



## Playing against no-regret players

Maurizio D 'Andrea

### ► To cite this version:

| Maurizio D 'Andrea. Playing against no-regret players. 2022. hal-03573153

**HAL Id: hal-03573153**

**<https://hal.science/hal-03573153>**

Preprint submitted on 15 Feb 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

---

# PLAYING AGAINST NO-REGRET PLAYERS

---

**Maurizio D’Andrea**

ANITI and Toulouse School of Economics, UT Capitole  
maurizio.dandrea@tse-fr.eu

February 15, 2022

## ABSTRACT

In increasingly different contexts, it happens that a human player has to interact with artificial players who make decisions following decision-making algorithms. How should the human player play against these algorithms to maximize his utility? Does anything change if he faces one or more artificial players? The main goal of the paper is to answer these two questions.

Consider  $n$ -player games in normal form repeated over time, where we call the human player optimizer, and the  $(n - 1)$  artificial players, learners. We assume that learners play no-regret algorithms, a class of algorithms widely used in online learning and decision-making. In these games, we consider the concept of Stackelberg equilibrium. In a recent paper, Deng, Schneider, and Sivan have shown that in a 2-player game the optimizer can always guarantee an expected cumulative utility of at least the Stackelberg value per round. In our first result, we show, with counterexamples, that this result is no longer true if the optimizer has to face more than one player. Therefore, we generalize the definition of Stackelberg equilibrium introducing the concept of correlated Stackelberg equilibrium. Finally, in the main result, we prove that the optimizer can guarantee at least the correlated Stackelberg value per round. Moreover, using a version of the strong law of large numbers, we show that our result is also true almost surely for the optimizer utility instead of the optimizer’s expected utility.

**Keywords** Normal Form Games · Stackelberg Equilibrium · No-Regret

## 1 Introduction

Consider a  $n$ -player finite game repeated over time. In practice in online learning, strategies are often computed through decision-making algorithms and a class of these algorithms widely used and deeply studied are the No-Regret algorithms. A No-Regret strategy ensures that looking at the history of the strategies played, there is no action that if it had been played at each round would have guaranteed a better payoff. How should the human player (Optimizer) behave to face No-Regret learners? Is there any gameplay for the optimizer to increase his utility? Does anything change if he faces one or more artificial players? Answering these questions is the main focus of the paper.

An important role in our work is played by the concept of Stackelberg equilibrium of the game. The Stackelberg version of our game is a game where the optimizer chooses his strategy first and communicates his choice; after observing the strategy of the optimizer, the learners play a Nash equilibrium. Moreover, we define the Stackelberg value as the worst/best utility for the optimizer in any Stackelberg equilibrium. In a recent paper [1] Deng, Schneider and Sivan have proved that in games with an optimizer and a learner, if the learner plays No-Regret the optimizer can always guarantee an expected utility of at least the Stackelberg value per round. Starting from this result, the first question we ask ourselves is: is the result valid even in a game with more than one learner? The answer to this question is No. Indeed, we show that it is possible to find a simple counterexample with an optimizer and two learners where the learners play No-Regret but the optimizer receives an expected utility strictly lower than the Stackelberg value per round.

Given this negative result, we extend the definition of Stackelberg equilibrium so that it fits better games with more than one No-Regret learner. We first recall well-known results for No-Regret strategies. In a game in normal form

where every player plays No-Regret, the empirical average distribution converges almost surely to the set of correlated equilibrium distributions. Hence, we define the correlated Stackelberg equilibrium as an equilibrium where the optimizer plays first and communicates his strategy, then the learners play a correlated equilibrium instead of a Nash equilibrium. We define the correlated Stackelberg value as the worst/best utility for the optimizer in any correlated Stackelberg equilibrium. Thus, in our main result, we prove that in a  $n$ -player game with an optimizer and  $n - 1$  learners, if the learners play No-Regret the optimizer can guarantee an expected utility of at least the correlated Stackelberg value per round. Finally using a version of the strong law of large numbers we show that our main result is also true almost surely for the utility of the optimizer and not only for the expected utility of the optimizer.

**Other related work** The notion of No-Regret learning is one of the most famous and deeply studied in online learning and decision making. The first notion of Regret, called External-Regret, was given by Hannan in [2] and it was subsequently taken up in numerous other papers (see [3, 4, 5, 6]). Later another stronger Regret notion called Internal-Regret was introduced earlier in [7]. Several No-Internal Regret algorithms have been created, for example in [8, 9]. This notation is linked with the concept of correlated equilibrium introduced for the first time by Aumann in [10]. The paper [11] of Hart and Mas-Colell prove that in a  $n$ -player game in normal form where every player plays No-Internal Regret the empirical average distribution converges almost surely to the set of correlated equilibrium distributions. Moreover, if every player plays No-External Regret the empirical average distribution converges almost surely to the Hanna number of papers that study the behavior of No-regret algorithm in long term games, see for example [12, 13].

