# Online Learning against Strategic Adversary

Doctoral Consortium

Le Cong Dinh
University of Southampton
United Kingdom
l.c.dinh@soton.ac.uk

## ABSTRACT

Our work considers repeated games in which one player has a different objective than others. In particular, we investigate repeated two-player zero-sum games where the column player not only aims to minimize her regret but also stabilize the actions. Suppose that while repeatedly playing this game, the row player chooses her strategy at each round by using a no-regret algorithm to minimize her regret. We develop a no-dynamic regret algorithm for the column player to exhibit last round convergence to a minimax equilibrium. We show that our algorithm is efficient against a large set of popular no-regret algorithms the row player can use, including the multiplicative weights update algorithm, general follow-the-regularized-leader and any no-regret algorithms satisfy a property so called "stability". Our algorithm can be applied to the game setting where the column player is also a designer of the system, and has full control over payoff matrices.

## KEYWORDS

Online Learning, Last Round Convergence, System Design

**ACM Reference Format:**
Le Cong Dinh. 2022. Online Learning against Strategic Adversary: Doctoral Consortium. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Auckland, New Zealand, May 9–13, 2022*, IFAAMAS, 2 pages.

## 1 INTRODUCTION

We consider a repeated two-player game setting, in which one player (say the column player) is also a designer of the system (i.e., she can design the payoffs for both players), and her opponent (the row player) is a strategic utility maximiser, who can efficiently learn to adapt their strategy to the column player's behaviour over time in order to achieve good total payoff. The goal of the column player is to guide her opponent into selecting a mixed strategy which is favourable for the system designer. In particular, she needs to achieve this by: (i) designing appropriate payoffs for both players; and (ii) strategically interacting with the row player during a sequence of plays in order to ensure her opponent converges to the desired behaviour. In this paper, we propose an approach for solving this problem, which consists of two components corresponding to (ii) and (i) respectively:

**Last round convergence in zero-sum games**. Once a zero-sum game has been chosen, the system designer must strategically incentivise the row player to converge to the desired behaviour through repeated play. Recall that, if both players commit to a no-regret algorithm at each phase of play, then payoffs will converge in expectation to the value of the game. Additionally, during this process, no exchange of information takes place, as both players require only the payoffs they observe in order to update their strategies [2, 5]. As a result, one may hope that, by adopting a no-regret algorithm, the system designer can naturally guide the row player to their minimax strategy. Unfortunately, convergence of average payoffs, in general, does not imply convergence in strategies. This issue is known as the last round convergence problem in the online learning literature, and has recently attracted a good amount of attention [3, 4]. In what follows, we propose a novel no-regret algorithm which leverages the information advantage of the system designer to guide the row player to its minimax strategy over time, under a mild set of assumptions.

**Games with a unique minimax solution**. To begin, the system designer must decide upon the payoffs for both players. For two-player zero-sum games, it is well known that, in the full information setting, the best payoff rational players can attain is their payoff at any minimax equilibrium [2]. Thus, a natural idea is to construct a zero-sum game, $\boldsymbol{A}$, in which the *only* minimax strategy available to the row player is the desired behaviour. In doing so, the system designer can hope that any reasonable player will eventually begin to play their minimax strategy, and thus adopt the desired behaviour. Construction of games with unique equilibrium solutions is a long running problem within the literature, beginning with the seminal work of Shapley, Karlin, and Bohnenblust [1]. In this work, we aim to provide zero-sum game constructions which offer the system designer more flexibility in terms of the payoff matrix they choose. Such flexibility may be useful when enforcing system payoffs correspond to costly real world actions.

For the remainder, we will describe each component of our approach in more detail, beginning with the last round convergence algorithm.

## 2 LAST ROUND CONVERGENCE IN TWO-PLAYER ZERO-SUM GAMES

Given a game matrix $\boldsymbol{A}$, we investigate how the system designer can guide the row player to its minimax strategy, under the assumption that the row player uses a no-regret algorithm.

