

# Markov Decision Processes

## CS 221

### Section 6

October 30, 2009

Today we will discuss several sample MDP problems. The solutions are included here, so you can work through them on your own if you like.

#### 1. MDPs with Random Stopping Times

Suppose we have a Markov Decision Process (MDP)  $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P_{sa}, \gamma, R)$ , where  $\mathcal{S}$  is a discrete state space with  $n$  states, and the rewards are discounted by a factor  $\gamma$ . (Recall that  $P_{sa}(s')$  is the “transition model” and  $R(s)$  is the reward function.) We can view this process as a game where we begin in some state  $s_0 \in \mathcal{S}$  and take turns selecting actions and transitioning to new states, accumulating rewards along the way. At the  $n$ th turn, we first receive some (discounted) reward,  $\gamma^n R(s)$ , for the current state  $s$ . Then, we select an action,  $a \in \mathcal{A}$  and transition, randomly, to a new state  $s'$  according to the probabilities,  $P_{sa}(s')$ . Since the discount factor is  $\gamma < 1$ , our rewards become smaller and smaller as the game goes on. (Hence, the optimal strategy will try to accumulate big rewards early.)

Now consider a slight modification of this game. At the start of each turn we receive an *undiscounted* reward,  $R(s)$ , and then flip a biased coin that lands **heads** with probability  $\epsilon$ ,  $0 < \epsilon \leq 1$ . If the coin lands **heads**, then the game is stopped and we are left with whatever reward we have accumulated thus far. Otherwise, we choose our action and we transition to the next state according to  $P_{sa}$ , as usual. We will now show that this new game can be expressed as an MDP. In addition, we’ll also show that the value of this game (i.e., the largest reward we expect to gain from playing it) is equivalent to the *discounted* reward in the original MDP,  $\mathcal{M}$ .

Define a new MDP,  $\tilde{\mathcal{M}} = (\tilde{\mathcal{S}}, \mathcal{A}, \tilde{P}_{sa}, 1, \tilde{R})$ . This MDP has the same action space as  $\mathcal{M}$ , but the discount factor is 1, and we have a different state space, transition model, and reward function. We’ll construct the MDP  $\tilde{\mathcal{M}}$  so that it is just like the MDP  $\mathcal{M}$ , but with some modifications to include the coin-flipping rules defined above.

In particular, we’re going to add a new state called the “sink” state, which we’ll denote  $e$ . If the coin toss comes up **heads**, then we’ll transition, always, to this state and remain there forever (accumulating 0 reward each turn). If the coin toss is **tails**, then we’ll just transition according to  $P_{sa}$  as before, with no chance of entering the sink state,  $e$ .

- (a) Complete the construction by specifying explicitly  $\tilde{\mathcal{S}}$ ,  $\tilde{P}_{sa}$ , and  $\tilde{R}$  for the new MDP,  $\tilde{\mathcal{M}}$ .

**Answer:** Let  $\tilde{\mathcal{S}} = \mathcal{S} \cup \{e\}$ , where  $e$  is the new sink state.

Now, let's assume we're in a state  $s \in \mathcal{S}$  (i.e.,  $s \neq e$ ), then we have:

$$\begin{aligned}\tilde{P}_{sa}(s'|\mathbf{heads}) &= \begin{cases} 1 & \text{if } s' = e, \\ 0 & \text{if } s' \in \mathcal{S}, \end{cases} \\ &\text{and} \\ \tilde{P}_{sa}(s'|\mathbf{tails}) &= \begin{cases} 0 & \text{if } s' = e, \\ P_{sa}(s') & \text{if } s' \in \mathcal{S}. \end{cases}\end{aligned}$$

Thus, using  $p(\mathbf{heads}) = \epsilon$ , we can derive:

$$\begin{aligned}\tilde{P}_{sa}(s') &= \tilde{P}_{sa}(s|\mathbf{heads})p(\mathbf{heads}) + \tilde{P}_{sa}(s|\mathbf{tails})p(\mathbf{tails}) \\ &= \begin{cases} \epsilon & \text{if } s' = e, \\ (1 - \epsilon) \cdot P_{sa}(s') & \text{if } s' \in \mathcal{S}. \end{cases}\end{aligned}$$

This covers the transition probabilities from any state in  $\mathcal{S}$  to any other state in  $\tilde{\mathcal{S}}$ . On the other hand, if we're in the sink state,  $s = e$ , then the transition model is:

$$\tilde{P}_{ea}(s') = \begin{cases} 1 & \text{if } s' = e, \\ 0 & \text{if } s' \in \mathcal{S}. \end{cases}$$

(That is, we transition back to  $e$  with probability 1, and can never escape.) This completely defines the transition model.

