POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Safely Guiding a No-Regret Learner to the Equilibrium

Tesi di Laurea Magistrale in
Computer Science and Engineering - Ingegneria Informatica

Author: **Francesco Emanuele Stradi**

Student ID: 944616
Advisor: Prof. Nicola Gatti
Co-advisors: Martino Bernasconi, Federico Cacciamani
Academic Year: 2021-22

Abstract

Is it possible to build an algorithm capable of teaching a complex rational agent, as a human, how to play a game?
To answer this question, it is necessary to explain what we mean by both human and teaching. As far as the latter is concerned, Game Theory proposes as solution the notion of Equilibrium, that is, a playing strategy which guarantees a certain level of reward whatever the opponent strategy is. Thus, learning how to play means reaching this equilibrium, while teaching means carrying the opponent to it.
In order to model the human properly, it is necessary to deal with at least two aspects: first, every person has different learning ability, second, an incentive to continue the game is needed to avoid a premature interruption of the learning path. The different learning abilities will be represented by the assumption that the opponent, which must be guided by our algorithm to the equilibrium, may employ an entire family of learning algorithms (technically, No-Regret); finally, the incentive to keep playing the game will be modeled by the safety property, which guarantees the opponent's reward to lie in an interval in every round. The interval is chosen as hyperparameter of the algorithm in order to prevent the human from getting bored, due to easy victories, or from giving up, due to tremendous defeats.
In this thesis we will show how, with proper assumptions, it is possible to build an algorithm that can guide the opponent to the equilibrium, without knowing the specific algorithm of the adversary and ensuring at the same time, not only the safety property, but also a sublinear learning time (namely, Dynamic Regret).
The problem will be tackled both in setting in which the "teacher" has a full feedback, and in setting in which he has a partial one, showing the differences in terms of theoretical results.

Keywords: Game Theory, Online Learning, Artificial Intelligence, Safety, Last Round Convergence

Abstract in lingua italiana

E' possibile costruire un algoritmo in grado di insegnare ad un agente razionale complesso, quale un umano, come approcciarsi ad un gioco?
E' innanzitutto doveroso chiarire cosa si intenda sia per umano che per insegnare. Per quanto concerne questo secondo aspetto, la Teoria dei Giochi propone come soluzione il concetto di Equilibrio, cioè una strategia di gioco tale da garantire un certo tipo di risultato qualsiasi sia la strategia del proprio avversario. Imparare a giocare significa dunque raggiungere questo equilibrio, mentre insegnare significa condurci il proprio avversario.
Per modellare in maniera consona un umano bisogna invece considerare almeno due aspetti: in primis, ogni persona ha capacità di apprendimento diverse, in secundis, è necessario che ci sia un incentivo a continuare il gioco, altrimenti il percorso di apprendimento verrebbe prematuramente interrotto. Le diverse capacità verranno declinate dall'assunzione che l'avversario, il quale deve essere condotto dal nostro algoritmo all'equilibrio, possa sfruttare non uno, bensì una famiglia di algoritmi di apprendimento (tecnicamente, algoritmi No-Regret); infine, l'incentivo a continuare il gioco verrà modellato dalla proprietà di safety, la quale garantisce che il risultato ottenuto dall'avversario in ogni partita giacia all'interno di un intervallo scelto come iperparametro dell'algoritmo, in modo da evitare che l'umano si annoi vincendo troppo facilmente o demorda perdendo gravemente.
In questa tesi mostreremo come, con le necessarie assunzioni, sia possibile costruire un algoritmo che conduca il proprio avversario all'equilibrio, senza conoscere l'algoritmo esatto dello sfidante e garantendo non solo la proprietà di safety, ma anche un tempo di apprendimento (tecnicamente, Regret dinamico) sublineare.
Il problema verrà affrontanto sia in situazioni in cui il "maestro" abbia un feedback totale, che un feedback parziale, mostrando le differenze in termini di garanzie teoriche.

Parole chiave: Teoria dei Giochi, Apprendimento Sequenziale, Intelligenza Artificiale, Safety, Convergenza all'Ultimo Turno

Contents

List of Figures

List of Tables

1 | Introduction

Algorithmic game theory and Online learning have recently contributed to significant achievements in the field of Artificial intelligence, leading to the deployment of artificial agents capable of defeating top professionals in several games such as chess [Campbell et al. 4], Go [Silver et al. 27] and poker [Brown and Sandholm 3]. So far, the great majority of multi-agent learning techniques developed to defeat human players does not take into account their peculiar behavior. This approach may lead the artificial agents not to adapt to the actual abilities of humans, generating an impressive gap in performances. Indeed, playing against a super-computer may not be endearing for most human players. As Egri-Nagy and Törmänen underline, it is "hopeless and frustrating to play against an AI, since it is practically impossible to win". In particular, humans are interested in repeatedly playing against an opponent when they are sufficiently engaged in the competition; thus, taking into account the actual human abilities becomes strictly necessary. The same reasoning holds if the aim of the algorithm is not defeating a human player, but teaching him how to reach an equilibrium. If the opponent/human is not engaged, he will not keep playing the game, interrupting the teaching dynamic. Different forms of engagement have been developed (for example, Abbasi et al. in computer games). In this thesis, we model the humans' engagement as a constraint over the utility the humans expect to receive. If this value is under a given threshold, the humans will get bored playing, as they lose too much and have no hope to win. Thus, they will drop out from the learning dynamic. Similarly, if such a value is above a given threshold, the humans will get bored playing, as they are winning too easily. Therefore, assuring engagement becomes fundamental when designing rational agents for humans; this is particularly true when the goal is not exploiting the opponent but educating him as in the world of serious games (see [Dörner et al. 13], [DeFalco et al. 10] for military, [Rossetti et al. 25] for transportation and [Wang et al. 28] for healthcare).

1.1. Goal

This thesis aims to develop algorithms capable of teaching human-like learners how to play games with strict competition (two-player zero-sum games) while interacting with them in an online fashion. Properly modeling a human presents many challenges; indeed, humans have different learning abilities, thus, we cannot make assumptions on the exact algorithm the opponent employs. Moreover, such algorithms must incentivize humans to keep playing the game since, in principle, they could interrupt the learning dynamic due to easy victories or catastrophic defeats. This incentive will be modeled through a constraint on the utility obtained by the players, namely, the per-round reward will always be bounded over an interval (please note that in zero-sum games, a bound on the utility of one of the player guarantees a bound on the utility of the opponent). As concerns the meaning of teaching, we want our algorithm to carry the human to the Nash Equilibrium (Minmax equilibrium for zero-sum games).

Our work will present the pseudo code of this type of algorithms and their theoretical guarantees in two different settings. In the first one, we will assume that players have expert feedback, namely, every player knows the reward he could have achieved playing any discrete distribution over his actions. In the latter, we will consider that the teacher can only observe the single action played by the human (the so-called partial semi-bandit feedback).

In the literature, different notions of learning have been proposed. The most common is the notion of *self-play learning*, in which an algorithm plays against copies of itself, with the average strategy converging to some equilibrium [Celli et al. 5, 6, Farina et al. 15]. The setting studied in this Thesis is substantially different as we do not assume to have control over the player algorithm and our objective is to achieve last round convergence as opposed to average strategy convergence.

1.2. Original Contribution

Our work starts by the framework proposed by Dinh et al. in their LRCA algorithm (3.4), that is, column player (in our thesis, the teacher) has full knowledge of the payoff matrix, and has as objective to drag his opponent (in our thesis, the human) to the Nash Equilibrium in Last Round. LRCA attains Last Round Convergence and No-Dynamic Regret property against the entire family of FTRL algorithm (2.5) in zero-sum games with fully-mixed equilibrium strategy for the row player.

We propose two versions of this algorithm: E-LRCA (algorithm 4.1) and PAUSE E-LRCA

(algorithm 5.1).
The first one deals with the Expert feedback setting and guarantees Last Round Convergence (see definition 3.2.2.1) and Sublinear Dynamic Regret (see definition 2.2.1.2) against the entire OMD family (2.8) in games with any kind of equilibrium (fully-mixed, partially-mixed, pure); in addition it guarantees safety (see definition 3.1.0.1) at each round, with a constraint on the upper bound of the safety region when there is not a fully-mixed equilibrium strategy for the row player.
The latter works in setting where the human/row player receives an expert feedback while the teacher receives the index of the action played by his opponent. In this case, PAUSE E-LRCA guarantees Last Round Convergence with high probability and Sublinear Dynamic Regret with respect to the value of the game (see definition 5.5.0.1) with high probability against the entire OMD family in games with fully-mixed equilibrium strategy for the row player (while, experimentally, these properties are valid even in absence of fully-mixed equilibrium); as concerns safety, it is guaranteed with high probability in case of fully-mixed equilibrium, otherwise, it is guaranteed with probability equal to one adding a constraint on the upper bound of the safety region.

The results are summarized in table 1.1.

Result Table

	Fully-mixed Equilibrium	**Not Fully-Mixed Equilibrium**
Expert Feedback	E-LRCA: • Safety • Last Round Convergence • Sublinear Dynamic Regret	E-LRCA: • Safety when $\|\|\boldsymbol{U}\boldsymbol{y}^*\|\|_\infty < \xi_2$ • Last Round Convergence • Sublinear Dynamic Regret
Partial Semi Bandit Feedback	PAUSE E-LRCA: • Safety with high probability • Last Round Convergence with high probability • Sublinear Dynamic Regret with respect to the MaxMin with high probability	PAUSE E-LRCA: • Safety when $\|\|\boldsymbol{U}\boldsymbol{y}^*\|\|_\infty < \xi_2$ • Experimental Last Round Convergence • Experimental Sublinear Dynamic Regret with respect to the MaxMin

Table 1.1: Table with the algorithms developed during the thesis and the final results obtained

1.3. Structure of the Thesis

The thesis is organized as follows:

- *Chapter 2* : In the preliminaries the necessary background to understand the thesis is highlighted. In addition we introduce the notation that will be employed in the core chapters.
- *Chapter 3* : Theoretical results obtained by other researchers in the area the thesis lies on are reported. In subsection 3.2.3 the algorithm which our thesis extends is shown.
- *Chapter 4* : Algorithm which deals with the expert feedback is reported, with theo-

rems/lemmas and related proofs. An entire section will be devoted for each property. [Core]

- *Chapter 5* : Algorithms which deal with the Partial Semi-Bandit feedback are reported, with theorems/lemmas and related proofs. An entire section will be devoted for each property. [Core]
- *Chapter 6* : Experiments related to all the algorithms reported in the core chapters of the thesis.
- *Chapter 7* : We show some interesting paths that future researchers may follows in order to improve our work.

2 | Preliminaries

In this chapter the background knowledge required to understand properly the thesis is reported.

2.1. Game Theory

2.1.1. Normal Form Game

In this subsection we present the normal form representation, also known as strategic form, arguably the most fundamental in game theory, as most other representations (e.g. Extensive Form Games) can be reduced to it. In particular we will focus mainly on two-player games.

Definition 2.1.1.1. *(Normal Form Game (Shoham and Leyton-Brown, 2008)) A (finite, two-person) normal-form game is a tuple* (N, A, u)*, where:*

- N *is a set of two players;*
- $A = A_1 \times A_2$*, where* A_1 *is a finite set of* n *actions available to player 1, while* A_2 *is a finite set of* m *actions available to player 2. Each vector* $(\boldsymbol{e}_1, \boldsymbol{e}_2) \in A$ *is called an action profile;*
- $u = (u_1, u_2)$ *where* $u_1 : A \to \mathbb{R}$ *is a real-valued utility (or payoff) function for player 1 and* $u_2 : A \to \mathbb{R}$ *is a real-valued utility (or payoff) function for player 2.*

A natural way to represent two-player normal form games is through a 2-dimensional payoff matrix; each row corresponds to a possible action for player 1, each column corresponds to a possible action for player 2 and each cell corresponds to one possible outcome. Each player's utility for an outcome is written in the cell corresponding to that outcome, with player 1's utility listed first.
We report here an example of Normal form game, known as the Prisoner Dilemma:

	C	D
C	-1,-1	-4,0
D	0,-4	-3,-3

There are some restricted classes of normal-form games that deserve special mention. The first is the class of common-payoff games. These are games in which, for every action profile, all players have the same payoff.
Now we present the category of games we will deal with during the thesis, that are constant sum games.

Definition 2.1.1.2. *(Constant Sum Game (Shoham and Leyton-Brown, 2008)) A two-player normal-form game is constant-sum if there exists a constant c such that for each strategy profile* $(\boldsymbol{e}_1, \boldsymbol{e}_2) \in A_1 \times A_2$ *it is the case that* $u_1(\boldsymbol{e}_1, \boldsymbol{e}_2) + u_2(\boldsymbol{e}_1, \boldsymbol{e}_2) = c$.

Throughout the entire thesis, we will always assume that c = 0, that is, we have a zero-sum game, a situation of pure competition; one player's gain comes at the expense of the other player. A classical example of a zero-sum game is the game of Matching Pennies. In this game, each of the two players has a penny and independently chooses to display either heads or tails. The two players then compare their pennies. If they are the same then player 1 pockets both, and otherwise player 2 pockets them. The payoff matrix is now shown:

	Heads	Tail
Heads	1,-1	-1,1
Tail	-1,1	1,-1

Another famous zero-sum game, that will be used in our final experiments, is Rock Paper Scissor, for which the payoff Matrix is:

	Rock	Paper	Scissor
Rock	0,0	-1,1	1,-1
Paper	1,-1	0,0	-1,1
Scissor	-1,1	1,-1	0,0

We have so far defined the actions available to each player in a game, but not yet his set of strategies or his available choices. Certainly one kind of strategy is to select a single action and play it. We call such a strategy a pure strategy. We call a choice of pure strategy for each agent a pure-strategy profile. Players could also follow another, less obvious type of strategy: randomizing over the set of available actions according to some

probability distribution. Such a strategy is called a mixed strategy. We define a mixed strategy for a normal-form game as follows.

Definition 2.1.1.3. *(Mixed Strategies (Shoham and Leyton-Brown, 2008)) Let (N, A, u) be a normal-form game, and for any set S let $\Pi(S)$ be the set of all probability distributions over S. Then the set of mixed strategies for player 1 is $\Delta_n = \Pi(A_1)$, while the set of mixed strategies for player 2 is $\Delta_m = \Pi(A_2)$, where Δ_d is the d-dimensional simplex.*

Definition 2.1.1.4. *(Mixed Strategy Profile (Shoham and Leyton-Brown, 2008)) The set of mixed-strategy profiles is simply the Cartesian product of the individual mixed-strategy sets, $\Delta_n \times \Delta_m$.*

Throughout the thesis, we will adopt the standard convention for two-player zero-sum games, that is, we define as $\boldsymbol{x}$ ($\in \Delta_n$) the strategy of the row player (first player): $\boldsymbol{x}$ is a vector of n dimension. We will refer as $\boldsymbol{x}(i)$ to the probability of the action $\boldsymbol{e}_i$. Similarly we will use $\boldsymbol{y}$ ($\in \Delta_m$) for the strategy of the column player (second player).

Definition 2.1.1.5. *(Support) The support (supp) of a mixed strategy $\boldsymbol{x}$ is the set of pure strategies $\{\boldsymbol{e}_i | \boldsymbol{x}(i) > 0\}$.*

Note that a pure strategy is a special case of a mixed strategy, in which the support is a single action. At the other end of the spectrum we have fully-mixed strategies. A strategy is fully-mixed if it has full support (namely, if it assigns every action a nonzero probability). We next define the payoffs of players given a particular strategy profile, since the payoff matrix defines those directly only for the special case of pure-strategy profiles. Formally, we define the expected utility as follows:

Definition 2.1.1.6. *(Expected Utility) Given a two-player normal-form game (N, A, u), the expected utility u_1 for player 1 of the mixed-strategy profile $(\boldsymbol{x}, \boldsymbol{y})$ is defined as:*

$$u_1(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{x}^\top \boldsymbol{U}_1 \boldsymbol{y}$$

where $\boldsymbol{U}_1$ is the payoff matrix of player 1, while the expected utility u_2 is defined as:

$$u_2(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{x}^\top \boldsymbol{U}_2 \boldsymbol{y}$$

where $\boldsymbol{U}_2$ is the payoff matrix of player 2.

Please note that we will deal with zero-sum games, which can be formalized by a single Payoff Matrix, with one of the two players that will try to minimize the utility of the

opponent, called the minimizer (in our case the row player), and the other that will try to maximize his utility (in our case the column player). From this:

Definition 2.1.1.7. *(Expected Utility in two-player zero-sum games) Given a two-player zero-sum normal-form game (N, A, u), the expected utility of the minimizer player given a mixed-strategy profile $(\boldsymbol{x}, \boldsymbol{y})$ is defined as:*

$$u_{min}(\boldsymbol{x}, \boldsymbol{y}) := -\boldsymbol{x}^\top \boldsymbol{U} \boldsymbol{y}$$

while the expected utility of the maximizer is defined as:

$$u_{max}(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{x}^\top \boldsymbol{U} \boldsymbol{y}$$

where $\boldsymbol{U}$ is the payoff matrix.

2.1.2. Nash Equilibrium

Once that we have defined what games in normal form are and what strategies are available to players in them, the question is how to reason about such games. Game theorists deal with this problem by identifying certain subsets of outcomes, called solution concepts, that are interesting in one sense or another. In this section we describe the most fundamental solution concepts: the Nash equilibrium.
Our first observation is that if an agent knew how the other were going to play, his strategic problem would become simple. Specifically, he would be left with the single-agent problem of choosing a utility-maximizing action, that is the problem of determining his best response.
Formally:

Definition 2.1.2.1. *(Best response (Shoham and Leyton-Brown, 2008)) Player 1's best response to the opponent strategy $\boldsymbol{y}$ is the strategy $\arg\max_{\boldsymbol{x} \in \Delta_n} \boldsymbol{x}^\top \boldsymbol{U}_1 \boldsymbol{y}$ (while the value of the best response is $f(\boldsymbol{y}) := \max_{\boldsymbol{x} \in \Delta_n} \boldsymbol{x}^\top \boldsymbol{U}_1 \boldsymbol{y}$). Player 2's best response to the opponent strategy $\boldsymbol{x}$ is the strategy $\arg\max_{\boldsymbol{y} \in \Delta_m} \boldsymbol{x}^\top \boldsymbol{U}_2 \boldsymbol{y}$ (while the value of the best response is $f(\boldsymbol{x}) := \max_{\boldsymbol{y} \in \Delta_m} \boldsymbol{x}^\top \boldsymbol{U}_2 \boldsymbol{y}$).*

Unfortunately, in general an agent will not know what strategy the other player plans to adopt. Thus, the notion of best response is not a solution concept; it does not identify an interesting set of outcomes in this general case. However, we can leverage the idea of best response to define what is arguably the most central notion in non cooperative game theory, the Nash equilibrium.

Definition 2.1.2.2. *(Nash Equilibrium (Nash, 1950)) Given a two-player normal form game, a strategy profile* $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ *is a Nash equilibrium if* $\boldsymbol{x}^*$ *is the best response for player 1 and* $\boldsymbol{y}^*$ *is the best response for player 2.*

Intuitively, a Nash equilibrium is a stable strategy profile: no agent would want to change his strategy if he knew what strategy the other agent was following.
We can divide Nash equilibria into two categories, strict and weak, depending on whether or not every agent's strategy constitutes a unique best response to the other agent's strategy.

Definition 2.1.2.3. *(Strict Nash (Shoham and Leyton-Brown, 2008)) Given a two-player normal form game, a strategy profile* $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ *is a strict Nash equilibrium if,* $u_1(\boldsymbol{x}^*, \boldsymbol{y}^*) > u_1(\boldsymbol{x}, \boldsymbol{y}^*)$ *for all* $\boldsymbol{x} \in \Delta_n$ *and* $u_2(\boldsymbol{x}^*, \boldsymbol{y}^*) > u_2(\boldsymbol{x}^*, \boldsymbol{y})$ *for all* $\boldsymbol{y} \in \Delta_m$.

Definition 2.1.2.4. *(Weak Nash (Shoham and Leyton-Brown, 2008)) Given a two-player normal form game, a strategy profile* $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ *is a weak Nash equilibrium if,* $u_1(\boldsymbol{x}^*, \boldsymbol{y}^*) \geq u_1(\boldsymbol{x}, \boldsymbol{y}^*)$ *for all* $\boldsymbol{x} \in \Delta_n$, $u_2(\boldsymbol{x}^*, \boldsymbol{y}^*) \geq u_2(\boldsymbol{x}^*, \boldsymbol{y})$ *for all* $\boldsymbol{y} \in \Delta_m$ *and* $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ *is not a strict Nash equilibrium.*

As concerns Nash equilibria, it is important to underline that:

Theorem 2.1. *(Nash, 1951) Every game with a finite number of players and action profiles has at least one Nash equilibrium.*

We need a couple of more definitions in order to introduce the specific setting in which our algorithms will work.

Definition 2.1.2.5. *(Minmax, two-player) In a two-player game, the minmax strategy for player 1 against player 2 is* $\operatorname{argmin}_{\boldsymbol{x}\in\Delta_n} \max_{\boldsymbol{y}\in\Delta_m} \boldsymbol{x}^\top \boldsymbol{U}_2 \boldsymbol{y}$*, and player 1 minmax value is* $\min_{\boldsymbol{x}\in\Delta_n} \max_{\boldsymbol{y}\in\Delta_m} \boldsymbol{x}^\top \boldsymbol{U}_2 \boldsymbol{y}$.

Definition 2.1.2.6. *(Maxmin, two-player) In a two-player game, the maxmin strategy for player 2 against player 1 is* $\operatorname{argmax}_{\boldsymbol{y}\in\Delta_m} \min_{\boldsymbol{x}\in\Delta_n} \boldsymbol{x}^\top \boldsymbol{U}_2 \boldsymbol{y}$*, and player 2 maxmin value is* $\max_{\boldsymbol{y}\in\Delta_m} \min_{\boldsymbol{x}\in\Delta_n} \boldsymbol{x}^\top \boldsymbol{U}_2 \boldsymbol{y}$.

Now we report a fundamental theorem that will be necessary for all the future reasoning:

Theorem 2.2. *(Minimax theorem (Neumann, 1928)) In any finite, two-player, zero-sum game, in any Nash equilibrium each player receives a payoff that is equal to both his maxmin value and his minmax value.*

The minmax theorem demonstrates that maxmin strategies, minmax strategies and Nash equilibria coincide in two-player, zero-sum games. In particular, the previous theorem allows us to conclude that in two-player, zero-sum games:

1. Each player's maxmin value is equal to his minmax value, called the value of the game v
2. For both players, the set of maxmin strategies coincides with the set of minmax strategies
3. Any maxmin strategy profile (or, equivalently, minmax strategy profile) is a Nash equilibrium. Furthermore, these are all the Nash equilibria. Consequently, all Nash equilibria have the same payoff vector.

Nash equilibria in zero-sum games can be viewed graphically as a saddle point in a high-dimensional space. At a saddle point, any deviation of the agent lowers his utility and increases the utility of the other agent. More formally we have that:

$$(\boldsymbol{x}^*, \boldsymbol{y}^*) = \arg\min_{\boldsymbol{x}\in\Delta_n} \arg\max_{\boldsymbol{y}\in\Delta_m} \boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{y} \tag{2.1}$$

with :

$$\min_{\boldsymbol{x}\in\Delta_n} \max_{\boldsymbol{y}\in\Delta_m} \boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{y} = v \tag{2.2}$$

Finally, if the equilibrium is fully-mixed we will have:

$$\boldsymbol{x}^{*\top}\boldsymbol{U} = [v, v, \ldots, v] \qquad \boldsymbol{U}\boldsymbol{y}^* = [v, v, \ldots, v]^\top \tag{2.3}$$

while if it is partially-mixed (assuming all the first actions in the support of the equilibrium):

$$\boldsymbol{x}^{*\top}\boldsymbol{U} = [v, \ldots, v, \alpha_1, \ldots, \alpha_l] \qquad \boldsymbol{U}\boldsymbol{y}^* = [v, \ldots, v, \beta_1, \ldots, \beta_k]^\top \tag{2.4}$$

with $\alpha_i < v \; \forall i \in \{1, \ldots, l\}$ and $\beta_j > v \; \forall j \in \{1, \ldots, k\}$.

2.2. Online Learning and OCO

2.2.1. Setting

In online convex optimization (OCO), an online player iteratively makes decisions. At each round, the outcomes associated with the choices are unknown to the player, that is, after committing to a decision, the decision maker suffers a loss: every possible decision

incurs a possibly different loss, which is unknown beforehand. The losses can be chosen by an adversary and depend on the action taken by the decision maker.

Please note that in order to make this approach meaningful, it is necessary to make some assumptions. First, the losses determined by the adversary should not be allowed to be unbounded, otherwise the adversary could keep decreasing the scale of the loss at each step, and never allow the algorithm to recover from the loss of the first step. Moreover, the decision set must be somehow bounded, otherwise the opponent could assign high loss to all the strategies chosen by the player indefinitely, while setting apart some strategies with zero loss.

The Online Convex Optimization framework models the decision set as a convex set in Euclidean space denoted $\mathcal{K} \subseteq \mathbb{R}^n$, while the costs are modeled as bounded convex functions over $\mathcal{K}$. It is important to highlight that there exists a strong connection between Game Theory (see section 2.1) and Online Convex Optimization; indeed, the OCO framework can be seen as a repeated game. Formally:
At iteration t, the online player chooses $\boldsymbol{x}_t \in \mathcal{K}$. After the player has committed to this choice, a convex cost function $\ell_t \in \mathcal{L} : \mathcal{K} \to \mathbb{R}$ is revealed. $\mathcal{L}$ is the bounded family of cost functions available to the adversary. The cost incurred by the online player is $\ell_t(\boldsymbol{x}_t)$, namely, the value of the cost function for the choice $\boldsymbol{x}_t$.
Let T denote the total number of game iterations, we define the regret of the decision maker to be the difference between the total cost he has incurred and that of the best fixed decision in hindsight.
Let Al be an algorithm for OCO, which maps a certain game history to a decision in the decision set. We formally define the regret of Al after T iterations as:

$$R_T^{Al} := \sum_{t=1}^{T} \ell_t(\boldsymbol{x}_t) - \min_{\boldsymbol{x} \in \mathcal{K}} \sum_{t=1}^{T} \ell_t(\boldsymbol{x}) \tag{2.5}$$

Intuitively, an algorithm performs well if its regret is sublinear as a function of T, that is $R_T^{Al} = o(T)$, since this implies that on the average the algorithm performs as well as the best fixed strategy in hindsight. More formally:

Definition 2.2.1.1. *(No-Regret) An algorithm Al is a no-regret algorithm (or has no-regret) if for every adversary,* $\lim_{T\to\infty} \frac{R_T^{Al}}{T} = 0$.

We then introduce a stronger notion of Regret, that is the Dynamic Regret. Please note that, in this case, for the sake of simplicity, we will introduce the Regret with respect to a reward and not with respect to a loss (goal is to maximize $\ell(\boldsymbol{x})$).

We define the Dynamic Regret as:

$$DR_T^{Al} := \sum_{t=1}^{T} \left(\max_{\boldsymbol{x} \in \mathcal{K}} \ell_t(\boldsymbol{x}) - \ell_t(\boldsymbol{x}_t) \right) \tag{2.6}$$

Definition 2.2.1.2. *(No-Dynamic Regret) An algorithm Al is a no-dynamic regret algorithm (or has the no-dynamic regret property) if* $\lim_{T\to\infty} \frac{DR_T^{Al}}{T} = 0$.

2.2.2. Follow the Leader

In an OCO setting of regret minimization, the most natural approach for the online player is to use at any time the optimal decision (namely, the best point in $\mathcal{K}$) in hindsight. Formally, let:

$$\boldsymbol{x}_{t+1} := \arg\min_{\boldsymbol{x} \in \mathcal{K}} \sum_{k=1}^{t} \ell_k(\boldsymbol{x})$$

This idea of strategy is known as Follow the Leader (FTL). Unfortunately, this strategy fails in worst-case, namely, the regret can be linear in the number of iterations, as the following example shows. Consider $\mathcal{K} = [-1, 1]$, let $\ell_1(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}$, and let ℓ_k for $k = 2, \ldots, T$ alternate between $-\boldsymbol{x}$ or $\boldsymbol{x}$. In this specific setting the FTL strategy will keep shifting between $\boldsymbol{x}_t = -1$ and $\boldsymbol{x}_t = 1$, always making the wrong choice. Follow the Leader strategy fails simply because it is unstable. In order to solve this issue we need to stabilize the method, that is, inserting a regularization term in the update.

2.2.3. Follow the Regularized Leader

In order to deal with the FTRL family of algorithms, which aims to stabilize follow the leader update, we need to introduce a mathematical concept that will make the understanding of the next sections easier.
We consider regularization functions, denoted $F : \mathcal{K} \to \mathbb{R}$, which are strongly convex and smooth. In particular:

Definition 2.2.3.1. *A differentiable function F is σ-strongly convex with $\sigma > 0$ if $\forall \boldsymbol{x}, \boldsymbol{y}$ belonging to its domain:*

$$F(\boldsymbol{y}) \geq F(\boldsymbol{x}) + \nabla F(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{\sigma}{2} ||\boldsymbol{y} - \boldsymbol{x}||^2$$

Now we are ready to introduce the main algorithm of this subsection; by adding a regu-

larization term to the previously described follow the leader, we obtain the FRTL (Follow the Regularized Leader) family of algorithms. FTRL is defined in Algorithm 2.1. The regularization function F is, as previously highlighted, assumed to be σ-strongly convex, smooth, and twice differentiable.

Algorithm 2.1 Follow the Regularized leader (FTRL)

1: Input: $\mu > 0$, regularization function F, and a convex compact set $\mathcal{K}$.
2: Let $\boldsymbol{x}_1 = \arg\min_{\boldsymbol{x}\in\mathcal{K}} F(\boldsymbol{x})$.
3: **for** $t = 1$ *to* T **do**
4: Predict $\boldsymbol{x}_t$
5: Observe the function ℓ_t.
6: update

$$\boldsymbol{x}_{t+1} := \text{argmin}_{\boldsymbol{x}\in\mathcal{K}}\, \boldsymbol{x}^\top \left(\sum_{k=1}^{t} \nabla \ell_k(\boldsymbol{x}_k) \right) + \frac{1}{\mu} F(\boldsymbol{x})$$

7: **end for**

We finally report the Regret result of FTRL:

Theorem 2.3. *(Hazan, 2019) The FTRL Algorithm attains for every $\boldsymbol{u} \in \mathcal{K}$ the following bound on the regret:*

$$R_T \leq 2\mu \sum_{t=1}^{T} ||\nabla \ell_t(\boldsymbol{x}_t)||_{*_t}^2 + \frac{F(\boldsymbol{u}) - F(\boldsymbol{x}_1)}{\mu}$$

If an upper bound on the local norms is known, i.e.$||\nabla \ell_t(\boldsymbol{x}_t)||_{*_t} \leq G_R$ for all times t, then we can further optimize over the choice of the learning rate μ to obtain:

$$R_T \leq 2 D_R G_R \sqrt{2T}$$

with D_R diameter of set $\mathcal{K}$.

2.2.4. Multiplicative Weight Update

As pointed out in the previous subsection, FTRL is a family of algorithms; based on the regularization function that is chosen, this family generates a huge variety of No-Regret algorithms.
In this subsection we show the derivation of Multiplicative Weight Update (and thus of its linear version). First of all, we choose the proper regularization function: to obtain

MWU we choose the negative entropy that is:

$$F(\boldsymbol{x}) := \sum_{i=1}^{n} \boldsymbol{x}(i) \ln \boldsymbol{x}(i)$$

With this choice of regularizer, we have:

$$\boldsymbol{x}_{t+1} := \text{argmin}_{\boldsymbol{x}\in\mathcal{K}}\, \boldsymbol{x}^\top \left(\sum_{k=1}^{t} \nabla \ell_k(\boldsymbol{x}_k) \right) + \frac{1}{\mu} \sum_{i=1}^{n} \boldsymbol{x}(i) \ln \boldsymbol{x}(i)$$

then we choose $\mathcal{K} = \Delta$, that is, assuming the convex set to be a simplex, from which:

$$\boldsymbol{x}_{t+1} := \text{argmin}_{\boldsymbol{x}\in\Delta}\, \boldsymbol{x}^\top \left(\sum_{k=1}^{t} \nabla \ell_k(\boldsymbol{x}_k) \right) + \frac{1}{\mu} \sum_{i=1}^{n} \boldsymbol{x}(i) \ln \boldsymbol{x}(i)$$

To compute the minimum of the above function we will use the method of Lagrange multipliers. We introduce a new parameter η and define the function:

$$g_\eta(\boldsymbol{x}) := g(\boldsymbol{x}) + \eta \left(a^\top \boldsymbol{x} - b \right)$$

with $g(\boldsymbol{x})$ objective of the optimization problem. Then, if $\boldsymbol{x}^*$ is a feasible minimizer of $g(\boldsymbol{x})$, then there is at least a value of η such that $\nabla g_\eta(\boldsymbol{x}^*) = 0$. Thus, it is possible to find all $\boldsymbol{x}, \eta$ such that $\nabla g_\eta(\boldsymbol{x}) = 0$, to remove the values of $\boldsymbol{x}$ such that $a^\top \boldsymbol{x} \neq b$, and finally to look at which of the remaining $\boldsymbol{x}$ minimizes $g(\boldsymbol{x})$. The constraint $\boldsymbol{x} \in \Delta$ can be rewritten as $\sum_{i=1}^{n} \boldsymbol{x}(i) = 1$, from which we consider the function:

$$\boldsymbol{x}^\top \left(\sum_{k=1}^{t} \nabla \ell_k(\boldsymbol{x}_k) \right) + \frac{1}{\mu} \sum_{i=1}^{n} \boldsymbol{x}(i) \ln \boldsymbol{x}(i) + \eta \left(\boldsymbol{x}^\top \mathbf{1} - 1 \right)$$

we compute the partial derivative of the previous expression with respect to $\boldsymbol{x}_i$:

$$\left(\sum_{k=1}^{t} \nabla \ell_k(\boldsymbol{x}_k)(i) \right) + \frac{1}{\mu} \left(1 + \ln \boldsymbol{x}(i) \right) + \eta$$

If we want the gradient to be zero we obtain:

$$\boldsymbol{x}(i) = e^{-1-\eta\mu-\mu\left(\sum_{k=1}^{t} \nabla \ell_k(\boldsymbol{x}_k)(i)\right)}$$

There is only one value of η that makes the solution a probability distribution which

corresponds to the result:

$$\boldsymbol{x}_{t+1}(i) = \frac{e^{-\mu \sum_{k=1}^{t} \nabla \ell_k(\boldsymbol{x}_k)(i)}}{\sum_{j=1}^{n} e^{-\mu \sum_{k=1}^{t} \nabla \ell_k(\boldsymbol{x}_k)(j)}}$$

from which we obtain the MWU algorithm 2.2

Algorithm 2.2 Multiplicative Weight Update (MWU)

1: Input: $\mu > 0$.
2: Let $\boldsymbol{x}_1 = [1/n, \ldots, 1/n]^\top$
3: **for** $t = 1$ *to* T **do**
4: Predict $\boldsymbol{x}_t$
5: Observe the function ℓ_t.
6: update

$$\boldsymbol{x}_{t+1}(i) := \boldsymbol{x}_t(i) \frac{e^{-\mu \nabla \ell_t(\boldsymbol{x}_t)(i)}}{\sum_{j=1}^{n} \boldsymbol{x}_t(j) e^{-\mu \nabla \ell_t(\boldsymbol{x}_t)(j)}}$$

7: **end for**

Please note that even if from a computational perspective, this algorithm appears much more efficient than the general update of FTRL, MWU attains similar guarantees in terms of Regret when a proper learning rate μ is chosen. For completeness we report the linear version of algorithm 2.2

Algorithm 2.3 Linear Multiplicative Weight Update (LMWU)

1: Input: $\mu > 0$.
2: Let $\boldsymbol{x}_1 = [1/n, \ldots, 1/n]^\top$
3: **for** $t = 1$ *to* T **do**
4: Predict $\boldsymbol{x}_t$
5: Observe the function ℓ_t.
6: update

$$\boldsymbol{x}_{t+1}(i) := \boldsymbol{x}_t(i) \frac{1 - \mu \nabla \ell_t(\boldsymbol{x}_t)(i)}{\sum_{j=1}^{n} \boldsymbol{x}_t(j) \left(1 - \mu \nabla \ell_t(\boldsymbol{x}_t)(j)\right)}$$

7: **end for**

2.2.5. Online Mirror Descent

Online Mirror Descent (OMD) is an iterative family of algorithms that computes the current decision using a simple gradient update rule and the previous decision. The

generality of the method stems from the update being carried out in a dual space, where the duality is defined by the choice of regularization: the gradient of the regularization function defines a mapping from $\mathbb{R}^n$ onto itself, which is a vector field. The gradient updates are then carried out in this vector field. In OMD, regularization transforms the space in which gradient updates are performed.

