expert $\gamma$’s loss is the minus logarithm of the wealth attained by using the same portfolio $\gamma$ starting with a unit capital. (Expert $\gamma$’s strategy is called a constant rebalanced portfolio strategy; it actually involves a great deal of trading.) The algorithm used by Cover and Ordentlich [6] is tantamount to AA applied to this pool of experts with $\beta = 1/e$, $\Sigma$ the identity function (when $\beta = 1/e$, every pseudoprediction is a real prediction in this example), $\mu$ the Lebesgue measure in $\Gamma$, and $p_0$ either the uniform or Dirichlet ($\frac{1}{2}, \ldots, \frac{1}{2}$) density. They obtain the following analogues of (1),

$$L_t \leq L_0(i) + (K - 1) \ln(t + 1) \quad \text{(Theorem 1)}$$

for $p_0$ the uniform density, and

$$L_t \leq L_0(i) + \frac{K - 1}{2} \ln(t + 1) + \ln 2 \quad \text{(Theorem 2)}$$

for $p_0$ the Dirichlet ($\frac{1}{2}, \ldots, \frac{1}{2}$) density.

**Remark.** Loss function (50) can take negative values, whereas the non-negativity of the loss function was one of our assumptions. We needed this assumption to define the notion of minimax substitute function; the identity function is the most natural substitute function when $\beta = 1/e$ in Example 9 but we cannot say that it is minimax in our sense. Following [16, Section 5], we can “normalize” the price relatives $\omega_{t,k}, \ k = 1, \ldots, K$, replacing them by $\omega_{t,k}^* := \omega_{t,k}/\max_{j=1, \ldots, K} \omega_{t,j}$. All theorems in [6] are invariant under such normalization; on the other hand, the loss function becomes non-negative and the identity function becomes a minimax loss function.

We already mentioned that Cover and Ordentlich [6] apply AA with $\beta = 1/e$, i.e., with learning rate $\eta = 1$. Experiments with NYSE data (historical stock market data from the New York Stock Exchange accumulated over a 22-year period) conducted by Helmbold et al. [16, Section 4] suggest, however, that a better choice would be, say, $\eta = 0.05$; they report that learning rates from 0.01 to 0.15 all achieved great wealth, greater than the wealth achieved by the universal portfolio algorithm (i.e., Cover and Ordentlich’s algorithm) and in many cases comparable to the wealth achieved by the best constant rebalanced portfolio. The algorithm used in [16] is different from (though closely related to) AA, and further experiments are needed to test the performance of AA with learning rates different from 1.

**Example 10 (Freund [13]).** In the games of Examples 3–5 (namely, the absolute loss, Brier, and logarithmic games) we can consider the following pool of experts. The experts are indexed by $\Theta = \Gamma = [0, 1]$; expert $\gamma \in [0, 1]$ always predicts with $\gamma$. The performance of AA for this pool of experts in those games is analyzed in Freund [13].

In Examples 9 and 10, AA was used for merging a pool of possible strategies for the learner. Another application of this idea is using AA as a universal method of coping with the problem of overfitting. The following is a typical example (other possible examples are predicting the next symbol in a sequence by fitting Markov chains of different orders, estimating a probability density using different smoothing factors, etc.).

**Example 11 (Approximating a Function by a Polynomial).** Consider the following scenario. At each trial $t = 1, 2, \ldots$:

1. The environment chooses $x_t \in [0, 1]$.
2. The learner makes a guess $\hat{y}_t \in [0, 1]$.
3. The environment chooses $y_t \in [0, 1]$.
4. The learner suffers loss $(\hat{y}_t - y_t)^2$.

So the learner’s task is to predict $y_t$ given $x_t$. Suppose he decided to do so by fitting a polynomial

