POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Safely Guiding a No-Regret Learner to the Equilibrium

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Francesco Emanuele Stradi

Advisor: Prof. Nicola Gatti

Co-advisor: Martino Bernasconi, Federico Cacciamani

Academic year: 2021-2022

1. Introduction

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).

1.2. Related Work

This thesis relies on the framework proposed by Dinh et al. [2021], which developed LRCA (Last Round Convergence in Asymmetric algorithm), an algorithm that achieves convergence (see definition 2.1) to minmax equilibrium against an entire family of No-Regret algorithms (namely, FTRL). Dinh et al. assume that the algorithm is employed by a player (in our thesis, the teacher) with full knowledge of the game, that is, he knows the equilibrium. To conclude, LRCA achieves a sublinear dynamic regret (see definition 2.3) in games where there exists a fully-mixed equilibrium strategy for the opponent.

1.3. Original Contribution

We propose two versions of LRCA algorithm: E-LRCA (algorithm 1) and PAUSE E-LRCA (algorithm 2). The first one deals with the Expert feedback setting and guarantees Last Round Convergence (see definition 2.1) and Sublinear

Dynamic Regret (see definition 2.3) against the entire Online Mirror Descent (OMD) family in games with any kind of equilibrium (fully-mixed, partially-mixed, pure); in addition it guarantees safety (see definition 2.2) 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 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 (see definition 2.1) with high probability and Sublinear Dynamic Regret with respect to the value of the game (see definition 2.4) 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.

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

2. Preliminaries

Consider a repeated two-player zero-sum game. This game is described by a $n \times m$ payoff matrix $\boldsymbol{U}$ scaled in $[0, 1]$. The rows and columns of $\boldsymbol{U}$ represent the pure strategies of the row and column players, respectively. We define the set of feasible strategies of the row player, at round t, by $\Delta_n := \{\boldsymbol{x}_t \in \mathbb{R}^n \mid \sum_{i=1}^n \boldsymbol{x}_t(i) = 1, \boldsymbol{x}_t(i) \geq 0 \; \forall i \in \{1, \dots, n\}\}$. The set of feasible strategies of the column player, denoted by Δ_m, is defined in a similar way. At round t, if the row (resp. column) player chooses a mixed strategy $\boldsymbol{x}_t \in \Delta_n$ (resp. $\boldsymbol{y}_t \in \Delta_m$), then the row player's payoff (or utility) is $-\boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{y}_t$, while the column player's payoff (or utility) is $\boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{y}_t$. Thus, the row (resp. column) player aims to minimise (resp. maximise) the quantity $\boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{y}_t$. We recall that in zero-sum games:

$$\max_{\boldsymbol{y} \in \Delta_m} \min_{\boldsymbol{x} \in \Delta_n} \boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{y} = \min_{\boldsymbol{x} \in \Delta_n} \max_{\boldsymbol{y} \in \Delta_m} \boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{y} = v \quad (1)$$

for some $v \in \mathbb{R}$. We call a point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ satisfying equation 1 the Minmax (or Maxmin) equilibrium of the game, that in zero-sum games is a Nash Equilibrium [2]. The equilibrium strategy $\boldsymbol{x}^*$ is fully-mixed if $\boldsymbol{x}^*(i) > 0 \; \forall i \in \{1, \dots, n\}$.
As specified in section 1 we developed algorithms for different settings. In section 3 both players have the so called expert feedback [3], 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 $\boldsymbol{y}_t$. In section 4 the column player will receive a semi-bandit feedback, namely, 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.
Throughout the entire thesis, the payoff matrix $\boldsymbol{U}$ is known by the column player (the teacher), that is, he perfectly knows the equilibrium (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). Next, we define the main properties the algorithms developed during the thesis will guarantee.

Definition 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.

Definition 2.2. *(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 $u(t)$ opponent's payoff at time t.*

We now introduce the notion of Dynamic Regret for the column player as:

$$DR_T := \sum_{t=1}^{T} \left(\max_{\boldsymbol{y}\in\Delta_m} \boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{y} - \boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{y}_t \right) \tag{2}$$

Definition 2.3. *(No-Dynamic Regret) An algorithm is no-dynamic regret (or has the no-dynamic regret property) if* $\lim_{T\to\infty} \frac{DR_T}{T} = 0$.

To conclude, we introduce the notion of Dynamic Regret with respect to the Maxmin value of the game:

$$DR_T^{eq} := \sum_{t=1}^{T} |\boldsymbol{x}_t^\top \boldsymbol{U}\boldsymbol{y}_t - v| \tag{3}$$

From which:

Definition 2.4. *(No-Dynamic Regret with respect to the MaxMin) An algorithm is no-dynamic regret with respect to the MaxMin if* $\lim_{T\to\infty} \frac{DR_T^{eq}}{T} = 0$.

