How to Combine Expert (or Novice) Advice when Actions Impact the Environment

Daniela Pucci de Farias[*]
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139
pucci@mit.edu

Nimrod Megiddo
IBM Almaden Research Center
650 Harry Road, K53-B2
San Jose, CA 95120
megiddo@almaden.ibm.com

Abstract

The so-called "experts algorithms" constitute a methodology for choosing actions repeatedly, when the rewards depend both on the choice of action and on the unknown current state of the environment. An experts algorithm has access to a set of strategies ("experts"), each of which may recommend which action to choose. The algorithm learns how to combine the recommendations of individual experts so that, in the long run, for any fixed sequence of states of the environment, it does as well as the best expert would have done relative to the same sequence. This methodology may not be suitable for situations where the evolution of states of the environment depends on past chosen actions, as is usually the case, for example, in a repeated non-zero-sum game.

A new experts algorithm is presented and analyzed in the context of repeated games. It is shown that asymptotically, under certain conditions, it performs as well as the best available expert. This algorithm is quite different from previously proposed experts algorithms. It represents a shift from the paradigms of regret minimization and myopic optimization to consideration of the long-term effect of a player's actions on the opponent's actions or the environment. The importance of this shift is demonstrated by the fact that this algorithm is capable of inducing cooperation in the repeated Prisoner's Dilemma game, whereas previous experts algorithms converge to the suboptimal non-cooperative play.

1 Introduction

Experts algorithms. A well-known class of methods in machine learning are the so-called *experts algorithms*. The goal of these methods is to learn from experience how to combine advice from multiple experts in order to make sequential decisions in an online environment. The general idea can be described as follows. An agent has to choose repeatedly from a given set of actions. The reward in each stage is a function of the chosen action and the choices of Nature or the environment (also referred to as the "adversary" or the "opponent"). A set of strategies $\{1, \ldots, r\}$ is available for the agent to choose from. We refer

[*] Work done while at IBM Almaden Research Center, San Jose, California.

to each such strategy as an “expert,” even though some of them might be simple enough to be called a “novice.” Each expert suggests a choice of an action based on the history of the process and the expert’s own choice algorithm. After each stage, the agent observes his own reward. An experts algorithm directs the agent with regard to which expert to follow in the next stage, based on the past history of actions and rewards.

Minimum Regret. A popular criterion in decision processes is called Minimum Regret (MR). Regret is defined as the difference between the reward that could have been achieved, given the choices of Nature, and what was actually achieved. An expert selection rule is said to minimize regret if it yields an average reward as large as that of any single expert, against any fixed sequence of actions chosen by the opponent. Indeed, certain experts algorithms, which at each stage choose an expert from a probability distribution that is related to the reward accumulated by the expert prior to that stage, have been shown to minimize regret [1, 2]. It is crucial to note though that, since the experts are compared on a sequence-by-sequence basis, the MR criterion ignores the possibility that different experts may induce different sequences of choices by the opponent. Thus, MR makes sense only under the assumption that Nature’s choices are independent of the decision maker’s choices.

Repeated games. We consider a multi-agent interaction in the form of a repeated game. In repeated games, the assumption that the opponent’s choices are independent of the agent’s choices is not justified, because the opponent is likely to base his choices of actions on the past history of the game. This is evident in nonzero-sum games, where players are faced with issues such as how to coordinate actions, establish trust or induce cooperation. These goals require that they take each other’s past actions into account when making decisions. But even in the case of zero-sum games, the possibility that an opponent has bounded rationality may lead a player to look for patterns to be exploited in the opponent’s past actions.

We illustrate some of aforementioned issues with an example involving the Prisoner’s Dilemma game.

The Prisoner’s Dilemma. In the single-stage Prisoner’s Dilemma (PD) game, each player can either cooperate (C) or defect (D). Defecting is better than cooperating regardless of what the opponent does, but it is better for both players if both cooperate than if both defect. Consider the repeated PD. Suppose the row player consults with a set of experts, including the “defecting expert,” who recommends defection all the time. Let the strategy of the column player in the repeated game be fixed. In particular, the column player may be very patient and cooperative, willing to wait for the row player to become cooperative, but eventually becoming non-cooperative if the row player does not seem to cooperate. Since defection is a dominant strategy in the stage game, the defecting expert achieves in each step a reward as high as any other expert against any sequence of choices of the column player, so the row player learns with the experts algorithm to defect all the time. Obviously, in retrospect, this seems to minimize regret, since for any fixed sequence of actions by the column player, constant defection is the best response. Obviously, constant defection is not the best response in the repeated game against many possible strategies of the column player. For instance, the row player would regret very much using the experts algorithm if he were told later that the column player had been playing a strategy such as Tit-for-Tat.[1]

In this paper, we propose and analyze a new experts algorithm, which follows experts judiciously, attempting to maximize the long-term average reward. Our algorithm differs from previous approaches in at least two ways. First, each time an expert is selected, it is followed for multiple stages of the game rather than a single one. Second, our algorithm takes

[1] The Tit-for-Tat strategy is to play C in the first stage, and later play in every stage whatever the opponent played in the preceding stage.

into account only the rewards that were actually achieved by an expert in the stages it was followed, rather than the reward that could have been obtained in any stage. Our algorithm enjoys the appealing simplicity of the previous algorithms, yet it leads to a qualitatively different behavior and improved average reward. We present two results:

1. A "worst-case" guarantee that, in any play of the game, our algorithm achieves an average reward that is asymptotically as large as that of the expert that did best in the rounds of the game when it was played. The worst-case guarantee holds without any assumptions on the opponent's or experts' strategies.
2. Under certain conditions, our algorithm achieves an average reward that is asymptotically as large as the average reward that could have been achieved by the best expert, had it been followed exclusively. The conditions are required in order to facilitate learning and for the notion of a "best expert" to be well-defined.