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_d x^d$$
to his data \((x_1, y_1), \ldots, (x_{t-1}, y_{t-1})\) and predicting with
\[
\hat{y}_t := \text{trunc}_{[0, 1]}(a_0 + a_1 x_t + \cdots + a_n x_t^n),
\]
where
\[
\text{trunc}_{[0, 1]} u := \begin{cases} 
0, & \text{if } u < 0, \\
1, & \text{if } u > 1, \\
\text{otherwise}; & \text{otherwise;}
\end{cases}
\]
he is unsure, however, what degree \(t\) to choose. Typically his predictive performance will suffer if he chooses \(t\) too big or too small (“overfitting” or “underfitting,” respectively, will occur). This problem is usually treated as a special case of the problem of model selection and has been extensively studied; popular approaches are, e.g., Rissanen’s [25]–[27] Minimum Description Length principle, Vapnik’s [29, 30] Structural Risk Minimization principle, and Wallace’s [34, 35] Minimum Message Length principle. We will consider an alternative approach to the problem of avoiding over- and underfitting: instead of picking the best, in some sense, model (i.e., degree \(i\) of the polynomial) we merge all possible models using AA. Therefore, we introduce the following countable pool of experts: at trial \(t\), expert \(i\), \(i = 1, 2, \ldots\), predicts with \(\text{trunc}_{[0, 1]} f(x)\), where \(f(x)\) is the polynomial of degree at most \(i\) that is the best least-squares approximation for the set of points \((x_1, y_1), \ldots, (x_{t-1}, y_{t-1})\). (If there exist more than one polynomials of degree at most \(i\) that are best least-squares approximations, we take the polynomial of the lowest possible degree; in particular, all experts \(i\) with \(i \geq t - 2\) make the same guess.) It is easy to see that the predictive performance of AA in this example is given by the following generalization of (1): for all \(t, i\),
\[
L_t \leq L_0(i) + \alpha(i) \ln \frac{1}{\pi_d(i)},
\]
where \(\pi_d(i)\) are the initial weights for the experts, \(\sum \pi_d(i) = 1\) (11) is obtained by taking \(\pi_d(i) := 1/n, i = 1, \ldots, n\). Putting \(\alpha(i) := e^{-2}\), we obtain (see Example 4 in Section 2 and recall that \(\alpha(i) = c(\beta)/\ln(1/\beta)\))
\[
L_t \leq L_0(i) + \frac{1}{2} \ln \frac{1}{\pi_d(i)}. \quad (51)
\]
A natural choice for the prior weights \(\pi_d(i)\) is Rissanen’s [25] universal prior for integers
\[
\pi_d(i) := \frac{1}{c i \log i \log \log i},
\]
where \(c\) is the natural logarithm and the product involves only terms greater than 1; Rissanen found that the normalizing constant is \(c \approx 2.865064\). Under this choice (51) becomes
\[
L_t \leq L_0(i) + 0.3466 \log* i + 0.5263, \quad (52)
\]
where \(\log* i := \log i + \log \log i + \cdots \) (the sum involving only positive terms), \(0.3466 \approx 1/(2 \log e)\), and \(0.5263 \approx \frac{1}{2} \ln 2.865064\). For example, (52) shows that if the environment selects points \((x_t, y_t)\) with
\[
y_t := \text{trunc}_{[0, 1]}(a_0 + a_1 x_t + a_2 x_t^2 + \eta_t),
\]
where \(a_0, a_1, a_2\) are constants and \(\eta_t\) are independent \(\mathcal{N}(0, 1)\) random variables representing Gaussian noise, the extra loss suffered by AA will be at most \(0.3466 \log 2 + 0.5263 = 0.8729 < 1\) as compared with the loss of the algorithm that knows a priori that the best degree for the approximating polynomial is 2. Two obvious drawbacks of our approach are that
- it is computationally less efficient than the “best-model” approaches: at step \(t\) we merge \(t - 2\) predictions (recall that experts \(t - 2, t - 3, \ldots\) make identical predictions);
- we need to assume that the prediction space \(\Gamma\) is compact (Assumption 1 of Section 1).
Its possible advantage might be better predictive performance.

Remark. The problem of choosing the prior distribution in the pool of experts is very important in applying AA. In Example 11 we took Rissanen’s universal prior for integers; a similar approach could also be applied in Examples 9 and 10 if we are willing to tolerate an extra additive constant in the cumulative loss \(L_t\) incurred by the learner.

APPENDIX B: PROOF OF LEMMAS 16, 17, AND 18

Proof of Lemma 16. We begin with (26). When \(\pi > \max \zeta\), we have for some \(e > 0\) and for all \(\zeta > 0\):
\[
\psi(\zeta) < e^{(\pi - e)} \zeta;
\]
\[
\ln \psi(\zeta) < (\pi - e) \zeta;
\]
\[
a^* - \ln \psi(\zeta) > e^*; \]
therefore, \(A(\pi) = \infty\). Analogously, when \(\pi < \min \zeta\), we have for some \(e > 0\) and all \(\zeta < 0\)
\[
\psi(\zeta) < e^{(\pi + e)} \zeta;
\]
\[
\ln \psi(\zeta) < (\pi + e) \zeta;
\]
\[
a^* - \ln \psi(\zeta) > -e^*;
\]
and we again obtain \(A(\pi) = \infty\).
Now we prove (27), and this will also complete the proof of (26) (since $A$ is convex). When $\alpha = \max \zeta$, we have $(d/d\zeta)(\alpha \zeta - \ln \psi(\zeta)) \geq 0$ (see (22)), and so

$$A(\alpha) = \lim_{\zeta \to \infty} \left( \alpha \zeta - \ln \psi(\zeta) \right) = \lim_{\zeta \to \infty} \left( \alpha \zeta - \ln(e^{\alpha} \operatorname{prob}\{\zeta = \alpha\}) \right) = -\ln \operatorname{prob}\{\zeta = \alpha\}.$$

