동적계획법과 강화학습 강의 노트

작성자 : 박영익(2021314079)

작성일시 : 2022.11.17

Lecture 9. Policy Gradient Method

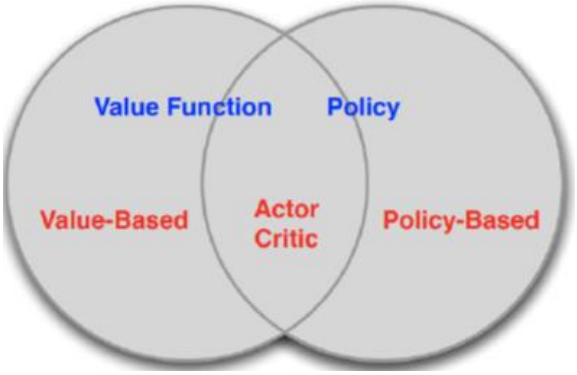
1. Value-based vs Policy-based RL

Value based RL

Value based RL은 value function을 이용하여 강화학습을 하는 방법이다. Action space가 discrete 하고 action이 deterministic 하여 value function을 정의하기 용이할 때 주로 사용되며, 각 state 에서의 action에 따른 value를 판단하여 최적의 action을 찾는다. 이때 policy는 value function으로 부터 자연스럽게 얻을 수 있으므로 implict policy라고 부른다. Value-based algorithm에는 Q-Learning, SALSA, DQN 등이 있다.

Poicy based RL

Value function 없이 policy 자체를 approximate 해서 policy 만을 이용하여 학습하는 알고리즘이다. 대표적인 알고리즘으로 Policy gradient가 있다. 이런 Policy-based RL은 Value-based RL에 비해 converge 속도가 빠르고 action space가 continuous 하거나 high-dimensional인 경우에 효과적이다.



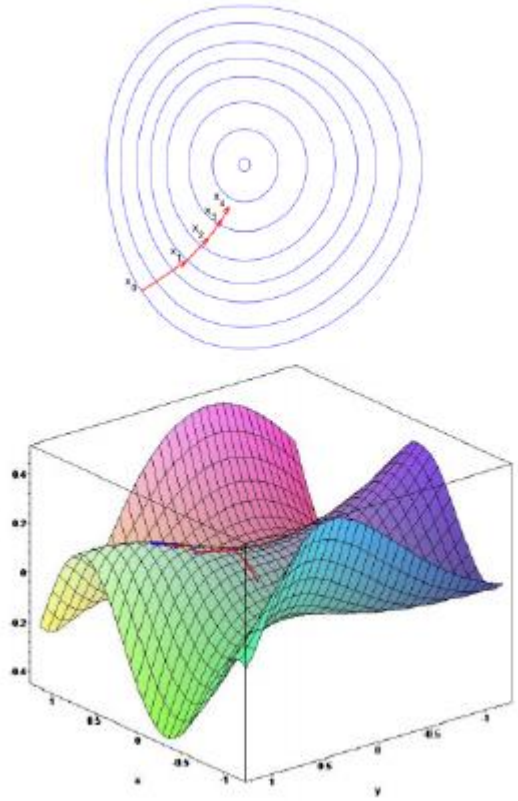
2. Policy Optimization & Policy Gradient Method

Policy Optimization

Policy based RL 을 위한 Objective function 에는 Expected reward 값을 사용한다. 이때, episodic environment 에는 start value 를 사용하고 continuing environment 에는 average value 혹은 average reward per time-step 을 사용한다.

Policy Gradient Method

위의 Objective function 을 최대화하기 위해 사용하는 방법으로 위의 objective function 의 gradient 를 direction 으로 하여 optimal 값을 찾는 방법이다.



[참고문헌]

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