The effectiveness of the algorithm is demonstrated by its performance in the repeated PD game, namely, it is capable of identifying the opponent's willingness to cooperate and it induces cooperative behavior.

The paper is organized as follows. The algorithm is described in section 2. A bound based on actual expert performance is presented in section 3. In section 4, we introduce and discuss an assumption about the opponent. This assumption gives rise to asymptotic optimality, which is presented in section 5.

2 The algorithm

We consider an "experts strategy" for the row player in a repeated two-person game in normal form. At each stage of the game, the row and column player choose actions $i \in I$ and $j \in J$, respectively. The row player has a reward matrix R, with entries $0 \leq R_{ij} \leq u$. The row player may consult at each stage with a set of experts $\{1, \ldots, r\}$, before choosing an action for the next stage. We denote by σ_e the strategy proposed by expert e, i.e., $\sigma_e = \sigma_e(h_s)$ is the proposed probability distribution over actions in stage s, given the history h_s. We refer to the row player as the agent and to the column player as the opponent.

Usually, the form of experts algorithms found in the literature is as follows. Denote by $M_e(s-1)$ the average reward achieved by expert e prior to stage s of the game[2]. Then, a reasonable rule is to follow expert e in stage s with a probability that is proportional to some monotone function of $M_e(s-1)$. In particular, when this probability is proportional to $\exp\{\eta_s M_e(s-1)\}$, for a certain choice of η_s, this algorithm is known to minimize regret [1, 2]. Specifically, by letting j_s $(s = 1, 2, \ldots)$ denote the observed actions of the opponent up to stage s, and letting σ_X denote the strategy induced by the experts algorithm, we have

$$\sum_{s'=1}^{s} \mathrm{E}[R(i, j_s) : i \sim \sigma_X(h_s)] \geq \sup_e \frac{1}{s} \sum_{s'=1}^{s} \mathrm{E}[R(i, j_s) : i \sim \sigma_e(h_s)] - o(s). \qquad (1)$$

The main deficiency of the regret minimization approach is that it fails to consider the influence of chosen actions of a player on the future choices of the opponent — the inequality (1) holds for any *fixed* sequence (j_s) of the opponent's moves, but does not account for the fact that different choices of actions by the agent may induce different sequences of the opponent. This subtlety is also missing in the experts algorithm we described above. At each

[2] In different variants of the algorithm and depending on what information is available to the row player, $M_e(s-1)$ could be either an estimate of the average reward based on reward achieved by expert e in the stages it was played, or the reward it could have obtained, had it been played in all stages against the same history of play of the opponent.

stage of the game, the selection of expert is based solely on how well various experts have, or could have, done so far. There is no notion of learning how an expert's actions affect the opponent's moves. For instance, in the repeated PD game described in the introduction, assuming that the opponent is playing Tit-for-Tat, the algorithm is unable to establish the connection between the opponent's cooperative moves and his own.

Based on the previous observations, we propose a new experts algorithm, which takes into account how the opponent reacts to each of the experts. The idea is simple: instead of choosing a (potentially different) expert at each stage of the game, the number of stages an expert is followed, each time it is selected, increases gradually. We refer to each such set of stages as an "iteration" of the algorithm. Following is the statement of the *Strategic Experts Algorithm* (SEA). The iteration number is denoted by i. The number of iterations during which expert e has been followed is denoted by N_e. The average payoff from iterations in which expert e has been followed is denoted by M_e.

Strategic Experts Algorithm (SEA):

1. For $e = 1, \ldots, r$, set $M_e = N_e = 0$. Set $i = 1$.
2. With probability $1/i$ perform an *exploration iteration*, namely, choose an expert e from the uniform distribution over $\{1, \ldots, r\}$; otherwise, perform an *exploitation iteration*, namely, choose an expert e from the uniform distribution over the set of experts e' with maximum $M_{e'}$.
3. Set $N_e = N_e + 1$. Follow expert e's instructions for the next N_e stages. Denote by $\tilde{R}$ the average payoff accumulated during the current iteration (i.e., these N_e stages), and set
$$M_e = M_e + \tfrac{2}{N_e+1}\,(\tilde{R} - M_e)\ .$$
4. Set $i = i + 1$ and go to step 2.

Throughout the paper, s will denote a stage number, and i will denote an iteration number. We denote by $M_1(i), \ldots, M_r(i)$ the values of the registers $M_1, \ldots, M_r$, respectively, at the end of iteration i. Similarly, we denote by $N_1(i), \ldots, N_r(i)$ the values of the registers $N_1, \ldots, N_r$, respectively, at the end of iteration i. Thus, $M_e(i)$ and $N_e(i)$ are, respectively, the average payoff accumulated by expert e and the total number of iterations this expert was followed on or before iteration i. We will also let $M(s)$ and $M(i)$ denote, without confusion, the average payoff accumulated by the algorithm in the first s stages or first i iterations of the game.

3 A bound based on actual expert performance

When the SEA is employed, the average reward $M_e(i)$ that was actually achieved by each available expert e is being tracked. It is therefore interesting to compare the average reward $M(s)$ achieved by the SEA, with the averages achieved by the various experts. The following theorem states that, in the long run, the SEA obtains almost surely at least as much as the actual average reward obtained by any available expert during the same play.

