
Autonomous Agents

Report Assignment 3: Multi-Agent Planning and Learning

Authors:

Agnes VAN BELLE (10363130),
Maaïke FLEUREN (10350470),
Norbert HEIJNE (10357769),
Lydia MENNES (10333843)

October 21, 2012

1 Introduction

This report has been written for the Master Artificial Intelligence course Autonomous Agents. This report will contain the answers, motivation and explanation for our implementations of the tasks we had to accomplish in our third assignment for this course. These tasks were centered around the topic of ‘Multi-Agent Planning and Learning’.

1.1 The environment

In all tasks there is assumed to be a grid world (of 11×11) with at least one predator and a prey in it. The agents can both move one tile forward each iteration. The direction they take (or if they move at all) is affected by probabilities (their policies). If they move over the edge of the grid they end up at the opposing side of the grid. We are focused on improving the decisions of all the agents, the predator(s) and the prey.

1.2 The state space representation

In this assignment, we used the efficient state space representation that we also used in the previous Assignment for the learning algorithms. To give a good understanding of our learning algorithms, which were built again using this efficient state space representation, we will once again explain how this representation works.

Figure 1(a) illustrates that there is a symmetry in the default state space, and thus that there were relatively much values redundantly computed. By using this symmetry in the default state space a much smaller state space was achieved.

Each state represents a distance between the prey and predator. These are represented in the lower left diagonal of a matrix, in which the x -axis is the relative horizontal distance in the MDP and the y -axis the relative vertical distance in the MDP. This matrix is shown in Figure 1(b). Combinations of positions of prey and predator for which the horizontal and vertical distances are equal are now treated equivalent. Also two combinations for which the horizontal distance in one equals the vertical distance in the other and vice versa are considered equal. In order to navigate through this state space different actions are required. These are: *horizontal retreat*, *horizontal approach*, *vertical retreat*, *vertical approach*, as illustrated in Figure 1(a), and of course the action *wait*. When interacting with the environment these actions are converted into corresponding actions in the real world. This only requires the relative direction of the prey (which is always

located at the centre, regardless of its coordinates) with respect to the predator. This is computed by using the difference in location of the prey and predator on the x - and y -axis.



(a) The 11×11 grid divided into eight symmetric pieces, with the corresponding possible moves which are also symmetric. (b) Colormap of V -values, the brighter the color the higher the corresponding V -value. The prey is always located on the (1, 1) coordinate in this state representation.

Figure 1: Illustration of the symmetry and corresponding values of the new state space representation

Note that this state representation works only for two agents (here we talked about one prey and one predator). For Independent Q-Learning (see Section 2), in which more agents will be used, we will require a more extensive state space. That state space is, however, still built upon this one.

1.3 Implementation details

This report will not be about our exact code and implementation details. However, a class diagram of our code is provided alongside the source code.

2 Independent Q-Learning

Q-Learning is an off-policy temporal difference control algorithm. Temporal difference methods can learn directly from raw experience without a model of the environment’s dynamics. Furthermore, it updates estimates based in part on other learned estimates, without waiting for a final outcome [3, pp. 133].

While the distinguishing feature of on-policy methods is that they estimate the value of a policy while using it for control, these two functions are separated in off-policy methods. The behavior policy, used to generate behavior, and the estimation policy, that is evaluated and improved, may in fact be unrelated. This separation is an advantage because the estimation policy may be deterministic (e.g., greedy), while the behavior policy can continue to sample all possible actions [3, pp. 126]. Its simplest form, *one-step Q-learning*, is defined by [3, pp. 148]:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right]$$

When using Q-Learning in a multi-agent setting independently for each agent, the method is called Independent Q-Learning (IQL). In IQL, for a single agent, all the other agents are modeled as part of the environment. While Q-Learning is guaranteed to converge, Independent Q-Learning is not guaranteed to converge. This is because the premise of the convergence property of Q-Learning is that the environment be stationary. This will not be the case in a multiagent setting in which agents are learning concurrently, since the agents base their policy upon the movements of the other agents which are also learning. That is, the other agents don’t act deterministically but are modeled as such. Independent Q-Learning can thus not be justified theoretically [4, pp. 57].

