Towards Well-Defined Multi-agent Reinforcement Learning*

Rinat Khoussainov

Department of Computer Science, University College Dublin
Belfield, Dublin 4, Ireland
rinat@ucd.ie

Abstract. Multi-agent reinforcement learning (MARL) is an emerging area of research. However, it lacks two important elements: a coherent view on MARL, and a well-defined problem objective. We demonstrate these points by introducing three phenomena, social norms, teaching, and bounded rationality, which are inadequately addressed by the previous research. Based on the ideas of bounded rationality, we define a very broad class of MARL problems that are equivalent to learning in partially observable Markov decision processes (POMDPs). We show that this perspective on MARL accounts for the three missing phenomena, but also provides a well-defined objective for a learner, since POMDPs have a well-defined notion of optimality. We illustrate the concept in an empirical study, and discuss its implications for future research.

1 Introduction

Multi-agent reinforcement learning (MARL) addresses the following question: "How can an agent learn to act optimally in an *a priori* unknown dynamic environment through trial-and-error interaction in the presence of other self-interested and possibly adapting agents?" There are numerous practical application domains for MARL, such as robotic soccer, electronic market places, self-managing computer systems, distributed Web search, or even military/counter-terrorism applications. While single-agent RL has been an active area of research in AI for many years, MARL is a relatively new field.

We do not have sufficient space in this paper to provide a detailed survey of the prior work in MARL, but we can point out two important issues in the current state of research: the lack of a coherent view on MARL; and the lack of a well-defined objective for a MARL algorithm. What behaviour should an agent learn in a multi-agent environment? In single-agent RL, this problem is well-defined, since the agent's performance depends only on its own actions. In a MA environment, the agent's performance may depend on simultaneous actions of others. Thus, a given agent would need to know what to expect from the other agents, and the optimal behaviour may change as the other agents also evolve (adapt) over time.

The question of a clear research agenda has been already emphasised by several RL researchers and is becoming a topic of much ongoing debate in the area [1]. The

* Thanks to Nicholas Kushmerick for helpful discussions and valuable comments. This research was supported by grant SFI/01/F.1/C015 from Science Foundation Ireland, and grant N00014-03-1-0274 from the US Office of Naval Research.

C. Bussler and D. Fensel (Eds.): AIMSA 2004, LNAI 3192, pp. 399–408, 2004.

contributions of this paper are as follows. We demonstrate the lack of a coherent view on MARL by introducing three phenomena inadequately addressed by the previous research: social norms, teaching, and bounded rationality. While these phenomena are not new to MARL, none of the previous works attempted to systematically account for all three together, and many papers ignore all of them. Based on the ideas of bounded rationality [2], we define a very broad class of "realistic MA learners", and show that MARL in this case is equivalent to learning in partially observable Markov decision processes (POMDPs). Such perspective on MARL accounts for the above phenomena, but also provides a well-defined objective for a learner, since POMDPs have a well-defined notion of optimality. Finally, we illustrate the proposed concept in an empirical study, and discuss its implications for future research.

2 Formal Framework and Definitions

The standard RL model [3] consists of the two main components: a learning agent and a dynamic environment. The agent is connected to its environment via perception and action signals. The interaction with the environment proceeds in steps. At each step, the agent receives as input some indication (observation) of the current state of the environment and then generates an output by choosing some action. The action changes the state of the environment, and the value of this state transition is communicated back to the agent using some scalar reinforcement signal (*reward*, or *payoff*).

In a MA scenario, there may be several agents acting in the same environment, and their actions may affect not only the state transitions, but also the rewards of each other. Formally, a MA environment is well modelled as a *stochastic game* (SG) [4]. A stochastic game is a tuple $\langle I, S, s_0, (A_i)_{i=1}^I, Z, (u_i)_{i=1}^I \rangle$, where I is the number of agents, also called *players* in the game, S is a set of the game states, s_0 is the initial state, and A_i is the set of actions available to player i. If a_i is the action selected by player i at some step, then $a = (a_i)_{i=1}^I$ is an *action profile* (or a *joint action*). A set $A = A_1 \times A_2 \times \ldots \times A_I$ is the set of all possible action profiles. $u_i : S \times A \to \mathbb{R}$ is a payoff function of player i. $Z : S \times A \times S \to [0, 1]$ is a stochastic state transition function, which for given state and action profile returns a probability distribution over the possible next states. That is, $Z(s, a, s')$ gives the probability that after performing joint action a in state s, the next state of the game will be s'.