Theorem 3.1.
$$\Pr\left(\liminf_{s\to\infty} M(s) \geq \max_e \liminf_{i\to\infty} M_e(i)\right) = 1\ . \tag{2}$$

Although the claim of Theorem 3.1 seems very close to regret minimization, there is an essential difference in that we compare the average reward of our algorithm with the average reward *actually achieved* by each expert in the stages when it was played, as opposed to the estimated average reward based on the whole history of play of the opponent.

Note that the bound (2) is merely a statement about the average reward of the SEA in comparison to the average reward achieved by each expert, but nothing is claimed about the limits themselves. Theorem 5.1 proposes an application of this bound in a case when an additional assumption about the experts' and opponent's strategies allows us to analyze convergence of the average reward for each expert. Another interesting case occurs when one of the experts plays a maximin strategy; in this case, bound (2) ensures that the SEA achieves at least the maximin value of the game. The same holds if one of the experts is a regret-minimizing experts algorithm, which is known to achieve at least the maximin value of the game.

The proof of Theorem 3.1 is based on several lemmas. The first lemma is a basic fact about the vanishing frequency of exploration iterations.

For every iteration i, let Z_i be a random variable such that $Z_i = 1$ if iteration i is an exploration iteration; otherwise, $Z_i = 0$. Also, let $c > 0$ be any constant and let e be any expert. Define random variable iteration numbers I_k $(k = 1, 2, \ldots)$ by

$$I_k = \min\{i \mid N_e(i) = k\}$$

(where $\min \emptyset = \infty$).

Lemma 3.1. *With probability one,*

$$\lim_{k\to\infty} \sup_n \left\{ \tfrac{1}{n} \sum_{i=I_k+1}^{I_k+n} Z_i \ : \ n \geq cN_e(I_k) \right\} = 0 \, .$$

Proof: Let k be fixed. We first show that the random variables $Z_{I_k+1}, Z_{I_k+2}, \ldots$ are conditionally independent given I_k; furthermore, for every $n \geq 1$,

$$\Pr(Z_{I_k+n} \mid I_k = j) = \frac{1}{j+n} \, .$$

For any iteration j, denote $N_e^j = (N_e(1), \ldots, N_e(j))$ and denote by W_j the set of all the possible sequences of values of N_e^j, which may occur during in the play, so that $I_k = j$. Explicitly,

$$W_j = W_j(k) = \{(w_1, \ldots, w_j) \mid w_{j-1} = k-1, \ w_j = k, w_{i-1} \leq w_i, \ i = 2, \ldots, j\} \, .$$

Note that if $N_e^j = w \in W_j(k)$, then $I_k = j$. Let $n \geq 1$ and let $a_i \in \{0,1\}^n$, $i = 1, \ldots, n$, be any numbers. For any j,

$$\begin{aligned}
&\Pr(Z_i = a_i, \ i = j+1, \ldots, j+n \mid I_k = j) \\
&= \sum_{w \in W_j} \Pr(Z_i = a_i, \ i = j+1, \ldots, j+n \mid I_k = j, \ N_e^j = w) \cdot \Pr(N_e^j = w \mid I_k = j) \\
&= \sum_{w \in W_j} \Pr(Z_i = a_i, \ i = j+1, \ldots, j+n \mid N_e^j = w) \cdot \Pr(N_e^j = w \mid I_k = j) \, .
\end{aligned}$$

Since the decision whether or not to perform an exploration iteration is independent of the number of times any expert has been previously followed, it follows that

$$\Pr(Z_i = a_i, \ i = j+1, \ldots, j+n \mid N_e^j = w) = \Pr(Z_i = a_i, \ i = j+1, \ldots, j+n) \, .$$

Thus,

$$\begin{aligned}
\Pr(Z_i = a_i, \ i = j+1, \ldots, j+n \mid I_k = j) \\
&= \prod_{i=j+1}^{j+n} \Pr(Z_i = a_i) \cdot \sum_{w \in W_j} \Pr(N_e^j = w \mid I_k = j) \\
&= \prod_{i=j+1}^{j+n} \Pr(Z_i = a_i)
\end{aligned}$$

We have $I_k \geq N_e(I_k) = k$, and therefore

$$z_{k,n} \equiv \tfrac{1}{n}\,\mathbf{E}\sum_{i=I_k+1}^{I_k+n} Z_i \leq \frac{1}{k}.$$

Let ϵ be any positive number. Using Hoeffding's inequality , we have for all $k \geq \frac{2}{\epsilon}$

$$\begin{aligned}
\Pr\left(\tfrac{1}{n}\sum_{i=I_k+1}^{I_k+n} Z_i \geq \epsilon | I_k\right) &= \Pr\left(\tfrac{1}{n}\sum_{i=I_k+1}^{I_k+n} Z_i - z_{k,n} \geq \epsilon - z_{k,n} | I_k\right) \\
&\leq \exp\left\{-2n(\epsilon - 1/k)^2\right\} \\
&\leq \exp\left\{-n\epsilon^2/2\right\} \;.
\end{aligned}$$

We conclude that

$$\Pr\left(\tfrac{1}{n}\sum_{i=I_k+1}^{I_k+n} Z_i \geq \epsilon\right) \leq \exp\left\{-\tfrac{1}{2}n\,\epsilon^2\right\}.$$

