Reinforcement Learning for Nash Equilibrium Generation (Extended Abstract)

David Cittern
Imperial College London, UK
david.cittern10@imperial.ac.uk

Abbas Edalat
Imperial College London, UK
a.edalat@imperial.ac.uk

ABSTRACT

We propose a new conceptual multi-agent framework which, given a game with an undesirable Nash equilibrium, will almost surely generate a new Nash equilibrium at some predetermined, more desirable pure action profile. The agent(s) targeted for reinforcement learn independently according to a standard model-free algorithm, using internally-generated states corresponding to high-level preference rankings over outcomes. We focus in particular on the case in which the additional reward can be considered as resulting from an internal (re-)appraisal, such that the new equilibrium is stable independent of the continued application of the procedure.

Categories and Subject Descriptors

I.2.11 [**Computing Methodologies**]: Distributed Artificial Intelligence—*Multiagent systems*

General Terms

Algorithms, Design, Experimentation, Theory

Keywords

Single and multi-agent learning techniques, Computational architectures for learning, Reward structures for learning

1. INTRODUCTION

One of the biggest challenges in game theory and multi agent systems is the problem of how independent and self-interested agents who do not communicate can be guided towards stable behaviour at some particular action profile that has been deemed desirable. Recently, reinforcement learning frameworks such as Intrinsically Motivated Reinforcement Learning (IMRL) have begun to distinguish between extrinsic reward (tied to task-related, extrinsic motivation) and intrinsic reward (generated according to fulfilment of agent-specific goals), with emotion a potential appraisal mechanism driving intrinsic reward generation [4]. Emotion-based intrinsic reward is supported by neuroscientific evidence suggesting that the orbitofrontal cortex and amygdala (circuits central to emotional processing) play a key role in the computation of reward predictions and errors, which are then projected to midbrain dopamine neurons for use in model-free reward prediction error signalling [3, p.361].

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May 4–8, 2015, Istanbul, Turkey.*

We propose a new conceptual framework in which an external agent can be considered as being able to influence the intrinsic reward generation (appraisal) process of agents partaking in a game, resulting in the emergence of a predetermined new, strict Nash Equilibrium (NE) in pure actions with probability 1. In contrast to [2] (which determines the minimum amount of additional reward required to induce cooperative behaviour in a Prisoner's Dilemma game played by stateless Q-learning agents), we consider agents that learn over a dynamic, internally generated state representation corresponding to a high-level preference ranking over outcomes. Although we don't consider the specifics of the appraisal/reappraisal mechanism here, our results provide motivation for implementations based on these principles.

2. FRAMEWORK

We consider the game in Fig. 1 with a set of 2 agents $A = \{\alpha_1, \alpha_2\}$, where the action set for agent α_1 is $B_1 = \{\beta_{11}, \beta_{12}\}$, and for agent α_2 is $B_2 = \{\beta_{21}, \beta_{22}\}$. The initial payoff matrices are U and V for α_1 and α_2 respectively. We assume that this game has an initial NE in pure actions at $NE_{initial} = (\beta_{12}, \beta_{22})$, i.e. that $U_{22} \geq U_{12}$ and $V_{22} \geq V_{21}$. Our goal is to generate a new, strict NE in pure actions at $NE_{target} = (\beta_{11}, \beta_{21})$. Either one or both of the agents $\alpha_i \in R \subseteq A$ will take the role of "reinforced" agents, and any remaining agent $\alpha_j \in A \setminus R$ will be a "reactive" agent.

We now introduce a source of additional reward, which is considered here as resulting from an internal (re-)appraisal mechanism (e.g. a process of self-reflection or self-therapy within a human agent, perhaps triggered by a particular environmental signal). From this point onwards, we refer to the source of additional reward simply as the "external agent", who does not partake in the game themself but deterministically exerts an influence over the reinforced agents. We assume that the reactive and reinforced agents do not communicate with each other, but the reinforced agent(s) are free to communicate with the external agent.

2.1 Reactive Agents

We assume that reactive agents are unaware of any change to the structure of the game being played and that they therefore continue to play according to their initial, static payoff matrix. Reactive agents maximise their expected payoff based on a probability distribution over the last t moves that the other agents have chosen. In this study we concentrate on the simple case of this iterated strategy for which $t = 1$, i.e. a best response to last move iterated strategy.