Now we need to define the reward function. If  $s \in \mathcal{S}$ , then we just get our usual reward. That is,  $\tilde{R}(s) = R(s)$ . But if we're in the sink state,  $e$ , then we receive no reward:  $\tilde{R}(e) = 0$ . This completes the definition of our MDP,  $\tilde{\mathcal{M}}$ .

- (b) Show that  $\tilde{V}^*(e) = 0$ , where  $\tilde{V}^*$  is the optimal value function for the MDP  $\tilde{\mathcal{M}}$ .

**Answer:** From our definitions above, the Bellman equation gives:

$$\begin{aligned}\tilde{V}^*(e) &= \tilde{R}(e) + \gamma \max_{a \in \mathcal{A}} \sum_{s' \in \tilde{\mathcal{S}}} \tilde{P}_{ea}(s') \cdot \tilde{V}^*(s') \\ &= 0 + \gamma \max_{a \in \mathcal{A}} \tilde{V}^*(e) \\ &= \gamma \tilde{V}^*(e) \\ \Rightarrow \tilde{V}^*(e) &= 0.\end{aligned}$$

- (c) Now show that when  $s \in \mathcal{S}$ , the Bellman equation for  $\tilde{\mathcal{M}}$  is the same as the Bellman equation for  $\mathcal{M}$ , but with a different discount factor.

**Answer:** Using the Bellman equation for  $\tilde{\mathcal{M}}$ , we have:

$$\begin{aligned}
 \tilde{V}^*(s) &= \tilde{R}(s) + 1 \cdot \max_{a \in \mathcal{A}} \sum_{s' \in \tilde{\mathcal{S}}} \tilde{P}_{sa}(s') \cdot \tilde{V}^*(s') \\
 &= R(s) + \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \tilde{P}_{sa}(s') \cdot \tilde{V}^*(s') + \tilde{P}_{sa}(e) \cdot \tilde{V}^*(e) \\
 &= R(s) + \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \tilde{P}_{sa}(s') \cdot \tilde{V}^*(s') \\
 &= R(s) + \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} (1 - \epsilon) \cdot P_{sa}(s') \cdot \tilde{V}^*(s') \\
 &= R(s) + (1 - \epsilon) \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P_{sa}(s') \cdot \tilde{V}^*(s').
 \end{aligned}$$

Compare this to the Bellman equation for  $\mathcal{M}$ :

$$V^*(s) = R(s) + \gamma \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P_{sa}(s') \cdot V^*(s')$$

We see that they are identical when  $\gamma = 1 - \epsilon$ . This gives us some intuition for how the discount factor affects the optimal value of the MDP: Discounting the game's rewards by a factor of  $\gamma$  is the same as playing without discounting, but where the probability that the game will end during the next turn is  $\epsilon = 1 - \gamma$ . So, if  $\gamma$  is small, this is like playing a game where we are very likely to get stopped soon—so we should try to gather as much reward as possible early on, rather than waiting for bigger rewards later.

## 2. MDPs: Decision-making with Model Errors

Suppose that we want to solve an MDP with a transition model given by  $P_{sa}(s')$ , but we don't have access to the true transition model. Instead, we only have an estimate of the model,  $\hat{P}_{sa}(s')$ , with the following property:

$$\sum_{s' \in \mathcal{S}} |\hat{P}_{sa}(s') - P_{sa}(s')| \leq \epsilon, \forall s, a.$$