## 2 Model

### 2.1 Games and Equilibria

In this paper we consider  $n$ -player games, where each player has a finite set of strategies. Let  $N = \{1, \dots, n\}$  be the set of players and let  $A_i$  be the finite set of actions of player  $i$ . Denote with  $A = \prod_{i=1}^n A_i$  the set of strategy profiles. For every player  $i$  we define  $u_i : A \rightarrow \mathbb{R}$  the utility function of player  $i$ , i.e. if the players play the strategy profile  $a \in A$  the payoff of player  $i$  is  $u_i(a)$ . Players can randomize their strategies, so we define with  $\Delta_i$  the set of mixed strategies for the player  $i$  and denote  $\Delta = \prod_{i \in N} \Delta_i$ . As usual we may linearly extend the utility functions to the set of mixed strategies. For the rest of the article, we will refer to the  $n$ -th player as the optimizer and we will call the remaining players Learners and we call  $G$  the game described above.

**Definition 2.1.** A strategy  $x_i \in \Delta_i$  is a Best-Reply for player  $i$  against  $x_{-i} \in \Delta_{-i}$ <sup>1</sup> if

$$x_i \in \arg \max_{z \in \Delta_i} u_i(z, x_{-i}).$$

We denote  $BR^i(x_{-i})$  the set of all the best replies of player  $i$  against the mixed action  $x_{-i} \in \Delta_{-i}$ .

We are now ready to define the notion of equilibrium that we consider in the game  $G$  where there are  $n - 1$  learners and 1 optimizer. This notion was introduced for the first time by Von Stackelberg in [14]. A Stackelberg equilibrium is an equilibrium where the followers play a Nash equilibrium once they have observed the optimizer's mixed strategy. We refer to an optimistic Stackelberg equilibrium when the followers play a Nash equilibrium that maximizes the optimizer's utility, and to a pessimistic Stackelberg equilibrium when they play a Nash equilibrium that minimizes the optimizer utility.

**Definition 2.2.** The profile of strategies  $(a, x_n) \in A_{-n} \times \Delta_n$  is a pure Optimistic Stackelberg Equilibrium (POSE) if

$$u_n(a, x_n) = \max_{y \in \Delta_n} \max_{z \in A_{-n}} \{u_n(z, y) \mid z_i \in BR^i(z_{-i}, y) \forall i = 1, \dots, n - 1\}.$$

We call the value  $V_{pure} := u_n(a, x_n)$  the pure Optimistic Stackelberg value of the game if such an equilibrium exists.

The profile of strategies  $(a, x_n) \in A_{-n} \times \Delta_n$  is a pure Pessimistic Stackelberg Equilibrium (PPSE) if

$$u_n(a, x_n) = \max_{y \in \Delta_n} \min_{z \in A_{-n}} \{u_n(z, y) \mid z_i \in BR^i(z_{-i}, y) \forall i = 1, \dots, n - 1\}.$$

We call the value  $v_{pure} := u_n(a, x_n)$  the pure Pessimistic Stackelberg value of the game if such an equilibrium exists.

However, in a game  $G$  with  $n > 2$  pure Stackelberg Equilibria may not exist. Therefore we extend the definition of Stackelberg equilibrium in a natural way so that its existence is guaranteed in any  $n$ -player game.

<sup>1</sup>We use the notation  $x_{-i}$  to say that we consider all elements except element  $i$ ; i.e.  $x_{-i} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ . Moreover we denote  $\Delta_{-i} = \prod_{j \neq i} \Delta_j$

**Definition 2.3.** The profile of strategies  $x \in \Delta$  is a mixed Optimistic Stackelberg Equilibrium (MOSE) if

$$u_n(x) = \max_{y \in \Delta_n} \max_{z \in \Delta_{-n}} \{u_n(z, y) \mid z_i \in BR^i(z_{-i}, y) \forall i = 1, \dots, n-1\}.$$

We call the value  $V_{mixed} := u_n(x)$  the mixed Optimistic Stackelberg value of the game.

The profile of strategies  $x \in \Delta$  is a mixed Pessimistic Stackelberg Equilibrium (MPSE) if

$$u_n(x) = \max_{y \in \Delta_n} \min_{z \in \Delta_{-n}} \{u_n(z, y) \mid z_i \in BR^i(z_{-i}, y) \forall i = 1, \dots, n-1\}.$$

We call the value  $v_{mixed} := u_n(x)$  the mixed Pessimistic Stackelberg value of the game.

**Remark 2.1.** The values defined above respect the following chain of inequalities

$$v_{mixed} \leq v_{pure} \leq V_{pure} \leq V_{mixed}.$$

Moreover for  $n = 2$  we have  $V_{pure} = V_{mixed}$  and  $v_{pure} = v_{mixed}$ .

