MATH4406 (Control Theory), HW4 (Unit 5) Infinite Horizon Discounted MDP – Part 1.

Prepared by Yoni Nazarathy, Last Updated: September 15, 2014

This homework project is about Infinite Horizon Discounted MDP. You will investigate MDPs on a birth-death structure.

- Use Chapter 6 of [Put94]. You can get the chapter on-line from the library.
- By now the programming aspect should be less of a challenge. But still allow enough time for this. **This assignment is very programming intensive.**
- Please make sure to present your results in a clear and organised manner. Numerical output results should always be well explained and documented. Labels on graphs, diagrams, tables etc...
- Hand-in all code (preferably as an appendix).

Consider an MDP on states $\mathcal{S} = \{-5, -4, \dots, 4, 5\}$. The action set for states $s \in \{-4, \dots, 4\}$ is $\mathcal{A}_s = \{-1, 1\}$. For state -5 the action set is $\mathcal{A}_{-5} = \{3\}$. For state 5 the action set is $\mathcal{A}_5 = -3$. The horizon is infinite and the objective is,

$$\max_{\pi \in \Pi^{MD}} \quad \lim_{T \to \infty} \mathbb{E}\,{}_s^{\pi}\,[\sum_{t=1}^{T} \lambda^{t-1}\, r(X_t, Y_t)],$$

where X_t is the state at time t, Y_t is the action at that time and $\lambda \in (0,1)$. The rewards are the product of the state and the action, namely, $r(s,a) = s \cdot a$. The transition probabilities for $s \in \{-4, \dots, 4\}$ are

$$p(j \,|\, s, -1) = \begin{cases} 3/4, & j = s-1, \\ 1/4, & j = s+1. \end{cases} \qquad \text{and} \qquad p(j \,|\, s, 1) = \begin{cases} 1/4, & j = s-1, \\ 3/4, & j = s+1. \end{cases}$$

For the boundary states, -5 and 5 the transition probabilities are

$$p(j \,|\, -5, 3) = \begin{cases} 1/2, & j = -5, \\ 1/2, & j = -4, \end{cases} \qquad \text{and} \qquad p(j \,|\, 5, -3) = \begin{cases} 1/2, & j = 4, \\ 1/2, & j = 5. \end{cases}$$

The boundary states are costly: Each time unit there costs 15 units and the expected number of time spent in these states is 2 time units. The states $\{1, 2, 3, 4\}$ yield a minor reward when $a = 1$ and incur a minor cost when $a = -1$. But by choosing $a = -1$ it becomes more likely to avoid reaching the border state 5. A similar (symmetric) situation holds for the states $\{-4, -3, -2, -1\}$.

Let's exhaustively analyse this example!!! Fun, no?

1. Think of how this may fit in some applied situation, where your controller is trying to regulate a system at a "set-point". Give (in a short a paragraph) a clear real-world example of this. Your description should try to account for the cost and transition structure. Be creative but precise.

2. Postulate further about this process and the associated control decisions. What would optimal policies look like? What role does λ play? What do you think is the optimal policy when $\lambda \approx 0$ (but > 0). What do you think is the optimal policy when $\lambda \approx 1$ (but < 1)? State your reasons clearly.

   The results of 3–7 below, should be presented together in a brief and precise manner – 2 pages maximum.

3. Write a function that finds the optimal policy by brute force enumeration of all (512) policies and solution of the policy evaluation equations, e.g. page 144, (6.1.6).

4. Write a function that finds the optimal policy using value iteration with specified $\epsilon > 0$ and a stopping criterion adapted to λ as in page 161, (6.3.3).

5. Write a function that finds the optimal policy using the policy iteration algorithm, e.g. page 174.

6. (Optional). Write a function that finds the optimal policy using linear programming.

7. Use 3–6 to find the optimal policy for a range of $\lambda \in (0, 1)$ in steps of 0.01 (i.e. run the algorithm 99 times). For the value iteration use a "small enough" epsilon. There should not be any discrepancies, if there are, investigate them and comment on them. Present the optimal policies (for each λ) in the neatest manner that you can think of. Briefly comment on the results.

8. (Optional – but a good enough result will count as credit towards future HW). Assume now that the problem is in a more general form, where the values, 5, 3, 1/2, 1/4 are represented by arbitrary parameters (at least some of them). Do you believe that the optimal policy has some threshold structure? If not, explain why. If so, try to prove this using value-iteration and induction. Suggestion: Keep 3, 1/2 and 1/4 as numbers. Only parameterise 5 by L, so the state space is $\{-L, \ldots, L\}$. Prove your result for this class of problems (indexed by L).

Good Luck.