It follows that

$$\begin{aligned}
\Pr\left(\sup_{n \geq cn_e(I_k)} \tfrac{1}{n}\sum_{i=I_k+1}^{I_k+n} Z_i > \epsilon\right) &= \Pr\left(\sup_{n \geq ck} \tfrac{1}{n}\sum_{i=I_k+1}^{I_k+n} Z_i > \epsilon\right) \\
&\leq \sum_{n=ck}^{\infty} \exp\left\{-\tfrac{1}{2}n\epsilon^2\right\} \\
&\leq 2\epsilon^{-2}\,\exp\left\{-\tfrac{1}{2}(ck-1)\,\epsilon^2\right\}.
\end{aligned}$$

Finally,

$$\sum_{k=\lceil\frac{2}{\epsilon}\rceil}^{\infty} \Pr\left(\sup_{n \geq cn_e(I_k)} \tfrac{1}{n}\sum_{i=k+1}^{k+n} Z_i > \epsilon\right) \;\leq\; \sum_{k=\lceil\frac{2}{\epsilon}\rceil}^{\infty} 2\epsilon^{-2}\exp\left\{-\tfrac{1}{2}(ck-1)\,\epsilon^2\right\} < \infty\;,$$

so by the Borel-Cantelli lemma, with probability one, for only finitely many value of k,

$$\sup\left\{\tfrac{1}{n}\sum_{i=I_k+1}^{I_k+n} Z_i \;:\; n \geq cn_e(I_k)\right\} > \epsilon\;.$$

The lemma follows since the latter holds for every $\epsilon > 0$. □

Lemma 3.2. *For every $\epsilon > 0$, any expert e and any two iterations $i < i'$, if*

$$M_e(i') \leq M_e(i) - \epsilon\;,$$

then

$$N_e(i') - N_e(i) \geq \frac{N_e(i)\epsilon}{3\bar{R}}\;.$$

Proof: Fix $\epsilon > 0$, e and $i < i'$. For simplicity, denote $N = N_e(i)$, and $v = M_e(i)$. Let $I_0 = i$, and for $j = 1, \ldots, k$, let I_j, denote the j^{th} iteration such that expert e is played after iteration i and on or before iteration i'. Then we have $N_e(I_j) = N + j$.
Since all payoffs of the game are nonnegative, we have

$$M_e(I_j) \geq M_e(I_{j-1})\,\frac{(N+j-1)(N+j)}{(N+j)(N+j+1)}\;.$$

A simple induction argument yields

$$M_e(I_k) \geq v \, \frac{N(N+1)}{(N+k)(N+k+1)} . \tag{3}$$

By hypothesis, we also have

$$v - \epsilon \geq M_e(I_k) \; . \tag{4}$$

Combining (3) and (4) and rearranging terms, we get that k must satisfy

$$k^2 + k\,(2N+1) - \frac{\epsilon}{v-\epsilon}\, N(N+1) \geq 0 \; . \tag{5}$$

Denoting $\bar{k} = \epsilon N/3u$,

$$\begin{aligned}\bar{k}^2 + \bar{k}(2N+1) \, - \frac{\epsilon}{v-\epsilon}\, N(N+1) &\leq^{(a)} \bar{k}^2 + \bar{k}(N+1) - \frac{\epsilon}{u}\, N(N+1) \\ &= \left(\frac{\epsilon}{u}\right)^2 \frac{N^2}{9} + \frac{\epsilon}{u} \cdot \frac{N(2N+1)}{3} - \frac{\epsilon}{u}\, N(N+1) \\ &\leq^{(b)} \frac{\epsilon}{u} \cdot N \left(\frac{N}{9} + \frac{2N+1}{3} - (N+1) \right) < 0 \; .\end{aligned}$$

Inequality (a) follows from $v \leq u$, and inequality (b) follows from $\epsilon \leq u$, which must hold since all payoffs are between 0 and u . We conclude that $k \geq \bar{k}$, and the lemma follows. □

The next lemma is the center piece in the proof of Theorem 3.1. Loosely speaking, after the game has been played long enough, if an expert e has average payoff significantly less than

$$\max_{e'} \liminf_{i \to \infty} M_{e'}(i) \; ,$$

then e must have been played during only a small fraction of the iterations. For a logical proposition ϕ, let $\delta[\phi] = 1$ if ϕ is true; otherwise $\delta[\phi] = 0$.

Lemma 3.3. *Let $\bar{e}$ be any expert and define a random variable V by*

$$V = \liminf_{i \to \infty} M_{\bar{e}}(i) \; .$$

Then, for every expert e and $\epsilon > 0$, with probability one,

$$\lim_{i \to \infty} \frac{N_e(i) \cdot \delta\left[M_e(i) \leq V - \epsilon\right]}{i} = 0 \; . \tag{6}$$

Proof: We let $Z_i = 1$ if iteration i was an exploration iteration; otherwise $Z_i = 0$. We use the phrase “Almost all histories” to represent “all histories except for a set of histories with zero probability.”

Let an expert e be fixed. We distinguish three types of histories of the game according to the asymptotic behavior of $M_e(i)$.

Type 1: $\liminf_{i \to \infty} M_e(i) > V - \epsilon$. For all histories of type 1, inequality (6) holds trivially.

Type 2: $\limsup_{i \to \infty} M_e(i) < V$. For histories of type 2, we use the following upper bound on $N_e(I)$:

$$N_e(I) \leq \sum_{i=1}^{I} \left\{ \, \delta\left[M_e(i) > M_{\bar{e}}(i)\right] \, + \, Z_i \, \right\} \; .$$

By the strong law of large numbers, with probability one,

$$\lim_{I \to \infty} \frac{1}{I} \sum_{i=1}^{I} Z_i = 0 \; ,$$

and for every history, if

$$\limsup_{i\to\infty} M_e(i) < \liminf_{i\to\infty} M_{\bar{e}}(i) \ ,$$

then

$$\sum_{i=1}^{\infty} \delta[\, M_e(i) > M_{\bar{e}}(i)] < \infty$$

and

$$\lim_{I\to\infty} \frac{1}{I} \sum_{i=1}^{I} \delta[\, M_e(i) > M_{\bar{e}}(i)] = 0 \ .$$

We conclude that inequality (6) holds for almost all histories of type 2.

