RL in Self-Driving Cars

Assignment 1

Snigdha Reddy

June 2025

1 Task 1

1. Q-Learning

Goal: Learn the optimal action-value function $Q^*(s, a)$, which estimates the maximum expected future reward starting from state s, taking action a, and following the optimal policy thereafter.

Update Rule:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right]$$

where α is the learning rate and γ is the discount factor.

Type: Off-policy, model-free.

2. Monte Carlo Control

Goal: Estimate the value function $Q^\pi(s, a)$ using full-episode returns and improve the policy iteratively.

Update Rule:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[G_t - Q(s, a) \right]$$

where $G_t = \sum_{k=0}^{T-t-1} \gamma^k r_{t+k+1}$ is the return.

Type: On-policy or off-policy, model-free, requires episodic tasks.

3. PPO (Proximal Policy Optimization)

Goal: Update a parameterized policy π_θ using a clipped surrogate objective to avoid large policy updates.

Objective Function:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right]$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$, and $\hat{A}_t$ is the advantage estimate.

Type: On-policy, policy gradient with trust-region constraints.

4. DDPG (Deep Deterministic Policy Gradient)

Goal: Learn a deterministic policy $\mu(s)$ for continuous action spaces using actor-critic architecture.

Critic Update:

$$Q(s_t, a_t) \leftarrow r_t + \gamma Q'(s_{t+1}, \mu'(s_{t+1}))$$

Actor Update (Policy Gradient):

$$\nabla_\theta J \approx \mathbb{E}\left[\nabla_a Q(s,a)\big|_{a=\mu(s)} \nabla_\theta \mu(s)\right]$$

Type: Off-policy, uses target networks and experience replay.

5. SAC (Soft Actor-Critic)

Goal: Learn a stochastic policy that maximizes both expected reward and entropy to encourage exploration.

Objective:

$$J(\pi) = \sum_t \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}\left[r(s_t,a_t) + \alpha \mathcal{H}(\pi(\cdot|s_t))\right]$$

where $\mathcal{H}(\pi) = -\mathbb{E}_{a\sim\pi}[\log \pi(a|s)]$ is the entropy term.

Type: Off-policy, entropy-regularized, continuous control.