Foolproof Cooperative Learning

***

Abstract

This paper extends the notion of learning equilibrium in game theory from matrix games to stochastic games. We introduce Foolproof Cooperative Learning (FCL), an algorithm that converges to a Tit-for-Tat behavior. It allows cooperative strategies when played against itself while being not exploitable by selfish players. We prove that in repeated symmetric games, this algorithm is a learning equilibrium. We illustrate the behavior of FCL on symmetric matrix and grid games, and its robustness to selfish learners.

1 Introduction

In William Golding's novel "Lord of the Flies", a group of children who survived an airplane crash try to establish rules on a desert island in order to avoid chaos. Unfortunately, they fail at forcing a cooperative solution and some of them start defecting, which results in a demented group behaviour. In this paper, we prevent such tragedies in learning algorithms by constructing a safe way to learn cooperation in unknown environments, without being exploitable by potentially selfish agents.

In multi-agent learning settings, environments are usually modeled by stochastic games (Shapley 1953). Multi-agent reinforcement learning (MARL) brings a framework to construct algorithms that aim to solve stochastic games where players individually or jointly search for an optimal decision-making policy to maximize a reward function. Individualist approaches mostly aim at reaching equilibrium, taking the best actions whatever the opponents behaviors are (Bowling and Veloso 2001; Littman 2001). Joint approaches aim at optimizing a cooperative objective and can be viewed as a single agent problem in a larger dimension (Claus and Boutilier 1998), but are easily exploited when one agent starts being individualist.

We focus on symmetric situations, making sure that no agent has an individual advantage. For example, this is the case on a desert island with a quantity of resources equally accessible to all agents. Moreover, we consider repeated games, modelling the recurrent possibility to start again the situation from the beginning. In the island resource example, repetitions could represent successive days or, at larger scale, 4-seasons cycles. In fact, any stochastic game where players have the same reward functions and dynamics is symmetric since all players are starting with the same chances. This applies to most of common-pool resource appropriation games.

In this context, we introduce Foolproof Cooperative Learning (FCL), a model-free learning algorithm that, by construction, converges to a Tit-for-Tat behaviour, cooperative against itself and retaliatory against selfish algorithms. We extend the notion of learning equilibrium (Brafman and Tennenholtz 2003) to stochastic games, describing a class of learning algorithms such that the best way to play against them is to adopt the same behaviour. We demonstrate that FCL is a learning equilibrium that forces a cooperative behaviour, and we empirically verify this claim with two-agents matrix games and grid-world repeated symmetric games.

When not given in the paper, the proofs of all stated results are provided in the appendix.

2 Definitions and Notations

An N-player stochastic game can be written as a tuple $(\mathcal{S}, (\mathcal{A}_i)_{i=1\dots N}, \mathcal{P}, \mu_0, (r_i)_{i=1\dots N})$, where $\mathcal{S}$ is the set of states, $\mathcal{A}_i$ the set of actions for player i, $\mathcal{P}$ the transition probability $(\mathcal{P}(\cdot|s, a_1 \dots a_N))$, μ_0 a distribution over states $(\mu(s^0))$, r_i the reward function for player i $(r_i(s, a_1 \dots a_N))$. We also assume bounded, deterministic reward functions and finite state and action spaces.

In a repeated stochastic game, a stochastic game (the stage game) is played and at each iteration, it continues with probability $\gamma \in [0, 1[$ or terminates and starts again according to μ_0. This is repeated an infinite number of times, and players have to maximize their average return during a stage game (Munoz de Cote and Littman 2008). Terminating with probability γ is equivalent to use a discount factor while playing a stage game.

A stationary strategy (or policy) for player i, $\pi_i(\cdot|s) \in \Pi_{\mathcal{A}_i}$, maps a state to a probability distribution over its set of possible actions. We note π_{-i} the product of all

players strategies but player i and $\boldsymbol{\pi} = \pi_1 \times \cdots \times \pi_N = \pi_i \times \pi_{-i}$ the product of all player strategies, called the strategy profile. Given opponents strategies π_{-i}, the goal for a rational player i is to find a strategy π_i^* that maximizes its average return $\mathcal{R}_i$ during a stage game:

$$\begin{aligned}\pi_i^* &= \underset{\pi_i}{\operatorname{argmax}}\, \mathcal{R}_i(\pi_i, \pi_{-i}) \\ &= \underset{\pi_i}{\operatorname{argmax}}\, \mathbb{E}_{\pi_i, \pi_{-i}, \mathcal{P}}\left[\sum_l \gamma^l r(s^l, a_i^l, a_{-i}^l)\right].\end{aligned}$$

The policy π_i^* depends on the opponents strategies and is called the best response for player i to π_{-i}. In general, we call strategy any process $\{\pi^t\}_t$ defining a stationary strategy for any stage t. The value of a player's non-stationary strategy $\{\pi^t\}_t$ is the average return over stage games, $\mathbb{E}_{t>0}[\mathcal{R}_i(\pi_i^t, \pi_{-i}^t)]$.

