Eindhoven University of Technology

2MMS50, Stochastic Decision Theory

Assignment 1: Markov Decision Processes

Due: Wednesday May 18th, 2022

Authors

Prof. B. Zwart
Luuc Vlieger

May, 2022

Please provide concise yet complete answers. Always give explanations to your answers.

Question 1: Rapid food delivery services

Most people are familiar with it nowadays; the current so-called 'Flitsbezorgers' or rapid food delivery services. The environment of these rapid food delivery services is very tough with having several competitors such as Getir and Gorillaz, while also having large supermarket brands Jumbo, Spar and Albert Heijn announcing that they will be joining the super competitive market. Investment costs are very high and the only way to be profitable in the future is by becoming the largest company of all competitors. Luckily, there are many investors who are willing to take the risk, and Flink now has a good budget for trying to obtain as much new customers as possible. So, the CEO of Flink has asked you, a consultant, to investigate what marketing strategy will be best way forward for Flink, that maximizes potential profits. There has already been research at the company and the company has estimated that each client corresponds to 5 euros of potential profits. You can add your potential profits to your so-far accumulated profits at the end of a campain.

There are multiple marketing strategies available. One can choose for social media advertisements, which takes 2 months, and costs €200.000 in total. The discount code campain takes 1 month and cost €100.000. Your starting budget is €200.000, and you are only allowed to start a new campain if your budget plus so-far accumulated potential profits are higher than the campain costs. You can also not start a new campain in between when you are busy with a 2-month social media campain.

Amount of new customers	Social Media Advertisement	Discount Codes
Successful campain	100.000	40.000
Unsuccessful campain	20.000	20.000

As each customer in the table above corresponds to €5 of potential profits, we can multiply the amount of customers by this amount to obtain the potential profit per strategy.
Now, when your previous campain was successful, the next campain has a higher probability of staying successful. The probability of some campain being successful after a previously successful campain equals 0.7. This implies that the probability of a successful campain becoming unsuccessful in the next point and time equals 0.3. The probability of some campain being unsuccessful after having an unsuccessful in the previous point in time equals 0.6.
Note that in the explanation above the previous campain does not have to be the same marketing strategy, so when you have an unsuccessful social media campain then the probability of having an unsuccessful discount code campain still equals 0.6. The CEO of Flink has told you that the campain prior to when you started was unsuccessful, so the probabilities of having a successful first campain and unsuccessful first campain are 0.4 and 0.6 respectively. Note again, that the goal will be to maximize potential profits. This equals the available budget plus the amount of customers acquired so far times the amount of potential profits per customer (€5). The entire marketing campain will take four months. Flink wants to know what should be the marketing strategy in order to maximize potential profits over these four months.

Question 1a

Formalize the above problem as a Markov Decision Problem. Also explain why the Markov property holds for this particular problem. Be sure to include:

1. The state space $\mathcal{I}$
2. The action space $\mathcal{A}$
3. The direct rewards $r^{\alpha}(\cdot)$
4. The transition probabilities $p^{\alpha}(\cdot,\cdot)$

Question 1b

Give the optimal policy as well as the expected reward function under this policy. Name the theorem that you use in order to compute these two.

Question 2: Optimal Investment Strategy

For this problem we introduce a discounted infinite horizon problem. The setting is as follows. We have discrete time $\mathcal{T} = \{0, 1, 2, 3, ...\}$. Say that we are shareholder in a company, and that this company follows a discrete process X_t with $X_t = \max\{-5, \min\{Y_t, 5\}\}$ for $t > 0$, where $Y_t = Y_{t-1} + \epsilon_t(Y_{t-1})$, with .

$$\epsilon_t(Y_{t-1}) = \begin{cases} 1 - \mathbb{1}(Y_{t-1} = 5) & \text{w.p. } 0.5 \\ -1 + \mathbb{1}(Y_{t-1} = -5) & \text{w.p. } 0.5 \end{cases} \tag{1}$$

When the company is doing well, when $X_t > 0$, we get $r(X_t) = X_t$ profit as a reward. When the company is doing badly, when $X_t < 0$, then the company will ask for support via a reinvestment of $r(X_t) = X_t$. Let $\alpha = 0.95$ be the discount factor.

At each point in time, we have the choice of staying a shareholder and obtaining reward $r(X_t) = X_t$ (note, this can be negative when $X_t < 0$), which we call action 0, or leaving the company as a shareholder at time T and obtaining $r(X_t) = 0$, $\forall t \geq T$, so gaining no utility for the rest of time, which we call action 1. So, at each point in time T we have the choice of staying or leaving the company, and gaining $r(X_T) = X_T$ rewards or $r(X_t) = 0$ for all $t \geq T$ respectively.

The goal is to maximize the total accumulated reward over time, which for a certain strategy s equals $V^{\mathbf{s}}_{0.95}(i) = \sum_{n=0}^{\infty} 0.95^n \cdot \mathbb{E}^{\mathbf{s}}[r^{A_n}(X_n)|X_0 = i]$. The question is what the optimal policy $\mathbf{f}^*$ will be in order to maximize this total accumulated reward.

The corresponding discounted functional equation is:

$$(T^*_{\alpha} v)(i) = \max\{0, r^0(i) + \alpha \cdot \sum_{j \in \mathcal{I}} v(j) \cdot p^0(i, j)\} \tag{2}$$

Question 2a

Give the following:

1. The state space $\mathcal{I}$

2. The action space $\mathcal{A}$
3. The direct rewards $r^{\mathbf{f}}(\cdot)$
4. The transition probabilities $p^{\mathbf{f}}(\cdot, \cdot)$

Also give the corresponding discounted functional equation as in 2.

Question 2b

Say we have the policy $\mathbf{f}_0$ of stopping when we obtain negative rewards $r^{\mathbf{f}_0}(X_t) = X_t < 0$ at some point in time.

Show that this strategy is optimal or not optimal by performing a step of policy iteration and showing that the improved policy is equal or better than this policy respectively. Give the policy $\mathbf{f}_0$, the value function $V_{0.95}^{\mathbf{f}_0}(\cdot)$ and the policy $\mathbf{f}_1$ after one iteration of policy iteration. Explain why the policy $\mathbf{f}_0$ is optimal or not optimal. Also give the long term average reward under the strategy $\mathbf{f}_0$, given that you are still part of the company as a shareholder.

Question 2c

Apply successive approximation to the above problem to obtain the optimal policy.

Question 2d

Say that the errors are changed to

$$\epsilon_t(Y_{t-1}) = \begin{cases} 1 - \mathbb{1}(Y_{t-1} = 5) & \text{w.p. } \rho \\ -1 + \mathbb{1}(Y_{t-1} = -5) & \text{w.p. } 1 - \rho \end{cases} \tag{3}$$

, where $\rho \neq 0.5$. Argue what will change to your answer to question 2b in the upcoming aspects. Analyze both the cases $\rho < 0.5$ and $\rho > 0.5$.

1. In what way does the value function $V_{0.95}^{\mathbf{f}_0}(\cdot)$ change compared to your answer in 2b?
2. Will the policy $\mathbf{f}_0$ still be optimal / still not be optimal?
3. Will the average long term reward given that you are still a shareholder better or worse than in question 2d?
4. Will the successive approximation algorithm still converge? Why / why not?