Stochastic Optimisation - Assessed Problem Sheet 2

Dom Hutchinson (170 1111)

December 22, 2020

Answer 1) (a)
Here I formulate the problem described in 1) as a *Finite-Horizon Markov Decision Problem.*

- *Decision Epoch* - Start of each month.
- *Time-Horizon* - $T = \{0, 1, \dots, N-1\}$.
- *Agent Actions.*

  Let Y_t denote the action taken by the agent in epoch t and be defined as the amount of product the company produces in epoch t.
- *Action-Space.*

  By the definition of Y_t and the fact that the company can produce at most m products in a month, we have that the action-space is

  $$A = \{0, \dots, m\}$$
- *System States.*

  Let X_t denote the system state in epoch t and be defined as the index of the unit price which each product was sold at in epoch $t-1$ (ie at the end of the previous month).
- *State -Space.*

  By the definition of X_t and since unit prices are indexed from 1 to n, we have that the state-space is

  $$S = \{1, \dots, n\}$$
- *Transition Probabilities.*

  Let $s, s' \in S, a \in A(s)$ and define transition probabilities as $p_t(s'|s,a) = \mathbb{P}(X_{t+1} = s'|X_t = s, Y_t = a)$. Then

  $$\begin{array}{rcl} p_t(s'|s,a) & = & p(s'|s,a) \\ & = & \pi(s'|s,a) \end{array}$$

  where $\pi(\cdot|\cdot,\cdot)$ is as defined in the question.
- *Immediate Rewards.*

  The immedaite reward R_t recieved in each epoch t depends on the unit price at the end of that epoch, but, due to my definition of system states, this value is stored in X_{t+1}, (not

X_t). This does not fit into the framework of Markovian decision problems so we take the expected reward in each epoch.

$$\begin{array}{rcl}
\mathbb{E}^\pi[R_t] &=& \mathbb{E}^\pi[r(X_t,Y_t)] \\
&=& \begin{cases} \mathbb{E}^\pi\left[Y_t\big(\beta(X_{t+1})-\gamma\big)\right] & \text{if } \beta(X_{t+1})\geq\gamma \\ \mathbb{E}^\pi\left[Y_t\big(\beta(X_{t+1})-\gamma\big)+\alpha Y_t\big(\beta(X_{t+1})-\gamma\big)\right] & \text{if } \beta(X_{t+1})<\gamma \end{cases} \\
&=& \begin{cases} \mathbb{E}^\pi\left[Y_t\beta\big(X_{t+1}-l\big)\right] & \text{if } X_{t+1}\geq l \\ \mathbb{E}^\pi\left[Y_t\beta\big(X_{t+1}-l\big)(1+\alpha)\right] & \text{if } X_{t+1}<l \end{cases} \text{by def. } \beta(\cdot),\gamma \\
&=& \mathbb{E}^\pi\left[Y_t\beta\big(X_{t+1}-l\big)\big(1+\alpha\mathbb{1}\{X_{t+1}<l\}\big)\right] \\
&=& \mathbb{E}^\pi\left[\mathbb{E}^\pi\left[Y_t\beta\big(X_{t+1}-l\big)\big(1+\alpha\mathbb{1}\{X_{t+1}<l\}\big)\right]\middle|X_t,Y_t\right] \text{ by Tower property} \\
&=& Y_t\beta\cdot\big(\mathbb{E}^\pi[X_{t+1}|X_t,Y_t]-l\big)\cdot\big(1+\alpha\mathbb{1}\{\mathbb{E}^\pi[X_{t+1}|X_t,Y_t]<l\}\big)
\end{array}$$

Answer 1) (b)

From the specification of the question, we know the following

$$\begin{array}{rcl}
T &=& \{0,1\} \\
A &=& \{0,1,2\} \\
S &=& \{1,2\} \\
A(s) &=& \{0,1,2\}\ \forall\ s\in S \\
r_t(s,a) &=& a\cdot\big(\mathbb{E}^\pi[X_{t+1}|s,a]-1\big)\cdot\big(1+0.3\cdot\mathbb{1}\{\mathbb{E}^\pi[X_{t+1}|s,a]<1\}\big) \text{ for } t\in T \\
r_2(s) &=& 0\ \forall\ s\in S
\end{array}$$

$\pi(1|s,a) =$

s\ a	0	1	2
1	0.6	0.3	0.1
2	0.2	0.6	0.7

$\pi(2|s,a) =$

s\ a	0	1	2
1	0.4	0.7	0.9
2	0.8	0.4	0.3

From this specification of the transition probabilities $p(s'|s,a)$ we can deduce the following values for the next expected system state

$\mathbb{E}^\pi(X_{t+1}|s,a) =$

s\a	0	1	2
1	1.6	1.3	1.1
2	1.7	1.4	1.3

Since $\mathbb{E}^\pi(X_{t+1}|s,a)>1=l$ for all $s\in S, a\in A(s)$ we can simplify the reward function to the following

$$r_t(s,a) = a(\mathbb{E}[X_{t+1}|s,a]-1) \text{ for } t\in T$$

$\implies r_t(s,a) =$

s\a	0	1	2
1	0	0.3	0.2
2	0	0.4	0.6

To find the optimal Markovian decision policy π^* for this problem, using the dynamic programming algorithm we compute the following terms in a backwards recursion through T

$$\begin{array}{rcl}
w_t^*(s,a) &:=& r_t(s,a)+\sum_{s'\in S}u_{t+1}^*(s')p(s'|s,a) \\
u_t^*(s) &=& \max_{a\in A(s)}(w_t^*) \\
d_t^*(s) &\in& \text{argmax}_{a\in A(s)}(w_t^*)
\end{array}$$

where $u_2^*(s):=r_2(s)=0\ \forall\ s\in S$.