Type 3: $\liminf_{i\to\infty} M_e(i) \le V - \epsilon$ and $\limsup_{i\to\infty} M_e(i) \ge V$. Let τ be any constant, $0 < \tau < \epsilon$. We denote $I_0 = 0$ and define the following random variables for $k = 1, 2, \ldots$:

$$\begin{aligned} I_k &= \min\left\{i > I_{k-1} : M_e(i) \le V - \epsilon + \tfrac{\tau}{2}\right\} \\ I_k^- &= \max\left\{i < I_k : M_e(i) \ge V - \tfrac{\epsilon}{2}\right\} \ . \end{aligned}$$

Note that $I_k^- = I_{k'}^-$ may hold for $k \ne k'$. Note also that, for all histories of type 3, I_k is finite for all k, and I_k^- tends to infinity as k tends to infinity. Since

$$M_e(I_k) \le V - \epsilon + \tfrac{\tau}{2} \le M_e(I_k^-) - \tfrac{\epsilon-\tau}{2} \ ,$$

Lemma 3.2 yields the following bound on $N_e(I_k)$:

$$\begin{aligned} N_e(I_k) &= N_e(I_k) - N_e(I_k^-) + N_e(I_k^-) \\ &\le \left(1 + \tfrac{6u}{\epsilon-\tau}\right) \left(N_e(I_k) - N_e(I_k^-)\right) \ . \end{aligned} \qquad (7)$$

Also,

$$N_e(I_k) - N_e(I_k^-) \le \sum_{i=I_k^-+1}^{I_k} \left\{\delta[\, M_{\bar{e}}(i) \le V - \tfrac{\epsilon}{2}\,] + Z_i\,\right\} \ . \qquad (8)$$

It follows from (7) and (8) that

$$\begin{aligned} \frac{N_e(I_k)}{I_k} &\le \left(1 + \tfrac{6u}{\epsilon-\tau}\right) \frac{N_e(I_k) - N_e(I_k^-)}{I_k - I_k^-} \\ &\le \left(1 + \tfrac{6u}{\epsilon-\tau}\right) \left[\frac{\sum_{i=I_k^-+1}^{I_k} \delta[\, M_{\bar{e}}(i) \le V - \frac{\epsilon}{2}\,]}{I_k - I_k^-} + \frac{\sum_{i=I_k^-+1}^{I_k} Z_i}{I_k - I_k^-} \right] \\ &\le \left(1 + \tfrac{6u}{\epsilon-\tau}\right) \frac{\sum_{i=I_k^-+1}^{I_k} \delta[\, M_{\bar{e}}(i) \le V - \frac{\epsilon}{2}\,]}{I_k - I_k^-} \qquad (9) \\ &\quad + \left(1 + \tfrac{6u}{\epsilon-\tau}\right) \sup_{n \ge N_e(I_k^-)(\epsilon-\tau)/(6u)} \frac{\sum_{i=I_k^-+1}^{I_k^-+n} Z_i}{n} \ , \end{aligned}$$

where in the last inequality we have applied Lemma 3.2 to bound

$$I_k - I_k^- \ge N_e(I_k) - N_e(I_k^-) \ge N_e(I_k^-)\tfrac{\epsilon-\tau}{6u} \ .$$

The term in (9) converges to zero for almost all histories of type 3 because I_k^- tends to infinity with k and, with probability one, $\delta[\, M_{\bar{e}}(i) \le V - \frac{\epsilon}{2}]$ converges to 0. Recall that for all histories of type 3, $N_e(I_k^-)$ tends to infinity with k. Moreover, since $M_e(I_k^- + 1) \ne M_e(I_k^-)$, it follows that $N_e(I_k^-) < N_e(I_k^- + 1)$, so that $I_k^- + 1$, $k = 1, 2, \ldots$

is a subsequence of the sequence of iteration numbers when expert e was selected. Thus, Lemma 3.1 can be applied to conclude that the second term in (9) converges to zero for almost all such histories. □

Our last lemma establishes a lower bound on the total number of stages played up to iteration i of the game.

Lemma 3.4. *For every iteration* i

$$\sum_{e=1}^{r} N_e(i)(N_e(i)+1) \geq i(i/r+1).$$

Proof: At each iteration i, $N_e(i)$ must satisfy

$$\sum_{e} N_e(i) = i.$$

Therefore, $\sum_{e=1}^{r} N_e(i)(N_e(i)+1)$ is bounded from below by the solution of the following quadratic minimization problem:

$$\begin{array}{ll} \min_x & \sum_e x_e(x_e+1) \\ \text{s.t.} & \sum_e x_e = i \; . \end{array}$$

Convexity and symmetry imply that the symmetric solution $x_e = \frac{i}{r}$ (for every e) is optimal. □

