Fundamentals of Reinforcement Learning

Master IASD

2021–2022

1 Markov Decision Process

A company is looking for an intern, and has very little time to organise interviews. Interviewing a candidate allows to discover her/his quality. Let us consider candidates can be of three types: suitable for the position, perfect for the position, or not a good fit for the position. The company has observed in the past that 50% of the candidates are suitable, 25% are not a good fit, and 25% are perfect for the position.

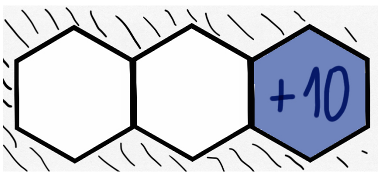
Given the time constraints, the company can organise at most two interviews and *must* decide whether to hire a candidate before interviewing the next one.

Hiring a suited candidate will earn a return of 50 for the company, the return will be of 200 for a perfect candidate. Making an interview costs 30 to the company. The company will *not* hire someone that is not a good fit and will prefer not to fill the position if is too expensive.

Question 1. *Model this problem as a Markov Decision Process (MDP): provide the description of all states, all actions, describe the transition function as well as the reward function (you can draw a graph representation or define the corresponding matrices). Do not make assumptions about the solution of this problem (one could change the costs or the probability distribution over the candidates' types).*

2 Policy Improvement

Let us consider an MDP with three states s_0, s_1 and s_2, shown from left to right on the graph below, and 6 actions. In s_0 and s_1, six actions are available: going *east*, going *west*, going *north east*, going *north west*, going *south east*, and going *south west*.



The transition function works as follows. When taking an action, we effectively go in that direction with a probability 0.7, otherwise, we slide slightly either on the right or on the left of the desired direction. For instance, if an action is going *north west*, the agent will actually go *north west* with probability 0.7, and it will end up going *north* with a probability 0.15, and *west* with a probability 0.15. If the action makes the agent hit the border off the mosaic (shaded area), the agent actually bounces back and remains in the same position. For instance, going *north west* in s_0, the agent is guaranteed to remain in s_0. If the agent take action *east* in s_0, it will end up in s_1 with probability 0.7 and it will remain in s_0 with probability 0.3.

The reward function is as follows: when the agent bounces back in the same state, it receives a penalty of 1 (i.e. a reward of -1). When the agent reaches state s_2, it receives 10 and the episode terminates. The discount factor is chosen as $\gamma = 0.9$

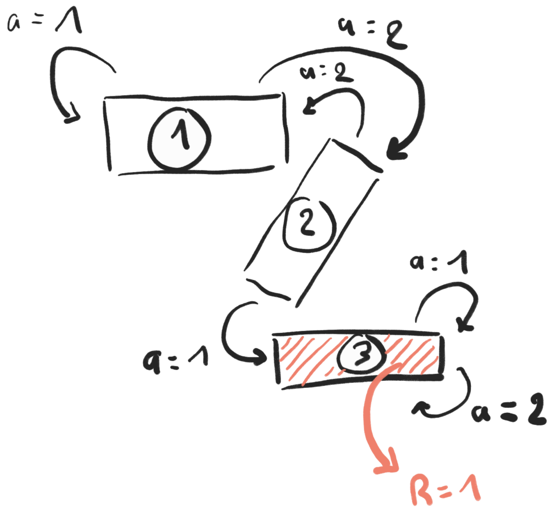
Question 2. *Let us assume that the policy π is $s_0 \mapsto$east and $s_1 \mapsto$north east. What system of equations should one solve to compute v_π?*

Question 3. *We computed for you the solution of this system of equations and the solution is* $v_\pi = \begin{pmatrix} 1.425 \\ 2.128 \\ 10 \end{pmatrix}$. *Show the existence of an improvement that results in a new improved policy* π'.

Question 4. *Is* π' *optimal?*

3 Policy Gradient

We consider the simple Markov decision process with three states and two actions per state depicted below.



Action 1 can be interpreted as "going left" and action 2 as "going right", all action outcomes are deterministic and state 3 is a terminal state which provides reward 1 for both actions (while the two other states provide 0 reward). We start in state 1, i.e., $S_0 = 1$ and focus on the undiscounted finite-horizon criterion for horizon $T = 3$, maximizing $v_\pi = \mathbb{E}_\pi[\sum_{i=0}^{2} R_{i+1}]$.