- *Time-Period* $t = 1$.

  In this period

$$\begin{aligned} w_1^*(s,a) &= r_1(s,a) + \sum_{s \in \{1,2\}} u_2^*(s')p(s'|s,a) \\ &= r_1(s,a) \end{aligned}$$

  Thus we can compute the following table of values for $w_1^*(s,a)$

$$w_1^*(s,a) = \begin{array}{c|ccc} \text{s}\backslash\text{a} & 0 & 1 & 2 \\ \hline 1 & 0 & 0.3 & 0.2 \\ 2 & 0 & 0.4 & 0.6 \end{array}$$

  The table below summarises the values of $u_1^*(s), d_1^*(s)$ given this information

s	$u_1^*(s)$	$d_1^*(s)$
1	0.3	1
2	0.6	2

- *Time-Period* $t = 0$.

  In this period

$$\begin{aligned} w_0^*(s,a) &= r_0(s,a) + \sum_{s' \in \{1,2\}} u_1^*(s')p(s'|s,a) \\ &= r_0(s,a) + 0.3 \cdot p(1|s,a) + 0.6 \cdot p(2|s,a) \end{aligned}$$

  Thus we can compute the following table of values for $w_0^*(s,a)$

$$w_0^*(s,a) = \begin{array}{c|ccc} \text{s}\backslash\text{ a} & 0 & 1 & 2 \\ \hline 1 & 0.42 & 0.57 & 0.5 \\ 2 & 0.54 & 0.7 & 0.9 \end{array}$$

  The table below summarises the values of $u_0^*(s), d_0^*(s)$ given this information

s	$u_0^*(s)$	$d_0^*(s)$
1	0.57	1
2	0.9	2

The optimal value function is $u_0^*(s)$ and the optimal Markovian policy

$$\pi^* := (d_1^*(s), d_2^*(s))$$

Answer 2) (a)

$$\begin{aligned} v^*(s) &= \max_{a \in A(s)} \left\{ r(s,a) + \alpha \sum_{s' \in S} v^*(s')p(s'|s,a) \right\} && \forall\ s \in S \\ \implies v^*(s) &\geq r(s,a) + \alpha \sum_{s' \in S} v^*(s')p(s'|s,a) && \forall\ s \in S, a \in A(s) \\ \implies 0 &\leq v^*(s) - r(s,a) + \alpha \sum_{s' \in S} v^*(s')p(s'|s,a) && \forall\ s \in S, a \in A(s) \\ &= u^*(s,a) \\ \implies 0 &\leq u^*(s,a) && \forall\ s \in S, a \in A(s) \end{aligned}$$

Answer 2) (b)
Let $\tilde{s} \in S$ and $\tilde{a} \in (A(\tilde{s}) \setminus D^*(\tilde{s}))$ (ie a is a sub-optimal action).
Suppose there is a non-negative probability that $\tilde{a}$ is chosen by our policy, given the system is in state $\tilde{s}$.

$$q(\tilde{a}|\tilde{s}) > 0$$

Since $\tilde{a} \notin D^*(\tilde{s})$ then

$$\begin{array}{rrcl} & v^*(\tilde{s}) & > & r(\tilde{s},\tilde{a}) + \alpha \sum_{s' \in S} v^*(s')p(s'|\tilde{s},\tilde{a}) \\ \implies & 0 & < & v^*(\tilde{s}) - r(\tilde{s},\tilde{a}) - \alpha \sum_{s' \in S} v^*(s')p(s'|\tilde{s},\tilde{a}) \\ & & = & u^*(\tilde{s},\tilde{a}) \\ \implies & 0 & < & u^*(\tilde{s},\tilde{a}) \end{array}$$

Since we assume $q(\tilde{a}|\tilde{s}) > 0$, we have that

$$u^*(\tilde{s},\tilde{a})q(\tilde{a},\tilde{s}) > 0 \ \forall \ \tilde{a} \in (A(\tilde{s}) \setminus D^*(\tilde{s}))$$