2.1 Implementation

For this assignment, we implemented Independent Q-learning and used it independently for all our agents, the predators and the prey. The agents learn concurrently (i.e. all agents do a move, then all update their estimated values), not sequentially (i.e. after one agent does a move, all other agents update their estimated values). When any two predators bump into each other, the prey wins (reward of 10) and the predators all lose (reward of -10). When a predator catches the prey, the prey loses and all predators win.

Each agent will have its own state representation and learning and will view the other agents as part of the environment. We used ϵ -greedy action selection, which behaves greedily most of the time, but with probability ϵ , instead select an action at random, uniformly, independently of the action-value estimates [3, pp. 28]. In this case, we used $\epsilon = 0.1$. We initiated the values of each Q-learning table optimistically with a value of 15 for all cells for the tables. Furthermore the learning rate α was set to 0.5 and the discount factor γ was set to 0.9.

2.1.1 State space

The state space is of crucial importance to Independent Q-Learning, because it grows exponentially with the number of agents. In the experiments described in the previous report, where we had just one prey and one predator, we initially used a state space that was an intuitive, yet cumbersome representation. We then changed the state space representation to a more efficient one in the second assignment, referred to as the ‘efficient’ state space, which led to a reduction of 697 times less states, resulting in just 21 different states. See Section 1.2 for details.

The default amount of states for this assignment would be 121^{p+1} for each agent, where p equals the number of predators. By using the efficient state space this has been reduced to $21 \cdot 121^{p-1}$ because the prey’s position is fixed in the state space representation and the predator to which this state space belongs moves along those 21 states, the rest is used to determine in which state the predator currently resides. In this assignment however the prey also learns, so to make use of our state space representation, a single predator will be used as reference point every time so that the predator is fixed at the position (0,0) in our state space and the prey would be moving around the 21 states just like a predator.

The reduction is increased further by pointing to the same state for each combination of current positions of the other predators, though this is only effective for more than two predators. A simple example would be a case three predators and a prey, the current predator (p_1) which has to find his current state would

take the position of the first other predator (p_2) and then the position of the second other predator (p_3) to locate his state in the state space. In this case it shouldn't matter which position, p_2 or p_3 , would be looked up first. In this case both p_2 then p_3 and p_3 then p_2 would point to the same state. This reduces the state space to:

$$\text{stateSpaceSize}(p) = \begin{cases} 21 & p = 1 \\ 21 \cdot 121 & p = 2 \\ \frac{21 \cdot 121^{p-1}}{(p-1)!} & p > 2 \end{cases}$$

2.2 Results

We analyzed what happened in the case of 2, 3 and 4 predators. For the case of one predator (and one prey), we have not used this case here, we used this case in the comparison with the Minimax-Q algorithm (in Section 3.4).

After producing the results it came to our attention that Q-learning did not yet have the prey trip chance implemented, for more than one predator it did not cause an infinite loop (the predators' policies were good enough to catch a prey that cannot trip). However, for the results in Section 3.4 this was implemented, where before it would cause the prey to stay away from a single predator indefinitely.

Even though we were using an efficient state space representation we ran out of time while trying to initialize one prey and four predators for the 11×11 grid world. We dedicated around 3.6 GB heap space for the JVM on a 64-bit machine with 4 processor cores of 2.67 Ghz, and under these settings it took more than 7 hours to declare and initialize four predators, the prey not being created and initialized yet. However, when using a 9×9 grid world instead of a 11×11 one, the five agents were in short time up and running. Therefore we will show the results for 2, 3 and 4 predators for this 9×9 grid world first, and later we will attempt to make a comparison with the 11×11 grid world.

2.2.1 Number of predators

We kept track of two measures:

- The number of time steps before the game was ended (by collision or catching)
- If the predators won the episode, or not

We ran each episode **200** times. Therefore these measures are displayed as an average value and a percentage value, respectively.

In Figure 2 are the means and standard deviations shown for a total of 5000 episodes, for the different numbers of predators.

We can see that in the case of four predators the percentage of times that the predators won an episode is the smallest. This is probably because of the higher chance of collision and more difficult coordination necessities. The variance for this measure for the case of four predators is also pretty high, also suggesting that a "common" and more or less converged policy is not developed yet (insofar as a true common converged policy is possible under IQL).

