Topics in RL – Project 2

Dr. Tejas Bodas, Dr. Harikumar, TAs

April 2024

Common Instructions

For each project, you should implement the any 5 algorithms (out of 7) using only Python to find the optimal policy:

1. DQN
2. policy gradient
3. actor-critic
4. TRPO
5. natural actor-critic
6. Rainbow
7. A3C

Also plot the instantaneous and cumulative regrets for the RL algorithms defined as follows.

Instantaneous episodic regret: It is the difference between the total discounted reward earned by your RL algorithm in the current episode and the expected cumulative discounted reward earned by the optimal policy in an episode (essentially $V_\alpha(s)$ is the episode started in state s). It measures how much reward you are losing choosing a sub-optimal policy.

Cumulative regret: At the current iteration/episode, this is running/cumulative sum of all the previous instantaneous episodic regrets till now.

Project 1: Inventory Management problem

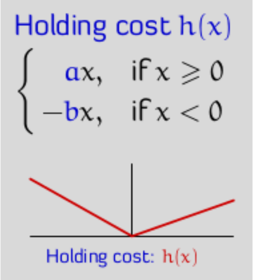
Retail stores stockpile products in warehouses to meet the random demand. Additional stocks are procured at regular intervals. Let X_t denote the amount of stock before the t-th procurement. In this example, time denotes the number of additional stock procurements. At time t, the store may procure an additional stock U_t $(\leq U)$ units for price p per unit. Thus the total procurement cost is pU_t.

The random demand W_t is i.i.d. with distribution P_W. The stock available at the next time is $X_{t+1} = X_t + U_t - W_t$, where a negative stock denotes backlogged demand.

The holding cost for the stock is given by $h(x)$ where a is the per-unit storage cost and b is the per-unit backlog cost.

Per-stage cost is $c(X_{t+1}, U_t) = h(X_{t+1}) + pU_t$. Find the optimal inventory control strategy to minimize the expected total cost over a finite horizon.

Parameters: Take P_W as uniform[0,10].



Reference: Page 14 of Aditya Mahajan Slides on Markov Decision Processes: Sequential decision-making with perfect observation

Project 4: Call options

An investor has a *call option* to buy one share of a stock at a fixed price p and has T days to *exercise* this option. For simplicity, assume that the investor makes a decision at the beginning of each day.

The investor may decide not to exercise the option but if he does exercise the option when the stock price is s, he effectively gets $(s - p)$.

Assume that the price of the stock varies with independent increments, *i.e.*, the price on day $t + 1$ is

$$S_{t+1} = S_t + W_t$$

where $\{W_t\}_{t \geq 1}$ is an i.i.d. process.

Your task is to create an agent that decides when to exercise (or not exercise) the call option in the given span of T days. Assume $p \in \mathbb{N}$ and assume W_t to be (discrete) uniformly distributed with endpoints $-\epsilon$ and $+\epsilon$ for some $\epsilon \in \mathbb{N}$. For example, for $\epsilon = 5$, $W_t \sim \mathcal{U}\{-5, 5\}$.

References:

- Call Option Wiki
- Discrete Uniform Distribution Wiki
- Aditya Mahajan Notes on Call Option
- Page 23 of Aditya Mahajan Slides on Markov Decision Processes: Sequential decision-making with perfect observation

Project 7: A House Selling Example

An individual wants to sell his house and an offer comes in at the beginning of each day. Assume that the successive offers are independent and an offer is j with probability P_j, $j = 0, 1, \ldots, N$ (here is j is ”ranking” of offers). We suppose, however, that any offer not immediately accepted is not lost but may be accepted at any later date. Also, a maintenance cost of C is incurred each day the house remains unsold.

The state at time t will be the largest offer received up to t (including the offer at t). Therefore, (i is currently the best offer ”rank”, j is the next day offer ”rank”)

$$P_{ij} = \begin{cases} 0 & j < i \\ \sum_{k=0}^{i} P_k & j = i \\ P_j & j > i \end{cases}$$

Assume N = 25, i = 10, and P to follow a uniform distribution. Find the optimal strategy.