Firstly, we show that a naïve approach, namely to repeatedly playing $\mathbf{y}^*$, will not always lead to the desired last round convergence (i.e., the row player will converge to $\mathbf{x}^*$):

Claim 1. *If* $support(\mathbf{x}) > 1$ *, then there is no guarantee that if the column player repeatedly plays* $\mathbf{y}^*$*, the row player will eventually converge to* $\mathbf{x}^*$*.*

*Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Auckland, New Zealand.* 

Given this result, we need to design a different game playing policy for the column player. More specifically, we require a policy which actively exploits the information advantage possessed by the column player. The LRCA algorithm, originally proposed by Dinh *et al.* [4], was designed with this goal in mind. On odd rounds, LRCA selects the minimax strategy in order to stabilise the trajectory of both players. Meanwhile, in even rounds, the LRCA algorithm exploits any weaknesses in the row player's strategy by moving in the direction of the highest payoff strategy in the previous round, where the rate of moving depends on how far the row player's strategy from the Nash equilibrium (i.e., $f(\boldsymbol{x}_{t-1}) - v$). The LRCA algorithm guarantees last round convergence (for both players) against a number of popular no-regret algorithms including the multiplicative weight update algorithm, online mirror descent, and the linear multiplicative weight update algorithm.

LRCA is described in detail by Algorithm 1 below:

**Algorithm 1:** Last Round Convergence with Asymmetry (LRCA) algorithm

**Input:** Current iteration $t$, past feedback $\boldsymbol{x}_{t-1}^\top A$ of the row player, minimax strategy $\boldsymbol{y}^*$ and value $v$ of the game.
**Output:** Strategy $\boldsymbol{y}_t$ for the column player
**if** $t = 2k-1,\ k \in \mathbb{N}$ **then**
  $\boldsymbol{y}_t = \boldsymbol{y}^*$
**if** $t = 2k,\ k \in \mathbb{N}$ **then**
  $\boldsymbol{e}_t := \text{argmax}_{\boldsymbol{e} \in \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots \boldsymbol{e}_m\}} \boldsymbol{x}_{t-1}^\top \boldsymbol{A}\boldsymbol{e}$
  $f(\boldsymbol{x}_{t-1}) := \max_{\boldsymbol{y} \in \Delta_m} \boldsymbol{x}_{t-1}^\top \boldsymbol{A}\boldsymbol{y}; \quad \alpha_t := \frac{f(\boldsymbol{x}_{t-1}) - v}{\max(\frac{n}{4}, 2)}$
  $\boldsymbol{y}_t := (1-\alpha_t)\boldsymbol{y}^* + \alpha_t \boldsymbol{e}_t$

More generally, we prove that LRCA also guarantees last round convergence when paired with *any* no-regret algorithm, as long as this no-regret algorithm possesses the "stability" property, as defined below:

*Definition 2.* A no-regret algorithm is *stable* if $\forall t : \mathbf{y}_t = \mathbf{y}^* \implies \mathbf{x}_{t+1} = \mathbf{x}_t$.

Note that a wide range of no-regret algorithms adhere to this property. For example, the class of Follow the Regularised Leader algorithms are stable. A proof of this fact is given in the full paper.

Theorem 3. *Assume that the row player follows a stable no-regret algorithm and n is the dimension of the row player's strategy. Then, by following LRCA, for any $\epsilon > 0$, there exists $l \in \mathbb{N}$ such that $\frac{\mathcal{R}_l}{l} = O(\frac{\epsilon^2}{n})$ and $f(\mathbf{x}_l) - v \le \epsilon$.*

Note that $(\mathbf{x}_l, \mathbf{y}^*)$ are $\epsilon$-Nash equilibria. For no-regret algorithms with the optimal regret bound $\mathcal{R}_l = O(l^{\frac{1}{2}})$, Theorem 3 guarantees that the row player will reach an $\epsilon$-Nash equilibrium in at most $O\left(\epsilon^{-4}\right)$ rounds. Thus, LRCA is successful in guiding the row player towards the desired behaviour, as long as the row player adopts a stable no-regret algorithm.

## 3 DESIGNING GAMES WITH A UNIQUE MINIMAX SOLUTION

Without loss of generality, suppose there exists a preferred strategy, $\mathbf{y}^*$, that the column player would like to play whilst at a minimax equilibrium. That is, the column player wishes not only to ensure that the row player's minimax strategy is $\mathbf{x}^*$, but also that their own minimax strategy is $\mathbf{y}^*$. Under this assumption, we observe that the strategy pair $(\mathbf{x}^*, \mathbf{y}^*)$ must form a minimax equilibrium of $\boldsymbol{A}$. Moreover, as previously mentioned, $\mathbf{x}^*$ must be the unique minimax strategy for the row player.

