동적계획법과 강화학습 강의 노트

작성자 : 황우진(2021314074)

작성일시 : 2022.10.06

5주차 강의노트

Model Free Policy Control-SARSA, Q-Learning

Model Free Policy Iteration

① Policy evaluation

- Estimate v_π by Monte-Carlo policy evaluation : $Q = q_\pi$

② Policy improvement

- ϵ-greedy policy improvement

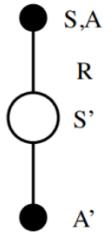
ϵ-Greedy Exploration

- ϵ : random action
- $1-\epsilon$: greedy action
- Improvement will go $v_{\pi'}(s) \geq v_\pi(s)$

GLIE(Greedy in the Limit with Infinite Exploration)

- $\lim_{k\to\infty} N_k(s,a) = \infty$: all state-action pairs explore infinite times
- ϵ reduces to zero at $\epsilon_k = \frac{1}{k}$

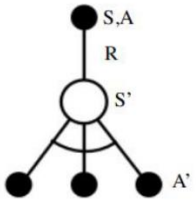
On-Policy Learning (SARSA)



$$Q(S,A) \leftarrow Q(S,A) + \alpha(R + \gamma Q(S',A') - Q(S,A))$$

- SARSA converges to the optimal action-value function under on-line policy

Off-Policy Learning (Q-Learning)



$$Q(S,A) \leftarrow (1-\alpha)Q(S,A) + \alpha(R + \gamma \max_{a'} Q(S',a'))$$

- No importance sampling is required
- Choose next action with behaviour policy $A_{t+1} \sim \mu(\cdot\,|S_t)$
- But consider alternative successor action $A' \sim \pi(\cdot\,|S_t)$
- Q-learning control converges to the optimal action-value function under off-line policy

[참고문헌]

- Sutton, Richard S., and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.