Before diving into the details of the algorithm, we introduce a mathematical definition.

Definition 2.2.5.1. *(Bregman Divergence) Denote by $B_F(\boldsymbol{x}||\boldsymbol{y})$ the Bregman divergence with respect to the function F (defined in subesction 2.2.3), defined as $B_F(\boldsymbol{x}||\boldsymbol{y}) = F(\boldsymbol{x}) - F(\boldsymbol{y}) - \nabla F(\boldsymbol{y})^\top(\boldsymbol{x} - \boldsymbol{y})$*

We now point out some property of the Bregman divergence:

1. Strict convexity in the first argument $\boldsymbol{x}$.
2. Non negativity: $B_F(\boldsymbol{x}||\boldsymbol{y}) \geq 0 \quad \forall \boldsymbol{x}, \boldsymbol{y}$.
3. $B_F(\boldsymbol{x}||\boldsymbol{y}) = 0$ if and only if $\boldsymbol{x} = \boldsymbol{y}$.
4. Asymmetry
5. If F σ-strongly convex: $B_F(\boldsymbol{x}||\boldsymbol{y}) \geq \frac{\sigma}{2}||\boldsymbol{x} - \boldsymbol{y}||^2$

There exists two versions of OMD algorithm: an agile and a lazy one. We report only the lazy version as it is the one that we will use in the thesis.

Algorithm 2.4 Online Mirror Descent (OMD)

1: Input: $\mu > 0$, regularization function $F(\boldsymbol{x})$, and a convex compact set $\mathcal{K}$.
2: Let $\boldsymbol{z}_1$ be such that $\nabla(\boldsymbol{z}_1) = 0$ and $\boldsymbol{x}_1 = \arg\min_{\boldsymbol{x}\in\mathcal{K}} B_F(\boldsymbol{x}||\boldsymbol{z}_1)$.
3: **for** $t = 1$ *to* T **do**
4: Predict $\boldsymbol{x}_t$
5: Observe the function ℓ_t.
6: update $\boldsymbol{z}_t$ according to the rule:

$$\nabla F(\boldsymbol{z}_{t+1}) = \nabla F(\boldsymbol{z}_t) - \mu \nabla \ell_t(\boldsymbol{x}_t)$$

7: Project according to B_F:

$$\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x}\in\mathcal{K}} B_F(\boldsymbol{x}||\boldsymbol{z}_{t+1})$$

8: **end for**

In order to compute the Regret of OMD is sufficient to show that for linear cost functions, the algorithm is equivalent to FTRL (algorithm 2.1).

Lemma 2.1. *(Equivalence OMD and FTRL (Hazan, 2019)) Let $\ell_1, \dots, \ell_t$ be linear cost functions. The lazy OMD and FTRL algorithms produce identical predictions.*

Proof. First, observe that the unconstrained minimum

$$\boldsymbol{x'} := \text{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} \boldsymbol{x}^\top \left(\sum_{k=1}^{t-1} \nabla \ell_k(\boldsymbol{x}_k) \right) + \frac{1}{\mu} F(\boldsymbol{x})$$

satisfies:

$$\nabla F(\boldsymbol{x'}) = -\mu \sum_{k=1}^{t-1} \nabla \ell_k(\boldsymbol{x}_k)$$

By definition, $\boldsymbol{z}_t$ also satisfies the above equation, but since $F(\boldsymbol{x})$ is strictly convex, there is only one solution for the above equation and thus $\boldsymbol{z}_t = \boldsymbol{x}'_t$. Hence,

$$\begin{aligned} B_F(\boldsymbol{x}||\boldsymbol{z}_t) &= F(\boldsymbol{x}) - F(\boldsymbol{z}_t) - \nabla F(\boldsymbol{z}_t)^\top (\boldsymbol{x} - \boldsymbol{z}_t) \\ &= F(\boldsymbol{x}) - F(\boldsymbol{z}_t) + \mu \sum_{k=1}^{t-1} \nabla \ell_k(\boldsymbol{x}_k)^\top (\boldsymbol{x} - \boldsymbol{z}_t) \end{aligned}$$

Since $F(\boldsymbol{z}_t)$ and $\sum_{k=1}^{t-1} \nabla \ell_k(\boldsymbol{x}_k)^\top \boldsymbol{z}_t$ are independent of $\boldsymbol{x}$, it follows that the Bregman is minimized at the point $\boldsymbol{x}$ that minimizes $F(\boldsymbol{x}) + \mu \sum_{k=1}^{t-1} \nabla \ell_k(\boldsymbol{x}_k)^\top \boldsymbol{x}_t$ over $\mathcal{K}$ which, in turn, concludes the proof. □

2.2.6. Learning in Games

The aim of this subsection is to translate the results obtained in previous sections in a game theoretic setting.

During the thesis we will deal with two zero-sum players as specified in section 2.1; the first player, also called row player, will be the minimizer with actions space $\mathcal{K} = \Delta_n$, where Δ_n represents the simplex built on the n actions, while the second player, the column one, will be the maximizer with actions space $\mathcal{K} = \Delta_m$. As previously specified, we will refer to the strategy of the minimizer at time t with the vector $\boldsymbol{x}_t$ while we will use $\boldsymbol{y}_t$ for the maximizer. Thus, we report all the algorithm seen in section 2.2, adapted to a game theoretic setting, so that during the core of the thesis the reader will be facilitated in the comprehension.

We start by the FTRL family:

Algorithm 2.5 Follow the Regularized leader (FTRL) on zero-sum games

1: Input: $\mu > 0$, regularization function F, and a simplex Δ_n.
2: Let $\boldsymbol{x}_1 = \arg\min_{\boldsymbol{x}\in\Delta_n} F(\boldsymbol{x})$.
3: **for** $t = 1$ *to* T **do**
4: Predict $\boldsymbol{x}_t$
5: update

$$\boldsymbol{x}_{t+1} := \operatorname{argmin}_{\boldsymbol{x}\in\Delta_n} \boldsymbol{x}^\top \left(\sum_{k=1}^{t} \boldsymbol{U}\boldsymbol{y}_k \right) + \frac{1}{\mu} F(\boldsymbol{x})$$

6: **end for**

then we proceed with the famous multiplicative weight update:

Algorithm 2.6 Multiplicative Weight Update (MWU) on zero-sum games

1: Input: $\mu > 0$.
2: Let $\boldsymbol{x}_1 = [1/n, \ldots, 1/n]^\top$
3: **for** $t = 1$ *to* T **do**
4: Predict $\boldsymbol{x}_t$
5: update

$$\boldsymbol{x}_{t+1}(i) := \boldsymbol{x}_t(i) \frac{e^{-\mu \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_t}}{\sum_{j=1}^{n} \boldsymbol{x}_t(j) e^{-\mu \boldsymbol{e}_j^\top \boldsymbol{U}\boldsymbol{y}_t}}$$

6: **end for**

and its linear version:

Algorithm 2.7 Linear Multiplicative Weight Update (LMWU) on zero-sum games

1: Input: $\mu > 0$.
2: Let $\boldsymbol{x}_1 = [1/n, \ldots, 1/n]^\top$
3: **for** $t = 1$ *to* T **do**
4: Predict $\boldsymbol{x}_t$
5: update

$$\boldsymbol{x}_{t+1}(i) := \boldsymbol{x}_t(i) \frac{1 - \mu \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_t}{\sum_{j=1}^{n} \boldsymbol{x}_t(j) \left(1 - \mu \boldsymbol{e}_j^\top \boldsymbol{U}\boldsymbol{y}_t\right)}$$

6: **end for**

In order to deal with MWU convergence we introduce the notion of KL divergence:

Definition 2.2.6.1. *(KL divergence (Kullback and Leibler, 1951)) The relative entropy or KL divergence between two vectors* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$ *in* Δ_n *is defined as* $KL(\boldsymbol{x}_1||\boldsymbol{x}_2) = \sum_{i=1}^{n} \boldsymbol{x}_1(i) \ln \frac{\boldsymbol{x}_1}{\boldsymbol{x}_2}$.

The Kullback-Leibler divergence is always non-negative. Furthermore $KL(\boldsymbol{x}_1||\boldsymbol{x}_2) = 0$ if and only if $\boldsymbol{x}_1 = \boldsymbol{x}_2$ almost everywhere.

Finally we present OMD:

Algorithm 2.8 Online Mirror Descent (OMD) on zero-sum games

1: Input: $\mu > 0$, regularization function $F(\boldsymbol{x})$, and the simplex Δ_n.
2: Let $\boldsymbol{z}_1$ be such that $\nabla(\boldsymbol{z}_1) = 0$ and $\boldsymbol{x}_1 = \arg\min_{\boldsymbol{x}\in\mathcal{K}} B_F(\boldsymbol{x}||\boldsymbol{z}_1)$.
3: **for** $t = 1$ *to* T **do**
4: Predict $\boldsymbol{x}_t$
5: update $\boldsymbol{z}_t$ according to the rule:

$$\nabla F(\boldsymbol{z}_{t+1}) = \nabla F(\boldsymbol{z}_t) - \mu \boldsymbol{U} \boldsymbol{y}_t$$

6: Project according to B_F:

$$\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x}\in\Delta_n} B_F(\boldsymbol{x}||\boldsymbol{z}_{t+1})$$

7: **end for**

We conclude with an important property of the algorithms presented so far:

Definition 2.2.6.2. *(Stability) A no-regret algorithm is stable if* $\forall t : \boldsymbol{y}_t = \boldsymbol{y}^* \implies \boldsymbol{x}_{t+1} = \boldsymbol{x}_t$ *when there exists a fully-mixed equilibrium strategy* $\boldsymbol{x}^*$.

3 | Related Works

In this chapter we introduce the two main properties that our algorithms guarantee: Safety and Last Round Convergence. In particular we show some of the main results obtained by the scientific community so far: part of these results will be used to build our final algorithms, others represent the state of the art in different settings with respect to the one on which our algorithms work.

3.1. Safety

The aim of this section is to explain what the Safety constraint is, why it is important and to give an example of how an online algorithm can guarantee this property.

Definition 3.1.0.1. *(Safety) Given two bounds ξ_1 and ξ_2 with $\xi_1 < \xi_2$, an Online algorithm applied to games guarantees safety if and only if $u(t) \in [\xi_1, \xi_2] \quad \forall t$, with $u(t)$ utility of the opponent at time t.*

As specified in the abstract, the safety property is fundamental in order to keep the opponent engaged, that is, our adversary will have an incentive to keep playing the game. In 2021 Bernasconi-de-Luca et al. developed an algorithm which guarantees the property. This paper mainly focuses on Extensive Form Games, but some of the techniques employed will be useful for the Normal Form setting (on which the thesis focuses).
Thus we report some theorems on which their algorithm relies:

Lemma 3.1. *(Lemma 3 (Devroye, 1983)) Let $\boldsymbol{x} \in \Delta_n$ and $i^1, \dots, i^t \in \{1, \dots, n\}$ be t indices of actions sampled independently according to $\boldsymbol{x}$. Then, for any $0 < \delta \leq 3\exp(-4n/5)$, it holds:*

$$\mathbb{P}\left(\sum_{i=1}^{n} |\bar{\boldsymbol{x}}_t(i) - \boldsymbol{x}(i)| \leq 5\sqrt{\frac{\ln(3/\delta)}{t}}\right) \geq 1 - \delta$$

where $\bar{\boldsymbol{x}}_t$ is the empiric frequency of the actions played by $\boldsymbol{x}$ player.

Lemma 3.2. *(Bernasconi-de-Luca et al. 2021) Let* $\boldsymbol{x} \in \Delta_n$ *and* $i^1, \dots, i^t \in \{1, \dots, n\}$ *be* t *indices of actions sampled independently according to* $\boldsymbol{x}$*. Then, for any* $0 < \delta \leq 3\exp(-4n/5)$*, it holds:*

$$\mathbb{P}\left(\bigcap_{i=1}^{n}\left\{|\bar{\boldsymbol{x}}_t(i) - \boldsymbol{x}(i)| \leq \frac{5}{2}\sqrt{\frac{\ln(3/\delta)}{t}}\right\}\right) \geq 1 - \delta$$

The reader is probably wondering why this trick is necessary; the idea is that the algorithms specified in previous sections are considering the case in which the agent receives a full feedback at the end of each round (the complete gradient is received), but during the thesis we will deal even with the case in which our opponent does not play a strategy, but an action sampled by that specific strategy (we will refer to it as partial semi-Bandit feedback).
These two theorems will be fundamental to estimate a set in which the strategy of the opponent (when it is fixed) lies with high probability and thus, to play a strategy that will be safe for any possible opponent strategy in that specific set.
We now report the algorithm and then we give a hint of how it works:

Algorithm 3.1 COX-UCB

1: **for** $t = 1$ *to* T **do**
2: Build confidence region Y_{t-1} from history of past feedback H_{t-1}
3: Build set of linear constraints characterizing engaging set X_t by exploiting Y_{t-1}
4: Play the game according to strategy $\boldsymbol{x}_t := \operatorname{argmax}_{\boldsymbol{x} \in \tilde{X}_t} \max_{\boldsymbol{y} \in Y_{t-1}} \boldsymbol{x}^\top \boldsymbol{U} \boldsymbol{y}$
5: update history H_t
6: **end for**

We underline few aspects of this algorithm that will be important for the comprehension of the thesis: the region Y_{t-1}, on which the opponent lies with high probability, is estimated using the Devroye Formula (Lemma 3.1), with a little modification to take into account the extensive form structure of the game. The safe set X_t is built by linear constraints such that all the strategies in the set are safe (with respect to the safe bounds $[\xi_1, \xi_2]$) given any strategy of the opponent in Y_{t-1}. The set $\tilde{X}_t$ is built starting from X_t in order to guarantee proper exploration. Finally, the strategy $\boldsymbol{x}_t$ is chosen optimistically given the sets $\tilde{X}_t$ and Y_{t-1}

3.2. Last Round Convergence

This section aims to highlight some of the most important results obtained so far in terms of convergence. We start by the well known result of average convergence of No-Regret learners and we conclude with the algorithm which is the actual starting point of the thesis.

3.2.1. Hannan Consistency and Average Convergence

We start by the definition of Hannan Consistent strategy:

Definition 3.2.1.1. *(Hannan Consistency) A strategy is said to be Hannan consistent if:*

$$\limsup_{T\to\infty} \frac{1}{T}\left(\sum_{t=1}^{T} \boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{y}_t - \min_{\boldsymbol{x}\in\Delta_n} \sum_{t=1}^{T} \boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{y}_t\right) = 0$$

This definition is fundamental to achieve following results in zero-sum repeated normal form games:

Theorem 3.1. *(Cesa-Bianchi and Lugosi, 2006) Assume that in a two-person zero-sum game the row player plays according to a Hannan-consistent strategy. Then:*

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{y}_t \leq v \quad whp$$

with v minmax value of the game.

The theorem shows that, regardless of what the opponent plays, if the row player plays according to a Hannan-consistent strategy, then his cumulative loss is guaranteed to be asymptotically not more than the value v of the game. An important corollary follows:

Corollary 3.1. *(Cesa-Bianchi and Lugosi, 2006) Assume that in a two-person zero-sum game, both players play according to some Hannan consistent strategy. Then:*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{y}_t = v \quad whp$$

Finally we can obtain the result on convergence:

Theorem 3.2. *(Average Convergence to Equilibria (Cesa-Bianchi and Lugosi, 2006)) If*

both players follow some Hannan consistent strategy, then it is also easy to see that the product distribution $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$*, formed by the (marginal) empirical distributions of play:*

$$\bar{\boldsymbol{x}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t \quad \textit{and} \quad \bar{\boldsymbol{y}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{y}_t$$

of the two players converges, almost surely, to the set of Nash equilibria $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ *of the game.*

The last theorem simply states that if two players (a min and a max one) play a zero-sum game repeatedly, following Hannan consistent procedures, their average strategy will converge to the minmax equilibrium of the game. It is easy to understand that even if it is a strong result, in many real world applications the instability of the strategy played round after round is a tremendous drawback. For example, considering the Market as a game and a company as the player, changing the (mixed) strategy will increase the cost of operation to implement the new mixed strategy (e.g., as a result of having to hire new equipment and employees). Therefore, the company would aim to maximise the revenue (namely, the average payoff) and reduce the cost of operation by having a stable strategy. The property of stability is given by the Last Round Convergence, which is a far stronger concept with respect to the average convergence.

3.2.2. Last Round Convergence in Self-Play

First, we give the definition of the property:

Definition 3.2.2.1. *(Last Round Convergence) A sequence of strategies* $\boldsymbol{x}_t$ *is convergent in last round if and only if:*

$$\lim_{t\to\infty} \boldsymbol{x}_t = \boldsymbol{x}^*$$

with $\boldsymbol{x}^*$ *equilibrium strategy of the player.*

In recent years much work has been done to build algorithms that can guarantee this property; mainly, they focus on a self-play setting. The idea behind self-play is that an agent plays against himself in order to learn a good policy. In two-player zero-sum games it means reaching an equilibrium. Formally, given a payoff matrix of a zero-sum game $\boldsymbol{U}$ we want to find the bilinear saddle point $(\boldsymbol{x}^*, \boldsymbol{y}^*) = \arg\min_{\boldsymbol{x}\in\Delta_n} \arg\max_{\boldsymbol{y}\in\Delta_m} \boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{y}$; thus we make the row player and column one employ the same algorithm and if Last Round Convergence is satisfied both players will converge to the optimal strategy (note that the first player will receive a loss at each round while the second player will receive a reward).

Before reporting the most famous algorithms that solve the convergence issue, we want to underline that these algorithms will be slight variations of those presented in section 2.2. The general idea to guarantee convergence in self-play is to try to predict the opponent's next action (strategy); the adversary's next move is generally predicted using the previous round gradient. That is the reason why these algorithms are called 'Optimistic' version of the standard OCO procedures.
We start with Optimistic Mirror Descent:

Algorithm 3.2 Optimistic Online Mirror Descent (OOMD) on zero-sum games

1: Input: $\mu > 0$.
2: **for** $t = 1$ *to* T **do**
3: Predict $\boldsymbol{x}_t$
4: update $\boldsymbol{x}_{t+1}$ according to the rule:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - 2\mu \boldsymbol{U}\boldsymbol{y}_t + \mu \boldsymbol{U}\boldsymbol{y}_{t-1}$$

5: **end for**

This algorithm has been shown to exhibit last round convergence in the unconstrained case [Daskalakis et al. 9], that is, strategies are not constrained to be in the simplex.
Few years later the optimistic version of MWU has shown to guarantee the property even in the constrained case [Daskalakis and Panageas 8]:

Algorithm 3.3 Optimistic Multiplicative Weight Update (OMWU) on zero-sum games

1: Input: $\mu > 0$.
2: Let $\boldsymbol{x}_1 = [1/n, \dots, 1/n]^\top$
3: **for** $t = 1$ *to* T **do**
4: Predict $\boldsymbol{x}_t$
5: update

$$\boldsymbol{x}_{t+1}(i) := \boldsymbol{x}_t(i) \frac{e^{-2\mu \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_t + \mu \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_{t-1}}}{\sum_{j=1}^{n} \boldsymbol{x}_t(j) e^{-2\mu \boldsymbol{e}_j^\top \boldsymbol{U}\boldsymbol{y}_t + \mu \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_{t-1}}}$$

6: **end for**

Finally, we underline that steps ahead in the convergence have been done even with a bandit feedback (see for example [Lin et al. 19]) and in Extensive Form Games (e.g. see [Lee et al. 18]).

3.2.3. Last Round Convergence in Asymmetric Setting

Now we can finally introduce the setting in which our algorithm will work on.
Please notice that as specified in the abstract and in section 1, the aim of our thesis is to teach a human-like learner, that is why the idea of learning an equilibrium in self-play does not make any sense. Thus, we consider a setting in which the column player (teacher) has full knowledge of the payoff matrix, while the opponent (human) can employ a family of No-Regret algorithms.
The idea of this asymmetric setting has been taken into account by Dinh et al. which proposed an algorithm capable to achieve last round convergence and sublinear dynamic regret against the entire FTRL family when there exists a fully-mixed equilibrium strategy for the row player (algorithm 2.5).
We report here the pseudocode of LRCA:

Algorithm 3.4 Last Round Convergence in Asymmetric algorithm (LRCA)

for $t = 1$ to T **do**
  if $t = 2k-1,\ k \in \mathbb{N}$ **then**
    $\boldsymbol{y}_t = \boldsymbol{y}^*$
  end if
  if $t = 2k,\ k \in \mathbb{N}$ **then**
    $\boldsymbol{e}_t := \text{argmax}_{e \in \{\boldsymbol{e}_1, \boldsymbol{e}_2, \dots \boldsymbol{e}_m\}} \boldsymbol{x}_{t-1}^\top \boldsymbol{U} \boldsymbol{e}; \quad f(\boldsymbol{x}_{t-1}) := \max_{\boldsymbol{y} \in \Delta_m} \boldsymbol{x}_{t-1}^\top \boldsymbol{U} \boldsymbol{y}$
    $\alpha_t := \dfrac{f(\boldsymbol{x}_{t-1}) - v}{\beta}$
    $\boldsymbol{y}_t := (1 - \alpha_t)\, \boldsymbol{y}^* + \alpha_t \boldsymbol{e}_t$
  end if
end for

To conclude we show the dynamic regret obtained by this algorithm:

Theorem 3.3. *(Dynamic Regret (Dinh et al., 2021)) Assume that the row player follows the above-mentioned no-regret type algorithms: MWU, LMWU, FTRL. If there exists a fully-mixed minmax strategy for the row player, then by following LRCA, the column player will achieve the no-dynamic regret property with the dynamic regret satisfying* $R_T \leq DR_T = \mathcal{O}\left(\sqrt{\log(n)} T^{3/4}\right)$. *Furthermore, in the case the row player uses a constant learning rate* μ, *we have* $DR_T = \mathcal{O}\left(\frac{n}{\sqrt{\mu}} T^{1/2}\right)$.

4 | Safe Guide with Expert Feedback

In this chapter we present the algorithm we developed for the expert feedback case: E-LRCA (4.1).

4.1. Assumptions and Setting

In this section we provide the general assumptions/setting and most relevant considerations which will help the reader to understand the results and the proofs of the rest of the chapter.

As specified in the introduction we are in a zero-sum repeated game (players play the same game for each round); the entries of the payoff matrix (which is positive for the column player) are scaled in $[0, 1]$ without loss of generality. Both players have the so called expert feedback, that is, the complete gradient is received by every player at the end of the round. To be precise, row player will receive $-\boldsymbol{U}\boldsymbol{y}_t$ after having played $\boldsymbol{x}_t$ while column player will receive $\boldsymbol{x}_t^\top \boldsymbol{U}$ after having played $\boldsymbol{y}_t$. The payoff matrix $\boldsymbol{U}$ is known by the column player (the teacher), that is, he perfectly knows the equilibrium (this is called Asymmetric information), while row player (the learner/human) employs an algorithm of the OMD family (which for linear losses, as in our setting, is equivalent to FTRL).

We present our algorithm for the column player which will guarantee:

1. *Safety* (see definition 3.1.0.1): this property will be guaranteed in different ways depending on the equilibrium the game has. In case of fully-mixed equilibrium strategy for the row player it will be possible to predict the opponent next strategy (see sections 4.2) so that safety can be obtained in an efficient manner; in case of not fully-mixed equilibrium, safety must be guaranteed for any possible strategy of the opponent, decelerating the teaching/learning procedure. Moreover, in the latter case, an assumption on the upper bound value ξ_2 will be necessary.

2. *Last Round Convergence* (see definition 3.2.2.1): results are equivalent for games with any kind of equilibrium.

3. *Sublinear Dynamic Regret* (see definition 2.2.1.2): in the case of not fully-mixed equilibrium strategy for the row player, the regret will be worse in terms of constants and it will not be possible to express it without exploiting the dynamic of the opponent's learning rate.

4.2. Algorithm

Algorithm 4.1 Engaged - Last Round Convergence in Asymmetric algorithm (E-LRCA)

for $t = 1$ to T **do**
  if $t = 2k - 1, k \in \mathbb{N}$ **then**
    $\boldsymbol{y}_t = \boldsymbol{y}^*$
  end if
  if $t = 2k, k \in \mathbb{N}$ **then**
    $\boldsymbol{e}_{t-1} := \text{argmax}_{e \in \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots \boldsymbol{e}_m\}} \boldsymbol{x}_{t-1}^\top \boldsymbol{U} \boldsymbol{e}; \quad f(\boldsymbol{x}_{t-1}) := \max_{\boldsymbol{y} \in \Delta_m} \boldsymbol{x}_{t-1}^\top \boldsymbol{U} \boldsymbol{y}$
    if game has a fully-mixed equilibrium **then**
      $\alpha_{new} = \dfrac{\xi_2 - v}{\boldsymbol{x}_{t-1}^\top \boldsymbol{U} \boldsymbol{e}_{t-1} - v}$
    end if
    if game has not a fully-mixed equilibrium **then**
      $\alpha_{new} = \min \left(\dfrac{\xi_2 - ||\boldsymbol{U}\boldsymbol{y}^*||_\infty}{||\boldsymbol{U}||_{max} - v}, \dfrac{\xi_1 - v}{||\boldsymbol{U}||_{min} - ||\boldsymbol{U}\boldsymbol{y}^*||_\infty} \right)$
    end if
    $\alpha_t := \min \left(\alpha_{new}, \dfrac{f(\boldsymbol{x}_{t-1}) - v}{\beta} \right)$
    $\boldsymbol{y}_t := (1 - \alpha_t) \boldsymbol{y}^* + \alpha_t \boldsymbol{e}_{t-1}$
  end if
end for

We underline the main ideas behind algorithm 4.1.

In odd rounds column player plays the equilibrium so that if the row player's equilibrium strategy is fully-mixed and the opponent algorithm is stable, it is possible to predict his next strategy (it will be the same as in the previous round). If row player's equilibrium strategy is not fully-mixed, playing the equilibrium will push the opponent to the support of the equilibrium, not invalidating final result of Last Round Convergence.

In even rounds column player computes the best response ($\boldsymbol{e}_{t-1}$) and the value of the best

response $f(\boldsymbol{x}_{t-1})$ at the previous round (note that if the equilibrium is fully-mixed we have $\boldsymbol{e}_{t-1} = \boldsymbol{e}_t$ and $f(\boldsymbol{x}_{t-1}) = f(\boldsymbol{x}_t)$). Then, column player plays a convex combination between the equilibrium and the best response of the previous round, built using a parameter α_t, which must be dependant on the distance between the opponent strategy and the equilibrium ($\alpha_t = \frac{f(\boldsymbol{x}_{t-1})-v}{\beta}$); in case this parameter would lead to an utility outside the safety bounds (checked by the min operator), we scale $\frac{f(\boldsymbol{x}_{t-1})-v}{\beta}$ by a factor $\gamma_t \in (0, 1]$ obtaining α_{new} (the multiplication $\gamma_t \frac{f(\boldsymbol{x}_{t-1})-v}{\beta}$ is implicit in the algorithm but shown in section 4.3).

To conclude, it is important to underline that the scaling factor γ_t depends on the equilibrium the game has; in case there exists a fully-mixed equilibrium, we find a γ_t such that the next round utility will be exactly the upper bound ξ_2, otherwise we need a γ_t that is safe for every strategy of the opponent (the smallest possible), which will lead to a deceleration of the teaching dynamic.

4.3. Safety

In this section we provide the two theorems related to safety property of algorithm 4.1. We start with the result in games where there exists a fully-mixed equilibrium strategy for the row player (theorem 4.1) and we conclude with the case in which the equilibrium is not fully-mixed (theorem 4.2). The main difference between the two statements is a constraint on the upper bound of the safety region ξ_2 in the second theorem (namely, $||\boldsymbol{U}\boldsymbol{y}^*||_\infty < \xi_2$).

Theorem 4.1. *Assume that the row player is following a no-regret stable learning algorithm, given two bounds ξ_1, ξ_2 on the Utility such that $v \in [\xi_1, \xi_2)$, if there exists a fully-mixed minmax equilibrium strategy (for the row player) and the column player follows E-LRCA (algorithm 4.1), the Utility of the column Player will be bounded in $[\xi_1, \xi_2]$ at each round (from which follows that the Utility of the row Player will be bounded in $[-\xi_2, -\xi_1]$ at each round).*

Proof. The proof is divided in two parts: one for the odd rounds, the other for the even ones.
In the case of odd rounds, the column player will choose to play $\boldsymbol{y}^*$, which for assumption of fully-mixed equilibrium means $\boldsymbol{U}\boldsymbol{y}^* = [v, ..., v]$. Thus, we have that for any strategy the row player could choose, the utility would be v, that is inside the safety bounds.
As concerns the even rounds, given that the no-regret algorithm of the row player is stable, we can predict his next strategy (it will be the same of the previous/odd round) so, if

$\alpha_t = \frac{f(\boldsymbol{x}_{t-1})-v}{\beta}$ will lead to an utility outside the bounds we use $\alpha_{new} = \gamma_t \frac{f(\boldsymbol{x}_{t-1})-v}{\beta}$ with $\gamma_t \in (0, 1]$ computed as follows:

$$\boldsymbol{x}_{t-1}^\top \boldsymbol{U} \left(\left(1 - \gamma_t \frac{f\left(\boldsymbol{x}_{t-1}\right) - v}{\beta} \right) \boldsymbol{y}^* + \gamma_t \frac{f\left(\boldsymbol{x}_{t-1}\right) - v}{\beta} \boldsymbol{e}_t \right) = \xi_2 \tag{4.1}$$

$$\gamma_t \frac{f\left(\boldsymbol{x}_{t-1}\right) - v}{\beta} \left(-\boldsymbol{x}_{t-1}^\top \boldsymbol{U} \boldsymbol{y}^* + \boldsymbol{x}_{t-1}^\top \boldsymbol{U} \boldsymbol{e}_t \right) = \xi_2 - \boldsymbol{x}_{t-1}^\top \boldsymbol{U} \boldsymbol{y}^*$$

$$\gamma_t = \frac{\left(\xi_2 - \boldsymbol{x}_{t-1}^\top \boldsymbol{U} \boldsymbol{y}^*\right) \beta}{\left(-\boldsymbol{x}_{t-1}^\top \boldsymbol{U} \boldsymbol{y}^* + \boldsymbol{x}_{t-1}^\top \boldsymbol{U} \boldsymbol{e}_t \right) \left(f(\boldsymbol{x}_{t-1}) - v \right)}$$

$$\gamma_t = \frac{(\xi_2 - v)\beta}{(f(\boldsymbol{x}_{t-1}) - v)^2}$$

Multiplying γ_t for the standard α_t we obtain the update in algorithm 4.1, which, given $\boldsymbol{x}_{t-1} = \boldsymbol{x}_t$ for odd $t-1$, will guarantee an utility equal ξ_2 as shown in equation 4.1.
Please note that the safety with respect to the lower bound ξ_1 is automatically guaranteed by the choice of playing a combination by the best response and the equilibrium. □

Theorem 4.2. *Given two bounds ξ_1, ξ_2 on the Utility such that $v \in (\xi_1, \xi_2)$ and $||\boldsymbol{U}\boldsymbol{y}^*||_\infty < \xi_2$, if there is not a fully-mixed minmax equilibrium strategy for the row player and the column player follows E-LRCA (algorithm 4.1), the Utility of the column Player will be bounded in $[\xi_1, \xi_2]$ at each round (from which follows that the Utility of the row Player will be bounded in $[-\xi_2, -\xi_1]$ at each round).*

Proof. The proof is divided in two parts: one for the odd rounds, the other for the even ones.
In the odd rounds, safety is guaranteed by the assumption:

$$||\boldsymbol{U}\boldsymbol{y}^*||_\infty < \xi_2$$

As concerns the even rounds the choice of α_{new} is equivalent to find a γ_t which guarantees safety for every strategy of the opponent (that is the same to choose the smallest γ_t possible). Formally for ξ_2:

$$\boldsymbol{x}_t^\top \boldsymbol{U} \left(\left(1 - \gamma_t \frac{f\left(\boldsymbol{x}_t\right) - v}{\beta} \right) \boldsymbol{y}^* + \gamma_t \frac{f\left(\boldsymbol{x}_t\right) - v}{\beta} \boldsymbol{e}_{t-1} \right) = \xi_2$$

where t is an even round. From here we find:

$$\gamma_t \frac{f\left(\boldsymbol{x}_{t-1}\right) - v}{\beta} \left(-\boldsymbol{x}_t^\top \boldsymbol{U} \boldsymbol{y}^* + \boldsymbol{x}_t^\top \boldsymbol{U} \boldsymbol{e}_{t-1} \right) = \xi_2 - \boldsymbol{x}_t^\top \boldsymbol{U} \boldsymbol{y}^*$$

$$\gamma_t = \frac{\left(\xi_2 - \boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{y}^*\right)\beta}{\left(-\boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{y}^* + \boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{e}_{t-1}\right)\left(f(\boldsymbol{x}_{t-1}) - v\right)}$$

which leads to the lower bound:

$$\gamma_t \geq \frac{(\xi_2 - ||\boldsymbol{U}\boldsymbol{y}^*||_\infty)\beta}{(||\boldsymbol{U}||_{max} - v)^2}$$

instead for ξ_1:

$$\boldsymbol{x}_t^\top \boldsymbol{U}\left(\left(1 - \gamma_t \frac{f\left(\boldsymbol{x}_{t-1}\right) - v}{\beta}\right)\boldsymbol{y}^* + \gamma_t \frac{f\left(\boldsymbol{x}_{t-1}\right) - v}{\beta}\boldsymbol{e}_{t-1}\right) = \xi_1$$

$$\gamma_t \frac{f\left(\boldsymbol{x}_{t-1}\right) - v}{\beta}\left(-\boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{y}^* + \boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{e}_{t-1}\right) = \xi_1 - \boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{y}^*$$

from which:

$$\begin{aligned}\gamma_t &= \frac{\left(\xi_1 - \boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{y}^*\right)\beta}{\left(-\boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{y}^* + \boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{e}_{t-1}\right)\left(f(\boldsymbol{x}_t) - v\right)} \\ &\geq \frac{(\xi_1 - v)\beta}{(||\boldsymbol{U}||_{max} - v)(||\boldsymbol{U}||_{min} - ||\boldsymbol{U}\boldsymbol{y}^*||_\infty)}\end{aligned}$$

from that we use $\gamma_t := \min\left(\frac{(\xi_2 - ||\boldsymbol{U}\boldsymbol{y}^*||_\infty)\beta}{(||\boldsymbol{U}||_{max} - v)^2}, \frac{(\xi_1 - v)\beta}{(||\boldsymbol{U}||_{max} - v)(||\boldsymbol{U}||_{min} - ||\boldsymbol{U}\boldsymbol{y}^*||_\infty)}\right)$ which is safe both with respect to the upper bound and with respect to the lower bound given any strategy of the opponent. □

Remark 4.1. *The reader may notice that this second result is somehow stronger than the first one; indeed, γ_t found in the last theorem guarantees safety for every strategy in the opponent simplex, while γ_t of theorem 4.1 only for a specific one. Nevertheless, as it will be shown in section 4.5, the second result will have a cost in terms of regret.*

4.4. Convergence

We summarize the main steps of this section to facilitate the comprehension.