Stochastic games are a generalisation of Markov decision processes (MDPs) [3] to multiple decision makers. In general, the environment can be non-deterministic, i.e. taking the same joint action in the same state may result in different state transitions and/or reinforcements. However, it is usually assumed that the probabilities of making state transitions or receiving specific rewards do not change over time. A *game history* in a stochastic game is a sequence of tuples $\langle s, a \rangle$, where a is the players' action profile (joint action) at some step of the game and s is the next state.

The goal of the agent is to find a decision rule (a *policy*) that maximises some long-term measure of the reinforcement. From the agent's point of view, there are two types of environments: *fully observable* environments and *partially observable* environments. In fully observable environments, the agent can reliably observe the game history at each step. In partially observable environments, agents have incomplete information about the

game. For example, an agent may not be able to distinguish between some states of the environment based on his observations, or he may not be able to observe actions of other players. An *observation function* Ω_i is a mapping from a set of possible elements of the game history to a set O_i of possible observations for player i, $\Omega_i : S \times A \times O_i \to [0, 1]$. $\Omega_i(s, a, o_i)$ is the probability that player i receives observation o_i when the game state changes to s after agents realise action profile a.

A policy (or strategy) of a player in a stochastic game is a function that maps possible sequences of observations of a player to probability distributions over the player's actions: $\Lambda_i : \mathcal{O}_i \times A_i \to [0, 1]$, where $\mathcal{O}_i$ is a set of possible observation sequences $\{(o_i(k))_k : o_i(k) \in O_i\}$ of player i. If the mapping is deterministic ($\Lambda_i : \mathcal{O}_i \to A_i$), then the strategy is called *deterministic*. A strategy of a player in a stochastic game is called *Markov* or *reactive* if it is a function only of the current observation $\Lambda_i : O_i \times A_i \to [0, 1]$ (i.e. if the player's action at a given step depends only on the current observation of the game at that step).

Let $(r_i(k))_{k=1}^K$ be a sequence of rewards received by the agent i over K steps of interaction with the environment. The *long-term reward* for a given sequence of rewards from separate steps can be calculated as a *discounted sum* $U_i^K = \sum_{k=1}^K \gamma^{k-1} r_i(k)$, where $0 < \gamma \leq 1$ is a discount factor, or as *average reward*: $U_i^K = 1/K \sum_{k=1}^K r_i(k)$. The case when the long-term performance is analysed over infinitely many steps is called the *infinite horizon* model. The long-term payoff in the infinite horizon case is evaluated as a limit for $K \to \infty$: $U_i^\infty = \lim_{K\to\infty} U_i^K$. The infinite-horizon discounted sum model has received the most attention in RL due to its mathematical tractability.

3 Optimality in Stochastic Games

Unlike in the single-agent case (MDPs), the optimal behaviour in stochastic games is in general opponent-dependent. That is, the long-term payoff of a player's strategy can only be evaluated given the strategies of the other players in the game. Therefore, the notion of optimality has to be replaced with the notion of *best response*. A player's strategy is the best response to a combination of strategies of other players (opponents) in a game, if it maximises the player's payoff for the given strategies of the opponents. Consequently, a strategy is optimal, if it is the best response to the combination of strategies used by other players in the game. A fundamental problem is that a player cannot know what strategies opponents will use, because he has no direct control over their behaviour (all players are independent in selecting their strategies). Moreover, the opponents' strategies may change over time as the opponents adapt their behaviour.

Game theory attempts to resolve this uncertainty using the concept of Nash equilibrium [5]. Nash equilibrium captures a situation in a game in which each player holds the correct expectations about the other players' behaviour and acts the best response to those expectations. Let $\hat{U}_i(\Lambda_1, \Lambda_2, \cdots, \Lambda_I)$ be the expected long-term payoff of player i for the given players' strategies combination $(\Lambda_j)_{j=1}^I$. A strategy combination $(\Lambda_j^*)_{j=1}^I$ is a Nash equilibrium, if $\hat{U}_i(\Lambda_1^*, \cdots, \Lambda_{i-1}^*, \Lambda_i, \Lambda_{i+1}^*, \cdots, \Lambda_I^*) \leq \hat{U}_i((\Lambda_j^*)_{j=1}^I)$, for any strategy Λ_i for all players i. That is, no player can increase its payoff by unilaterally deviating from the equilibrium. The idea behind Nash equilibrium is that if there was a mechanism that correctly predicts for a player what strategies his opponents will use, then

the main requirement for such a mechanism would be that it should not be self-defeating (i.e. knowing the predictions of this mechanism players should not have incentives to contradict it). Thus a valid prediction must be a Nash equilibrium.

