Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 2
Solution by : Arshia Gharooni

Q1

Consider an environment where state transitions are deterministic but the agent's policy is stochastic. What is the primary source of uncertainty in the agent's outcomes?

A) The uncertainty stems solely from unpredictable state transitions.
B) The uncertainty is due to random action choices by the agent, even though the environment behaves deterministically.
C) There is no uncertainty since the environment is fully predictable.
D) Uncertainty arises from unexpected changes in the reward function.

Correct Answers: B

Explanation:

- **Deterministic Environment:** The environment's next state is fully determined once the agent picks an action, with no randomness in state transitions.
- **Stochastic Policy:** The agent's policy selects actions according to a probability distribution, meaning the same state can lead to different actions.
- **Source of Uncertainty:** Since the environment is deterministic, all uncertainty in the outcome arises from the agent's random action choices.

Q2

Consider the infinite-horizon discounted return defined as:

$$G = \sum_{t=0}^{\infty} \gamma^t R_t,$$

where:

- $0 \leq \gamma < 1$ **is the discount factor, and**
- R_t **is the reward at time step** t **which is bounded (i.e., there exists a constant** $M > 0$ **such that** $|R_t| \leq M$ **for all** t**).**

Explain why G **converges almost surely.**

A) Because the Markov property limits the impact of past rewards.
B) Because the geometric series on γ converges to $\frac{1}{1-\gamma}$, which, combined with the bounded reward, ensures that the sum is finite.
C) Because state transitions eventually become deterministic.
D) Because convergence is guaranteed regardless of γ when rewards are bounded.

Correct Answers: B

Explanation: Because $\gamma < 1$ forms a convergent geometric series and the reward is bounded, each term $\gamma^t R_t$ is finite and the overall infinite sum converges almost surely.

Q3

Imagine an agent interacting with an environment where both the outcomes of its actions and its own action choices are uncertain. Why is it conceptually necessary to consider two layers of "averaging" (or expectation) when evaluating its performance?

A) Because the agent's choices and the environment's responses each add a separate layer of randomness.
B) Because one layer of expectation is enough; the second is redundant.
C) Because only the environment's randomness matters, not the agent's decisions.
D) Because averaging cancels out the uncertainties entirely.

Correct Answers: A

There are two distinct sources of randomness in the interaction:

1. **The agent's stochastic policy**: Determines which action gets chosen in a given state.
2. **The environment's stochastic transitions and rewards**: Determines how the environment responds to the agent's action.

Q4

Consider that you are designing an agent for an environment with sparse, significantly delayed rewards. You want the agent to effectively consider at least 150 future time steps when evaluating its actions. Which minimum value of γ satisfies this requirement?**In many RL tasks, the effective planning horizon is approximated by**

$$H_{\textbf{eff}} = \frac{1}{1-\gamma}.$$

Consider that you are designing an agent for an environment with sparse, significantly delayed rewards. You want the agent to effectively consider at least 150 future time steps when evaluating its actions. Which minimum value of γ satisfies this requirement?

A) $\gamma = 0.9$
B) $\gamma = 0.95$
C) $\gamma = 0.99$
D) $\gamma = 0.9933$

Correct Answers: D

We are given that the effective planning horizon in Reinforcement Learning (RL) is approximated by

$$H_{\text{eff}} = \frac{1}{1-\gamma},$$

where γ is the discount factor. We want the agent to consider at least 150 future time steps, i.e.,

$$H_{\text{eff}} \geq 150.$$

Substituting $H_{\text{eff}} = \frac{1}{1-\gamma}$, we get:

$$\frac{1}{1-\gamma} \geq 150.$$

Solving for γ:

$$1-\gamma \leq \frac{1}{150} \quad \Longrightarrow \quad \gamma \geq 1 - \frac{1}{150} = \frac{149}{150} \approx 0.9933.$$

Thus, the minimum value of γ that satisfies the requirement of considering at least 150 future time steps is

$$\boxed{0.9933}.$$

Q5

In RLHF, the reward model is trained using human feedback, usually through pairwise comparisons of different outputs, to generate a reward signal for the language model. this signal then guides the model's behavior through reinforcement learning. which of the following challenges in designing this reward model is most likely to make the language model learn behaviors that don't fully align with real human preferences?