In order to allow rewarding or retaliation strategies, we only consider games where all players are aware of all opponents actions and rewards, and receive a signal each time the game is reset. We also admit players to share information with some opponents in order to organize joint retaliation actions or joint explorations. Moreover, we only consider *Repeated Symmetric Games* (RSG):

Definition 1 (Repeated Symmetric game (RSG))**.** *An N-player repeated stochastic game is symmetric if, for any stationary strategy profile $(\pi_1 \dots \pi_N)$ and for any permutation ψ over players:*

$$\forall 1 \leq i \leq N,\ \mathcal{R}_{\psi(i)}(\pi_i, \pi_{-i}) = \mathcal{R}_i(\pi_{\psi(i)}, \pi_{\psi(-i)}).$$

This generalizes the definition for symmetric N-player matrix games (Dasgupta, Maskin, and others 1986) to stochastic games where players utilities are replaced by average returns[1]. In this paper, we use the concept of N-cyclic permutations to construct specific strategies:

Definition 2 (N-cyclic permutation)**.** *A permutation σ is N-cyclic if for all $i, j \in \{1 \dots N\}$, there is k such that $\sigma^k(i) = j$.*

2.1 Nash equilibrium

A Nash equilibrium describes a stationary strategy profile $\boldsymbol{\pi}^* = \pi_1^* \times \cdots \times \pi_N^*$, such that no player can individually deviate and increase its payoff (Nash 1951):

$$\forall 1 \leq i \leq N,\ \forall \pi_i \in \Pi_{\mathcal{A}_i},\ \mathcal{R}_i(\pi_i, \pi_{-i}^*) \leq \mathcal{R}_i(\pi_i^*, \pi_{-i}^*).$$

Note that in a symmetric game, for any Nash equilibrium with returns $(\mathcal{R}_1 \dots \mathcal{R}_N)$ and for any permutation σ over players, there is another Nash equilibrium with returns $(\mathcal{R}_{\sigma(1)} \dots \mathcal{R}_{\sigma(N)})$. This definition can be extended to non-stationary strategies using expected return over stage games: no players can individually deviate from an equilibrium non-stationary strategy and increase its average return over stage games ($\mathbb{E}_{t>0}[x^t] = \mathbb{E}[\sum_{t>0} x^t]$ stands for the average over stages):

$$\begin{aligned}&\forall 1 \leq i \leq N,\ \forall \{\pi_i^t \in \Pi_{\mathcal{A}_i}\}_t, \\ &\mathbb{E}_{t>0}[\mathcal{R}_i(\pi_i^t, \pi_{-i}^{t,*})] \leq \mathbb{E}_{t>0}[\mathcal{R}_i(\pi_i^{t,*}, \pi_{-i}^{t,*})].\end{aligned}$$

As $\mathbb{E}_{t>0}[\mathcal{R}_i(\pi_i^t, \pi_{-i}^t)] = \mathcal{R}_i(\pi_i, \pi_{-i})$ for stationary strategy profiles, any stationary strategy equilibrium is still an equilibrium among non-stationary processes.

3 Cooperative strategies

We call cooperative any strategy (not necessary stationary) that maximizes a common quantity $\hat{\mathcal{R}} = f(\mathcal{R}_1 \dots \mathcal{R}_N)$. Usual examples are strategies that maximize the sum, the product or the minimum of players returns. In RSGs, the strategy that maximizes the minimum of player returns is particularly interesting as it coincides with the egalitarian solutions (Kalai 1977) to the Bargaining problem (Nash Jr 1950) and is easy to determine. In this paper, we refer to this strategy as the *egalitarian strategy*. An important property of RSGs is the fact that egalitarian solutions can always be obtained by repeatedly applying an N-cyclic permutation on a stationary strategy that maximizes the sum of players returns.

Theorem 1. *Let π_i^Σ be a stationary strategy for player i that maximizes the sum of players returns in an N-player RSG, σ an N-cyclic permutation over players, and t indexing the repeated stage games. Then, the strategy $\boldsymbol{\pi}^t = (\pi_{\sigma^t(1)}^\Sigma \dots \pi_{\sigma^t(N)}^\Sigma)$ (where $\sigma^t = \sigma \circ \cdots \circ \sigma$ t times) is an egalitarian strategy.*

3.1 Tit-for-Tat

Given a stochastic game, one player i can learn a strategy $\pi_i^{r,j}$ that retaliates when another player j deviates from a target strategy. If a retaliation is smaller than the reward obtained by the player while deviating, the strategy can be repeated until the retaliation is larger than this reward in total. In that case, the target strategy is said enforceable: if all player are accorded to retaliate when a player deviates from a strategy profile and if the retaliation is strong enough, no player can improve its payoff by individually deviating from the strategy profile. If opponents actions are part of the observable state and if the target strategy profile and the dynamics are deterministic, it becomes possible to construct a stationary strategy that retaliates when a player does not play according to the profile. If the retaliation lasts forever after the first deviation, the strategy is by construction a Nash equilibrium (Osborne and Rubinstein 1994). However, we are more interested in finished retaliations since it gives a chance to a selfish learning agent to learn the target strategy. Such processes are called Tit-for-Tat (TFT) and are known to induce cooperation in repeated social dilemma (Axelrod and Hamilton

[1] Actually, the definition initially given: $\forall i, \mathcal{R}_i(\pi_i, \pi_{-i}) = \mathcal{R}_{\psi(i)}(\pi_{\psi(i)}, \pi_{\psi(-i)})$ (Dasgupta, Maskin, and others 1986) is incorrect in the sens that symmetries are not independent of player identities, which is not the case if the right-hand return is indexed with the inverse permutation instead (Vester 2012).