## 2.2 No-Regret

From here on we are interested in the game  $G$  repeated over time. Hence, we denote by  $a_i^t$  the realization at time  $t$  of the player  $i$ . Let  $a^t = (a_1^t, \dots, a_n^t)$  and we assume that players' utilities are additive over time. A strategy for the player  $i$  is an element  $\sigma_i = (\sigma_i^t)_{t \geq 1}$  where, for each  $t \in \mathbb{N}$ ,  $\sigma_i^t$  is a map from  $(A)^{t-1}$  to  $\Delta_i$ . Denote by  $\sigma = (\sigma_1, \dots, \sigma_n) \in \Sigma$  the profile of strategies. For each stage  $t$ ,  $u_i(a_1^t, \dots, a_n^t)$  is the random realized payoff at stage  $t$  for the player  $i$ . Finally, we denote with  $\mathbb{P}_\sigma$  the probability induced by the infinite sequences  $(a^t)_{t \geq 0}$ .

To describe a possible optimal behavior of the players in these types of repeated games the no-regret algorithms are the most commonly used.

**Definition 2.4.** A strategy  $\sigma_i$  of the learner  $i$  has No-Regret if for every profile of strategies  $\sigma_{-i}$ ,

$$\left( \max_{a \in A_i} \frac{1}{M} \sum_{t=1}^M (u_i(a, a_{-i}^t) - u_i(a^t)) \right)_+ \xrightarrow{M \rightarrow \infty} 0 \quad \mathbb{P}_\sigma \text{ a.s.}^2$$

**Definition 2.5.** A strategy  $\sigma_i$  of the learner  $i$  has No-Internal Regret if for every profile of strategies  $\sigma_{-i}$ ,

$$\max_{a, b \in A_i} \frac{1}{M} \sum_{\substack{t \in \{1, \dots, M\}, \\ a_i^t = a}} (u_i(b, a_{-i}^t) - u_i(a, a_{-i}^t)) \xrightarrow{M \rightarrow \infty} 0 \quad \mathbb{P}_\sigma \text{ a.s.}$$

Intuitively, No-Regret guarantees that strategies perform well against the best possible action, while No-Internal Regret guarantees that strategies perform as well as the best possible action over each subset of rounds where the same action is played. Thus, it is clear that the set of No-Internal Regret strategies is a subset of the set of No-Regret strategies.

We conclude this section with two well known results for No-Regret and No-Internal Regret strategies. For every  $t$ , let  $Z_t \in \Delta(A)$  be the empirical average distribution of the  $n$ -tuples of strategies played until time  $t$ , that is for every  $a \in A$

$$Z_t(a) := \frac{1}{t} |\{s \leq t : a^s = a\}|. \quad (1)$$

**Theorem 2.6** (Hart and Mas-Colell[11]). Let  $G$  be a game and let  $CED(G)$  be the set of correlated equilibrium distribution of the game  $G$ ; that is

$$CED(G) := \left\{ \Phi \in \Delta(A) : \sum_{a_{-i} \in A_{-i}} \Phi(a) u_i(a'_i, a_{-i}) \leq \sum_{a_{-i} \in A_{-i}} \Phi(a) u_i(a_i, a_{-i}), \quad \forall i \in N, a_i, a'_i \in A_i \right\}. \quad (2)$$

Then, if each player of the game  $G$  follows some No-Internal Regret procedure, the distance from the empirical distribution of moves to  $CED(G)$  converges a.s. to 0.

**Theorem 2.7.** Let  $G$  be a game and let  $\mathbb{H}(G)$  be the Hannan set of the game  $G$ ; that is

$$\mathbb{H}(G) := \left\{ \Phi \in \Delta(A) : \sum_{a \in A} \Phi(a) u_i(a'_i, a_{-i}) \leq \sum_{a \in A} \Phi(a) u_i(a), \quad \forall i \in N, a'_i \in A_i \right\}. \quad (3)$$

Then, if each player of the game  $G$  follows some No-Regret procedure, the distance from the empirical distribution of moves to the Hannan set  $\mathbb{H}(G)$  converge a.s. to 0.

<sup>2</sup>We use the notation  $(f)_+ = \max\{f, 0\}$

### 3 Preliminary Results

We start presenting a result given by Deng, Schneider, and Sivan in [1] which shows how in a two-player game, the optimizer can achieve an average utility per round arbitrarily close to the pure optimistic Stackelberg value against a no-regret learner.

**Definition 3.1.** A strategy  $a_i \in A_i$  is very weakly dominated if there exists  $x_i \in \Delta(A_i - \{a_i\})$  such that for all  $a_{-i} \in A_{-i}$ ,  $u_i(x_i, a_{-i}) \geq u_i(a_i, a_{-i})$ .

**Theorem 3.2** (Theorem 4 [1]). Let  $G$  be a game with two players ( $n = 2$ ) repeated  $M$  times. Let  $V_{\text{pure}}$  be the pure optimistic Stackelberg value of the game  $G$ . Assume that the learner doesn't have very weakly dominated strategies. If the learner is playing no-regret, then for every  $\epsilon > 0$  the optimizer can guarantee at least  $(V_{\text{pure}} - \epsilon)M - o(M)$  expected cumulative utility.

