CSC 2534— Decision Making Under Uncertainty

Assignment 2 — Solutions

Craig Boutilier — Fall 2014

1. **[30 points]**

   There are several ways to justify the optimal poliies and value functions (e.g., just running policy or value iteration). Finding the indifference probabilities, however, requires that you compute an explicit expression for the value of the two policies, a or b. For this reason, the way I would have done this is simply to write out the value equations for the two policies and compare their actual values at state s_1 when the stickiness and risky probabilities are set as required. This is very simple, since this MDP only has two (deterministic) policies, and each policy has only three or four reachable states. The equations are as follows:

   For policy a:

   $$v_1 = 0.95(0.1v_1 + 0.9v_2) \tag{1}$$
   $$v_2 = 0.95(p_S v_2 + (1 - p_S)v_4) \tag{2}$$
   $$v_4 = 10 + 0.95(0.1v_4 + 0.9v_1) \tag{3}$$

   For policy b:

   $$v_1 = 0.95(0.1v_1 + 0.9v_3) \tag{4}$$
   $$v_3 = 0.95(0.1v_3 + p_R v_5 + (0.9 - p_R)v_4) \tag{5}$$
   $$v_4 = 10 + 0.95(0.1v_4 + 0.9v_1) \tag{6}$$
   $$v_5 = -10 + 0.95(0.1v_5 + 0.9v_1) \tag{7}$$

   (a) Fix $p_S = 0.2$.

   With $p_R = 0.01$, we solve the policy a equations and obtain: $v_1 = 60.26, v_2 = 63.78, v_4 = 67.98$. The policy b equations give: $v_1 = 61.52, v_3 = 65.12, v_5 = 47.07, v_4 = 69.17$. The optimal policy is b (take the risky path).

   With $p_R = 0.03$, the policy a equations are the same as above. The b equations give: $v_1 = 58.72, v_3 = 62.16, v_5 = 44.43, v_4 = 66.52$. The optimal policy is a (take the sticky path).

   To find the indifference level for p_R, we simply leave p_R variable in the b equations,

and add the constraint that $v_1 = 60.26$ (it's value under a). Solving for p_R gives $p_R = 0.018989$.

(b) We do the same as above with $p_S = 0.6$.

With $p_R = 0.1$, policy a gives: $v_1 = 43.67, v_2 = 46.23, v_4 = 52.31$. Policy b gives: $v_1 = 48.93, v_3 = 51.80, v_5 = 35.18, v_4 = 57.28$. The optimal policy is b (take the risky path).

With $p_R = 0.2$, the policy a equations are the same as above. The b equations give: $v_1 = 34.95, v_3 = 37.00, v_5 = 21.97, v_4 = 44.07$. The optimal policy is a (take the sticky path).

The indifference level for p_R is 0.13762.

2. **[40 points (adds up to 42, so possible 2pt "bonus"]**

(a) **[4]** One stage is insufficient since it prevents one form using a conditional plan/policy in which one first does a test and then determines a treatment based on the result of the test. This would require at least two stages, and such policies are optimal in belief states in which testing has not been done and their is sufficient uncertainty regarding the patient's disease.

Three (or more) stages are not necessary because: (i) once testing is completed, any further testing adds negative value to a policy; and (ii) once a treatment is applied, any further treatment add negative value to a policy.

(b) **[0]** To be ignored (as instructed).

(c) **[8]** There are six one-stage plans, one for each possible action: Null, M1, M2, M3, T1, T2. Plans M1, M2, M3 and Null are all useful:

- M1 is optimal in any state where Y holds (i.e., belief state where X is true with probability 1) or any belief state where Y is quite likely.
- M2 is optimal in any state where X holds (i.e., belief state where X is true with probability 1) or any belief state where X is quite likely.
- M3 is optimal in belief states where X and Y are each reasonably likley.
- Null is optimal in any state where Tr (treated) holds or belief state where is it quite likely.

The two test plans (T1 and T2) are pointwise dominated by the Null action: in any state s the Null action has a cost of zero and leads to the same terminal state s (receiving the terminal reward at s). In any state s a test action has a cost of -2 and leads to the same terminal state s (again with the same terminal reward). So the each test action has a total reward that is 2 less than that of Null, at ant state s, with one stage to go.

(d) **[10]** Recall that you were to assume that $\Pr(T1 = H|Tr) = 1.0$ and $\Pr(T2 = Y|Tr) = 1.0$. The ten α-vectors (over the first four states) are as follows:

- α_1: [16, 14.4, 3, -102]
- α_2: [16.3, 0.8, 5, -102]
- α_3: [17.2, 8, 4, -102]
- α_4: [-1.1, 14.6, 7.6, -102]
- α_5: [12.9, 0.2, 7.4, -102]
- α_6: [7.3, 8.2, 7.8, -102]
- α_7: [0, 3.6, 3, -102]
- α_8: [-2, 18, 4, -102]
- α_9: [0, 20, 6, -100]
- α_{10}: [12, 12, 8, -100]

(e) **[3]** Vector α_6 is pointwise dominated by α_{10}. Vector α_7 is pointwise dominated by α_1 (or α_3 or α_{10} or α_6). Vector α_8 is pointwise dominated by α_9.