However, the predictive power of the Nash equilibrium is limited by two problems. The first problem is the possible multiplicity of equilibria. When there are multiple equilibria in a game, players need to reach a coordinated equilibrium outcome to play optimally. The second problem is the assumption that all players in the game have equal reasoning capabilities. If one player can analyse the game to correctly predict the behaviour of other players, then the other players also have these abilities to analyse the game. That is, we assume that all players are equally "clever" (equally rational) and possess the same necessary information about the game to derive their expectations. Generally speaking, this assumption may not always hold in practice.

4 Previous Work

Many learning methods in SGs focused on extending traditional RL algorithms like Q-learning [3] to MA settings. Most of results, however, are for a limited class of single-state SGs, called *repeated games*. Unlike in single-agent RL, the long-term payoff of an action in a SG depends on the strategies of opponents, so the learner has to make some assumptions about the opponents' behaviour. In *Minimax-Q* [6], a player learns to select at each step the action that maximises the reward the player can guarantee. However, this strategy works well only in zero-sum games (two-player games where the sum of players' payoffs at every step is zero). *Nash-Q* [7] extends Minimax-Q to general-sum games by choosing actions that are a part of some Nash equilibrium of the constituent one-step game. Since there may be many Nash equilibria in a game, the analysis is restricted to special cases of games with globally optimal points and saddle points, where equilibrium selection is simplified.

Opponent modellers try to model the opponents' strategies from interaction experience and construct a best-response strategy to those models. Examples of this approach are joint action learners (JALs) [8] and finite automata modellers [9]. These methods can work well when the opponents' strategies are fixed, but are questionable when players learn. *PHC* and *WoLF-PHC* [10] are two algorithms for MARL in fully observable domains based on a policy search approach. PHC is essentially an extension of Q-learning to non-deterministic Markov polices. WoLF PHC uses variable learning rate to encourage convergence of the algorithm to a Nash equilibrium in self-play.

5 Multi-agent Learning Revisited

There are several phenomena that are addressed inadequately in MARL. We subdivide them into social norms, teaching considerations, and bounded rationality. While these issues are not entirely new to MARL, none of the previous approaches provides a complete and systematic coverage of them, and too frequently they are simply ignored.

Social norms. SGs have a rich structure of equilibrium behaviour that may be interpreted in terms of a "social norm" [5]. The idea is that players can sustain mutually desirable outcomes, if their strategies involve "punishing" any player whose behaviour

Table 1. Prisoner's dilemma

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

Table 2. The row player prefers to "teach" the column player that her strategy is a

	b	$\bar{b}$
a	1,0	3,1
$\bar{a}$	2,2	4,0

is undesirable. This is regulated by the "folk theorems"[1] originally proposed for single-state SGs, called *repeated games*. Two-player repeated games can be described by a game matrix that specifies the players payoffs for each possible joint action. Consider the example of repeated Prisoner's dilemma (RPD) [5] in Table 1. At each step of the game, players select one of the two possible actions C or D and receive the respective payoffs as specified in the table. At the next step, the same game repeats.

A one-step Prisoner's dilemma has a unique Nash equilibrium where both players choose C. Indeed, no matter what action the opponent selects, the player is always better off choosing C. In the case of RPD however, the outcome (D, D) can also be sustained as equilibrium, if each player punishes the opponent for deviation by also switching to C. If both players follow such strategy, then it is better for them to play (D, D) repeatedly, receiving payoff of 3 at each step, rather than getting the higher deviation payoff of 4 in a single period, but receiving a lower payoff of 1 in all subsequent steps. The formal results below are derived for infinite-horizon average payoff games (called *limit of means games*), though similar results exist for discounted games too.