2.2 Reinforced Agents

The reinforced agents (either one or both of the agents) must change the value that they place on individual outcomes in order for a new pure action NE to emerge. Each reinforced agent therefore plays

		α_2	
		β_{21}	β_{22}
α_1	β_{11}	U_{11}, V_{11}	U_{12}, V_{12}
	β_{12}	U_{21}, V_{21}	U_{22}, V_{22}

Figure 1: Initial game with a NE in pure actions at (β_{12}, β_{22})

a dynamic game in which its payoff matrix changes according to a payoff reinforcement rule. For $U, V \in \mathbb{R}^{+2\times2}$ we define $U \equiv V$ iff $U_{mn} < U_{m'n'} \Leftrightarrow V_{mn} < V_{m'n'}$ and $U_{mn} = U_{m'n'} \Leftrightarrow V_{mn} = V_{m'n'}$. A complete set of equivalence classes for $\equiv$ is contained in $\mathbb{N}_4^{+2\times2}$, where $\mathbb{N}_4^+ = \{1, 2, 3, 4\}$, and for convenience we use this representation. Let $M \in \mathbb{R}^{+2\times2}$ be the current payoff matrix for $\alpha_i \in R$ (the "M-state"). We introduce the canonical representation of M under $\equiv$ by $[M] := M_{/\equiv} \in E \subset \mathbb{N}_4^{+2\times2}$ which we call the "Q-state", where $E = \{X \in \mathbb{N}_4^{+2\times2} \mid \min_{mn}(X_{mn}) = 1, \forall m, n : (X_{mn} = 1 \text{ or } \exists\, m', n' : X_{m'n'} = X_{mn} - 1)\}$. The Q-state $[M]$ is thus a dense ranking over α_i's payoff matrix M, and the current state for α_i is given by the (M-state, Q-state) tuple $(M, [M])$.

We say that a reinforced agent α_i plays a "reinforced game", which is defined by the state transition system and transition rules given in Fig. 2. The state transition system is a 4-tuple, defined fully by the state space, α_i's initial state $(M^0, [M^0])$, reinforcement set η_i and reinforcement parameter $r_i > 1$. The initial state consists of α_i's initial payoff matrix M^0 (α_i's initial M-state) along with its equivalence $[M^0]$ (α_i's initial Q-state). The reinforcement set η_i is the set of outcomes that will trigger reinforcements in α_i's payoff matrix M. A multiplicative reinforcement of magnitude $r_i > 1$ will be applied to M_{pq} following every occurrence of an action-combination outcome $(\beta_{1p}, \beta_{2q}) \in \eta_i$, resulting in M' as in Fig. 2 (ii). Whilst it would be possible to consider other types of reinforcements (e.g. an additive rule or convergent series), we employ a multiplicative factor as the simplest case and note that it has the desirable property of inducing proportional payoff increments.

Reinforced agents use a model-free Q-learning algorithm, which has a biological basis in the brain's dopaminergic reward system and allows us to capture an anticipation of future reward for deviating from the initial NE for different types of agents with differing discount factors. $Q : E \times B_i \to \mathbb{R}$ gives a Q value for each action of the reinforced agent α_i associated with a particular Q-state, and is initialised under the assumption that the opposing agent will play $NE_{initial}$ (i.e. $Q^0([M], \beta_{ij}) = [M]_{iq}$ for $NE_{initial} = (\beta_{ip}, \beta_{kq})$). Following the choice of action $\beta_{ij} \in B_i$ in the current Q-state $[M]$, we use the conventional single-agent update rule $Q([M], \beta_{ij}) \leftarrow Q([M], \beta_{ij}) + \ell(D(M, \beta_{ij}) + \delta_i \max_{\beta_{iq}} Q(s, \beta_{iq}) - Q([M], \beta_{ij}))$ where $s \in \{[M], [M']\}$ (according to Fig. 2), and $0 \leq \delta_i < 1$ is α_i's discount factor. The learning rate is $0 < \ell([M], \beta_{ij}) = (n([M], \beta_{ij}))^{-1} \leq 1$ where $n([M], \beta_{ij})$ equals the number of times action β_{ij} has been chosen in Q-state $[M]$. Whilst convergence would occur more quickly for a stateless framework, the focus here is on dynamic, internally generated state representations, for which we consider the simplest case consisting of a high-level preference ranking over outcomes.

