Stochastic Optimisation - Problem Sheet 4

Dom Hutchinson

November 22, 2020

Question 1)
Consider a Markov Decision Process with finite state-space S, finite action-space A, time-horizon T and transition probabilities $\{p_t(s'|s,a)\}$. Let $A(s)$ be the set of actions available at state s.

Assume a *History Dependent Randomised* policy $\pi \in HR(T)$ is used with decision probability $q_t(a|s_{0:t}, a_{0:t-1})$ is applied at epoch $t \in T$.

For $t \geq 1$, $s_{0,t} := (s_0, \dots, s_t) \in S^{t+1}$, $a_{0:t} := (a_0, \dots, a_t) \in A^{t+1}$ compute the following in terms of the marginal distribution of X_0, the transition probabilities $p_t(s'|s,a)$ and the decision probabilities $q_t(a|s_{0:t}, a_{0:t-1})$.

i). $\mathbb{P}^\pi(X_{0:t} = s_{0,t}, Y_{0,t} = a_{0:t})$.

ii). $\mathbb{P}^\pi(X_{0:t} = s_{0,t})$.

iii). $\mathbb{P}^\pi(Y_{0,t} = a_{0:t})$.

Answer 1) i)

$$\begin{array}{rl}
& \mathbb{P}^\pi(X_{0:t} = s_{0:t}, Y_{0:t} = a_{0:t}) \\
= & \mathbb{P}(X_0 = s_0) \displaystyle\prod_{k=1}^{t} \frac{\mathbb{P}^\pi(X_{0:k+1} = s_{0:k+1}, Y_{0:k} = a_{0:k})}{\mathbb{P}^\pi(X_{0:k} = s_{0:k}, Y_{0:k} = a_{0:k})} \cdot \frac{\mathbb{P}^\pi(X_{0:k+1} = s_{0:k}, Y_{0:k} = a_{0:k})}{\mathbb{P}^\pi(X_{0:k} = s_{0:k}, Y_{0:k-1} = a_{0:k-1})} \\
= & p_{X_0}(s_0) \displaystyle\prod_{k=1}^{t} \mathbb{P}^\pi(X_{k+1} = s_{k+1}|X_{0:k} = s_{0:k}, Y_{0:k} = a_{0:k}) \cdot \mathbb{P}^\pi(Y_k = a_k|X_{0:k} = s_{0:k}, Y_{0:k-1} = a_{0:k-1}) \\
= & p_{X_0}(s_0) \displaystyle\prod_{k=1}^{t} \mathbb{P}^\pi(X_{k+1} = s_{k+1}|X_k = s_k, Y_k = a_k) \cdot \mathbb{P}^\pi(Y = a_k|X_k = s_k, Y_{k-1} = a_{k-1}) \text{ by Markov Property} \\
= & p_{X_0}(s_0) \displaystyle\prod_{k=1}^{t} p_k(s_{k+1}|s_k, a_k) q_k(a_k|s_{0:k}, a_{0:k-1})
\end{array}$$

Answer 1) ii)

$$
\begin{array}{rl}
& \mathbb{P}^\pi(X_{0:t}=s_{0:k}) \\
= & \mathbb{P}(X_0=s_0)\prod_{k=1}^t \mathbb{P}^\pi(X_k=s_k|X_{0:k-1}=s_{0:k-1}) \text{ by Bayes Rule} \\
= & p_{X_0}(s_0)\prod_{k=1}^t \mathbb{P}^\pi(X_k=s_k|X_{k-1}=s_{k-1}) \text{ by Markov Property} \\
= & p_{X_0}(s_0)\prod_{k=1}^t \left(\sum_{a\in A(s_k)} \mathbb{P}^\pi(X_k=s_k,Y_{k-1}=a|X_{k-1}=s_{k-1})\right) \text{ by Marginalisation} \\
= & p_{X_0}(s_0)\prod_{k=1}^t \left(\sum_{a\in A(s_k)} \mathbb{P}^\pi(X_k=s_k|Y_{k-1}=a,X_{k-1}=s_{k-1})\mathbb{P}(Y_{k-1}a|X_{k-1}=s_{k-1})\right) \text{ by Bayes Rule} \\
= & p_{X_0}(s_0)\prod_{k=1}^t \left(\sum_{a\in A(s_k)} p_k(s_k|s_{k-1},a)q_k(a|s_{k-1})\right)
\end{array}
$$

Answer 1) iii)

$$
\begin{array}{rl}
& \mathbb{P}^\pi(Y_{0:t}=a_{0:t}) \\
= & \sum_{s_{0:t}\in S^{t+1}} \mathbb{P}^\pi(X_{0:t}=s_{0:t},Y_{0:t}=a_{0:t}) \text{ by Marginalisation} \\
= & \sum_{s_{0:t}\in S^{t+1}} p_{X_0}(s_0)\prod_{k=1}^t p_k(s_{k+1}|s_k,a_k)q_k(a_k|s_{0:k},a_{0:k-1}) \text{ by } \texttt{1) i)}
\end{array}
$$

Question 3)
Consider a Markov Decision Process with finite state-space S, finite action-space $A:=\{a(1),\dots,a(M)\}$, time-horizon T and transition probabilities $\{p_t(s'|s,a)\}$. Let $A(s)$ be the set of actions available at state s.