Reference: Applied Probability Models with Optimization Application, Sheldon Ross (Chapter 6).

Project 13: Portfolio Optimization with single asset

Consider the portfolio optimization problem at this link. You will have to consider ne assets in your portfolio that you have to optimize. This is discussed in chapter 3. In the illustrated examples, where are only 3 weights $\{-1, 0, 1\}$ that are considered. You can consider -0.5 and 0.5 in addition as well.

Project 14: Finite inventory pricing

Assume you operate a chartered plane of K seats with a booking window of T days.i.e., the plane leaves from A to B after every T days. Bookings can be made only for that particular segment and for the next departing flight and not for future departures. As an operator, you want to come up with an optimal pricing policy $p_t, t = 1, 2, \ldots, T$ that will maximize your total revenue over T days. Time is discrete, and on day t, the demand for seats at price p_t, denoted by $d(p_t)$ is a random variable supported on positive integers. Some examples are Bernoulli, Poisson, Geometric random variables. Note that if available seats are less than the demand on a day, then some demand is lost (but there is no cost for lost sales). First perform VI or PI to get the optimal pricing policy. Now assume that the demand function is unknown and use RL algorithms to see if the optimal policy is learned.

CS7.603: Topics in Reinforcement Learning - Final Project 2024

Instructions:

- **Average reward versus episode plot is needed for all the algorithms.**
- **Plots of optimal (the best one after the completion of training) control input trajectory and corresponding state trajectory must be given for all the algorithms.**
- **Mention the architecture of Neural Networks used for function approximation with a Table citing the hyper-parameters used.**
- **Assume any data if found missing and mention that in the presentation.**
- **Presentation slides must be submitted after the necessary modifications suggested during the project presentation.**

Q.1) The model equations of the pendulum shown in Figure 1 is given below.

$$ml^2\frac{d^2\theta(t)}{dt^2} + b\frac{d\theta(t)}{dt} + mglsin(\theta(t)) = u(t) \tag{1}$$

Here m is the mass in kg, l is the length in m, g is the acceleration due to gravity, b is the coefficient of air friction, $\theta(t)$ is the angle from vertical and $u(t)$ is the control input torque.
Find the optimal control input that takes the pendulum from $\theta(0) = 0$ radians to $\theta(t_f) = \pi$ radians? (Here, the final time t_f is a free variable.)
Formulate an appropriate cost function to be minimized?
System paramters are given by m=1.5 kg, l=0.8 m, b=0.2
Find the optimal control when m, l and b are unknown using
a) Deep Q-Learning (consider discrete action space)
b) DDPG
c) PPO
d) SAC

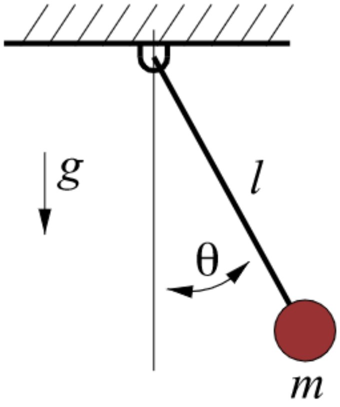


Figure 1: Pendulum

Q.2) The dynamics of the cart-pole system shown in Figure 2 is given below. Here M and m_p is the mass (kg) of the cart and pole respectively. The linear displacement (in m) of the cart is denoted by x, g is the acceleration due to gravity, θ is the angular displacement (radians) of

the pole of length L (in m) and F_x is the input force applied to the cart (in N).
Find the optimal control input (F_x) that takes the pole from the initial angular position i) $\theta(0) = \frac{\pi}{3}$, ii) $\theta(0) = \frac{\pi}{6}$, iii) $\theta(0) = \frac{\pi}{2}$ to the desired angular position $\theta(t_f) = 0$? (Here, the final time t_f is a free variable.)
M= 40 Kg, m_p= 2 kg, $L = 0.75\ m$. **Find the optimal control when system parameters are unknown using**

a) Deep Q-Learning (consider discrete action space)
b) DDPG
c) PPO
d) SAC

$$\ddot{\theta} = \frac{-m_p L \sin\theta\cos\theta\dot{\theta}^2 + (M + m_p)\, g\sin\theta + \cos\theta F_x}{(M + m_p(1 - \cos^2\theta))\, L}$$

$$\ddot{x} = \frac{-m_p L \sin\theta\dot{\theta}^2 + m_p g \sin\theta\cos\theta + F_x}{M + m_p(1 - \cos^2\theta)}$$

