Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

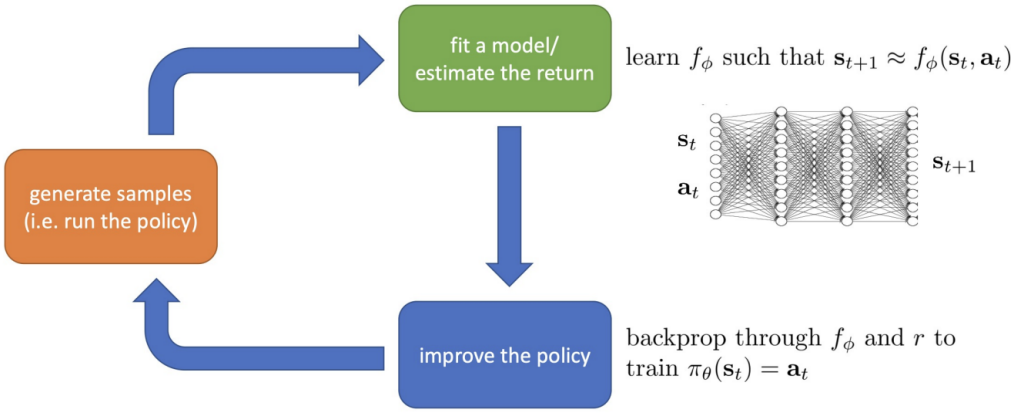
Quiz Solutions - Lecture 1

Solution by : Benyamin Naderi

Q1

Describe the RL Loop completely, clarify the functionality of each block and explain why RL is different from supervised learning?

first of all lets clarify why we need a framework, **Reinforcement learning is a framework for solving control tasks (also called decision problems) by building agents that learn from the environment by interacting with it through trial and error and receiving rewards (positive or negative) as unique feedback.** so lets conclude this way that we aim to build an agent-environment frame work to model our problem. in this framework we as described in the picture first of all we have an agent makes some change in the environment by choosing an action , then evaluates how well this action was executed. form these interactions between agent and environment we collect data (aka trajectories) and the corresponding reward. using this sampled data we can now fit a model that describes us how our environment actually works. this model by estimates the reward to go on by capturing the input data generated by the latter block (state , action), afterwards there is a possibility to learn which action in which state to take and maximize the expected return by doing so and using backpropagation throughout the model we can finally learn a policy model an evaluate taken actions this model outputs a conditional distribution over actions taken in a specific state for lets ensemble what we have leaned so far in the picture below:



as discussed in the first lecture in supervised learning the ground truth is known in advance and the training data is static and iid whereas in reinforcement learning a series of trial and error(search) is performed , the best action(policy) is known in prior data is **dynamic** and **non-iid** and a series of actions is needed in the learning process and they are chosen in the way that the long-term reward is maximized in future. in a nutshell Supervised Learning is about the generalization of the knowledge given by the supervisor (training data) to use in an uncharted area (test data). It is based on instructive feedback where the agent is provided with correct actions (labels) to take given a situation (features).Reinforcement Learning is about learning through interaction by trial-and-error. There is no instructive feedback but only evaluative feedback that evaluates the action taken by an agent by informing how good the action taken was instead of saying the correct action to take.

Q2

A warehouse uses autonomous robots to transport packages from storage shelves to a delivery station. Each robot moves in a grid-like warehouse environment where it can navigate around obstacles and pick up/deliver packages. Your task is to model this problem in an RL framework by defining :

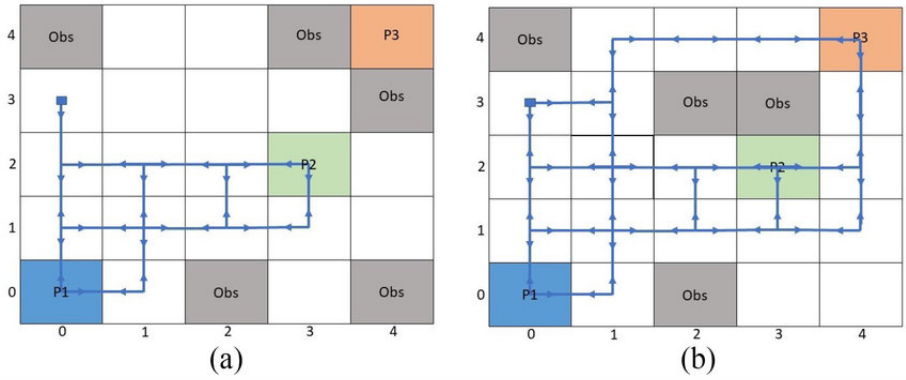
a) Draw a picture describing the warehouse rules please be informed that irrelevant drawings doesn't get any credit. Help: check out the grid world model problem

b) Represent the state space and actions space.