1981). Theorem 2 states that in an RSG, if the target is an egalitarian strategy, there is always a stationary way to retaliate and therefore one can always construct a TFT strategy:

Theorem 2. *In an RSG, let* $\boldsymbol{\pi}^{r,j} = \operatorname{argmin}_{\pi_{-j}} \operatorname{argmax}_{\pi_j} \mathcal{R}_j(\pi_j, \pi_{-j}) = \pi_j^{r,j} \times \pi_{-j}^{r,j}$*, and* $\boldsymbol{\pi}^*$ *a egalitarian strategy (not necessary stationary). Then,* $\boldsymbol{\pi}^{r,j}$ *is a retaliation strategy with respect to* $\boldsymbol{\pi}^*$*:*

$$\forall 1 \leq j \leq N, \forall \pi_j \in \Pi_{A_j},$$
$$\mathcal{R}_j(\pi_j, \pi_{-j}^{r,j}) \leq \mathbb{E}_{t\geq 0}[\mathcal{R}_j(\pi_j^{*,t}, \pi_{-j}^{*,t})].$$

For a player j, we note $V_j^c = \mathbb{E}_{t\geq 0}[\mathcal{R}_j(\pi_j^{*,t}, \pi_{-j}^{*,t})]$ its average return when all players cooperate, $V_j^r = \mathcal{R}_j(\boldsymbol{\pi}^{r,j})$ its best average return when others retaliate and $V_j^d = \max_{\pi_j} \mathbb{E}_{t\geq 0}[\mathcal{R}_j(\pi_j, \pi_{-j}^{*,t})]$ its best average return by defecting. When a single retaliation is too small so it is still worth defecting for a selfish player, the retaliation must be repeated. The minimal number or retaliation repeats can be given by (see the proof of Thm. 3 below):

$$K_j = \left\lceil \frac{V_j^d - V_j^c}{V_j^c - V_j^r} \right\rceil. \tag{1}$$

In the edge case where $V_j^c = V_j^r$, the retaliation strategy must be employed endless, but the cooperative objective is not affected (this is the case in rock–paper–scissors). Let $\{\boldsymbol{\pi}_{\text{TFT}}^t\}_t$ be the (non-stationary) strategy that follows $\boldsymbol{\pi}^*$ if all players cooperate, or repeat $\boldsymbol{\pi}^{r,j}$ over K_j stage games if a player j deviates from $\boldsymbol{\pi}^*$. By construction, $\{\boldsymbol{\pi}_{\text{TFT}}^t\}_t$ is a Nash equilibrium.

Theorem 3. $\{\boldsymbol{\pi}_{TFT}^t\}_t$ *is a Nash equilibrium.*

Proof. Since $K_j \geq \frac{V_j^d - V_j^c}{V_j^c - V_j^r}$ and $V_j^c \geq V_j^r$, we write:

$$K_j(V_j^c - V_j^r) \geq V_j^d - V_j^c,$$

which gives:

$$V_j^c \geq \frac{1}{K_j + 1}(V_j^d + K_j V_j^r).$$

On the left, this is the average return over stages of an always cooperating player, on the right this is the average return over stages of any deviating player. Therefore, for any $\{\pi_j^t\}_t \neq \{\boldsymbol{\pi}_{\text{TFT}}^t\}_t$:

$$\mathbb{E}_{t\geq 0}[\mathcal{R}_j(\boldsymbol{\pi}_{\text{TFT}}^t)] \geq \mathbb{E}_{t\geq 0}[\mathcal{R}_j(\pi_j^t, \pi_{\text{TFT}_{-j}}^t)].$$

□

4 Learning algorithm

In this paper, we define *learning algorithms* as follows:

Definition 3. *A learning algorithm (for player i) is a random process* $A_i = \{\pi_i^T\}_T$ *conditioned, at any stage* $T > 0$*, by the historic*

$$\mathcal{H}^T = \{\{s^l, a_i^l, a_{-i}^l, r_i^l, r_{-i}^l\}_{l\in t}\}_{t<T}$$

of all states, actions and rewards encountered up to stage $T-1$ *(*$l \in t$ *stands for the* l*-th transition belonging to stage* t*).*

The algorithm profile $\boldsymbol{A} = (A_1 \dots A_N)$ is the set of all players algorithms. We will note $A_i(t) = \pi_i^t$.

