Solutions for HW3

Yunhai Han
Department of Mechanical and Aerospace Engineering
University of California, San Diego
y8han@eng.ucsd.edu

May 6, 2020

Contents

1 Problem 1

1.1 Problem formulation

Consider an MDP with state space $X = \{a, b, c\}$ and control action space $U = \{1, 2, u_T\}$ where u_T is a terminal action that you can only apply at the goal state $X_G = \{c\}$, so that $p_{cc}(u_T) = 1$ and $p_{xy}(u_T) = 0$ for all other $y, x \neq c$. Other state transition probabilities are given in Figure [1] In addition, suppose that the stage cost function is $\ell(x, u, x') = 1$ for all x, x' and $u \neq u_T$ and that $\ell(c, u_T) = 0$.

Pick the initial policy as follows $\pi(a) = 1, \pi(b) = 1$ and $\pi(c) = u_T$, and implement the policy iteration algorithm "by hand".

1.2 solution

From the figure, we can write down the transition probability for each control action.

$$p_{aa}(u=1) = \frac{1}{3}, p_{ab}(u=1) = \frac{1}{3}, p_{ac}(u=1) = \frac{1}{3}; p_{ba}(u=1) = \frac{1}{3}, p_{bb}(u=1) = \frac{1}{3}, p_{bc}(u=1) = \frac{1}{3}$$

$$p_{aa}(u=2) = 0, p_{ab}(u=2) = \frac{1}{2}, p_{ac}(u=2) = \frac{1}{2}; p_{ba}(u=2) = \frac{1}{4}, p_{bb}(u=2) = 0, p_{bc}(u=2) = \frac{3}{4}$$

Also, from the title, we are given that:$p_{cc}(u_T) = 1$ and $p_{xy}(u_T) = 0$ for all other $y, x \neq c$. These two transition probabilities make the whole graph complete.

Besides, from the title, the stage cost function is $\ell(x, u, x') = 1$ for all x, x' and $u \neq u_T$ and that $\ell(c, u_T) = 0$.

Using the policy iteration equations:

Let π by a policy, a policy evaluation step calculates the solution to:

$$J(i) = \sum_j p_{ij}(\pi(i))(\ell(i, \pi(i), j) + \gamma J(j)), \quad i = 1, 2, \ldots, n$$

Denote by $J^\pi(i)$ its solution. A policy improvement step updates

$$\pi'(i) = \text{argmin}_{u \in U} \sum_j p_{ij}(u)\left(\ell(i, u, j) + \gamma J^\pi(j)\right)$$

or, also, $\boldsymbol{T}_{\pi'} \boldsymbol{J}^\pi = \boldsymbol{T} J^\pi$

We set γ = 1. Besides, we know that if we repeat these steps many times, the policy would converge to the optimal one:

$$J^\pi = T_\pi J^\pi$$

Iteration 1

$$J^{\pi_0}(a) = \sum_j p_{aj}(\pi_0(a))(1 + \gamma J^{\pi_0}(j))$$

$$J^{\pi_0}(b) = \sum_j p_{bj}(\pi_0(b))(1 + \gamma J^{\pi_0}(j))$$

$$J^{\pi_0}(c) = \sum_j p_{cj}(\pi_0(c))(1 + \gamma J^{\pi_0}(j))$$

Solve these equations, we could choose any value for $J^{\pi_0}(c)$. Without any confusion, we could always set $J^{\pi_0}(c) = 0$ and obtain the other two uniquely:$J^{\pi_0}(a) = 3,^{\pi_0}(b) = 3$. Then, solve for $\pi^1(a), \pi^1(b), \pi^1(c)$.

$$\pi^1(a) = \arg\max_u(TJ^{\pi_0}) = \arg\max(3(u=1), 2(u=2)) = 2$$
$$\pi^1(b) = \arg\max_u(TJ^{\pi_0}) = \arg\max(3(u=1), \frac{7}{4}(u=2)) = 2$$
$$\pi^1(c) = u_T(\text{always})$$

Iteration 2

$$J^{\pi_1}(a) = \sum_j p_{aj}(\pi_1(a))(1+\gamma J^{\pi_1}(j))$$
$$J^{\pi_1}(b) = \sum_j p_{bj}(\pi_1(b))(1+\gamma J^{\pi_1}(j))$$
$$J^{\pi_1}(c) = \sum_j p_{cj}(\pi_1(c))(1+\gamma J^{\pi_1}(j))$$

