Multi-agent learning

Gerard Vreeswijk, Intelligent Systems Group, Computer Science Department, Faculty of Sciences, Utrecht University, The Netherlands.
Research questions

1. Are there differences between
   (a) Independent Learners (IL) Agents that attempt to learn
       i. The values of single actions (single-action RL).
   (b) Joint Action Learners (JAL) Agents that attempt to learn both
       i. The values of joint actions (multi-action RL).
       ii. The behaviour employed by other agents (Fictitious Play).

2. Are RL algorithms guaranteed to converge in multi-agent settings? If so, do they converge to equilibria? Are these equilibria optimal?

3. How are rates of convergence and limit points influenced by the system structure and action selection strategies?

Claus et al. address some of these questions in a limited setting, namely, A repeated cooperative two-player multiple-action game in strategic form.
Cited work


The paper on which this presentation is mostly based on.


Mainly the result that Q-learning converges to the optimum action-values with probability one as long as all actions are repeatedly sampled in all states and the action-values are represented discretely.


Q-learning

- The general version of Q-learning is multi-state and amounts to continuously updating the various $Q(s,a)$ with

$$r(s,a,s') + \gamma \cdot \max_a Q(s',a) \quad (1)$$

- In the present setting, there is only one state (namely, the stage game $G$) so that (1) reduces to

$$r(s,a,s)$$

which may be abbreviated to $r(a)$ or even $r$.

- Single-state reinforcement learning rule:

$$Q_{\text{new}}(a) = (1 - \lambda)Q_{\text{old}}(a) + \lambda \cdot r$$

- Two sufficient conditions for convergence in Q-learning (Watkins, Dayan, 1992):
  1. Parameter $\lambda$ decreases through time such that $\sum_t \lambda$ is divergent and $\sum_t \lambda^2$ is convergent.
  2. All actions are sampled infinitely often.
Exploitive vs. non-exploitive exploration

Convergence on Q-learning does not depend on the exploration strategy used. (It is just that all actions must be sampled infinitely often.)

**Non-exploitive exploration**  This is like what happens in the $\epsilon$-part of $\epsilon$-greedy learning.

**Exploitive exploration**  Even during exploration, there is a probabilistic bias to exploring optimal actions.

*Example.* Boltzmann exploration (a.k.a. soft max, mixed logit, or quantal response function):

$$\frac{e^{Q(a)/T}}{\sum_{a'} e^{Q(a')/T}}$$

with $T > 0$.

Letting $T \to 0$ establishes convergence conditions (1) and (2) as mentioned above (Watkins, Dayan, 1992).
Independent Learning (IL)

- A MARL algorithm is an independent learner (IL) algorithm if the agents learn Q-values for their individual actions.
- Experiences for agent $i$ take the form $\langle a_i, r(a_i) \rangle$ where $a_i$ is the action performed by $i$ and $r(a_i)$ is a reward for action $a_i$.
- Learning is based on

$$Q_{\text{new}}(a) = (1 - \lambda)Q_{\text{old}}(a) + \lambda \cdot r(a)$$

ILs perform their actions, obtain a reward and update their Q-values without regard to the actions performed by other agents.

- Typical conditions for Independent Learning:
  - An agent is unaware of the existence of other agents.
  - It cannot identify other agent’s actions, or has no reason to believe that other agents are acting strategically.

Of course, even if an agent can learn through joint actions, it may still choose to ignore information about the other agents’ behaviour.
Joint-Action Learning (JAL)

- **Joint Q-values** are estimated rewards for joint actions.
  For a $2 \times 2$ game an agent would have to maintain $Q(T, L)$, $Q(T, R)$, $Q(B, L)$, and $Q(B, R)$.

- Row can only influence $T, B$ but not opponent’s actions $L, R$.
  Let $a_i$ be an action of player $i$. A complementary joint action profile is a set of joint actions $a_{-i}$ such that $a = a_i \cup a_{-i}$ is a complete joint action profile.