In Lemma 4.1 we prove that the γ_t used to generate the parameter ($\alpha_t = \gamma_t \frac{f(\boldsymbol{x}_{t-1}) - v}{\beta}$) of the convex combination will be always greater than 0; it is necessary to avoid a premature interruption of the teaching dynamic.

In Theorem 4.3 we show how, if there exists a fully-mixed equilibrium strategy for the row player, potentially every stable No-Regret Learner can reach any point of the simplex while playing against our algorithm.

In Lemma 4.2 and Lemma 4.3 we prove that when the equilibrium is played against MWU,

LMWU or OMD, if the strategy of the opponent will change, it can only move towards the support of the equilibrium (if the equilibrium strategy for the row player is fully-mixed, the opponent would trivially keeps playing the same strategy).
In Lemma 4.4 we lower bound the convergence step towards the equilibrium between odd rounds when opponent employs MWU, LMWU.
Finally, in Theorem 4.4 we lower bound the convergence step towards the equilibrium between odds rounds when opponent employs OMD and prove the Last Round Convergence for both MWU, LMWU and OMD (which for linear losses is equivalent for FTRL).

Lemma 4.1. *Given a two-player zero-sum game in normal form, described by the matrix* $\boldsymbol{U}$ *and two bounds* ξ_1, ξ_2 *with* $v \in (\xi_1, \xi_2)$*, it is possible to find a positive lower bound* γ_{min} *s.t. all the* γ_t *found by E-LRCA (algorithm 4.1) are greater or equal w.r.t* γ_{min}*.*

Proof. In the case of game in which the fully-mixed equilibrium is not present, γ_{min} is the one employed at each round (when the standard update $\alpha_t = \frac{f(\boldsymbol{x}_{t-1})-v}{\beta}$ is not safe) as shown in theorem 4.2; thus we have:

$$\gamma_{min} = \min\left(\frac{(\xi_2 - ||\boldsymbol{U}\boldsymbol{y}^*||_\infty)\beta}{(||\boldsymbol{U}||_{max} - v)^2}, \frac{(\xi_1 - v)\beta}{(||\boldsymbol{U}||_{max} - v)(||\boldsymbol{U}||_{min} - ||\boldsymbol{U}\boldsymbol{y}^*||_\infty)}\right) \tag{4.2}$$

which is greater than zero for $v \in (\xi_1, \xi_2)$ and $||\boldsymbol{U}\boldsymbol{y}^*||_\infty < \xi_2$ (assumption made in theorem 4.2). Note that in games with meaningful payoffs $||\boldsymbol{U}||_{max} \neq ||\boldsymbol{U}||_{min} \neq v$, nevertheless, it is not a necessary assumption for the functioning of the algorithm.

In case of of fully-mixed equilibrium we recall the solution found in theorem 4.1 from which:

$$\gamma_t = \frac{(\xi_2 - v)\beta}{(f(\boldsymbol{x}_{t-1}) - v)^2}$$

which is always greater or equal than:

$$\gamma_{min} = \frac{(\xi_2 - v)\beta}{(||\boldsymbol{U}||_{max} - v)^2} \tag{4.3}$$

that is greater than zero for $v \in [\xi_1, \xi_2)$. Again, note that in game with meaningful payoffs $||\boldsymbol{U}||_{max} \neq v$, nevertheless, it is not a necessary assumption for the functioning of the algorithm. □

Theorem 4.3. *Assume that the row player follows a stable no-regret algorithm and there exists a fully-mixed minmax equilibrium strategy for the row player. Then, by following E-LRCA with* ξ_1, ξ_2 *s.t.* $v \in [\xi_1, \xi_2)$*, for any* $\epsilon > 0$ *there exists* $t \in \mathbb{N}$ *such that* $f(\boldsymbol{x}_t) - v \leq \epsilon$*.*

Proof. We proceed in order to find a contradiction. Suppose there exists $\epsilon > 0$ such that:

$$f(\boldsymbol{x}_t) - v > \epsilon, \forall t \in \mathbb{N} \tag{4.4}$$

Recall that for Algorithm E-LRCA (algorithm 4.1) we have $\boldsymbol{y}_{2k-1} = \boldsymbol{y}^*$. Thus, following assumption 4.4:

$$\alpha_{2k} = \gamma_{2k} \frac{f(\boldsymbol{x}_{2k-1}) - v}{\beta} > \gamma_{2k} \frac{\epsilon}{\beta} \geq \gamma_{min} \frac{\epsilon}{\beta}$$

By the stability property (see definition 2.2.6.2), as $\boldsymbol{y}_{2k-1} = \boldsymbol{y}^*$, we have that $\boldsymbol{x}_{2k-1} = \boldsymbol{x}_{2k}$. Following the update rule of the algorithm and using γ_{min} as found in lemma 4.1:

$$\begin{aligned}
\boldsymbol{x}_{2k}^\top \boldsymbol{U} \boldsymbol{y}_{2k} &= \boldsymbol{x}_{2k-1}^\top \boldsymbol{U} \left((1 - \alpha_{2k}) \boldsymbol{y}^* + \alpha_{2k} \boldsymbol{e}_{2k} \right) \\
&\geq (1 - \alpha_{2k}) v + \alpha_{2k} f(\boldsymbol{x}_{2k-1}) && \text{(4.5a)} \\
&> (1 - \alpha_{2k}) v + \alpha_{2k} (v + \epsilon) && \text{(4.5b)} \\
&\geq v + \gamma_{min} \frac{\epsilon^2}{\beta}
\end{aligned}$$

Where inequality 4.5a is true by definition of Nash Equilibrium and where inequality 4.5b comes from the assumption 4.4. We then have:

$$\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t^\top \boldsymbol{U} \boldsymbol{y}_t \geq \frac{v + \left(v + \gamma_{min} \frac{\epsilon^2}{\beta} \right)}{2} = v + \gamma_{min} \frac{\epsilon^2}{2\beta}$$

We also note that, from the definition of the value of the game, we have:

$$\min_i \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t = \min_i \boldsymbol{e}_i^\top \boldsymbol{U} \frac{\sum_{t=1}^{T} \boldsymbol{y}_t}{T} \leq v$$

Thus, we have:

$$\lim_{T \to \infty} \min_i \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t - \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t^\top \boldsymbol{U} \boldsymbol{y}_t \leq v - \left(v + \gamma_{min} \frac{\epsilon^2}{2\beta} \right) = -\gamma_{min} \frac{\epsilon^2}{2\beta}$$

which contradicts the definition of a no-regret algorithm. □

Lemma 4.2. *In two-player zero-sum games where there is not a fully-mixed equilibrium strategy for the row player, assume column player plays the equilibrium and row one employs MWU or LMWU algorithm, if row player's strategy will change, it will move towards the support of the equilibrium.*

Proof. Recall the update formula of MWU (algorithm 2.6):

$$\boldsymbol{x}_{t+1}(i) = \boldsymbol{x}_t(i)\frac{e^{-\mu_t \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_t}}{Z_t}$$

Z_t is just a normalization factor, so we study what happens to the numerator when $\boldsymbol{y}^*$ is played by the column player. By property of the Nash, $\boldsymbol{e}_i\boldsymbol{U}\boldsymbol{y}^* = v \ \forall i \in supp$ while $\boldsymbol{e}_i\boldsymbol{U}\boldsymbol{y}^* > v \ \forall i \notin supp$ as shown in equation 2.4. From that we obtain:

$$\frac{e^{-\mu_t \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}^*}}{Z_t} > \frac{e^{-\mu_t \boldsymbol{e}_j^\top \boldsymbol{U}\boldsymbol{y}^*}}{Z_t} \qquad \forall i \in supp, \quad \forall j \notin supp$$

which means that if the strategy will change, it can only move towards the support of the equilibrium. The same reasoning hold for LMWU (algorithm 2.7). □

Lemma 4.3. *In two-player zero-sum games where there is not a fully-mixed equilibrium strategy for the row player, assume column player plays the equilibrium and the row one employs OMD (FTRL) algorithm with distance generating function $F(\boldsymbol{x})$ σ-strongly convex and fixed learning rate μ, if row player's strategy will change, it will move towards the support of the equilibrium.*

Proof. Let $\boldsymbol{e}$ be a strategy in the support of the equilibrium of the row player (i.e. $\boldsymbol{e} := [1, 0, \dots, 0]$). Denote by $B_F(\boldsymbol{x}_t||\boldsymbol{z}_t)$ the Bregman divergence between the current row player's strategy and lazy update of OMD as described in algorithm 2.8; then define $D_t(\boldsymbol{e}) := (B_F(\boldsymbol{e}||\boldsymbol{z}_t) - B_F(\boldsymbol{x}_t||\boldsymbol{z}_t))\frac{1}{\mu}$, following properties of strongly convex function we have:

$$\begin{aligned}
D_t(\boldsymbol{e}) &= \left(F(\boldsymbol{e}) - F(\boldsymbol{z}_t) - \nabla F(\boldsymbol{z}_t)^\top(\boldsymbol{e} - \boldsymbol{z}_t) - \left(F(\boldsymbol{x}_t) - F(\boldsymbol{z}_t) - \nabla F(\boldsymbol{z}_t)^\top(\boldsymbol{x}_t - \boldsymbol{z}_t)\right)\right)\frac{1}{\mu} \\
&= \frac{1}{\mu}F(\boldsymbol{e}) - \frac{1}{\mu}F(\boldsymbol{z}_t) + (\boldsymbol{e} - \boldsymbol{z}_t)^\top \sum_{k=1}^{t-1} \boldsymbol{U}\boldsymbol{y}_k - \left(\frac{1}{\mu}F(\boldsymbol{x}_t) - \frac{1}{\mu}F(\boldsymbol{z}_t) + (\boldsymbol{x}_t - \boldsymbol{z}_t)^\top \sum_{k=1}^{t-1} \boldsymbol{U}\boldsymbol{y}_k\right) \\
&= \frac{1}{\mu}F(\boldsymbol{e}) - \frac{1}{\mu}F(\boldsymbol{x}_t) + (\boldsymbol{e} - \boldsymbol{x}_t)^\top \sum_{k=1}^{t-1} \boldsymbol{U}\boldsymbol{y}_k \\
&\geq \frac{\sigma}{2\mu}\left\|\boldsymbol{e} - \boldsymbol{x}_t\right\|^2
\end{aligned}$$

We prove the Theorem showing that $D_t(\boldsymbol{e})$ is decreasing.
Formally:

$$D_t\left(\boldsymbol{e}\right) - D_{t+1}\left(\boldsymbol{e}\right) \geq 0 \quad \forall t \quad odd$$

From the definition of D_t we have:

$$D_t(\boldsymbol{e}) - D_{t+1}(\boldsymbol{e}) = D_t(\boldsymbol{x}_{t+1}) + \boldsymbol{x}_{t+1}^\top \boldsymbol{U}\boldsymbol{y}_t - \boldsymbol{e}^\top \boldsymbol{U}\boldsymbol{y}_t$$

recalling then $\boldsymbol{y}_t = \boldsymbol{y}^*$ we obtain:

$$D_t(\boldsymbol{e}) - D_{t+1}(\boldsymbol{e}) \geq \frac{\sigma}{2\mu}||\boldsymbol{x}_{t+1} - \boldsymbol{x}_t||^2 + \boldsymbol{x}_{t+1}^\top \boldsymbol{U}\boldsymbol{y}^* - \boldsymbol{e}^\top \boldsymbol{U}\boldsymbol{y}^*$$

given that $\boldsymbol{e}$ is in the support and recalling $\boldsymbol{x}_{t+1}^\top \boldsymbol{U}\boldsymbol{y}^* \geq v$:

$$D_t(\boldsymbol{e}) - D_{t+1}(\boldsymbol{e}) \geq \frac{\sigma}{2\mu}||\boldsymbol{x}_{t+1} - \boldsymbol{x}_t||^2 + v - v \geq 0$$

□

Remark 4.2. *Please note that if there exists a fully-mixed equilibrium strategy for the row player, any stable No-Regret learner (see definition 2.2.6.2) will not change his strategy when the equilibrium is played by the opponent (in our case, the column player).*

We proceed (lower) bounding the convergence step of the row player between odd rounds both for MWU, LMWU and OMD, and finally, showing the convergence.

Lemma 4.4. *Assume that the row player follows the MWU or LMWU algorithm with a non-increasing learning rate μ_t such that there exists $t' \in \mathbb{N}$ with $\mu_{t'} \leq \frac{1}{3}$. If the column player follows E-LRCA with $\beta \geq 2$ then*

$$KL(\boldsymbol{x}^*||\boldsymbol{x}_{2k-1}) - KL(\boldsymbol{x}^*||\boldsymbol{x}_{2k+1}) \geq \frac{1}{2}\mu_{2k}\alpha_{2k}(f(\boldsymbol{x}_{2k-1}) - v) \quad \forall k \in \mathbb{N}: \quad 2k \geq t'$$

where KL denotes the KL divergence (definition 2.2.6.1).

Proof. We start by bounding the KL for MWU. Following the Definition 2.2.6.1 we have:

$$\begin{aligned}
&KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{2k+1}\right) - KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{2k-1}\right) \\
&= \left(KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{2k+1}\right) - KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{2k}\right)\right) + \left(KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{2k}\right) - KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{2k-1}\right)\right) \\
&= \left(\sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\boldsymbol{x}^*(i)}{\boldsymbol{x}_{2k+1}(i)}\right) - \sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\boldsymbol{x}^*(i)}{\boldsymbol{x}_{2k}(i)}\right)\right) + \\
&\left(\sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\boldsymbol{x}^*(i)}{\boldsymbol{x}_{2k}(i)}\right) - \sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\boldsymbol{x}^*(i)}{\boldsymbol{x}_{2k-1}(i)}\right)\right) \\
&= \left(\sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\boldsymbol{x}_{2k}(i)}{\boldsymbol{x}_{2k+1}(i)}\right)\right) + \left(\sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\boldsymbol{x}_{2k-1}(i)}{\boldsymbol{x}_{2k}(i)}\right)\right)
\end{aligned}$$

Due to update rule of the multiplicative weights update (algorithm 2.6) we have:

$$\begin{aligned}
&KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{2k+1}\right) - KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{2k-1}\right) \\
&= \left(\mu_{2k}\boldsymbol{x}^{*\top}\boldsymbol{U}\boldsymbol{y}_{2k} + \ln\left(Z_{2k}\right)\right) + \left(\mu_{2k-1}\boldsymbol{x}^{*\top}\boldsymbol{U}\boldsymbol{y}_{2k-1} + \ln\left(Z_{2k-1}\right)\right) \\
&\leq \left(\mu_{2k}v + \ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{2k}(i) e^{-\mu_{2k}\boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_{2k}}\right)\right) + \left(\mu_{2k-1}v + \ln\left(Z_{2k-1}\right)\right) \qquad (4.6a) \\
&= \left(\mu_{2k}v + \ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{2k-1}(i) e^{-\mu_{2k-1}\boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_{2k-1}} e^{-\mu_{2k}\boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_{2k}}\right) - \ln\left(Z_{2k-1}\right)\right) \\
&+ \left(\mu_{2k-1}v + \ln\left(Z_{2k-1}\right)\right)
\end{aligned}$$

where inequality 4.6a comes from the definition of Nash Equilibrium. Thus:

$$\begin{aligned}
&KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{2k+1}\right) - KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{2k-1}\right) \\
&\leq \left(\mu_{2k}v + \ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{2k-1}(i) e^{-\mu_{2k-1}\boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}^*} e^{-\mu_{2k}\boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_{2k}}\right)\right) + \mu_{2k-1}v \\
&\leq \left(\mu_{2k}v + \ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{2k-1}(i) e^{-\mu_{2k-1}v} e^{-\mu_{2k}\boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_{2k}}\right)\right) + \mu_{2k-1}v \\
&= \mu_{2k}v + \ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{2k-1}(i) e^{-\mu_{2k}\boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_{2k}}\right) \qquad (4.7a)
\end{aligned}$$

where inequality 4.7a comes still from the definition of the Nash. Then, using the update rule of E-LRCA ($\boldsymbol{y}_{2k} = (1-\alpha_{2k})\,\boldsymbol{y}^* + \alpha_{2k}\boldsymbol{e}_{2k-1}$) we obtain:

$$
\begin{aligned}
&KL\left(\boldsymbol{x}^{*} \| \boldsymbol{x}_{2k+1}\right)-KL\left(\boldsymbol{x}^{*} \| \boldsymbol{x}_{2k-1}\right) \\
&\leq \mu_{2k} v+\ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{2k-1}(i) e^{-\mu_{2k} \boldsymbol{e}_{i}^{\top} \boldsymbol{U} \boldsymbol{y}_{2k}}\right) \\
&=\mu_{2k} v+\ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{2k-1}(i) e^{-\mu_{2k} \boldsymbol{e}_{i}^{\top} \boldsymbol{U}\left((1-\alpha_{2k}) \boldsymbol{y}^{*}+\alpha_{2k} \boldsymbol{e}_{2k-1}\right)}\right) \\
&\leq \mu_{2k} v+\ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{2k-1}(i) e^{-\mu_{2k}\left((1-\alpha_{2k}) v+\boldsymbol{e}_{i}^{\top} \boldsymbol{U}\left(\alpha_{2k} \boldsymbol{e}_{2k-1}\right)\right)}\right) \qquad (4.8a) \\
&\leq \mu_{2k} \alpha_{2k} v+\ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{2k-1}(i) e^{-\mu_{2k} \alpha_{2k} \boldsymbol{e}_{i}^{\top} \boldsymbol{A} \boldsymbol{e}_{2k-1}}\right) \\
&\leq \mu_{2k} \alpha_{2k} v+\ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{2k-1}(i)\left(1-\left(1-e^{-\mu_{2k} \alpha_{2k}}\right) \boldsymbol{e}_{i}^{\top} \boldsymbol{U} \boldsymbol{e}_{2k-1}\right)\right) \qquad (4.8b) \\
&=\mu_{2k} \alpha_{2k} v+\ln\left(1-\left(1-e^{-\mu_{2k} \alpha_{2k}}\right) {\boldsymbol{x}_{2k-1}}^{\top} \boldsymbol{U} \boldsymbol{e}_{2k-1}\right) \\
&\leq \mu_{2k} \alpha_{2k} v-\left(1-e^{-\mu_{2k} \alpha_{2k}}\right) \boldsymbol{x}_{2k-1}^{\top} \boldsymbol{U} \boldsymbol{e}_{2k-1} \qquad (4.8c) \\
&=\mu_{2k} \alpha_{2k} v-\left(1-e^{-\mu_{2k} \alpha_{2k}}\right) f\left(\boldsymbol{x}_{2k-1}\right)
\end{aligned}
$$

where inequality 4.8a is due to definition of the Nash and where inequalities 4.8b, 4.8c come from $\beta^{x} \leq 1-(1-\beta) x \quad \forall \beta \geq 0 \;\; x \in[0,1]$ and $\ln(1-x) \leq -x \quad \forall x<1$.

To conclude, we obtain:

$$
\begin{aligned}
&KL\left(\boldsymbol{x}^{*} \| \boldsymbol{x}_{2k+1}\right)-KL\left(\boldsymbol{x}^{*} \| \boldsymbol{x}_{2k-1}\right) \\
&\leq \mu_{2k} \alpha_{2k} v-\left(1-e^{-\mu_{2k} \alpha_{2k}}\right) f\left(\boldsymbol{x}_{2k-1}\right) \\
&\leq \mu_{2k} \alpha_{2k} v-\left(1-\left(1-\mu_{2k} \alpha_{2k}+\frac{1}{2}\left(\mu_{2k} \alpha_{2k}\right)^{2}\right)\right) f\left(\boldsymbol{x}_{2k-1}\right) \qquad (4.9a) \\
&=-\mu_{2k} \alpha_{2k}\left(f\left(\boldsymbol{x}_{2k-1}\right)-v\right)+\frac{1}{2}\left(\mu_{2k} \alpha_{2k}\right)^{2} f\left(\boldsymbol{x}_{2k-1}\right) \\
&\leq-\mu_{2k} \alpha_{2k}\left(f\left(\boldsymbol{x}_{2k-1}\right)-v\right)+\frac{1}{2} \mu_{2k} \alpha_{2k} \mu_{2k} \frac{f\left(\boldsymbol{x}_{2k-1}\right)-v}{f\left(\boldsymbol{x}_{2k-1}\right)} f\left(\boldsymbol{x}_{2k-1}\right) \qquad (4.9b) \\
&\leq-\mu_{2k} \alpha_{2k}\left(f\left(\boldsymbol{x}_{2k-1}\right)-v\right)+\frac{1}{2} \mu_{2k} \alpha_{2k}\left(f\left(\boldsymbol{x}_{2k-1}\right)-v\right) \qquad (4.9c) \\
&=-\frac{1}{2} \mu_{2k} \alpha_{2k}\left(f\left(\boldsymbol{x}_{2k-1}\right)-v\right) \leq 0
\end{aligned}
$$

where inequality 4.9a is due to $e^{x} \leq 1+x+\frac{1}{2} x^{2} \quad \forall \boldsymbol{x} \in(-\infty, 0]$, inequality 4.9b comes from the definition of α_{t}:

$$
\alpha_{t}=\gamma_{t} \frac{f\left(\boldsymbol{x}_{t-1}\right)-v}{\beta} \leq \frac{f\left(\boldsymbol{x}_{t-1}\right)-v}{\beta}, \beta \geq 2, f\left(\boldsymbol{x}_{2k-1}\right) \leq 1
$$

and inequality 4.9c comes from the choice of k s.t. $\mu_{2k} \leq 1$.

We now proceed bounding the KL divergence for the LMWU case. We recall the learning rate assumption:

$$\exists t \in \mathbb{N} \text{ such that } \mu_t \leq \frac{1}{3} \text{ and } \sum_{i=t}^{\infty} \mu_i = \infty$$

Using the update of LMWU (algorithm 2.7) we obtain:

$$\frac{\boldsymbol{x}_{m+1}(1)}{\boldsymbol{x}_m(1)} : \ldots : \frac{\boldsymbol{x}_{m+1}(n)}{\boldsymbol{x}_m(n)} = \left(1 - \mu_m \boldsymbol{e}_1^\top \boldsymbol{U} \boldsymbol{y}_m\right) : \ldots : \left(1 - \mu_m \boldsymbol{e}_n^\top \boldsymbol{U} \boldsymbol{y}_m\right) \forall m$$

Take m equal t and rearranging the equations we obtain:

$$\frac{\boldsymbol{x}_{t+1}(1)}{\boldsymbol{x}_{t-1}(1)} : \frac{\boldsymbol{x}_{t+1}(2)}{\boldsymbol{x}_{t-1}(2)} : \ldots : \frac{\boldsymbol{x}_{t+1}(n)}{\boldsymbol{x}_{t-1}(n)} =$$

$$= \left(1 - \mu_t \boldsymbol{e}_1^\top \boldsymbol{U} \boldsymbol{y}_t\right)\left(1 - \mu_{t-1} \boldsymbol{e}_1^\top \boldsymbol{U} \boldsymbol{y}_{t-1}\right) : \ldots : \left(1 - \mu_t \boldsymbol{e}_n^\top \boldsymbol{U} \boldsymbol{y}_t\right)\left(1 - \mu_{t-1} \boldsymbol{e}_n^\top \boldsymbol{U} \boldsymbol{y}_{t-1}\right)$$

which implies:

$$\boldsymbol{x}_{t+1}(i) = \frac{\boldsymbol{x}_{t-1}(i)\left(1 - \mu_t \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t\right)\left(1 - \mu_{t-1} \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_{t-1}\right)}{\sum_{j=1}^{n} \boldsymbol{x}_{t-1}(j)\left(1 - \mu_t \boldsymbol{e}_j{}^\top \boldsymbol{U} \boldsymbol{y}_t\right)\left(1 - \mu_{t-1} \boldsymbol{e}_j^\top \boldsymbol{U} \boldsymbol{y}_{t-1}\right)} \quad \forall i \in 1, 2, \ldots n$$

For even t, $\boldsymbol{y}_{t-1} = \boldsymbol{y}^*$. For any i such that $\boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}^* = v$, that is, for every action in the support of the equilibrium, we have:

$$\begin{aligned}
\frac{\boldsymbol{x}_{t+1}(i)}{\boldsymbol{x}_{t-1}(i)} &= \frac{\left(1 - \mu_{t-1} \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}^*\right)\left(1 - \mu_t \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t\right)}{\sum_{j=1}^{n} \boldsymbol{x}_{t-1}(j)\left(1 - \mu_{t-1} \boldsymbol{e}_j^\top \boldsymbol{U} \boldsymbol{y}^*\right)\left(1 - \mu_t \boldsymbol{e}_j^\top \boldsymbol{U} \boldsymbol{y}_t\right)} \\
&= \frac{\left(1 - \mu_{t-1} v\right)\left(1 - \mu_t \boldsymbol{e}_i{}^\top \boldsymbol{U} \boldsymbol{y}_t\right)}{\sum_{j=1}^{n} \boldsymbol{x}_{t-1}(j)\left(1 - \mu_{t-1} \boldsymbol{e}_j^\top \boldsymbol{U} \boldsymbol{y}^*\right)\left(1 - \mu_t \boldsymbol{e}_j^\top \boldsymbol{U} \boldsymbol{y}_t\right)} \\
&= \frac{\left(1 - \mu_t \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t\right)}{\sum_{j=1}^{n} \boldsymbol{x}_t(j) \frac{1 - \mu_{t-1} \boldsymbol{e}_j{}^\top \boldsymbol{U} \boldsymbol{y}^*}{1 - \mu_{t-1} v}\left(1 - \mu_t \boldsymbol{e}_j^\top \boldsymbol{U} \boldsymbol{y}_t\right)} \\
&\geq \frac{\left(1 - \mu_t \boldsymbol{e}_i{}^\top \boldsymbol{U} \boldsymbol{y}_t\right)}{\sum_{j=1}^{n} \boldsymbol{x}_{t-1}(j)\left(1 - \mu_t \boldsymbol{e}_j{}^\top \boldsymbol{U} \boldsymbol{y}_t\right)}
\end{aligned}$$

where the last inequality is due to definition of Nash Equilibrium. We also have that any j such that $\boldsymbol{e}_j^\top \boldsymbol{U} \boldsymbol{y}^* > v$ is outside the support of the equilibrium, namely $\boldsymbol{x}^*(j) = 0$.

Therefore, we proceed:

$$\begin{aligned} KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{t-1}\right) - KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{t+1}\right) &= \sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\boldsymbol{x}_{t+1}(i)}{\boldsymbol{x}_{t-1}(i)}\right) \\ &\geq \sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\left(1-\mu_t \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_t\right)}{\sum_{j=1}^{n} \boldsymbol{x}_{t-1}(j)\left(1-\mu_t \boldsymbol{e}_j^\top \boldsymbol{U}\boldsymbol{y}_t\right)}\right) \\ &= \sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\left(1-\mu_t \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_t\right)}{1-\mu_t \boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}_t}\right) \end{aligned}$$

Recalling that $\ln(x) \geq (x-1) - (x-1)^2 \quad \forall x \geq 0.5$ we obtain:

$$\begin{aligned} KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{t-1}\right) - KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{t+1}\right) \geq & \sum_{i=1}^{n} \boldsymbol{x}^*(i) \left(\frac{\left(1-\mu_t \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_t\right)}{1-\mu_t \boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}_t} - 1 - \left(\frac{\left(1-\mu_t \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_t\right)}{1-\mu_t \boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}_t} - 1\right)^2\right) \\ = & \frac{\mu_t\left(\boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}_t - \boldsymbol{x}^{*\top} \boldsymbol{U}\boldsymbol{y}_t\right)}{1-\mu_t \boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}_t} - \sum_{i=1}^{n} \boldsymbol{x}^*(i) \frac{\mu_t^2\left(\boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}_t - \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_t\right)^2}{\left(1-\mu_t \boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}_t\right)^2} \end{aligned}$$

Now, by update rule of the algorithm ($\boldsymbol{y}_t = (1-\alpha_t)\,\boldsymbol{y}^* + \alpha_t \boldsymbol{e}_{t-1}$) and recalling that $\boldsymbol{x}^*(j) = 0$ if $\boldsymbol{e}_j$ outside the support of the equilibrium, we can simplify the last equation and use the Cauchy theorem to obtain:

$$\begin{aligned} & KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{t-1}\right) - KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{t+1}\right) \\ & \geq \frac{\mu_t\left(1-\alpha_t\right)\left(\boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}^* - v\right)}{1-\mu_t \boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}_t} - \sum_{i=1}^{n} \boldsymbol{x}^*(i) \frac{2\mu_t^2\left(1-\alpha_t\right)^2\left(\boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}^* - v\right)^2}{\left(1-\mu_t \boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}_t\right)^2} + \\ & + \frac{\mu_t\alpha_t\left(\boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{e}_{t-1} - \boldsymbol{x}^{*\top} \boldsymbol{U}\boldsymbol{e}_{t-1}\right)}{1-\mu_t \boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}_t} - \sum_{i=1}^{n} \boldsymbol{x}^*(i) \frac{2\mu_t^2\alpha_t^2\left(\boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{e}_{t-1} - \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{e}_{t-1}\right)^2}{\left(1-\mu_t \boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}_t\right)^2} \end{aligned} \tag{4.10a}$$

Given $\mu_t \leq \frac{1}{3}$:

$$\frac{\mu_t\left(1-\alpha_t\right)\left(\boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}^* - v\right)}{1-\mu_t \boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}_t} - \sum_{i=1}^{n} \boldsymbol{x}^*(i) \frac{2\mu_t^2\left(1-\alpha_t\right)^2\left(\boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}^* - v\right)^2}{\left(1-\mu_t \boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}_t\right)^2} \geq 0$$

and:

$$\frac{\left(\boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{e}_{t-1} - \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{e}_{t-1}\right)^2}{\left(1-\mu_t \boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}_t\right)^2} \leq \frac{1}{\left(1-\mu_t\right)\left(1-\mu_t \boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}_t\right)}$$

From inequality 4.10a, we obtain:

$$\begin{aligned} &KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{t-1}\right) - KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{t+1}\right) \\ &\geq \frac{\mu_t\alpha_t\left(\boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{e}_{t-1} - \boldsymbol{x}^{*\top}\boldsymbol{U}\boldsymbol{e}_{t-1}\right)}{1-\mu_t\boldsymbol{x}_{t-1}^\top\boldsymbol{U}\boldsymbol{y}_t} - \frac{2\mu_t^2\alpha_t^2}{(1-\mu_t)\left(1-\mu_t\boldsymbol{x}_{t-1}^\top\boldsymbol{U}\boldsymbol{y}_t\right)} \end{aligned} \tag{4.11a}$$

We exploit the definition of α_t (that is $\alpha_t = \gamma_t\frac{f(\boldsymbol{x}_{t-1})-v}{\beta} \leq \frac{f(\boldsymbol{x}_{t-1})-v}{\beta}$) to have:

$$\alpha_t \leq \frac{\boldsymbol{x}_{t-1}^\top\boldsymbol{U}\boldsymbol{e}_{t-1} - v}{2} \leq \frac{\boldsymbol{x}_{t-1}^\top\boldsymbol{U}\boldsymbol{e}_{t-1} - \boldsymbol{x}^{*\top}\boldsymbol{U}\boldsymbol{e}_{t-1}}{2}$$

from which, with $\mu_t \leq \frac{1}{3}$ we have:

$$\frac{1}{2}\frac{\mu_t\alpha_t\left(\boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{e}_{t-1} - \boldsymbol{x}^{*\top}\boldsymbol{U}\boldsymbol{e}_{t-1}\right)}{1-\mu_t\boldsymbol{x}_{t-1}^\top\boldsymbol{U}\boldsymbol{y}_t} \geq \frac{2\mu_t^2\alpha_t^2}{(1-\mu_t)\left(1-\mu_t\boldsymbol{x}_{t-1}^\top\boldsymbol{U}\boldsymbol{y}_t\right)} \tag{4.12}$$

