Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 16

Solution by : Amirhossein Asadi

Q1

In the derivation of performance bounds for policy improvement, how does the advantage function $A^{\pi}(s,a)$ facilitate the estimation of the expected return difference between two policies π and π'?

A) By computing the return difference using total value estimates, without explicitly accounting for state-action visitation patterns of either policy.
B) By acting as a reference for action-value comparisons, allowing variance reduction in policy evaluation during learning updates.
C) By expressing the benefit of specific actions under π, supporting estimation of return difference using $A^{\pi}(s,a)$ and the distribution of π'.
D) By introducing entropy-based criteria that influence exploration behavior, indirectly affecting return estimation across policies.

Correct Answer: C

The advantage function $A^{\pi}(s,a)$ plays a critical role in formalizing the difference in expected return between two policies π and π'. Specifically, it allows the decomposition of the return difference $J(\pi') - J(\pi)$ into an expectation over the advantage values of policy π, weighted by the visitation distribution and action probabilities of policy π'. This is formalized as:

$$J(\pi') - J(\pi) = \mathbb{E}_{s\sim d^{\pi'}, a\sim\pi'}\left[A^{\pi}(s,a)\right]$$

Here, $d^{\pi'}$ denotes the discounted state visitation distribution under policy π'. The advantage function quantifies how much better or worse an action is compared to the average behavior of policy π, enabling a tractable and interpretable estimation of policy improvement bounds. This formulation is central in algorithms such as TRPO and PPO, which rely on performance difference lemmas to derive theoretically grounded policy update rules.

Q2

Why might one prefer using the Kullback-Leibler (KL) divergence over total variation distance when measuring the difference between two probability distributions in RL?

A) KL divergence is symmetric, whereas total variation distance is not.
B) KL divergence is easier to compute in high-dimensional spaces.
C) KL divergence penalizes differences in low-probability events more heavily.
D) Total variation distance cannot be used with continuous action spaces.

Correct Answer: C

In RL, KL divergence is often preferred over total variation distance for measuring the difference between probability distributions, particularly in the context of policy optimization. One key reason is that KL divergence assigns a greater penalty to mismatches in low-probability events. This property is useful in guiding conservative updates in policy-based algorithms, as it discourages large deviations in action probabilities, especially in regions where the reference policy assigns low probability. This characteristic is particularly exploited in algorithms like TRPO and PPO, where the KL divergence is used to bound the deviation between successive policies, promoting stable and reliable learning.

Q3

Which of the following algorithms combines both value-based and policy-based methods in reinforcement learning?

A) Q-learning
B) Actor-Critic
C) Monte Carlo methods
D) Deep Q-Network

Correct Answer: B

The Actor-Critic algorithm integrates both value-based and policy-based approaches in RL. In this framework, the *actor* is responsible for selecting actions based on a policy, while the *critic* evaluates these actions by estimating value functions. This combination allows the algorithm to leverage the strengths of both methods: the actor updates the policy in the direction suggested by the critic, and the critic updates the value function based on feedback from the environment. This synergy facilitates more stable and efficient learning, especially in environments with continuous action spaces.

Q4

Discuss the importance of bounding the total variation distance between successive policies in policy improvement. How does this constraint affect the convergence and performance of RL algorithms?

Bounding the TVD between successive policies is crucial for ensuring stable and reliable policy improvement in RL. Large deviations between consecutive policies can lead to significant changes in the distribution of visited states and the actions taken, potentially destabilizing learning and causing divergence. By constraining this distance, algorithms promote smoother transitions, preserving previously learned behaviors and enhancing convergence stability. This approach is particularly vital in trust-region methods like TRPO and PPO, where constraints on policy divergence ensure monotonic improvement and prevent performance collapse.

Q5

How can coupling methods and transport inequalities be used to derive mixing-time bounds for Markov chains in RL, particularly leveraging total variation distance?

Coupling methods bound the mixing time by measuring how quickly two coupled Markov chains converge. The total variation distance between the chain's distribution at time t and its stationary distribution is bounded by the probability that the chains have not coupled. Transport inequalities like Pinsker's inequality help relate total variation to KL divergence:

$$D_{TV}(P, Q) \leq \sqrt{\frac{1}{2} D_{KL}(P \| Q)}$$

This connection allows us to derive convergence bounds using more tractable quantities like KL, which are widely used in RL algorithms.

Q6

In analyzing the TVD between two probability distributions p and q, why is it beneficial to define a joint distribution $\gamma(s, s')$ such that $\mathbb{P}(s = s') = 1 - \epsilon$?

A) To facilitate the computation of the Kullback-Leibler divergence between p and q by aligning their supports through a shared joint distribution.
B) To construct a coupling that directly relates the total variation distance to the probability of disagreement between s and s' under γ.
C) To simplify the optimization landscape in policy gradient methods by ensuring smoother transitions between state distributions.
D) To ensure that the support of p and q is identical, enabling straightforward comparison and analysis of their differences.

Correct Answer: B

Defining a joint distribution $\gamma(s, s')$ such that $\mathbb{P}(s = s') = 1-\epsilon$ is a classic technique in probability theory known as *coupling*. This approach is instrumental in relating the total variation distance (TVD) between two distributions p and q to the probability that two random variables drawn from these distributions differ. Specifically, the TVD can be expressed as:

$$\delta(p, q) = \inf_{\gamma} \mathbb{P}_{(s,s') \sim \gamma}(s \neq s')$$

where the infimum is taken over all couplings γ of p and q. By constructing such a coupling where $\mathbb{P}(s = s') = 1 - \epsilon$, we directly relate the TVD to the probability of disagreement between s and s' under γ. This method is particularly useful in deriving bounds on mixing times of Markov chains and in analyzing convergence properties in reinforcement learning algorithms.