Using the same method, the solutions are:$J^{\pi_1}(c) = 0, J^{\pi_1}(a) = \frac{12}{7}, J^{\pi_1}(b) = \frac{10}{7}$. Then, solve for $\pi^2(a), \pi^2(b), \pi^2(c)$.

$$\pi^2(a) = \arg\max_u(TJ^{\pi_1}) = \arg\max(\frac{43}{21}(u=1), \frac{12}{7}(u=2)) = 2$$
$$\pi^2(b) = \arg\max_u(TJ^{\pi_1}) = \arg\max(\frac{43}{21}(u=1), \frac{10}{7}(u=2)) = 2$$
$$\pi^2(c) = u_T(\text{always})$$

Hence, you could see the optimal policy converges to $\pi(a) = 2, \pi(b) = 2, \pi(c) = u_T$(Because in this iteration, you could see $J^{\pi_2} = T_\pi J^{\pi_1}$).

2 Problem 2

2.1 Problem formulation

In the previous problem, implement Q-value iteration by hand by calculating the Q-factors. What differences do you find between Q-value iteration and policy iteration?

2.2 solution

Here are the procedures of this algorithm:

Initialize $Q_0(i,u) = 0$ for each i.

for k in $\{0, \dots, N-1\}$:

for all states i in X, and all u in U :

$$Q_{k+1}(i,u) = \sum_j p_{ij}(u)\left(\ell(i,u,j) + \alpha \min_v Q(j,v)\right)$$

Can be repeated until convergence

$$|Q_{k+1}(i,u) - Q_k(i,u)| \le \varepsilon$$

We set $\alpha = 1$.

First Initialize $Q_0(i, u) = 0$ for each i.

iteration 1

$$\begin{aligned} Q_1(a, u = 1) = 1; Q_1(a, u = 2) = 1 \\ Q_1(b, u = 1) = 1; Q_1(b, u = 2) = 1 \\ Q_1(c, u = u_T) = 0 \end{aligned}$$

iteration 2

$$\begin{aligned} Q_2(a, u = 1) = \frac{5}{3}; Q_2(a, u = 2) = \frac{3}{2} \\ Q_2(b, u = 1) = \frac{5}{3}; Q_2(b, u = 2) = \frac{5}{4} \\ Q_2(c, u = u_T) = 0 \end{aligned}$$

iteration 3

$$\begin{aligned} Q_3(a, u = 1) = 1.9167; Q_3(a, u = 2) = 1.6250 \\ Q_3(b, u = 1) = 1.9167; Q_3(b, u = 2) = 1.3750 \\ Q_3(c, u = u_T) = 0 \end{aligned}$$

iteration 4

$$\begin{aligned} Q_4(a, u = 1) = 2; Q_4(a, u = 2) = 1.6875 \\ Q_4(b, u = 1) = 2; Q_4(b, u = 2) = 1.4063 \\ Q_4(c, u = u_T) = 0 \end{aligned}$$

If we choose $\varepsilon = 0.1$, the iteration terminates.

iteration 5

$$\begin{aligned} Q_5(a, u = 1) = 2.0313; Q_5(a, u = 2) = 1.7031 \\ Q_5(b, u = 1) = 2.0313; Q_5(b, u = 2) = 1.4219 \\ Q_5(c, u = u_T) = 0 \end{aligned}$$

If we choose $\varepsilon = 0.05$, the iteration terminates.

The subsequent computations are omitted and you can imagine even with smaller ε, the difference is less than it.

Finally, we recover $J_N(i), \pi_N(i)$ for each i from these Q-values:

$$\begin{aligned} \text{compute } J_N(j) = \min_v Q_N(j, v) \text{ for all } j \\ \text{compute } \pi_N(i) = \operatorname{argmin}_v Q_N(j, v) \text{ for all } j \end{aligned} \tag{1}$$

$$\begin{aligned} J_5(a) = 1.7031; \pi_5(a) = 2 \\ J_5(b) = 1.4219; \pi_5(b) = 2 \\ J_5(c) = 0; \pi_5(c) = u_T \end{aligned}$$

The difference between Q-value iteration and policy iteration is:

1 Policy iteration would update policy during each iteration until it get converged. On the contrary, in Q-value iteration, the optimal policy would be obtained directly after the final iteration.

2 The computational complexity for the two algorithms are different.

3 Problem 3

3.1 Problem formulation

Formulate a linear program to find the optimal cost to go of the previous problem. The optimal solution to the problem satisfies the Bellman equation, solve for J^* directly in the equation. Extra credit (3 points): Show how to find the J^* value from the KKT optimality conditions of a linear program.