The agent wants to apply a *History Dependent Randomised* decision rule with decision probability $q_t(a|s_{0:t},a_{0:t-1})$ at epoch $t\in T\setminus\{0\}$. To do so, the agent realised on the following Monte-Carlo procedure:

i). Sample a random number $U_t\sim\text{Uniform}[0,1]$ which is independent of $X_{0:t},Y_{0:t-1}$.

ii). Set $Y_t=\Phi_t(U_t|X_{0:t},Y_{0:t-1})$ where for $u\in\mathbb{R}$ $\Phi(\cdot|s_{0,t},a_{0:t-1})$ is defined as

$$
\Phi(u|s_{0,t},a_{0:t-1})=\begin{cases} a(1) & \text{if } u\leq q_t(a_1|s_{0:t},a_{0:t-1}) \\ a(k) & \text{if } k\in(1,M) \\ & \text{and } u\in\left(\sum_{j=1}^{k-1}q_t(a(j)|s_{0:t},a_{0:t-1}),\sum_{j=1}^{k-1}q_t(a(k)|s_{0:t},a_{0:t-1})\right] \\ a(M) & \text{otherwise}\end{cases}
$$

Show that using this process, the agent indeed implements the decision rule with decision function $q_t(a|s_{0:t},a_{0:t-1})$. Moreover, show that

$$
\mathbb{P}(Y_t=a|X_{0:t}=s_{0:t},Y_{0:t-1}=a_{0:t-1})=q_t(a|s_{0:t},a_{0:t-1})
$$

for all $a\in A,\ s_{0:t}\in S^{t+1}$ and $a_{0:t-1}\in A^t$.

Answer 3)
For ease of notation let $p_{Y_t}(a) := \mathbb{P}(Y_t = a | X_{0:t} = s_{0:t}, Y_{0:t-1} = a_{0:t-1})$.
In order to show that when the agent uses the procedure defined in the question they do indeed implement the decision rule with decision function $q_t(a|s_{0:t}, a_{0:t-1})$, I shall show that

$$p_{Y_t}(a) = q_t(a|s_{0:t}, a_{0:t-1})$$

By the definition of Y_t, we have

$$p_{Y_t}(a) = \mathbb{P}(\Phi_t(U_t|s_{0:t}, a_{0:t-1}) = a|X_{0:t} = s_{0:t}, Y_{0:t-1} = a_{0:t-1})$$

By the definition of $\Phi_t(\cdot)$, we have three separate cases

$$p_{Y_t}(a) = \begin{cases} \mathbb{P}(U_t \leq q_t(a(1)|s_{0:t}, a_{0:t-1})) & \text{if } a = a(1) \\ \mathbb{P}\left(U_t \in \left(\sum_{j=1}^{k-1} q_t(a(j)|s_{0:t}, a_{0:t-1}), \sum_{j=1}^{k} q_t(a(j)|s_{0:t}, a_{0:t-1})\right]\right) & \text{if } a = a(k) \text{ with } k \in [2, M-1] \\ \mathbb{P}\left(U_t > \sum_{j=1}^{M-1} q_t(a(j)|s_{0:t}, a_{0:t-1})\right) & \text{if } a = a(M) \end{cases}$$

Since U_t is a uniform distribution on $[0,1]$ and $q_t(\cdot)$ is a probability distribution, we can restate these cases as

$$p_{Y_t}(a) = \begin{cases} q_t(a(1)|s_{0:t}, a_{0:t-1}) & \text{if } a = a(1) \\ \sum_{j=1}^{k} q_t(a(j)|s_{0:t}, a_{0:t-1}) - \sum_{j=1}^{k-1} q_t(a(j)|s_{0:t}, a_{0:t-1}) & \text{if } a = a(k) \text{ with } k \in [2, M-1] \\ 1 - \sum_{j=1}^{M-1} q_t(a(j)|s_{0:t}, a_{0:t-1})) & \text{if } a = a(M) \end{cases}$$

By the fact that $q_t(\cdot)$ is a probability distribution, and the definition of summation, the cases can be further simplified to

$$p_{Y_t}(a) = \begin{cases} q_t(a(1)|s_{0:t}, a_{0:t-1}) & \text{if } a = a(1) \\ q_t(a(k)|s_{0:t}, a_{0:t-1}) & \text{if } a = a(k) \text{ with } k \in [2, M-1] \\ q_t(a(M)|s_{0:t}, a_{0:t-1}) & \text{if } a = a(M) \end{cases}$$

This final set of cases are just a partition of the decision function, and thus can be simplified to

$$p_{Y_t}(a) = q_t(a|s_{0:t}, a_{0:t-1}) \ \forall \ a \in A$$

$\square$