Analogously, in the case $\alpha = \min \zeta$ we have $(d/d\zeta)(\alpha \zeta - \ln \psi(\zeta)) \leq 0$, and again

$$A(\alpha) = \lim_{\zeta \to -\infty} \left( \alpha \zeta - \ln \psi(\zeta) \right) = \lim_{\zeta \to -\infty} \left( \alpha \zeta - \ln(e^{\alpha} \operatorname{prob}\{\zeta = \alpha\}) \right) = -\ln \operatorname{prob}\{\zeta = \alpha\}.$$

Now let $\vartheta > 0$; we will prove (28). For each $\alpha \in ]\min \zeta$, $\max \zeta$ there exists $\zeta = \zeta(\alpha)$ such that $E(\eta - \alpha) = \vartheta$, where $\eta = \eta(\zeta)$ is the simple random variable defined by (21). By (22), at $\zeta = \zeta(\alpha)$ we have $\alpha = \psi'(\zeta)/\psi(\zeta) = (\ln \psi)'(\zeta)$ and, thus, $(d/d\zeta)(\alpha \zeta - \ln(\zeta)) = 0$. Since $\ln(\psi)' > 0$ (see (23) and recall that $\vartheta > 0$ and, thus, $\vartheta > 0$), $\zeta(\alpha)$ is the value of $\zeta$ where the supremum in (25) is attained. Since $\zeta(\alpha)$ is the inverse to the smooth function $(\ln \psi)'$ with positive derivative, $\zeta(\alpha)$ is also a smooth function.

It is easy to see that $\zeta(E_{\alpha}) = 0$ and, since

$$A(\alpha) = \alpha \zeta(\alpha) - \ln \psi(\zeta(\alpha)), \quad (53)$$

$A(E_{\alpha}) = 0$. Furthermore, (53) implies

$$A'(\alpha) = \zeta(\alpha) + \alpha \zeta'(\alpha) = \zeta'(\alpha) = \zeta(\alpha)$$

and, therefore (see (23)),

$$A'(E_{\alpha}) = \zeta(E_{\alpha}) = 0,$$

$$A''(E_{\alpha}) = \zeta'(E_{\alpha}) = \frac{1}{(\ln \psi)'(0)} = \frac{1}{\vartheta \zeta}.$$

The case $\zeta = 0$ follows from (26) and (27).

Since $A$ is a Young–Fenchel transform, it is convex and closed, and, hence, continuous at $\min \zeta$ and $\max \zeta$.

Proof of Lemma 17. Without loss of generality we assume $N > 0$. First we consider the case $\alpha \in ]\min \zeta$, $\max \zeta[$. Let $\zeta = \zeta(\alpha)$ be the value where the supremum in (25) is attained; in the proof of Lemma 16 we saw that it is indeed attained and $\eta = \eta(\zeta)$ satisfies

$$E(\eta - \alpha) = 0. \quad (54)$$

Put $H_N := Z_N - \alpha N$ (see (24)). We have, by Lemma 15,

$$\operatorname{prob}\{S_N \leq \alpha N\} = \sum_{x \leq \alpha N} \operatorname{prob}\{S_N = x\} = \sum_{x \leq \alpha N} \psi_N(\zeta) e^{-\vartheta \zeta} \operatorname{prob}\{Z_N = x\}.$$

Putting $x := z - \alpha N$ (so $z = x + \alpha N$), we obtain

$$\operatorname{prob}\{S_N \leq \alpha N\} = \sum_{z \leq x \leq \vartheta N} \psi_N(\zeta) e^{-\vartheta \zeta} \operatorname{prob}\{Z_N = x\} = \sum_{z \leq x \leq \vartheta N} e^{-\vartheta(z - \alpha N)} \operatorname{prob}\{H_N = x\} = e^{-\vartheta \alpha N} \sum_{z \leq x \leq \vartheta N} e^{-\vartheta \zeta} \operatorname{prob}\{H_N = x\}.$$

Therefore, we are only required to prove that

$$\sum_{z \leq x} e^{-\vartheta \zeta} \operatorname{prob}\{H_N = x\} \geq \frac{1}{C \sqrt{N + 1}},$$

where $H_N$ is the sum of $N$ independent simple random variables with mean $0$ (see (54)). The difficulty here is that it is possible that $\zeta < 0$, and then $e^{-\vartheta \zeta}$ shrinks very fast as $x$ decreases. But it is easy to see that it suffices to prove

$$\operatorname{prob}\{C_1 < H_N \leq C_2\} \geq \frac{1}{C_3 \sqrt{N}}, \quad (55)$$

for some constants $C_1 < C_2 \leq 0$, $C_3 > 0$ and for large $N$.