- Opponent’s actions can be estimated through forecast by, e.g., fictitious play:
  \[
  f_i(a_{-i}) = \text{Def } \Pi_{j \neq i} \phi^j(a_{-i})
  \]
  where $\phi^j(a_{-i})$ is $i$’s empirical distribution of $j$’s actions on $a_{-i}$.

- The expected value of an individual action is the sum of joint Q-values, weighed by the estimated probability of the associated complementary joint action profiles:
  \[
  EV(a_i) = \sum_{a_{-i} \in A_{-i}} Q(a_i \cup a_{-i}) f_i(a_{-i})
  \]
Comparing Independent and Joint-Action Learners

Case 1: the coordination game

\[
\begin{pmatrix}
T & (10, 0) \\
B & (0, 10)
\end{pmatrix}
\]

- A JAL is able to distinguish Q-values of different joint actions \( a = a_i \cup a_{-i} \).
- However, its ability to use this information is circumscribed by the limited freedom of its own actions \( a_i \in A_i \).
- A JAL maintains beliefs \( f(a_i) \) about the strategy being played by other agents through fictitious play, and plays a softmax best response.

A JAL computes singular Q-values by means of explicit belief distributions on joint Q-values. Thus,

\[
EV(a_i) = \sum_{a_{-i} \in A_{-i}} Q(a_i \cup a_{-i}) f_i(a_{-i})
\]

is more or less the same as the Q-values learned by ILs.
- Thus even though a JAL may be fairly sure of the relative Q-values of its joint actions, it seems it cannot really benefit from this.
Figure 1: Convergence of coordination for ILs and JALs (averaged over 100 trials).
Comparing Independent and Joint-Action Learners

Case 1: the coordination game

\[
\begin{bmatrix}
T & 10 & 0 \\
B & 0 & 10 \\
\end{bmatrix}
\]

- A JAL is able to distinguish Q-values of different joint actions \( a = a_i \cup a_{-i} \).
- However, its ability to use this information is circumscribed by the limited freedom of its own actions \( a_i \in A_i \).
- A JAL maintains beliefs \( f(a_i) \) about the strategy being played by other agents (fictitious play) and plays a softmax best response.

A JAL computes single Q-values by means of explicit belief distributions on joint Q-values.

Thus,

\[
EV(a_i) = \sum_{a_{-i} \in A_{-i}} Q(a_i \cup a_{-i}) f_i(a_{-i})
\]

is more or less the same as the Q-values learned by ILs.

- Thus even though a JAL may be fairly sure of the relative Q-values of its joint actions, it seems it cannot really benefit from this.
Case 2: Penalty game

\[
\begin{pmatrix}
L & M & R \\
T & 10 & 0 & k \\
C & 0 & 2 & 0 \\
B & k & 0 & 10 \\
\end{pmatrix}
\]

Suppose penalty \( k = -100 \).

The following stories are entirely symmetrical for Row and Column.

**IL** 1. Initially, Column explores.
2. Therefore, Row will find \( T \) and \( B \) on average very unattractive, and will converge to \( C \).
3. Therefore, Col will find \( T \) and \( B \) slightly less attractive, and will converge to \( C \) as well.

**JAL** 1. Initially, Column explores.
2. Therefore Row gives low EV to \( T \) and \( B \). Plays \( C \) the most.
3. Convergence to \( (C, M) \).
Figure 2:
Likelihood of convergence to opt. equilibrium as a function of penalty $k$ (100 trials).
Case 3: Climbing game

- Initially, the two learners are almost always going to begin to play the non-equilibrium strategy profile $(B, R)$.
- Once they settle at $(B, R)$, and as long as exploration continues, Row will soon find $C$ to be more attractive, so long as Col continues to primarily choose $R$.
- Once the non-equilibrium point $(C, R)$ is attained, Col tracks Row’s move and begins to perform action $M$. Once this equilibrium is reached, the agents remain there.
- This phenomenon will obtain in general, allowing one to conclude that the multiagent Q-learning schemes we have proposed will converge to equilibria almost surely.
Figure 3:

A’s strategy in climbing game.
Figure 4: B’s strategy in climbing game.
Figure 5:
Joint actions in climbing game.
Being asymptotically myopic

Recall from fictitious and Bayesian play the notion of a **predictive strategy** with forecast function $f_i : H \rightarrow \Delta(A_{-i})$ and behaviour rule $g_i : H \rightarrow \Delta(A_i)$.

- A forecast $f_i$ is said to be **asymptotically empirical** if it converges to the empirical frequencies of play with probability one.

- A behaviour rule $g_i$ is said to be **asymptotically myopic** if the loss from player $i$’s choice of action at every history given $g_i^t$ goes to zero as $t$ proceeds:

  $$u_i(f_i^t, g_i^t) \nearrow \max\{u_i(f_i^t, a_i) \mid a_i \in A_i\}$$

  as $t \to \infty$, where $u_i$ denotes expected payoff.

  - Being asymptotically myopic is **less demanding** than a behaviour rule that assigns positive probability only to pure strategies that eventually come close to maximising expected payoff.
Asymp. empiricism and myopia imply convergence to Nash

Being asymptotically myopic includes:

- Strategies that incur a large loss with a small probability (regardless of opponents’ play).
- Strategies that incur a small loss with a large probability (regardless).
- A combination of both.

Definition. A joint action profile $a$ is called stable if joint behaviour is a limit point of $a$ with probability one.

Proposition. (Fudenberg and Kreps, 1993, p. 343): Let forecast $f_i$ be asymptotically empirical, and let behaviour rule $g_i$ be asymptotically myopic. Then every stable joint action profile is a Nash equilibrium.

Proof (Outline.) Suppose $a$ is stable. Then empirical frequencies eventually converge (!) to $a$. Because $f_i$ is asymptotically empirical, so will the $f_i$, for all $i$, with probability one. Convergence of $f_i$ together with asymptotic myopia of $g_i$ implies that the $g_i$ converge as well. This situation is in effect a Nash equilibrium.  \[\square\]
### Sufficient conditions for asymptotic behaviour

1. The learning rate $\lambda$ decreases over time such that $\sum_t \lambda$ is divergent and $\sum_t \lambda^2$ is convergent.
   - Required for convergence in Q-learning.

2. Each agent samples each of its actions infinitely often.
   - Required for convergence in Q-learning.

3. The probability $P^i_t(a)$ of agent $i$ choosing action $a$ is nonzero.
   - Ensures (2), and ensures that agents explore with positive probability at all times.

4. Agents become full exploiters with probability one eventually:

   $$\lim_{t \rightarrow \infty} P^i_t(X_t) = 0,$$

   where $X_t$ is a random variable denoting the event that $(f_i, g_i)$ prescribe a sub-optimal action.
Myopic heuristics

Optimistic Boltzmann (OB): For agent $i$, action $a_i \in A_i$, let

$$\text{MaxQ}(a_i) = \text{Def} \max \prod_{-i} Q(\prod_{-i}, a_i).$$

Choose actions with Boltzmann exploration (another exploitive strategy would suffice) using $\text{MaxQ}(a_i)$ as the value of $a_i$.

Weighted OB (WOB): Explore using Boltzmann using factors

$$\text{MaxQ}(a_i) \cdot \text{Pr(optimal match } \prod_{-i} \text{ for } a_i).$$

Combined: Let

$$C(a_i) = \rho \text{MaxQ}(a_i) + (1 - \rho)\text{EV}(a_i),$$

for some $0 \leq \rho \leq 1$. Choose actions using Boltzmann exploration with $C(a_i)$ as value of $a_i$. 
Figure 6:
Sliding average reward in the penalty game.