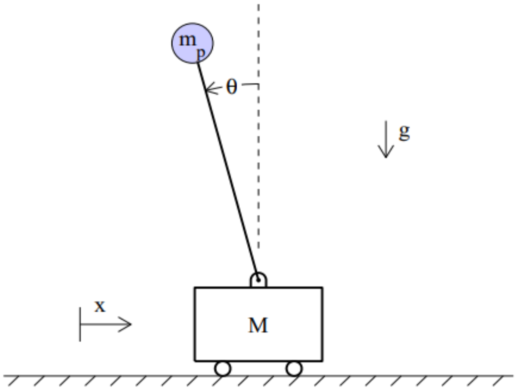


Figure 2: Cart-Pole System

Q.3) The state-space model of a ground robot is given below. Here $(x,\, y)$ is the position, $(v_x,\, v_y)$ is the velocity and $(a_x,\, a_y)$ is the input acceleration. The coefficient of friction is given by $\mu = 0.1$.

$$\dot{x} = v_x \tag{2}$$

$$\dot{v}_x = a_x - \mu\, v_x \tag{3}$$

$$\dot{y} = v_y \tag{4}$$

$$\dot{v}_y = a_y - \mu\, v_y \tag{5}$$

Find the optimal control input to track a desired trajectory as given below if the initial position and velocity are given by (0,0) and (1,-1) respectively.

$$x_d(t) = 5\,cos(0.4t), \;\; y_d(t) = 5\,\sin(0.4t) \tag{6}$$

The maximum value of speed and acceleration is given by 5 m/s and 2 m/s^2 respectively. Use the following algorithms.
a) Deep Q-Learning (consider discrete action space)
b) DDPG
c) PPO
d) SAC

Q.4): Two identical ground robots $G1$ and $G2$ are shown in Figure (3) with $O1$ to $O7$ representing the static obstacles. The size of the arena is 10 m × 10 m. The initial positions of $G1$ and $G2$ are given by (3,9) and (8,2) respectively. Their destination coordinates are given by (8,1) and (3,10) respectively. The state-space model of the ground robots are given below with a maximum acceleration input of 1 m/s^2. The coefficient of friction is given by $\mu = 0.2$.

$$\dot{x} = v_x \tag{7}$$

$$\dot{v}_x = a_x - \mu\, v_x \tag{8}$$

$$\dot{y} = v_y \tag{9}$$

$$\dot{v}_y = a_y - \mu\, v_y \tag{10}$$

All the obstacles are circular in shape with a diameter of 1 m. The coordinates of static obstacles are given by O1=(3,7), O2=(5,7), O3=(2,5), O4=(5,5), O5=(8,5), O6=(5,3), O7=(7,3). Find the optimal control that drives the ground robots to their destination in minimum time maintaining a minimum distance of 0.2 m from the static obstacles and 0.5 m from each other.
Find the optimal control when system parameters are unknown using
a) Deep Q-Learning (consider discrete action space)
b) DDPG
c) PPO
d) SAC

Q.5): The stat-space model of the longitudinal dynamics of a fixed-wing UAV (Uncrewed Aerial Vehicle) is given below.