c) What rewards or penalties should be used to encourage efficient package delivery show them on your world model?

d) What would you prefer for this setup value based methods or policy-gradient based explain why?

a)this world model can be described as a grid world model where the states are related to the location of the robot in terms of 2-Dimensional axis in order to visualize this world take a loot at the picture below:



b) the states can be described as a 5 by 5 matrix and in each state there a 4 possible options for action selection:

$$M = \langle S, A, P, R, \gamma \rangle$$

- **State Space** (S):
  The environment consists of a 5×5 grid, so the state space is:

  $$S = \{(x, y) \mid x, y \in \{0, 1, 2, 3, 4\}\}$$

  The total number of states is $|S| = 25$.
- **Action Space** (A):
  The agent can take one of four possible actions:

  $$A = \{\text{up } (\uparrow), \text{down } (\downarrow), \text{left } (\leftarrow), \text{right } (\rightarrow)\}$$

  If an action moves the agent out of bounds, it remains in the same state.
- **Transition Probability** ($P : S \times A \times S \to [0, 1]$):
  The transition function defines the probability of moving from one state to another given an action. In a deterministic setting:

  $$P(s'|s, a) = \begin{cases} 1, & \text{if } s' \text{ is the expected next state given } s \text{ and } a, \\ 0, & \text{otherwise.} \end{cases}$$

c) In a stochastic setting, the agent may move in an unintended direction with some probability.
Reward Function ($R : S \times A \to \mathbb{R}$):
The agent receives rewards based on transitions:

- $R(s, a) = -1$ (default step cost).
- $R(s, a) = +10$ if s' is a goal state.
- $R(s, a) = -5$ if the agent collides with an obstacle.

c) Since the actions space is discrete and sample efficiency is critical value based methods like Q-learning is preferred. it's often more stable compared to policy-based methods.

Q3

What is the main difference between Inverse Reinforcement Learning (IRL) and Imitation Learning?

A) IRL infers the reward function that explains the expert's behavior, while Imitation Learning directly replicates the expert's actions without inferring the underlying reward structure.
B) Both methods require the expert to explicitly provide the reward function for the agent to learn.
C) Imitation Learning cannot generalize to new situations outside the expert's demonstrations, while IRL can extrapolate to unseen states by understanding the reward function.
D) IRL ignores the expert's demonstrations and relies entirely on the agent's trial-and-error exploration to learn the reward function.

Correct Answers: A

Inverse Reinforcement Learning (IRL): The goal is to **infer the reward function** that explains the expert's behavior. Once the reward function is learned, it can be used to train an agent in different settings, allowing for better generalization. **Imitation Learning (IL):** The agent directly learns to **mimic** the expert's actions without trying to infer the reward structure.
Why not the other options?

- **B) Incorrect** – Neither IRL nor IL requires the expert to explicitly provide a reward function; IRL infers it, while IL bypasses it entirely.
- **C) Partially true but misleading** – IL has generalization challenges, but IRL's ability to generalize depends on how well the inferred reward function models real-world dynamics. It doesn't *guarantee* better extrapolation.
- **D) Incorrect** – IRL **relies on expert demonstrations** to infer the reward function; it does not ignore them.

Q4

Given two sampled trajectories of actions and states, what can we infer about the environment?

A) The optimal policy
B) The entire transition dynamics
C) A partial understanding of state transitions
D) The exact reward function

Correct Answers: C

we can observe some transitions, but not the entire transition dynamics of the environment. This gives us a partial understanding of how states change based on actions, but we cannot infer the full transition probabilities or guarantee knowledge of all possible transitions.

Q5

which of the following correctly seperates value based and policy based Reinforcement Learning?