A vector (x_i) is called a feasible long-term payoff profile in a limit of means SG with initial state s_0, if there is a combination of players' strategies (Λ_i), such that $x_i = \hat{U}_i(s_0, (\Lambda_i)) = \lim_{K\to\infty} 1/K \sum_{k=1}^{K} \hat{r}_i(s_0, (\Lambda_i), k)$, where $\hat{r}_i(s_0, (\Lambda_i), k)$ is the expected payoff of player i at stage k for the given initial state and players' strategies. Let Λ_{-i} denote a strategy profile of all players except i. We define the *minimax payoff* $\mu_i(s_0)$ of player i in a limit of means SG with initial state s_0 as $\mu_i(s_0) = \min_{\Lambda_{-i}} \max_{\Lambda_i} \hat{U}_i(s_0, (\Lambda_i, \Lambda_{-i}))$. Essentially, the strategy profile Λ_{-i} corresponding to the minimax is the most severe punishment that other players can inflict on i for the given initial state. A payoff profile (x_i) is *strictly enforceable* in SG with initial state s_0, if it is feasible and $x_i > \mu_i(s_0)$ for all i. Intuitively, enforceable profiles are the outcomes where each player can be punished, and so threats of punishment can sustain the outcome as an equilibrium.

Proposition 1 (Folk theorem for limit of means stochastic games)
Define the following assumptions: (A1) *– the set of feasible long-term average payoff profiles is independent of the initial state;* (A2) *– the long-term average minimax payoff of each player is independent of the initial state. For a given limit of means stochastic game and assumptions (A1)–(A2), any strictly enforceable long-term average payoff profile is a Nash equilibrium payoff profile of the game [11].*

Informally, assumption (A2) tells us that the punishment is always effective for all players, while assumption (A1) tells that the desired payoff profile can be achieved from any state (hence, it can be achieved after punishment as well). Since repeated games have

[1] The term used in the game theory community, since their originator is apparently unknown.

only a single state, and the feasible payoff profiles in a limit of means repeated game are the same as in the constituent one-shot game, any strictly enforceable payoff profile in the constituent one-shot game is a Nash equilibrium payoff profile of the repeated game [12]. An example of such a punishing strategy for RPD is *Tit-for-Tat* [13]. Tit-for-Tat starts by playing D and then plays the same action that the opponent in the previous period. The optimal strategy against Tit-for-Tat is to always play D.

Folk theorems (especially for repeated games) have been known in game theory for a long while. However, they received very little attention in MARL. Implementing punishing strategies requires the players to be able to condition their behaviour on a possibly long game history. At the same time, most MARL algorithms focused on learning only reactive policies which may be inadequate for sustaining social norms.

Teaching. Learning in games has been studied extensively in game theory as well. However, in game theory learning was used as an alternative way to explain the concept of equilibrium as a long-term outcome arising out of a process in which less than fully rational players search for optimality over time. An important point here is that actions of a player influence the learning and, hence, the future play of the opponents. That is, a player is learning himself and simultaneously teaching other players. Thus, in such environments the players ought to consider not only how their opponents may play in the future, but also how players' current behaviour will affect the future play of the opponents. As a result, a player's strategy may become to "teach" the opponents to play a best response to a particular action by playing that action over and over.

Consider an example repeated game from [14] described by Table 2. Since action $\bar{a}$ dominates a for the row player, a row player who ignores considerations of the repeated play will choose $\bar{a}$ as her strategy. Consequently, the column player will eventually learn to play b, because b maximises his payoff for the given strategy of the row player. Hence, the learning process will converge to outcome $(\bar{a}, b)$, where the row player's payoff is 2. However, if the row player is patient and knows that the column player "naively" chooses his strategy to maximise his own payoff given the history of the row player's actions, then the row player can do better by always playing a. This will lead the column player to choose $\bar{b}$ as his strategy, yielding a payoff of 3 to the row player. Essentially, a "sophisticated" and patient player facing a naive opponent can develop a "reputation" leading the learning process to a desired outcome.

Most of game theory, however, ignores these teaching considerations, explicitly or implicitly relying on a model in which the incentive to try to alter the future play of opponents is negligible. One class of such models that make the teaching considerations negligible are *large populations*. In large population models, opponents for each period are chosen from a large population of players making interaction relatively anonymous. Unlike game theory, MARL should take into account the fact that the actions of a given agent influence the learning and, hence, the future behaviour of other agents in the same environment. However, most existing MARL algorithms either ignore these teaching considerations, or use them only implicitly, for example to analyse convergence properties of the learning process in self-play (i.e. when opponents use the same learning algorithm). As shown in [15], teaching can actually allow an agent to exploit his opponents and achieve better individual performance.