4.1 Multi-agent learning

Reinforcement learning provides a class of algorithms that aim at maximizing an agent's return. Out of all of them, our interest concerns Q-learning approaches (Watkins and Dayan 1992) for three reasons: they are model-free, off-policy and they are guaranteed to converge in finite state and action spaces. In a game $\mathcal{G}$, for a player i and given opponents policy π_{-i}, the basic idea is to learn a Q-function that approximates, for all states and actions, the average return starting from playing this action at this point while using the best strategy. Ideally, the Q-function Q_i associated with player i's policy that maximizes its return holds:

$$Q_i(s, a_i, a_{-i}) = r_i(s, a_i, a_{-i}) + \gamma \sum_{s'} \mathcal{P}(s'|s, a_i, a_{-i}) \max_{a_i'} Z_i(a_i', s', \pi_{-i})$$

where $Z_i(a_i, s, \pi_{-i}) = \sum_{a_{-i}'} \pi_{-i}(a_{-i}|s') Q_i(s', a_i', a_{-i}')$ is the expected value for gent i given its opponent policy. Q-learning algorithms are constructed in order to progressively approximate the Q-function without approximating the problem dynamics $\mathcal{P}$ and reward functions r, and without knowing the decision process that generated the historic buffer (in contrast, for example, to policy gradient algorithms (Williams 1992)). In finite state and action spaces, the approximation is obtained by successively applying the updates:

$$Q_i^{t+1}(s^t, a_i^t, a_{-i}^t) = Q_i^t(s^t, a_i^t, a_{-i}^t) + \alpha_t \left(r_i^t + \gamma \max_{a_i} Z_i(a_i, s^{t+1}, \pi_{-i}) - Q_i(s^t, a_i^t, a_{-i}^t) \right),$$

where α_t is the learning rate. However, when the opponent policy is not fixed, maximizing the Q-function with respect to actions is no longer an improvement of the policy (the response of the opponents to this deterministic policy can decrease the average player's return). MARL provides several alternative greedy improvements. For example, a defensive player can expect opponents to minimize its Q-function (*minimax* Q-learning). In that case, a greedy improvement of the policy to evaluate the value of a new state is obtained by solving the linear problem (Littman 1994):

$$\pi_i^{greedy}(.|s) = \operatorname*{argmax}_{\pi_i} \min_{a_{-i}} \sum_{a_i} \pi_i(a_i|s) Q_i(s, a_i, a_{-i}) \tag{2}$$
$$= \operatorname*{argmax}_{\pi_i} \min_{a_{-i}} Z_{-i}(a_{-i}, s, \pi_i)$$

and the corresponding Q-learning update becomes:

$$Q_i^{t+1}(s^t, a_i^t, a_{-i}^t) = Q_i^t(s^t, a_i^t, a_{-i}^t) + \alpha_t \left(r_i^t + \gamma \max_{\pi_i} \min_{a_{-i}} Z_{-i}(a_{-i}, s, \pi_i) - Q_i(s^t, a_i^t, a_{-i}^t) \right).$$

4.2 Learning equilibrium

We extend the notion of learning equilibrium (Brafman and Tennenholtz 2003) to repeated stochastic games as follows.

Definition 4 (Learning equilibrium)**.** *Let $\mathcal{G}$ be a set of stochastic games. An algorithm profile $\boldsymbol{A}^* = (A_1^* \dots A_N^*)$ is a learning equilibrium for $\mathcal{G}$ if, for any game $g \in \mathcal{G}$, there is a stage T_g such that, for any player i and any learning algorithm A_i:*

$$\mathbb{E}_{t>T_g}\left[\mathcal{R}_i\left(A_i(t), A_{-i}^*(t)\right)\right] \leq \mathbb{E}_{t>T_g}\left[\mathcal{R}_i\left(A_i^*(t), A_{-i}^*(t)\right)\right]$$

Consequently, just like Nash equilibrium for the choice of a strategy, no player can individually follow an alternative algorithm and increase its asymptotic score. However, one important difference is the fact that a learning algorithm is not defined with respect to a particular game, but a set of games.

We may think that a process always playing a Nash equilibrium of the given game ($\pi_i^t = \pi_i^*$ for all t) is a learning equilibrium. However, such a process requires an initial knowledge about the dynamics and the reward functions of the game and can't be obtained from a process starting with an empty condition. Therefore, it can't be described as a learning algorithm. For the same reason, a TFT process is not a learning equilibrium. However, we may construct learning algorithms that asymptotically behave as a TFT or always play a Nash equilibrium. This is the key idea of FCL.

5 Foolproof cooperative learning

As we are interested in forced cooperation, we are looking for a learning algorithm profile that converges to a TFT process, retaliating if a player deviates from a cooperative strategy. Since the objective of a cooperative strategy is a common quantity and TFT processes are symmetric, such a convergence can be obtained if all players are using the same algorithm. FCL, as described in Alg. 1 (for a player i), has the property to converge to such a behavior when played by all players. In an N-player game, each FCL player approximates $2N+1$ Q-functions: one associated with the cooperative policy that maximizes the sum of all players (Q^c), N associated with retaliation policies preventing any defection from other players j (Q_j^r), and N associated with each opponent's best response to the cooperative strategy (Q_j^d). At each played stage game, FCL will play according to a egalitarian cooperative strategy (learned through Q^c) unless one of the opponents deviates from that strategy. In case of an opponent's defection, all FCL agents will agree on a joint retaliation according to the *minimax* strategy (learned through Q_j^r with Eq.(2)) for K stages according to Eq.(1). In order to allow exploration, a deterministic process $\phi(t)$ is used to decide, at each time t, between exploration and exploitation. We design ϕ as a known realization of a random process such that explorations are endless ($\forall T\,,\, \exists t > T, \mathbb{P}[\phi(t) = \textbf{True}] > 0$), but becomes rare enough with time so the probability of explorations tends to zero ($\forall \epsilon > 0\,,\, \exists T_\epsilon\,,\, \forall t > T_\epsilon, \mathbb{P}[\phi(t) = \textbf{True}] < \epsilon$). This can be implemented using a pseudo-random process with a fixed seed, known by all FCL players. At exploration stages, all agents are allowed to perform any action without being accused of defection. In a way, this algorithm can be seen as a disentangled version of Friend-or-Foe Q-learning (FFQ) (Littman 2001) which learns to play cooperatively if an opponent is cooperative, or defensively if the opponent is defective with a single Q-function. However, FFQ can't learn a TFT behavior as it is either always cooperative, or always defensive. The following theorem 4 describe the asymptotically behaviour of FCL in RSGs. Theorem 5 states that FCL defines a learning equilibrium for RSGs.