Substituting inequality 4.12 in 4.11a we obtain:

$$\begin{aligned} KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{t-1}\right) - KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{t+1}\right) &\geq \frac{1}{2}\frac{\mu_t\alpha_t\left(\boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{e}_{t-1} - \boldsymbol{x}^{*\top}\boldsymbol{U}\boldsymbol{e}_{t-1}\right)}{1-\mu_t\boldsymbol{x}_{t-1}^\top\boldsymbol{U}\boldsymbol{y}_t} \\ &\geq \frac{1}{2}\frac{\mu_t\alpha_t\left(\boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{e}_{t-1} - v\right)}{1-\mu_t\boldsymbol{x}_{t-1}^\top\boldsymbol{U}\boldsymbol{y}_t} \\ &\geq \frac{\mu_t\alpha_t\left(\boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{e}_{t-1} - v\right)}{2} \\ &\geq 0 \quad \forall t = 2k \end{aligned}$$

□

Theorem 4.4. *Assume that the row player follows OMD (FTRL) with σ-strongly convex distance generating function $F(\boldsymbol{x})$, with fixed learning rate such that $\mu \leq 1$ and $\sigma \geq 1$ or MWU, LMWU with $\mu_t \leq 1/3$. Then if the column player follows the Algorithm E-LRCA with $\beta \geq n^2$ and ξ_1, ξ_2 s.t. $v \in (\xi_1, \xi_2)$ and $||\boldsymbol{U}\boldsymbol{y}^*||_\infty < \xi_2$, there will be last round convergence to the minmax equilibrium.*

Proof. Let $\boldsymbol{x}^*$ be a minmax equilibrium of the row player. Denote by $B_F(\boldsymbol{x}_t||\boldsymbol{z}_t)$ the Bregman divergence between the current row player's strategy and lazy update of OMD (algorithm 2.8); then define $D_t(\boldsymbol{x}^*) := (B_F(\boldsymbol{x}^*||\boldsymbol{z}_t) - B_F(\boldsymbol{x}_t||\boldsymbol{z}_t))\frac{1}{\mu}$, following properties

of strongly convex function we have:

$$\begin{aligned}
D_t(\boldsymbol{x}^*) &= \left(F(\boldsymbol{x}^*) - F(\boldsymbol{z}_t) - \nabla F(\boldsymbol{z}_t)^\top(\boldsymbol{x}^* - \boldsymbol{z}_t) - \left(F(\boldsymbol{x}_t) - F(\boldsymbol{z}_t) - \nabla F(\boldsymbol{z}_t)^\top(\boldsymbol{x}_t - \boldsymbol{z}_t)\right)\right)\frac{1}{\mu} \\
&= \frac{1}{\mu}F(\boldsymbol{x}^*) - \frac{1}{\mu}F(\boldsymbol{z}_t) + (\boldsymbol{x}^* - \boldsymbol{z}_t)^\top \sum_{k=1}^{t-1} \boldsymbol{U}\boldsymbol{y}_k - \left(\frac{1}{\mu}F(\boldsymbol{x}_t) - \frac{1}{\mu}F(\boldsymbol{z}_t) + (\boldsymbol{x}_t - \boldsymbol{z}_t)^\top \sum_{k=1}^{t-1} \boldsymbol{U}\boldsymbol{y}_k\right) \\
&= \frac{1}{\mu}F(\boldsymbol{x}^*) - \frac{1}{\mu}F(\boldsymbol{x}_t) + (\boldsymbol{x}^* - \boldsymbol{x}_t)^\top \sum_{k=1}^{t-1} \boldsymbol{U}\boldsymbol{y}_k \\
&\geq \frac{\sigma}{2\mu}\left\|\boldsymbol{x}^* - \boldsymbol{x}_t\right\|^2
\end{aligned}$$

Thus, if $D_t(\boldsymbol{x}^*)$ converges to 0 then we have $\boldsymbol{x}_t$ converges to $\boldsymbol{x}^*$. We will prove that

$$D_{t-1}\left(\boldsymbol{x}^*\right) - D_{t+1}\left(\boldsymbol{x}^*\right) \geq \gamma_t \frac{\left(f\left(\boldsymbol{x}_{t-1}\right) - v\right)^2}{2n^2} \quad \forall t = 2k$$

From the definition of the Bregman we have:

$$\begin{aligned}
&D_{t-1}\left(\boldsymbol{x}^*\right) - D_{t+1}\left(\boldsymbol{x}^*\right) \\
&= D_{t-1}(\boldsymbol{x}_{t+1}) + \boldsymbol{x}_{t+1}^\top \boldsymbol{U}\left(\boldsymbol{y}_{t-1} + \boldsymbol{y}_t\right) - \boldsymbol{x}^{*\top}\boldsymbol{U}\left(\boldsymbol{y}_{t-1} + \boldsymbol{y}_t\right) \\
&\geq \frac{\sigma}{2\mu}\left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1}\right\|^2 + \boldsymbol{x}_{t+1}^\top \boldsymbol{U}\left(\boldsymbol{y}_{t-1} + \boldsymbol{y}_t\right) - \boldsymbol{x}^{*\top}\boldsymbol{U}\left(\boldsymbol{y}_{t-1} + \boldsymbol{y}_t\right) \qquad (4.13a)
\end{aligned}$$

From definition of minmax equilibrium in zero-sum games we have $\boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{y}^* \geq \boldsymbol{x}^{*\top}\boldsymbol{U}\boldsymbol{y}^* = v \ \ \forall \boldsymbol{x} \in \Delta_n$. Thus, given $\boldsymbol{y}_{t-1} = \boldsymbol{y}^*$ for an even t, we obtain:

$$\begin{aligned}
&D_{t-1}\left(\boldsymbol{x}^*\right) - D_{t+1}\left(\boldsymbol{x}^*\right) \\
&\geq \frac{\sigma}{2\mu}\left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1}\right\|^2 + \left(\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\right)^\top \boldsymbol{U}\boldsymbol{y}_t \\
&= \frac{\sigma}{2\mu}\left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1}\right\|^2 + \left(\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\right)^\top \boldsymbol{U}\left(\left(1 - \alpha_t\right)\boldsymbol{y}^* + \alpha_t \boldsymbol{e}_{t-1}\right) \qquad (4.14a) \\
&\geq \frac{\sigma}{2\mu}\left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1}\right\|^2 + \alpha_t\left(\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\right)^\top \boldsymbol{U}\boldsymbol{e}_{t-1} \qquad (4.14b) \\
&= \frac{\sigma}{2\mu}\left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1}\right\|^2 + \alpha_t\left(\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1}\right)^\top \boldsymbol{U}\boldsymbol{e}_{t-1} + \alpha_t\left(\boldsymbol{x}_{t-1} - \boldsymbol{x}^*\right)^\top \boldsymbol{U}\boldsymbol{e}_{t-1} \\
&\geq \frac{\sigma}{2\mu}\left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1}\right\|^2 - \alpha_t\left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1}\right\|\left\|\boldsymbol{U}\boldsymbol{e}_{t-1}\right\|_* + \alpha_t\left(f\left(\boldsymbol{x}_{t-1}\right) - v\right) \qquad (4.14c)
\end{aligned}$$

Inequalities 4.14a and 4.14b come from the update rule of E-LRCA, while inequality 4.14c comes from the definition of dual norm. Bounding the dual norm with the dimension of

the vector in inequality 4.14c we obtain:

$$\begin{aligned}
&D_{t-1}(\boldsymbol{x}^*) - D_{t+1}(\boldsymbol{x}^*) \\
&\geq \frac{\sigma}{2\mu}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1}\|^2 - n\alpha_t\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1}\| + \alpha_t(f(\boldsymbol{x}_{t-1}) - v) \\
&= \left(\sqrt{\frac{\sigma}{2\mu}}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1}\| - \frac{n\alpha_t}{2\sqrt{\frac{\sigma}{2\mu}}}\right)^2 + \alpha_t(f(\boldsymbol{x}_{t-1}) - v) - \frac{n^2\alpha_t^2\mu}{2\sigma} \\
&\geq \alpha_t(f(\boldsymbol{x}_{t-1}) - v) - \frac{n^2\alpha_t^2\mu}{2\sigma} \geq \alpha_t(f(\boldsymbol{x}_{t-1}) - v) - \frac{n^2\alpha_t^2}{2}
\end{aligned} \tag{4.15a}$$

Now, from E-LRCA we have

$$\alpha_t = \gamma_t \frac{f(\boldsymbol{x}_{t-1} - v)}{n^2}$$

then inequality 4.15a implies, as $\gamma_t^2 \leq \gamma_t$:

$$D_{t-1}(\boldsymbol{x}^*) - D_{t+1}(\boldsymbol{x}^*) \geq \frac{\alpha_t}{2}(f(\boldsymbol{x}_{t-1}) - v) = \gamma_t \frac{(f(\boldsymbol{x}_{t-1}) - v)^2}{2n^2} \geq 0 \;\; \forall t = 2k \tag{4.16}$$

Once that we have bounded the distance we proceed proving the convergence.

We have showed that the sequence of $D_{2k-1}(\boldsymbol{x}^*)$ is non-increasing (same has been done for the KL of MWU, LMWU in lemma 4.4). As the sequence is bounded below by 0, it has a limit for any minmax equilibrium strategy $\boldsymbol{x}^*$.

Then, we will prove that $\forall \epsilon > 0, \exists h \in \mathbb{N}$ such that following E-LRCA for the column player and OMD algorithm for the row player, the row player will play strategy $\boldsymbol{x}_h$ at round h and $f(\boldsymbol{x}_h) - v \leq \epsilon$. In particular, we prove this by contradiction. That is, suppose that $\exists \epsilon > 0$ such that $\forall h \in \mathbb{N}, \;\; f(\boldsymbol{x}_h) - v > \epsilon$. Then $\forall k \in \mathbb{N}$

$$\alpha_{2k}(f(\boldsymbol{x}_{2k-1}) - v) = \gamma_{2k}\frac{(f(\boldsymbol{x}_{2k-1}) - v)^2}{n^2} > \gamma_{2k}\frac{\epsilon^2}{n^2} \geq \gamma_{min}\frac{\epsilon^2}{n^2} > 0$$

Let k vary from $\lceil \frac{t'}{2} \rceil$ to T in equation 4.16. By summing over k, we obtain:

$$\begin{aligned}
D_{2T+1}(\boldsymbol{x}^*) &\leq D_{t'}(\boldsymbol{x}^*) - \frac{1}{2}\sum_{k=\lceil \frac{t'}{2} \rceil}^{T} \alpha_{2k}(f(\boldsymbol{x}_{2k-1}) - v) \\
&\leq D_{t'}(\boldsymbol{x}^*) - \frac{1}{2}\gamma_{min}\frac{\epsilon^2}{n^2}\sum_{k=\lceil \frac{t'}{2} \rceil}^{T} 1
\end{aligned}$$

Since $\lim_{T\to\infty}\sum_{k=\lceil\frac{t'}{2}\rceil}^{T} 1 = \infty$ and $D_{T+1}(\boldsymbol{x}^*) \geq 0$, which contradicts our assumption about $\forall h \in \mathbb{N},\ \ f(\boldsymbol{x}_h) - v > \epsilon$.

Please note that the same reasoning holds for MWU and LMWU since $\lim_{T\to\infty}\sum_{k=\lceil\frac{t'}{2}\rceil}^{T}\mu_{2k} = \infty$.

Now, we take a sequence of $\epsilon_k > 0$ such that $\lim_{k\to\infty}\epsilon_k = 0$. Then for each k, there exists $\boldsymbol{x}_{t_k} \in \Delta_n$ such that $v \leq f(\boldsymbol{x}_{t_k}) \leq v + \epsilon_k$. As Δ_n is a compact set and $\boldsymbol{x}_{t_k}$ is bounded then following the Bolzano-Weierstrass theorem, there is a convergence subsequence $\boldsymbol{x}_{\bar{t}_k}$. The limit of that sequence, $\boldsymbol{x}^*$, is a minmax equilibrium strategy of the row player (since $f(\boldsymbol{x}^*) = f(\lim_{k\to\infty}\boldsymbol{x}_{\bar{t}_k}) = \lim_{k\to\infty} f(\boldsymbol{x}_{\bar{t}_k}) = v$). Combining with the fact that$D_{2k-1}(\boldsymbol{x}^*)$ is non-increasing and $D_t(\boldsymbol{x}^*) = 0$ at convergence, we have $\lim_{k\to\infty} D_{2k-1}(\boldsymbol{x}^*) = 0$. We also note that from lemma 4.3, we have $\lim_{k\to\infty} D_{2k}(\boldsymbol{x}^*) = 0$ as well. Subsequently, $\lim_{t\to\infty} D_t(\boldsymbol{x}^*) = 0$, which concludes the proof.
The same reasoning holds for MWU and LMWU. □

4.5. Regret

First we report a fundamental lemma that will be useful for the Regret computation of theorem 4.5.

Lemma 4.5. *(Lemma 6.7 Orabona, 2019) Let B_F the Bregman divergence w.r.t. $F : X \to \mathbb{R}$ and assume F to be σ-strongly convex with respect to $||\cdot||$ in V. Let $V \subseteq X$ a non-empty closed convex set, then $\forall \boldsymbol{u} \in V$ following inequality holds:*

$$\mu(\boldsymbol{x}_t - \boldsymbol{u})^\top \boldsymbol{U}\boldsymbol{y}_t \leq B_F(\boldsymbol{u}||\boldsymbol{x}_t) - B_F(\boldsymbol{u}||\boldsymbol{x}_{t+1}) + \frac{\mu^2}{2\sigma}||\boldsymbol{U}\boldsymbol{y}_t||_*^2$$

We proceed showing the result for games without fully-mixed equilibrium strategy for the row player.

Theorem 4.5. *Assume that the row player follows the above-mentioned no-regret type algorithms: MWU, LMWU, FTRL, OMD with constant learning rate $\mu = 1/\sqrt{T}$; then by following E-LRCA, the column player will achieve the no-dynamic regret property with the dynamic regret satisfying $DR_T = \mathcal{O}\left(\frac{n^2}{\sqrt{\gamma_{min}}}T^{3/4}\right)$ in games without fully-mixed minmax equilibrium strategy for the row player.*

Proof. First we recall the bound to $f(\boldsymbol{x}_{2k-1}) - v$ in the case OMD (which for linear losses is equivalent to FTRL). Assume that $\max_{\boldsymbol{x}\in\Delta_n} F(\boldsymbol{x}) = 1$ without loss of generality.

Following the proof of Theorem 4.4 we have:

$$\sum_{k=1}^{T/2} \left(f\left(\boldsymbol{x}_{2k-1}\right) - v\right)^2 \leq \frac{2n^2}{\mu\gamma_{min}}$$

which implies:

$$\sum_{k=1}^{T/2} \left(f\left(\boldsymbol{x}_{2k-1}\right) - v\right) \leq \frac{n}{\sqrt{\mu\gamma_{min}}} T^{1/2} \tag{4.17}$$

Then we start by the definition of dynamic Regret:

$$DR_T := \sum_{k=1}^{T} \max_{\boldsymbol{y}\in\Delta} \boldsymbol{x}_k^\top \boldsymbol{U}\boldsymbol{y} - \boldsymbol{x}_k^\top \boldsymbol{U}\boldsymbol{y}_k$$

and we decompose the regret in the two different kinds of round:

$$DR_T := \sum_{k=1}^{T/2} \left(\max_{\boldsymbol{y}\in\Delta} \boldsymbol{x}_{2k}^\top \boldsymbol{U}\boldsymbol{y} - \boldsymbol{x}_{2k}^\top \boldsymbol{U}\boldsymbol{y}_{2k} + \max_{\boldsymbol{y}\in\Delta} \boldsymbol{x}_{2k-1}^\top \boldsymbol{U}\boldsymbol{y} - \boldsymbol{x}_{2k-1}^\top \boldsymbol{U}\boldsymbol{y}_{2k-1} \right)$$

inserting the right update for each kind of round:

$$\begin{aligned} DR_T &= \sum_{k=1}^{T/2} \left(\max_{\boldsymbol{y}\in\Delta} \boldsymbol{x}_{2k}^\top \boldsymbol{U}\boldsymbol{y} - \boldsymbol{x}_{2k}^\top \boldsymbol{U}\left((1-\alpha_{2k})\boldsymbol{y}^* + \alpha_{2k}\boldsymbol{e}_{2k-1}\right) + \max_{\boldsymbol{y}\in\Delta} \boldsymbol{x}_{2k-1}^\top \boldsymbol{U}\boldsymbol{y} - \boldsymbol{x}_{2k-1}^\top \boldsymbol{U}\boldsymbol{y}^* \right) \\ &= \sum_{k=1}^{T/2} \left(\max_{\boldsymbol{y}\in\Delta} \boldsymbol{x}_{2k}^\top \boldsymbol{U}\boldsymbol{y} - (1-\alpha_{2k})\boldsymbol{x}_{2k}^\top \boldsymbol{U}\boldsymbol{y}^* - \alpha_{2k}\boldsymbol{x}_{2k}^\top \boldsymbol{U}\boldsymbol{e}_{2k-1} + (f(\boldsymbol{x}_{2k-1}) - v) \right) \end{aligned}$$

recalling that $\boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{y}^* \geq v$:

$$\begin{aligned} &DR_T \\ &\leq \sum_{k=1}^{T/2} \left(\max_{\boldsymbol{y}\in\Delta} \boldsymbol{x}_{2k}^\top \boldsymbol{U}\boldsymbol{y} - (1-\alpha_{2k})v - \alpha_{2k}\boldsymbol{x}_{2k}^\top \boldsymbol{U}\boldsymbol{e}_{2k-1} + (f(\boldsymbol{x}_{2k-1}) - v) \right) \\ &= \sum_{k=1}^{T/2} \left((1-\alpha_{2k}) \max_{\boldsymbol{y}\in\Delta} \boldsymbol{x}_{2k}^\top \boldsymbol{U}\boldsymbol{y} - (1-\alpha_{2k})\, v + \alpha_{2k} \left(\max_{\boldsymbol{y}\in\Delta} \boldsymbol{x}_{2k}^\top \boldsymbol{U}\boldsymbol{y} - \boldsymbol{x}_{2k}^\top \boldsymbol{U}\boldsymbol{e}_{2k-1} \right) + (f(\boldsymbol{x}_{2k-1}) - v) \right) \end{aligned}$$

we notice that $\max_{\boldsymbol{y}\in\Delta} \boldsymbol{x}_{2k}^\top \boldsymbol{U}\boldsymbol{y} - \boldsymbol{x}_{2k}^\top \boldsymbol{U}\boldsymbol{e}_{2k-1} \leq 1$:

$$\begin{aligned} DR_T &\leq \sum_{k=1}^{T/2} \left((1-\alpha_{2k}) \max_{\boldsymbol{y}\in\Delta} \boldsymbol{x}_{2k}^\top \boldsymbol{U}\boldsymbol{y} - (1-\alpha_{2k})v + \alpha_{2k} + (f(\boldsymbol{x}_{2k-1}) - v) \right) \\ &\leq \sum_{k=1}^{T/2} \left((1-\alpha_{2k}) \left(\max_{\boldsymbol{y}\in\Delta} \boldsymbol{x}_{2k}^\top \boldsymbol{U}\boldsymbol{y} - v \right) + 2(f(\boldsymbol{x}_{2k-1}) - v) \right) \\ &\leq \sum_{k=1}^{T/2} \left(\left(||\boldsymbol{x}_{2k}^\top \boldsymbol{U}||_\infty - ||\boldsymbol{x}_{2k-1}^\top \boldsymbol{U}||_\infty \right) + 3(f(\boldsymbol{x}_{2k-1}) - v) \right) \end{aligned}$$

By inverse triangle inequality and by previous bound of $f(\boldsymbol{x}_{2k-1}) - v$ in inequality 4.17:

$$\begin{aligned} DR_T &\leq \sum_{k=1}^{T/2} \left(||\boldsymbol{U}^\top (\boldsymbol{x}_{2k} - \boldsymbol{x}_{2k-1})||_\infty \right) + \mathcal{O}\left(\frac{n}{\sqrt{\mu\gamma_{min}}} T^{\frac{1}{2}} \right) \\ &\leq \sum_{k=1}^{T/2} (||\boldsymbol{U}||_1 ||\boldsymbol{x}_{2k} - \boldsymbol{x}_{2k-1}||_\infty) + \mathcal{O}\left(\frac{n}{\sqrt{\mu\gamma_{min}}} T^{\frac{1}{2}} \right) \\ &\leq \sum_{k=1}^{T/2} (n||\boldsymbol{x}_{2k} - \boldsymbol{x}_{2k-1}||_\infty) + \mathcal{O}\left(\frac{n}{\sqrt{\mu\gamma_{min}}} T^{\frac{1}{2}} \right) \\ &\leq \sum_{k=1}^{T/2} (n||\boldsymbol{x}_{2k} - \boldsymbol{x}_{2k-1}||_2 + \mathcal{O}\left(\frac{n}{\sqrt{\mu\gamma_{min}}} T^{\frac{1}{2}} \right) \\ &= \sum_{k=1}^{T/2} (n||\boldsymbol{x}_{2k-1} - \boldsymbol{x}_{2k}||_2 + \mathcal{O}\left(\frac{n}{\sqrt{\mu\gamma_{min}}} T^{\frac{1}{2}} \right) \end{aligned}$$

By property of the Bregman: $(B_F(\boldsymbol{x}_{2k-1}||\boldsymbol{x}_{2k}) \geq \frac{\sigma}{2}||\boldsymbol{x}_{2k-1} - \boldsymbol{x}_{2k}||_2^2)$

$$DR_T \leq \sum_{k=1}^{T/2} \left(n\sqrt{\frac{2}{\sigma} B_F(\boldsymbol{x}_{2k-1}||\boldsymbol{x}_{2k})} \right) + \mathcal{O}\left(\frac{n}{\sqrt{\mu\gamma_{min}}} T^{\frac{1}{2}} \right)$$

Following lemma 4.5 with $\boldsymbol{u} = \boldsymbol{x}_t$ and for $\sigma \geq 1$:

$$DR_T \leq \sum_{k=1}^{T/2} \left(n\sqrt{\frac{2}{\sigma}\mu^2 ||\boldsymbol{U}\boldsymbol{y}_{2k-1}||_*^2} \right) + \mathcal{O}\left(\frac{n}{\sqrt{\mu\gamma_{min}}} T^{\frac{1}{2}} \right)$$

given that $\max_{\boldsymbol{y}\in\Delta_m} ||\boldsymbol{U}\boldsymbol{y}||_* \leq n$:

$$DR_T \leq \mathcal{O}\left(n^2 \mu T \right) + \mathcal{O}\left(\frac{n}{\sqrt{\mu\gamma_{min}}} T^{\frac{1}{2}} \right)$$

which for $\mu = 1/\sqrt{T}$:

$$DR_T = \mathcal{O}\left(\frac{n^2}{\sqrt{\gamma_{min}}}T^{3/4}\right)$$

□

The result in the case of fully-mixed equilibrium is presented as corollary of the previous theorem.

Corollary 4.1. *Assume that the row player follows the above-mentioned no-regret type algorithms: MWU, LMWU, FTRL, OMD. If there exists a fully-mixed minmax equilibrium strategy for the row player, then by following E-LRCA, the column player will achieve the no-dynamic regret property with the dynamic regret satisfying* $DR_T = \mathcal{O}\left(\frac{\sqrt{\log(n)}}{\sqrt{\gamma_{min}}}T^{3/4}\right)$ *for MWU and LMWU. Furthermore, in the case the row player uses a constant learning rate* μ*, we have* $DR_T = \mathcal{O}\left(\frac{n}{\sqrt{\mu\gamma_{min}}}T^{1/2}\right)$.

Proof. We start computing the Regret when row player employs MWU or LMWU algorithm. In odd rounds $2k-1$, the dynamic regret of the column player is:

$$\begin{aligned} DR_{2k-1} &= \max_{i\in\{1,\dots,m\}} \boldsymbol{x}_{2k-1}^\top \boldsymbol{U}\boldsymbol{e}_i - \boldsymbol{x}_{2k-1}^\top \boldsymbol{U}\boldsymbol{y}^* \\ &\leq f(\boldsymbol{x}_{2k-1}) - v \end{aligned}$$

In even rounds $2k$, for the existence of the fully-mixed minmax equilibrium of the row player, we have $\boldsymbol{x}_{2k} = \boldsymbol{x}_{2k-1}$. Combining the case of odd and even round, we derive

$$DR_T \leq 2\sum_{k=1}^{T/2}\left(f(\boldsymbol{x}_{2k-1}) - v\right)$$

Now, from Lemma 4.4 we have:

$$\frac{1}{2}\gamma_{2k}\mu_{2k}\frac{\left(f\left(\boldsymbol{x}_{2k-1}\right)-v\right)^2}{2} \leq KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{2k-1}\right) - KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{2k+1}\right)$$

from which in the case $n \geq 8$:

$$\sum_{k=1}^{T/2}\gamma_{2k}\mu_{2k}\left(f\left(\boldsymbol{x}_{2k-1}\right)-v\right)^2 \leq 4KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_1\right) \leq 4\ln(n)$$

Using the Cauchy-Schwarz inequality, we obtain:

$$\sum_{k=1}^{T/2} \left(f\left(\boldsymbol{x}_{2k-1}\right) - v\right) \leq 2\sqrt{\ln(n)} \sqrt{\sum_{k=1}^{T/2} \frac{1}{\gamma_{2k}\mu_{2k}}}$$

from which:

$$DR_T \leq 4\sqrt{\ln(n)} \sqrt{\sum_{k=1}^{T/2} \frac{1}{\gamma_{2k}\mu_{2k}}}$$

Exploiting γ_{min} as founded in Lemma 4.1 (for the fully-mixed case), if the row player follows a decreasing learning rate $\mu_k = \sqrt{8\ln(n)/k}$ (Cesa-Bianchi and Lugosi 7) we obtain:

$$\begin{aligned} DR_T &\leq 4\sqrt{\ln(n)} \sqrt{\sum_{k=1}^{T/2} \frac{1}{\gamma_{min}\sqrt{8\ln(n)/k}}} \\ &\leq \frac{\ln(n)^{1/4}}{\sqrt{\gamma_{min}}} T^{3/4} \\ &= \mathcal{O}\left(\frac{\sqrt{\log(n)}}{\sqrt{\gamma_{min}}} T^{3/4}\right) \end{aligned}$$

We continue the proof in the case of OMD (which is equivalent to FTRL for linear losses). We assume assume that $\max_{\boldsymbol{x}\in\Delta_n} F(\boldsymbol{x}) = 1$. From Theorem 4.4 we have:

$$\sum_{k=1}^{T/2} \left(f\left(\boldsymbol{x}_{2k-1}\right) - v\right)^2 \leq \frac{2n^2}{\mu\gamma_{min}}$$

By simple math:

$$\sum_{k=1}^{T/2} \left(f\left(\boldsymbol{x}_{2k-1}\right) - v\right) \leq \frac{n}{\sqrt{\mu\gamma_{min}}} T^{1/2}$$

From which:

$$DR_T \leq \frac{2n}{\sqrt{\mu\gamma_{min}}} T^{1/2}$$

that is:

$$DR_T = \mathcal{O}\left(\frac{n}{\sqrt{\mu\gamma_{min}}} T^{1/2}\right)$$

□

We point out some final remarks about the Regret computation.

Remark 4.3. *If there exists a fully-mixed strategy for the row player, it is possible to obtain a sublinear regret without specifying the learning rate μ.*

Remark 4.4. *In Corollary 4.1 we have a specific computation of the regret for MWU and LMWU; nevertheless, recall that MWU can be derived by OMD (which for linear losses is equivalent to FTRL).*

Remark 4.5. *The reader should notice that the regret scales with $1/\sqrt{\gamma_{min}}$. Since γ_{min} is smaller when there is not a fully-mixed equilibrium strategy for the row player (see lemma 4.1), this leads to a deceleration of the learning/teaching procedure for this specific setting.*

5 | Safe Guide with Partial Semi-Bandit Feedback

In this chapter we present the algorithm we developed for the partial semi-bandit feedback case: PAUSE E-LRCA (5.1).

5.1. Assumptions and Setting

In this section we provide the general assumptions/setting and most relevant considerations which will help the reader to understand the results and the proofs of the rest of the chapter.

As specified in the introduction we are in a zero-sum repeated game (players play the same game for each round); the entries of the payoff matrix (which is positive for the column player) are scaled in $[0, 1]$ without loss of generality. Row player has expert feedback, that is, the complete gradient is received. Thus, row player will receive $-\boldsymbol{U}\boldsymbol{y}_t$ after having chosen $\boldsymbol{x}_t$. Column player will instead receive a semi-bandit feedback, that is, the index of the action played by the opponent (sampled according to discrete distribution $\boldsymbol{x}_t$). We will refer to this feedback as "Partial Semi-Bandit", or simply "Partial Bandit" feedback. The payoff matrix $\boldsymbol{U}$ is known by the column player (the teacher), that is, he perfectly knows the equilibrium (this is called Asymmetric information) while row player (the learner/human) employs an algorithm of the OMD family (which for linear losses, as in our setting, is equivalent to FTRL).

We present our algorithm for the column player which will guarantee:

1. *Safety* (see definition 3.1.0.1): this property will be guaranteed in different ways depending on the equilibrium the game has. In case of fully-mixed equilibrium strategy for the row player, it will be possible to predict the opponent next strategy (see sections 5.2) with high probability so that safety can be obtained in an efficient manner; in case of not fully-mixed equilibrium, safety must be guaranteed for any possible strategy of the opponent, decelerating the learning procedure of

the opponent, but at the same time obtaining safety with probability equal to one. Moreover, in the latter case, an assumption on the upper bound value ξ_2 will be necessary.

2. *Last Round Convergence* (see definition 3.2.2.1) *with high probability*: for games with fully-mixed equilibrium strategy for the row player; experimental convergence in other games will be shown in chapter 6.

3. *Sublinear Dynamic Regret with respect to the MaxMin* (see definition 5.5.0.1) *with high probability*: for games with fully-mixed equilibrium strategy for the row player; experimental Sublinear Dynamic Regret with respect to the Maxmin in other games will be shown in chapter 6.

5.2. Algorithm

Algorithm 5.1 Engaged - Last Round Convergence in Asymmetric algorithm with partial semi-Bandit feedback (PAUSE E-LRCA)

for $t = 1$ to T **do**
  Play $\boldsymbol{y}_t = \boldsymbol{y}^*$ for $K(t) := \ln\left(\frac{3}{\delta}\right) t^{\lambda}$ times
  Compute $\bar{\boldsymbol{x}}_{K(t)}$ as the average of the $K(t)$ samples of the row player strategy
  Build $\tilde{X}_t$ using Devroye formula and flattening expansion
  $\boldsymbol{e}_t := \text{argmax}_{e \in \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots \boldsymbol{e}_m\}} \max_{x \in \tilde{X}_t} \boldsymbol{x}^\top \boldsymbol{U} \boldsymbol{e}$
  $f_{max}(\boldsymbol{x}_t) := \max_{e \in \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots \boldsymbol{e}_m\}} \max_{x \in \tilde{X}_t} \boldsymbol{x}^\top \boldsymbol{U} \boldsymbol{e}$
  $\boldsymbol{x}_{min} := \text{argmin}_{x \in \tilde{X}_t} \boldsymbol{x}^\top \boldsymbol{U} \boldsymbol{e}_t$
  if game has a fully-mixed equilibrium **then**
    if $\boldsymbol{x}_{min}^\top \boldsymbol{U} \boldsymbol{e}_t < \xi_1$ **then**
      $\alpha := \min\left(\frac{\xi_2 - v}{f_{max}(\boldsymbol{x}_t) - v}, \frac{\xi_1 - v}{\boldsymbol{x}_{min}^\top \boldsymbol{U} \boldsymbol{e}_t - v}\right)$
    end if
    if *not* **then**
      $\alpha := \frac{\xi_2 - v}{f_{max}(\boldsymbol{x}_t) - v}$
    end if
  end if
  if game has not a fully-mixed equilibrium **then**
    $\alpha = \min\left(\frac{\xi_2 - ||\boldsymbol{U}\boldsymbol{y}^*||_\infty}{||\boldsymbol{U}||_{max} - v}, \frac{\xi_1 - v}{||\boldsymbol{U}||_{min} - ||\boldsymbol{U}\boldsymbol{y}^*||_\infty}\right)$
  end if
  $\alpha_t := \min\left(\frac{f_{max}(\boldsymbol{x}_t) - v}{\beta}, \alpha\right)$
  $\boldsymbol{y}_t := (1 - \alpha_t)\, \boldsymbol{y}^* + \alpha_t \boldsymbol{e}_t$
end for

We underline the main ideas behind algorithm 5.1 (a part from safety result we consider the equilibrium to be fully-mixed, as the results for other kinds of equilibrium are mainly experimental).

Column player plays the equilibrium $K(t)$ times in order to estimate the opponent strategy with high probability. Please note that due to fully-mixed equilibrium, row player will keep choosing the same strategy, which implies that round after round we are collecting data from the same discrete distribution. Devroye formula (see lemma 3.1 and lemma 3.2) and flattening expansion (see section 5.3) allows to build the confidence region where row player's strategy lies with high probability. Then, column player computes the optimistic best response ($\boldsymbol{e}_t$) and the optimistic value of the best response $f_{max}(\boldsymbol{x}_t)$ with respect to

the estimated region in order to play a convex combination between $\boldsymbol{e}_t$ and the equilibrium, built using a parameter α_t.
α_t must be dependant on the distance between the optimistic value of the best response and the value of the equilibrium ($\alpha_t = \frac{f_{max}(\boldsymbol{x}_t)-v}{\beta}$), but, in case this parameter would lead to an utility outside the safety bounds (checked by the min operator), we scale $\frac{f_{max}(\boldsymbol{x}_t)-v}{\beta}$ by a factor $\gamma_t \in (0, 1]$ obtaining α (the multiplication $\gamma_t \frac{f_{max}(\boldsymbol{x}_t)-v}{\beta}$ is implicit in the algorithm but shown in section 5.3). As for the expert feedback algorithm, the scaling factor γ_t depends on the equilibrium the game has; in case there exists a fully-mixed equilibrium strategy for the row player, a γ_t such that the next round utility will be safe with respect to every strategy in the confidence interval is chosen, otherwise a γ_t safe for every strategy in the opponent simplex (the smallest γ possible) is needed. In the latter case, the choice of γ_t will lead to a deceleration of the teaching dynamic.

We then show a similar version of the algorithm which guarantees a monotonic convergence with high probability (not only at the limit as in the previous case), that is, distance measure (e.g. KL divergence) from current strategy of the row player and the equilibrium will not increase round after round. Regret has not been computed for this version as $K(t)$ cannot be estimated; indeed, in algorithm 5.2 the equilibrium is played until a condition is not met. The meaning of this condition will be clear after the proofs of section 5.4, but the idea is that column player will keep estimating the opponent strategy until he is not sure that his next strategy will make the opponent move towards the equilibrium. Please refer to chapter 6 for the difference in the experimental results between algorithm 5.1 and algorithm 5.2.