Proof of Theorem 3.1: For every $\epsilon > 0$ and every iteration i,

$$\begin{aligned} R(i) &= \frac{\sum_e N_e(i)(N_e(i)+1)M_e(i)}{\sum_e N_e(i)(N_e(i)+1)} \\ &\geq V - \epsilon - u \cdot \frac{\sum_e N_e(i)(N_e(i)+1)\,\delta[\,M_e(i) \leq V-\epsilon\,]}{\sum_e N_e(i)(N_e(i)+1)} \\ &\geq V - \epsilon - u \cdot \frac{\sum_e N_e(i)(N_e(i)+1)\,\delta[\,M_e(i) \leq V-\epsilon\,]}{i(\frac{i}{r}+1)} \; , \qquad (10) \end{aligned}$$

where the second inequality follows from Lemma 3.4. From Lemma 3.3, we conclude that, with probability one, the second term in (10) converges to zero. Since this holds for any ϵ,

$$\liminf_{i\to\infty} M(i) \geq \max_e \liminf_{i\to\infty} M_e(i).$$

Finally, let s be any stage. Let I_s be the largest iteration number I such that

$$\sum_{i=1}^{I} N_{e_i}(i) \leq s \; .$$

Note that

$$s \leq \sum_{i=1}^{I_s+1} N_{e_i}(i) \leq \frac{(I_s+1)(I_s+2)}{2} \; ,$$

so that I_s tends to infinity with s. The average reward at stage s satisfies

$$\begin{aligned} M(s) &\geq M(I_s)\,\frac{1}{2s}\sum_e N_e(I_s)(N_e(I_s)+1) \\ &\geq \tfrac{1}{2}M(I_s)\,\frac{\sum_e N_e(I_s)(N_e(I_s)+1)}{\sum_e N_e(I_s)(N_e(I_s)+1)/2 + N_{e_{I_s+1}}(I_s)+1} \\ &\geq M\left(I_s\right)\frac{1}{1+2/N_{e_{I_s+1}}}\ . \end{aligned}$$

The theorem follows by taking liminf on both sides as s tends to infinity. Note that for every e, with probability one, $N_e(i)$ tends to infinity with i. □

4 The flexible opponent

In general, it is impossible for an experts algorithm to guarantee, against an unknown opponent, a reward close to that of the best available expert. It is easy to construct examples which prove this impossibility.

Example: Repeated Matching Pennies. In the Matching Pennies (MP) game, each of the player and the adversary has to choose either H ("Heads") or T ("Tails"). If the choices match, the player loses 1; otherwise, he wins 1. A possible strategy for the adversary in the repeated MP game is:

Adversary: Fix a positive integer s and a string $\sigma^s \in \{H,T\}^s$. In each of the first s stages, play the 50:50 mixed strategy. In each of the stages $s+1, s+2, \ldots$, if the sequence of choices of the player during the first s stages coincided with the string σ^s, then play T; otherwise, play the 50:50 mixed strategy.

Suppose each available expert e corresponds to a strategy of the form:

Expert: Fix a string $\sigma_e \in \{H,T\}^s$. During the first s stages play according to σ_e. In each of the stages $s+1, s+2, \ldots$, play H.

Suppose an expert e^* with $\sigma_{e^*} = \sigma^s$ is available. Then, in order for an experts algorithm to achieve at least the reward of e^*, it needs to follow the string σ^s precisely during the first s stages. Of course, without knowing what σ^s is, the algorithm cannot play it with probability one, nor can it learn anything about it during the play.

In view of the repeated MP example, some assumption about the opponent must be made in order for the player to be able to learn how to play to against that opponent. The essence of the difficulty with the above strategy of the opponent is that it is not flexible — the player has only one chance to guess who the best expert is and thus cannot recover from a mistake. Here, we introduce the assumption of *flexibility* as a possible remedy to that problem. Under the assumption of flexibility, the SEA achieves an average reward that is asymptotically as high as what the best expert could be expected to achieve.

Definition 4.1 (Flexibility). *(i) An opponent playing strategy $\pi(s)$ is said to be* flexible with respect to expert e ($e = 1, \ldots, r$) *if there exist constants μ_e, $\tau > 0.25$ and c such that for every stage s_0, every possible history h_{s_0} at stage s_0 and any number of stages s,*

$$\mathbf{E}\left[\ \left|\tfrac{1}{s}\textstyle\sum_{s=s_0+1}^{s_0+s} R(a_e(s), b(s)) - \mu_e\right|\ :\ a_e(s) \sim \sigma_e(h_s),\ b(s) \sim \pi(h_s)\right] \leq \frac{c}{s^\tau}$$

(ii) Flexibility with respect to a set of experts is defined as flexibility with respect to every member of the set.

In words, the expected average reward during the s stages between stage s_0 and stage s_0+s converges (as s tends to infinity) to a limit that does not depend on the history of the play prior to stage s_0.

Example 4.1 : Finite Automata. In the literature on "bounded rationality", players are often modelled as finite automata. A *probabilistic automaton strategy* (PAS) is specified by a tuple $\mathcal{A} = \langle M, O, A, \sigma, P \rangle$, where $M = \{1, \ldots, m\}$ is the finite set of internal states of the automaton, A is the set of possible actions, $\mathcal{O}$ is the set of possible outcomes, $\sigma_i(a)$ is the probability of choosing action a while in state i ($i = 1, \ldots, m$) and $P^o = (P^o_{ij})$ ($1 \le i, j \le m$) is the matrix of state transition probabilities, given an outcome $o \in \mathcal{O}$. Thus, at any stage of the game, the automaton picks an action from a probability distribution associated with its current state and transitions into a new state, according to a probability distribution which depends on the outcome of the stage game. If both the opponent and an expert play PASs, then a Markov chain is induced over the set of pairs of the respective internal states. If this Markov chain has a single class of recurrent states, then the flexibility assumption holds. Note that we do not limit the size of the automata; a larger set of internal states implies a slower convergence of the average rewards, but does not affect the asymptotic results for the SEA.

