

RL and Planning Under Uncertainty (ANU, Sem2, 2008)
Tutorial 4: Review Questions
Tutorial Instructor: Scott Sanner

FINISH READING SUTTON & BARTO!!!

Questions:

1. Before you specify an algorithm to perform a task, what should you do? Why can this help?
2. What is the general 6-tuple that can be used to specify most agent-based problems? Can you solve such a problem if the states are unknown?
3. What are allowable finite-horizon objectives? What are allowable infinite horizon objectives? Do any of these yield stationary optimal policies? Which of the infinite-horizon objectives place more emphasis on the future?
4. Multiagent problems:
 - a. How do you model the action space and transition function for multiagent problems?
 - b. What is a good objective if the problem is 2-agent adversarial?
 - c. An objective when n-agents are purely self-interested?
 - d. What can you modify in the 6-tuple to encourage cooperation?
5. Why should properties that hold for any one-shot multiagent problem extend to any finite horizon multiagent sequential decision problem?
6. How would you model the objective for a problem where $T(s,a,s') \in [lb_{sas}, ub_{sas}]$?
7. Which of the following defines partial observability?
 - a. T is possibilistic (only know possible outcomes)
 - b. Z: $S \rightarrow O$ is not a bijective relation
 - c. Z is unknown to agent
8. Can partially observable problems be solved optimally by tracking an expectation over the state (i.e., belief state) and taking the action for the most likely state? If not, give a counterexample where this approach will not yield an optimal solution.
9. You have a choice between (a) a dollar amount drawn uniformly from [25,75] or (b) a dollar amount drawn uniformly from [0,100]. If I run an infinite (or very large number) of these trials, do you have a preference of (a) vs. (b)? If I only run one trial, do you have a preference of (a) vs. (b)? Can you formalize

a single objective that reflects this behavior?

10. If we only have preferences over lotteries (i.e., probabilistic outcomes), and we assume the four foundational axioms in the OneShot lecture (completeness, transitivity, independence, and continuity), can we assign values to each outcome such any expectation over these outcomes respects all lottery preferences? What happens if I scale these values by a positive affine transform? Does expected utility over these values still respect preferences?
11. Why should properties that hold for any one-shot multiagent problem extend to any finite horizon multiagent sequential decision problem? Even if partially observable? Can bandits optimally solve finite horizon problems? Caveats?
12. Define the value function for an MDP. Derive the value iteration update to compute the value function for a finite horizon h of decisions. Also derive the Q-function variant of this update. Show that the error decreases by γ on each iteration. What happens to the optimal value function and policy as $h \rightarrow \infty$?
13. Define policy iteration. What are its convergence properties?
14. Why are exploration policies required for convergence of all reinforcement learning algorithms? What are possible (dis)advantages of ϵ -greedy and Boltzmann exploration? Why are exploration policies not required in model-based methods?
15. Define the greedy policy for a value function assuming T & R are known. How to derive the greedy policy in a model-free RL environment?
16. Can we do (Q-based) value iteration in a model-free environment? What is this algorithm similar to? On-policy or off-policy? Convergence properties?
17. Define generalized policy iteration (GPI). Under what conditions is it guaranteed to converge? Define at least two ways to use sampled experience to estimate the value function for a given policy.
18. What are some (dis)advantages of $TD(\lambda < 1)$ vs. MC?
19. Assume a sum of weighted Gaussians as the function approximation method. Under what conditions on possible free parameters is this a linear-value approximation? Non-linear? Derive the gradient descent update rule for all parameters in the non-linear case.
20. Prove that the eligibility traces in function approximation generalize the fully enumerated state case.
21. If $TD(\lambda)$ control-learning with linear FA is underperforming, what should you do: (a) add new features, (b) switch to non-linear FA, or (c) switch to MC?
22. In model-based RL, do we need to throw out V/Q , T , or R when the policy is updated?