That is, the sum of the absolute values of the errors in our model is bounded by  $\epsilon$ . Recall that  $V^\pi(s)$  is the expected sum of rewards if we begin in state  $s$  and act according to policy  $\pi$ . In particular,  $V^\pi(s)$  and  $\hat{V}^\pi(s)$  satisfy the following:

$$\begin{aligned}
 V^\pi(s) &= R(s) + \gamma \sum_{s' \in \mathcal{S}} P_{s\pi(s)}(s') V^\pi(s'), \\
 \hat{V}^\pi(s) &= R(s) + \gamma \sum_{s' \in \mathcal{S}} \hat{P}_{s\pi(s)}(s') \hat{V}^\pi(s').
 \end{aligned}$$

So, for any policy  $\pi$ ,  $V^\pi$  is the expected sum of rewards when acting according to  $\pi$  using the true model, while  $\hat{V}^\pi$  is the estimated value according to our estimated model,  $\hat{P}_{sa}$ . Finally, define  $\|V\|_\infty = \max_{s \in \mathcal{S}} |V(s)|$ .

Show that if you use any policy  $\pi$ , the error in its expected reward computed from the estimated model is bounded. More specifically, show that

$$\|\hat{V}^\pi - V^\pi\|_\infty \leq \frac{\gamma\epsilon}{1-\gamma} \|V^\pi\|_\infty$$

for *any* policy  $\pi$ . You may find the “triangle inequality” useful:  $|\sum_i a_i| \leq \sum_i |a_i|$ .

**Answer:** We will start with the left-hand side of the statement we are trying to prove, and manipulate it until we get the right-hand side.

$$\|\hat{V}^\pi - V^\pi\|_\infty \equiv \max_s |\hat{V}^\pi(s) - V^\pi(s)| \quad (1)$$

$$= \max_s \left| R(s) + \gamma \sum_{s' \in S} \hat{P}_{s\pi(s)}(s') \hat{V}^\pi(s') - R(s) - \gamma \sum_{s' \in S} P_{s\pi(s)}(s') V^\pi(s') \right| \quad (2)$$

$$= \gamma \max_s \left| \sum_{s' \in S} \hat{P}_{s\pi(s)}(s') \hat{V}^\pi(s') - P_{s\pi(s)}(s') V^\pi(s') \right| \quad (3)$$

$$= \gamma \max_s \left| \sum_{s' \in S} \hat{P}_{s\pi(s)}(s') \hat{V}^\pi(s') - P_{s\pi(s)}(s') V^\pi(s') \right. \\ \left. + \left[ \hat{P}_{s\pi(s)}(s') V^\pi(s') - \hat{P}_{s\pi(s)}(s') V^\pi(s') \right] \right| \quad (4)$$

$$= \gamma \max_s \left| \sum_{s' \in S} \hat{P}_{s\pi(s)}(s') (\hat{V}^\pi(s') - V^\pi(s')) + (\hat{P}_{s\pi(s)}(s') - P_{s\pi(s)}(s')) V^\pi(s') \right| \quad (5)$$

$$\leq \gamma \max_s \sum_{s' \in S} \left| \hat{P}_{s\pi(s)}(s') (\hat{V}^\pi(s') - V^\pi(s')) + (\hat{P}_{s\pi(s)}(s') - P_{s\pi(s)}(s')) V^\pi(s') \right| \quad (6)$$

$$\leq \gamma \max_s \sum_{s' \in S} \left| \hat{P}_{s\pi(s)}(s') \hat{V}^\pi(s') - V^\pi(s') \right| + \left| \hat{P}_{s\pi(s)}(s') - P_{s\pi(s)}(s') \right| V^\pi(s') \quad (7)$$

$$= \gamma \max_s \sum_{s' \in S} \hat{P}_{s\pi(s)}(s') \left| \hat{V}^\pi(s') - V^\pi(s') \right| + \left| \hat{P}_{s\pi(s)}(s') - P_{s\pi(s)}(s') \right| |V^\pi(s')| \quad (8)$$

$$\leq \gamma \max_s \sum_{s' \in S} \hat{P}_{s\pi(s)}(s') \|\hat{V}^\pi - V^\pi\|_\infty + \left| \hat{P}_{s\pi(s)}(s') - P_{s\pi(s)}(s') \right| \|V^\pi\|_\infty \quad (9)$$

$$= \gamma \max_s \left[ \|\hat{V}^\pi - V^\pi\|_\infty \sum_{s' \in S} \hat{P}_{s\pi(s)}(s') + \|V^\pi\|_\infty \sum_{s' \in S} \left| \hat{P}_{s\pi(s)}(s') - P_{s\pi(s)}(s') \right| \right] \quad (10)$$

$$\leq \gamma \max_s \left[ \|\hat{V}^\pi - V^\pi\|_\infty \cdot 1 + \|V^\pi\|_\infty \cdot \epsilon \right] \quad (11)$$