Theorem 4. *Assume $\mathcal{S}$ and $\mathcal{A}_i$ are finite spaces and the opponents are exploring all possible state-action couples infinitely many times. Then, FCL converges to a TFT behavior forcing the egalitarian cooperative strategy in RSGs.*

Theorem 5. *FCL is a learning equilibrium for RSGs.*

6 Experiments

Despite our theoretical claims are established for any number of agents, we restrict our experiments to games involving two players. We first explore the case of three well known repeated symmetric matrix games: Iterated Prisoners Dilemma (IPD), Iterated Chicken (ICH) and Rock-Paper-Scissors (RPS). Table 1 shows the payoff matrices. Then, we investigate larger state spaces with grid games known to induce coordination problems and social dilemma (Munoz de Cote and Littman 2008). We introduced a new grid game, closer to the concept of limited resource appropriation: the Temptation game. In Temptation, making a movement to the sides can be seen as taking immediately the resource, while making a movement to the bottom can be seen as waiting for the winter. All grid games are described in details in Table 2. In order to verify that FCL is a learning equilibrium, we compare the score obtained by FCL and by selfish learning algorithm, Q-learning and policy-gradient (PG), against FCL. Indeed, we expect a learning equilibrium performing better than any other algorithm when the opponents are following the learning equilibrium.

6.1 Implementation details

We implemented FCL using a state-dependent learning rate $\alpha_t = (\sum_{l<t} \delta\{s^l = s^t\})^{-1}$ that counts the number of state visits, and exploration $\phi(t) = \{X_t > \epsilon d^t\}$ where

Algorithm 1 FCL for player i.

input List of counters $k_j = 0\ \forall j$ to repeat retaliations, exploration process $\phi(s,t)$, N-cyclic permutation σ, learning rate sequence $\{\alpha_t\}_t$, initial (arbitrary) functions Q^c, $\{Q_i^d\}_{i=1\ldots N}$ and $\{Q_i^r\}_{i=1\ldots N}$, initial state s.

for stages $t = 1$ **to** $+\infty$ **do**
  while stage continue **do**
    if $K_j = 0 \forall j$ **then**
      if $\phi(t)$ **then**
        Explore $a_i \sim \mathcal{U}(\mathcal{A}_i)$ with uniform probability
      else
        Take action $a_i = \text{argmax}_{a_{\sigma^t(i)}} \max_{a_{-\sigma^t(i)}} Q^c(s, a_i, a_{-i})$
      end if
    else
      Randomly select an agent j such that $K_j > 0$
      Take action $a_i \sim \text{argmin}_{\pi_{-j}} \max_{a_j} \sum_{a_{-j}} \pi_{-j}(a'_{-j}|s) Q_j^r(s, a_j, a_{-j})$
      $k_j \leftarrow k_j - 1$
    end if
    Observe a_{-i} and new state s', receive reward $r_i = r_i(s, a_i, a_{-i})$ and observe r_{-i}
    $Q^{c\prime} \leftarrow \max_{a'_i} \max_{a'_{-i}} Q^c(s', a'_i, a'_{-i})$
    $Q^c(s, a_i, a_{-i}) = Q^c(s, a_i, a_{-i}) + \alpha_t \left(\sum_{1 \le j \le N} r_j + \gamma Q^{c\prime} - Q^c(s, a_i, a_{-i}) \right)$
    for all other agents $j \neq i$ **do**
      $V_j^r(s') \leftarrow \min_{\pi_{-j}} \max_{a'_j} \sum_{a'_{-j}} \pi_{-j}(a'_{-j}|s') Q_j^r(s', a'_j, a'_{-j})$
      $V_j^d(s') \leftarrow \max_{a'_j} Q_j^d(s', a'_j, \text{argmax}_{a_{-j}} \max_{a'_j} Q^c(s', a_j, a_{-j}))$
      $Q_j^r(s, a_j, a_{-j}) = Q_j^r(s, a_j, a_{-j}) + \alpha_t \left(r_j + \gamma V_j^r(s') - Q_j^r(s, a_j, a_{-j}) \right)$
      $Q_j^d(s, a_j, a_{-j}) = Q_j^d(s, a_j, a_{-j}) + \alpha_t \left(r_j + \gamma V_j^d(s') - Q_j^d(s, a_j, a_{-j}) \right)$
      $K_j \leftarrow \left\lceil \frac{V_j^d(s') - V^c(s')}{V^c(s') - V_j^r(s')} \right\rceil$
      if not $\phi(t)$ **and** $a_j \neq \text{argmax}_{a_{\sigma^t(j)}} \max_{a_{-\sigma^t(j)}} Q^c(s, a_j, a_{-j})$ **then**
        $k_j \leftarrow k_j + K_j$
      end if
    end for
    $s \leftarrow s'$
  end while
end for