In order to satisfy both conditions, the support of $\mathbf{y}^*$ must be greater than equal to the support of $\mathbf{x}^*$. For if this is not the case, then it is provably impossible to construct a matrix $\boldsymbol{A}$ with minimax equilibrium $(\mathbf{x}^*, \mathbf{y}^*)$ whilst guaranteeing the uniqueness of the minimax strategy $\mathbf{x}^*$ [1]. From now on, without loss of generality, we shall assume that for any strategy with support $k$, that the first $k$ entries are nonzero.

If the support of $\mathbf{y}^*$ is greater than the support of $\mathbf{x}^*$, then we construct $\boldsymbol{A}$ according to Theorem 4. The other case, in which the support of $\mathbf{y}^*$ is equal to the support of $\mathbf{x}^*$ is dealt with in the full version of the paper. Note that, in both theorems, $\mathbf{y}^*$ is not necessarily the unique minimax strategy for the column player, whilst $\mathbf{x}^*$ is the unique minimax strategy for the row player.

Theorem 4. *Let $\mathbf{x} \in \Delta_n$, $\mathbf{y} \in \Delta_m$ such that $k = support(\mathbf{x}) < l = support(\mathbf{y})$. Let the matrix $\boldsymbol{A}$ be of the form*

$$\boldsymbol{A} = \begin{bmatrix} a_1 & \alpha_2 & \ldots & \alpha_k & \beta_1 & \ldots & \beta_1 \\ \alpha_1 & a_2 & \ldots & \alpha_k & \beta_2 & \ldots & \beta_2 \\ \ldots & \ldots & \ldots & \ldots & & & \\ \alpha_1 & \alpha_2 & \ldots & a_k & \beta_k & \ldots & \beta_k \\ \alpha_1 - z & \alpha_2 - z & \ldots & \alpha_k - z & v & \ldots & v \\ \ldots & \ldots & \ldots & \ldots & & & \\ \alpha_1 - z & \alpha_2 - z & \ldots & \alpha_k - z & v & \ldots & v \end{bmatrix}$$

*where the parameters of $\boldsymbol{A}$ satisfy*

$$0 < v_1 < v\bar{y}, \qquad \bar{y} = \sum_{i=k+1}^{l} \mathbf{y}_i, \quad z = \frac{v\bar{y} - v_1}{\sum_{i=1}^{k} \mathbf{y}_i},$$

$$\beta_i = v, \quad \alpha_i = v + \frac{x_i(v\bar{y} - v_1)}{\mathbf{y}_i}, \quad a_i = \alpha_i - \frac{v\bar{y} - v_1}{\mathbf{y}_i} \ \forall i \in [k].$$

*then $\mathbf{x}$ is the unique minimax strategy for the row player in the zero-sum game described by $\boldsymbol{A}$.*

In Theorem 4, the parameters $a_i$ and $z$ ensure that $\mathbf{x}$ is the unique minimax strategy for the row player, even in the case where the support of $\mathbf{x}$ is less than $n$. Meanwhile, the parameters $\beta_i$ ensure that $\mathbf{y}$ is a minimax strategy for the column player. Lastly, the parameter $\gamma$ ensures that all entries of $\boldsymbol{A}$ are nonnegative (although this is not strictly required).

## REFERENCES


[1] HF Bohnenblust, S Karlin, and LS Shapley. 1950. Solutions of discrete, two-person games. *Contributions to the Theory of Games* 1 (1950), 51–72.
[2] Nicolo Cesa-Bianchi and Gábor Lugosi. 2006. *Prediction, learning, and games*. Cambridge university press.
[3] C Daskalakis and Ioannis Panageas. 2019. Last-Iterate Convergence: Zero-Sum Games and Constrained Min-Max Optimization. In *10th Innovations in Theoretical Computer Science (ITCS) conference, ITCS 2019*.
[4] Le Cong Dinh, Tri-Dung Nguyen, Alain B. Zemhoho, and Long Tran-Thanh. 2021. Last Round Convergence and No-Dynamic Regret in Asymmetric Repeated Games. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory (Proceedings of Machine Learning Research, Vol. 132)*, Vitaly Feldman, Katrina Ligett, and Sivan Sabato (Eds.). PMLR, 553–577.
[5] Yoav Freund and Robert E Schapire. 1999. Adaptive game playing using multiplicative weights. *Games and Economic Behavior* 29, 1-2 (1999), 79–103.