$$\dot{u} = -qw + \frac{0.5V_a^2}{m}(C_L\,sin(\alpha) - C_D\,cos(\alpha) + C_{L\delta_e} sin(\alpha)\,\delta_e) - 9.81\,sin(\theta) + \frac{T}{m} \tag{11}$$

$$\dot{w} = qu + \frac{0.5V_a^2}{m}(-C_L\,cos(\alpha) - C_D\,sin(\alpha) - C_{L\delta_e} cos(\alpha)\,\delta_e) + 9.81\,cos(\theta) \tag{12}$$

$$\dot{q} = \frac{0.5V_a^2}{J_{yy}}(C_{m0} + C_{m\alpha}\alpha + 0.25C_{mq}\frac{q}{V_a} + C_{m\,\delta_e}\,\delta_e) \tag{13}$$

$$\dot{\theta} = q \tag{14}$$

$$\dot{h} = -u\,sin(\theta) + w\,cos(\theta) \tag{15}$$

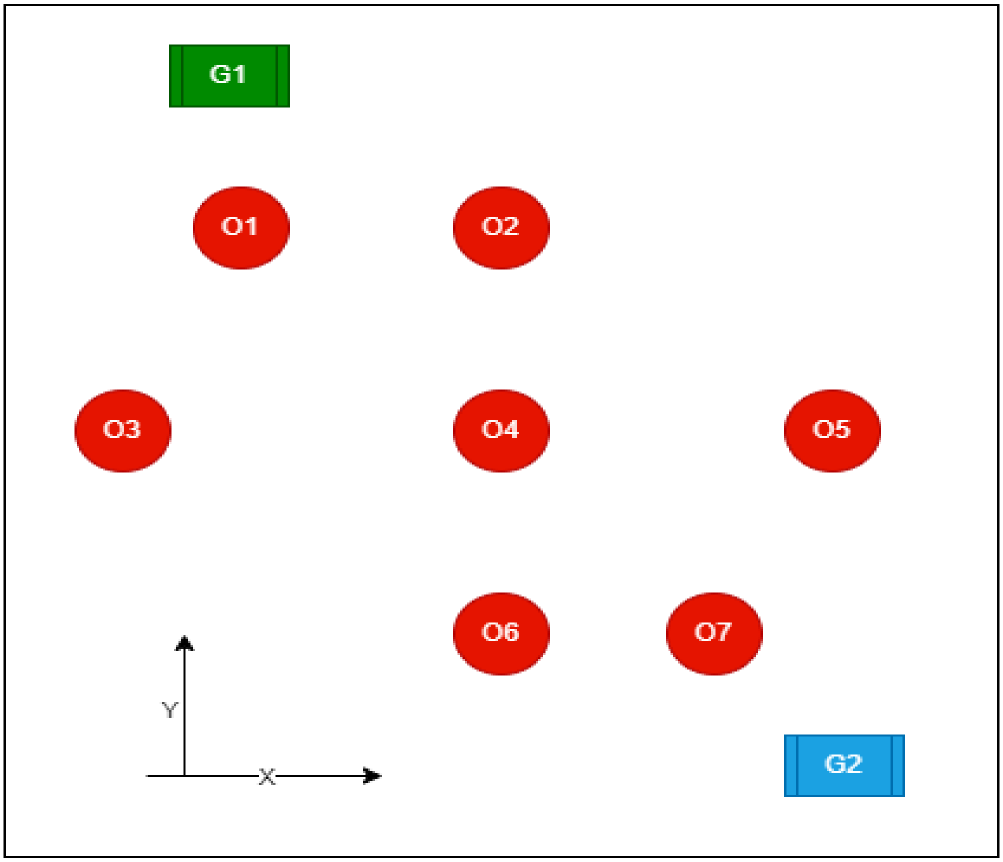


Figure 3: Robot Navigation

States are given by $X = [u,\ w,\ q,\ \theta,\ h]^T$ and control input is given by $U = [T,\ \delta_e]^T$. Here $tan(\alpha) = \frac{w}{u}$, $V_a = \sqrt(u^2 + w^2)$.
The UAV parameters are given below.
$m = 1.56$ Kg, $J_{yy} = 0.0576$ Kg-m^2, $C_L = 3.5\alpha + 0.09$, $C_D = 0.2\alpha + 0.016$, $C_{L\delta_e} = 0.27$, $C_{m0} = -0.02$, $C_{m\alpha} = -0.57$, $C_{mq} = -1.4$, $C_{m\,\delta_e} = -0.32$.