The case of there being three predators performs almost the same as the case of there being four predators on the average number of time steps needed for termination of an episode. However for the case of there being three predators the episode endings are more often in the predators' favor. It seems, that three predators catch the prey less often than when there are two predators, but they do need less time steps for it.

On the other hand, the case of there being two predators performs surprisingly well on the percentage of times the predators won, but needs a lot of time steps for that.

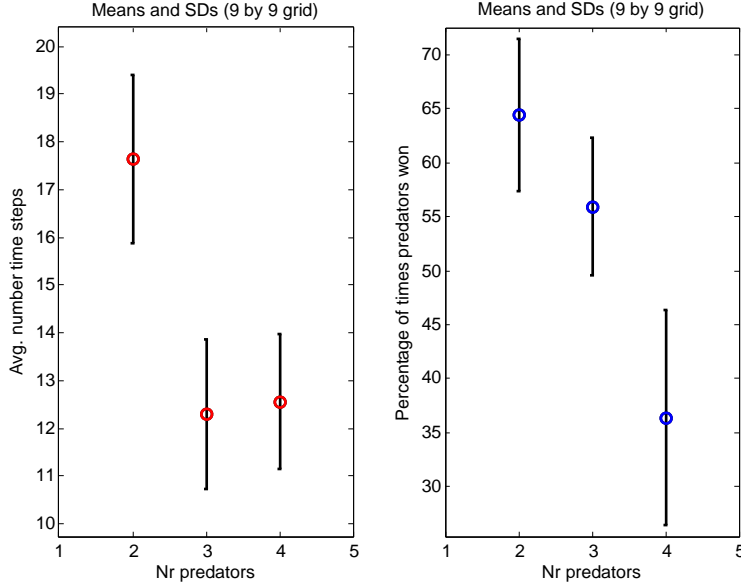


Figure 2: Means and SD's for performance measures for IQL, for 5000 episodes

Now it would be interesting to see the performance over time and also if IQL actually converges for these cases, since IQL is not guaranteed to. In Tables 1 and 2 we show the mean and variance again but now also the last value (i.e. at the 5000th episode, run 200 times) for the corresponding measure.

Nr. predators	Variance	Mean	Last value
2	50.109	64.482	62
3	41.104	55.941	62.5
4	99.92	36.331	60

Table 1: Percentage of times predators won (9×9 grid)

Nr. predators	Variance	Mean	Last value
2	3.08	17.64	16.61
3	2.435	12.297	12.155
4	1.997	12.562	12.89

Table 2: Average number of time steps (9×9 grid)

We can immediately see that the case of four predators seems to move a large amount up for the measure of the percentage of times they won. Its mean value is 36.331 but its end value is 60. The case for two predators seems to have “converged” long before, with a mean close to its last value.

The variance of the case of four predators, for the measure of the percentage of times they won, is quite large. This can mean the agents either learn very slow or because of large fluctuations in the distribution of “skill” over the prey and predators over time. In Figure 3 we show a plot of the percentage of times the predators won for each predator number setting. We indeed see that in the setting of four predators the learning simply occurs slowly.

In Figure 3 we also see that the predator performance slowly goes down between the 500th and 2000th episode for the case of two predators. Apparently the predators learned to optimize their performance before

the prey could, but then the prey learned their policy, and after that escapes more often. The predators apparently can't make a much better policy after that. For the settings of three and four predators, it seems the prey was the first to learn a (up until a certain time) good policy.

As for convergence, the average number of time steps used per episode seems to converge for all three settings, as can also be seen in Figure 4 (and Table 2). The percentage of times the predators won, seems to converge for 5000 episodes only for the settings of two and three predators. The value for four predators seems not stable yet, as can be seen in Figure 3 (and Table 1).