3.2 solution

LP formulation to find J^* : Let η_0 be any probability vector. Then, J^* solves

$$\max_J \eta_0^\top J$$

$$J(i) \leq \sum_j p_{ij}(u)(\ell(i,u,j) + \gamma J(j))$$

for all i, u this is equivalent to

$$\begin{array}{l} \max_J \eta_0^\top J \\ J \leq TJ \end{array}$$

This is also equivalent to

$$\begin{array}{l} \min_J \eta_0^\top J \\ J \geq TJ \end{array}$$

Without any confusion, we could simply set $\eta_0^T = (1, \cdots, 1)$(the number of ones is equal to the number of states).

In this problem, the linear programming problem is:

$$\begin{aligned} \min_J \eta_0^\top J = \min_J J_1 + J_2 + J_3 & \\ J_1 \geq & \min(\frac{1}{3}(1+J_1) + \frac{1}{3}(1+J_2) + \frac{1}{3}(1+J_3), \frac{1}{2}(1+J_2) + \frac{1}{2}(1+J_3)) \\ J_2 \geq & \min(\frac{1}{3}(1+J_1) + \frac{1}{3}(1+J_2) + \frac{1}{3}(1+J_3), \frac{1}{4}(1+J_1) + \frac{3}{4}(1+J_3)) \\ J_3 \geq & J_3 \\ J_1, J_2, J_3 \geq 0 & \end{aligned}$$

Indeed, in order to solve this optimization problem, we can solve four optimization problems with four different set of constraints.

- $$\begin{aligned} \min_J \eta_0^\top J = \min_J J_1 + J_2 + J_3 & \\ J_1 \geq & \frac{1}{3}(1+J_1) + \frac{1}{3}(1+J_2) + \frac{1}{3}(1+J_3) \\ J_2 \geq & \frac{1}{3}(1+J_1) + \frac{1}{3}(1+J_2) + \frac{1}{3}(1+J_3) \\ J_3 \geq & J_3 \\ J_1, J_2, J_3 \geq 0 & \end{aligned}$$

- $$\begin{aligned}\min_J \eta_0^\top J = \min_J & \, J_1 + J_2 + J_3 \\ J_1 \geq & \frac{1}{3}(1+J_1) + \frac{1}{3}(1+J_2) + \frac{1}{3}(1+J_3) \\ J_2 \geq & \frac{1}{4}(1+J_1) + \frac{3}{4}(1+J_3)) \\ J_3 \geq & J_3 \\ J_1, J_2, J_3 \geq 0 & \end{aligned}$$

- $$\begin{aligned}\min_J \eta_0^\top J = \min_J & \, J_1 + J_2 + J_3 \\ J_1 \geq & \frac{1}{2}(1+J_2) + \frac{1}{2}(1+J_3) \\ J_2 \geq & \frac{1}{3}(1+J_1) + \frac{1}{3}(1+J_2) + \frac{1}{3}(1+J_3) \\ J_3 \geq & J_3 \\ J_1, J_2, J_3 \geq 0 & \end{aligned}$$

- $$\begin{aligned}\min_J \eta_0^\top J = \min_J & \, J_1 + J_2 + J_3 \\ J_1 \geq & \frac{1}{2}(1+J_2) + \frac{1}{2}(1+J_3) \\ J_2 \geq & \frac{1}{4}(1+J_1) + \frac{3}{4}(1+J_3) \\ J_3 \geq & J_3 \\ J_1, J_2, J_3 \geq 0 & \end{aligned}$$

We could compare the different results and pick up the minimum one.

Using *Matlab* to solve each linear programming problem. The final solution is:$J_1 = 1.7143, J_2 = 1.4286, J_3 = 0$(the forth one above).

The next question is to show how to find the J^* value from the KKT optimality conditions of a linear programming.

