REINFORCEMENT LEARNING

Homework 1

Author
Paolo Barba, 1885324
Sapienza, University of Rome

Exercise 1

Derive the formula to get the minimum number of iterations of Value Iteration that are needed if we want an error on the quality of the policy that is at most ϵ.

The formula we want to derive is the following:

$$i \geq \frac{\ln\left(\frac{2}{\epsilon(1-\gamma)^2}\right)}{1-\gamma}$$

where:

- i : Number of iterations. $i \in \mathbb{N}$
- γ : The discount factor. $\gamma \in [0\ ;1]$
- ϵ : The error want to achieve.

The goal is to achieve the $||V^{\pi_i}(s) - V^*(s)|| \leq \epsilon$.

$$\begin{aligned}
\|V^{\pi_t}(s) - V^*(s)\| &= Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^*(s)) \\
&= Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^t(s)) + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s)) \\
&= \gamma \mathbb{E}_{s \sim P(s, \pi^t(s))} \left(V^{\pi_t}(s') - V^*(s')\right) + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s)) \\
&\geq \gamma \mathbb{E}_{s \sim P(s, \pi^t(s)}\big(V^{\pi_t}(s') - V^*(s')\big) + Q^*(s, \pi^t(s)) - Q^t(s, \pi^t(s)) + Q^t(s, \pi^*(s)) - Q^*(s, \pi^*(s)) \\
&\geq \gamma \mathbb{E}_{s \sim P(s, \pi^t(s))} \left(V^{\pi_t}(s') - V^*(s')\right) - 2\gamma^t \left\|Q^0 - Q^*\right\|
\end{aligned}$$

Explanation of the above series of equations and inequalities

- First row: by applying the definition of Q value.
- Second row: by adding and subtracting the same quantity, which is $Q^*(s, \pi^t(s))$ the equality holds.
- Third row: by applying the definition of Q value, as the expected value of the next state following the policy π^i computed for both Q^{π^t} and Q^*.
- Fourth row: since $\pi^t(s)$ is the policy that maximizes Q^T we can notice that $-Q^t(s, \pi^t(s)) + Q^t(s, \pi^*(s))$ is a negative number and then we can introduce the greater inequality.
- Fifth row: in general $|||Q^{t+1} - Q^*||| \leq \gamma^{t+1} |||Q^0 - Q^*|||$ since Q^* is a fixed point solution. In the row, we are considering the opposite sign of the formula twice, which implies a greater inequality.

By iterating the process i times we will end up with $(V^{\pi_t}(s') - V^*(s')) \approx \frac{1}{1-\gamma}$ multiplied by γ^i

So we can switch the problem to find the i such that:

$$\frac{2\gamma^i}{(1-\gamma)}||Q^0 - Q^*|| \leq \epsilon$$

where Q^0 is a fixed value we can set to 0 for simplicity.

First by adding and subtracting 1 and then, since $1+x \leq e^x \; \forall x \in \mathbb{R}$ and $Q^*$is in the range$(0, \frac{1}{1-\gamma})$ we can introduce the lower inequality.

$$\begin{aligned}\frac{2\gamma^i}{(1-\gamma)}||Q^*|| &= \frac{2(1-(1-\gamma^i)}{(1-\gamma)}||Q^*|| \\ &\leq \frac{2e^{-(1-\gamma)i}}{(1-\gamma)^2}\end{aligned}$$

Finally we have to resolve the following inequality for i.

$$\frac{2e^{-(1-\gamma)i}}{(1-\gamma)^2} \leq \epsilon$$

Multiply both sides by $(1-\gamma)^2$ to get rid of the denominator:

$$2e^{-(1-\gamma)i} \leq \epsilon(1-\gamma)^2$$

Divide both sides by 2:

$$e^{-(1-\gamma)i} \leq \frac{\epsilon(1-\gamma)^2}{2}$$

Take the natural logarithm of both sides:

$$-(1-\gamma)i \leq \ln\left(\frac{\epsilon(1-\gamma)^2}{2}\right)$$

Now, divide both sides by $-(1-\gamma)$, noting that $(1-\gamma)$ is positive since γ is in the range $[\ 0 \ \ ; 1]$:

$$i \geq \frac{\ln\left(\frac{\epsilon(1-\gamma)^2}{2}\right)}{-(1-\gamma)}$$

Since $ln(\frac{1}{x}) = -ln(x)$ the solution for i is:

$$i \geq \frac{\ln\left(\frac{2}{\epsilon(1-\gamma)^2}\right)}{1-\gamma}$$

■

Exercise 2

Given the following environment settings:

