Reinforcement Learning 2015/2016: Tutorial 7

1. **[Model-based RL]** RL learning aims at related tasks of optimising the value function, the policy and the behaviour based on the reward signal. The rewards is used only locally (in simple RL algorithms), such that models are used in order to use the experience of the agents in the environment more efficiently. Model learning faces analogous challenges: The model should be correct (w.r.t. the data available at a given time), it should not or overfit but generalise beyond the data, but not over-generalise and it should focus on relevant data (i.e. on regions of high reward or high regret) although "relevance" might change during learning. Asymptotically, the model is (ideally) on the one hand perfect and on the other hand not needed. Discuss the interaction of the aspects of the learning the action and the aspects of learning a model. It might make sense to remember here the Actor-Critic architecture.

2. **[Dyna-Q]** The nonplanning method looks particularly poor in Figure 8.6 because it is a one-step method; a method using eligibility traces would do better. Do you think an eligibility trace method could do as well as the Dyna method? Explain why or why not.[from Sutton&Barto]

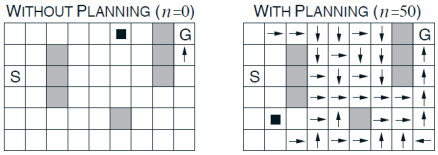

Figure 8.6: Policies found by planning and nonplanning Dyna-Q agents halfway through the second episode. The arrows indicate the greedy action in each state; no arrow is shown for a state if all of its action values are equal. The black square indicates the location of the agent.

3. **[Dyna-Q]** Why did the Dyna agent with exploration bonus, Dyna-Q+, perform better in the first phase as well as in the second phase of the blocking and shortcut experiments? Careful inspection of Figure 8.8 reveals that the difference between Dyna-Q+ and Dyna-Q narrowed slightly over the first part of the experiment. What is the reason for this? [from Sutton&Barto]

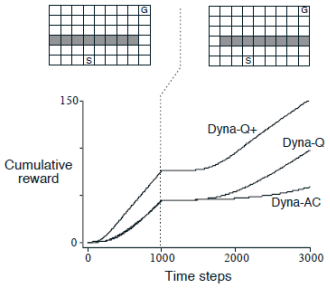

Figure 8.7: Average performance of Dyna agents on a blocking task. The left environment was used for the first 1000 steps, the right environment for the rest. Dyna-Q+ is Dyna-Q with an exploration bonus that encourages exploration. Dyna-AC is a Dyna agent that uses an actor–critic learning method instead of Q-learning.

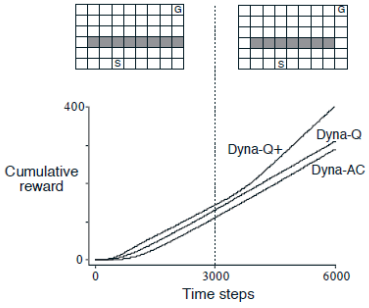

Figure 8.8: Average performance of Dyna agents on a shortcut task. The left environment was used for the first 3000 steps, the right environment for the rest.

4. **[Dyna-Q]** Dyna-Q+ uses exploration bonuses, e.g. of the form $r + \kappa\sqrt{n}$. This is a kind of an intrinsic reward. Discuss advantages and problems in connection to intrinsic rewards. Consider also the aspect of evolutionary learning.

5. **[Dyna-Q]** Prioritised sweeping (see Sutton&Barto section 8.4 in 2nd ed.): While Dyna agents select state-action pairs uniformly at random from the previously experienced pairs, it might be more efficient to use a non-uniform probability distribution. Why? Which state-action pairs should be preferred? Discuss the role of a goal states in this context.

6. **[MORL]** Under what conditions is a Pareto front not-connected? How can a MORL agent reach other connectivity components of the Pareto front? Can a scalarised MORL algorithm reach all points on the Pareto front by testing all combinations of weights in a weighted sum of the value function?

7. **[MORL]** How can policy gradient methods be adapted to the MORL problem?

8. **[MORL]** Instead of using MORL, it may be possible to provide the agent with appropriate state information, e.g. instead of a spatial state and rewards for (a) reaching a goal and (b) keeping batteries charged, we could design a the state that contains both information about the battery level and about the spatial position. Compare the two variants.

9. **[MARL]** Apply RL to the iterated prisoners dilemma. Discuss several scenarios: (i) Two prisoners, (ii) simultaneous plays between pairs randomly selected from many prisoners, (iii) prisoners are situated in a plane, plays with neighbours (iv) different state definitions in the group. Discuss also the effect of details of the RL algorithm.

10. **[Applications]** Assume your are heading a large team of researches and technicians working on building a humanoid robot. Motivate your coworkers to use RL in various tasks related to the project. Consider hierarchical approaches, specify sub-tasks and required resources. Discuss alternatives. What options for hybrid algorithms might be interesting here?

11. **[Applications]** Assume your are heading a large team of researches and technicians working in robot soccer. Motivate your coworkers to use RL in various tasks related to the project. Consider hierarchical approaches, specify sub-tasks and required resources. Discuss in particular team-level learning tasks and the role of models.