3. Safe Guide with Expert Feedback

3.1. Algorithm

We underline the main ideas behind algorithm 1.

In odd rounds column player plays the equilibrium so that, if the row player's equilibrium strategy is fully-mixed, it is possible to predict his next strategy (for stability, 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 towards 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).

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.

Algorithm 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**
    $\boldsymbol{e}_{t-1} := \text{argmax}_{e\in\{\boldsymbol{e}_1,\boldsymbol{e}_2,\dots\boldsymbol{e}_m\}} \boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{e}$
    $f(\boldsymbol{x}_{t-1}) := \max_{\boldsymbol{y}\in\Delta_m} \boldsymbol{x}_{t-1}^\top \boldsymbol{U}\boldsymbol{y}$
    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

3.2. Theoretical Results

We start stating the result in terms of safety. The requirement $||\boldsymbol{U}\boldsymbol{y}^*||_\infty < \xi_2$ is necessary only for games in which there is not a fully-mixed equilibrium strategy for the row player, otherwise it is implicit in taking $v \in (\xi_1, \xi_2)$. Moreover, for games with fully-mixed equilibrium, we can choose $v \in [\xi_1, \xi_2)$. To conclude, note that if the utility of the column player is bounded in $[\xi_1, \xi_2]$, the utility of the row player will be bounded in $[-\xi_2, -\xi_1]$.

Theorem 3.1. *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 column player follows E-LRCA (algorithm 1), the Utility of the column Player will be bounded in $[\xi_1, \xi_2]$ at each round.*

We proceed with the convergence result, which requires similar assumptions to theorem 3.1.

Theorem 3.2. *Assume that the row player follows an algorithm of the OMD family, then if the column player follows the Algorithm E-LRCA with 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.*

We conclude with the Dynamic Regret, stating the main theorem and then reporting the corollary for games with fully-mixed equilibrium strategy for the row player. In the latter, it is possible to express the Regret without exploiting the dynamic of the opponent learning rate μ.

Theorem 3.3. *Assume that the row player follows an algorithm of the OMD family, 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 strategy for the row player.*

Corollary 3.1. *Assume that the row player follows an algorithm of the OMD family. If there exists a fully-mixed minmax 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 $DR_T = \mathcal{O}\left(\frac{\sqrt{\log(n)}}{\sqrt{\gamma_{min}}} T^{3/4}\right)$. Furthermore, in the case the row player uses a constant learning rate μ, we have $DR_T = \mathcal{O}\left(\frac{n}{\sqrt{\mu\gamma_{min}}} T^{1/2}\right)$.*

4. Safe Guide with Partial Semi-Bandit Feedback

4.1. Algorithm

Algorithm 2 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 2 (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 at 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. Then, column player computes the optimistic best response ($\boldsymbol{e}_t$) and 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). 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 is needed, leading to a deceleration of the teaching dynamic.

4.2. Theoretical Results

We start with the safety result. Note that in this subsection, theorems will be related only to games with fully-mixed equilibrium strategy for the row player. As concerns safety, it is still guaranteed for games without fully-mixed equilibrium strategy for the row player, but the theorem is the same as seen in the previous section. Due to Semi-Bandit feedback, the next theorems will be valid with high probability.

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 PAUSE E-LRCA (algorithm 2), the Expected Utility of the column Player will be bounded in $[\xi_1, \xi_2]$ at each round with high probability.*

We proceed with the convergence result, which requires similar assumptions to theorem 4.1.

Theorem 4.2. *Assume that the row player follows an algorithm of the OMD family 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 2) with 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.*

To conclude, we compute the Dynamic Regret with respect to the Maxmin value of the game. Note that the result is dependant on the parameter λ used in the computation of $K(t)$ in algorithm 2. For $0 < \lambda < 2$, the Regret is sublinear with high probability.

Theorem 4.3. *Assume that the row player follows an algorithm of the OMD family. Then by following PAUSE E-LRCA (algorithm 2), the column player will achieve the no-dynamic regret (with respect to the MaxMin) property with the dynamic regret satisfying $DR_T^{eq} = \mathcal{O}\left(\frac{n}{\sqrt{\mu\gamma_{min}}} T^{\max\left(-\lambda/2+1, \frac{1}{2}+\frac{-\lambda/2+1}{2}\right)}\right)$ in games with fully-mixed equilibrium strategy for the row player, with high probability.*

5. Experiments

We highlight significant experiments for PAUSE E-LRCA (algorithm 2) in a game where there is not a fully-mixed equilibrium strategy for the row player, in order to show the empiric properties explained in table 1. As benchmarks for our experiments, we implemented a not-safe version of the algorithm (namely, PAUSE LRCA).