The goal of this section is to show that this result is false in a game with more than two players. To prove it we show how to construct a class of no-internal regret strategies in the following specific case.

**Example 3.1.** Let  $G$  be a 2-player game  $G$  with set of actions  $A_1 = \{T, B\}$ ,  $A_2 = \{L, R\}$  and utility function for player 1 given by the matrix

$$\begin{array}{cc} & \begin{array}{cc} L & R \end{array} \\ \begin{array}{c} T \\ B \end{array} & \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \end{array}$$

Then a strategy  $\sigma_1$  of player 1 with no-internal regret can be constructed as follows. Play arbitrarily at stage 1. At the end of stage  $t$ , for  $t \geq 1$ , compute the vector of regret  $\bar{r}^t = \begin{pmatrix} 0 & \bar{r}_{T,B}^t \\ \bar{r}_{B,T}^t & 0 \end{pmatrix}$  with

$$\bar{r}_{T,B}^t = \frac{1}{t} \sum_{\substack{m \in \{1, \dots, t\}, \\ a_1^m = T}} (\mathbb{1}_{a_2^m = R} - \mathbb{1}_{a_2^m = L})$$

and

$$\bar{r}_{B,T}^t = \frac{1}{t} \sum_{\substack{m \in \{1, \dots, t\}, \\ a_1^m = B}} (\mathbb{1}_{a_2^m = L} - \mathbb{1}_{a_2^m = R}).$$

And at stage  $t + 1$ , play  $T$  with probability  $p \in [0, 1]$  such that

$$p \cdot (\bar{r}_{T,B}^t)_+ = (1 - p) \cdot (\bar{r}_{B,T}^t)_+. \quad (4)$$

Now we are ready to prove that Theorem 3.2 is false for  $n > 2$ .

**Counterexample 3.1.** Consider a 3-player game where  $A_1 = \{T, B\}$ ,  $A_2 = \{L, R\}$ ,  $A_3 = \{E\}$  and the payoff matrix is

$$\begin{array}{cc} & \begin{array}{cc} L & R \end{array} \\ \begin{array}{c} T \\ B \end{array} & \begin{pmatrix} (1, 1, 0) & (0, 0, 0) \\ (0, 0, -1) & (1, 1, 1) \end{pmatrix}. \end{array}$$

We can define a no-regret strategy  $\sigma_1$  for Player 1 as in the Example 3.1 with the assumption that at the first step he plays  $T$  and if  $p$  in (4) is arbitrary he chooses  $p = 1$ . Moreover, we can repeat the argument for player 2 reversing the role of  $T$  with  $L$  and  $B$  with  $R$ . Thus we have  $\sigma_1$  and  $\sigma_2$  no-regret strategies respectively for player 1 and player 2. Then

$$\begin{aligned} r_{T,B}^1 &= -1 \quad \text{and} \quad r_{B,T}^1 = 0 \Rightarrow \sigma_1^2 = T \\ r_{L,R}^1 &= -1 \quad \text{and} \quad r_{R,L}^1 = 0 \Rightarrow \sigma_2^2 = L. \end{aligned}$$

Repeating the reasoning we have  $\sigma_1^t = T$ ,  $\sigma_2^t = L$  for every  $t \geq 1$ . Hence

$$\mathbb{E} \left( \sum_{t=1}^M u_3(a_1^t, a_2^t, E) \right) = \sum_{t=1}^M u_3(T, L, E) = 0.$$

However in this game  $V_{\text{pure}} = 1$ , so Theorem 3.2 is false for  $n > 2$ .

The natural question that now arises is: can Theorem 3.2 be true for one of the other Stackelberg values defined in 2.2? To answer this question, we will modify the previous counterexample so that the optimizer can not guarantee the mixed Pessimistic Stackelberg value per round.

**Counterexample 3.2.** Consider a 3-player game  $G$  as in the Counterexample 3.1 but with payoffs given by

$$\begin{array}{cc} & L & R \\ \begin{array}{c} T \\ B \end{array} & \begin{pmatrix} (1, 1, 0) \\ (0, 0, -1) \end{pmatrix} & \begin{pmatrix} (0, 0, 1) \\ (1, 1, 0) \end{pmatrix} \end{array}.$$