Algorithm 5.2 Engaged - Last Round Convergence in Asymmetric algorithm with partial semi-Bandit feedback with non-increasing KL (PAUSE E-LRCA)

for $t = 1$ to T **do**
  Play $\boldsymbol{y}_t = \boldsymbol{y}^*$ until $2\tilde{f}(\boldsymbol{x}_t) - f_{max}(\boldsymbol{x}_t) > v$
  Compute $\bar{\boldsymbol{x}}_{K(t)}$ as the average of the received samples of the row player strategy
  Build $\tilde{X}_t$ using Devroye formula and flattening expansion
  $\boldsymbol{e}_t := \text{argmax}_{e \in \{\boldsymbol{e}_1, \boldsymbol{e}_2, \dots \boldsymbol{e}_m\}} \max_{x \in \tilde{X}_t} \boldsymbol{x}^\top \boldsymbol{U} \boldsymbol{e}$
  $f_{max}(\boldsymbol{x}_t) := \max_{e \in \{\boldsymbol{e}_1, \boldsymbol{e}_2, \dots \boldsymbol{e}_m\}} \max_{x \in \tilde{X}_t} \boldsymbol{x}^\top \boldsymbol{U} \boldsymbol{e}$
  $\tilde{f}(\boldsymbol{x}_t) := \min_{x \in \tilde{X}_t} \boldsymbol{x}^\top \boldsymbol{U} \boldsymbol{e}_t$
  if game has a fully-mixed equilibrium **then**
    if $\tilde{f}(\boldsymbol{x}_t) < \xi_1$ **then**
      $\alpha := \min\left(\frac{\xi_2 - v}{f_{max}(\boldsymbol{x}_t) - v}, \frac{\xi_1 - v}{\tilde{f}(\boldsymbol{x}_t) - v}\right)$
    end if
    if *not* **then**
      $\alpha := \frac{\xi_2 - v}{f_{max}(\boldsymbol{x}_t) - v}$
    end if
  end if
  if game has not a fully-mixed equilibrium **then**
    $\alpha = \min\left(\frac{\xi_2 - ||\boldsymbol{U}\boldsymbol{y}^*||_\infty}{||\boldsymbol{U}||_{max} - v}, \frac{\xi_1 - v}{||\boldsymbol{U}||_{min} - ||\boldsymbol{U}\boldsymbol{y}^*||_\infty}\right)$
  end if
  $\alpha_t := \min\left(\frac{f_{max}(\boldsymbol{x}_t) - v}{\beta}, \alpha\right)$
  $\boldsymbol{y}_t := (1 - \alpha_t)\, \boldsymbol{y}^* + \alpha_t \boldsymbol{e}_t$
end for

5.3. Safety

In this section we provide the two main theorems related to safety of our algorithms with partial semi-bandit feedback. First, we introduce how the confidence set in which the opponent's strategy lies with high probability is built, then, we prove the property for games where there exists a fully-mixed equilibrium for the row player; finally, we conclude with the case in which equilibrium is not fully-mixed.

As underlined in the previous section we will use $K(t)$ to refer to the number of rounds in which equilibrium has been played consecutively, that are the rounds in which the opponent plays a fixed mixed strategy. Dealing with normal form games, row player strategy is defined as a probability distribution $\boldsymbol{x} \in \Delta_n$, where n denotes the number of actions available for the row player. In this case, column player observes $K(t)$ indices of actions

$i^1, \dots, i^{K(t)}$ sampled independently according to $\boldsymbol{x}$. Then, a natural estimator for $\boldsymbol{x}$ is the empirical frequency of actions $\bar{\boldsymbol{x}}_{K(t)} \in \Delta_n$, defined so that $\bar{\boldsymbol{x}}_{K(t)}(i) := \frac{1}{K(t)} \sum_{k=1}^{K(t)} \mathbf{1}\left\{i^k = i\right\}$ for every $i \in \{1, \dots, n\}$. By noticing that $K(t)\bar{\boldsymbol{x}}_{K(t)}$ is a random variable following a multinomial distribution with parameters $K(t)$ and $\boldsymbol{x}$, that is $K(t)\bar{\boldsymbol{x}}_{K(t)} \sim \mathcal{M}(K(t); \boldsymbol{x})$, lemma 3.1 can be used to derive the desired confidence intervals for the probabilities $\boldsymbol{x}(i)$. Then we can exploit lemma 3.2 to refine the previous result by giving bounds that hold for each component of $\boldsymbol{x}$ separately.
From that we build the first main component of our set:

$$X_{Dev_t} := \left\{ \boldsymbol{x} \in \Delta_n : \left|\bar{\boldsymbol{x}}_{K(t)}(i) - \boldsymbol{x}(i)\right| \leq \frac{5}{2}\sqrt{\frac{\ln(3/\delta)}{K(t)}} \quad \forall i \in \{1, \dots, n\} \right\} \tag{5.1}$$

At round t = 1, this estimation is the best we can achieve; fortunately, for the next ones, we can exploit the known dynamic of the row player in order to achieve a smaller set with high probability. Let's start making the assumption that the no-regret algorithm used by the row player is known, together with its learning rate; from that there exists a maximum step size s_t:

$$||\boldsymbol{x}_t - \boldsymbol{x}_{t-1}|| \leq s_t$$

Let's see how the maximum step size can be computed for MWU (algorithm 2.6), that is finding an upper bound for:

$$|\boldsymbol{x}_t(i) - \boldsymbol{x}_{t-1}(i)| = \left| \boldsymbol{x}_{t-1}(i) \frac{e^{-\mu \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_{t-1}}}{Z_{t-1}} - \boldsymbol{x}_{t-1}(i) \right| \tag{5.2}$$

Note that $\boldsymbol{U}\boldsymbol{y}_{t-1}$ is known, as it is the action played by the column player. Thus, the only unknown parameter is the value of the current strategy of the opponent. Aiming to find an upper bound to the step size, it is sufficient to take the row strategy in the confidence set $X_{Dev_{t-1}}$ that would maximize the column player strategy, namely $\tilde{\boldsymbol{x}}_{t-1} := \arg\max_{\boldsymbol{x} \in X_{Dev_{t-1}}} \boldsymbol{x}^\top \boldsymbol{U} \boldsymbol{y}_{t-1}$, from which:

$$s_t = \max_{i \in \{1, \dots, n\}} \tilde{\boldsymbol{x}}_{t-1}(i) \frac{e^{-\mu \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_{t-1}}}{Z_{t-1}} - \tilde{\boldsymbol{x}}_{t-1}(i) \tag{5.3}$$

Now we built the flattening expansion of the set $X_{Dev_{t-1}}$ that is defined as:

$$X_{s_t} := \left\{ \boldsymbol{x} \in \Delta_n : |\boldsymbol{x}(i) - \boldsymbol{w}(i)| \leq s_t \quad \forall i \in \{1, \dots, n\}, \forall \boldsymbol{w} \in X_{Dev_{t-1}} \right\} \tag{5.4}$$

From that we define:

$$\tilde{X}_t := \left\{ \boldsymbol{x} \in \Delta_n : \left| \bar{\boldsymbol{x}}_{K(t)}(i) - \boldsymbol{x}(i) \right| \leq \frac{5}{2} \sqrt{\frac{\ln(3/\delta)}{K(t)}} \quad \forall i \in \{1, \dots, n\} \right\}$$

$$\bigcap$$

$$\left\{ \boldsymbol{x} \in \Delta_n : |\boldsymbol{x}(i) - \boldsymbol{w}(i)| \leq s_t \quad \forall i \in \{1, \dots, n\}, \forall \boldsymbol{w} \in X_{Dev_{t-1}} \right\}$$

Please note the we encounter a loss in the probability related to the new set; in particular:

$$\mathbb{P}(\boldsymbol{x}_t \in X_{Dev_t}) = 1 - \delta \tag{5.5}$$

$$\mathbb{P}(\boldsymbol{x}_t \in X_{s_t}) = 1 - \frac{\delta}{e^{\frac{4}{5}(K(t-1))s_t\left(\frac{1}{5}s_t + \sqrt{\frac{\ln 3/\delta}{K(t-1)}}\right)}} \tag{5.6}$$

$$\mathbb{P}(\boldsymbol{x}_t \in X_{Dev_t} \cap \boldsymbol{x}_t \in X_{s_t}) = 1 - \delta \left(1 + \frac{1}{e^{\frac{4}{5}(K(t-1))s_t\left(\frac{1}{5}s_t + \sqrt{\frac{\ln 3/\delta}{K(t-1)}}\right)}}\right) = \mathbb{P}\left(\boldsymbol{x}_t \in \tilde{X}_t\right) \tag{5.7}$$

Procedure to build the probability of X_s is reported in appendix A.

A final remark about the flattening expansion.

Remark 5.1. *Building the final intersection set is useful to estimate with greater precision (with high probability) the strategy of the row player, which means reaching faster the convergence (experimentally). Nevertheless, the next proofs will be still be valid using only the Devroye estimation. Indeed, the flattening expansion can be built only by making strong assumptions such as knowing the opponent algorithm and the opponent learning rate. These assumptions may be not reasonable when playing against a human.*

Now we focus on the safety theorems:

Theorem 5.1. *Assume that the row player is following a no-regret stable learning algorithm, given two bounds ξ_1, ξ_2 on the Utility such that $v \in (\xi_1, \xi_2)$, if there exists a fully-mixed minmax equilibrium strategy for the row player and the column player follows PAUSE E-LRCA (algorithm 5.1 and algorithm 5.2), the Expected Utility of the column Player will be bounded in $[\xi_1, \xi_2]$ at each round with high probability (from which follows that the Expected Utility of the row Player will be bounded in $[-\xi_2, -\xi_1]$ at each round with high probability).*

Proof. The proof is divided in two parts: one for the rounds in which equilibrium is played, the other for the rounds in which update is performed.
In the first case, the column player will choose to play $\boldsymbol{y}^*$, which for assumption of fully-mixed equilibrium means $\boldsymbol{U}\boldsymbol{y}^* = [v, ..., v]$. Thus, we have that for any strategy the row player could choose, the utility would be v, that is inside the safety bounds.
As concerns the update rounds, we want to play a strategy $\boldsymbol{y}_t$ safe for every $\boldsymbol{x}_t \in \tilde{X}_t$. Thus, if the standard parameter $\alpha_t = \frac{f_{max}(\boldsymbol{x}_t - v)}{\beta}$ is not safe with high probability, we found a scale factor γ_t with the following procedure.
First we define:

$$(\boldsymbol{x}_{max}, \boldsymbol{e}) := \arg\max_{\boldsymbol{x} \in \tilde{X}_t} \arg\max_{\boldsymbol{e} \in \Delta_m} \boldsymbol{x}^\top \boldsymbol{U} \boldsymbol{e}$$

and

$$\boldsymbol{x}_{min} := \arg\min_{\boldsymbol{x} \in \tilde{X}_t} \boldsymbol{x}^\top \boldsymbol{U} \boldsymbol{e}$$

then we find the α_{max} value which guarantees safety in upper bound with respect to $\boldsymbol{x}_{max}$ and consequently with respect to any other strategy of the opponent in $\tilde{X}_t$ (with the same procedure followed for expert feedback in section 4.3) that is:

$$\alpha_{max} := \frac{(\xi_2 - v)}{\boldsymbol{x}_{max}^\top \boldsymbol{U} \boldsymbol{e}_t - v} = \frac{(\xi_2 - v)}{f_{max}(\boldsymbol{x}_t) - v}$$

which is equivalent to:

$$\alpha_{max} := \gamma_t \frac{f_{max}(\boldsymbol{x}_t) - v}{\beta}$$

with

$$\gamma_t := \frac{(\xi_2 - v)\beta}{(f_{max}(\boldsymbol{x}_t) - v)^2}$$

Now we make the same reasoning for the lower bound we obtain:

$$\alpha_{min} := \frac{(\xi_1 - v)}{\boldsymbol{x}_{min}^\top \boldsymbol{U} \boldsymbol{e}_t - v}$$

which is equivalent to:

$$\alpha_{min} := \gamma_t \frac{f_{max}(\boldsymbol{x}_t) - v}{\beta}$$

with

$$\gamma_t := \frac{(\xi_1 - v)\beta}{\left(\boldsymbol{x}_{min}^\top \boldsymbol{U} \boldsymbol{e}_t - v\right)\left(f_{max}(\boldsymbol{x}_t) - v\right)}$$

and finally play $\boldsymbol{y_t} := (1 - \alpha_t)\boldsymbol{y}^* + \alpha_t \boldsymbol{e}_t$ using the smallest between the three way of computing α described before.
Note if $\boldsymbol{x}_{min}^\top \boldsymbol{U} \boldsymbol{e}_t \geq \xi_1$, computation of α_{min} must be skipped, as we would be safe in lower

bound playing our optimistic best response.
Next we have to prove the safety with respect to the upper bound when α_{min} is chosen and the safety with respect to the lower bound when α_{max} is used, which leads to two situations situations ($\forall \boldsymbol{x} \in \tilde{X}_t$):

1. $\alpha_{max} < \alpha_{min}$:

$$\boldsymbol{x}^\top \boldsymbol{U} \left((1-\alpha_{max})\boldsymbol{y}^* + \alpha_{max}\boldsymbol{e}\right)$$
$$= v(1-\alpha_{max}) + \alpha_{max}\boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{e}$$

   with high probability we have:

$$\geq v(1-\alpha_{max}) + \alpha_{max}\boldsymbol{x}_{min}^\top \boldsymbol{U}\boldsymbol{e}$$

   given $\boldsymbol{x}_{min}^\top \boldsymbol{U}\boldsymbol{e} < v$ and $\alpha_{max} < \alpha_{min}$ we have:

$$\geq v(1-\alpha_{min}) + \alpha_{min}\boldsymbol{x}_{min}^\top \boldsymbol{U}\boldsymbol{e}$$

   that by construction of α_{min}

$$\geq \xi_1 \quad whp$$

2. $\alpha_{min} < \alpha_{max}$:

$$\boldsymbol{x}^\top \boldsymbol{U} \left((1-\alpha_{min})\boldsymbol{y}^* + \alpha_{min}\boldsymbol{e}\right)$$
$$= v(1-\alpha_{min}) + \alpha_{min}\boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{e}$$

   with high probability we have:

$$\leq v(1-\alpha_{min}) + \alpha_{min}\boldsymbol{x}_{max}^\top \boldsymbol{U}\boldsymbol{e}$$

   given $\boldsymbol{x}_{max}^\top \boldsymbol{U}\boldsymbol{e} > v$ and $\alpha_{min} < \alpha_{max}$ we have:

$$\leq v(1-\alpha_{max}) + \alpha_{max}\boldsymbol{x}_{max}^\top \boldsymbol{U}\boldsymbol{e}$$

   that by construction of α_{max}

$$\leq \xi_2 \quad whp$$

□

Theorem 5.2. *Given two bounds ξ_1, ξ_2 on the Utility such that $v \in (\xi_1, \xi_2)$ and $||\boldsymbol{U}\boldsymbol{y}^*||_\infty < \xi_2$, if there is not a fully-mixed minmax equilibrium strategy for the row player and the column player follows PAUSE E-LRCA (algorithm 5.1 or algorithm 5.2), the Utility of the column Player will be bounded in $[\xi_1, \xi_2]$ at each round (from which follows that the Utility of the row Player will be bounded in $[-\xi_2, -\xi_1]$ at each round).*

Proof. Proof is the same as the one with expert feedback (see section 4.3). □

5.4. Convergence

We summarize the main steps of this section in order to facilitate the comprehension.

In Lemma 5.1 we prove that the γ_t used to generate the parameter ($\alpha_t = \gamma_t \frac{f_{max}(\boldsymbol{x}_t) - v}{\beta}$) of the convex combination will be always greater than 0; it is necessary to avoid a premature interruption of the teaching dynamic.
In Lemma 5.2 we show how $K(t)$ must be chosen in order to guarantee Last Round Convergence.
In Theorem 5.3 we show how potentially every stable No-Regret Learner can reach any point of the simplex while playing against our algorithm, with high probability.
In Lemma 5.3 we lower bound the convergence step towards the equilibrium between update rounds when opponent employs MWU or LMWU.
In Theorem 5.4 we prove the Last Round Convergence with high probability for both MWU and LMWU.
Finally, in Theorem 5.5 we lower bound the convergence step towards the equilibrium between update rounds when opponent employs OMD and prove the Last Round Convergence with high probability.

Lemma 5.1. *Given a two-player zero-sum game in normal form, described by the matrix $\boldsymbol{U}$, and two bounds $[\xi_1, \xi_2]$ with $v \in (\xi_1, \xi_2)$, it is possible to find a positive lower bound γ_{min} s.t. all the γ_t found by PAUSE E-LRCA (algorithm 5.1 and algorithm 5.2) are greater or equal w.r.t γ_{min}.*

Proof. From Theorem 5.1 and following the same argument that in the expert case, in game with fully-mixed equilibrium strategy for the row player we obtain:

$$\gamma_{min} = \min\left(\frac{(\xi_2 - v)\beta}{(||\boldsymbol{U}||_{max} - v)^2}, \frac{(\xi_1 - v)\beta}{(||\boldsymbol{U}||_{min} - v)(||\boldsymbol{U}||_{max} - v)}\right) \tag{5.8}$$

which is greater than zero for $v \in (\xi_1, \xi_2)$. Note that in game with meaningful payoffs

$||\boldsymbol{U}||_{max} \neq ||\boldsymbol{U}||_{min} \neq v$, nevertheless, it is not a necessary assumption for the functioning of the algorithm.

In games where there is not an fully-mixed equilibrium strategy for the row player, γ_{min} is equivalent to the γ_t used at each round. Please note that as convergence will be proved only for games with fully-mixed equilibrium, this last result is not necessary. □

Lemma 5.2. *In order to guarantee Last Round Convergence following PAUSE E-LRCA (algorithm 5.1), it is necessary to choose $K(t)$ s.t. $\lim_{t\to\infty} K(t) = \infty$*

Proof. Using the same notation of the expert feedback case we call $f(\boldsymbol{x}_t)$ the value of the best response at time t. Trivially, this value is unknown at every step. However, it is possible to estimate at each step a set $\tilde{X}_t$ in which the strategy of the opponent lies with high probability. We use this set in order to compute the optimistic best response's value, defined as $f_{max}(\boldsymbol{x}_t)$. Thus, with high probability:

$$f_{max}(\boldsymbol{x}_t) = f(\boldsymbol{x}_t) + \sigma^{K(t)}$$

in which $\sigma^{K(t)}$ is a value dependant on the number of times $(K(t))$ equilibrium rounds have been played consecutively before the update round. Now, unfortunately both $f(\boldsymbol{x}_t)$ and $f_{max}(\boldsymbol{x}_t)$ are not sufficient to deeply describe the dynamic of the algorithm; therefore we introduce the value $f'(\boldsymbol{x}_t)$, that is the expected reward column player would obtain playing the optimistic best response. Again note that, with high probability:

$$f'(\boldsymbol{x}_t) = f(\boldsymbol{x}_t) - \rho^{K(t)}$$

in which $\rho^{K(t)}$ is a value dependant on the number of times $(K(t))$ equilibrium rounds have been played consecutively before the update round.
From these definitions we notice that there exists an interval in which column player's utility will lie with high probability; in order to achieve last round convergence is important to nullify the distance $\rho^{K(t)} + \sigma^{K(t)}$ at the limit, otherwise the optimistic best response would lead to instability in row player's dynamic. Formally, we want:

$$\lim_{t\to+\infty} \tilde{X}_t = \{\boldsymbol{x}_t\} \quad whp$$

which guarantees:

$$\lim_{t\to+\infty} \rho^{K(t)} + \sigma^{K(t)} = 0 \quad whp$$

Last inequality is true if and only if $\lim_{t\to\infty} K(t) = \infty$, which concludes the proof. □

Theorem 5.3. *Assume that the row player follows a stable no-regret algorithm and there exists a fully-mixed minmax equilibrium strategy for the row player. Then, by following PAUSE E-LRCA with ξ_1, ξ_2 s.t. $v \in (\xi_1, \xi_2)$, for any $\epsilon > 0$, there exists $t \in \mathbb{N}$ such that and $f(\boldsymbol{x}_t) - v \leq \epsilon$ with high probability.*

Proof. We will proceed to find a contradiction. Suppose there exists $\epsilon > 0$ such that:

$$f(\boldsymbol{x}_t) - v > \epsilon, \forall t \in \mathbb{N}$$

from which:

$$f_{max}(\boldsymbol{x}_t) - v > \epsilon, \forall t \in \mathbb{N} \quad whp \tag{5.9}$$

Then, by definition of parameter α_t and using the result on γ_{min} found in lemma 5.1 we have:

$$\alpha_t = \gamma_t \frac{f_{max}(\boldsymbol{x}_t) - v}{\beta} > \gamma_t \frac{\epsilon}{\beta} \geq \gamma_{min} \frac{\epsilon}{\beta}$$

Following the update rule of Algorithm PAUSE E-LRCA (algorithm 5.1 and algorithm 5.2) we have, with high probability:

$$\begin{aligned}
\boldsymbol{x}_t^\top \boldsymbol{U} \boldsymbol{y}_t &= \boldsymbol{x}_t^\top \boldsymbol{U} \left((1 - \alpha_t) \boldsymbol{y}^* + \alpha_t \boldsymbol{e}_t \right) \\
&\geq (1 - \alpha_t) v + \alpha_t f'(\boldsymbol{x}_t) && \text{(5.10a)} \\
&> (1 - \alpha_t) v + \alpha_t \left(v + \epsilon - \rho^{K(t)} \right) && \text{(5.10b)} \\
&\geq v + \gamma_{min} \frac{\epsilon \left(\epsilon - \rho^{K(t)} \right)}{\beta}
\end{aligned}$$

Where inequality 5.10a is due to the definition of Nash Equilibrium and where inequality 5.10b comes from the assumption 5.9. We also note that, from the definition of the value of the game, we have:

$$\min_i \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t = \min_i \boldsymbol{e}_i^\top \boldsymbol{U} \frac{\sum_{t=1}^{T} \boldsymbol{y}_t}{T} \leq v$$

Thus, we have (with the choice of $K(t)$ shown in lemma 5.2):

$$\lim_{T \to \infty} \min_i \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t - \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t^\top \boldsymbol{U} \boldsymbol{y}_t \leq v - \left(v + \gamma_{min} \frac{\epsilon^2}{\beta} \right) = -\gamma_{min} \frac{\epsilon^2}{\beta}$$

which contradicts the definition of a no-regret algorithm. □

We then lower bound the convergence step for MWU and LMWU with a similar procedure

with respect to the expert case.

Lemma 5.3. *Assume that the row player follows the MWU or LMWU algorithm with a non-increasing learning rate μ_t such that there exists $t' \in \mathbb{N}$ with $\mu_{t'} \leq \frac{1}{3}$. If there exists a fully-mixed minmax equilibrium strategy for the row player and the column player follows PAUSE E-LRCA with $\beta \geq 2$ then*

$$KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_{K(t)_t}\right) - KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_{1_{t+1}}\right) \geq \frac{1}{2}\mu_t \alpha_t \left(2f'\left(\boldsymbol{x}_t\right) - f_{max}\left(\boldsymbol{x}_t\right) - v\right) \quad \forall t \in \mathbb{N}: \quad t \geq t'$$

where KL denotes the KL divergence, $\boldsymbol{x}_{K(t)_t}$ is the strategy of the row player at round t after the last equilibrium have been played and $\boldsymbol{x}_{1_{t+1}}$ is the strategy of the row player when the first equilibrium has been played at round $t+1$.

Proof. We start bounding the KL divergence for MWU. By the definition of KL (definition 2.2.6.1) we have:

$$\begin{aligned}
&KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_{1_{t+1}}\right) - KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_{K(t)_t}\right) \\
&= \left(KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_{1_{t+1}}\right) - KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_t\right)\right) + \left(KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_t\right) - KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_{K(t)_t}\right)\right) \\
&= \left(\sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\boldsymbol{x}^*(i)}{\boldsymbol{x}_{1_{t+1}(i)}}\right) - \sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\boldsymbol{x}^*(i)}{\boldsymbol{x}_t(i)}\right)\right) + \\
&\left(\sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\boldsymbol{x}^*(i)}{\boldsymbol{x}_t(i)}\right) - \sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\boldsymbol{x}^*(i)}{\boldsymbol{x}_{K(t)_t}(i)}\right)\right) \\
&= \left(\sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\boldsymbol{x}_t(i)}{\boldsymbol{x}_{1_{t+1}}(i)}\right)\right) + \left(\sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\boldsymbol{x}_{K(t)_t}(i)}{\boldsymbol{x}_t(i)}\right)\right)
\end{aligned}$$

Due to update rule of the multiplicative weights update (algorithm 2.6) we have:

$$\begin{aligned}
&KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_{1_{t+1}}\right) - KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_{K(t)_t}\right) \\
&= \left(\mu_t \boldsymbol{x}^{*\top} \boldsymbol{U} \boldsymbol{y}_t + \ln\left(Z_t\right)\right) + \left(\mu_{K(t)_t} \boldsymbol{x}^{*\top} \boldsymbol{U} \boldsymbol{y}_{K(t)_t} + \ln\left(Z_{K(t)_t}\right)\right) \\
&\leq \left(\mu_t v + \ln\left(\sum_{i=1}^{n} \boldsymbol{x}_t(i) e^{-\mu_t \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t}\right)\right) + \left(\mu_{K(t)_t} v + \ln\left(Z_{K(t)_t}\right)\right) \qquad (5.12a) \\
&= \left(\mu_t v + \ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{K(t)_t}(i) e^{-\mu_{K(t)_t} \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_{K(t)_t}} e^{-\mu_t \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t}\right) - \ln\left(Z_{K(t)_t}\right)\right) \\
&+ \left(\mu_{K(t)_t} v + \ln\left(Z_{K(t)_t}\right)\right)
\end{aligned}$$

where Inequality 5.12a comes from the definition of Nash Equilibrium. Thus:

$$\begin{aligned}
&KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{1_{t+1}}\right) - KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{K(t)_t}\right) \\
&\leq \left(\mu_t v + \ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{K(t)_t}(i) e^{-\mu_{K(t)_t}\boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}^*} e^{-\mu_t \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_t}\right)\right) + \mu_{K(t)_t} v \\
&\leq \left(\mu_t v + \ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{K(t)_t}(i) e^{-\mu_{K(t)_t} v} e^{-\mu_t \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_t}\right)\right) + \mu_{K(t)_t} v \qquad (5.13a) \\
&= \mu_t v + \ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{K(t)_t}(i) e^{-\mu_t \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_t}\right)
\end{aligned}$$

where inequality 5.13a comes still from the definition of the Nash. Then, using the update rule of Algorithm PAUSE E-LRCA (algorithm 5.1 and algorithm 5.2) that is, $\boldsymbol{y}_t = (1-\alpha_t)\,\boldsymbol{y}^* + \alpha_t \boldsymbol{e}_t$ we obtain:

$$\begin{aligned}
&KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{1_{t+1}}\right) - KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{K(t)_t}\right) \\
&\leq \mu_t v + \ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{K(t)_t}(i) e^{-\mu_t \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{y}_t}\right) \\
&= \mu_t v + \ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{K(t)_t}(i) e^{-\mu_t \boldsymbol{e}_i^\top \boldsymbol{U}((1-\alpha_t)\boldsymbol{y}^* + \alpha_t \boldsymbol{e}_t)}\right) \\
&\leq \mu_t v + \ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{K(t)_t}(i) e^{-\mu_t\left((1-\alpha_t)v + \boldsymbol{e}_i^\top \boldsymbol{U}(\alpha_t \boldsymbol{e}_t)\right)}\right) \qquad (5.14a) \\
&\leq \mu_t \alpha_t v + \ln\left(\sum_{i=1}^{n} \boldsymbol{x}_{K(t)_t}(i)\left(1 - \left(1 - e^{-\mu_t \alpha_t}\right)\boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{e}_t\right)\right) \qquad (5.14b) \\
&= \mu_t \alpha_t v + \ln\left(1 - \left(1 - e^{-\mu_t \alpha_t}\right)\boldsymbol{x}_{K(t)_t}{}^\top \boldsymbol{U}\boldsymbol{e}_t\right) \\
&\leq \mu_t \alpha_t v - \left(1 - e^{-\mu_t \alpha_t}\right)\boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U}\boldsymbol{e}_t \qquad (5.14c) \\
&= \mu_t \alpha_t v - \left(1 - e^{-\mu_t \alpha_t}\right) f'\left(\boldsymbol{x}_t\right)
\end{aligned}$$

where inequality 5.14a is due to definition of the Nash and where inequalities 5.14b and 5.14c come from $\beta^x \leq 1 - (1-\beta)x \quad \forall \beta \geq 0 \;\; x \in [0,1]$ and $\ln(1-x) \leq -x \quad \forall x < 1$.

To conclude, we obtain:

$$
\begin{aligned}
& KL\left(\boldsymbol{x}^{*} \| \boldsymbol{x}_{1_{t+1}}\right)-KL\left(\boldsymbol{x}^{*} \| \boldsymbol{x}_{K(t)_{t}}\right) \\
& \leq \mu_{t} \alpha_{t} v-\left(1-e^{-\mu_{t} \alpha_{t}}\right) f^{\prime}\left(\boldsymbol{x}_{t}\right) \\
& \leq \mu_{t} \alpha_{t} v-\left(1-\left(1-\mu_{t} \alpha_{t}+\frac{1}{2}\left(\mu_{t} \alpha_{t}\right)^{2}\right)\right) f^{\prime}\left(\boldsymbol{x}_{t}\right) && \text{(5.15a)} \\
& =-\mu_{2 k} \alpha_{2 k}\left(f^{\prime}\left(\boldsymbol{x}_{t}\right)-v\right)+\frac{1}{2}\left(\mu_{t} \alpha_{t}\right)^{2} f^{\prime}\left(\boldsymbol{x}_{t}\right) \\
& \leq-\mu_{t} \alpha_{t}\left(f^{\prime}\left(\boldsymbol{x}_{t}\right)-v\right)+\frac{1}{2} \mu_{t} \alpha_{t} \mu_{t} \frac{f_{max}\left(\boldsymbol{x}_{t}\right)-v}{f^{\prime}\left(\boldsymbol{x}_{t}\right)} f^{\prime}\left(\boldsymbol{x}_{t}\right) && \text{(5.15b)} \\
& \leq-\mu_{t} \alpha_{t}\left(f^{\prime}\left(\boldsymbol{x}_{t}\right)-v\right)+\frac{1}{2} \mu_{t} \alpha_{t}\left(f_{max}\left(\boldsymbol{x}_{t}\right)-v\right) && \text{(5.15c)} \\
& =-\frac{1}{2} \mu_{t} \alpha_{t}\left(2 f^{\prime}\left(\boldsymbol{x}_{t}\right)-f_{max}\left(\boldsymbol{x}_{t}\right)-v\right) \leq 0
\end{aligned}
$$

where inequality 5.15a is due to $e^{x} \leq 1+x+\frac{1}{2} x^{2} \quad \forall \boldsymbol{x} \in[-\infty, 0]$, inequality 5.15b comes from the definition of α_t:

$$
\alpha_{t}=\gamma_{t} \frac{f_{max}\left(\boldsymbol{x}_{t}\right)-v}{\beta} \leq \frac{f_{max}\left(\boldsymbol{x}_{t}\right)-v}{\beta}, \beta \geq 2, f^{\prime}\left(\boldsymbol{x}_{t}\right) \leq 1
$$

and inequality 4.9c comes from the choice of t so that $\mu_t \leq 1$. Please note the final inequality is true if:

$$
\begin{aligned}
2 f^{\prime}\left(\boldsymbol{x}_{t}\right)-f_{max}\left(\boldsymbol{x}_{t}\right) & \geq v \\
2\left(f\left(\boldsymbol{x}_{t}\right)-\rho^{K(t)}\right)-f\left(\boldsymbol{x}_{t}\right)-\sigma^{K(t)}-v & \geq 0 \\
2 f\left(\boldsymbol{x}_{t}\right)-2 \rho^{K(t)}-f\left(\boldsymbol{x}_{t}\right)-\sigma^{K(t)}-v & \geq 0 \\
f\left(\boldsymbol{x}_{t}\right)-v-2 \rho^{K(t)}-\sigma^{K(t)} & \geq 0 \\
f\left(\boldsymbol{x}_{t}\right)-v & \geq 2 \rho^{K(t)}+\sigma^{K(t)} && \text{(5.16a)}
\end{aligned}
$$

We now proceed bounding the KL divergence for the LMWU (algorithm 2.7) case. We recall the learning rate assumption:

$$
\exists t \in \mathbb{N} \text{ such that } \mu_{t} \leq \frac{1}{3} \text{ and } \sum_{i=t}^{\infty} \mu_{i}=\infty
$$

Using the update rule of LMWU we obtain:

$$
\frac{\boldsymbol{x}_{m+1}(1)}{\boldsymbol{x}_{m}(1)}: \ldots: \frac{\boldsymbol{x}_{m+1}(n)}{\boldsymbol{x}_{m}(n)}=\left(1-\mu_{m} \boldsymbol{e}_{1}^{\top} \boldsymbol{U} \boldsymbol{y}_{m}\right): \ldots:\left(1-\mu_{m} \boldsymbol{e}_{n}^{\top} \boldsymbol{U} \boldsymbol{y}_{m}\right) \forall m
$$

Take m equal t and rearranging the equations we obtain:

$$\frac{\boldsymbol{x}_{1_{t+1}}(1)}{\boldsymbol{x}_{K(t)_t}(1)} : \frac{\boldsymbol{x}_{1_{t+1}}(2)}{\boldsymbol{x}_{K(t)_t}(2)} : \ldots : \frac{\boldsymbol{x}_{1_{t+1}}(n)}{\boldsymbol{x}_{K(t)_t}(n)} =$$

$$= \left(1 - \mu_t \boldsymbol{e}_1^\top \boldsymbol{U} \boldsymbol{y}_t\right)\left(1 - \mu_{K(t)_t} \boldsymbol{e}_1^\top \boldsymbol{U} \boldsymbol{y}_{K(t)_t}\right) : \ldots : \left(1 - \mu_t \boldsymbol{e}_n^\top \boldsymbol{U} \boldsymbol{y}_t\right)\left(1 - \mu_{K(t)_t} \boldsymbol{e}_n^\top \boldsymbol{U} \boldsymbol{y}_{K(t)_t}\right)$$

which implies:

$$\boldsymbol{x}_{1_{t+1}}(i) = \frac{\boldsymbol{x}_{K(t)_t}(i)\left(1 - \mu_t \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t\right)\left(1 - \mu_{K(t)_t} \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_{K(t)_t}\right)}{\sum_{j=1}^{n} \boldsymbol{x}_{K(t)_t}(j)\left(1 - \mu_t \boldsymbol{e}_j{}^\top \boldsymbol{U} \boldsymbol{y}_t\right)\left(1 - \mu_{K(t)_t} \boldsymbol{e}_j^\top \boldsymbol{U} \boldsymbol{y}_{K(t)_t}\right)} \quad \forall i \in 1, 2, \ldots n$$

