Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 1

Solution by : Benyamin Naderi

Q1

Check False Statements.

A) In Model-Based RL actions are taken by optimizing a policy network.
B) Model-Based methods just needs transition dynamics to solve the optimization problem.
C) Model-Based methods learns transition dynamics by sampling trajectories.
D) Model-Free methods are better than Model-Based methods in terms of sample efficiency.

A) In Model-Based RL actions are taken by optimizing a policy network. It's wrong , in Model-Based RL there is a planning for a sequence of actions chosen w.r.t transitoin dynamics by solving a constrained optimization problem for maximizing return. and the collected data is used for training the transition model.

B) Model-Based methods just needs transition dynamics to solve the optimization problem. It's Wrong , we need first state distribution as well.

C) Model-Based methods learns transition dynamics by sampling trajectories. True , in Model-Based methods by sampling trajectories we can obtain a surrogate model of transition dynamics, afterwards figure out choosing actions

D) Model-Free methods are better than Model-Based methods in terms of sample efficiency. It's wrong , Vice-versa , Model-Based methods are much more sample efficient since sampled trajectories estimate the transition dynamics and we don't need to train a policy network anymore.

Q2

Imagine training a robot to walk in a simulation. The robot must explore new movements (e.g., hopping, shuffling) but also exploit efficient gaits to maximize speed. SAC is a popular algorithm for this task. Why is SAC particularly effective here?

A) It uses a fixed exploration rate (e.g., -greedy) to balance randomness.

B) It maximizes expected reward while also encouraging random actions via entropy bonuses.

C) It uses a single Q-network to estimate values, reducing computational cost.

Correct Answers: B

B is True. SAC adds an entropy term to the reward, incentivizing the robot to try diverse actions (exploration) while still optimizing for high rewards (exploitation). This avoids getting stuck in suboptimal gaits.

Q3

A delivery drone is tasked with navigating to a destination.

- **Drone A:** Plans its path using a pre-loaded map and weather forecast but ignores real-time wind changes.
- **Drone B:** Continuously adjusts its route using live sensor data and a physics model to predict turbulence.

A) Drone A is model-based closed-loop; it prioritizes speed over accuracy.
B) Drone B is model-based open-loop; it relies on initial predictions.
C) Drone A is model-based open-loop; it follows a fixed plan without feedback.
D) Drone B is model-based closed-loop; it uses real-time data and a model to adapt.

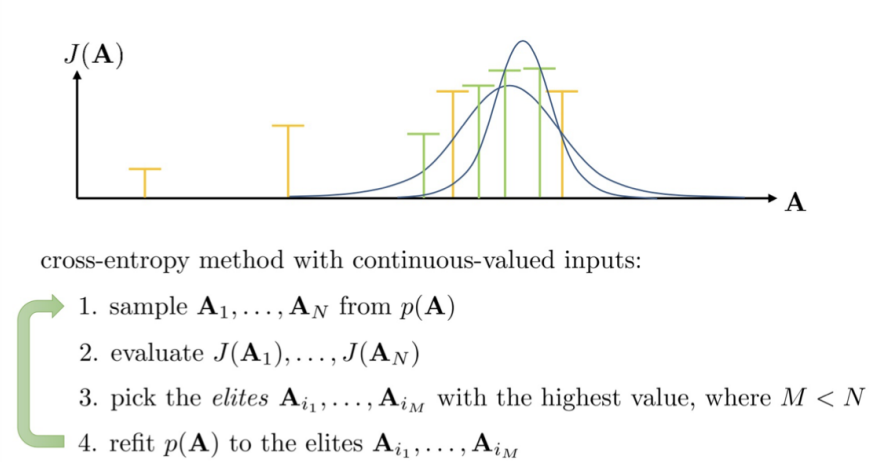
Correct Answers: D

The correct answer is D since , Drone B uses feedback (live sensors) and a model (physics for turbulence predictions) to dynamically adjust, defining model-based closed-loop control.

- A/C) Drone A lacks feedback, making it open-loop, even with a model.
- B) Drone B is closed-loop, not open-loop.

Q4

Explain how CEM guesses a probable optimizer for a stochastic optimization problem for Closed-loop setup in model-based RL?



The Cross-Entropy Method (CEM) in a closed-loop model-based reinforcement learning (RL) setup iteratively refines a probability distribution over actions to efficiently solve stochastic optimization problems. These actions represent potential strategies for interacting with the environment. The top M action sequences (elites) with the highest values are retained. These elites guide the search toward high-reward regions of the action space. The distribution is updated to match the statistics of the elites. In the next timestep, the agent repeats the process from the current state of the environment, incorporating real-time feedback. This closed-loop mechanism allows continuous policy refinement based on the latest observations.

Q5

Explain why in PPO objective it takes a clips over the Importance sampling term?

$$\theta' = \arg\max_{\theta'} \mathbb{E}_{s\sim\mu_\theta, a\sim\pi_\theta}\left[\min\left(\frac{\pi_{\theta'}(a|s)}{\pi_\theta(a|s)}A^{\pi_\theta}(s,a),\ \text{clip}\left(\frac{\pi_{\theta'}(a|s)}{\pi_\theta(a|s)}, 1-\epsilon, 1+\epsilon\right)A^{\pi_\theta}(s,a)\right)\right]$$

This prevents excessively large or harmful policy changes, ensuring stable training by balancing exploration and exploitation while avoiding performance collapse from overconfident updates.

Q6

Consider the SAC objective function, which maximizes both expected reward and policy entropy: What happens to the optimal policy as grows higher?

A) The policy becomes deterministic, favoring high-reward actions exclusively.

B) The policy converges to a uniform distribution over actions, ignoring rewards.

C) The policy prioritizes maximizing entropy while ignoring rewards entirely.

D) The policy becomes unstable due to conflicting reward and entropy terms.

Correct Answers: B

as it grows the entropy term dominates the objective. To maximize entropy, the policy becomes uniform (all actions equally likely), effectively ignoring rewards.

- A) Deterministic policies minimize entropy, which contradicts
- C) While entropy is prioritized, the policy doesn't "ignore" rewards—it just becomes indifferent to them.
- D) The trade-off is well-defined; instability arises only with poor tuning.