In equation (1) we used the definition of the max norm. In equation (2) we expanded the definition of  $\hat{V}^\pi(s)$  and  $V^\pi(s)$ , according to the equations shown in the problem statement. In equation (3), we used the fact that  $R(s) - R(s) = 0$ . We also combined the two summations and factored  $\gamma$  outside of the max. In equation (4) we add 0 (in the form of  $\left[ \hat{P}_{s\pi(s)}(s') V^\pi(s') - \hat{P}_{s\pi(s)}(s') V^\pi(s') \right]$ ). In equation (5) we rearrange terms, factoring a  $\hat{P}_{s\pi(s)}(s')$  out of two terms and a  $\hat{V}^\pi(s')$  out of the other two terms. In equation (6) we use the “triangle inequality” which was said to be useful. In equation (7) we use the same inequality again, this time on the terms inside the summation. In equation (8) we use the fact that  $\hat{P}_{s\pi(s)}(s') \geq 0$  for all  $s'$ , since it is a probability distribution. We also used the fact that  $|ab| = |a||b|$ . In equation (9) we used (twice) the fact that

$$V^\pi(s') \leq \max_{s''} |V^\pi(s'')| \equiv \|V^\pi\|_\infty.$$

In equation (10) we separated the summation into two summations, and factored out the max-norm terms, since they are constant. In equation (11) we used the fact that  $\sum_{s' \in S} \hat{P}_{s\pi(s)}(s') = 1$ , since again  $\hat{P}$  is a probability distribution. We also used the fact given to us in the problem statement concerning the error of our model:

$$\sum_{s' \in S} |\hat{P}_{sa}(s') - P_{sa}(s')| \leq \epsilon, \forall s, a.$$

Summarizing the resulting inequality, we have:

$$\|\hat{V}^\pi - V^\pi\|_\infty \leq \gamma \|\hat{V}^\pi - V^\pi\|_\infty + \gamma \epsilon \|V^\pi\|_\infty.$$

Collecting the  $\|\hat{V}^\pi - V^\pi\|_\infty$  terms on the left, and then dividing by  $1 - \gamma$  gives:

$$\|\hat{V}^\pi - V^\pi\|_\infty \leq \frac{\gamma \epsilon}{1 - \gamma} \|V^\pi\|_\infty.$$

### 3. MDPs: Reward Functions<sup>1</sup>

In class we discussed Markov Decision Problems (MDPs) formulated with a reward function  $R(s)$  just over states. Sometimes MDPs are formulated with a reward function  $R(s, a)$  that also depends on the action taken or a reward function  $R(s, a, s')$  that also depends on the outcome state.

(a) Write the Bellman updates for these formulations.

**Answer:** The formulation of the Bellman update in class was:

$$V(s) := R(s) + \gamma \max_a \sum_{s'} P_{sa}(s') V(s') \quad (12)$$

These update equations essentially perform a “one-step lookahead” on the MDP to compute the values resulting from the current optimal action—the first term  $R(s)$  gives the immediate reward, and the second term gives the best expected discounted reward, assuming the value function  $V$  is accurate. Thus, if the reward achieved in state  $s$  is also a function of the action  $a$  chosen in the state  $s$ , we get:

$$V(s) := \max_a \left( R(s, a) + \gamma \sum_{s'} P_{sa}(s') V(s') \right) \quad (13)$$

Similarly, if the reward achieved in state  $s$  is a function of the action  $a$  chosen and the state  $s'$  reached, we get:

$$V(s) := \max_a \sum_{s'} P_{sa}(s') (R(s, a, s') + \gamma V(s')) \quad (14)$$

Equation (14) is the most general; (13) and (12) can be derived by moving the reward function outside the sum and max operations.

<sup>1</sup>Problem taken from Russell & Norvig, Pg. 647.