The initial values of states are $X(t = 0) = [9.96,\ 0.87,\ 0,\ 0.0873,\ 50]^T$ and control input $U(t = 0) = [1.0545,\ -0.2179]^T$.
The desired values of the state after 10 seconds are given by $X_d(t = 10) = [9.85,\ 1.74,\ 0,\ 0.1745,\ 62]^T$.
Find the optimal control input with constraints $0 \leq T \leq 4$, $0 \leq \alpha \leq 0.2618$, $5 \leq V_a \leq 15$ and $-0.4363 \leq \delta_e \leq 0.1745$.
Use data-driven approach with a sampling time of 0.02 seconds.
a) Deep Q-Learning (consider discrete action space)
b) DDPG
c) PPO
d) SAC

Q.6): The stat-space model of a quadrotor UAV (Uncrewed Aerial Vehicle) is given below. Here $\mu = 0.05$, m= 1 Kg and g=9.81 m/s^2. Convention for (x, y, z) is given by x facing North, y facing East and z facing Down.

$$\dot{x} = v_x \tag{16}$$

$$\dot{v}_x = a_x - \mu v_x \tag{17}$$

$$\dot{y} = v_y \tag{18}$$

$$\dot{v}_y = a_y - \mu v_y \tag{19}$$

$$\dot{z} = v_z \tag{20}$$

$$\dot{v}_z = a_z - \mu v_z \tag{21}$$

where

$$a_x = (-0.7071\, cos\phi\, sin\theta - 0.7071\, sin\phi)\frac{T}{m} \tag{22}$$

$$a_y = (-0.7071\, cos\phi\, sin\theta - 0.7071\, sin\phi)\frac{T}{m} \tag{23}$$

$$a_z = g - (cos\phi\, cos\theta)\frac{T}{m} \tag{24}$$

The states are given by $X = [x,\, v_x,\, y,\, v_y,\, z,\, v_z]^T$ and the control inputs are given by $U = [T,\, \phi,\, \theta]^T$

Find the optimal control input to track a desired trajectory as given below if the initial position and velocity are given by (0,0,0) and (1,-1,0) respectively.

$$x_d(t) = 5\, cos(1.2t), \;\; y_d(t) = 5\, \sin(1.2t), \;\; z_d(t) = -20 \tag{25}$$

The constraint $0 \leq T \leq 20$ has to be satisfied.

Use data-driven approach with a sampling time of 0.02 seconds.
a) Deep Q-Learning (consider discrete action space)
b) DDPG
c) PPO
d) SAC

Q.7): The stat-space model of a quadrotor UAV (Uncrewed Aerial Vehicle) is given below. Here $\mu = 0.05$, m= 1 Kg and g=9.81 m/s^2. Convention for (x, y, z) is given by x facing North, y facing East and z facing Down.

$$\dot{x} = v_x \tag{26}$$

$$\dot{v}_x = a_x - \mu v_x \tag{27}$$

$$\dot{y} = v_y \tag{28}$$

$$\dot{v}_y = a_y - \mu v_y \tag{29}$$

$$\dot{z} = v_z \tag{30}$$

$$\dot{v}_z = a_z - \mu v_z \tag{31}$$

where

$$a_x = (cos\phi\, sin\theta)\frac{T}{m} \tag{32}$$

$$a_y = (-sin\phi)\frac{T}{m} \tag{33}$$

$$a_z = g - (cos\phi\, cos\theta)\frac{T}{m} \tag{34}$$

The states are given by $X = [x,\, v_x,\, y,\, v_y,\, z,\, v_z]^T$ and the control inputs are given by $U = [T,\, \phi,\, \theta]^T$

Find the optimal control input for the UAV of dimensions 0.3 m × 0.3 m × 0.6 m (L×W×H) to pass through a window of dimensions 0.2 m × 0.28 m × 1.0 m (L× W×H). The initial position and velocity of the UAV are given by (0,0,0) and (0,0,0) respectively. The coordinates of the centre of the window are given by (2.1 m, 0, -0.3 m).

The constraint $0 \leq T \leq 20$ has to be satisfied.

Use data-driven approach with a sampling time of 0.02 seconds.
a) Deep Q-Learning (consider discrete action space)
b) DDPG
c) PPO
d) SAC