Let $\lambda(m, \delta)$ be the normal distribution with mean $m$ and variance $\delta$. We know that $H_N$ is distributed approximately as $\lambda(0, N\sigma^2)$, where $\sigma^2 := \frac{\operatorname{var} \eta}{\psi'(0)}$ (note that $\sigma^2 > 0$ under our current assumption $\alpha \in ]\min \zeta$, $\max \zeta[$]. Of course, (55) would be true were $H_N$ distributed exactly as $\lambda(0, N\sigma^2)$; we will use the fact that the distribution of $H_N$ is very close to $\lambda(0, N\sigma^2)$; see Feller [10], Section XVI.4. Let us first assume that $\eta$ is not a lattice random variable. Choose $C_1$ and $C_2$ arbitrarily (with the only restriction $C_1 < C_2 \leq 0$) and put $H^*_N := H_N/(\sigma \sqrt{N})$; Theorem XVI.4.1 of Feller [10] allows us to transform the left-hand side of (55) as
follows \((C_4, C_5, \ldots \text{ are some constants})\):
\[
\begin{align*}
\Pr\{C_1 < H_N \leq C_2\} & = \Pr\left(\frac{C_1}{\sqrt{N}} < H_N \leq \frac{C_2}{\sqrt{N}}\right) \\
& = \frac{C_2}{\sqrt{N}} \Phi\left(\frac{1}{2}\right) + \frac{C_1}{\sqrt{N}} \Phi\left(\frac{1}{2}\right) + o\left(\frac{1}{\sqrt{N}}\right) \\
& = C_2 \Phi\left(\frac{1}{2}\right) + C_1 \Phi\left(\frac{1}{2}\right) + o\left(\frac{1}{\sqrt{N}}\right)
\end{align*}
\]
(where \(C_7 > 0\)). If \(\eta\) is a lattice random variable, this argument must be slightly modified: as \(C_1\) (resp. \(C_2\)) we take the second largest (resp. the largest) non-positive middle point of the lattice of \(H_N\); the reference to Feller’s Theorem XVI.4.1 is replaced by reference to Theorem XVI.4.2. (Now \(C_1, C_2, \ldots\) depend on \(N\); however, \(C_2 - C_1\) does not depend on \(N\) and \(C_7\) is bounded below.) This completes the proof for the case \(\eta \in \min \xi, \max \xi\).

It remains to consider the case \(\eta \notin \min \xi, \max \xi\). By Lemma 16, when \(\eta \notin \min \xi, \max \xi\), we have \(A_\eta(\pi) = \infty\)
and, thus, (29) holds trivially. When \(\pi \in \min \xi, \max \xi\), we have \(A_\eta(\pi) = -\ln \Pr\{\xi = \pi\}\), and (29) follows from the obvious inequality
\[
\Pr\{S_N \leq xN\} \geq \Pr\{\xi = \pi\}^N.
\]

Proof of Lemma 18. Let \(\xi^*\) be the simple random variable defined by
\[
\Pr\{\xi^* = y\} = \frac{1}{w(\xi)} \Pr\{\xi = y\}, \quad \forall y \in \mathbb{R}.
\]

It is easy to see that
\[
A_\xi = A_{\xi^*} - \ln w(\xi).
\]

Now we can deduce from Lemma 17
\[
\begin{align*}
\Pr\{S_N \leq xN\} & = (w(\xi))^N \Pr\{S_N \leq xN\} \\
& \geq (w(\xi))^N \frac{1}{C \sqrt{N} + 1} \exp\left(-N A_{\xi^*}(\xi)\right) \\
& = \frac{1}{C \sqrt{N} + 1} \exp\left(-N A_{\xi^*}(\pi) + N \ln w(\xi)\right) \\
& = \frac{1}{C \sqrt{N} + 1} \exp\left(-N A_\xi(\pi)\right),
\end{align*}
\]
where \(S_N^\xi := \xi^* + \cdots + \xi^*_N\) and \(\xi^*_1, \ldots, \xi^*_N\) are independent and distributed as \(\xi^*\).

REFERENCES

33. V. Vovk, An optimality property of the weighted majority algorithm, unpublished manuscript.