

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 MPDs 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.)

The easiest solution is to change the given optimal policy, shown on the left below, to the one on the right.

→	→	
→	→	↑

→	→	
→	↑	↑

To see that this second policy is optimal, we only need to show that its long term expected discounted total is the same as the total for the first policy. This is easy, since the policies only differ on one square, and we are told that the long term total is 100 in both the upper-middle and lower-right squares.

2. From Tom Mitchell’s book chapter “Reinforcement Learning”, exercise 13.3 on page 388. First, let’s adopt a more compact representation of a tic-tac-toe board. For example, instead of

		X	
O			
			X

we shall write this board as $(-, X, -, O, -, -, -, -, X)$. We are only interested in the boards where there are either an equal number of X ’s as O ’s, or whether there is exactly one more X than O . We also are only interested in the boards where neither of the players has already won.

The states of our MDP are

- ★ (b, agent) , where b is a board; indicating that the next move is by the agent we are modelling
 - ★ $(b, \text{opponent})$, where b is a board; indicating that the next move is by the opponent.
 - ★ IP; indicating improper play
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 states s ,

$$V^\pi(s) = \text{reward}(s) + \gamma \max_{\alpha} \sum_t \text{go}(s, \alpha, t) V^\pi(t) \tag{1}$$

Since π is unimprovable, for all states s ,

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

Thus for all s we have

$$\max_{\alpha} \sum_t \operatorname{go}(s, \alpha, t) V^{\pi}(t) = \sum_t \operatorname{go}(s, \pi(s), t) V^{\pi}(t)$$

Multiplying by γ on both sides and also adding $\operatorname{reward}(s)$, we have

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

But the right hand quantity above is exactly $V^{\pi}(s)$. And so we see that

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

That is, we have (1) from above.

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 $V^{\sigma}(s) \geq V^{\pi_2}(s)$. Let

$$\sigma(s) = \begin{cases} \pi_1(s) & \text{if } V^{\pi_1}(s) > V^{\pi_2}(s) \\ \pi_2(s) & \text{if } V^{\pi_2}(s) \geq V^{\pi_1}(s) \end{cases}$$

We want to show that $V^{\sigma} \geq V^{\pi_1}$ and $V^{\sigma} \geq V^{\pi_2}$. For this, let V^{max} be the vector whose s th entry is the maximum of $V^{\pi_1}(s)$ and $V^{\pi_2}(s)$. Clearly $V^{max} \geq V^{\pi_1}$ and $V^{max} \geq V^{\pi_2}$. So, all we have to do is to show that $V^{\sigma} \geq V^{max}$.

We claim that $R + \gamma P^{\sigma} V^{max} \geq V^{max}$. To see this, we take the two cases on s : either $V^{\pi_1}(s) > V^{\pi_2}(s)$, or else $V^{\pi_2}(s) \geq V^{\pi_1}(s)$. In the first case, $\sigma(s) = \pi_1(s)$ by our definition above, and also $V^{max}(s) = V^{\pi_1}(s)$. And so

$$\begin{aligned} & R(s) + \gamma \sum_t \operatorname{go}(s, \sigma(s), t) V^{max}(t) \\ &= R(s) + \gamma \sum_t \operatorname{go}(s, \pi_1(s), t) V^{max}(t) \\ &\geq R(s) + \gamma \sum_t \operatorname{go}(s, \pi_1(s), t) V^{\pi_1}(t) \\ &= V^{\pi_1}(s) \\ &= V^{max}(s) \end{aligned}$$

This ends our proof of the claim in the first case. The second case is similar, but we use π_2 instead of π_1 .

So now we have our claim. Now we repeat the argument of the Policy Improvement Theorem.

By rearranging things, $R \geq (1 - \gamma P^{\sigma}) V^{max}$. And now it follows that

$$\begin{aligned} V^{\sigma} &= (I - \gamma P^{\sigma})^{-1} R \\ &\geq (I - \gamma P^{\sigma})^{-1} (1 - \gamma P^{\sigma}) V^{max} \\ &= V^{max} \end{aligned}$$

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.

Let π be unimprovable. To see that π is optimal, we take an arbitrary policy σ and show that $V^\sigma \leq V^\pi$.

Since π is unimprovable, we see that for all s ,

$$\sum_t \text{go}(s, \sigma(s), t) V^\pi(t) \leq \sum_t \text{go}(s, \pi(s), t) V^\pi(t).$$

In other words $P^\sigma V^\pi \leq P^\pi V^\pi$. Thus

$$R = V^\pi - \gamma P^\pi V^\pi \leq V^\pi - \gamma P^\sigma V^\pi = (I - \gamma P^\sigma) V^\pi.$$

And so we have our desired conclusion, as follows:

$$V^\sigma = (I - \gamma P^\sigma)^{-1} R \leq (I - \gamma P^\sigma)^{-1} (I - \gamma P^\sigma) V^\pi = V^\pi.$$

This concludes half of what we want to do in this problem, and so we turn to the second half.

Let π be optimal; we prove that π is unimprovable. If we run the Policy Iteration algorithm starting with π , we get a sequence of policies

$$\pi = \pi_0 < \pi_1 < \dots < \pi_n.$$

Since π is optimal, the sequence must just consist of π alone. That is, at the first improvement step, the improved version π' must be \equiv (equivalent) to π . So the output of the Policy Iteration algorithm is π itself. But this means that π is unimprovable.

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 ,

$$x_s = \text{reward}(s) + \gamma \max_\alpha \sum_t \text{go}(s, \alpha, t) x_t \quad (2)$$

Define a policy π by

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

Prove that π is unimprovable. Here is the proof. The definition of π tells us that for all s ,

$$\max_\alpha \sum_t \text{go}(s, \alpha, t) x_t = \sum_t \text{go}(s, \pi(s), t) x_t. \quad (3)$$

So when we run one step of the Policy Improvement Algorithm to get an improvement π' , we take $\pi'(s) = \pi(s)$ for all s . And this just says that π is unimprovable.

At this point, it is useful to also prove that for all s , $V^\pi(s) = x_s$. To do this, recall that the numbers $V^\pi(s)$ are the unique ones such that for all s ,

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

That is, the $V^\pi(s)$ numbers are the unique solution to a linear system. To show that the $V^\pi(s)$ numbers are the same as the x_s numbers, we'll show that the x 's satisfy the condition that uniquely defines the $V^\pi(s)$ numbers. That is, we'll show that

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

But this last equation follows from the assumption (2) and the fact in (3).

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

Let π be any policy, and let π' be an unimprovable policy such that $\pi' \geq \pi$. (This π' exists by what we did in class: run the Policy Improvement algorithm on π .) Problem 3 shows that the numbers $V^{\pi'}(s)$ solve the Bellman equation. So we know that the Bellman equation has at least one solution.

To see that it has *only* one solution, let the numbers x_s be a solution to the Bellman equation. We shall show that for all s , $x_s = V^{\pi'}(s)$. Let σ be as given in Problem 6. Then σ is an unimprovable policy, and for all s , $V^\sigma(s) = x_s$. By problem 5, σ and π' are both optimal. Since π' is optimal, we have that for all s , we have $V^{\pi'}(s) \geq V^\sigma(s)$. Since σ is optimal, we also have that for all s , $V^\sigma(s) \geq V^{\pi'}(s)$. Since these are all numbers, we have $V^{\pi'}(s) = V^\sigma(s)$. And so

$$x_s = V^\sigma(s) = V^{\pi'}(s).$$

This proves what we want.