Note that $\boldsymbol{y}_{K(t)_t} = \boldsymbol{y}^*$ in PAUSE E-LRCA (algorithm 5.1 and algorithm 5.2). For any i such that $\boldsymbol{e}_i{}^\top \boldsymbol{U} \boldsymbol{y}^* = v$, that is, for every action in the support of the equilibrium, we have:

$$\begin{aligned}
\frac{\boldsymbol{x}_{1_{t+1}}(i)}{\boldsymbol{x}_{K(t)_t}(i)} &= \frac{\left(1 - \mu_{K(t)_t} \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}^*\right)\left(1 - \mu_t \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t\right)}{\sum_{j=1}^{n} \boldsymbol{x}_{K(t)_t}(j)\left(1 - \mu_{K(t)_t} \boldsymbol{e}_j^\top \boldsymbol{U} \boldsymbol{y}^*\right)\left(1 - \mu_t \boldsymbol{e}_j^\top \boldsymbol{U} \boldsymbol{y}_t\right)} \\
&= \frac{\left(1 - \mu_{K(t)_t} v\right)\left(1 - \mu_t \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t\right)}{\sum_{j=1}^{n} \boldsymbol{x}_{K(t)_t}(j)\left(1 - \mu_{K(t)_t} \boldsymbol{e}_j^\top \boldsymbol{U} \boldsymbol{y}^*\right)\left(1 - \mu_t \boldsymbol{e}_j^\top \boldsymbol{U} \boldsymbol{y}_t\right)} \\
&= \frac{\left(1 - \mu_t \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t\right)}{\sum_{j=1}^{n} \boldsymbol{x}_{K(t)_t}(j) \frac{1 - \mu_{K(t)_t} \boldsymbol{e}_j^\top \boldsymbol{U} \boldsymbol{y}^*}{1 - \mu_{K(t)_t} v}\left(1 - \mu_t \boldsymbol{e}_j^\top \boldsymbol{U} \boldsymbol{y}_t\right)} \\
&\geq \frac{\left(1 - \mu_t \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t\right)}{\sum_{j=1}^{n} \boldsymbol{x}_{K(t)_t}(j)\left(1 - \mu_t \boldsymbol{e}_j{}^\top \boldsymbol{U} \boldsymbol{y}_t\right)}
\end{aligned}$$

where the last inequality is due to definition of Nash Equilibrium. We also have any j such that $\boldsymbol{e}_j^\top \boldsymbol{U} \boldsymbol{y}^* > v$ is outside the support of the equilibrium, namely $\boldsymbol{x}^*(j) = 0$. Therefore, we proceed:

$$\begin{aligned}
KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_{K(t)_t}\right) - KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_{1_{t+1}}\right) &= \sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\boldsymbol{x}_{1_{t+1}}(i)}{\boldsymbol{x}_{K(t)_t}(i)}\right) \\
&\geq \sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\left(1 - \mu_t \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t\right)}{\sum_{j=1}^{n} \boldsymbol{x}_{K(t)_t}(j)\left(1 - \mu_t \boldsymbol{e}_j^\top \boldsymbol{U} \boldsymbol{y}_t\right)}\right) \\
&= \sum_{i=1}^{n} \boldsymbol{x}^*(i) \ln\left(\frac{\left(1 - \mu_t \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t\right)}{1 - \mu_t \boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t}\right)
\end{aligned}$$

Recalling that $\ln(x) \geq (x-1) - (x-1)^2 \quad \forall x \geq 0.5$ we obtain:

$$\begin{aligned}
&KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_{K(t)_t}\right) - KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_{1_{t+1}}\right) \\
\geq &\sum_{i=1}^{n} \boldsymbol{x}^*(i) \left(\frac{\left(1 - \mu_t \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t\right)}{1 - \mu_t \boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t} - 1 - \left(\frac{\left(1 - \mu_t \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t\right)}{1 - \mu_t \boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t} - 1 \right)^2 \right) \\
= &\frac{\mu_t \left(\boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t - \boldsymbol{x}^{*\top} \boldsymbol{U} \boldsymbol{y}_t \right)}{1 - \mu_t \boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t} - \sum_{i=1}^{n} \boldsymbol{x}^*(i) \frac{\mu_t^2 \left(\boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t - \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}_t \right)^2}{\left(1 - \mu_t \boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t\right)^2}
\end{aligned}$$

Now, by update rule of the algorithm ($\boldsymbol{y}_t = (1-\alpha_t)\,\boldsymbol{y}^* + \alpha_t \boldsymbol{e}_t$) and recalling that $\boldsymbol{x}^*(j) = 0$ if $\boldsymbol{e}_j$ outside the support of the equilibrium, we can simplify the last equation and use the Cauchy theorem to obtain:

$$\begin{aligned}
&KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_{K(t)_t}\right) - KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_{1_{t+1}}\right) \geq \\
&\frac{\mu_t (1-\alpha_t) \left(\boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}^* - v \right)}{1 - \mu_t \boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t} - \sum_{i=1}^{n} \boldsymbol{x}^*(i) \frac{2\mu_t^2 (1-\alpha_t)^2 \left(\boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}^* - v \right)^2}{\left(1 - \mu_t \boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t\right)^2} \\
&+ \frac{\mu_t \alpha_t \left(\boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{e}_t - \boldsymbol{x}^{*\top} \boldsymbol{U} \boldsymbol{e}_t \right)}{1 - \mu_t \boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t} - \sum_{i=1}^{n} \boldsymbol{x}^*(i) \frac{2\mu_t^2 \alpha_t^2 \left(\boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{e}_t - \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{e}_t \right)^2}{\left(1 - \mu_t \boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t\right)^2}
\end{aligned} \tag{5.17a}$$

Given $\mu_t \leq \frac{1}{3}$:

$$\frac{\mu_t (1-\alpha_t) \left(\boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}^* - v \right)}{1 - \mu_t \boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t} - \sum_{i=1}^{n} \boldsymbol{x}^*(i) \frac{2\mu_t^2 (1-\alpha_t)^2 \left(\boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}^* - v \right)^2}{\left(1 - \mu_t \boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t\right)^2} \geq 0$$

and:

$$\frac{\left(\boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{e}_t - \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{e}_t \right)^2}{\left(1 - \mu_t \boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t\right)^2} \leq \frac{1}{(1-\mu_t)\left(1 - \mu_t \boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t\right)}$$

From inequality 5.17a, we obtain:

$$\begin{aligned}
&KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_{K(t)_t}\right) - KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_{1_{t+1}}\right) \\
&\geq \frac{\mu_t \alpha_t \left(\boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{e}_t - \boldsymbol{x}^{*\top} \boldsymbol{U} \boldsymbol{e}_t \right)}{1 - \mu_t \boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t} - \frac{2\mu_t^2 \alpha_t^2}{(1-\mu_t)\left(1 - \mu_t \boldsymbol{x}_{K(t)_t}^\top \boldsymbol{U} \boldsymbol{y}_t\right)}
\end{aligned} \tag{5.18a}$$

for $\beta \geq 2$ and $\mu_t \leq \frac{1}{3}$:

$$\frac{1}{2}\frac{\mu_t\alpha_t\left(f_{max}(\boldsymbol{x}_t)-v\right)}{1-\mu_t\boldsymbol{x}_{K(t)_t}^{\top}\boldsymbol{U}\boldsymbol{y}_t} \geq \frac{2\mu_t^2\alpha_t^2}{(1-\mu_t)\left(1-\mu_t\boldsymbol{x}_{K(t)_t}^{\top}\boldsymbol{U}\boldsymbol{y}_t\right)} \tag{5.19}$$

Substituting inequality 5.19 in 5.18a we obtain:

$$\begin{aligned} KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{K(t)_t}\right) - KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{1_{t+1}}\right) &\geq \frac{1}{2}\frac{\mu_t\alpha_t\left(2\left(\boldsymbol{x}_{K(t)_t}^{\top}\boldsymbol{U}\boldsymbol{e}_t - \boldsymbol{x}^{*\top}\boldsymbol{U}\boldsymbol{e}_t\right) - \left(f_{max}(\boldsymbol{x}_t)-v\right)\right)}{1-\mu_t\boldsymbol{x}_{K(t)_t}^{\top}\boldsymbol{U}\boldsymbol{y}_t} \\ &\geq \frac{1}{2}\frac{\mu_t\alpha_t\left(2\left(\boldsymbol{x}_{K(t)_t}^{\top}\boldsymbol{U}\boldsymbol{e}_t - v\right) - \left(f_{max}(\boldsymbol{x}_t)-v\right)\right)}{1-\mu_t\boldsymbol{x}_{K(t)_t}^{\top}\boldsymbol{U}\boldsymbol{y}_t} \\ &= \frac{\mu_t\alpha_t\left(2\boldsymbol{x}_{K(t)_t}^{\top}\boldsymbol{U}\boldsymbol{e}_t - 2v + v - f_{max}(\boldsymbol{x}_t)\right)}{2} \end{aligned}$$

that is:

$$KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{K(t)_t}\right) - KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{1_{t+1}}\right) \geq \frac{1}{2}\mu_t\alpha_t(2f'(\boldsymbol{x}_t) - f_{max}(\boldsymbol{x}_t) - v) \geq 0$$

last inequality is true at the same condition of MWU. □

Now we prove the convergence to the minmax equilibrium for MWU and LMWU, with high probability.

Theorem 5.4. *Assume that the row player follows the MWU or LMWU algorithm with a non-increasing learning rate μ_t such that $\lim_{T\to\infty}\sum_{t=1}^{\top}\mu_t = \infty$ and there exists $t' \in \mathbb{N}$ with $\mu_{t'} \leq \frac{1}{3}$. If the column player plays PAUSE E-LRCA with ξ_1, ξ_2 s.t. $v \in (\xi_1, \xi_2)$ then there will be last round convergence to the minmax equilibrium with high probability in games where there exists a fully-mixed minmax equilibrium strategy $\boldsymbol{x}^*$ for the row player.*

Proof. Let $\boldsymbol{x}^*$ be a fully-mixed minmax equilibrium strategy of the row player. Since μ_t is a non-increasing learning rate, there exists t' such that $\mu_t \leq \frac{1}{3}$ for all $t \geq t'$. Following lemma 5.3 , for all $t \in \mathbb{N}$ such that $t \geq t'$, we have:

$$KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{K(t)_t}\right) - KL\left(\boldsymbol{x}^*\|\boldsymbol{x}_{1_{t+1}}\right) \geq \frac{1}{2}\mu_t\alpha_t\left(2f'\left(\boldsymbol{x}_t\right) - f_{max}\left(\boldsymbol{x}_t\right) - v\right) \quad \forall t \in \mathbb{N}: \quad t \geq t'$$

That is, for stability property (definition 2.2.6.2) of the row player No-Regret algorithm, the same as stating:

$$KL\left(\boldsymbol{x}^{*}\|\boldsymbol{x}_{t}\right)-KL\left(\boldsymbol{x}^{*}\|\boldsymbol{x}_{t+1}\right)\geq\frac{1}{2}\mu_{t}\alpha_{t}\left(2f'\left(\boldsymbol{x}_{t}\right)-f_{max}\left(\boldsymbol{x}_{t}\right)-v\right)\quad\forall t\in\mathbb{N}:\quad t\geq t' \tag{5.20}$$

Thus, the sequence $KL\left(\boldsymbol{x}^{*}\|\boldsymbol{x}_{t}\right)$ is non-increasing with high probability, given a right choice of $K(t)$. As the sequence is bounded below by 0 , it has a limit for the minmax equilibrium strategy $\boldsymbol{x}^{*}$. Since t' is a finite number and $\sum_{t=1}^{\infty}\mu_{t}=\infty$, we have $\sum_{t=t'}^{\infty}\mu_{t}=\infty$. Thus:

$$\lim_{T\to\infty}\sum_{k=t'}^{T}\mu_{t}=\infty$$

We will prove that $\forall\epsilon>0,\exists h\in\mathbb{N}$ such that following PAUSE E-LRCA for the column player and MWU or LMWU algorithm for the row player, the row player will play strategy $\boldsymbol{x}_{h}$ at round h and $f\left(\boldsymbol{x}_{h}\right)-v\leq\epsilon$ with high probability.

We will proceed to find a contradiction. Thus, suppose that $\exists\epsilon>0$ such that $\forall h\in\mathbb{N},\ \ f\left(\boldsymbol{x}_{h}\right)-v>\epsilon$. Then $\forall t\in\mathbb{N}$

$$\alpha_{t}\left(f\left(\boldsymbol{x}_{t}\right)-v\right)\geq\gamma_{t}\frac{\left(f\left(\boldsymbol{x}_{t}\right)-v\right)^{2}}{\beta}>\gamma_{t}\frac{\epsilon^{2}}{\beta}\geq\gamma_{min}\frac{\epsilon^{2}}{\beta}>0\quad whp$$

Let t vary from t' to T in equation 5.20. By summing over t, we obtain:

$$\begin{aligned}KL\left(\boldsymbol{x}^{*}\|\boldsymbol{x}_{T}\right)&\leq KL\left(\boldsymbol{x}^{*}\|\boldsymbol{x}_{t'}\right)-\frac{1}{2}\sum_{t=t'}^{T}\mu_{t}\alpha_{t}\left(2f'\left(\boldsymbol{x}_{t}\right)-f_{max}\left(\boldsymbol{x}_{t}\right)-v\right)\\&\leq KL\left(\boldsymbol{x}^{*}\|\boldsymbol{x}_{t'}\right)-\frac{1}{2}\gamma_{min}\frac{\epsilon}{\beta}\sum_{t=t'}^{T}\mu_{t}\left(\epsilon-2\rho^{K(t)}-\sigma^{K(t)}\right)\end{aligned}$$

Since $\lim_{T\to\infty}\sum_{t=t'}^{T}\mu_{t}\left(\epsilon-2\rho^{K(t)}-\sigma^{K(t)}\right)=\infty$ (both $\rho^{K(t)}$ and $\sigma^{K(t)}$ converge to 0) and $KL\left(\boldsymbol{x}^{*}\|\boldsymbol{x}_{T}\right)\geq 0$, which contradicts our assumption about $\forall h\in\mathbb{N},\ \ f\left(\boldsymbol{x}_{h}\right)-v>\epsilon$.
From this point on, the proof is different for the two versions of the algorithm.

In case of the last round convergence version with non-decreasing KL (algorithm 5.2), the constraint $2f'(\boldsymbol{x}_{t})-f_{max}(\boldsymbol{x}_{t})\geq v$ that is equivalent to $\epsilon\geq 2\rho^{K(t)}+\sigma^{K(t)}$ is satisfied at each round with high probability. Thus we take a sequence of $\epsilon_{k}>0$ such that $\lim_{k\to\infty}\epsilon_{k}=0$. Then for each k, there exists $\boldsymbol{x}_{t_{k}}\in\Delta_{n}$ such that $v\leq f\left(\boldsymbol{x}_{t_{k}}\right)\leq v+\epsilon_{k}$. As Δ_{n} is a compact set and $\boldsymbol{x}_{t_{k}}$ is bounded then following the Bolzano-Weierstrass theorem, there is a convergence subsequence $x_{\bar{t}_{k}}$. The limit of that sequence, $\boldsymbol{x}^{*}$, is the minmax equilibrium

strategy of the row player (since $f(\boldsymbol{x}^*) = f(\lim_{k\to\infty} \boldsymbol{x}_{\bar{t}_k}) = \lim_{k\to\infty} f(\boldsymbol{x}_{\bar{t}_k}) = v$). Combining with the fact that $KL(\boldsymbol{x}^*\|\boldsymbol{x}_t)$ is non-increasing for $t \geq t'$ and $KL(\boldsymbol{x}^*\|\boldsymbol{x}^*) = 0$, we have $\lim_{t\to\infty} KL(\boldsymbol{x}^*\|\boldsymbol{x}_t) = 0$, which concludes the proof.

In the case of the standard version (algorithm 5.1), we proceed in a similar way, that is, we still take a sequence of $\epsilon_k > 0$ such that $\lim_{k\to\infty} \epsilon_k = 0$. Then for each k, there exists $\boldsymbol{x}_{t_k} \in \Delta_n$ such that $v \leq f(\boldsymbol{x}_{t_k}) \leq v + \epsilon_k$. Again, there is a convergence subsequence $\boldsymbol{x}_{\bar{t}_k}$. The limit of that sequence, $\boldsymbol{x}^*$, is the minmax equilibrium strategy of the row player. In this new version of the algorithm we have no guarantee that the KL is not increasing, but we know that as $t \to \infty$ the monotonicity of $KL(\boldsymbol{x}^*\|\boldsymbol{x}_t)$ is guaranteed for $\boldsymbol{x}_t$ that are nearer to the equilibrium with high probability which implies $\lim_{t\to\infty} KL(\boldsymbol{x}^*\|\boldsymbol{x}_t) = 0$. □

We conclude the section with the result of last round convergence with high probability against OMD.

Theorem 5.5. *Assume that the row player follows OMD (FTRL) with σ-strongly convex distance generating function $F(\boldsymbol{x})$, with fixed learning rate such that $\mu \leq 1$ and $\sigma \geq 1$ and that there exists a fully-mixed minmax equilibrium strategy for the row player. Then if the column player follows the Algorithm PAUSE E-LRCA (algorithm 5.1 and algorithm 5.2) with $\beta \geq n^2$ and ξ_1, ξ_2 s.t. $v \in (\xi_1, \xi_2)$, there will be last round convergence to the minmax equilibrium with high probability.*

Proof. Let $\boldsymbol{x}^*$ be a minmax equilibrium of the row player. Denote by $B_F(\boldsymbol{x}_t||\boldsymbol{z}_t)$ the Bregman divergence between the current row player's strategy and lazy update of OMD (algorithm 2.8); then define $D_t(\boldsymbol{x}^*) := (B_F(\boldsymbol{x}^*||\boldsymbol{z}_t) - B_F(\boldsymbol{x}_t||\boldsymbol{z}_t))\frac{1}{\mu}$, following properties of strongly convex function we have:

$$
\begin{aligned}
D_t(\boldsymbol{x}^*) &= \big(F(\boldsymbol{x}^*) - F(\boldsymbol{z}_t) - \nabla F(\boldsymbol{z}_t)^\top(\boldsymbol{x}^* - \boldsymbol{z}_t) - \big(F(\boldsymbol{x}_t) - F(\boldsymbol{z}_t) - \nabla F(\boldsymbol{z}_t)^\top(\boldsymbol{x}_t - \boldsymbol{z}_t)\big)\big)\frac{1}{\mu} \\
&= \frac{1}{\mu}F(\boldsymbol{x}^*) - \frac{1}{\mu}F(\boldsymbol{z}_t) + (\boldsymbol{x}^* - \boldsymbol{z}_t)^\top \sum_{k=1}^{t-1} \boldsymbol{U}\boldsymbol{y}_k - \left(\frac{1}{\mu}F(\boldsymbol{x}_t) - \frac{1}{\mu}F(\boldsymbol{z}_t) + (\boldsymbol{x}_t - \boldsymbol{z}_t)^\top \sum_{k=1}^{t-1} \boldsymbol{U}\boldsymbol{y}_k\right) \\
&= \frac{1}{\mu}F(\boldsymbol{x}^*) - \frac{1}{\mu}F(\boldsymbol{x}_t) + (\boldsymbol{x}^* - \boldsymbol{x}_t)^\top \sum_{k=1}^{t-1} \boldsymbol{U}\boldsymbol{y}_k \\
&\geq \frac{\sigma}{2\mu}\|\boldsymbol{x}^* - \boldsymbol{x}_t\|^2
\end{aligned}
$$

Thus, if $D_t(\boldsymbol{x}^*)$ converges to 0 then we have $\boldsymbol{x}_t$ converges to $\boldsymbol{x}^*$. We will prove that

$$
D_t(\boldsymbol{x}^*) - D_{t+1}(\boldsymbol{x}^*) \geq \gamma_t \frac{(f_{max}(\boldsymbol{x}_t) - v)(2f'(\boldsymbol{x}_t) - f_{max}(\boldsymbol{x}_t) - v)}{2n^2} \quad \forall t
$$

From the definition of $D_t(\boldsymbol{x})$ we have:

$$\begin{aligned}
&D_{K(t)_t}(\boldsymbol{x}^*) - D_{1_{t+1}}(\boldsymbol{x}^*) \\
&= \left(D_{K(t)_t}(\boldsymbol{x}_{1_{t+1}}) + \boldsymbol{x}_{1_{t+1}}^\top \boldsymbol{U}\left(\boldsymbol{y}_{K(t)_t} + \boldsymbol{y}_t\right)\right) - \boldsymbol{x}^{*\top}\boldsymbol{U}\left(\boldsymbol{y}_{K(t)_t} + \boldsymbol{y}_t\right) \\
&\geq \frac{\sigma}{2\mu}\left\|\boldsymbol{x}_{1_{t+1}} - \boldsymbol{x}_{K(t)_t}\right\|^2 + \boldsymbol{x}_{1_{t+1}}^\top \boldsymbol{U}\left(\boldsymbol{y}_{K(t)_t} + \boldsymbol{y}_t\right) - \boldsymbol{x}^{*\top}\boldsymbol{U}\left(\boldsymbol{y}_{K(t)_t} + \boldsymbol{y}_t\right) \qquad (5.21a)
\end{aligned}$$

as usual we take 1_{t+1} as the first round in which we play the equilibrium at time $t+1$ and $K(t)_t$ as the last time in which we play the equilibrium at time t. From definition of minmax equilibrium in zero-sum games we have $\boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{y}^* \geq \boldsymbol{x}^{*\top}\boldsymbol{U}\boldsymbol{y}^* = v \;\; \forall \boldsymbol{x} \in \Delta_n$. Thus, given $\boldsymbol{y}_{K(t)_t} = \boldsymbol{y}^*$, we obtain:

$$\begin{aligned}
&D_{K(t)_t}(\boldsymbol{x}^*) - D_{1_{t+1}}(\boldsymbol{x}^*) \geq \frac{\sigma}{2\mu}\left\|\boldsymbol{x}_{1_{t+1}} - \boldsymbol{x}_{K(t)_t}\right\|^2 + \left(\boldsymbol{x}_{1_{t+1}} - \boldsymbol{x}^*\right)^\top \boldsymbol{U}\boldsymbol{y}_t \\
&= \frac{\sigma}{2\mu}\left\|\boldsymbol{x}_{1_{t+1}} - \boldsymbol{x}_{K(t)_t}\right\|^2 + \left(\boldsymbol{x}_{1_{t+1}} - \boldsymbol{x}^*\right)^\top \boldsymbol{U}\left(\left(1-\alpha_t\right)\boldsymbol{y}^* + \alpha_t \boldsymbol{e}_t\right) \qquad (5.22a) \\
&\geq \frac{\sigma}{2\mu}\left\|\boldsymbol{x}_{1_{t+1}} - \boldsymbol{x}_{K(t)_t}\right\|^2 + \alpha_t\left(\boldsymbol{x}_{1_{t+1}} - \boldsymbol{x}^*\right)^\top \boldsymbol{U}\boldsymbol{e}_t \qquad (5.22b) \\
&= \frac{\sigma}{2\mu}\left\|\boldsymbol{x}_{1_{t+1}} - \boldsymbol{x}_{K(t)_t}\right\|^2 + \alpha_t\left(\boldsymbol{x}_{1_{t+1}} - \boldsymbol{x}_{K(t)_t}\right)^\top \boldsymbol{U}\boldsymbol{e}_t + \alpha_t\left(\boldsymbol{x}_{K(t)_t} - \boldsymbol{x}^*\right)^\top \boldsymbol{U}\boldsymbol{e}_t \\
&\geq \frac{\sigma}{2\mu}\left\|\boldsymbol{x}_{1_{t+1}} - \boldsymbol{x}_{K(t)_t}\right\|^2 - \alpha_t\left\|\boldsymbol{x}_{1_{t+1}} - \boldsymbol{x}_{K(t)_t}\right\|\left\|\boldsymbol{U}\boldsymbol{e}_t\right\|_* + \alpha_t\left(f'(\boldsymbol{x}_t) - v\right) \qquad (5.22c)
\end{aligned}$$

Inequalities 5.22a and 5.22b come from the update rule of PAUSE E-LRCA, while inequality 5.22c comes from the definition of dual norm. Bounding the dual norm with the dimension of the vector in inequality 5.22c we obtain:

$$\begin{aligned}
&D_{K(t)_t}(\boldsymbol{x}^*) - D_{1_{t+1}}(\boldsymbol{x}^*) \\
&\geq \frac{\sigma}{2\mu}\left\|\boldsymbol{x}_{1_{t+1}} - \boldsymbol{x}_{K(t)_t}\right\|^2 - n\alpha_t\left\|\boldsymbol{x}_{1_{t+1}} - \boldsymbol{x}_{K(t)_t}\right\| + \alpha_t\left(f'(\boldsymbol{x}_t) - v\right) \\
&= \left(\sqrt{\frac{\sigma}{2\mu}}\left\|\boldsymbol{x}_{1_{t+1}} - \boldsymbol{x}_{K(t)_t}\right\| - \frac{n\alpha_t}{2\sqrt{\frac{\sigma}{2\mu}}}\right)^2 + \alpha_t\left(f'(\boldsymbol{x}_t) - v\right) - \frac{n^2\alpha_t^2\mu}{2\sigma} \\
&\geq \alpha_t\left(f'(\boldsymbol{x}_t) - v\right) - \frac{n^2\alpha_t^2\mu}{2\sigma} \geq \alpha_t\left(f'(\boldsymbol{x}_t) - v\right) - \frac{n^2\alpha_t^2}{2} \qquad (5.23a)
\end{aligned}$$

Now, from PAUSE E-LRCA (algorithm 5.1 and algorithm 5.2) we have:

$$\alpha_t = \gamma_t \frac{f_{max}(\boldsymbol{x}_t) - v}{n^2}$$

then inequality 5.23a implies, as $\gamma_t^2 < \gamma_t$:

$$D_{K(t)_t}(\boldsymbol{x}^*) - D_{1_{t+1}}(\boldsymbol{x}^*) \geq \frac{\alpha_t}{2}\left(2f'(\boldsymbol{x}_t) - f_{max}(\boldsymbol{x}_t) - v\right)$$
$$= \gamma_t \frac{\left(f_{max}(\boldsymbol{x}_t) - v\right)\left(2f'(\boldsymbol{x}_t) - f_{max}(\boldsymbol{x}_t - v\right)}{2n^2} \quad \forall t$$

That for stability property (see definition 2.2.6.2) is:

$$D_t(\boldsymbol{x}^*) - D_{t+1}(\boldsymbol{x}^*) \geq \gamma_t \frac{\left(f_{max}(\boldsymbol{x}_t) - v\right)\left(2f'(\boldsymbol{x}_t) - f_{max}(\boldsymbol{x}_t) - v\right)}{2n^2} \quad \forall t$$

With the same reasoning of theorem 5.4, we obtain the last round convergence result with high probability. □

5.5. Regret

Before entering the details of this section we need to provide few definitions which have not been discussed in the preliminaries. First, we introduce a different version of the Dynamic Regret, the so called Dynamic Regret with respect to the Maxmin value, formally:

$$DR_T^{eq} := \sum_{t=1}^{T} |\boldsymbol{x}_t^\top \boldsymbol{U} \boldsymbol{y}_t - v| \tag{5.24}$$

where v is the maxmin (minmax) value of the game. From which:

Definition 5.5.0.1. *(No-Dynamic Regret with respect to the MaxMin) An algorithm is no-dynamic regret with respect to the MaxMin if* $\lim_{T\to\infty} \frac{DR_T^{eq}}{T} = 0$.

Then we define the so called Pompeiu-Hausdorff distance:

Definition 5.5.0.2. *(Pompeiu-Hausdorff Distance (Rockafellar and Wets, 1998)) Let* $(\mathbb{R}^n, d)$ *be a metric space, for* $C, D \subset \mathbb{R}^n$ *closed and nonempty, the Pompeiu-Hausdorff distance between* C *and* D *is the quantity:*

$$d_H(C, D) := \sup_{\boldsymbol{x}\in\mathbb{R}^n} |d_C(\boldsymbol{x}) - d_D(\boldsymbol{x})|$$

where $d_A(\boldsymbol{x}) := \inf_{\boldsymbol{y}\in A} d(\boldsymbol{x}, \boldsymbol{y})$.

To conclude we report a Lemma that will be useful in the computation of the final regret.

Lemma 5.4. *(Hausdorff distance convergence) If Devroye set (see lemma 3.1 and lemma*

3.2) is used to estimate opponent strategy with high probability and thus $d_H\left(\tilde{X}_t, \{\boldsymbol{x}_t\}\right) = \mathcal{O}\left(1/\sqrt{K(t)}\right)$*, we will have that*

$$\rho^{K(t)} + \sigma^{K(t)} = f_{max}(\boldsymbol{x}_t) - f'(\boldsymbol{x}_t) \leq 3d_H\left(\tilde{X}_t, \{\boldsymbol{x}_t\}\right)$$

from which :

$$2\rho^{K(t)} + \sigma^{K(t)} \leq 2\left(\rho^{K(t)} + \sigma^{K(t)}\right) \leq 6d_H\left(\tilde{X}_t, \{\boldsymbol{x}_t\}\right) \leq 30\sqrt{\frac{\ln(3/\delta)}{K(t)}}$$

Finally, we report the computation of the Dynamic Regret with respect to the value of the game.

Theorem 5.6. *Assume that the row player follows the above-mentioned no-regret type algorithms: MWU, LMWU, FTRL, OMD with constant learning rate* μ*; then by following PAUSE E-LRCA (algorithm 5.1) with* $K(t) = \ln(3/\delta)\, t^{\lambda}$ *and* $0 \leq \lambda < 2$*, the column player will achieve the no-dynamic regret (with respect to the MaxMin) property with the dynamic regret (with respect to the MaxMin) satisfying* $DR_T^{eq} = \mathcal{O}\left(\frac{n}{\sqrt{\mu\gamma_{min}}} T^{\max\left(-\lambda/2+1, \frac{1}{2}+\frac{-\lambda/2+1}{2}\right)}\right)$ *for OMD (FTRL) and* $DR_T^{eq} = \mathcal{O}\left(\frac{\sqrt{\log(n)}}{\sqrt{\mu\gamma_{min}}} T^{\max\left(-\lambda/2+1, \frac{1}{2}+\frac{-\lambda/2+1}{2}\right)}\right)$ *for MWU and LMWU, in games with fully-mixed minmax equilibrium strategy for the row player, with high probability.*

From the definition of Dynamic Regret with respect to the equilibrium we have:

$$\begin{aligned}
DR_T^{eq} &= \sum_{t=1}^{T} |\boldsymbol{x}_t^{\top} \boldsymbol{U} \boldsymbol{y}_t - v| \\
&\leq \sum_{t=1}^{T} max\{(f(\boldsymbol{x}_t) - f'(\boldsymbol{x}_t)), (f(\boldsymbol{x}_t) - v)\} \\
&\leq \sum_{t=1}^{T} max\{(f_{max}(\boldsymbol{x}_t) - f'(\boldsymbol{x}_t)), (f_{max}(\boldsymbol{x}_t) - v)\} \quad whp \\
&\leq \sum_{t=1}^{T} (f_{max}(\boldsymbol{x}_t) - f'(\boldsymbol{x}_t)) + \sum_{t=1}^{T} (f_{max}(\boldsymbol{x}_t) - v)
\end{aligned}$$

we start with the computation of the regret of the first term in last inequality bounding

(using result of lemma 5.4):

$$\sum_{t=1}^{T}(f_{max}(\boldsymbol{x}_t) - f'(\boldsymbol{x}_t)) = \sum_{t=1}^{T}\left(\sigma^{K(t)} + \rho^{K(t)}\right) \leq \sum_{t=1}^{T} 3d_H\left(\tilde{X}_t, \{\boldsymbol{x}_t\}\right)$$

The Hausdorff distance of our estimation of the row player is upper bounded by $2\frac{5}{2}\sqrt{\frac{\ln 3/\delta}{K(t)}}$ with $0 < \delta \leq 3\exp -4n/5$, from that:

$$\sum_{t=1}^{T}(f_{max}(\boldsymbol{x}_t) - f'(\boldsymbol{x}_t)) \leq \sum_{t=1}^{T}\frac{15\sqrt{\ln(3/\delta)}}{\sqrt{K(t)}} = \sum_{t=1}^{T}\frac{15}{\sqrt{t^{\lambda}}}$$

where the last inequalities are due to $K(t) = \ln(3/\delta)t^{\lambda}$. Bounding the summation with the integral:

$$\sum_{t=1}^{T}(f_{max}(\boldsymbol{x}_t) - f'(\boldsymbol{x}_t)) \leq 1 + \frac{15}{-\frac{\lambda}{2}+1}T^{-\lambda/2+1} = \mathcal{O}\left(T^{-\lambda/2+1}\right) \tag{5.25}$$

where last equality is true for $0 \leq \lambda < 2$.
To reason about $f_{max}(\boldsymbol{x}_t) - v$ (for MWU and LMWU) we follow lemma 5.3 result, from which we have (for $\beta \geq 2$):

$$\begin{aligned}
KL(\boldsymbol{x}^*\|\boldsymbol{x}_t) - KL(\boldsymbol{x}^*\|\boldsymbol{x}_{t+1}) &\geq \frac{1}{2}\mu_t\alpha_t\left(2f'(\boldsymbol{x}_t) - f_{max}(\boldsymbol{x}_t) - v\right) \\
&= \frac{1}{2}\mu_t\alpha_t\left(2f(\boldsymbol{x}_t) - 2\rho^{K(t)} - f_{max}(\boldsymbol{x}_t) - v\right) \\
&= \frac{1}{2}\mu_t\alpha_t\left(2f_{max}(\boldsymbol{x}_t) - 2\sigma^{K(t)} - 2\rho^{K(t)} - f_{max}(\boldsymbol{x}_t) - v\right) \\
&= \frac{1}{4}\mu_t\gamma_t\left(f_{max}(\boldsymbol{x}_t) - v\right)\left(f_{max}(\boldsymbol{x}_t) - v - 2\sigma^{K(t)} - 2\rho^{K(t)}\right) \\
&= \frac{1}{4}\mu_t\gamma_t\left((f_{max}(\boldsymbol{x}_t) - v)^2 - \left(2\rho^{K(t)} + 2\sigma^{K(t)}\right)(f_{max}(\boldsymbol{x}_t) - v)\right) \\
&\geq \frac{1}{4}\mu_t\gamma_t\left((f_{max}(\boldsymbol{x}_t) - v)^2 - \left(2\rho^{K(t)} + 2\sigma^{K(t)}\right)\right)
\end{aligned}$$