Since the objective function is convex(no Hessian matrix) and all the constraints are linear, it is sufficient to find a KKT point which must be optimal. Thus, we could write the KKT conditions according to following optimization problem.

$$\begin{aligned}\min_J \eta_0^\top J = \min_J & \, J_1 + J_2 + J_3 \\ J_1 \geq & \frac{1}{2}(1+J_2) + \frac{1}{2}(1+J_3) \\ J_2 \geq & \frac{1}{4}(1+J_1) + \frac{3}{4}(1+J_3) \\ J_3 \geq & J_3 \\ J_1, J_2, J_3 \geq 0 & \end{aligned}$$

or

$$\begin{aligned}\max_J \eta_0^\top J = \max_J & \, J_1 + J_2 + J_3 \\ J_1 \leq & \frac{1}{2}(1+J_2) + \frac{1}{2}(1+J_3) \\ J_2 \leq & \frac{1}{4}(1+J_1) + \frac{3}{4}(1+J_3) \\ -J_1, -J_2, -J_3 \leq 0 & \end{aligned}$$

To write the KKT conditions, observe the following:

$$\nabla z = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad \nabla g_1 = \begin{bmatrix} 1 \\ -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix} \quad \nabla g_2 = \begin{bmatrix} \frac{1}{4} \\ 1 \\ -\frac{3}{4} \end{bmatrix} \quad \nabla g_3 = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} \quad \nabla g_4 = \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} \quad \nabla g_5 = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}$$

We can now write the KKT conditions for this problem as: Primal Feasibility:

$$\begin{cases} J_1 - \frac{1}{2}J_2 - \frac{1}{2}J_3 \leq 1 \\ -\frac{1}{4}J_1 + J_2 - \frac{3}{4} \leq 1 \\ J_1, J_2, J_3 \geq 0 \end{cases}$$

Dual Feasibility:

$$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \lambda_1 \begin{bmatrix} 1 \\ -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix} - \lambda_2 \begin{bmatrix} \frac{1}{4} \\ 1 \\ -\frac{3}{4} \end{bmatrix} - \lambda_3 \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} - \lambda_4 \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} - \lambda_5 \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5 \geq 0$$

$$\text{Complementary Slackness: } \begin{cases} \lambda_1 \left(J_1 - \dfrac{1}{2}J_2 - \dfrac{1}{2}J_3 - 1 \right) = 0 \\ \lambda_2 \left(-\dfrac{1}{4}J_1 + J_2 - \dfrac{3}{4}J_3 - 1 \right) = 0 \\ \lambda_3 \left(-J_1 \right) = 0 \\ \lambda_4 \left(-J_2 \right) = 0 \\ \lambda_5 \left(-J_3 \right) = 0 \end{cases}$$

Consider Dual Feasibility for a moment. I can expand the matrices to obtain a system of equations:

$$\begin{aligned} 1 - \lambda_1 - \frac{1}{4}\lambda_2 + \lambda_3 &= 0 \\ 1 + \frac{1}{2}\lambda_1 - \lambda_2 + \lambda_4 &= 0 \\ 1 + \frac{1}{2}\lambda_1 + \frac{3}{4}\lambda_2 + \lambda_5 &= 0 \\ \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5 &\geq 0 \end{aligned}$$

or:

$$\begin{aligned} \lambda_1 + \frac{1}{4}\lambda_2 - \lambda_3 &= 1 \\ -\frac{1}{2}\lambda_1 + \lambda_2 - \lambda_4 &= 1 \\ -\frac{1}{2}\lambda_1 - \frac{3}{4}\lambda_2 - \lambda_5 &= 1 \\ \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5 &\geq 0 \end{aligned}$$

Indeed, from the problem formulation, we could conclude $J_3 = 0$. Hence, λ_5 can be safely removed and the new KKT constrains are:

$$\begin{aligned} \lambda_1 + \frac{1}{4}\lambda_2 - \lambda_3 &= 1 \\ -\frac{1}{2}\lambda_1 + \lambda_2 - \lambda_4 &= 1 \\ \lambda_1, \lambda_2, \lambda_3, \lambda_4 &\geq 0 \end{aligned}$$

since $\lambda_3, \lambda_4 \geq 0$, they act like surplus variables and we can write the Dual Feasibility as:

$$\begin{cases} \lambda_1 + \frac{1}{4}\lambda_2 \geq 1 \\ -\frac{1}{2}\lambda_1 + \lambda_2 \geq 1 \\ \lambda_1, \lambda_2 \geq 0 \end{cases}$$

It would now suffice to find values for $J_1, J_2 (J_3 = 0), \lambda_1, \lambda_2, \lambda_3,$ and λ_4 the satisfy the KKT conditions and we could solve the linear programming problem. The solutions are:

$$J_1 = 1.7143 \quad J_2 = 1.4286 \quad J_3 = 0 \quad \lambda_1 = 0.6667 \quad \lambda_2 = 1.3333 \quad \lambda_3 = \lambda_4 = 0$$