Bounded rationality. The type of rationality usually assumed in game theory is perfect, logical, *deductive* rationality. Such deductive rationality presumes that all deci-

sion makers possess full knowledge of the game, unlimited abilities to analyse it, and can perform the analysis without mistakes. *Bounded rationality* [2] explicitly assumes that the reasoning capabilities of decision makers are limited, and therefore, they do not necessarily behave optimally in the game-theoretic sense. Bounded rationality proposes *inductive* instead of deductive reasoning.

Theoretical reasons for bounded rationality can be subdivided into knowledge limitations (i.e. players may not have sufficient knowledge of the game to compute an equilibrium) and computational limitations. For example, there is a large body of research on equilibrium selection in game theory using various principles and criteria [16]. However, a generic assumption that players are payoff maximisers is not sufficient to select a unique outcome. Also, it may ultimately require characterising all Nash equilibria of a game. The task is NP-hard even for simple one-shot games and even given complete information about the game [17]. In real world, agents may have physical limitations as well (e.g. faulty acting parts) preventing them from implementing certain strategies.

The main implication of bounded rationality is that even when the deductive game-theoretic reasoning gives an unambiguous answer (e.g. the game has a unique Nash equilibrium), there is still a possibility that opponents will not be able or willing to realise it, and so playing the equilibrium strategy will not be optimal. Though one could expect that machine learning should rely on inductive reasoning, still many MARL algorithms specifically focus on learning some equilibria of a game, resulting in quite limited application scenarios and unclear supporting motivation [1]. Only recently RL researchers have started to recognise and address this issue [10].

6 Learning with Realistic Agents

The discussion in the previous section indicates the lack of two important elements in MARL: a coherent view incorporating the described phenomena, and a well-defined objective (a problem statement) for a learner. What should an agent learn in a MA environment, or what behaviour is optimal? We propose here a novel perspective on MARL that gives a possible answer to these questions.

Multi-agent learning as a partially observable MDP. In general, one can view a reinforcement learner as two interacting components: a behaviour policy that is responsible for action selection, and an update algorithm that uses the learner's past experience to modify the policy so as to improve its long-term performance (see Figure 1). In particular, the update algorithm can use information about game observations, rewards, and actions performed by the agent, as well as the dynamics of the current policy and some prior knowledge of the game (if available).

So far, we have focused on learning a policy that yields the best performance in the given environment. However, Figure 1 clearly illustrates that a reinforcement learner as a whole is nothing but a decision rule mapping observations (including reward observations) onto the agent's actions. That is, learning is simply a special form of acting where actions are conditioned on the learning history. A reinforcement learning algorithm can essentially be viewed as a strategy in a "super" game, where players' actions are possible learning strategies, and payoffs are related to the outcomes of the learning process with a given combination of strategies.

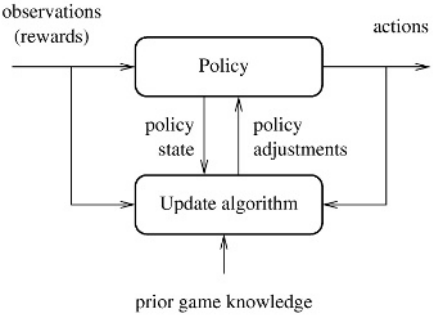

Fig. 1. Inside a reinforcement learner

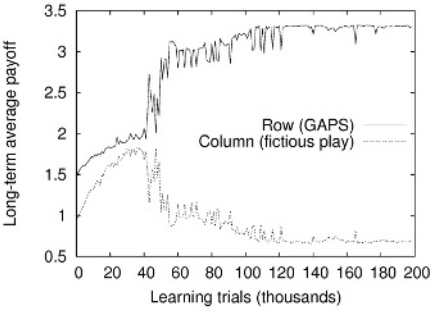

Fig. 2. Learning in the "teaching" game

What assumptions can we make about possible learning strategies? Perhaps the most generic assumption we can make is that strategies should be *computable* (i.e. implementable by a Turing machine). The second assumption we make is that strategies should be implementable by realistic agents. Realistic agents can only have a finite amount of memory available to them (equivalently, the corresponding Turing machines can only have finite tapes). Taking into account that a Turing machine with a finite tape is expressively equivalent to a finite state automaton (FSA), we obtain that realistic learning strategies should be representable by a (possibly non-deterministic) FSA.