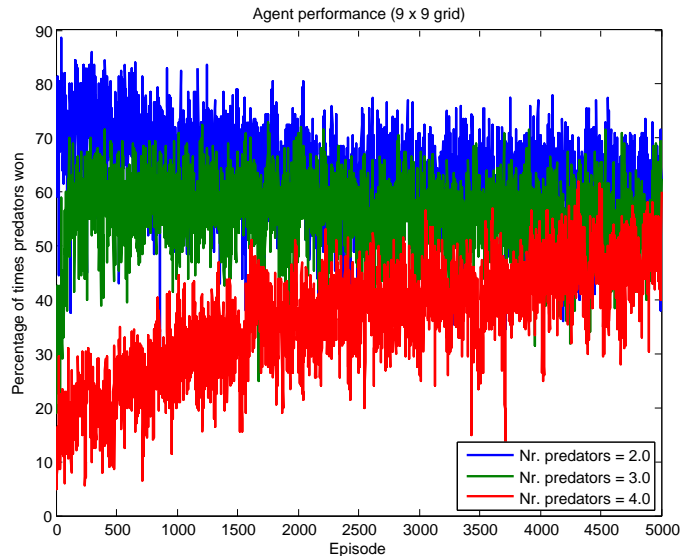


Figure 3: Percentage of times the predators won for the three predator number settings

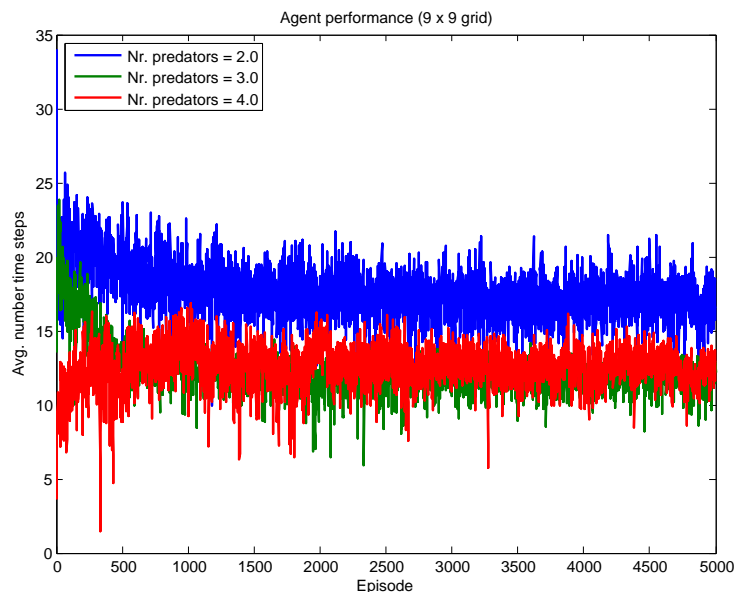


Figure 4: Average nr. of time steps before an episode ending, for the three predator number settings

2.2.2 Using an 11×11 grid

The results for using a larger, 11×11 grid seem roughly similar to those in Section 2.2.1 where we used a 9×9 grid. Compare Figure 5 with Figure 3 and Figure 6 with Figure 4. The differences seem more pronounced for the first 1000 episodes. Also, the percentage of times that the predators won, is higher for the case of three predators in the 11×11 setting, probably because of less collision issues.

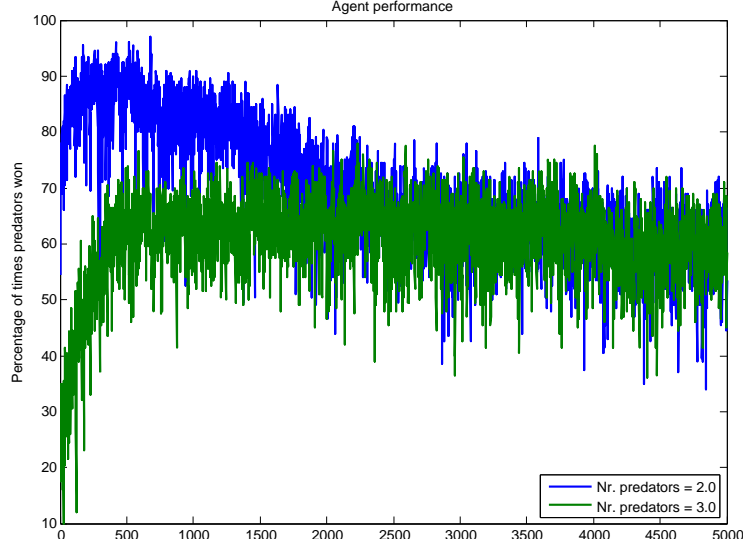


Figure 5: Percentage of times the predators won for the two predator number settings

