E EXAMPLE

Here we consider a simple example on the SockFarm operator's MDP. Suppose there are 10 products in total, i.e $|\mathcal{P}| = 10$ and the operator has 5 accounts to control, i.e. $|\mathcal{A}| = 5$, and $T = 10$.

The *initial state* as $s_0 = (0, \{dp_{s_0}(a) \mid a \in \mathcal{A}\}, \emptyset, \emptyset) \in S$ and $dp_{s_0}(a) = 0.2, \forall a \in \mathcal{A}$. The terminal states are states where either $t_s = 10$ or $dp_s(a) = 1$ for all $a \in \mathcal{A}$. The set of detected accounts at state s is $D(s)$, which is $\emptyset$ to start with in the initial state. The set of new requests added at state s is $\mathbf{REQ}^+(s)$ which is also empty at the initial state.

In our example ,we have $s_0 = (0, \{dp_{s_0}(a) \mid a \in \mathcal{A}\}, \emptyset, \emptyset) \in S$. After this we go to state $s_1 = (1, \{dp_{s_1}(a) \mid a \in \mathcal{A}\}, \emptyset, \emptyset)$ since no new requests have arrived and the detection probabilities remain unchanged since no reviews have been posted because $A(s_0) = \emptyset$, i.e., $\mathcal{TR}(s_0, \alpha = \emptyset, s_1) = 1$, $R(s_0, \alpha = \emptyset) = 0$, $\pi(s_0) = \alpha = \emptyset$. Similarly, at $t = 2$, we just reach $s_2 = (2, \{dp_{s_2}(a) \mid a \in \mathcal{A}\}, \emptyset, \emptyset)$ with everything else invariant. At $t = 3$, $\mathbf{r_1} = (p_1, 3, 5, 10)$ comes with probability 0.4. Then s_2 with $\pi(s_2) = \alpha = \emptyset$ will transition to one of the following two states:

$$s_3' = (3, \{dp_{s_3'}(a) \mid a \in \mathcal{A}\}, \{\mathbf{r_1}\}, \emptyset)$$
$$s_3'' = (3, \{dp_{s_3''}(a) \mid a \in \mathcal{A}\}, \emptyset, \emptyset)$$

with the following probability, respectively:

$$\mathcal{TR}(s_2, \alpha = \emptyset, s_3') = Pr(\{\mathbf{r_1}\}) \times 1 = 0.4$$
$$\mathcal{TR}(s_2, \alpha = \emptyset, s_3'') = (1 - Pr(\{\mathbf{r_1}\})) \times 1 = 0.6.$$

Here, $dp_{s_3''}(a) = dp_{s_3'}(a) = dp_{s_2}(a) = 0.2, \forall a \in \mathcal{A}$. Now we consider s_3' with policy $\pi(s_3') = \alpha_3' = \{(a_1, p_1), (a_2, p_1), (a_3, p_1)\}$ and the conditioned detection probabilities:

$$Pr((1.0, 0.6, 0.4, 0.2, 0.2) \mid (0.2, 0.2, 0.2, 0.2, 0.2), \alpha_3') = 0.3$$
$$Pr((0.6, 0.6, 0.4, 0.2, 0.2) \mid (0.2, 0.2, 0.2, 0.2, 0.2), \alpha_3') = 0.7,$$

which are determined by the detection algorithm. Then we have the following reward function:

$$R(s_3', \alpha_3') = (0.3 \times 0 + 0.7 \times 10) - (0.3 \times 2 + 0.7 \times 0) - 3 = 3.4,$$

where the SockFarm operator posts each review with cost 1, has one detected account with cost 2 (with probability 0.3), finishes the request with reward 10 (with probability 0.7).

At $t = 4$, $\mathbf{r_2} = (p_2, 3, 5, 10)$ comes with probability 0.4. Then, $\pi(s_3') = \alpha_3'$ will make the operator transit to one of the following four states:

$$s_{4,1} = (4, (1, 0.6, 0.4, 0.2, 0.2), \{(p_1, 1, 5, 10), \mathbf{r_2}\}, \{(a_2, p_1), (a_3, p_1)\})$$
$$s_{4,2} = (4, (1, 0.6, 0.4, 0.2, 0.2), \{(p_1, 1, 5, 10)\}, \{(a_2, p_1), (a_3, p_1)\})$$
$$s_{4,3} = (4, (0.6, 0.6, 0.4, 0.2, 0.2), \{\mathbf{r_2}\}, \{\emptyset\})$$
$$s_{4,4} = (4, (0.6, 0.6, 0.4, 0.2, 0.2), \emptyset, \emptyset\}),$$

where request $\mathbf{r}_1$ in $s_{4,1}$ and $s_{4,2}$ has not been finished because a_1 is detected, and a_1 can not be used anymore. The transition function is:

$$\mathcal{TR}(s_3', \alpha_3', s_{4,1}) = 0.4 \times 0.3 = 0.12$$
$$\mathcal{TR}(s_3', \alpha_3', s_{4,2}) = 0.6 \times 0.3 = 0.18$$
$$\mathcal{TR}(s_3', \alpha_3', s_{4,3}) = 0.4 \times 0.7 = 0.28$$
$$\mathcal{TR}(s_3', \alpha_3', s_{4,4}) = 0.6 \times 0.7 = 0.42.$$

Consider the following policies:

$$\pi(s_{4,1}) = \alpha_{4,1} = \{(a_4, p_1), (a_3, p_2), (a_4, p_2), (a_5, p_2)\}$$
$$\pi(s_{4,2}) = \alpha_{4,2} = \{(a_4, p_1)\}$$
$$\pi(s_{4,3}) = \alpha_{4,3} = \{(a_3, p_2), (a_4, p_2), (a_5, p_2)\}$$
$$\pi(s_{4,4}) = \alpha_{4,4} = \emptyset.$$

Suppose the detection probability for each account is not changed after taking these actions at the corresponding state. Then the requests are finished at the corresponding state with following rewards:

$$R(s_{4,1}, \alpha_{4,1}) = 10 + 10 - 0 - 4 = 16$$
$$R(s_{4,2}, \alpha_{4,2}) = 10 - 0 - 1 = 9$$
$$R(s_{4,3}, \alpha_{4,3}) = 10 - 0 - 3 = 7$$
$$R(s_{4,4}, \alpha_{4,4}) = 0$$

After that, no more requests will come, and the operator will not post any reviews in the following states. Suppose, after s_3'', $\mathbf{r}_2$ is finished successfully, represented by $V(s_3'') = 2.8$. Then the expected reward for the policy π is ($\gamma = 1$):

$$\begin{aligned}
V(\pi) = {} & R(s_0, \alpha_0) + R(s_1, \alpha_1) + R(s_2, \alpha_2) + 0.6 \times V(s_3'') \\
& + 0.4 \times R(s_3', \alpha_3') \\
& + 0.4 \times (0.12 \times R(s_{4,1}, \alpha_{4,1}) + 0.18 \times R(s_{4,2}, \alpha_{4,2}) \\
& \quad + 0.28 \times R(s_{4,3}, \alpha_{4,3}) + 0.42 \times R(s_{4,4}, \alpha_{4,4})) \\
= {} & 0 + 0 + 0 + 0.6 \times 2.8 + 0.4 \times 3.4 \\
& + 0.4 \times (0.12 \times 16 + 0.18 \times 9 + 0.28 \times 7 + 0.42 \times 0) \\
= {} & 1.68 + 1.36 + 0.4 \times (1.92 + 1.62 + 1.96 + 0) \\
= {} & 5.24.
\end{aligned}$$