It is easy to see that from a single agent's point of view a Markov environment augmented with several FSAs remains Markov, since at each step the immediate reward and the next state of the environment depend only on the agent's action, and the current combined state of the environment and FSAs. Therefore, learning in such environment (i.e. MA learning with realistic agents) becomes equivalent to learning in a *partially observable MDP* (POMDP). The MDP should be partially observable, because it is unlikely that an agent will be able to observe state transitions inside other learners.

Discussion. Treating MARL as learning in POMDPs allows us to systematically account for all three previously discussed phenomena. Indeed, social norms become just a kind of history-dependent behaviour in a POMDP, and it is well known that in POMDPs history-dependent policies can achieve better performance [18]. Getting high rewards in MDPs requires an agent to bring the environment into "well-paying" states. Since now other learners are just a part of the environment from a given agent's point of view, teaching can be viewed as changing the state of other agents to get better rewards. Finally, bounded rationality served as a defining assumption for the realistic learners in our model. The proposed perspective on MARL provides a well-defined objective for a MA learner, because POMDPs have a well-defined notion of optimality

Similar considerations have been proposed in [1], where the authors put forward three research agendas: distributed AI (designing an adaptive procedure for decentralised control of a set of agents); equilibrium agenda (investigating whether given learning strategies are in equilibrium); and AI agenda (finding the best learning strategy against a fixed class of opponents). By defining such a fixed class of opponents as "realistic learners", we have merged these three agendas. Equilibrium strategies and optimal strategies are equivalent in POMDPs. The distributed AI agenda can be viewed as common-payoff

games, where all agents have the same payoff functions, and thus it ultimately reduces to finding the best strategy in the corresponding POMDP. The proposed concept does not make MARL easier, but it does provide a coherent view on the previous approaches and possible future directions. From the POMDP perspective, MA learning algorithms that try to take into account the possibility that other agents' strategies may change over time, simply try to account (in somewhat specific ways) for the hidden state of the environment.

Empirical illustration. To illustrate the proposed concept on a concrete example, we used the teaching game in Table 2, where the column player uses the following simple learning algorithm (also known as fictious play in game theory [14]). It assumes that the row player's strategy is a stationary probability distribution over his actions. The learner tries to empirically estimate this distribution and plays the best response to the current estimation. Let $C(x)$ be the number of times the row player selected action x in the past. Then the column player estimates the probability $\Pr(a)$ that the row player selects a as $C(a)/(C(a) + C(\bar{a}))$, and the expected payoff for playing action b as $2\Pr(a)$. Similar calculations are done for the $\bar{b}$ action. Hence, the strategy of the column player is to choose b if $2C(\bar{a}) > C(a)$, and $\bar{b}$ otherwise.

The row player used the GAPS algorithm [19] originally developed for learning in POMDPs. GAPS belongs to the class of policy search algorithms which essentially perform search in a (usually restricted) space of possible policies for the one that gives the highest long-term reward [3]. In GAPS, the learner plays a parameterised strategy represented by a non-deterministic FSA, where the parameters are the probabilities of actions and state transitions. The automaton's inputs are game observations, the outputs are the player's actions. GAPS implements stochastic gradient ascent in the space of policy parameters. After each learning trial, parameters of the policy are updated by following the reward gradient. The advantage of GAPS is that it can learn history-dependent policies, since past observations can be memorised in the FSA state.

Each learning trial consisted of 300 steps, and the GAPS policy had 4 states. The players' observations consisted of the action taken by their opponent at the previous step. We used average payoff to evaluate the long-term performance of the players. As pointed out in Section 5, a clever player can teach the column learner that his strategy is a by repeatedly playing it, and thus encourage the column player to choose $\bar{b}$. This behaviour yields the row player the average payoff of 3 instead of the one-shot Nash equilibrium payoff of 2 for $(\bar{a}, b)$. Figure 2 shows the learning curves for both players.

