Empirical Study of Q-Learning Based Elemental Hose Transport Control

Jose Manuel Lopez-Guede, Borja Fernandez-Gauna, Manuel Graña, Ekaitz Zulueta

HAIS 2011

Outline

1. Introduction
2. Experimental system design
3. Experimental results
4. Conclusions
5. Future work

Introduction

- Interest in control algorithms for Linked MultiComponent Robotic Systems (L-MCRS).
- Inherent complexity due to the dynamics of the physical links.
- This complexity adds on the difficulty of dealing with collections of independent robots.
- We have started applying different control approaches:
  - based on analytical detailed models,
  - some using simplified spring-like linking-elements.

- Prototypical case: hose transportation problem with a single mobile robot at the tip of the hose.
- This is a simple formulation of the problem, and the results can serve as a starting point for further generalization.
- We have proposed Q-Learning as the basic approach to learn controllers for this system through experience.
- Q-Learning is a Reinforcement Learning algorithm able to learn from on-line experience without requiring accurate knowledge of the environment.

- Given some sensorial data to characterize the world state, an action selection policy selects an action and stores information regarding the quality of the action taken as qualified by a predefined reward system.
- Results were provided by the accurate simulation of L-MCRS developed by our group based on the Geometrically Exact Dynamic Splines (GEDS) approach to build dynamical model of uni-dimensional objects.
- This paper extends the results giving further insight into the effect of the different learning parameters by the conducted exhaustive experiments.

Experimental system design

- Elemental hose system:
  - One hose segment attached to a fixed end (the source).
  - The tip is transported by a mobile robot attached to it.
  - The fixed end is set as the middle of the configuration space.
  - The task for the robot is to bring the tip of the hose to a destination.
  - The working space where the tip-of-the-hose robot moves is a square of size 2×2 m^2.

Q-Learning definitions

- State: we have defined the state using three alternative models: $S = (P_r, P_d, i)$, $S = (P_r, P_d, i, c)$ and $S = (P_r, P_d, i, P_1, P_2)$, where
  - $P_r = (x_r, y_r)$ is the actual position of the tip-of-the-hose robot.
  - $P_d = (x_d, y_d)$ is the desired position of the tip-of-the-hose robot, the goal.
  - i is a binary variable that indicates if the line $\overline{P_r P_d}$ intersects the hose. $i = 1$ means that there is an intersection.
  - c is a binary variable that indicates if the box with corners P_r and P_d intersects the hose. $c = 1$ means that there is an intersection,
  - $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ are two points of the hose that are uniformly distributed from one end to the other end.

- Working space discretization:
  - Two different discretization steps of 0,5 m. and 0,2 m.
  - It determines the cardinality of the universe of states.
  - It determines the precision of the coordinates of the points P_r, P_d, P_1 and P_2.
  - Our working space is, thus, partitioned into 16 and 100 boxes respectively.
- Final state: the final state can be of three kinds.
  - Goal: the tip of the hose reaches the goal.
  - Failure: the tip of the hose is blocked in its advance by the hose itself.
  - Inconclusive: the simulation ends without reaching any of the aforementioned states.

- Actions:
  - We can only interact with the scenario using the mobile robot to change the position of the tip-of-the-hose.
  - We have chosen a small set of only four actions: $A = \{North, South, East, West\}$
  - The robot will move in a direction for a length equivalent to the size of the resolution box.
- Action selection: ε-greedy policy, with $\varepsilon = 0,2$. We choose the action a with this criterion:

$$a \leftarrow \begin{cases} \max_{a'} Q(s, a') & \text{with probability } (1-\varepsilon) \\ any\ a' \in A & \text{with probability } \varepsilon \end{cases}.$$

- Reward system: We have used several reward systems.
  - Reward systems 10 and 20: both give a positive reward when reaching the goal, negative when failing and nothing if the end state is inconclusive.
  - Reward system 50: only gives positive reward when reaching the goal.
  - The remaining reward systems give positive reward when reaching the goal, negative when failing and for the inconclusive states, a function of the actual distance between the hose tip and the goal. In some cases the reward function is also function of the binary variables c and i.

- α: $[0 < \alpha \leq 1]$, as we suppose that we are working in a deterministic environment we can assume that the value of this parameter is 1, so the Q-table update expression is: $Q(s,a) \leftarrow r + \gamma max_{a'} Q(s',a')$.
- γ: $[0 < \gamma \leq 1]$, we have set this value to 0,9.

Experimental results

- Systematic exploration of the combinations of state, reward system and discretization step.
- Numerical values of the results with different training time (expressed in terms of training episodes) in order to compare the learning of the same systems varying this parameter.
- For each combination, we show the results obtained in the test phase with 100 different initial configurations.

	state model				
	$S=(P_y,P_d,i)$		$S=(P_y,P_d,i,c)$		$S=(P_r,P_d,i,P_1,P_2)$
reward system	$\triangle s$ $0'5\,m.$	$\triangle s$ $0'2\,m.$	$\triangle s$ $0'5\,m.$	$\triangle s$ $0'2\,m.$	$\triangle s$ $0'5\,m.$
10	6.410	1.740	5.920	1.410	9.830
20	6.410	460	6.490	370	10.260
30	6.070	1.900	6.700	1.510	12.360
40	4.530	780	4.440	650	12.480
50	7.330	2.260	7.540	1.660	22.620
60	22.650	1.010	20.470	750	
70	23.290	1.090	20.450	620	
80	34.490	2.350	31.270	1.940	
90	37.970	2.810	36.430	1.980	