As this holds for all such $\tilde{a}$, their sum is strictly positive

$$\sum_{a \in A(\tilde{s})} u^*(\tilde{s},s)q(a|\tilde{s}) > 0$$

Answer 2) (c)

$$\begin{array}{rl} & \mathbb{E}^\pi \left[v^*(X_t) - \alpha v^*(X_{t+1}) - r(X_t,Y_t)\right] \\ = & \mathbb{E}^\pi \left[v^*(X_t)\right] - \alpha\mathbb{E}^\pi \left[v^*(X_{t+1})\right] - \mathbb{E}^\pi \left[r(X_t,Y_t)\right] \\ = & \mathbb{E}^\pi \left[v^*(X_t)\right] - \alpha\mathbb{E}^\pi \left[\mathbb{E}^\pi \left[v^*(X_{t+1})|X_t,Y_t\right]\right] - \mathbb{E}^\pi \left[r(X_t,Y_t)\right] \text{ by tower property} \\ = & \mathbb{E}^\pi \left[v^*(X_t)\right] - \alpha\mathbb{E}^\pi \left[\sum_{s' \in S} v^*(s')p(s'|X_t,Y_t)\right] - \mathbb{E}^\pi \left[r(X_t,Y_t)\right] \\ = & \mathbb{E}^\pi \left[v^*(X_t) - \alpha \sum_{s' \in S} v^*(s')p(s'|X_t,Y_t) - r(X_t,Y_t)\right] \\ = & \mathbb{E}^\pi \left[u^*(X_t,Y_t)\right] \text{ by definition} \\ = & \mathbb{E}^\pi \left[\mathbb{E}^\pi \left[u^*(X_t,Y_t)|X_t\right]\right] \text{ by tower property} \\ = & \mathbb{E}^\pi \left[\sum_{a \in A(X_t)} u^*(X_t,a)q(a|X_t)\right] \text{ by def. conditional expectation} \\ \geq & 0 \end{array}$$

The inequality is due to $u^*(s,a) \geq 0$ and $q(a|s) \geq 0$ for all $s \in S, a \in A(s)$ by 2) (a) and the definition of probability distributions.

Answer 2) (d)

$$\begin{aligned}
&\sum_{t=0}^{\infty}\alpha^t\mathbb{E}^\pi\left[v^*(X_t)-\alpha v^*(X_{t+1})-r(X_t,Y_t)\right]\\
=&\sum_{t=0}^{\infty}\alpha^t\mathbb{E}^\pi\left[v^*(X_t)-\alpha v^*(X_{t+1})\right]-\sum_{t=0}^{\infty}\alpha^t\mathbb{E}^\pi\left[r(X_t,Y_t)\right]\\
=&\mathbb{E}^\pi\left[\sum_{t=0}^{\infty}\alpha^t\big(v^*(X_t)-\alpha v^*(X_{t+1})\big)\right]-\mathbb{E}^\pi\left[\sum_{t=0}^{\infty}\alpha^t r(X_t,Y_t)\right]\\
=&\mathbb{E}^\pi\left[v^*(X_0)\right]-\mathbb{E}^\pi\left[\sum_{t=0}^{\infty}\alpha^t r(X_t,Y_t)\right]
\end{aligned}$$

The last step is due to all terms, except $v^*(X_0)$, in the summation cancelling.

Answer 2) (e)

Consider this result

$$\begin{aligned}
&=\mathbb{E}^\pi\left[v^*(X_0)-\alpha v^*(X_1)-r(X_0,Y_0)\right]\\
&=\mathbb{E}^\pi\left[\sum_{a\in A(X_0)}u^*(X_0,a)q(a|X_0)\right]\text{ by 2) (c)}\\
&=\mathbb{E}^\pi\left[\sum_{a\in D^*(s)}u^*(X_0,a)q(a|X_0)\right]\text{ since }\pi\text{ is optimal}^{[1]}\\
&=0\text{ as }u^*(s,a)=0\ \forall\ a\in D^*(s)
\end{aligned}$$

Using this result and 2) (c)

$$\begin{aligned}
0&=\mathbb{E}^\pi\left[\sum_{a\in A(X_0)}u^*(X_0,a)q(a|X_0)\right]\\
&=\sum_{s\in S}\left(\sum_{a\in A(s)}u^*(s,a)q(a|s)\right)\mathbb{P}(X_0=s)\quad[1]
\end{aligned}$$

Note that $u^*(s,a)\geq 0$ for all $s\in S,a\in A(s)$ by 2) (a), and $q(a|s)\geq 0,\mathbb{P}(X_0=s)\geq 0$ by the defintion of probability distributions. Thus

$$\left(\sum_{a\in A(s)}u^*(s,a)q(a|s)\right)\mathbb{P}(X_0=s)\geq 0\ \forall\ s\in S$$

As each term of the outer summation in [1] is non-negative and sum to 0, all the terms must be 0.

$$\left(\sum_{a\in A(s)}u^*(s,a)q(a|s)\right)\mathbb{P}(X_0=s)=0\ \forall\ s\in S$$

Since the question assumes that $\mathbb{P}^\pi(X_0=s)>0$ for all $s\in S$, it must be that

$$\sum_{a\in A(s)}u^*(s,a)q(a|s)=0\ \forall\ s\in S$$

This contradicts the result in 2) (b), thus we can conclude that the conditions of 2) (b) are violated. Meaning, I can conclude that

$$q(a|s)=0\ \forall\ s\in S,a\in\big(A(s)\setminus D^*(s)\big)$$

[1]Since π is optimal $q(a|s)=0$ if $a\notin D^*(s)$