

Q520: Homework on MDP and Related Matters
Due, Tuesday, February 28

Problems 3–7 look like a lot of work, but they are all short arguments. I wrote them in a specific order, so I think it will be easiest to do them in that order. In particular, you may use one problem in doing the next ones. You also *may* use one problem in doing *previous* problems in the list, but if you do that, please be sure that at the end of the day you’re not *reasoning circularly*. This would be bankrupt reasoning, like bankrupt financial dealing: using a Visa card to pay off a MasterCard, and then turning around and doing it the opposite way.

1. From Tom Mitchell’s book chapter “Reinforcement Learning”, exercise 13.1 on page 388. (Please note that the book formulates MDPs a little differently than I did. It gives rewards on the actions, not on the states. You’ll want to keep this in mind as you read the chapter.)
2. From Tom Mitchell’s book chapter “Reinforcement Learning”, exercise 13.3 on page 388. I may not get to the the Q-learning algorithm until next week, but this is not so important. What you want to do is to formulate the problem of learning an optimal strategy in tic-tac-toe in MDP terms. The last part of the problem asks about running the Q-learning algorithm on your model, or rather on a different model. You can wait to do this, or else read the example of Q-learning in the text.
3. Let π be an unimprovable policy. (This means that the Policy Improvement algorithm would take π and return π itself in one step.) Show that the value function V^π gives a solution to the Bellman equations. That is, show that for all This means that that for all states s ,

$$V^\pi(s) = \text{reward}(s) + \gamma \max_{\alpha} \sum_t \text{go}(s, \alpha, t) V^\pi(t)$$

[This should not be hard once you understand the notation and the policy improvement material.]

4. Let M be an MDP. Let π_1 and π_2 be any two policies on this MDP. Show that there is a policy σ such that $\sigma \geq \pi_1, \pi_2$. This means that for all states s , $V^\sigma(s) \geq V^{\pi_1}(s)$ and $V^\sigma(s) \geq V^{\pi_2}(s)$.
5. A policy π is *optimal* if for all policies π' , $\pi \geq \pi'$. This means that for all states s , $V^\pi(s) \geq V^{\pi'}(s)$. Prove that π is unimprovable if and only if π is optimal.
6. Suppose we had numbers x_s , one for each state, and suppose that these numbers happened to solve the Bellman equation. That is, suppose that for all s ,

$$x_s = \text{reward}(s) + \gamma \max_{\alpha} \sum_t \text{go}(s, \alpha, t) x_t$$

Define a policy π by

$$\pi(s) = \operatorname{argmax}_{\alpha} \sum_t \text{go}(s, \alpha, t) x_t.$$

Prove that π is unimprovable.

7. Prove that the Bellman equation for M has exactly one solution.