In this game it is easy to see that  $v_{mixed} = 0$ . Assume that player 1 and player 2 play no-regret as in the Counterexample 3.1 with the difference that if at the stage  $t$  the choice of  $p$  in (4) is arbitrary, player 1 play

$$p = \begin{cases} 1 & \text{if } \sigma^{t-1} = (B, L) \\ 0 & \text{if } \sigma^{t-1} = (T, L) \text{ or } \sigma^{t-1} = (B, R) \\ \text{arbitrary} & \text{otherwise} \end{cases}$$

and Player 2 plays

$$p = \begin{cases} 1 & \text{if } \sigma^{t-1} = (B, R) \text{ or } \sigma^{t-1} = (B, L) \\ 0 & \text{if } \sigma^{t-1} = (T, L) \\ \text{arbitrary} & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} r_{T,B}^1 &= -1 \quad \text{and} \quad r_{B,T}^1 = 0 \Rightarrow \sigma_1^2 = B \\ r_{L,R}^1 &= -1 \quad \text{and} \quad r_{R,L}^1 = 0 \Rightarrow \sigma_2^2 = R. \\ r_{T,B}^2 &= -\frac{1}{2} \quad \text{and} \quad r_{B,T}^2 = -\frac{1}{2} \Rightarrow \sigma_1^3 = B \\ r_{L,R}^2 &= -\frac{1}{2} \quad \text{and} \quad r_{R,L}^2 = -\frac{1}{2} \Rightarrow \sigma_2^3 = L. \\ r_{T,B}^3 &= -\frac{1}{3} \quad \text{and} \quad r_{B,T}^3 = 0 \Rightarrow \sigma_1^4 = T \\ r_{L,R}^3 &= 0 \quad \text{and} \quad r_{R,L}^3 = -\frac{1}{3} \Rightarrow \sigma_2^4 = L, \end{aligned}$$

and the process repeats. Hence, this couple of no-regret strategies induce the following sequence of action:  $(T, L), (B, R), (B, L), (T, L), (B, R), (B, L), \dots$  and the cumulative payoff for the optimizer is

$$\frac{1}{M} \mathbb{E} \left( \sum_{t=1}^M u_3(a_1^t, a_2^t, E) \right) = -\frac{1}{3} < v_{mixed}.$$

## 4 Main Result

In the counterexamples 3.1 and 3.2 the learners have played a 2 player game and their No-Internal Regret behavior has generated a sequence of actions whose frequency represents a correlated equilibrium of the game. Therefore, in this section we generalize the definition of Stackelberg equilibrium using the concept of correlated equilibrium instead of Nash equilibrium and we show that the optimizer can guarantee the value of this *correlated Stackelberg equilibrium*.

**Definition 4.1.** The profile of strategies  $(\gamma, \alpha) \in \Delta(A_{-n}) \times \Delta_n$  is a *correlated Pessimistic Stackelberg Equilibrium (CPSE)* if

$$u_n(\gamma, \alpha) = \max_{y \in \Delta_n} \min_{\Phi \in CED(G_y)} u_n(\Phi, y),$$

Where  $G_y$  is the  $n - 1$  player game in which the payoffs are given by  $u_i(\cdot, y)$ ; in this game players do not observe realizations of actions of player  $n$ . We call the value  $v_{corr} := u_n(\gamma, \alpha)$  the *correlated Pessimistic Stackelberg value* of the game.

Similarly, we can define the Hannan Stackelberg equilibrium as follow.

**Definition 4.2.** The profile of strategies  $(\gamma, \alpha) \in \Delta(A_{-n}) \times \Delta_n$  is an *Hannan Pessimistic Stackelberg Equilibrium (HPSE)* if

$$u_n(\gamma, \alpha) = \max_{y \in \Delta_n} \min_{\Phi \in \mathbb{H}(G_y)} u_n(\Phi, y),$$

We call the value  $v_h := u_n(\gamma, \alpha)$  the *Hannan Pessimistic Stackelberg value* of the game.

**Theorem 4.3.** Let  $v_{\text{corr}}$  be the correlated pessimistic Stackelberg value of the game  $G$ . If the learners are playing no-internal regret, then for every  $\epsilon > 0$  the optimizer can guarantee at least  $(v_{\text{corr}} - \epsilon)M$  expected cumulative utility for  $M$  large enough.

To prove Theorem 4.3 we recall the version of the strong law of large numbers for uncorrelated random variables.

**Theorem 4.4 (Strong Law of Large Number).** Let  $\{X_t\}_{t \in \mathbb{N}}$  be a sequence of random variable such that  $\mathbb{E}[X_t] = 0$  and  $\text{Var}[X_t] \leq C < \infty$ . If for every  $t, s$  in  $\mathbb{N}$  with  $t \neq s$ ,  $\mathbb{E}[X_t X_s] = 0$  then

$$\frac{\sum_{t=1}^M X_t}{M} \xrightarrow{M \rightarrow \infty} 0 \quad \text{a.s.}$$

**Corollary 4.5.** Let  $G$  be a  $n$ -player game. Let  $\sigma_n$  be the strategy of player  $n$  that consists of playing  $\alpha \in \Delta_n$  in every period. For each profile of strategies  $\sigma_{-n}$  of the other players, and for every  $i \in N$

$$\frac{1}{M} \sum_{t=1}^M (u_i(a_{-n}^t, a_n^t) - u_i(a_{-n}^t, \alpha)) \xrightarrow{M \rightarrow \infty} 0 \quad \mathbb{P}_\sigma \text{ a.s.}$$

Finally, we are ready to prove our main result.