- (b) Show how an MDP with reward function  $R(s, a, s')$  can be transformed into a different MDP with reward function  $R(s, a)$ , such that optimal policies in the new MDP correspond exactly to optimal policies in the original MDP.

**Answer:** Suppose the original MDP  $M = (S, A, P_{sa}, R, \gamma)$ , where  $R$  depends on the current state, the action taken and state transitioned to. Define a new MDP  $\tilde{M} = (S, A, P_{sa}, \tilde{R}, \gamma)$  by defining the reward function as:  $\tilde{R}(s, a) = \sum_{s'} P_{sa}(s') R(s, a, s')$ . The MDPs  $M$  and  $\tilde{M}$  will have the same optimal value function, as any action  $a$  in any state  $s$  leads to the same final state  $s'$  and also gives the same expected reward in both MDPs. It follows that optimal policies in  $M$  will correspond to optimal policies in  $\tilde{M}$ . [More formally, using the notation from class:

$$\begin{aligned} \tilde{\pi}^*(s) &= \arg \max_a \left( \tilde{R}(s, a) + \gamma \sum_{s'} P_{sa}(s') \tilde{V}^*(s) \right) \\ &= \arg \max_a \sum_{s'} (P_{sa}(s') R(s, a, s') + \gamma P_{sa}(s') V^*(s)) = \pi^*(s) \end{aligned}$$

where the second step follows using the definition of  $\tilde{R}$  and the fact that the optimal value function in the two MDPs is equal.]

- (c) Now do the same to convert MDPs with  $R(s, a)$  into MDPs with  $R(s)$ .

**Answer:** Suppose the original MDP  $M = (S, A, P, R, \gamma)$ , where  $R$  depends on the current state and the action taken. Define a new MDP  $\tilde{M} = (\tilde{S}, \tilde{A}, \tilde{P}, \tilde{R}, \tilde{\gamma})$  as follows:

- $\tilde{S}$  consists of the set  $S \cup S \times A$ . In other words, each state  $\tilde{s}$  in the new MDP is represented either as  $s$  or using a tuple  $(s, a)$ , where  $s \in S$  and  $a \in A$ .
- $\tilde{A}$  allows all actions from  $A$  if the state  $\tilde{s} \in S$ ; if  $\tilde{s} \in S \times A$ , then only a special action  $a_0$  is allowed (and it will turn out to not matter). Our goal is to break down a transition  $s \xrightarrow{a} s'$  in the original MDP to a two step transition  $s \xrightarrow{a} (s, a) \xrightarrow{a_0} s'$ . The point of the intermediate state is that it implicitly records the action to be taken, and this will allow us to represent the reward just using the state.
- The transition function for the first of two steps above is deterministic, and the state is obtained by appending the action taken to the current state. For the second step, the transition function applies the original transition function  $P_{sa}$  to sample a new state  $s'$ ; note that both  $s$  and  $a$  are known implicitly in this intermediate state.
- The first of two intermediate steps has zero reward. The second step has a reward  $\tilde{R}(\tilde{s}) = R(s, a)/\sqrt{\gamma}$ , where  $\tilde{s} = (s, a)$  in our representation.
- The discount factor should be such that receiving a one-step reward  $r$  should be equivalent (in the discounted reward sense) to the two-step rewards 0 and  $r/\sqrt{\gamma}$ . This can be achieved by setting  $\tilde{\gamma} = \sqrt{\gamma}$ . To see this, suppose a sequence of rewards in the original MDP was  $r_1, r_2, r_3, \dots$ . Then, the discounted reward in the new MDP is the same as the original MDP:

$$0 + \sqrt{\gamma} \frac{r_1}{\sqrt{\gamma}} + \sqrt{\gamma}^2 \cdot 0 + \sqrt{\gamma}^3 \frac{r_2}{\sqrt{\gamma}} + \dots = r_1 + \gamma r_2 + \gamma^2 r_3 + \dots$$

Two steps of the new MDP thus produce the same resulting state and discounted reward as one step of the old MDP. Thus, these MDPs will have corresponding optimal policies.

**Note:** Don't worry if you don't completely understand the details of the proofs in parts (b) and (c).