X_t is a pseudo-random uniform sample between 0 and 1 with a fixed seed, ϵ the initial threshold and d a decay parameter close to one. The closer is d to one, the longer lasts the exploration. For selfish Q-learning, we used a similar learning rate and exploration process, however with different seeds and decay parameters. The policy gradient was implemented with a tabular representation and Adam gradient descent with learning rate 0.1. Since matrix games are not sequential and since grid games were automatically reset after 30 steps, we could use a discount factor $\gamma = 1$ to estimate value functions. In practice, we found that adding 1 to the minimal number of retaliation repeats given in Eq. 1 significantly improves the robustness to selfish learners. In iterated matrix games, since they do not require large explorations, we used $\epsilon = 0.5$ and $d = 0.9$ for both selfish Q-learning and FCL. We used $\epsilon = 1$ and $d = 0.995$ in grid games.

6.2 Results

Figure 1 displays our results with the three matrix games IPD, ICH and RPS. Figure 2 displays our results on grid games. As expected, the score of selfish learners was never higher than the score of FCL, when the opponent is FCL. Except in RCP, defection conduced to less reward than cooperation because of retaliations. In RCP, FCL found the only way to retaliate by infinitely playing randomly against selfish learners, resulting in an average of 0 reward for all players, equivalent to the reward for cooperation. This illustrates the fact that FCL is a learning equilibrium, since no algorithm performs better than FCL against FCL. Consequently, FCL was never exploited by selfish learners while being cooperative in self-play.

Table 1: Payoff matrices used for IPD, ICH and RPS.

IPD

	Coop.	Defect
Coop.	(-1,-1)	(-3,0)
Defect	(0,-3)	(-2,-2)

ICH

	Straight	Swerve
Straight	(-3,-3)	(0,-2)
Swerve	(-2,0)	(-1,-1)

RPS

	Rock	Paper	Scissors
Rock	(0,0)	(-1,1)	(1,-1)
paper	(1,-1)	(0,0)	(-1,1)
Scissors	(-1,1)	(1,-1)	(0,0)

Table 2: Grid games. A is the starting position of one player, B is the starting position of the other. At each turn, both players simultaneously select one action among going up, down, left, right or stay. When reward cells with \$ symbol are reached by one player, the player obtains the corresponding reward and the game is immediately reset. $\$_{A:X}$ means that only player A gets the reward X when reaching the cell, $\$_X$ means that any player gets reward X when reaching the cell, and $\$_{X,Y}$ means that the player who reach the cell gets X and the other gets Y (if the other player reach another rewarding cell, the rewards are summed). Two players can not be on the same cell at the same time and they can not cross each other. In case of conflict, one player reaches the cell and the other stays with probability 0.5. Grey cells are walls and are not reachable.

(a) Grid prisoners dilemma

				$\$_{100}$				
$\$_{A:100}$			A		B			$\$_{B:100}$

(b) Compromise

		$\$_{B:100}$				$\$_{A:100}$		
		A				B		

(c) Coordination

$\$_{B:100}$		$\$_{A:100}$
A		B

(d) Temptation

$\$_{20,-10}$	A	B	$\$_{20,-10}$
$\$_{40,-20}$			$\$_{40,-20}$

7 Related work

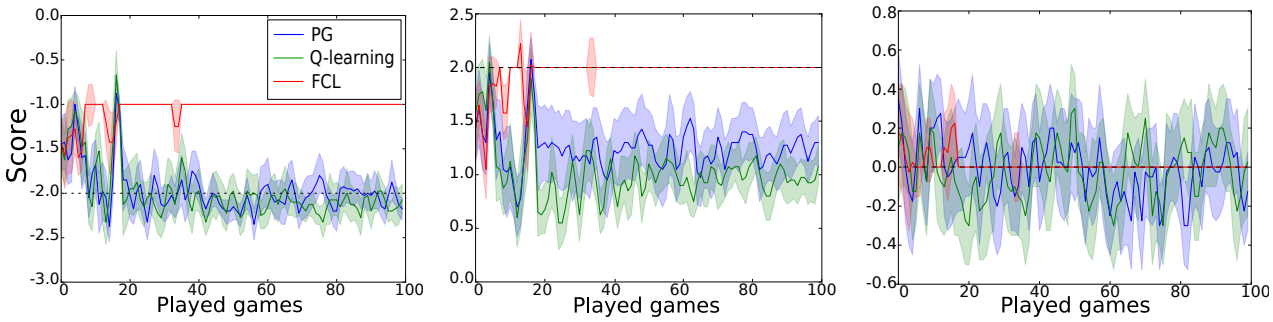
Learning cooperative behaviours in a multi-agent setting is a vast field of research, and various approaches depend on assumptions about the type of games, the type and number of agents, the type of cooperation and the initial knowledge.