Example 4.2 : Bounded dependence on the history. The number of possible histories at stage s grows exponentially with s. Thus, it is reasonable to assume that the choice of action would be based not on the exact detail of the history but rather on the empirical distribution of past actions or patterns of actions. If the opponent is believed not to be stationary, then discounting previous observations by recency may be sensible. For instance, if the frequency of play of action j by the opponent is relevant, the player might condition his choice at stage $s+1$ on the quantities $\tau_j = \sum_{s'=1}^{s} \beta^{s-s'} \delta_{jj_s}$ where $\beta < 1$ and δ is the Kronecker delta. In this case, only actions j_s at stages s that are relatively recent have a significant impact on τ_j. Therefore strategies based on τ_j should exhibit behavior similar to that of bounded recall, and lead to flexibility in the same circumstances as the latter.

5 A bound based on expected expert performance

In this section we show that if the opponent is "flexible" with respect to the available experts, then the SEA achieves almost surely an average payoff that is asymptotically as large as what the best expert could achieve against the same opponent.

Theorem 5.1. *If an opponent π is flexible with respect to the experts $1, \ldots, r$, then the average payoff up to stage s, $M(s)$, satisfies*

$$\Pr\left(\liminf_{s\to\infty} M(s) \ge \max_e \mu_e\right) = 1 \ .$$

Theorem 5.1 follows from Lemma 5.2, stated and proven below, and Theorem 3.1.

Flexibility comes into play as a way of ensuring that the value of following any given expert is well-defined, and can eventually be estimated as long as the SEA follows that expert a sufficiently many times. In other words, flexibility ensures that there is a best expert to be learned, and that learning can effectively occur because actions taken by other experts, which could affect the behavior of the opponent, are eventually forgotten by the latter.

Before Lemma 5.2, we present the following auxiliary lemma, which establishes a simple fact about the number of iterations each expert is selected.

Lemma 5.1. *With probability one, each expert e is selected infinitely many iterations.*

Fix an arbitrary expert strategy σ_e, and consider the event Z_{ie} defined as: “at iteration i, expert e is selected from the uniform distribution over the set of all experts.” Clearly, for a fixed e, the event: “Infinitely many events Z_{ie} occur” is a subset of the event “Expert e is selected infinitely many times,” so it suffices to show that the former event has probability one. The events Z_{ie} are independent and $\Pr(Z_{ie}) = \frac{1}{ir}$. Hence,

$$\sum_{i=1}^{\infty} \Pr(Z_{ie}) = \frac{1}{r} \sum_{i=1}^{\infty} \frac{1}{i} = \infty,$$

and by the Borel-Cantelli lemma, with probability one, infinitely many Z_{ie} s occur. □

We now present Lemma 5.2, which shows that, under the flexibility assumption, the average reward achieved by each expert is asymptotically almost surely the same as the reward that would have been achieved by the same expert, had he been the only available expert.

Lemma 5.2. *If the opponent is flexible with respect to expert e, then with probability one, $\lim_{i\to\infty} M_e(i) = \mu_e$.*

Proof: Let e be any expert. By Lemma 5.1, with probability one, expert e is followed during infinitely many iterations. Let $I_j = I_j(e)$, $(j = 1, 2, \ldots)$ be the iteration numbers in which e is followed. Denote by $\tilde{R}(I_j)$ the average payoff during iteration I_j, and denote $\epsilon_{I_j} = \tilde{R}(I_j) - \mu_e$. Note that $|\epsilon_{I_j}| \le u$. By Markov’s inequality, for every $\epsilon > 0$,

$$\Pr(|M_e(I_j) - \mu_e| > \epsilon) \le \frac{\mathbf{E}[(M_e(I_j) - \mu_e)^4]}{\epsilon^4} \ .$$

If we could show that

$$\sum_{j=1}^{\infty} \mathbf{E}[(M_e(I_j) - \mu_e)^4] < \infty \ , \tag{11}$$

then we could conclude, by the Borel-Cantelli lemma, that with probability one, the inequality $|M_e(I_j) - \mu_e| > \epsilon$ holds only for finitely many values of j. This, in turn, would imply that, with probability one, $\lim_{i\to\infty} M_e(i) = \mu_e$. We will show that, if the opponent is flexible with respect to expert e, as n tends to infinity we have

$$\mathbf{E}[(M_e(I_n) - \mu_e)^4] = O(n^{-1-\nu}), \tag{12}$$

for some $\nu > 0$, which suffices for (11).
We have

$$\begin{aligned}
\tfrac{n^4(n+1)^4}{16}\, \mathbf{E}[(M_e(I_n) - \mu_e)^4] &= \mathbf{E}\left[\left(\textstyle\sum_{j=1}^{n} j\, \epsilon_{I_j}\right)^4\right] \\
&= \sum_{j} j^4 \mathbf{E}[\epsilon_{I_j}^4] + 4 \sum_{j\ne k} j^3 k\, \mathbf{E}[\epsilon_{I_j}^3 \epsilon_{I_k}] + 6 \sum_{j\ne k} j^2 k^2 \mathbf{E}[\epsilon_{I_j}^2 \epsilon_{I_k}^2] \\
&\quad + 12 \sum_{j\ne k\ne \ell} j^2 kl\, \mathbf{E}[\epsilon_{i_j}^2 \epsilon_{I_k} \epsilon_{I_\ell}] + 24 \sum_{j\ne k\ne \ell\ne m} jklm\, \mathbf{E}[\epsilon_{I_j} \epsilon_{I_k} \epsilon_{I_\ell} \epsilon_{I_m}] \\
&\le n^5\, u^4 + 10 n^6 u^4 + 12 n^4 \sum_{j\ne k\ne \ell} \mathbf{E}[\, \epsilon_{I_j}^2 \cdot |\epsilon_{I_k} \epsilon_{I_\ell}|\,] \\
&\quad + 24 n^4 \sum_{j\ne k\ne \ell\ne m} \mathbf{E}[\, |\epsilon_{I_j} \epsilon_{I_k} \epsilon_{I_\ell} \epsilon_{I_m}|\,] \ .
\end{aligned}$$