Table: Total episodes of the training phase (thousands of episodes)

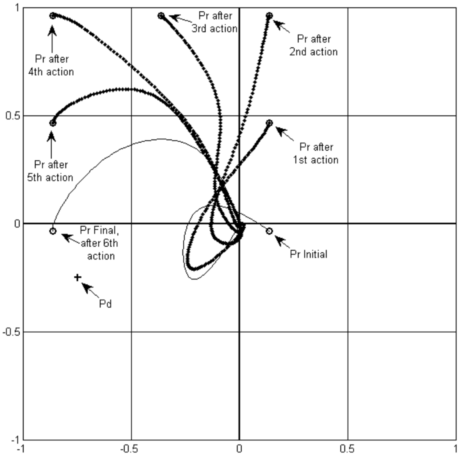

Figure: Episode where the tip of the hose robot reaches the goal

- The best results: reward system code 20 with the state defined as $S = (P_r, P_d, i, P_1, P_2)$.
  - Success rate, 77% of the test episodes.
  - The 2% of the test episodes concluded because the maximum allowed step count was reached.
  - The 21% of the test episodes failed either because the robot collided with the hose or because the whole system reached a non-feasible position.
- Increasing the training episodes the result in the test phase improve in two ways:
  - The number of episodes that finish with success have experienced a slight increase.
  - The number of episodes that fail decrease to increase the number of those that finish because the maximum allowed step count was reached.

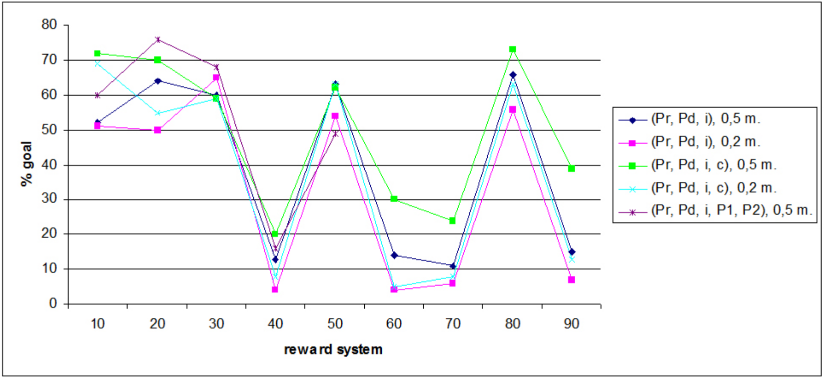

Figure: Percentage of successful runs (reaching goal) obtained in test phase with each reward system and state definition over a hundred simulations per combination of parameters.

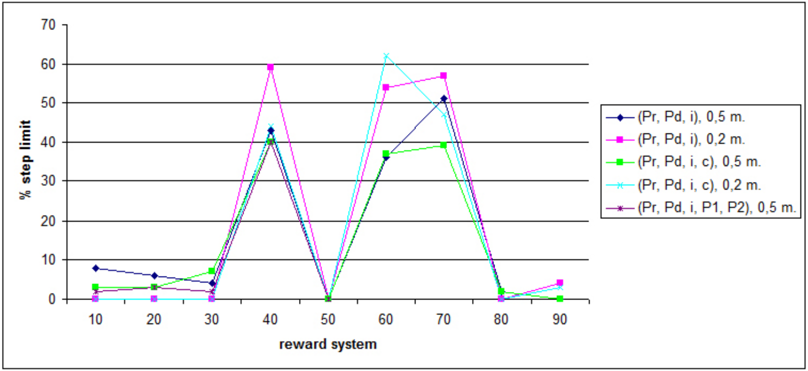

Figure: Percentage of runs terminated because they reached the limit number of steps obtained in test phase with each reward system and state definition over a hundred simulations per combination of parameters.

Conclusions

- Approximation of a controller for a hose transportation problem in a L-MCRS multiagent system using Reinforcement Learning methods (Q-learning).
- The work in this paper is restricted to a single robot moving.
- We have tested a number of combinations of the model, the reward systems and the step size used in the discretization process.
- Results of the training computational experiment are good for some combinations.

- We have paid much attention to the time needed to get a reasonably good training in terms of episodes.
- Viewing the experimental data we can realize:
  - Investing much more time in the training phase, the results improve, but the ratio of that improvement is not linear in relation to the computational effort.
  - One of the biggest issue while conducting experiments was the large duration of simulations:
    - Mainly because the hose model's computation requirements,
    - Also due to the huge number of episodes needed to explore before learned Q-table exploitation can yield good results.

Future work

- Optimize the state-action space representation, while keeping the most important information required for the learning purpose.
- Apply the learned control strategy on real robots to further validate the results.
- Long term research will deal with learning control strategies for a collection of robots attached along the hose.
- We will evaluate the application of hierarchical decomposition techniques.
- Application of alternative knowledge modeling paradigms.

Thanks for your attention.