A) value based methods rely on q functions to estimate future rewards, while policy based methods optimize actions directly.

B) policy based methods use q learning, and value based ones use gradient ascent.

C) policy based methods dont need rewards, while value based ones do.

D) value based rl is totally deterministic, while policy based RL is fully stochastic.

Correct Answers: A

Value-based methods estimate the value of states or state-action pairs (like Q-values) to inform future decisions, aiming to maximize reward.

Q6

Suppose a robot is playing soccer! If it only receives a reward when it scores a goal, what problem arises?

A) The robot stays stuck in defense.

B) The reward arrives too late, and the robot doesn't know which actions were useful.

C) The robot attacks excessively.

D) None of the above.

Correct Answers: A

When a robot only receives a reward for scoring a goal, the feedback is sparse and delayed. This makes it difficult for the robot to learn which specific actions contributed to the goal, as the reward signal comes only after a long sequence of actions.

Q7

In the Actor-Critic method, what exactly does the "Critic" do?

A) It directly selects the action.

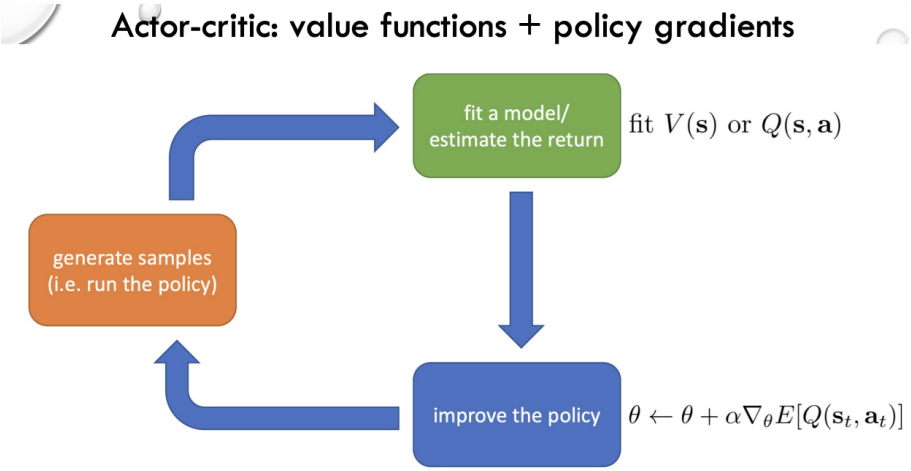
B) It evaluates the Actor's performance and provides feedback.

C) It controls the environment's reward.

D) It is only used to display data.

Correct Answers: B

first of all lets dig into the actor-critic method:



In the Actor-Critic method, the Actor is responsible for selecting actions based on the current policy, while the Critic evaluates the actions taken by the Actor by estimating the value function (or sometimes the advantage function). The Critic provides feedback (usually in the form of a value function or TD-error) to help the Actor improve its policy over time. The Critic essentially guides the Actor by telling it how good or bad its actions were in terms of expected future rewards.

Q8

Explain the RL objective in value based methods in terms the value function and first state distribution and describe how it is related the expected sum of discounted rewards.

- **Value Function** ($V^\pi(s)$): The expected sum of discounted rewards starting from state s under policy π:

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s \right],$$

- **Initial State Distribution** ($\mu(s)$): The probability distribution over starting states s.

The **RL Objective** is to find a policy π^* that maximizes:

$$J(\pi) = \mathbb{E}_{s\sim\mu} \left[V^\pi(s) \right].$$

This is equivalent to optimizing the average value of $V^\pi(s)$ across states sampled from $\mu(s)$.

Connection to Expected Discounted Rewards

The objective $J(\pi)$ expands to:

$$J(\pi) = \mathbb{E}_{s\sim\mu} \left[\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s \right] \right],$$

which simplifies to the **expected total discounted reward** over trajectories:

$$J(\pi) = \mathbb{E}_{\pi,\mu} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right].$$

Q9

In one word, what does Imitation Learning do?

A) Predict!

B) Infer!

C) Exploit!

D) Copy!

Correct Answers: D! it simply copies :)

Imitation Learning learns a policy by mimicking or copying the behavior of an expert, typically by observing the expert's actions and then trying to replicate them in similar situations.