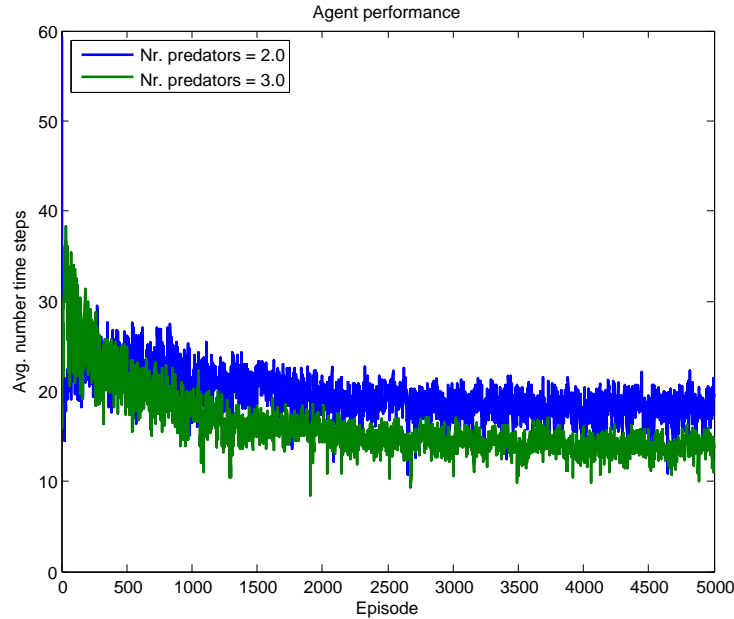


Figure 6: Average nr. of time steps before an episode ending, for the two predator number settings

3 Minimax-Q algorithm

The Minimax-Q algorithm can be used for planning in a Markov Game. It is an implementation of the general Minimax principle for Markov Games such as the prey-predator scenario in this assignment. This algorithm needs a strictly competitive scenario and is therefore implemented for the one prey and one predator setting.

3.1 Minimax principle and translation to Markov Games

The Minimax principle is stated in [1] as: “Behave so as to maximize your reward in the worst case”. This means that the policy should be such that the agent takes the action resulting in the highest expected reward under the assumption that the opponent will also take the action that results in the highest expected reward for the opponent.

The value of a state $s \in S$ in a Markov Game is:

$$V(s) = \arg \max_{\pi \in PD(A)} \arg \min_{o \in O} \sum_{a \in A} Q(s, a, o) \pi_a$$

where

$$Q(s, a, o) = R(s, a, o) + \gamma \sum_{s'} T(s, a, o, s') V(s')$$

Choosing the action that maximizes the reward in the worst case is than the policy that is guaranteed to result in an expected score of V no matter which action the opponent chooses, for the maximal value of V . This can be written down in constraints where there is a constraint for each action the opponent can take:

$$\begin{aligned} \pi_1 \cdot Q(s, a_1, o_1) + \dots + \pi_n \cdot Q(s, a_n, o_1) &\geq V \\ \pi_1 \cdot Q(s, a_1, o_2) + \dots + \pi_n \cdot Q(s, a_n, o_2) &\geq V \\ &\dots \\ \pi_1 \cdot Q(s, a_1, o_n) + \dots + \pi_n \cdot Q(s, a_n, o_n) &\geq V \end{aligned}$$

Of course the probabilities of π are bounded by the constraints that they cannot be negative and they need to add up to one. After adding those constraints:

$$\begin{aligned} \pi_1 + \dots + \pi_n &= 1 \\ \pi_1 &\geq 0 \\ &\dots \\ \pi_n &\geq 0 \end{aligned}$$

The maximal value for V that can be found for which all constraints hold is stored as V for the next sweep. The resulting values for π for each state are the policy used by the agent.

3.2 Implementation

The maximal V -value for which the constraints above hold can be found using linear programming. For this purpose we used the Linear Programming methods of the JAVA-library Apache Commons Math, which can be found on [2]. The agent keeps track of a table of V -values to be able to calculate $Q(s, a, o)$ in order to find the right values for π . This can be done a number of times (or sweeps as they are called) as in Value Iteration. The resulting policy is used to behave in the environment to either catch the prey or avoid the predator.