*Proof Theorem 4.3.* Let  $(\gamma, \alpha) \in \Delta(A_{-n}) \times \Delta_n$  be a correlated Stackelberg equilibrium. Fix  $\epsilon > 0$ . Suppose that at each round the optimizer plays the strategy  $\alpha$ . Since the followers play no internal regret; for every  $i \in N$  and every  $a, b \in A_i$

$$\frac{1}{M} \sum_{t=1}^M \mathbb{1}_{a_i^t=a} (u_i(b, a_{-i}^t) - u_i(a, a_{-i}^t)) \leq o(1) \quad \mathbb{P}_\sigma \text{ a.s.} \quad (5)$$

Fix  $a, b \in A_i$ , then <sup>3</sup>

$$\begin{aligned} \frac{1}{M} \sum_{t=1}^M \mathbb{1}_{a_i^t=a} (u_i(b, a_{-i,n}^t, \alpha) - u_i(a, a_{-i,n}^t, \alpha)) &= \frac{1}{M} \sum_{t=1}^M \mathbb{1}_{a_i^t=a} (u_i(b, a_{-i,n}^t, \alpha) - u_i(b, a_{-i}^t)) \\ &\quad + \frac{1}{M} \sum_{t=1}^M \mathbb{1}_{a_i^t=a} (u_i(b, a_{-i}^t) - u_i(a, a_{-i}^t)) \\ &\quad + \frac{1}{M} \sum_{t=1}^M \mathbb{1}_{a_i^t=a} (u_i(a, a_{-i}^t) - u_i(a, a_{-i,n}^t, \alpha)) \\ &= \frac{1}{M} \sum_{t=1}^M (v_i(a_{-n}^t, \alpha) - v_i(a^t)) \quad (6) \\ &\quad + \frac{1}{M} \sum_{t=1}^M \mathbb{1}_{a_i^t=a} (u_i(b, a_{-i}^t) - u_i(a, a_{-i}^t)) \quad (7) \\ &\quad + \frac{1}{M} \sum_{t=1}^M (w_i(a^t) - w_i(a^t, \alpha)) \quad (8) \end{aligned}$$

Where  $v_i(a^t) := \mathbb{1}_{a_i^t=a} u_i(b, a_{-i}^t)$  and  $w_i(a^t) := \mathbb{1}_{a_i^t=a} u_i(a, a_{-i}^t)$ . Now, (6) and (8) go to 0  $\mathbb{P}_\sigma$  a.s. by Corollary 4.5, and using the inequality (5) we obtain that for every  $i \in N$  and for every  $a, b \in A_i$

$$\frac{1}{M} \sum_{t=1}^M \mathbb{1}_{a_i^t=a} (u_i(b, a_{-i,n}^t, \alpha) - u_i(a, a_{-i,n}^t, \alpha)) \leq o(1) \quad \mathbb{P}_\sigma \text{ a.s.}$$

---

<sup>3</sup>We use the notation  $a_{-i,n}$  to say that we consider all elements except element  $i$  and  $n$ ; i.e.  $a_{-i,n} = (a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_{n-1})$ .

Hence, denoting with  $Z_M \in \Delta(A_{-n})$  the empirical distribution; for every  $i \in N$  and for every  $a, b \in A_i$

$$\sum_{a_{-i,n} \in A_{-i,n}} Z_M(a, a_{-i,n})(u_i(b, a_{-i,n}, \alpha) - u_i(a, a_{-i,n}, \alpha)) \leq o(1) \quad \mathbb{P}_\sigma \text{ a.s.}$$

So,

$$\text{dist}(Z_M, \text{CED}(G_\alpha)) \xrightarrow{M \rightarrow \infty} 0 \quad \mathbb{P}_\sigma \text{ a.s.}$$

By dominated convergence theorem

$$\mathbb{E}_\sigma[\text{dist}(Z_M, \text{CED}(G_\alpha))] \xrightarrow{M \rightarrow \infty} 0;$$

therefore for every  $\delta > 0$  there exist  $M_0$  such that

$$\mathbb{E}[\text{dist}(Z_M, \text{CED}(G_\alpha))] \leq \delta \quad \forall M \geq M_0,$$

Denote by  $\Psi_t$  the projection of  $Z_t$  to the set  $\text{CED}(G_\alpha)$ , then

