HW05: MDP II & RL I *

Alex Clemmer, u0458675

February 28, 2012

1 AIMA Problem 17.10

Assumptions I make: First off, while the Norvig Russell text was helpful, the Sutton Barto book has a much clearer formulation of policy iteration. I will be using that formulation instead. Since they're both required reading, they should either be equivalent, or simultaneously correct.

1. By "qualitatively" I assume you are asking what we can tell without explicitly determining the optimal policy.

   We know for sure that this is a non-optimal policy. The reason is, in state 2, we will incur a lot more cost than in state 1, and because it will take a while to finally get to state 3, it is much better to spend a couple turns moving to state 1 and then repeatedly executing b.

   The second thing is less certain, but it is probably the case that the optimal policy is going to be to execute action b on state 1 and action a on state 2.

2. The algorithm calls for a heuristic parameter θ, but for reasonable choices, it doesn't matter what you pick, as it always results in the same thing and converges with two policy evaluations. The first total iteration works out as follows. I wrote a Python script to trudge through it, and it will change with values of theta, but for my values it workes as so:

Policy Evaluation

$\pi(s)$	possible actions and their values a	"best" action
b	$[(a:-1.8),(b:-0.9)]$	b
b	$[(a:-1.20),(b:-1.8)]$	a

Value Iteration

s	v ("old" value)	$V(s)$ ("updated" value)
state 1	-0.9	0
state 2	-1.92	0

   In the first policy iteration, the perscribed policy is not the same as our policy, and the result is that we must value iterate. Value iteration after first iteration gives a new $\Delta = 1.9128$, which is less than our θ, so we stop value iteration. Now that we have new values, it's time to evaluate policy again.

Policy Evaluation

$\pi(s)$	possible actions and their values a	"best" action
b	$[(a:-3.49),(b:-1.70)]$	b
a	$[(a:-2.29),(b:-3.50)]$	a

*CS 5300 AI; Spring 2012 Instructor: Jur van den Berg, University of Utah

This completes our process, and we have converged on a policy. Let s_1 stand for state 1, and let s_2 stand for state 2:

$$\pi := \{(s_1 \to a), (s_2 \to b)\} \tag{1}$$

3. The problem implies there are problems. Maybe there are with the Norvig Russell formulation, but the Sutton Barto version of the algorithm actually runs a lot *faster*: there is usually one traversal of the value iteration loop total, regardless of our γ. In contrast, if recommended actions for both b, convergence can take quite awhile depending on gamma.

   The optimal policy does not depend on γ in this case.

2 Mission to Mars II

There are a few ways to do this one, especially since the Russell/Norvig book does it one way and the Sutton/Barto book does it another way.

Policy Evaluation

s	a	$Q_0(s,a)$	$Q_1(s,a)$	$Q_2(s,a)$	$Q_3(s,a)$	$Q_4(s,a)$	$Q_5(s,a)$
$cool$	$slow$	0	2.4	2.4	2.4	2.4	2.4
$cool$	$fast$	0	0	7.296	13.296	12.1196	12.1196
$warm$	$slow$	0	0	0	0	0	8.94273
$warm$	$fast$	0	0	0	0	0	0

3 AIMA Problem 21.2

The main reason this is the case is the ADP agent will be biased against catastrophic outcomes, since it weights them the same. This biases it against exploration, which can cause the transition model to be different form the actual MDP. The result is that the policy π may be optimal, but the transition function may have hidden the propensity to seek rewards, and therefore by that argument it may be improper.

The case of $\gamma = 1$ will then fail because it will essentially view rewards as additive. However if we wait for the policy to be terminal, it is still possible to end up with a policy that is optimal even if the transition function is awful, and so policy evaluation won't fail.