Question 5. *Show that the optimal policy is such that* $v^* = 2$.

Question 6. *Show that fixed deterministic policies, i.e., such that* $\pi(1|s) = 1$ *or* $\pi(2|s) = 1$ *for all states* $s \in \{1, 2, 3\}$, *are such that* $v_\pi = 0$.

We now want to find the optimum policy among the family of constant randomized policies such that $\pi_\theta(2|s) = \theta \in (0, 1)$ for all $s \in \{1, 2, 3\}$, using policy gradient computations.

Question 7. *Show that*

$$\frac{dv_\theta}{d\theta} = \mathbb{E}_\theta\left[\left(\sum_{i=0}^{2} R_{i+1}\right)\left(\sum_{j=0}^{2} \frac{d\log \pi_\theta(A_j|S_j)}{d\theta}\right)\right] \tag{1}$$

Question 8. *Among the* 8 *possible sequences* (A_0, A_1, A_2) *of actions, show that there are only 3 of them that correspond to non-zero cumulative rewards and compute for each of them:*

$\sum_{i=0}^{2} R_{i+1}$	$\sum_{i=0}^{2} \frac{d\log \pi_\theta(A_i\|S_i)}{d\theta}$	*the probability of the sequence* (A_0, A_1, A_2)

Question 9. *Show using* (1) *that*

$$\frac{dv_\theta}{d\theta} = 3\theta^2 - 8\theta + 3$$

and give the value of $\theta \in (0, 1)$ *that corresponds to the optimal constant randomized policy.*

4 Multiple Play Bandit

In recommendation applications, it may be desirable to recommend bundle of products. Here we consider bandit algorithms suitable for recommending a pair of two distinct items.

Let $\theta_1, \dots, \theta_K$ denote unknown parameter values in $[0, 1]$, which will be assumed to be all distinct, i.e., such that $\theta_j \neq \theta_k$. At each time t, we are allowed to select a pair $A_t = (j, k)$ of items, where $1 \leq j \leq K$, $1 \leq k \leq K$ and $j \neq k$. Given A_t, the observed reward X_t satisfies:

$$\mathbb{E}[X_t | A_t = (j, k)] = \theta_j + \alpha\theta_k$$

where $0 < \alpha < 1$ is a known parameter. We will assume that the rewards X_t take their values in $[0, 1]$.

Question 10. *Define precisely the set of possible actions in this model. If the parameters $\theta_1, \dots, \theta_K$ were known, what action would maximize the expected reward?*

Question 11. *Write the expected regret up to horizon T as a function of the parameters and of the expected counts $\mathbb{E}[N_{(j,k)}(T)] = \sum_{t=1}^{T} \mathbb{P}[A_t = (j, k)]$.*

A first approach consists in using the standard UCB algorithm on the set of all possible actions.

Question 12. *Describe the UCB algorithm applied to this problem.*

Recall that for a J-armed bandit, the expected regret of UCB satisfies

$$\mathbb{E}[R_T] \leq \sum_{\substack{j=1 \\ j \neq j^*}}^{J} C \frac{\log T}{\Delta_j} + O(1)$$

where C is a constant and Δ_j denotes the gap between arm j and the optimal arm j^*.

Question 13. *Use this result to obtain a bound on the performance of UCB when applied to the multiple play model. How does the performance depend on the horizon T and on the number of items K? Intuitively, do you believe these dependencies to be optimal?*

Question 14. *A different way of proceeding consists in using a bandit algorithm suitable for linear bandits. Show that the multiple play bandit may be represented as a linear bandit model using a fixed set of $K(K-1)$ context vectors (to be defined) of dimension K.*

5 Best Arm Selection

Consider a two arm Gaussian bandit model with arm distributions $\nu_1 = \mathcal{N}(\mu_1, \sigma^2)$ and $\nu_2 = \mathcal{N}(\mu_2, \sigma^2)$. It is recalled that *(i)* the $\mathcal{N}(\mu, \sigma^2)$ distribution has probability density function $p(x) = 1/(\sqrt{2\pi}\sigma) \exp[-(x - \mu)^2/(2\sigma^2)]$; *(ii)* if X follows a $\mathcal{N}(\mu, \sigma^2)$ distribution,

$$\mathbb{P}(X < x) \leq e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

when $x < \mu$. We will denote by $\Delta = |\mu_1 - \mu_2|$ the gap between the two arms.