$$\begin{aligned} \mathbb{E} \left[ \frac{1}{M} \sum_{t=1}^M u_n(a_{-n}^t, a_n^t) \right] &= \mathbb{E} \left[ \frac{1}{M} \sum_{t=1}^M u_n(a_{-n}^t, \alpha) \right] = \mathbb{E} \left[ \sum_{a_{-n} \in A_{-n}} Z_M(a_{-n}) u_n(a_{-n}, \alpha) \right] \\ &\geq \mathbb{E} \left[ \sum_{a_{-n} \in A_{-n}} (Z_M(a_{-n}) - \Psi_M(a_{-n}) + \Psi_M(a_{-n})) u_n(a_{-n}, \alpha) \right] \\ &\geq \mathbb{E} \left[ \sum_{a_{-n} \in A_{-n}} \Psi_M(a_{-n}) u_n(a_{-n}, \alpha) \right] - \mathbb{E} \left[ \sum_{a_{-n} \in A_{-n}} (\Psi_M(a_{-n}) - Z_M(a_{-n})) u_n(a_{-n}, \alpha) \right] \\ &\geq \mathbb{E}[u_n(\Psi_M, \alpha)] - \mathbb{E} \left[ \sum_{a_{-n} \in A_{-n}} |\Psi_M(a_{-n}) - Z_M(a_{-n})| \cdot |u_n(a_{-n}, \alpha)| \right] \\ &\geq \mathbb{E}[u_n(\Psi_M, \alpha)] - \|u_n\|_\infty \mathbb{E} \left[ \sum_{a_{-n} \in A_{-n}} |\Psi_M(a_{-n}) - Z_M(a_{-n})| \right] \\ &\geq v_{\text{corr}} - \|u_n\|_\infty \mathbb{E}[\text{dist}(Z_M, \text{CED}(G_\alpha))] \geq v_{\text{corr}} - \|u_n\|_\infty \delta. \end{aligned}$$

We conclude choosing  $\delta = \frac{\epsilon}{\|u_n\|_\infty}$ . □

Now, if the learners play No-Regret instead of No-Internal regret, using the same argument we can prove that the optimizer can guarantee at least the Hannan pessimistic Stackelberg value.

**Theorem 4.6.** *Let  $v_h$  be the Hannan pessimistic Stackelberg value of the game  $G$ . If the learners are playing No-Regret, then for every  $\epsilon > 0$  the optimizer can guarantee at least  $(v_h - \epsilon)M$  expected cumulative utility for  $M$  large enough.*

*Proof.* Let  $(\gamma, \alpha) \in \Delta(A_{-n}) \times \Delta_n$  be an Hannan Stackelberg equilibrium. Fix  $\epsilon > 0$ . Suppose that at each round the optimizer plays the strategy  $\alpha$ . Since the followers play no regret; for every  $i \in N$  and every  $a \in A_i$

$$\frac{1}{M} \sum_{t=1}^M (u_i(a, a_{-i}^t) - u_i(a_i^t, a_{-i}^t)) \leq o(1) \quad \mathbb{P}_\sigma \text{ a.s.} \quad (9)$$

By Corollary 4.5 and inequality (9); repeating the passage in Theorem 4.3, we obtain that for every  $i \in N$  and for every  $a \in A_i$

$$\frac{1}{M} \sum_{t=1}^M (u_i(a, a_{-i,n}^t, \alpha) - u_i(a_i, a_{-i,n}^t, \alpha)) \leq o(1) \quad \mathbb{P}_\sigma \text{ a.s.}$$

Hence, denoting with  $Z_M \in \Delta(A_{-n})$  the empirical distribution; for every  $i \in N$  and for every  $a \in A_i$

$$\sum_{a_{-n} \in A_{-n}} Z_M(a_{-n})(u_i(a, a_{-i,n}, \alpha) - u_i(a_{-n}, \alpha)) \leq o(1) \quad \mathbb{P}_\sigma \text{ a.s.}$$



So,

$$\text{dist}(Z_M, \mathbb{H}(G_\alpha)) \xrightarrow{M \rightarrow \infty} 0 \quad \mathbb{P}_\sigma \text{ a.s.}$$

By dominated convergence theorem

$$\mathbb{E}[\text{dist}(Z_M, \mathbb{H}(G_\alpha))] \xrightarrow{M \rightarrow \infty} 0.$$

Denote by  $\Psi_t$  the projection of  $Z_t$  to the set  $\mathbb{H}(G_\alpha)$ , then repeating the inequality used in the proof of Theorem 4.3

$$\begin{aligned} \mathbb{E} \left[ \frac{1}{M} \sum_{t=1}^M u_n(a_{-n}^t, a_n^t) \right] &\geq \mathbb{E}[u_n(\Psi_M, \alpha)] - \|u_n\|_\infty \mathbb{E}[\text{dist}(Z_t, \mathbb{H}(G_\alpha))] \geq \\ &\geq v_h - \|u_n\|_\infty \mathbb{E}[\text{dist}(Z_t, \mathbb{H}(G_\alpha))] \geq v_h - \|u_n\|_\infty \delta. \end{aligned}$$

We conclude choosing  $\delta = \frac{\epsilon}{\|u_n\|_\infty}$ .  $\square$

We conclude observing that Corollary 4.5 allows us to adapt the proof of Theorem 4.3 and Theorem 4.6 to the payoff instead of the expected payoff.