By definition, the number of stages during iteration I_k is at least k. Hence,

$$\mathbf{E}[\, |\epsilon_{I_k}|\,] \le c k^{-\tau} \ .$$

Assume, without loss of generality, that $k < \ell$. The number of stages during iteration I_ℓ is at least ℓ. Denote by H^ℓ the history at the beginning of I_ℓ. By the definition of flexibility, for any realization of H^ℓ,

$$\mathbf{E}[\,|\epsilon_{I_\ell}| \text{ given } H^\ell\,] \le c\ell^{-\tau}\ .$$

It follows that

$$\begin{aligned}\mathbf{E}[\,\epsilon_{I_j}^2 \cdot |\epsilon_{I_k}\epsilon_{I_\ell}|\,] &\le u^2\,\mathbf{E}[\,|\epsilon_{I_k}\epsilon_{I_\ell}|\,] \\ &= u^2\,\mathbf{E}[\,|\epsilon_{I_k}|] \cdot \mathbf{E}[\,|\epsilon_{I_\ell}| \text{ given } \epsilon_{I_k}\,] \\ &\le u^2\,\mathbf{E}[\,|\epsilon_{I_k}|] \cdot \max_{H^\ell} \mathbf{E}[\,|\epsilon_{I_\ell}| \text{ given } H^\ell\,] \\ &\le u^2\,c^2 k^{-\tau}\ell^{-\tau}\ .\end{aligned}$$

Assuming, without loss of generality, $\tau < 1$ and using the inequality

$$\sum_{j=1}^{n} j^{-\tau} \le 1 + \frac{n^{1-\tau}}{1-\tau}$$

we get

$$\begin{aligned}\frac{1}{n^4}\sum_{j\ne k\ne \ell} \mathbf{E}[\,\epsilon_{I_j}^2 \cdot |\epsilon_{I_k}\epsilon_{I_\ell}|\,] &\le \frac{c^2u^2}{n^3}\left(\sum_{j=1}^{n} j^{-\tau}\right)^2 \\ &\le \frac{c^2u^2}{n^3}\left(1+\frac{n^{1-\tau}}{1-\tau}\right)^2 = O(n^{-(1+2\tau)})\ .\end{aligned}$$

Finally, a similar analysis gives

$$\mathbf{E}[\,|\epsilon_{I_j}\epsilon_{I_k}\epsilon_{I_\ell}\epsilon_{I_m}|\,] \le c^4 j^{-\tau}k^{-\tau}\ell^{-\tau}m^{-\tau},$$

and

$$\begin{aligned}\frac{1}{n^4}\sum_{j\ne k\ne \ell\ne m} \mathbf{E}[\,|\epsilon_{I_j}\epsilon_{I_k}\epsilon_{I_\ell}\epsilon_{I_m}|\,] &\le \frac{c^4}{n^4}\left(\sum_{j=1}^{n} j^{-\tau}\right)^4 \\ &\le \frac{c^4}{n^4}\left(1+\frac{n^{1-\tau}}{1-\tau}\right)^4 = O(n^{-4\tau})\ .\end{aligned}$$

Since $\tau > 0.25$, condition (12) holds, and the lemma follows. □

Proof of Theorem 5.1: Theorem 5.1 follows immediately from Lemma 5.2 and Theorem 3.1.

Example 5.1 : Repeated Prisoner's Dilemma revisited. Consider playing the repeated PD game against an opponent who plays Tit-for-Tat, and suppose there are only two experts: "Always defect" (AD) and "Always cooperate" (AC). Thus, AC induces cooperation in every stage and yields a payoff higher than AD, which induces defection in every stage of the game except the first one. It is easy to verify that Tit-for-Tat is flexible with respect to the experts AC and AD. Therefore, Theorem 5.1 holds and the SEA achieves an average payoff at least as much as that of AC. By contrast, as mentioned in the introduction, in order to minimize regret, the standard experts algorithm must play D in almost every stage of the game, and therefore achieves a lower payoff.

References

[1] Auer, P., Cesa-Bianchi, N., Freund, Y. & Schapire, R.E. (1995) Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proc. 36th Annual IEEE Symp. on Foundations of Computer Science*, pp. 322–331, Los Alamitos, CA: IEEE Computer Society Press.

[2] Freund, Y. & Schapire, R.E. (1999) Adaptive game playing using multiplicative weights. *Games and Economic Behavior* **29**:79–103.

[3] Foster, D. & Vohra, R. (1999) Regret and the on-line decision problem. *Games and Economic Behavior* **29**:7–35.

[4] Fudenberg, D. & Levine, D.K. (1997) *The Theory of Learning in Games.* Cambridge, MA: The MIT Press.

[5] Littlestone, N. & Warmuth, M.K. (1994) The weighted majority algorithm. *Information and Computation* **108** (2):212–261.