Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Summary of Lecture 13: Value-Based Theory

Summarized By: Amirhossein Asadi

Thompson sampling

- After discussing action selection strategies such as Epsilon-Greedy and Upper Confidence Bound (UCB), we now consider **Thompson Sampling**, an approach to decision-making under uncertainty. Unlike methods based on point estimates of expected rewards, Thompson Sampling maintains a probability distribution over the potential reward of each action.This method has been shown to perform effectively in practice, often achieving near-optimal regret in stochastic environments.
- In essence, Thompson Sampling relies on a **Bayesian estimation**, where we aim to estimate the expected reward for each action k, denoted as $\theta_k = \mathbb{E}[R^k]$. Given the observed rewards from past actions, we compute the **posterior distribution** over θ_k, expressed as $P(\theta_k \mid R_t^{a(1)} = r_1, \ldots, R_t^{a(T)} = r_T)$, which captures our updated belief about the expected reward after observing data.
- Thompson Sampling is conceptually similar to the UCB approach in that both aim to balance exploration and exploitation. However, while UCB constructs a confidence interval around point estimates to guide action selection, Thompson Sampling directly models the entire posterior distribution of the expected rewards, allowing for a more probabilistic and principled exploration strategy.
- After estimating the posterior distributions of the expected rewards, the simplest action selection method is **sampling**: draw one sample from each of the k distributions and select the action with the highest sample.
- Computing the exact posterior distribution is often intractable in real-world scenarios. However, when the prior and likelihood are **conjugate**, the posterior belongs to the same family as the prior, making the update process analytically straightforward. A common example is using a Beta prior with a Bernoulli likelihood.
- Unlike UCB, which takes a more rigid or deterministic approach to uncertainty through confidence bounds, Thompson Sampling deals with uncertainty in a much **softer and more probabilistic** manner. In fact, if Bayesian inference were computationally trivial, Thompson Sampling would be an ideal strategy for action selection.

Bellman's Optimality Equation

- Bellman's Optimality Equation provides a fundamental recursive characterization of the value function under an optimal policy. It serves as the cornerstone for many algorithms in RL.

  Assume a **stochastic reward function**:

$$\Pr(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a), \quad \forall s, s' \in \mathcal{S},\ r \in \mathcal{R},\ a \in \mathcal{A}$$

  This is abbreviated as:

$$p(s', r \mid s, a)$$

- The optimal action-value function $q_*(s, a)$ is defined as:

$$\begin{aligned} q_*(s,a) &= \max_\pi \mathbb{E}\left[G_t \mid S_t = s, A_t = a\right] \\ &= \max_\pi \mathbb{E}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right] \\ &= \mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a\right] + \gamma \max_\pi \mathbb{E}\left[G_{t+1} \mid S_t = s, A_t = a\right] \end{aligned}$$

Define the return G_t as:

$$G_t = \sum_{t'=t+1}^{\infty} \gamma^{t'-t-1} R_{t'}$$

- For the expected immediate reward:

$$\mathbb{E}[R_{t+1} \mid S_t = s, A_t = a] = \sum_r \sum_{s'} r \cdot p(s', r \mid s, a)$$

Recall the law of total expectation:

$$\mathbb{E}[f(X, Y)] = \sum_x \mathbb{E}[f(X, Y) \mid X = x] \cdot \mathbb{P}(X = x)$$

For the expected return:

$$\begin{aligned}
\mathbb{E}[G_{t+1} \mid S_t = s, A_t = a] &= \sum_{s',a'} p(s', a' \mid s, a) \cdot \mathbb{E}[G_{t+1} \mid S_{t+1} = s', A_{t+1} = a'] \\
&= \sum_{s',a'} p(s' \mid s, a) \cdot p(a' \mid s', s, a) \cdot \mathbb{E}[G_{t+1} \mid S_{t+1} = s', A_{t+1} = a'] \\
&= \sum_{s',a'} p(s' \mid s, a) \cdot \pi(a' \mid s') \cdot q_\pi(s', a') \\
&= \sum_{s'} p(s' \mid s, a) \sum_{a'} \pi(a' \mid s') \cdot q_\pi(s', a')
\end{aligned}$$

- Since the optimal policy chooses the action that maximizes the action-value, we can move the $\max$ inside the summation:

$$q_*(s, a) = \sum_r r \sum_{s'} p(s', r \mid s, a) + \gamma \max_\pi \sum_{s'} p(s' \mid s, a) \sum_{a'} \pi(a' \mid s') q_\pi(s', a')$$

$$\Rightarrow \quad q_*(s, a) = \sum_r r \sum_{s'} p(s', r \mid s, a) + \gamma \max_\pi \sum_{s'} p(s' \mid s, a) \max_{a'} q_\pi(s', a')$$

- The expression $\sum_{a'} \pi(a' \mid s') q_\pi(s', a')$ is a **convex combination** of action-values, since $\pi(a' \mid s') \in [0, 1]$ and $\sum_{a'} \pi(a' \mid s') = 1$.

By the property of convex combinations:

$$\sum_{a'} \pi(a' \mid s') q_\pi(s', a') \leq \max_{a'} q_\pi(s', a')$$

So, the maximum over all policies is achieved by a deterministic policy that selects the action with the highest q-value:

$$\max_\pi \sum_{a'} \pi(a' \mid s') q_\pi(s', a') = \max_{a'} q_*(s', a')$$