When the game's dynamics is initially known and in two-player settings, Kalais' bargaining solution can be obtained by mixing dynamic and linear programming. Therefore, a polynomial-time algorithm can be used to solve repeated matrix games (Littman and Stone 2005), as well as repeated stochastic games (Munoz de Cote and Littman 2008). Since a bargaining solution is always better than a *minimax* strategy (the disagreement point) (Osborne and Rubinstein 1994), a cooperative equilibrium is immediately given. An alternative to our cooperate or retaliate architecture consists in choosing between maximizing oneself reward (being competitive) or maximizing a cooperative reward, for example by inferring opponents intentions (Kleiman-Weiner et al. 2016). The novelty of our approach is an online setting witch does not require the dynamics nor the reward function in order to construct a foolproof cooperative behaviour.

In games inducing social dilemmas and when the dynamics is accessible as an oracle, cooperative solutions can also be obtained by self-play and then applied to define a TFT behaviour forcing cooperation (Lerer and Peysakhovich 2017), even when opponent actions are unknown, since in that case the reward function already brings sufficient information (Peysakhovich and Lerer 2018). Here again, they use an offline procedure witch does not apply to our purely online setting.

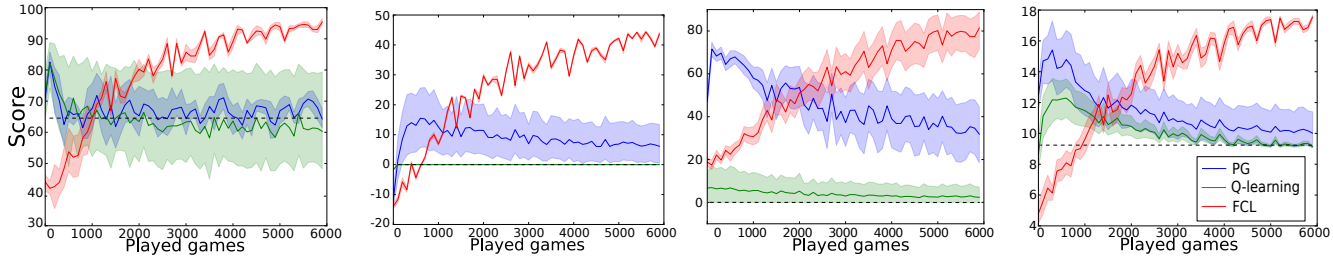
Closer to our setting, when the dynamics is unknown, online MARL can extract cooperative solutions in some non-cooperative games, and particularly in restricted resource appropriation (Pérolat et al. 2017). Using alternative objectives based on all players reward functions and their propensity to cooperate or defect improves and generalizes the emergence of cooperation in non-cooperative games and limits the risk of being exploited by purely selfish agents (Hughes et al. 2018). Regarding these approaches, one advantage of FCL is to disentangle the cooperative and the retaliating policies so it can always switch from one behaviour to the opposite without a forgetting and re-learning phase.

A similar approach, called Learning with Opponent Learning Awareness (LOLA), consists in modelling the strategies and the learning dynamics of opponents as part of the environment's dynamics and to derive the gradient of the average return's expectation (Foerster et al. 2018). If LOLA has no guaranties of convergence, a recent improvement of the gradient computation, which interpolates between first and second-order derivations, is proved to converge to a local optimums (Letcher et al. 2018). Although such agents are purely selfish, empirical results show that they are able to shape each others learning trajectories and to cooperate in prisoners dilemma. A limitation of this approach towards building learning equilibrium is the strong assumption regarding the opponents learning algorithms, supposed to perform policy gradient. Also, this approach differs to our goal since LOLA is selfish and aims at shaping an opponent's behavior (in 2-player settings) while FCL is cooperative



(a) Iterated prisoners dilemma (b) Iterated chicken (c) Rock paper scissors

Figure 1: Matrix games. Average scores over 20 runs obtained by two standard RL algorithms and FCL, playing against FCL. In IPD and ICH, after some iterations selfish behaviours, as induced by Q-learning and PG, start being sub-optimal because of FCL retaliations and accumulate less return than a cooperative behaviours, as induced by FCL against itself. In RPS, FCL learns to play with a uniform distribution against selfish algorithms so their average score is null. Black dotted line represents the average score after convergence of two selfish agent playing against themselves (the *minimax* solution).



(a) Prisoners dilemma (b) Compromise (c) Coordination (d) Temptation

Figure 2: Grid games. Average scores over 20 runs obtained by two standard RL algorithms and FCL, playing against FCL. After some iterations, selfish behaviours, as induced by Q-learning and PG, start being sub-optimal because of FCL retaliations and accumulate less return than a cooperative behaviour, as induced by FCL against itself. Black dotted line represents the average score after convergence of two selfish agents playing against themselves (the *minimax* solution).

but retaliates in response to selfish agents (in N-player settings).