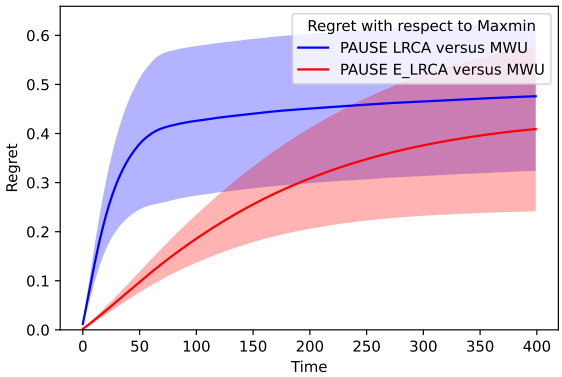

Figure 1: Dynamic Regret with respect to the maxmin of the column player in game with a partially-mixed equilibrium

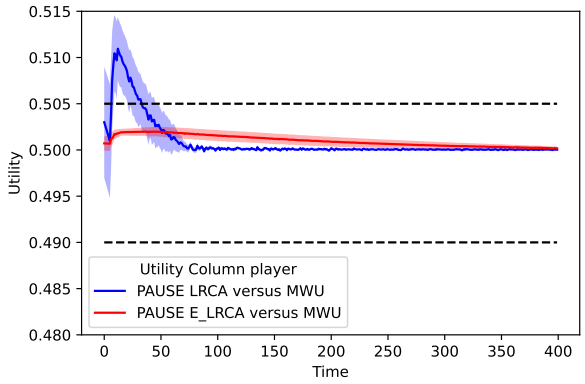

Figure 2: Expected Utility of the column player with the safety bounds in game with a partially-mixed equilibrium

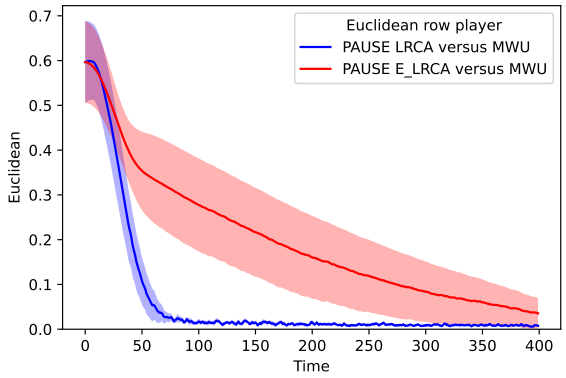

Figure 3: Euclidean distance from the equilibrium of the row player's strategy in game with a partially-mixed equilibrium

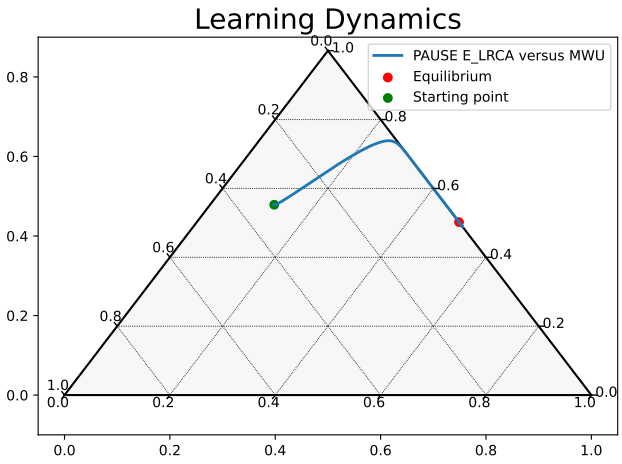

Figure 4: Row player's strategy dynamics on the simplex in game with a partially-mixed equilibrium

6. Conclusions

Convergence to Equilibria has often been studied in a self-play setting, that is, an agent aims to compute the equilibrium playing repeatedly against himself. We switched this perspective 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 types of game in order to show the empiric validity of our algorithms; in the case of bandit feedback, we showed that PAUSE E-LRCA (algorithm 2) achieves good performances even in setting (not fully-mixed equilibrium) where the results are not theoretically supported.

References

[1] Le Cong Dinh, Tri-Dung Nguyen, Alain B. Zemhoho, and Long Tran-Thanh. Last round convergence and no-dynamic regret in asymmetric repeated games. In Vitaly Feldman, Katrina Ligett, and Sivan 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`.

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

[3] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511546921.