(f) **[14]** Here are some example belief states (yours may not be the same, but should have similar "structure" or approximate ratios over the three states s_1, s_2, s_3. Belief state b_1 is one in which vector α_i is optimal. The table shows the expected value of each of the seven plans/vectors at each of the seven belief states

Belief State	$\Pr(s_1)$	$\Pr(s_2)$	$\Pr(s_3)$	α_1	α_1	α_1	α_1	α_1	α_1	α_{10}
b_1	0.5	0.5	0	**15.2**	8.55	12.6	6.75	6.55	10	12
b_2	0.5	0	0.5	9.5	**10.65**	10.6	3.25	10.15	3	10
b_3	1	0	0	16	16.3	**17.2**	-1.1	12.9	0	12
b_4	0	0.2	0.8	5.28	4.16	4.8	**9**	5.96	8.8	8.8
b_5	0.41	0	0.59	8.33	9.633	9.412	4.033	**9.655**	3.54	9.64
b_9	0	1	0	14.4	0.8	8	14.6	0.2	**20**	12
b_{10}	0.25	0.25	0.5	9.1	6.775	8.3	7.175	6.975	8	**10**

(g) **[3]** The optimal plan for s_4 is "Null; Null."

3. **[30 points]**

(a) We need to set depth

$$d \geq \log_\beta(\frac{(1-\beta)\varepsilon}{R^+}) - 1$$

This can be derived by observing that the error is bounded by

$$\beta^{d+1}\frac{R^+}{1-\beta}$$

which we prove inductively (an somewhat less rigorous, but intuitive, convincing justification would suffice).

That this relation holds for depth $d = 0$ is seen by observing that $V^0(s) = R(s)$, while

$$V^*(s) = \max_a R(s) + \beta \sum_t \Pr(s, a, t) V^*(t)$$

Since $V^*(t)$ is bounded by $\frac{R^+}{1-\beta}$, the relation holds.

Assume it holds for depth $d - 1$. Then

$$V^d(s) = \max_a Q^d(s, a) = \max_a R(s) + \beta \sum_t \Pr(s, a, t) V^{d-1}(t)$$

By the inductive hypothesis, the error in the Q-estimates, $Q^*(a) - Q^d(a)$, is less than $\beta\beta^d \frac{R^+}{1-\beta}$. The only other source of error is if some action a_d maximizes $Q^d(s, a)$ while a different action a_* maximizes $Q^*(s, a)$ (so that $V^*(s) = Q^*(s, a)$). But if this is the case,

$$\begin{aligned} Q^*(s, a_*) &\leq Q^d(s, a_*) + \beta^{d+1} \frac{R^+}{1-\beta} \\ &\leq Q^d(s, a_d) + \beta^{d+1} \frac{R^+}{1-\beta} \\ &\leq V^d(s) + \beta^{d+1} \frac{R^+}{1-\beta} \end{aligned}$$

Thus $V^*(s) \leq V^d(s) + \beta^{d+1} \frac{R^+}{1-\beta}$ (and by definition must at least as great as $V^d(s)$.

(b) We need to set depth

$$d \geq \log_\beta \frac{\varepsilon}{\delta}$$

This can be derived by observing that the error is bounded by

$$\beta^d \delta$$

using an inductive argument similar to the one above. The difference in the β^d vs. β^{d+1} terms has to do with the fact that in part (a) we estimate value at the leaves using R. This means that the error at the leaves in this w.r.t. to the true value function is discounted by β, since they must agree on the initial term $R(t)$ (for leaf state t).

(c) One simple way to prune the search tree in a way that exploits the heuristic function, specifically, its accuracy parameter δ is as follows. As we build the tree, we evaluate the Q-values of each action using its immediate successors and $\widetilde{V}$. We know that this value $\widetilde{Q}(s, a)$ is within δ of $Q(s, a)$ (this is for any interior node of the tree, not just the start state). If $\widetilde{Q}(s, a) + \delta$ is less than the lower bound on the values of any other action a' at state s, we need not expand the tree below the immediate successors of action a.

Note that this lower bound can be $\widetilde{Q}(s, a')$ for each a' whose successors haven't ye been evaluated, or could be a more refined estimate of $Q(s, a')$, if we;ve expanded the tree below a'. The precise mechanism will depend on how one expands the tree (e.g., depth-

first, breadth-first, in a heuristically chosen order, etc.)