Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 13
Solution by : Amirhossein Asadi

Q1

In the context of MAB problems, both UCB and Thompson Sampling algorithms aim to balance exploration and exploitation. However, they differ fundamentally in their approaches to managing uncertainty. Which of the following statements best captures this distinction?

A) UCB and Thompson Sampling both rely on Bayesian inference; however, UCB updates beliefs about action rewards using conjugate priors, whereas Thompson Sampling employs non-conjugate priors, resulting in different computational complexities.
B) UCB prioritizes actions with the highest upper confidence bounds, leading to a more aggressive exploration strategy, while Thompson Sampling selects actions based solely on prior knowledge, potentially limiting exploration.
C) UCB assumes a fixed reward distribution for each action and updates estimates deterministically, whereas Thompson Sampling assumes dynamic reward distributions that evolve based on observed data.
D) UCB constructs deterministic confidence intervals to estimate the potential of each action, while Thompson Sampling utilizes stochastic sampling from a posterior distribution to guide action selection.

Correct Answer: D

UCB and Thompson Sampling are both strategies designed to address the exploration-exploitation trade-off in decision-making scenarios like the MAB problem. UCB achieves this by calculating upper confidence bounds for the expected rewards of each action, effectively creating a deterministic confidence interval. Actions are then selected based on these bounds, favoring those with higher potential rewards. In contrast, Thompson Sampling adopts a probabilistic approach by maintaining a posterior distribution over the expected rewards and selecting actions based on random samples drawn from these distributions. This fundamental difference highlights how UCB relies on deterministic estimations, while Thompson Sampling leverages stochastic sampling to manage uncertainty.

Q2

In the context of Thompson Sampling for MAB problems, employing a conjugate prior is particularly advantageous because:

A) It ensures that the posterior distribution remains in the same family as the prior, facilitating analytical tractability and simplifying the update process.
B) It allows for the use of non-informative priors, thereby eliminating bias in the initial stages of learning.
C) It enables the modeling of complex, multimodal reward distributions without increasing computational complexity.
D) It guarantees faster convergence to the true reward distribution by reducing variance in posterior estimates.

Correct Answer: A

In Bayesian inference, a conjugate prior is chosen such that the resulting posterior distribution is of the same functional form as the prior. This property is particularly beneficial in Thompson Sampling, as it allows for straightforward analytical updates of the posterior distribution after observing new data. For example, when dealing with Bernoulli rewards, selecting a Beta distribution as the prior ensures that the posterior remains a Beta distribution, simplifying computations and maintaining consistency in the probabilistic model. This approach enhances computational efficiency and supports real-time learning in dynamic environments.

Q3

How does Thompson Sampling balance the exploration-exploitation trade-off in multi-armed bandit problems?

By sampling from the posterior distribution of each action's reward and selecting the action with the highest sampled value, integrating both exploration and exploitation.

Q4

In RL, the effective application of the Bellman Optimality Equation can be hindered under certain conditions. Which of the following scenarios presents the most significant challenge?

A) The environment has a finite state space with deterministic transitions and known reward functions.
B) The environment's transition dynamics and reward functions are unknown and must be estimated through interaction.
C) The agent operates with a high discount factor (1), placing substantial emphasis on future rewards over immediate ones.
D) The agent's policy is stochastic, assigning probabilities to multiple actions in each state rather than selecting a single deterministic action.

Correct Answer: B

The Bellman Optimality Equation relies on precise knowledge of the environment's transition probabilities and reward functions to compute optimal policies. When these components are not explicitly known, they must be estimated through interactions with the environment, introducing uncertainty and potential inaccuracies. This estimation process complicates the direct application of the Bellman Optimality Equation and often necessitates the use of approximate methods or model-free RL algorithms to learn optimal policies.

Q5

In a MDP with finite states and actions, Assuming the transition dynamics and reward functions are known, which of the following statements best describes the implication of the Bellman Optimality Equation for policy iteration methods?

A) Policy iteration will converge to the optimal policy in a finite number of steps, as each iteration ensures a guaranteed improvement in the policy's expected return and exploits the structure provided by the Bellman equation.
B) Policy iteration may converge to a locally optimal solution, especially in environments with large or complex state spaces, where the Bellman Optimality Equation does not always guarantee global convergence without further assumptions.
C) The Bellman Optimality Equation allows policy iteration to proceed by defining value-based updates, but also enables direct derivation of the optimal policy when transition and reward models are fully known and deterministic.
D) Policy iteration uses the Bellman Optimality Equation for iterative policy updates, but convergence depends on factors like the discount rate and starting policy.

Correct Answer: D

Policy iteration involves iterative processes of policy evaluation and improvement. While it often converges to the optimal policy in finite MDPs, this convergence is influenced by factors such as the discount factor () and the initial policy choice. A discount factor close to 1 places more emphasis on future rewards, which can affect convergence speed and stability. Additionally, the choice of initial policy can impact the number of iterations required for convergence. Therefore, while the Bellman Optimality Equation provides the foundation for policy iteration, practical convergence depends on these specific conditions.