We are interested in algorithms that select the best arm, i.e. the one with the highest expectation, with probability at least $1 - \delta$, where δ is a pre-specified maximal probability of error.

We first consider a deterministic allocation rule such that for a time-horizon T, that is assumed to be even, one has $N_1(T) = N_2(T) = T/2$, with the following decision rule used at time T:

$$\begin{cases} \textbf{Select arm 1} \text{ If } \bar{X}_1(T) > \bar{X}_2(T) \\ \textbf{Select arm 2} \text{ Otherwise} \end{cases}$$

Question 15. *Show that $\bar{X}_1(T) - \bar{X}_2(T)$ follows a $\mathcal{N}(\Delta, 4\sigma^2/T)$ distribution when $\mu_1 > \mu_2$.*

Question 16. *Deduce from what precedes that the previous algorithm selects the best arm with probability at least* $1-\delta$ *when*

$$T \geq \frac{8\sigma^2}{\Delta^2} \log \frac{1}{\delta}$$

When Δ is known, it may be possible to reach a decision earlier by the following decision rule:

$$\begin{cases} \textbf{Select arm 1 } \text{If } \bar{X}_1(T) > \bar{X}_2(T) + 4\sigma^2 \log(1/\delta)/(\Delta T) \\ \textbf{Select arm 2 } \text{If } \bar{X}_2(T) > \bar{X}_1(T) + 4\sigma^2 \log(1/\delta)/(\Delta T) \\ \textbf{Do not make any decision } \text{otherwise} \end{cases}$$

Question 17. *Show that the probability that the above algorithm selects the sub-optimal arm is upper bounded by* δ*.*

Question 18. *Conversely, show that the above algorithm selects the best arm with probability at least 1/2 whenever*

$$T \geq \frac{4\sigma^2}{\Delta^2} \log \frac{1}{\delta}$$

We now want to find related results for more general algorithms using lower bound arguments. Recall that for any sequential algorithm and any bandit model one has

$$\sum_{k=1}^{K} \mathrm{KL}(\nu_k, \nu'_k) \mathbb{E}_\nu[N_k(T)] \geq d(\mathbb{P}_\nu(E), \mathbb{P}_{\nu'}(E))$$

where $\mathrm{KL}(\nu, \nu') = \mathbb{E}_\nu[\log(\nu(X)/\nu'(X))]$ denotes the Kullback-Leibler divergence between two different distribution ν and ν', $d(p,q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$ is the Bernoulli Kullback-Leibler divergence and E is any event. We will admit that the above inequality also holds true when T is a random stopping time.

Question 19. *Show that when* ν *and* ν' *correspond respectively to the* $\mathcal{N}(\mu, \sigma^2)$ *and* $\mathcal{N}(\mu', \sigma^2)$ *distributions, one has*

$$\mathrm{KL}(\nu', \nu') = \frac{(\mu - \mu')^2}{2\sigma^2}$$

Question 20. *In the two arms case, assuming that our algorithm is such that at (a possibly random) time* T *it selects the correct arm with probability of error smaller than* δ *for any model, show by considering the changes of distribution* $\{\mu'_1 = \mu_1, \mu'_2 = \mu_1 + \epsilon\}$ *and* $\{\mu'_1 = \mu_2 - \epsilon, \mu'_2 = \mu_2\}$ *where* ϵ *is a positive quantity, and letting* ϵ *tend to zero, that*

$$\mathbb{E}_\nu[N_1(T)] \geq \frac{2\sigma^2}{\Delta^2} d(\delta, 1-\delta) \quad \textit{and} \quad \mathbb{E}_\nu[N_2(T)] \geq \frac{2\sigma^2}{\Delta^2} d(\delta, 1-\delta)$$