**Theorem 4.7.** *Let  $v_{corr}$  ( $v_h$ ) be the correlated (Hannan) pessimistic Stackelberg value of the game  $G$ . If the learners are playing No-Internal Regret (No-Regret), then for  $M$  large enough the optimizer can guarantee a.s. at least  $Mv_{corr} - o(M)$  ( $Mv_h - o(M)$ ) cumulative utility.*

*Proof.* We prove the theorem in the case of no-internal regret. The case of no-regret is analogous. Let  $(\gamma, \alpha) \in \Delta(A_{-n}) \times \Delta_n$  be a correlated Stackelberg equilibrium. Suppose that at each round the optimizer plays the strategy  $\alpha$ . Since the followers play no-internal regret; as we have shown in the Theorem 4.3

$$\text{dist}(Z_M, \text{CED}(G_\alpha)) \xrightarrow{M \rightarrow \infty} 0 \quad \mathbb{P}_\sigma \text{ a.s.} \quad (10)$$

Now

$$\begin{aligned} \frac{1}{M} \sum_{t=1}^M u_n(a_{-n}^t, a_n^t) &= \frac{1}{M} \sum_{t=1}^M (u_n(a_{-n}^t, a_n^t) - u_n(a_{-n}^t, \alpha)) + \frac{1}{M} \sum_{t=1}^M u_n(a_{-n}^t, \alpha) \\ &\geq v_{corr} + \frac{1}{M} \sum_{t=1}^M (u_n(a_{-n}^t, a_n^t) - u_n(a_{-n}^t, \alpha)) - \|u_n\|_\infty \cdot \text{dist}(Z_M, \text{CED}(G_\alpha)) \\ &= v_{corr} - o(1) \quad \mathbb{P}_\sigma \text{ a.s.} \end{aligned}$$

Where in the last equality we use Corollary 4.5 and (10).  $\square$

## 5 Acknowledgements

I am grateful to my advisors Jérôme Renault and Fabien Gensbittel for helpful discussions. This research has benefited from the financial support of the ANR (Programme d'Investissement d'Avenir ANR-17-EURE-0010), and the AI Interdisciplinary Institute ANITI, which is funded by the French "Investing for the Future - PIA3" program under the Grant agreement ANR-19-PI3A-0004.

## References

- [1] Yuan Deng, Jon Schneider, and Balasubramanian Sivan. Strategizing against no-regret learners. *CoRR*, abs/1909.13861, 2019.
- [2] James Hannan. *4. Approximation to rayes risk in repeated play*, pages 97–140. Princeton University Press, 2016.
- [3] Dean P. Foster and Rakesh V. Vohra. A randomization rule for selecting forecasts. *Oper. Res.*, 41(4):704–709, 1993.
- [4] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. In *Proceedings of the 30th Annual Symposium on Foundations of Computer Science, SFCS '89*, page 256–261, USA, 1989. IEEE Computer Society.
- [5] Yoav Freund and Robert E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In Paul Vitányi, editor, *Computational Learning Theory*, pages 23–37, Berlin, Heidelberg, 1995. Springer Berlin Heidelberg.

- [6] Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, may 1997.
- [7] Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):279–390, 1998.
- [8] Nicolò Cesa-Bianchi and Gábor Lugosi. Potential-based algorithms in on-line prediction and game theory. *Mach. Learn.*, 51(3):239–261, jun 2003.
- [9] Andreu Mas-Colell Sergiu Hart. *Simple Adaptive Strategies: From Regret-Matching to Uncoupled Dynamics*. World Scientific Series in Economic Theory. World Scientific Publishing Company, 2013.
- [10] Robert J. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1):67–96, 1974.
- [11] Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. 1997.
- [12] Lampros Flokas, Emmanouil-Vasileios Vlatakis-Gkaragkounis, Thanasis Lianas, Panayotis Mertikopoulos, and Georgios Piliouras. No-regret learning and mixed nash equilibria: They do not mix, 2020.
- [13] Sergiu Hart and Andreu Mas-Colell. Uncoupled dynamics do not lead to nash equilibrium. *American Economic Review*, 93(5):1830–1836, December 2003.
- [14] Heinrich Von Stackelberg. Market structure and equilibrium. *Springer Science & Business Media*, 2010.
- [15] Joyee Deb, Julio Gonzalez-Diaz, and Jerome Renault. Uniform folk theorems in repeated anonymous random matching games. *Games and Economic Behavior*, 100(C), 2016.
- [16] R. Laraki, J. Renault, and S. Sorin. *Mathematical Foundations of Game Theory*. Springer, Cham, 2019.
- [17] Tim Roughgarden. Algorithmic game theory lecture 17: No-regret dynamics. 2013.
- [18] Sergiu Hart and Andreu Mas-Colell. A general class of adaptive strategies. *J. Econ. Theory*, 98(1):26–54, 2001.
- [19] D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.
- [20] Robert J. Aumann. Correlated equilibrium as an expression of bayesian rationality. *Econometrica*, 55(1):1–18, 1987.
- [21] Stefano Coniglio, Alberto Marchesi, and Nicola Gatti. Computing a pessimistic stackelberg equilibrium with multiple followers: the mixed-pure case. *Algorithmica*, November 2019.