ISyE 4232 Spring 2014

H. Ayhan

Solutions to Homework 7

1. For simplicity we will use the shorthand $s_1 = 1$, $s_2 = 2$, $s_3 = 3$, $s_4 = 4$. We have a multi-chain model with a total of 4 possible policies. Let's look at each one in detail.

   (a) Set $d^1(1) = a_{11}, d^1(2) = a_{21}, d^1(3) = a_{31}, d^1(4) = a_{41}$. Here states 1 and 3 are transient, and states 2 and 4 are absorbent (each forms a recurrent class). If we start at state 1 or 3 we are absorbed into state 2. Therefore
   $$g_{d^1}(1) = g_{d^1}(2) = g_{d^1}(3) = g_{d^1}(4) = 1$$

   (b) Set $d^2(1) = a_{12}, d^2(2) = a_{21}, d^2(3) = a_{31}, d^2(4) = a_{41}$. Here states 1 and 3 are transient, and states 2 and 4 are absorbent (each forms a recurrent class). If we start at state 1 or 3 we are absorbed into state 4. Therefore
   $$g_{d^2}(1) = g_{d^2}(2) = g_{d^2}(3) = g_{d^2}(4) = 1$$

   (c) Set $d^3(1) = a_{11}, d^3(2) = a_{22}, d^3(3) = a_{31}, d^3(4) = a_{41}$. Here states 1, 2 and 3 for a recurrent class with stationary distribution $\pi = (1/3, 1/3, 1/3)$, and state 4 is another recurrent class. Therefore
   $$g_{d^3}(1) = g_{d^3}(2) = g_{d^3}(3) = \frac{1}{3}(4+2+1) = \frac{7}{3} \quad \text{and} \quad g_{d^3}(4) = 1$$

   (d) Set $d^4(1) = a_{12}, d^4(2) = a_{22}, d^4(3) = a_{31}, d^4(4) = a_{41}$. Here states 1, 2 and 3 are transient, and state 4 is absorbent. If we start at state 1, 2 or 3 we are absorbed into state 4. Therefore
   $$g_{d^4}(1) = g_{d^4}(2) = g_{d^4}(3) = g_{d^4}(4) = 1$$

   As we can see $d^3(s) \geq d^k(s)$ for all $s \in S$ and $k = 1, 2, 4$. Therefore $d^* = d^3$ is the optimal policy.

2. Define the state space $\mathcal{S} = \{1, 2\}$, where 1 = *low* and 2 = *high*, NOTE: This is opposite from Homework 5. Define the action space $\mathcal{A} = \{1, 2\}$, where 1 = *Do Nothing*, 2 = *Advertise*. First calculate the expected immediate rewards
   $$\begin{aligned} r(1,1) &= 7 \times 0.2 - 2 \times 0.8 = -0.2 \\ r(2,1) &= 10 \times 0.5 + 4 \times 0.5 = 7 \\ r(1,2) &= 3 \times 0.4 - 5 \times 0.6 = -1.8 \\ r(2,2) &= 7 \times 0.8 + 6 \times 0.2 = 6.8 \end{aligned}$$

   Let us begin with the policy iteration. We will start the algorithm at some arbitrary policy, say $d_0(1) = 1, d_0(2) = 1$. Now for the policy evaluation step we need to solve the following linear system,
   $$\begin{aligned} 7 &= g^0 - 0.5h^0(1) + 0.5h^0(2) \\ -0.2 &= g^0 + 0.2h^0(1) - 0.2h^0(2) \end{aligned}$$

We can set $h^0(1) = 0$, then solving this system we get $g^0 = 1.857$ and $h^0(2) = 10.286$. Now for policy improvement:

$$\begin{aligned}
d_1(1) &= \arg\max_{a \in A_1}\{-0.2 + (0.2 * 10.286 + 0.8 * 0)\ ,\ -1.8 + (0.4 * 10.286 + 0.6 * 0)\} \\
&= \arg\max_{a \in A_1}\{1.85\ ,\ 2.31\} \\
&= 2 \\
d_1(2) &= \arg\max_{a \in A_2}\{7 + (0.5 * 10.286 + 0.5 * 0)\ ,\ 6.8 + (0.8 * 10.286 + 0.2 * 0)\} \\
&= \arg\max_{a \in A_2}\{12.14\ ,\ 15.28\} \\
&= 2
\end{aligned}$$

So we have $d_1(1) = 2, d_1(2) = 2$, therefore we keep going. Solve:

$$\begin{aligned}
6 &= g^1 - 0.2h^1(1) + 0.2h^1(2) \\
-1.8 &= g^1 + 0.4h^1(1) - 0.4h^1(2)
\end{aligned}$$

We can set $h^1(1) = 0$, then solving this system we get $g^1 = 3.93$ and $h^1(2) = 14.33$. Now for policy improvement:

$$\begin{aligned}
d_2(1) &= \arg\max_{a \in A_1}\{-0.2 + (0.2 * 14.33 + 0.8 * 0)\ ,\ -1.8 + (0.4 * 14.33 + 0.6 * 0)\} \\
&= \arg\max_{a \in A_1}\{2.66\ ,\ 3.93\} \\
&= 2 \\
d_2(2) &= \arg\max_{a \in A_2}\{7 + (0.5 * 14.33 + 0.5 * 0)\ ,\ 6.8 + (0.8 * 14.33 + 0.2 * 0)\} \\
&= \arg\max_{a \in A_2}\{14.166\ ,\ 18.28\} \\
&= 2
\end{aligned}$$

So we have $d_1(1) = d_2(1) = 2, d_1(2) = d_2(2) = 2$, therefore the optimal policy is to always advertise.

Now for the LP method get that the LP primal is given by:

$$\text{Minimize} \quad g$$

subject to

$$\begin{aligned}
g + h(1) - (0.8h(1) + 0.2h(2)) &\geq -0.2 \\
g + h(1) - (0.6h(1) + 0.4h(2)) &\geq -1.8 \\
g + h(2) - (0.5h(1) + 0.5h(2)) &\geq 7 \\
g + h(2) - (0.2h(1) + 0.8h(2)) &\geq 6.8
\end{aligned}$$

And the LP dual is

$$\text{Maximize} \quad -0.2 * x(1,1) - 1.8 * x(1,2) + 7 * x(2,1) + 6.8 * x(2,2)$$

subject to

$$\begin{aligned}
x(1,1) + x(1,2) - (.2 * x(1,1) + .4 * x(1,2) + .5 * x(2,1) + .8 * x(2,2)) &= 0 \\
x(2,1) + x(2,2) - (.8 * x(1,1) + .6 * x(1,2) + .5 * x(2,1) + .2 * x(2,2)) &= 0 \\
x(1,1) + x(1,2) + x(2,1) + x(2,2) &= 1 \\
x(s,a) &\geq 0 \quad \forall s \in S, a \in A
\end{aligned}$$

We code the LP dual into Xpress, and the results are:

```
Begin running model
Optimal value =3.933
x(1,1)=0
x(1,2)=0.3333
x(2,1)=0
x(2,2)=0.6666
End running model
```

From this we conclude the optimal policy is $d^*(1) = 2, d^*(2) = 2$. That is, always advertise.