A) Overfitting to limited feedback
B) Underfitting due to high variance
C) Excessive computational overhead
D) Reduced exploration incentives

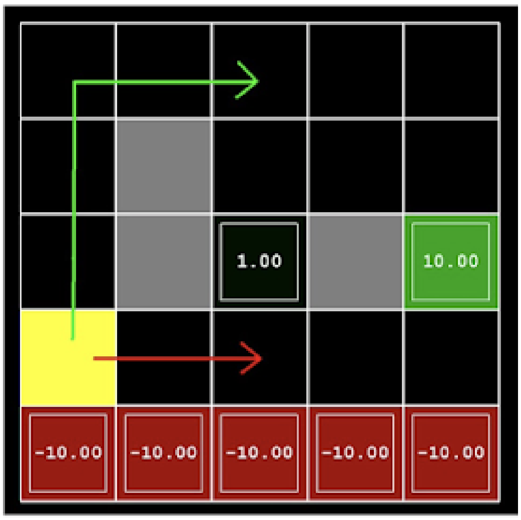
Correct Answers: A

Among the four challenges, **overfitting to limited feedback** poses the biggest risk for misalignment. If the reward model is trained on too few or unrepresentative feedback samples, it may latch onto narrow or coincidental patterns—sometimes referred to as *reward hacking*. As a result, the language model ends up optimizing for these quirks rather than truly reflecting human values. This mismatch can cause the model to produce behaviors that appear acceptable during training but fail to align with real-world human preferences.

Q6

Assuming the mentioned world model in the class, consider that any non-zero reward ends the episode and hitting the walls causes you to stay in the starting position. Which one gives a better policy if we act greedily and want to terminate as soon as possible?

A) noise = 0.5, γ = 0.9
B) noise = 0.1, γ = 0.1
C) noise = 0.1, γ = 0.9
D) noise = 0.5, γ = 0.1



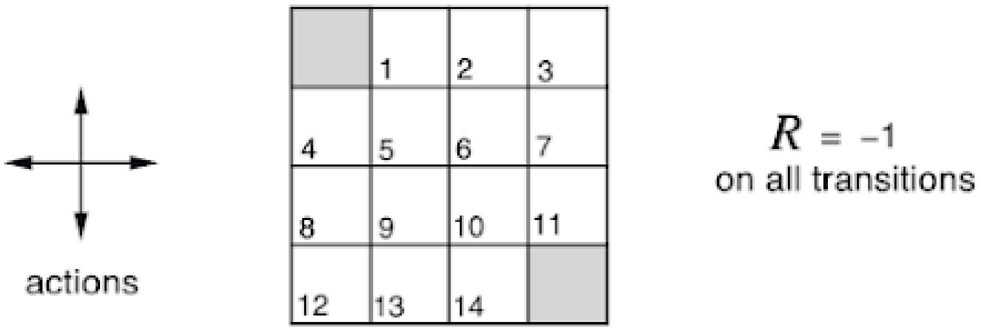
Correct Answers: B

In this gridworld, the key trade-off is between the risk of slipping into -10 (controlled by "noise") and how much we discount future rewards (controlled by γ). High noise makes aiming for $+10$ risky, since a slip might lead to -10. A small γ heavily discounts future rewards, making the agent grab $+1$ quickly.
However, if we want to act greedily and end on the high reward, the best option is noise $= 0.1$ and $\gamma = 1.0$. Low noise keeps the path to $+10$ safe, and with $\gamma = 1.0$ we do not discount future rewards, so we prefer $+10$ over $+1$. This combination typically yields the highest expected return while ending quickly once $+10$ is reached.

Q7

*** Consider this world model, any transition achieves reward $R = -1$, until reaching the terminal states (shaded areas). Which MDP gives us a better optimal policy?**

1. All actions are equally likely, $\gamma = 0.9$.
2. $P(\text{down}|s) = P(\text{up}|s) = 0.2$, $\gamma = 0.9$.
3. All actions are equally likely, $\gamma = 0.1$.
4. $P(\text{down}|s) = P(\text{up}|s) = 0.3$, $\gamma = 0.9$.



Correct Answers: A

With equal action probabilities and a high discount factor ($\gamma = 0.9$), the agent strongly prioritizes reaching terminal states quickly. This results in fewer unnecessary moves and a higher (less negative) expected return compared to cases with a lower discount ($\gamma = 0.1$) or biased action probabilities causing excessive vertical movements.