Learning equilibrium solutions have been constructed for repeated matrix games (Brafman and Tennenholtz 2003; Ashlagi, Monderer, and Tennenholtz 2006). However, these solutions would not easily adapt to stochastic games, one main reason being the fact that exploration becomes infinite, while it only requires $A \times N$ steps in N-agents matrix games with A different actions. Consequently, after a finished phase of exploration in matrix games, the deterministic payoff matrix is known and they can extract a Nash Equilibrium to exploit. Note that the restriction to symmetric games seems recurrent in learning equilibrium literature (Brafman and Tennenholtz 2005; Tennenholtz and Zohar 2009). In repeated congestion games, it is even possible to construct a class of asymmetric games that does not admit any learning equilibrium, hence the importance of the symmetry assumption.

8 Conclusion

We introduced FCL, a model-free learning algorithm that, by construction, converges to a TFT behaviour, cooperative against itself and retaliating against selfish algorithms. We proposed a definition for learning equilibrium, describing a class of learning algorithms such that the best way to play against it is to adopt the same behaviour. We demonstrated that FCL is a learning equilibrium that forces a cooperative behaviour, and we empirically verified this claim with two-agents matrix games and grid-world repeated symmetric games.

Our approach could be improved by facilitating opponent's learning of the optimal cooperative response and by using faster learning approaches. It could also be adapted to larger dimensions such as continuous state spaces and partially observed settings with function approximation by replacing tabular Q-learning with deep Q-learning (Mnih et al. 2015). In that perspective, the main limitation relies on the necessity to compute the minimax strategy using a linear programming approach.

References

Ashlagi, I.; Monderer, D.; and Tennenholtz, M. 2006. Robust learning equilibrium. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Axelrod, R., and Hamilton, W. D. 1981. The evolution of cooperation. *science* 211(4489):1390–1396.

Bowling, M., and Veloso, M. 2001. Rational and convergent learning in stochastic games. *Proceedings of the International joint conference on artificial intelligence*.

Brafman, R. I., and Tennenholtz, M. 2003. Efficient learning equilibrium. *Advances in Neural Information Processing Systems*.

Brafman, R. I., and Tennenholtz, M. 2005. Optimal efficient learning equilibrium: Imperfect monitoring in symmetric games. *Proceedings of the National Conference on Artificial Intelligence*.

Claus, C., and Boutilier, C. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. *Proceedings of the Association for the Advancement of Artificial Intelligence*.

Dasgupta, P.; Maskin, E.; et al. 1986. The existence of equilibrium in discontinuous economic games. *Review of Economic Studies*.

Foerster, J.; Chen, R. Y.; Al-Shedivat, M.; Whiteson, S.; Abbeel, P.; and Mordatch, I. 2018. Learning with opponent-learning awareness. *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*.

Hughes, E.; Leibo, J. Z.; Phillips, M.; Tuyls, K.; Dueñez-Guzman, E.; Castañeda, A. G.; Dunning, I.; Zhu, T.; McKee, K.; Koster, R.; et al. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in Neural Information Processing Systems*.

Kalai, E. 1977. Proportional solutions to bargaining situations: interpersonal utility comparisons. *Econometrica: Journal of the Econometric Society*.

Kleiman-Weiner, M.; Ho, M. K.; Austerweil, J. L.; Littman, M. L.; and Tenenbaum, J. B. 2016. Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. *Proceedings of Annual Conference of the Cognitive Science Society*.

Lerer, A., and Peysakhovich, A. 2017. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068*.

Letcher, A.; Foerster, J.; Balduzzi, D.; Rocktäschel, T.; and Whiteson, S. 2018. Stable opponent shaping in differentiable games. *Proceedings of the International Conference on Learning Representations*.

Littman, M. L., and Stone, P. 2005. A polynomial-time nash equilibrium algorithm for repeated games. *Decision Support Systems*.

Littman, M. L. 1994. Markov games as a framework for multi-agent reinforcement learning. *Proceedings of the International Conference on Machine Learning*.

Littman, M. L. 2001. Friend-or-foe q-learning in general-sum games. *Proceeding of the International Conference on Machine Learning*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature*.

Munoz de Cote, E., and Littman, M. L. 2008. A polynomial-time Nash equilibrium algorithm for repeated stochastic games. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Nash Jr, J. F. 1950. The bargaining problem. *Econometrica: Journal of the Econometric Society*.

Nash, J. 1951. Non-cooperative games. *Annals of mathematics*.

Osborne, M. J., and Rubinstein, A. 1994. *A course in game theory*. MIT press.

Pérolat, J.; Leibo, J. Z.; Zambaldi, V.; Beattie, C.; Tuyls, K.; and Graepel, T. 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. *Advances in Neural Information Processing Systems*.

Peysakhovich, A., and Lerer, A. 2018. Consequentialist conditional cooperation in social dilemmas with imperfect information. *Proceedings of the International Conference on Learning Representations*.

Shapley, L. S. 1953. Stochastic games. *Proceedings of the national academy of sciences*.

Tennenholtz, M., and Zohar, A. 2009. Learning equilibria in repeated congestion games. *Proceedings of The International Conference on Autonomous Agents and Multiagent Systems*.

Vester, S. 2012. *Symmetric Nash Equilibria*. Ph.D. Dissertation, Master thesis from Ecole Normale Superieure de Cachan.

Watkins, C. J., and Dayan, P. 1992. Q-learning. *Machine Learning*.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*.