	1	2	3	4	5
Minimax Predator vs Random Prey	9923	9432	9517	9545	9572
Random Predator vs Minimax Prey	282	5000000	5000000	5000000	5000000
Minimax Predator vs Minimax Prey	221	5000000	5000000	5000000	5000000

Table 3: Number of steps after each sweep for the three matches using a learningrate of 0.8

3.3 Results

Three “matches” have been held using the Minimax-Q algorithm. The first match is a Random Prey versus a Minimax Predator, the second one a Random Predator versus a Minimax Prey and finally a Minimax Prey versus a Minimax Predator. After each sweep of planning 25 runs in the actual environment are held and the number of steps necessary to catch the prey is recorded. This averaged can then be compared for each sweep and between the different matches. The results for the first five sweeps can be found in 3, the pattern found here does not change when increasing the number of sweeps. The V -values for the Minimax Predator can be found in table 4 and for the Minimax Prey in table 5 . The resulting policies can be found in Appendix B.

As can be seen in the tables with the V -values these look as would be expected based on their definition as described above. For the predator the values increase as the predator is closer to the prey. However for the prey all values are 0. This seems strange but remember that the V values are a linear combination of $Q(s, a, o)$ ’s and their weights are the corresponding π ’s. In order to maximize the value of V the π ’s that are corresponding to those $Q(s, a, o)$ ’s that are negative due to the predator catching the prey, will become zero. The maximal reward the prey can have is zero (not getting caught) and therefore the maximal value for V is also zero.

Unfortunately, the resulting policies make no sense. There are loops present, where the prey or predator keeps jumping forward and backward and the policy does not direct the predator towards the prey or the prey away from the predator. Whether this is due to a bug in the implementation (which we have not been able to find) or to the fact that there are multiple solutions to the constraints above which maximize V is uncertain. If there are multiple solution for two states not all solutions fit necessarily well together, this might cause the bad policies. i.e. for different states the policy might have converged to a different equilibrium.

The effect of this can be found in the number of steps it takes to catch the prey. For the first match, where the Minimax Predator tries to catch the Random Prey, there is no improvement in the number of steps at all. This is due to the resulting policy. For the second match, Random Predator versus Minimax prey, the number of steps needed becomes larger than 5 million (at which point the run is ended due to death of the predator by exhaustion) after the first sweep. The same applies to the third match, Minimax Predator versus Minimax Prey. It is expected that for these two matches the number of steps needed would be larger, since the prey now is an intelligent agent. However, due to the fact that the prey trips one of every five moves, it would be expected that with optimal policies it would take the predator fifty moves to catch the prey. The prey and predator start ten moves apart from each other and each time the prey trips the predator can move in on the prey one move, which happens every five moves, hence fifty moves are approximately needed to catch a Minimax Prey with an optimal policy.

0,0000					
10,0000	4,3644				
4,3644	1,7596	0,6019			
1,7596	0,6019	0,1624	0,0328		
0,6019	0,1624	0,0328	0,0047	0,0005	
0,0968	0,0328	0,0047	0,0005	0,0000	0,0000

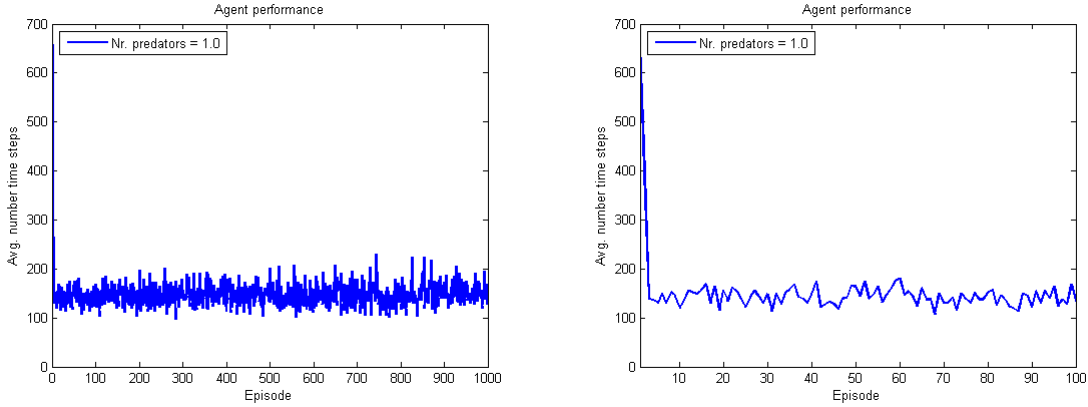
Table 4: V -values of Minimax Predator using a learningrate of 0.8

