DRL - HW 2

Nadav Shoham 315789115
Uri Rusanov

December 14, 2024

1 Monte-Carlo Policy Gradient (REINFORCE)

1.1 What does the value of the advantage estimate reflect?

The advantage estimate reflects how much better or worse taking a particular action in a given state is compared to the expected outcome defined by the baseline (typically the value function). It quantifies the relative benefit of choosing that action.

1.2 Why is it better to follow the gradient computed with the advantage estimate instead of just the return itself?

Using the advantage estimate reduces the variance in the policy gradient updates without introducing bias. Subtracting the baseline (e.g., value function) focuses the updates on meaningful deviations from the baseline, ensuring that updates are more stable and effective.

1.3 What is the prerequisite condition for the equation to hold true?

The baseline $b(s)$ must be an unbiased estimate of the return, ensuring that $\mathbb{E}_{\pi_\theta}[b(s)]$ is well-defined.

Proof The equation holds Expand the expectation over the policy distribution:

$$\mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a \mid s) b(s)\right] = \sum_s d^\pi \sum_a \pi_\theta(a \mid s) \nabla_\theta \log \pi_\theta(a \mid s) b(s).$$

Where d^π is the stationary distribution of the policy. Using the property of the log-gradient:

$$\nabla_\theta \log \pi_\theta(a \mid s) \pi_\theta(a \mid s) = \nabla_\theta \pi_\theta(a \mid s),$$

the sum becomes:

$$\sum_s d^\pi b(s) \sum_a \nabla_\theta \pi_\theta(a \mid s).$$

Probabilities sum to 1 so their gradients sum to 0:

$$\sum_a \nabla_\theta \pi_\theta(a \mid s) = 0.$$

Thus, the entire term sums to 0:

$$\mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t) b(s_t)\right] = 0.$$

2 Advantage Actor-Critic

2.1 Why is using the TD-error of the value function practically the same as using the advantage estimate?

The TD-error is an unbiased estimate of the advantage function A_t.
Over multiple steps, the expectation of the TD-error aligns with the advantage estimate. Specifically:

$$\mathbb{E}[\delta_t | s_t, a_t] = A(s_t, a_t),$$

Proof

To prove the equivalence, we expand and analyze both expressions:

$$A(s_t, a_t) = Q(s_t, a_t) - V(s_t) = \mathbb{E}[r_t + \gamma V(s_{t+1})|s_t, a_t] - V(s_t).$$
$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t).$$

Taking the expectation of δ_t conditioned on s_t and a_t:

$$\begin{aligned}\mathbb{E}[\delta_t|s_t, a_t] &= \mathbb{E}[r_t|s_t, a_t + \gamma V(s_{t+1})|s_t, a_t - V(s_t)].\\ &= \mathbb{E}[r_t + \gamma V(s_{t+1})|s_t, a_t] - \mathbb{E}[V(s_t)].\end{aligned}$$

Substituting $\mathbb{E}[r_t + \gamma V(s_{t+1})|s_t, a_t] = Q(s_t, a_t)$ and $\mathbb{E}[V(s_t)] = V(s_t)$, we get:

$$\mathbb{E}[\delta_t|s_t, a_t] = Q(s_t, a_t) - V(s_t) = A(s_t, a_t).$$

2.2 Explain the actor and the critic roles in the model.

The actor parameterizes the policy and is responsible for selecting actions based on the current policy. It optimizes the policy directly using feedback. The critic evaluates the actions chosen by estimating the value function $V(s_t)$. It provides feedback to the actor by computing the TD-error or advantage, guiding the policy updates.