- States : $\{s_i : i \in \{1 \dots 7\}\}$
- Reward :

$$r(s,a) = \begin{cases} 0.5 & \text{if } s = s_1 \\ 5 & \text{if } s = s_7 \\ 0 & \text{otherwise} \end{cases}$$

- Dynamics : $p(s_6|s_6, a_1) = 0.3, p(s_7|s_6, a_1) = 0.7$
- Policy: $\pi(s) = a_1 \forall s \in S$
- Value function at iteration $k = 1$:

$$v_k = [0.5, 0, 0, 0, 0, 0, 5]$$

- Discount Factor : $\gamma = 0.9$

Compute $V_{k+1}(s_6)$ following value iteration algorithm.
For updating the Value we should compute the following formula:

$$V_{k+1}(s) = Q(s; argmax_{a \in A} Q(s,a))$$

Since we have the set of actions $\mathbb{A} = \{a_1\}$ we just have to compute $Q(s, a_1)$.

$V_{k+i}(s_6) = Q(s_6, a_1) = r(s_6, a_1) + \gamma \sum_{s' \in S} p(s'|s,a) V_k(s') = 0 + 0.9(0.3 \cdot 0 + 0.7 \cdot 5) = 0.9 \cdot (3.5) = 3.15$

■

Report Code Exercises

Policy iteration

For the policy iteration exercise we are ask to compute the *reward*, *check feasibility* and *transition probabilities* functions.

The *reward function* is set to 1 if s corresponds to the goal state ($[envsize-1, envsize-1]$), otherwise it is 0.

The *check feasibility* function checks the feasibility of transitioning from state s to state s'. It ensures that s' is within the bounds of the environment. The obstacles (defined by the obstacles matrix) are considered feasible as said in the `classroom discussion chat`.

The *transition probabilities* function computes the transition probabilities for each possible next state given an action a taken from state s. It uses the check feasibility function to determine which states are reachable, if the new state is not reachable the action lead to the current state. Since the action is stochastic with action space $\mathbb{A} = \{\text{LEFT , DOWN , RIGHT UP}\}$ coded as $\{0, 1, 2, 3\}$ the probability of taking action $a' \in \mathbb{A}$ depends on the specific action a as the following:

$$\mathbb{P}(a'|a) = \begin{cases} \frac{1}{3} \text{ if } a' = & a \\ \frac{1}{3} \text{ if } a' = & mod(\frac{a+1}{4}) \\ \frac{1}{3} \text{ if } a' = & mod(\frac{a+3}{4}) \end{cases}$$

ilqr

For the ilqr exercise we are ask to compute some formulas:

$$p_t = q_t + K_t^T \left(R_t k_t + r_t\right) + \left(A_t + B_t K_t\right)^T p_{t+1} + \left(A_t + B_t K_t\right)^T P_{t+1} B_t k_t$$

$$P_t = Q_t + K_t^T R_t K_t + \left(A_t + B_t K_t\right)^T P_{t+1} \left(A_t + B_t K_t\right)$$

These formulas are used to compute the value function and gain matrices at each time step during the optimization process.

$$\dot{\theta}_{\text{new}} = \dot{\theta} + \left(1.5 \left(\frac{g}{l}\right) \sin(\theta) + \frac{3u}{ml^2}\right) \cdot \text{dt}$$

$$\theta_{\text{new}} = \theta + \dot{\theta}_{\text{new}} \cdot \text{dt}$$

These formulas describe the dynamics of the pendulum system. They are used to compute the new state of the pendulum based on the given dynamics equations.

$$\text{control} = k_t + K_t \cdot (\hat{x}(t) - x(t))$$

This formula represents the calculation of the control input at time step t, with the difference between the estimated state $\hat{x}(t)$ and the actual state $x(t)$.

■