The reward α_i receives for choosing action β_{ij} in state $(M, [M])$ is $D(M, \beta_{ij})$, which is either a reinforced or non-reinforced payoff. In particular, if action-combination outcome $(\beta_{ij}, \beta_{pq}) \in \eta_i$ has just occurred then $D(M, \beta_{ij}) = r_i M_{jq}$. Alternatively, if $(\beta_{ij}, \beta_{pq}) \notin \eta_i$ has just occurred then $D(M, \beta_{ij}) = M_{jq}$. We employ a simple softmax action selection rule $P(\beta_{ij} | [M]) = k_i^{Q([M],\beta_{ij})} / \sum_j k_i^{Q([M],\beta_{ij})}$ with exploration parameter $k_i > 1$

$$\boxed{\left(\mathbb{R}^{+2\times2} \times E,\ \left(M^0,\ \left[M^0\right]\right),\ \eta_i,\ r_i\right)} \quad \text{(i)}$$

$$\boxed{\begin{array}{l} (M, [M]) \xrightarrow{(\beta_{1p}, \beta_{2q})} (M, [M]) \ \text{ if } (\beta_{1p}, \beta_{2q}) \notin \eta_i \\ (M, [M]) \xrightarrow{(\beta_{1p}, \beta_{2q})} (M', [M']) \ \text{ if } (\beta_{1p}, \beta_{2q}) \in \eta_i \\ \text{with } M'_{mn} = \begin{cases} r_i\, M_{mn} & \text{if } m = p \text{ and } n = q \\ M_{mn} & \text{otherwise} \end{cases} \end{array}} \quad \text{(ii)}$$

Figure 2: (i) State transition system describing the reinforced game for agent α_i (ii) Transition rules. Description in text.

for α_i, such that reinforced agents choose their actions according to a path-dependent, non-stationary stochastic process.

2.3 Convergence

In the full paper [1] we prove the following convergence criteria and provide simulatory results based on a child-parent game, along with extended criteria and proofs for n-agents. For reactive agent α_1 and reinforced agent α_2, the convergence criterion is $U_{11} > U_{21}$ and $\eta = \{NE_{target}\} \cup \zeta$ with $\zeta \subseteq \{(\beta_{12}, \beta_{21})\}$, i.e. the target NE outcome must be reinforced, and the deviation from the initial NE by α_2 can optionally also additionally be reinforced. If $U_{11} < U_{21}$ then by definition a new NE cannot be generated at $NE_{target} = (\beta_{11}, \beta_{21})$, since α_1's payoff matrix does not change. If $U_{11} = U_{21}$ then the dynamics will depend on how α_1 discriminates between outcomes with equal payoffs, although any new NE generated at NE_{target} will not be a strict NE. As an example, single agent reinforcement will lead both agents to a new universally-preferred coordinated NE in the battle of the sexes game.

For some games the above condition does not hold on the initial payoff matrix (e.g. the Prisoner's Dilemma and Snowdrift games, where the desirable new NE is at the coordinated cooperation outcome). For such games we can instead reinforce both agents in order to almost surely guarantee the emergence of the new desirable NE, where agents α_1 and α_2 have states $(U, [U]), (V, [V]) \in \mathbb{R}^{+2\times2} \times E$ respectively. Assume again that we start with the game in Fig. 1 with an initial NE in pure actions at (β_{12}, β_{22}), then the convergence criterion for the generation of a new strict NE in pure actions at (β_{11}, β_{21}) is $\eta_1 = \{NE_{target}\} \cup \zeta_1$ with $\zeta_1 \subseteq \{(\beta_{11}, \beta_{22})\}$ and $\eta_2 = \{NE_{target}\} \cup \zeta_2$ with $\zeta_2 \subseteq \{(\beta_{12}, \beta_{21})\}$, i.e. the target NE outcome must be reinforced for both agents, and the independent deviation from the initial NE by each individual agent can optionally also additionally be reinforced for that respective agent.

REFERENCES

[1] D. Cittern and A. Edalat. Reinforcement learning for nash equilibrium generation. http://www.doc.ic.ac.uk/~dec10/papers/rlneg.pdf.

[2] K. Moriyama et al. Cooperation-eliciting prisoner's dilemma payoffs for reinforcement learning agents. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014.

[3] E. T. Rolls. *Emotion and decision-making explained*. Oxford University Press, 2013.

[4] P. Sequeira et al. Emotion-based intrinsic motivation for reinforcement learning agents. In *Affective Computing and Intelligent Interaction*, pages 326–336. Springer, 2011.