As we can see, the GAPS learner managed to achieve even higher average payoff close to 3.3. A detailed investigation of the players' behaviour shows that the row player found a strategy that was following a 3-step loop, playing a in two periods and $\bar{a}$ in the third one. Since $C(a) = 2C(\bar{a})$ for such strategy, the column player always selects $\bar{b}$. The row player receives the long-term average payoff of $(3 + 3 + 4)/3 = 3.3\cdots$. Thus, not only GAPS has taught the column player to choose $\bar{b}$, but it also exploited this behaviour to further increase the payoff. However, unlike e.g. PHC-Exploiter in [15], GAPS knew neither the game being played, nor the learning algorithm used by the opponent or whether the opponent was learning at all. Of course, this success was only possible because GAPS learns stateful (non-Markov) policies that could account for the hidden state of the column learner in the resulting POMDP. While this is a very simple example, still it provides a good demonstration for the proposed MA learning concept.

7 Conclusions

In this paper, we used the ideas of bounded rationality to define MARL as learning in a POMDP. Such perspective on MARL allows us to account for many phenomena inadequately addressed by the previous research. Most importantly, the proposed concept gives a well-defined objective for a learner. Unfortunately, POMDPs are intractable in general, so the concept in itself is not a "silver bullet" for the MARL problem. However, it provides a clear structure and direction for future efforts, which in our opinion should focus on learning how to deal with the hidden environment state, rather than explicitly trying to converge to some Nash equilibrium in self-play. For example, it emphasises the importance of learning history-dependent policies, or trying to infer the hidden environment state, e.g. by maintaining beliefs about opponents.

One can imagine that in real world settings, learners may not only change their strategies, but also become more "computationally clever", e.g. by acquiring more memory. Thus, an interesting direction for future work is to investigate such extensions to the proposed framework.

References

1. Shoham, Y., Grenager, T., Powers, R.: Multi-agent reinforcement learning: A critical survey. Tech.rep., Stanford University (2003)
2. Simon, H.: Models of Man. Social and Rational. John Wiley and Sons (1957)
3. Sutton, R., Barto, A.: Reinforcement Learning: An Introduction. MIT Press (1998)
4. Filar, J., Vrieze, K.: Competitive Markov Decision Processes. Springer Verlag (1997)
5. Osborne, M., Rubinstein, A.: A Course in Game Theory. The MIT Press (1999)
6. Littman, M.: Markov games as a framework for multi-agent reinforcement learning. In: Proc. of the 11th Intl. Conf. on Machine Learning. (1994)
7. Hu, J., Wellman, M.P.: Nash Q-learning for general-sum stochastic games. Journal of Machine Learning Research **4** (2003)
8. Claus, C., Boutilier, C.: The dynamics of reinforcement learning in cooperative multiagent systems. In: Proc. of the 15th AAAI Conf. (1998)
9. Carmel, D., Markovitch, S.: Learning models of intelligent agents. In: Proc. of the 13th AAAI Conf. (1996)
10. Bowling, M.: Multiagent Learning in the Presence of Agents with Limitations. PhD thesis, Carnegie Mellon University (2003)
11. Dutta, P.K.: A folk theorem for stochastic games. J. of Economic Theory **66** (1995)
12. Rubinstein, A.: Equilibrium in supergames. In: Essays in Game Theory. Springer-Verlag (1994)
13. Axelrod, R.: The Evolution of Cooperation. Basic Books (1984)
14. Fudenberg, D., Levine, D.K.: The Theory of Learning in Games. The MIT Press (1998)
15. Chang, Y., Kaelbling, L.P.: Playing is believing: The role of beliefs in multi-agent learning. In: Advances in Neural Information Processing Systems. Volume 14., The MIT Press (2001)
16. Harsanyi, J., Selton, R.: A General Theory of Equilibrium Selection in Games. The MIT Press (1988)
17. Conitzer, V., Sandholm, T.: Complexity results about Nash equilibria. In: Proc. of the 18th Intl. Joint Conf. on AI. (2003)
18. Singh, S., Jaakkola, T., Jordan, M.: Learning without state-estimation in partially observable Markovian decision processes. In: Proc. of the 11th Intl. Conf. on Machine Learning. (1994)
19. Peshkin, L., Meuleau, N., Kim, K.E., L.Kaelbling: Learning to cooperate via policy search. In: Proc. of the 16th Conf. on Uncertainty in AI. (2000)