Let's now bound:

$$\sum_{t=1}^{T}(f_{max}(\boldsymbol{x}_t) - v)$$

thus:

$$\sum_{t=1}^{T}\frac{1}{4}\mu_t\gamma_t\left((f_{max}(\boldsymbol{x}_t) - v)^2 - \left(2\rho^{K(t)} + 2\sigma^{K(t)}\right)\right) \leq KL(\boldsymbol{x}^*\|\boldsymbol{x}_1)$$

$$\sum_{t=1}^{T} \mu_t \gamma_t \left(\left(f_{max}\left(\boldsymbol{x}_t\right) - v \right)^2 - \left(2\rho^{K(t)} + 2\sigma^{K(t)} \right) \right) \leq 4KL\left(\boldsymbol{x}^* \| \boldsymbol{x}_1\right)$$

for $n \geq 8$:

$$\sum_{t=1}^{T} \mu_t \gamma_t (f_{max}\left(\boldsymbol{x}_t\right) - v)^2 \leq 4\ln(n) + \sum_{t=1}^{T} \mu_t \gamma_t \left(2\rho^{K(t)} + 2\sigma^{K(t)} \right)$$

with Cauchy–Schwarz inequality:

$$\sum_{t=1}^{T} (f_{max}\left(\boldsymbol{x}_t\right) - v) \leq \sqrt{4\ln(n) + \sum_{t=1}^{T} \mu_t \gamma_t \left(2\rho^{K(t)} + 2\sigma^{K(t)} \right)} \sqrt{\sum_{t=1}^{T} \frac{1}{\mu_t \gamma_t}}$$

with fixed learning rate, using γ_{min} as found in lemma 5.1 and given $\gamma_t \leq 1$:

$$\begin{aligned} \sum_{t=1}^{T} (f_{max}\left(\boldsymbol{x}_t\right) - v) &\leq \sqrt{4\ln(n) + \sum_{t=1}^{T} \mu (2\rho^{K(t)} + 2\sigma^{K(t)})} \sqrt{\sum_{t=1}^{T} \frac{1}{\mu \gamma_{min}}} \\ &\leq \sqrt{4\ln(n) + \sum_{t=1}^{T} \mu (2\rho^{K(t)} + 2\sigma^{K(t)})} \frac{1}{\sqrt{\mu \gamma_{min}}} T^{\frac{1}{2}} \end{aligned}$$

Due to result on Hausdorff distance $d_H\left(\tilde{X}_t, \{\boldsymbol{x}_t\}\right)$ in lemma 5.4:

$$\sum_{t=1}^{T} (f_{max}\left(\boldsymbol{x}_t\right) - v) \leq \sqrt{4\ln(n) + \sum_{t=1}^{T} \mu \frac{30\sqrt{\ln\left(3/\delta\right)}}{\sqrt{K(t)}}} \frac{1}{\sqrt{\mu \gamma_{min}}} T^{\frac{1}{2}}$$

For the choice of $K(t) = \ln\left(3/\delta\right) t^{\lambda}$:

$$\begin{aligned} \sum_{t=1}^{T} (f_{max}\left(\boldsymbol{x}_t\right) - v) &\leq \sqrt{4\ln(n) + \sum_{t=1}^{T} \mu \frac{30}{\sqrt{t^{\lambda}}}} \frac{1}{\sqrt{\mu \gamma_{min}}} T^{\frac{1}{2}} \\ &\leq \sqrt{4\ln(n) + \sum_{t=1}^{T} \mu \frac{30}{t^{\lambda/2}}} \frac{1}{\sqrt{\mu \gamma_{min}}} T^{\frac{1}{2}} \\ &\leq \sqrt{4\ln(n) + 1 + \frac{30}{-\frac{\lambda}{2} + 1} \mu T^{-\frac{\lambda}{2}+1}} \frac{1}{\sqrt{\mu \gamma_{min}}} T^{\frac{1}{2}} \end{aligned}$$

given $\mu < 1/3$ as assumption of lemma 5.3:

$$\sum_{t=1}^{T}(f_{max}\left(\boldsymbol{x}_t\right) - v) = \mathcal{O}\left(\frac{\sqrt{\log(n)}}{\sqrt{\mu\gamma_{min}}}T^{\frac{1}{2}+\frac{-\lambda/2+1}{2}}\right) \tag{5.26}$$

Finally summing result 5.25 with result 5.26:

$$DR_T^{eq} \leq \mathcal{O}\left(T^{-\lambda/2+1}\right) + \mathcal{O}\left(\frac{\sqrt{\log(n)}}{\sqrt{\mu\gamma_{min}}}T^{\frac{1}{2}+\frac{-\lambda/2+1}{2}}\right) = \mathcal{O}\left(\frac{\sqrt{\log(n)}}{\sqrt{\mu\gamma_{min}}}T^{\max\left(-\lambda/2+1,\frac{1}{2}+\frac{-\lambda/2+1}{2}\right)}\right)$$

This means that for $K(t) = \ln\left(3/\delta\right)t$:

$$DR_T^{eq} = \mathcal{O}\left(\frac{\sqrt{\log(n)}}{\sqrt{\mu\gamma_{min}}}T^{3/4}\right)$$

while for $K(t) = \ln\left(3/\delta\right)t^{2/3}$:

$$DR_T^{eq} = \mathcal{O}\left(\frac{\sqrt{\log(n)}}{\sqrt{\mu\gamma_{min}}}T^{5/6}\right)$$

For the OMD (FTRL) case, we exploit the result of theorem 5.5 from which can be derived:

$$\begin{aligned} D_t\left(\boldsymbol{x}^*\right) - D_{t+1}\left(\boldsymbol{x}^*\right) &\geq \gamma_t\frac{\left(f_{max}\left(\boldsymbol{x}_t\right) - v\right)\left(2f'\left(\boldsymbol{x}_t\right) - f_{max}(\boldsymbol{x}_t) - v\right)}{2n^2} \\ &\geq \gamma_t\frac{\left(\left(f_{max}\left(\boldsymbol{x}_t\right) - v\right)^2 - \left(2\rho^{K(t)} + 2\sigma^{K(t)}\right)\right)}{2n^2} \end{aligned}$$

using γ_{min} as found in lemma 5.1:

$$\sum_{t=1}^{T}\left(\left(f_{max}\left(\boldsymbol{x}_t\right) - v\right)^2 - \left(2\rho^{K(t)} + 2\sigma^{K(t)}\right)\right) \leq \frac{2n^2}{\mu\gamma_{min}}$$

from which:

$$\sum_{t=1}^{T}(f_{max}\left(\boldsymbol{x}_t\right) - v)^2 \leq \frac{2n^2}{\mu\gamma_{min}} + \sum_{t=1}^{T}\left(2\rho^{K(t)} + 2\sigma^{K(t)}\right)$$

using Cauchy-Schwarz:

$$\sum_{t=1}^{T}(f_{max}\left(\boldsymbol{x}_t\right) - v) \leq \sqrt{\frac{2n^2}{\mu\gamma_{min}} + \sum_{t=1}^{T}(2\rho^{K(t)} + 2\sigma^{K(t)})}T^{\frac{1}{2}}$$

with the same reasoning of MWU:

$$DR_T^{eq} \leq \mathcal{O}\left(T^{-\lambda/2+1}\right) + \mathcal{O}\left(\frac{n}{\sqrt{\mu\gamma_{min}}} T^{\frac{1}{2}+\frac{-\lambda/2+1}{2}}\right) = \mathcal{O}\left(\frac{n}{\sqrt{\mu\gamma_{min}}} T^{\max\left(-\lambda/2+1, \frac{1}{2}+\frac{-\lambda/2+1}{2}\right)}\right)$$

which concludes the proof.

6 | Experiments

In this chapter we report experiments which show the theoretical guarantees presented previously. General remarks:

1. Experiments are run against MWU (2.6) with a fixed learning rate of $\mu = 1$.
2. As benchmark we coded the standard version of LRCA (algorithm 3.4); this will be useful to show the difference in terms of safety and regret.
3. As concerns the experiments with partial semi-bandit feedback, we implemented a not-safe version of the algorithm (namely, PAUSE LRCA) which will substitute standard LRCA as a benchmark.
4. Standard python libraries have been used; to solve the linear programs we chose Pulp, a good trade-off between user-friendliness and efficiency.
5. For every experiment, we will report a plot for the Regret of the column player, one for the KL (convergence measure) of the row player, and one for the Utility obtained by the column player (with the safety bounds). When experiments are based on small games we will report the simplex with the convergence of the row player strategy.

6.1. Expert Feedback with Fully-Mixed Equilibrium

In this section we report experiments on games with fully-mixed equilibrium and where both players have the expert feedback.

We start by Rock Paper Scissor scaled in $[0, 1]$; formally the payoff matrix (of the column player) is:

	Rock	Paper	Scissor
Rock	0.5	1	0
Paper	0	0.5	1
Scissor	1	0	0.5

where the value v is 0.5.

In figure 6.1 we note that safety property leads to a deceleration of the convergence, as it was clear from the Dynamic Regret result in section 4.5; indeed the regret of LRCA is smaller than the E-LRCA one.

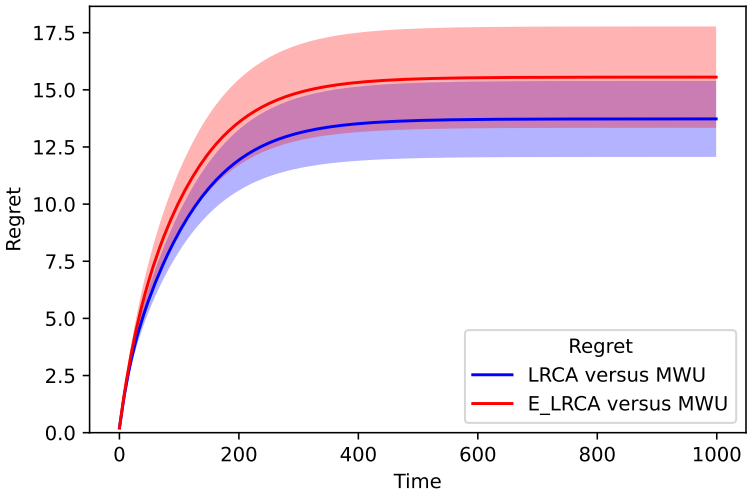

Figure 6.1: Dynamic Regret of the column player in Rock Paper Scissor game

In figure 6.2 it is shown how the safety constraints are never violated by E-LRCA, differently from the not-safe version of the algorithm.

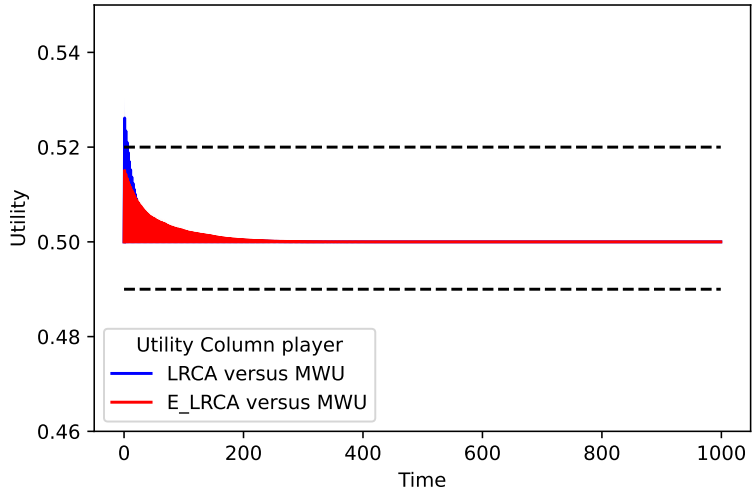

Figure 6.2: Utility of the column player in Rock Paper Scissor game with the safety bounds

Again, in figure 6.3 we see how the convergence is slower due to safety.

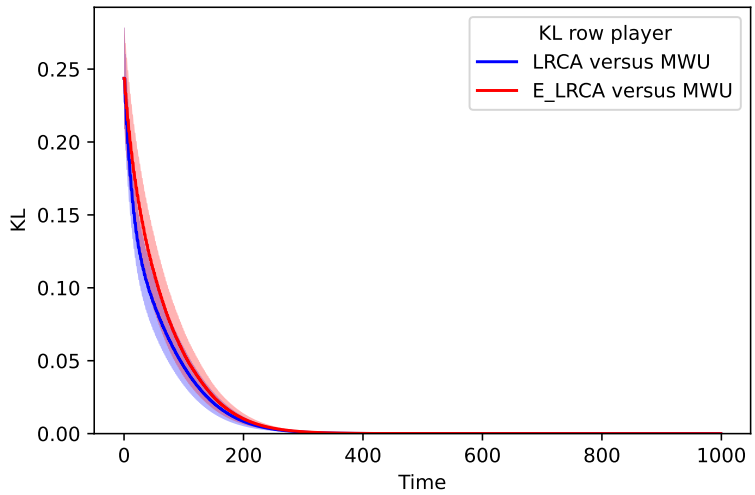

Figure 6.3: KL of the row player in Rock Paper Scissor game

Finally, in figure 6.4 we report the dynamic of the row player's strategy in the simplex; indeed, it converges to the equilibrium $\boldsymbol{x}_t = [1/3, 1/3, 1/3]$

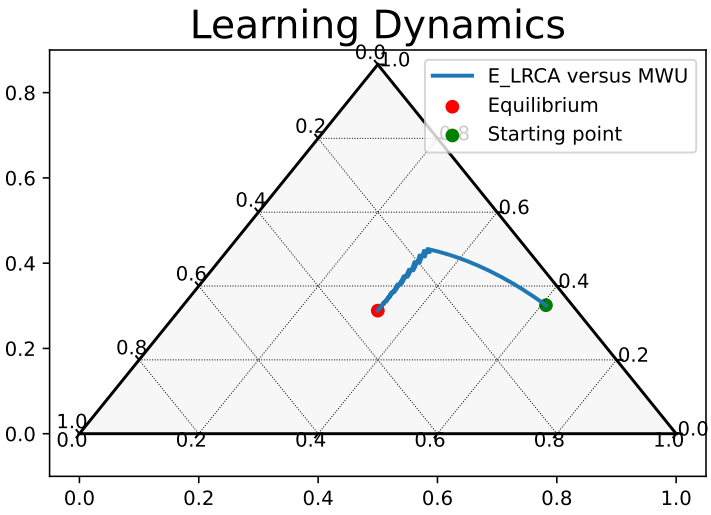

Figure 6.4: Row player's strategy in Rock Paper Scissor game

To conclude this section, we report results obtained in a skewed matching pennies game, namely, a version of the game where the equilibrium is not $\boldsymbol{x}^* = [1/2, 1/2]$. Formally:

	A	B
A	0	2/5
B	3/5	0

where the value v is $6/25$.

Since the safety bounds are tighter with respect to the previous examples, the result in terms of regret will seem worse.

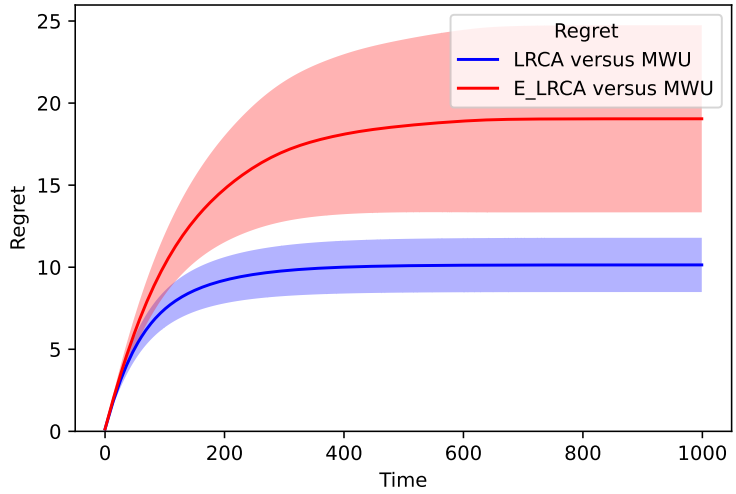

Figure 6.5: Dynamic Regret of the column player in the skewed matching pennies game

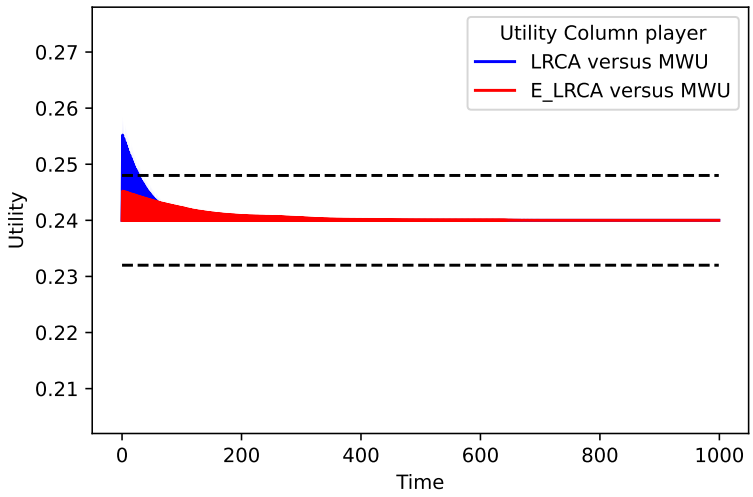

Figure 6.6: Utility of the column player in the skewed matching pennies game with the safety bounds

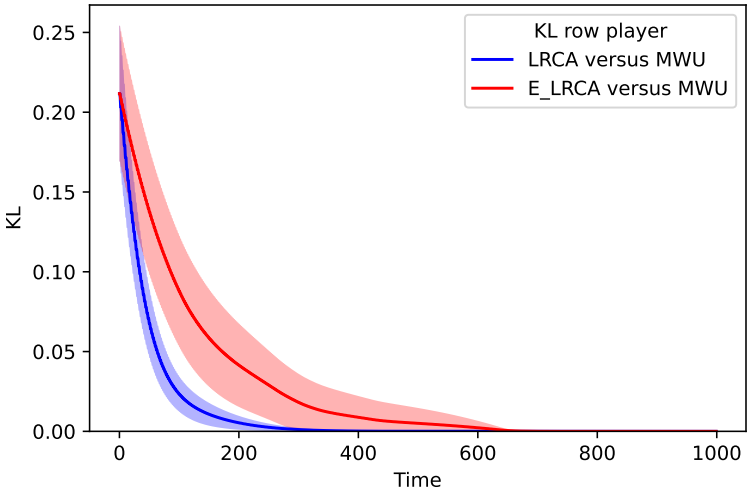

Figure 6.7: KL of the row player in the skewed matching pennies game

6.2. Partial Semi-Bandit Feedback with Fully-Mixed Equilibrium

In this section we report the experiments related to the partial semi-bandit feedback in games with fully-mixed equilibrium. We start by the result for PAUSE E-LRCA with linear $K(t)$.

In figure 6.8 we observe that the Dynamic Regret with respect to the equilibrium is better in the safe version of the algorithm; it is reasonable as we perform actions that are nearer to the equilibrium, which means that the utility (in the first rounds) is close to the value of the game v.

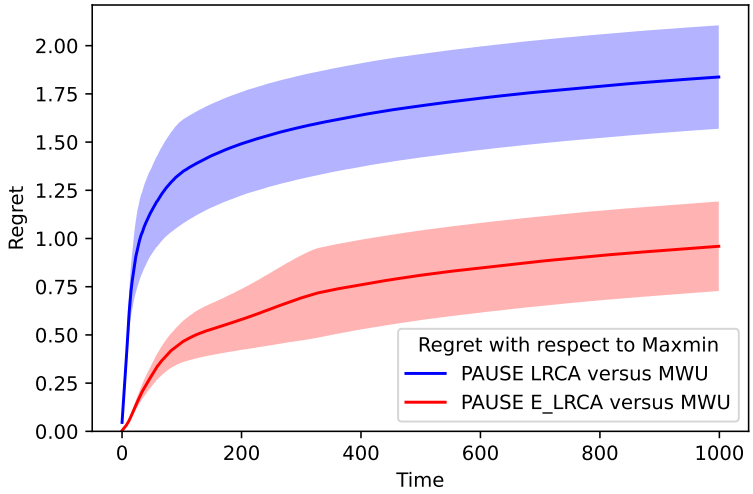

Figure 6.8: Dynamic Regret with respect to the MaxMin value of the column player in Rock Paper Scissor game

In figure 6.9 we notice that the utility obtained by the column player (it is clearer for the non safe version of the algorithm) can be smaller than the value of the game; this never happens when both players have expert feedback as the opponent strategy is always well estimated.

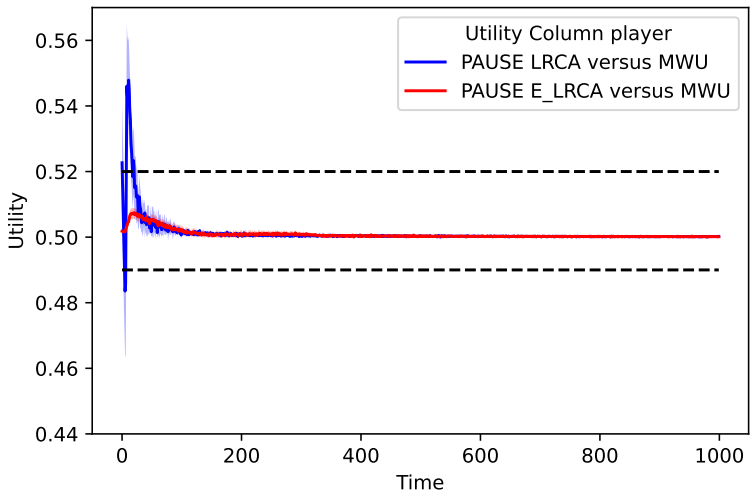

Figure 6.9: Expected Utility of the column player in Rock Paper Scissor game with the safety bounds

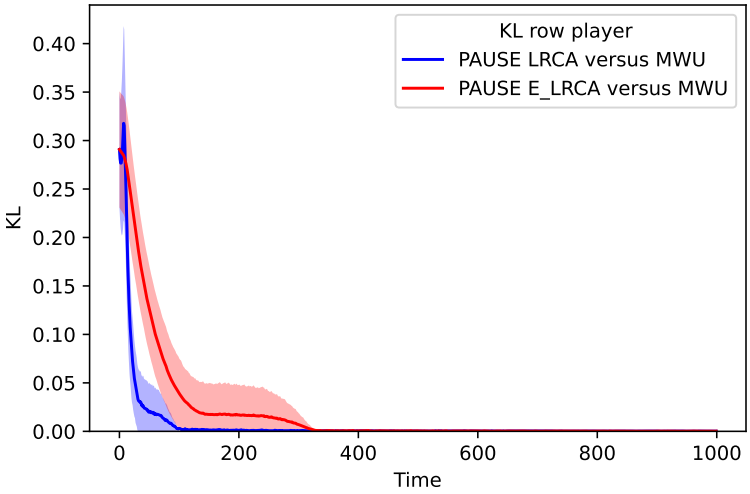

Figure 6.10: KL of the row player in Rock Paper Scissor game

Finally, in figure 6.11, it is shown that algorithm 5.1 does not guarantee a round by round improvement in terms of convergence, which was clear by the final inequality in lemma 5.3.

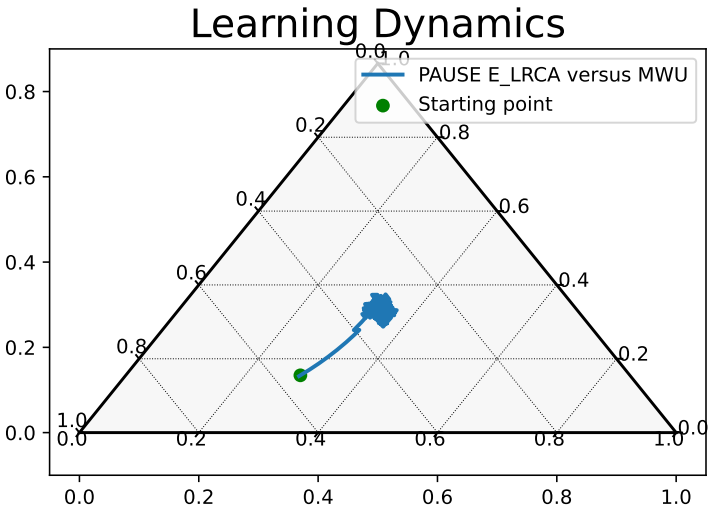

Figure 6.11: Row player's strategy in Rock Paper Scissor game

We show the results for PAUSE E-LRCA with $K(t) = \ln \frac{3}{\delta} t^{2/3}$. We observe that the regret of the safe version is smaller than in linear $K(t)$ case (for the first rounds); it is again reasonable as the algorithm has a worse estimation of the opponent strategy, which will lead PAUSE E-LRCA to play near the equilibrium.

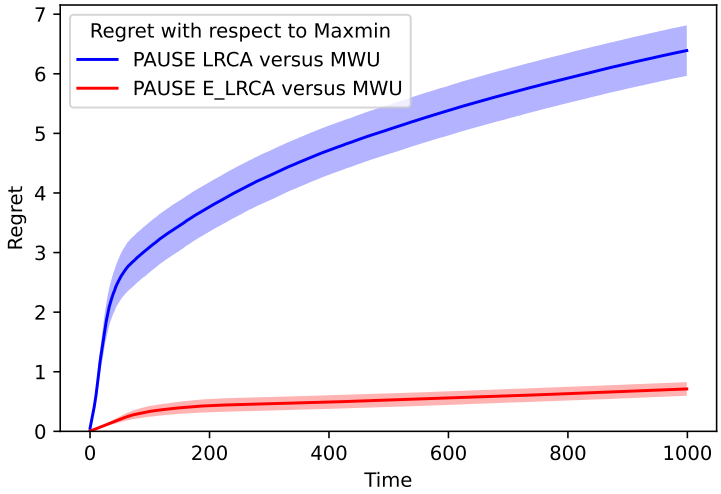

Figure 6.12: Dynamic Regret with respect to the MaxMin value of the column player in Rock Paper Scissor game

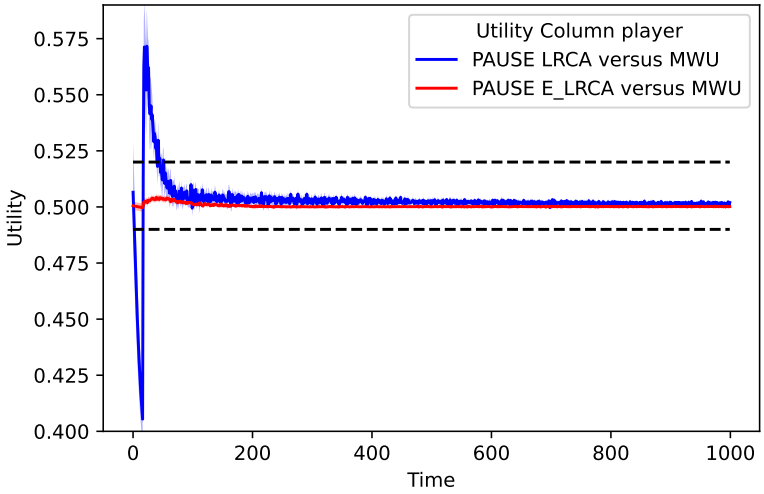

Figure 6.13: Expected Utility of the column player in Rock Paper Scissor game with the safety bounds

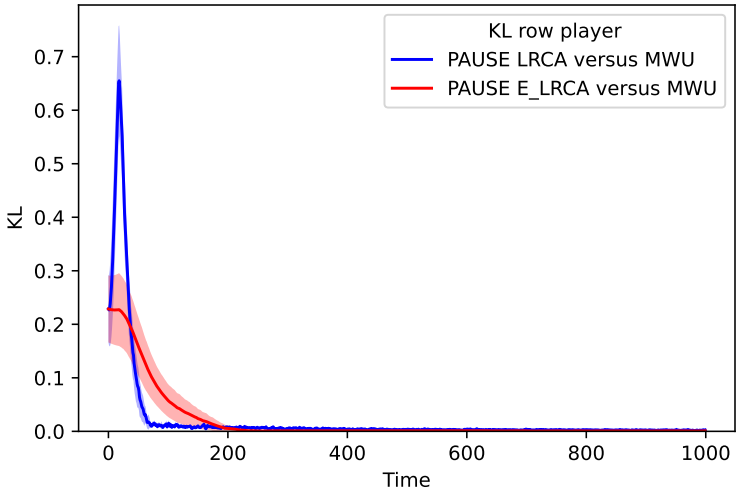

Figure 6.14: KL of the row player in Rock Paper Scissor game

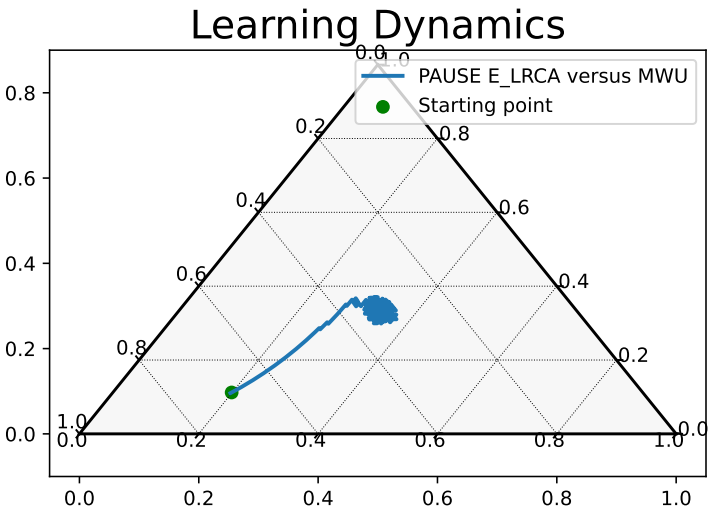

Figure 6.15: Row player's strategy in Rock Paper Scissor game

We proceed with the results in different games (for linear $K(t)$):

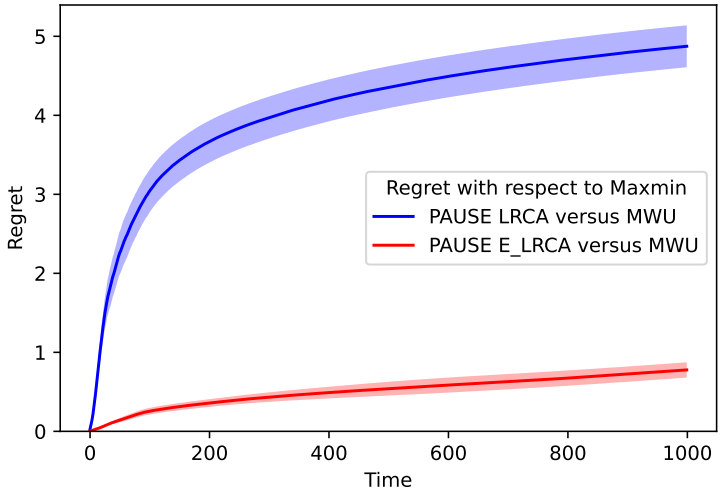

Figure 6.16: Dynamic Regret with respect to the MaxMin value of the column player in the bigger version of Rock Paper Scissor game

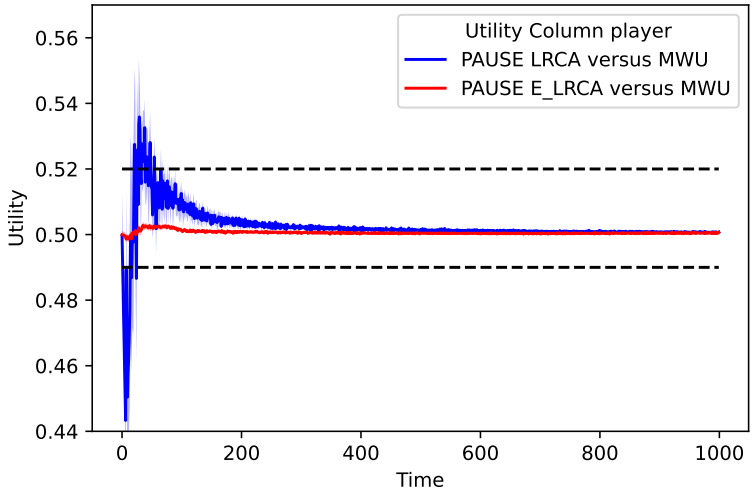

Figure 6.17: Expected Utility of the column player in the bigger version Rock Paper Scissor game with the safety bounds

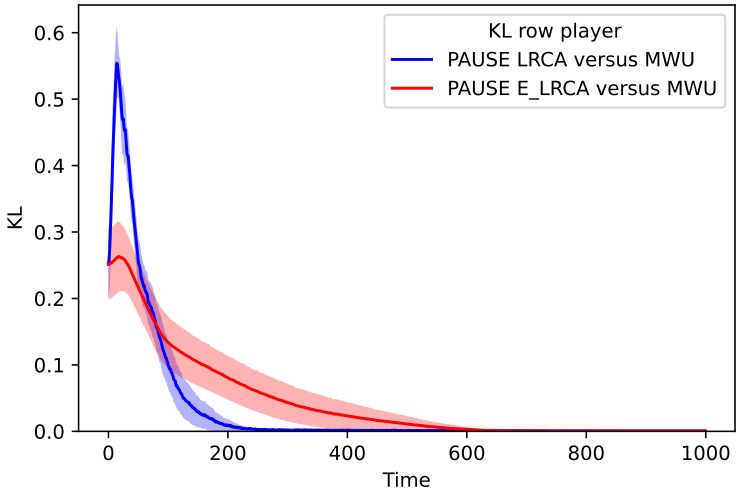

Figure 6.18: KL of the row player in the bigger version Rock Paper Scissor game

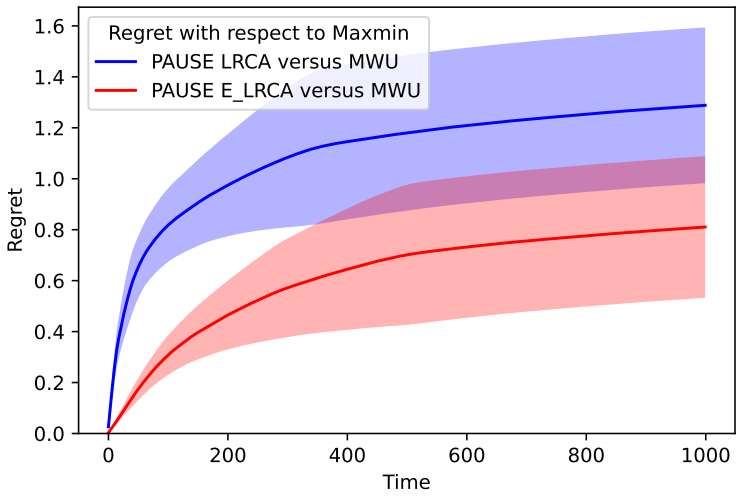

Figure 6.19: Dynamic Regret with respect to the MaxMin value of the column player in the skewed matching pennies game

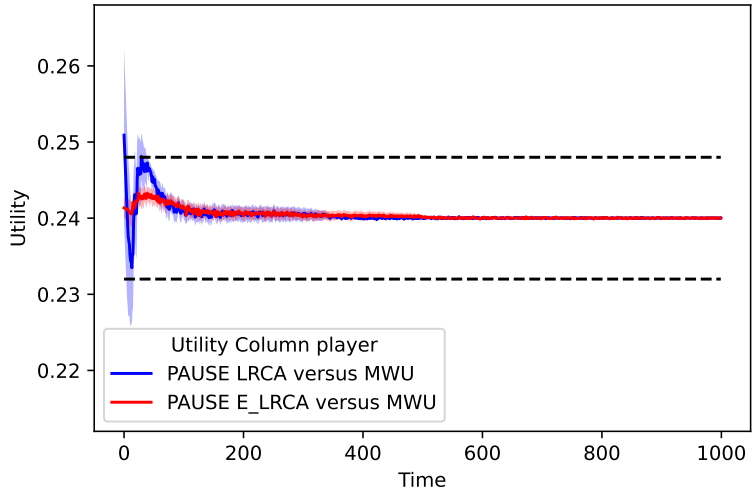

Figure 6.20: Expected Utility of the column player in the skewed matching pennies game with the safety bounds

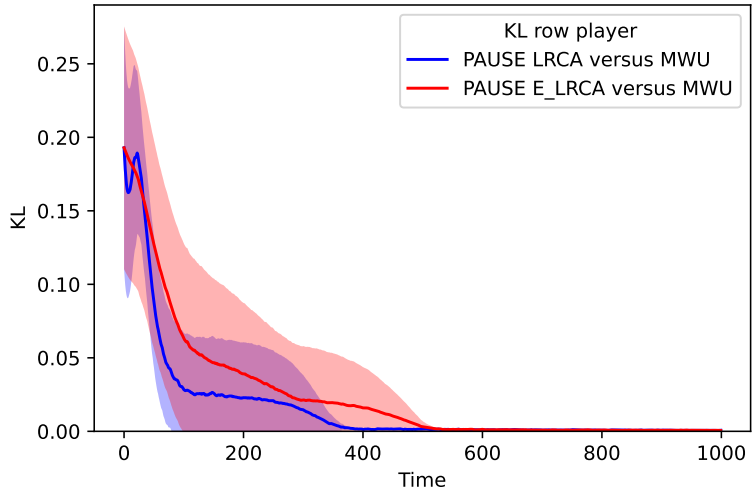

Figure 6.21: KL of the row player in the skewed matching pennies game

To conclude we show the results for algorithm 5.2, namely, the PAUSE E-LRCA version in which the KL decreases round after round. The reader will see that, in terms of convergence, the plots are similar to the ones where both players have Expert Feedback.

