

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

$$A = \{0, \dots, m\}$$

- *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

$$S = \{1, \dots, n\}$$

- *Transition Probabilities*.

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

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

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

- *Immediate Rewards* The immediate reward r_t received 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 not encoded into the system state until epoch $t + 1$. This does not fit into the framework of

Markovian decision problems so we take the expected reward in each epoch.

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

Answer 1) (b)

From the specification of the question, we know the following

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

$$\begin{aligned}
 \pi(1|s, a) &= \begin{array}{c|ccc} s \backslash a & 0 & 1 & 2 \\ \hline 1 & 0.6 & 0.3 & 0.1 \\ 2 & 0.2 & 0.6 & 0.7 \end{array} \\
 \pi(2|s, a) &= \begin{array}{c|ccc} s \backslash a & 0 & 1 & 2 \\ \hline 1 & 0.4 & 0.7 & 0.9 \\ 2 & 0.8 & 0.4 & 0.3 \end{array}
 \end{aligned}$$

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

$$\mathbb{E}^\pi(X_{t+1}|s, a) = \begin{array}{c|ccc} s \backslash a & 0 & 1 & 2 \\ \hline 1 & 1.6 & 1.3 & 1.1 \\ 2 & 1.7 & 1.4 & 1.3 \end{array}$$

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

$$\begin{aligned}
 r_t(s, a) &= a(\mathbb{E}[X_{t+1}|s, a] - 1) \text{ for } t \in T \\
 \implies r_t(s, a) &= \begin{array}{c|ccc} s \backslash a & 0 & 1 & 2 \\ \hline 1 & 0 & 0.3 & 0.2 \\ 2 & 0 & 0.4 & 0.6 \end{array}
 \end{aligned}$$

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

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

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 $w_1^*(s, a)$

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

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

$$\begin{array}{c|cc} s & u_1^*(s) & d_1^*(s) \\ \hline 1 & 0.3 & 1 \\ 2 & 0.6 & 2 \end{array}$$

- *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 $w_0^*(s, a)$

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

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

$$\begin{array}{c|cc} s & u_0^*(s) & d_0^*(s) \\ \hline 1 & 0.57 & 1 \\ 2 & 0.9 & 2 \end{array}$$

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

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

Answer 2) (a)

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

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} chosen to be played

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

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

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

Due to our assumption about $q(\tilde{a}|\tilde{s})$, we have that

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

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

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

Answer 2) (c)

$$\begin{aligned} &\mathbb{E}^\pi [v^*(X_t) - \alpha v^*(X_{t+1}) - r(X_t, Y_t)] \\ = &\mathbb{E}^\pi [v^*(X_t)] - \alpha \mathbb{E}^\pi [v^*(X_{t+1})] - \mathbb{E}^\pi [r(X_t, Y_t)] \\ = &\mathbb{E}^\pi [v^*(X_t)] - \alpha \mathbb{E}^\pi [\mathbb{E}^\pi [v^*(X_{t+1})|X_t, Y_t]] - \mathbb{E}^\pi [r(X_t, Y_t)] \\ = &\mathbb{E}^\pi [v^*(X_t)] - \alpha \mathbb{E}^\pi \left[\sum_{s' \in S} v^*(s') p(s'|X_t, Y_t) \right] - \mathbb{E}^\pi [r(X_t, Y_t)] \\ = &\mathbb{E}^\pi \left[v^*(X_t) - \alpha \sum_{s' \in S} v^*(s') p(s'|X_t, Y_t) - r(X_t, Y_t) \right] \\ = &\mathbb{E}^\pi [u^*(X_t, Y_t)] \text{ by definition} \\ = &\mathbb{E}^\pi [\mathbb{E}^\pi [u^*(X_t, Y_t)|X_t]] \text{ by tower property} \\ = &\mathbb{E}^\pi \left[\sum_{a \in A(X_t)} u^*(X_t, a) q(a|X_t) \right] \text{ by def. conditional expectation} \\ \geq &0 \text{ due to 2) (b)} \end{aligned}$$

Answer 2) (d)

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

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

Answer 2) (e)

Consider this result

$$\begin{aligned}
 &= \mathbb{E}^\pi [v^*(X_0) - \alpha v^*(X_1) - r(X_0, Y_0)] \\
 &= \mathbb{E}^\pi \left[\max_{Y_0 \in A(X_0)} \left(r(X_0, Y_0) + \alpha \sum_{s' \in S} v^*(s') p(s'|X_0, Y_0) \right) - \alpha v^*(X_1) - r(X_0, Y_0) \right] \\
 &= \max_{Y_0 \in A(X_0)} \left(r(X_0, Y_0) + \alpha \sum_{s' \in S} v^*(s') p(s'|X_0, Y_0) \right) - \mathbb{E}^\pi [r(X_0, Y_0) + \alpha v^*(X_1)] \\
 &= \max_{Y_0 \in A(X_0)} \left(r(X_0, Y_0) + \alpha \sum_{s' \in S} v^*(s') p(s'|X_0, Y_0) \right) - \max_{Y_0 \in A(X_0)} (r(X_0, Y_0) + \alpha \mathbb{E}^\pi [v^*(X_1)]) \text{ since } \pi \text{ is optimal} \\
 &= \max_{Y_0 \in A(X_0)} \left(r(X_0, Y_0) + \alpha \sum_{s' \in S} v^*(s') p(s'|X_0, Y_0) \right) - \max_{Y_0 \in A(X_0)} \left(r(X_0, Y_0) + \alpha \sum_{s' \in S} v^*(s') p(s'|X_0, Y_0) \right) \\
 &= 0
 \end{aligned}$$

Using this result and 2) (c)

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

By 2) (a), $u^*(s, a) \geq 0 \forall s \in S, a \in A(s)$, and $q(a|s) \mathbb{P}(X_0 = s) \geq 0$ by the definition of probability distributions.

Thus all terms in the outer summation are non-negative. Further, as they sum to 0, all terms in the outer summation are 0.

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

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

This means that the decision probability distribution $q(a|s)$ of our policy π does not satisfy 2) (b).

Thus, under our policy there does not exist any $s \in S, a \in (A(s) \setminus D^*(s))$ where $q(a|s) > 0$.

Meaning, I can conclude that

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

□