0,0000					
0,0000	0,0000				
0,0000	0,0000	0,0000			
0,0000	0,0000	0,0000	0,0000		
0,0000	0,0000	0,0000	0,0000	0,0000	
0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Table 5: V-values of Minimax Prey using a learningrate of 0.8

3.4 Comparison with Independent Q-Learning

A policy resulting from Independent Q-Learning (see Section 2) can be compared to a policy found with Minimax-Q for the case in which there is just one predator and one prey. We ran IQL under these settings for 1000 episodes and constructed policies for the predator and the prey. These policies can be found in Appendix C (the corresponding policies for Minimax-Q are in Appendix B). The number of steps taken by the predator to reach the prey is shown in figure 7. Comparing the policies of Q-learning and Minimax-Q shows that for the most part they are both straightforward, prey moves away from the predator and the predator moves toward the prey. The exception is the states next to the prey, where Minimax-Q does not move directly towards the prey. The reason is that Minimax-Q responds to the policy of the prey. Q-learning however uses episodic learning and randomness (ϵ -greedy action selection and prey trip chance) to learn the policy. Because of this, Q-learning cannot respond as precise as Minimax-Q. Even though this is true, the average number of time steps needed (150 according to figure 7(a)) to catch the prey is lower for Q-learning than for Minimax-Q as seen in table 3.



(a) Displays the average number of time steps needed to catch the prey for the first 1000 episodes.

(b) Close-up of figure 7(a).

Figure 7: Plots that display the number of time steps needed, averaged over 200 runs for each episode, for a single predator using Q-learning to catch a prey that also uses Q-learning.

4 Conclusion

With these results at hand we can conclude a few things. In the case of independent Q-learning (IQL), two or more predators was enough to catch a prey that uses IQL without it have a chance to trip. This was not the case with only one predator, where the prey would stay away from the predator indefinitely. Not only could they catch a non-tripping prey, more predators would result in faster capture, though it seems that from two predators onward they converge to approximately the same amount of time steps required before the prey is captured. The emphasis lies on approximately since the IQL does not guarantee convergence, even though our results converge quite reasonably. Q-learning is fit for Markov Decision Processes (MDP),

however this is a multi-agent setting which means that it is a Markov Game, where the MDP is a special case of a Markov Game.

Generating a state space for four predators on a 11×11 grid took a large amount of time and a large amount of RAM, 3.6 GB and over 7 hours to declare only the predator states. However, using a smaller grid (9×9) was much more feasible for four predators. Requiring about 5 times less states than the normal grid size as shown below.

$\frac{21 \cdot 121^{4-1}}{(4-1)!} \approx 6200464$	$\frac{15 \cdot 81^{4-1}}{(4-1)!} \approx 1328603$
11×11 grid	9×9 grid

Minimax-Q (MmQ) on the other hand showed some bugs, such as the possibility of infinite loops in the policy and the predator not catching the prey. If the MmQ is implemented correctly then the IQL results in a better policy, requiring less time steps to catch the prey. But sadly due to the possibility of bugs in the MmQ code, we cannot conclude anything in regards to comparing IQL and MmQ. Only that the the MmQ is restricted to two agents and is not constrained by memory size, where as IQL is constrained by memory size and is not restricted to only two agents.

A.2 Policy for prey

[illegible]

Table 7: Policy resulting from Minimax-Q for the prey

B.2 Policy for prey

[illegible]

Table 9: Policy resulting from IQL for the prey (2500 episodes)

References

- [1] M.L. Littman, *Markov games as a framework for multi-agent reinforcement learning*. In Proceedings of the Eleventh International Conference on Machine Learning, pp. 157-163 San Francisco, CA. Morgan Kaufmann. (1994)
- [2] Commons Math: The Apache Commons Mathematics Library (2012),
<http://commons.apache.org/math/>.
- [3] R.S. Sutton and A.G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, Massachusetts: The MIT press. (1998)
- [4] Vlassis, *A Concise Introduction to Multiagent Systems and Distributed Artificial Intelligence*. San Rafael, CA: Morgan & Claypool Publishers. ISBN 978-1-59829-526-9. (2008)