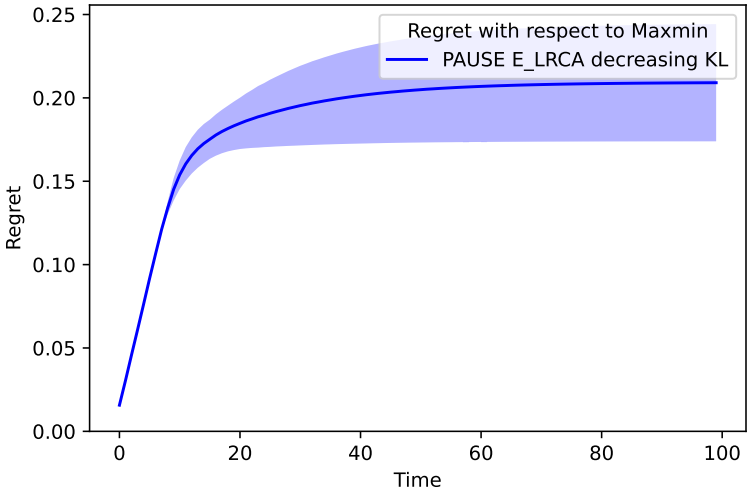

Figure 6.22: Dynamic Regret with respect to the MaxMin value of the column player in Rock Paper Scissor game

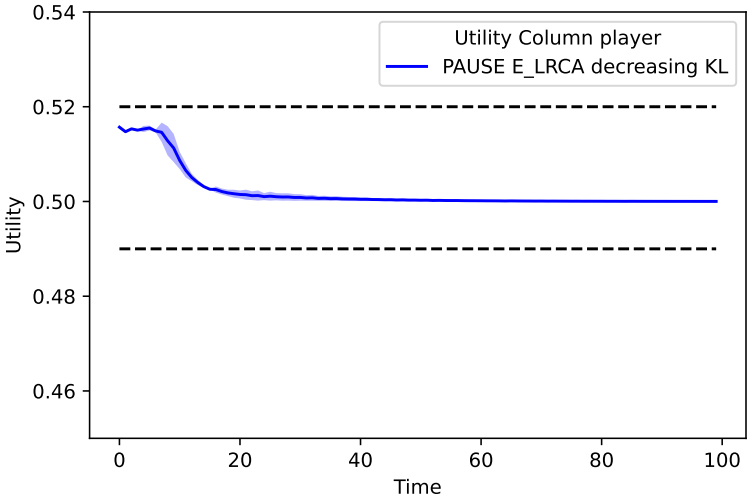

Figure 6.23: Expected Utility of the column player in Rock Paper Scissor game with the safety bounds

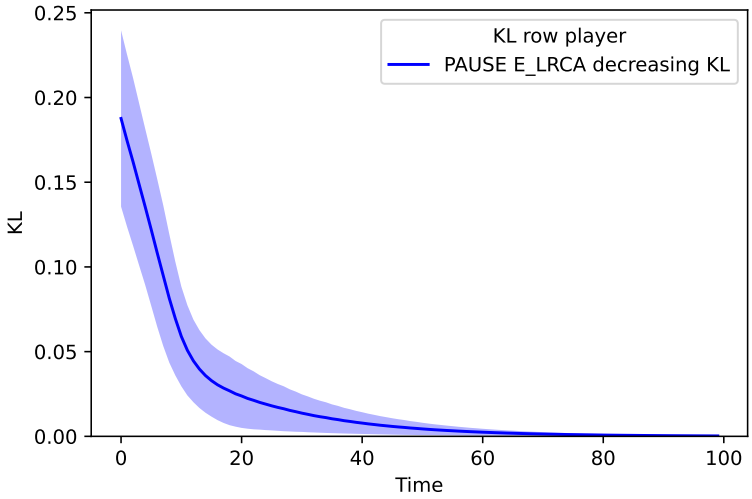

Figure 6.24: KL of the row player in Rock Paper Scissor game

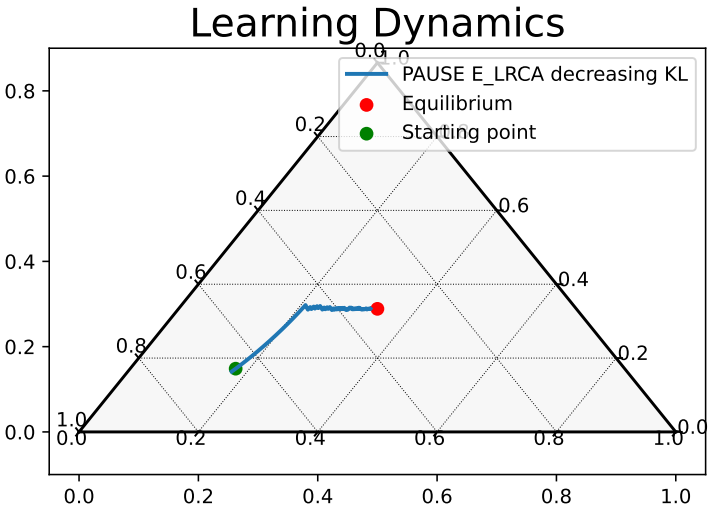

Figure 6.25: Row player's strategy in Rock Paper Scissor game

6.3. Comparison between different Feedback with Fully-Mixed Equilibrium

In this section we highlight the difference in terms of Dynamic Regret between the algorithms with expert feedback and the ones with partial semi-bandit feedback. In figure 6.27 the pink regret is linear since PAUSE LRCA uses a fixed $K(t) = 1$ (kind of a greedy approach).

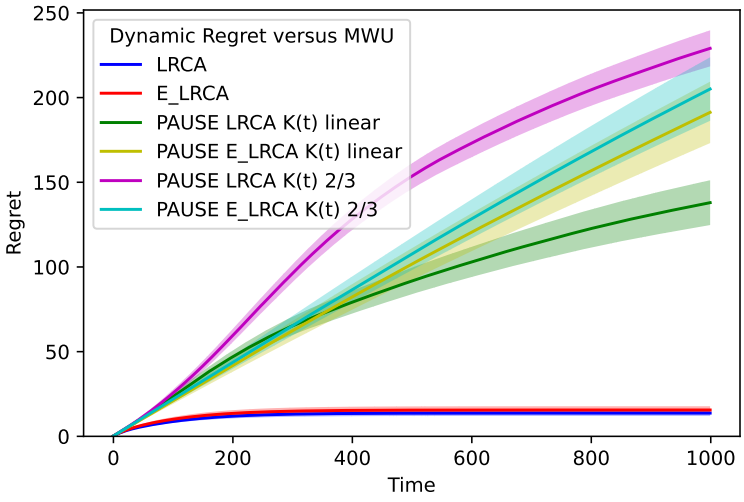

Figure 6.26: Dynamic Regret of the column player in Rock Paper Scissor game

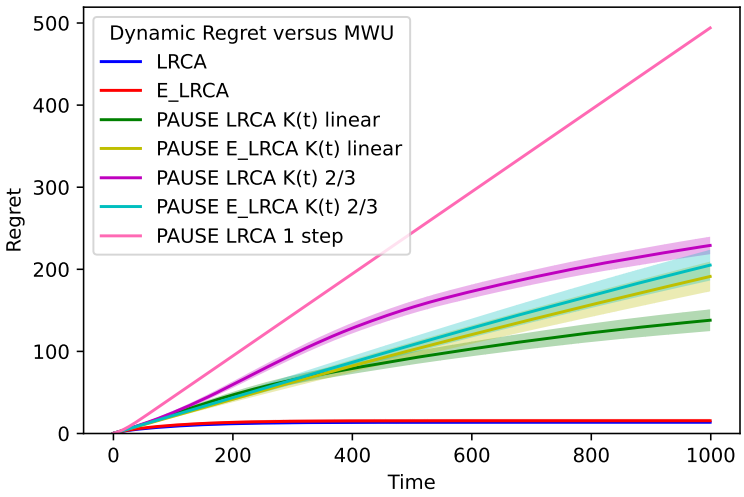

Figure 6.27: Dynamic Regret of the column player in Rock Paper Scissor game in comparison with Linear Regret

6.4. Expert Feedback with Partially-Mixed Equilibrium

In this section we show experiments with expert feedback in a game with partially-mixed equilibrium, specifically:

	A	B	C	D
A	0.25	0.75	0.3	0.32
B	0.75	0.25	0.3	0.32
C	0.501	0.501	0.94	0.05
D	0.502	0.502	0.044	0.94

where the value v is 0.5.

The convergence is decelerated not only by the choice of γ_{min} but also by the absence of fully-mixed equilibrium which forces the algorithm to play $\gamma_t = \gamma_{min}$ at each round.

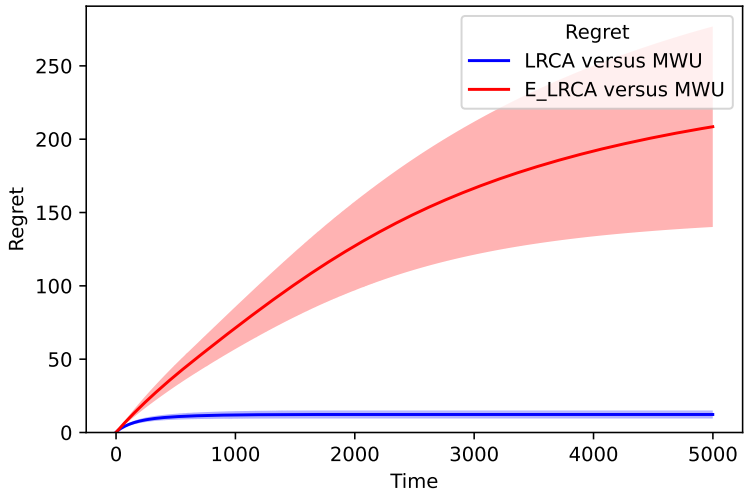

Figure 6.28: Dynamic Regret of the column player in game with a partially-mixed equilibrium

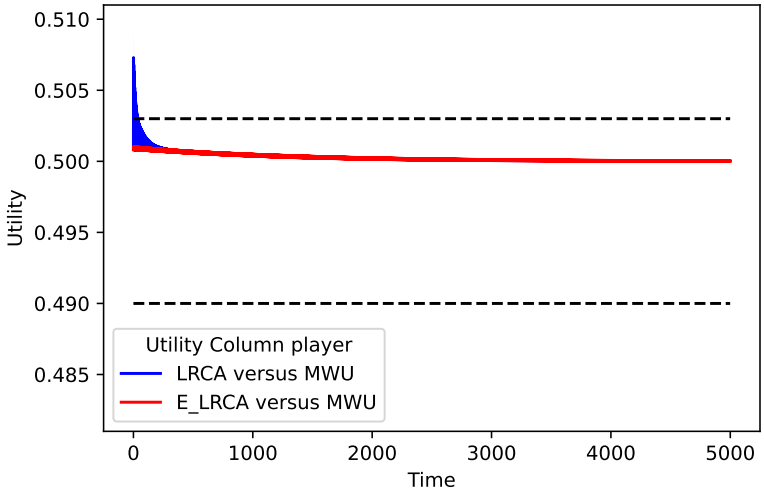

Figure 6.29: Utility of the column player in game with a partially-mixed equilibrium with the safety bounds

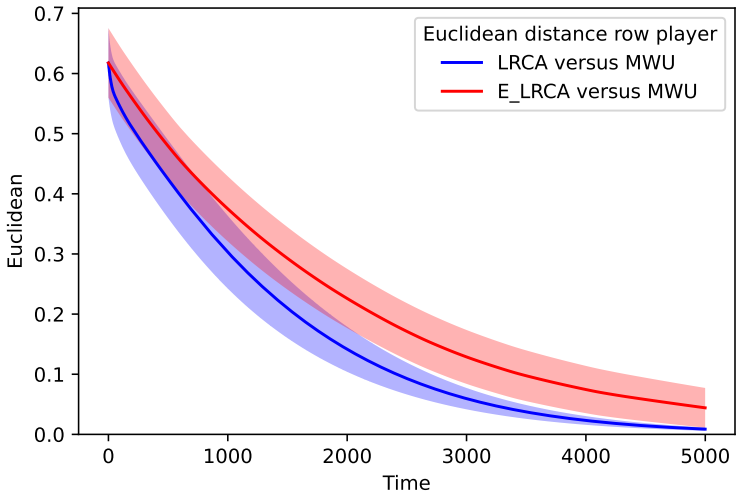

Figure 6.30: Euclidean distance between row player's strategy and the Equilibrium in game with a partially-mixed equilibrium

6.5. Partial Semi-Bandit Feedback with Partially-Mixed Equilibrium

In this last section we report the experimental results of PAUSE E-LRCA in games with partially-mixed equilibrium. We start by a game with four actions, that is:

	A	B	C	D
A	0.25	0.75	0.3	0.32
B	0.75	0.25	0.3	0.32
C	0.501	0.501	0.94	0.05
D	0.502	0.502	0.044	0.94

where the value v is 0.5.

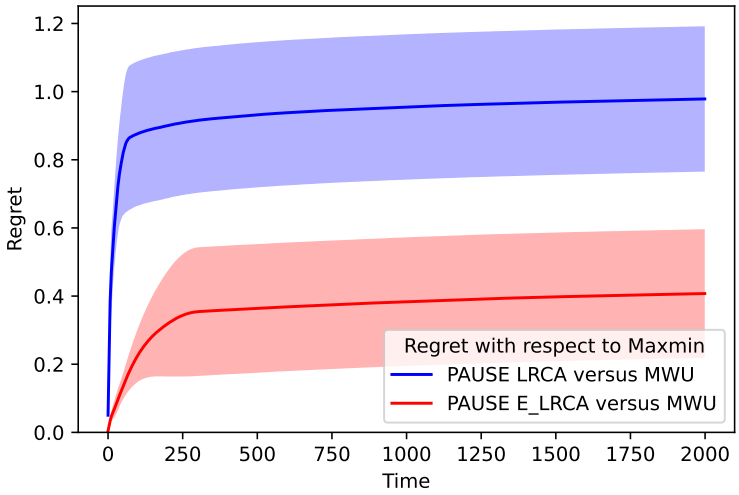

Figure 6.31: Dynamic Regret with respect to the maxmin of the column player in game with a partially-mixed equilibrium

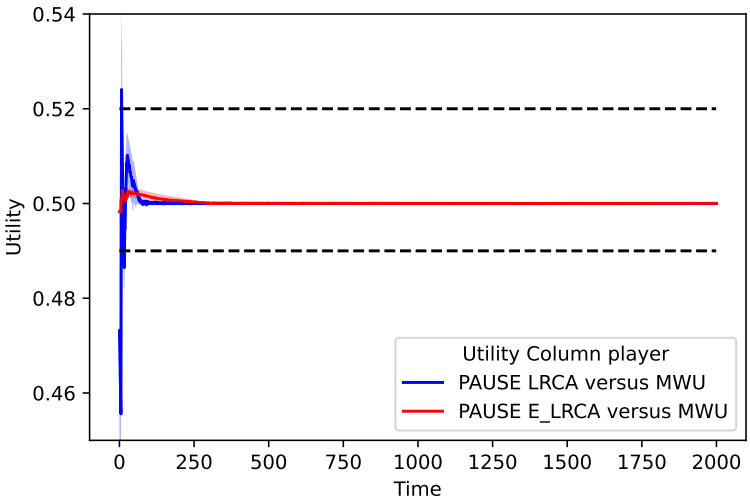

Figure 6.32: Expected Utility of the column player in game with a partially-mixed equilibrium with the safety bounds

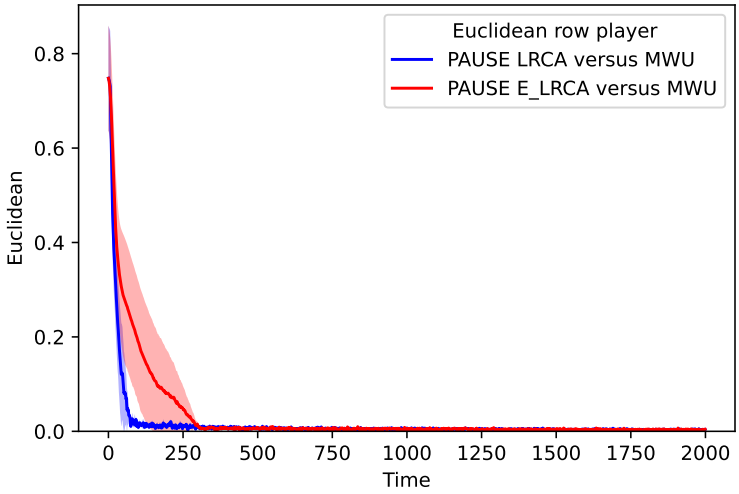

Figure 6.33: Euclidean distance between row player's strategy and the Equilibrium in game with a partially-mixed equilibrium

We conclude with the results in a three actions game with payoff matrix (for the column player):

	A	B	C
A	0.25	0.75	0.3
B	0.75	0.25	0.3
C	0.501	0.501	0.5

where the value v is 0.5.

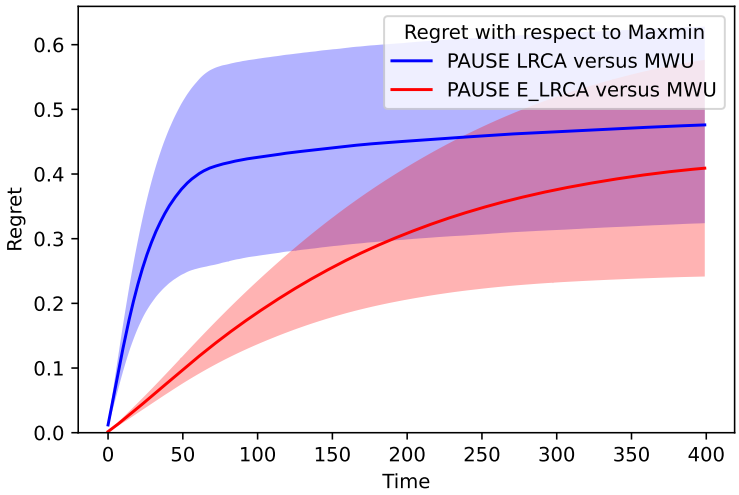

Figure 6.34: Dynamic Regret with respect to the maxmin of the column player in game with a partially-mixed equilibrium

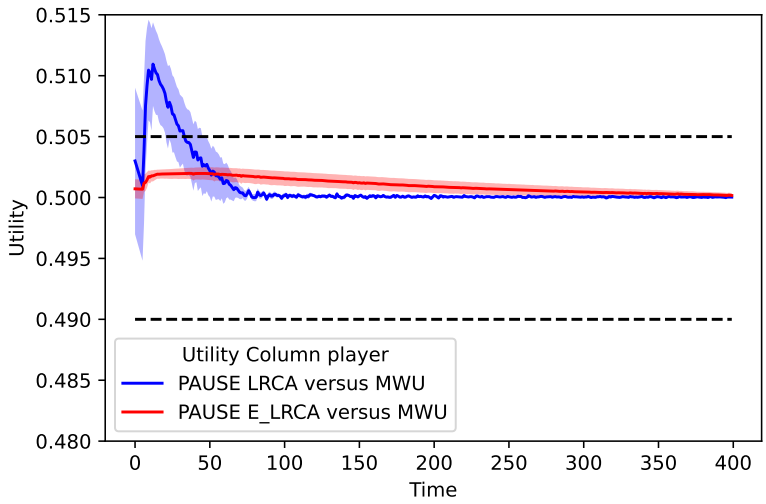

Figure 6.35: Expected Utility of the column player in game with a partially-mixed equilibrium with the safety bounds

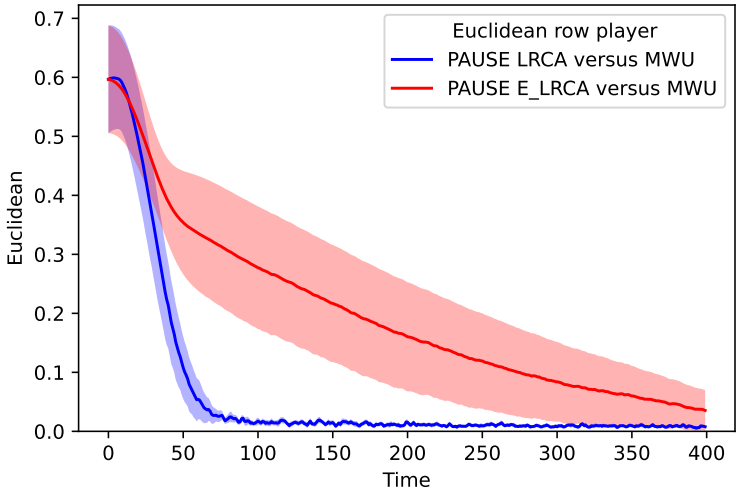

Figure 6.36: Euclidean distance between row player's strategy and the Equilibrium in game with a partially-mixed equilibrium

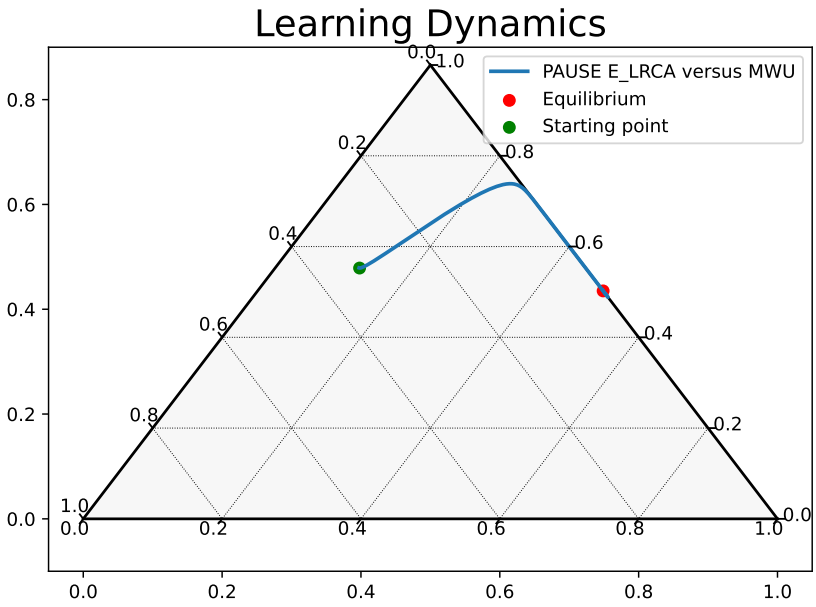

Figure 6.37: Row player's strategy in game with a partially-mixed equilibrium

7 | Conclusions and Future Developments

7.1. Conclusions

Convergence to Equilibria has often been studied in a self-play setting, that is, an agent aims to compute the equilibrium by playing repeatedly against himself. We switched this perspective by developing algorithms capable of making the opponent converge to the Nash of the game, without making assumptions on the exact algorithm the adversary employs. This framework is particularly useful when the opponents are human-like learners, which, by definition, may have different learning abilities. In addition, we introduced safety property in order to guarantee engagement of the human.

To summarize, we developed two algorithms capable of teaching a human-like learner with different feedback (expert and partial semi-bandit) which guarantee, with proper assumptions and in different manners, Safety, Last Round Convergence and Sublinear Dynamic Regret against one of the most famous family of No-Regret learning algorithms, the Online Mirror Descent.

In conclusion, we ran experiments on different type of games in order to show the empiric validity of our algorithm; in the case of bandit feedback, we showed that PAUSE E-LRCA (algorithm 5.1) achieves good performances even in setting (not fully-mixed equilibrium) where the results are not theoretically supported.

7.2. Future Works

There are mainly two paths that future researchers may follow in order to extend our work.

It would be interesting to understand how the framework introduced by this thesis could be adapted for general sum games, that is, games in which a good strategy for one player is not necessarily a bad one for his opponent. In this case, both players may converge

to a Nash that is not a trade-off between their possible rewards, but a solution which is optimal on both sides. From some perspectives, this setting could shift the learner/teacher paradigm, as the learning dynamic of the human would be convenient even for the teacher; indeed, in zero-sum games the teaching dynamic leads to a decreasing of the teacher utility, while in general sum, it could be not strictly necessary.

Moreover, future works may deal with relaxing the normal form game assumption, and extend our framework to extensive form games, which encompass more information such as sequentiality of actions or randomness elements in the game, which are essential in card games as Poker, Bridge etc.

To conclude, it could be interesting to extend our framework in case where even the human has a semi-bandit feedback, even if, differently from the previous proposals, it would involve a complete change of our algorithm.

Bibliography

[1] A. Abbasi, D. H. Ting, and H. Hlavacs. Engagement in games: Developing an instrument to measure consumer videogame engagement and its validation. *International Journal of Computer Games Technology*, 2017:1–10, 01 2017. doi: 10.1155/2017/7363925.

[2] M. Bernasconi-de-Luca, F. Cacciamani, S. Fioravanti, N. Gatti, A. Marchesi, and F. Trovò. Exploiting opponents under utility constraints in sequential games. *Advances in Neural Information Processing Systems*, 34, 2021.

[3] N. Brown and T. Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359:418 – 424, 2018.

[4] M. Campbell, A. Hoane, and F. hsiung Hsu. Deep blue. *Artificial Intelligence*, 134(1):57–83, 2002. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(01)00129-1. URL `https://www.sciencedirect.com/science/article/pii/S0004370201001291`.

[5] A. Celli, A. Marchesi, T. Bianchi, and N. Gatti. Learning to correlate in multi-player general-sum sequential games. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] A. Celli, A. Marchesi, G. Farina, and N. Gatti. No-regret learning dynamics for extensive-form correlated equilibrium. *Advances in Neural Information Processing Systems*, 33:7722–7732, 2020.

[7] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511546921.

[8] C. Daskalakis and I. Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. *arXiv preprint arXiv:1807.04252*, 2018.

[9] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

[10] J. A. DeFalco, J. P. Rowe, L. Paquette, V. Georgoulas-Sherry, K. Brawner, B. W.

Mott, R. S. Baker, and J. C. Lester. Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence in Education*, 28(2):152–193, 2018. ISSN 1560-4292. URL https://www.learntechlib.org/p/19059.

[11] L. Devroye. The equivalence of weak, strong and complete convergence in l1 for kernel density estimates. *The Annals of Statistics*, pages 896–904, 1983.

[12] L. C. Dinh, T.-D. Nguyen, A. B. Zemhoho, and L. Tran-Thanh. Last round convergence and no-dynamic regret in asymmetric repeated games. In V. Feldman, K. Ligett, and S. Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 553–577. PMLR, 16–19 Mar 2021. URL https://proceedings.mlr.press/v132/dinh21a.html.

[13] R. Dörner, S. Göbel, W. Effelsberg, and J. Wiemeyer. *Serious Games: Foundations, Concepts and Practice*. Springer International Publishing, 2016. ISBN 9783319406121. URL https://books.google.it/books?id=nQ7pDAAAQBAJ.

[14] A. Egri-Nagy and A. Törmänen. The game is not over yet—go in the post-alphago era. *Philosophies*, 5(4):37–0, 2020. ISSN 2409-9287. doi: 10.3390/philosophies5040037. URL https://www.mdpi.com/2409-9287/5/4/37.

[15] G. Farina, A. Celli, A. Marchesi, and N. Gatti. Simple uncoupled no-regret learning dynamics for extensive-form correlated equilibrium. *arXiv preprint arXiv:2104.01520*, 2021.

[16] E. Hazan. Introduction to online convex optimization. *CoRR*, abs/1909.05207, 2019. URL http://arxiv.org/abs/1909.05207.

[17] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[18] C.-W. Lee, C. Kroer, and H. Luo. Last-iterate convergence in extensive-form games. *Advances in Neural Information Processing Systems*, 34, 2021.

[19] T. Lin, Z. Zhou, W. Ba, and J. Zhang. Optimal no-regret learning in strongly monotone games with bandit feedback. *arXiv preprint arXiv:2112.02856*, 2021.

[20] J. Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951. ISSN 0003486X. URL http://www.jstor.org/stable/1969529.

[21] J. F. Nash. Equilibrium points in n-person games. *Proc. of the National Academy of Sciences*, 36:48–49, 1950.

[22] J. v. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100: 295–320, 1928. URL `http://eudml.org/doc/159291`.

[23] F. Orabona. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2019. URL `http://arxiv.org/abs/1912.13213`.

[24] R. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer Verlag, Heidelberg, Berlin, New York, 1998.

[25] R. J. Rossetti, J. E. Almeida, Z. Kokkinogenis, and J. Gonçalves. Playing transportation seriously: Applications of serious games to artificial transportation systems. *IEEE Intelligent Systems*, 28(4):107–112, 2013. doi: 10.1109/MIS.2013.113.

[26] Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, USA, 2008. ISBN 0521899435.

[27] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503, 2016. URL `http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html`.

[28] R. Wang, S. Demaria, A. Goldberg, and D. Katz. A systematic review of serious games in training health care professionals. *Simulation in healthcare : journal of the Society for Simulation in Healthcare*, 11, 11 2015. doi: 10.1097/SIH.0000000000000118.

A | Appendix A

We report here the procedure to build the probability found in section 5.3 (that is, solving equation A.1 with respect to the variable x):

$$\frac{5}{2}\sqrt{\frac{\ln\frac{3}{\delta}}{K(t)}} + s_{t+1} = \frac{5}{2}\sqrt{\frac{\ln\frac{3}{x}}{K(t)}} \tag{A.1}$$

By simple math:

$$\frac{2}{5}s_{t+1} + \sqrt{\frac{\ln\frac{3}{\delta}}{K(t)}} = \sqrt{\frac{\ln\frac{3}{x}}{K(t)}}$$

$$\left(\frac{2}{5}s_{t+1} + \sqrt{\frac{\ln\frac{3}{\delta}}{K(t)}}\right)^2 = \frac{\ln\frac{3}{x}}{K(t)}$$

$$K(t)\left(\frac{2}{5}s_{t+1} + \sqrt{\frac{\ln\frac{3}{\delta}}{K(t)}}\right)^2 = \ln\frac{3}{x}$$

$$e^{K(t)\left(\frac{2}{5}s_{t+1}+\sqrt{\frac{\ln\frac{3}{\delta}}{K(t)}}\right)^2} = \frac{3}{x}$$

$$x = \frac{3}{e^{K(t)\left(\frac{2}{5}s_{t+1}+\sqrt{\frac{\ln\frac{3}{\delta}}{K(t)}}\right)^2}} \tag{A.2}$$

Let's decompose the problem and work on the exponent of the denominator:

$$K(t)\left(\frac{4}{25}s_{t+1}^2 + \frac{\ln 3/\delta}{K(t)} + \frac{4}{5}s_{t+1}\sqrt{\frac{\ln 3/\delta}{K(t)}}\right) =$$

$$= \frac{4}{25}K(t)s_{t+1}^2 + \ln 3/\delta + \frac{4}{5}K(t)s_{t+1}\sqrt{\frac{\ln 3/\delta}{K(t)}}$$

Now going upwards:

$$e^{\frac{4}{25}K(t)s_{t+1}^2}\frac{3}{\delta}e^{\frac{4}{5}K(t)s_{t+1}\sqrt{\frac{\ln 3/\delta}{K(t)}}} =$$

$$= \frac{3}{\delta}e^{\frac{4}{5}K(t)s_{t+1}\left(\frac{1}{5}s_{t+1}+\sqrt{\frac{\ln 3/\delta}{K(t)}}\right)}$$

Finally, substituting the result in equation A.2 we obtain:

$$x = \frac{3}{\frac{3}{\delta}e^{\frac{4}{5}K(t)s_{t+1}\left(\frac{1}{5}s_{t+1}+\sqrt{\frac{\ln 3/\delta}{K(t)}}\right)}}$$

which implies:

$$x = \frac{\delta}{e^{\frac{4}{5}K(t)s_{t+1}\left(\frac{1}{5}s_{t+1}+\sqrt{\frac{\ln 3/\delta}{K(t)}}\right)}}$$

πoθεῖ μεν , ***εχθαιρει δε′***, ***βουλεται δ′ εχειν***
«Lo brama, lo detesta, desidera averlo»
[La città di Atene su Alcibiade, "le Rane" (Aristofane)]

Ringraziamenti

Alla mia famiglia che in questi intensi anni mi ha sempre sostenuto; in particolare alla mia mamma Aurelia, papà Silvio, a mio nonno Ernesto e alla mia nonna Franca che spero sarebbe fiera di vedere i traguardi da me raggiunti.

Francesco