Notes for RL

Jinze LI

June 13, 2024

Abstract

Notes

1 Outline

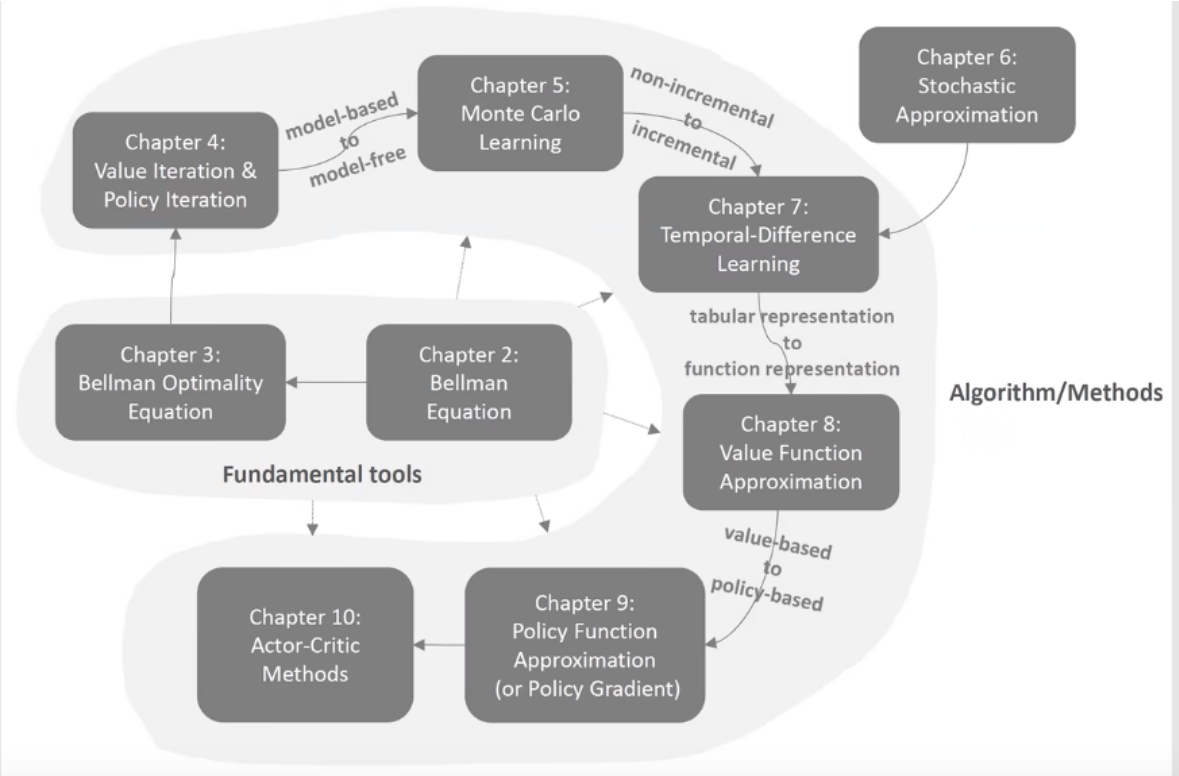

Figure 1: Enter Caption

2 Value Function Iteration

2.1 MP (Markov Process)

$$V(s) = E[G_t|s_t = s] = E[R_{t+1} + \gamma R_{t+2} + \dots |s_t = s] \tag{1}$$

$$= E[R_{t+1} + \gamma V(s')|s_t = s] \tag{2}$$

$$= R(s) + \gamma \sum_{s' \in S} P(s'|s)V(s') \tag{3}$$

$$\tag{4}$$

2.2 MRP (Markov Reward Process)

2.3 MDP (Markov Decision Process)

MDP = MRP + Action

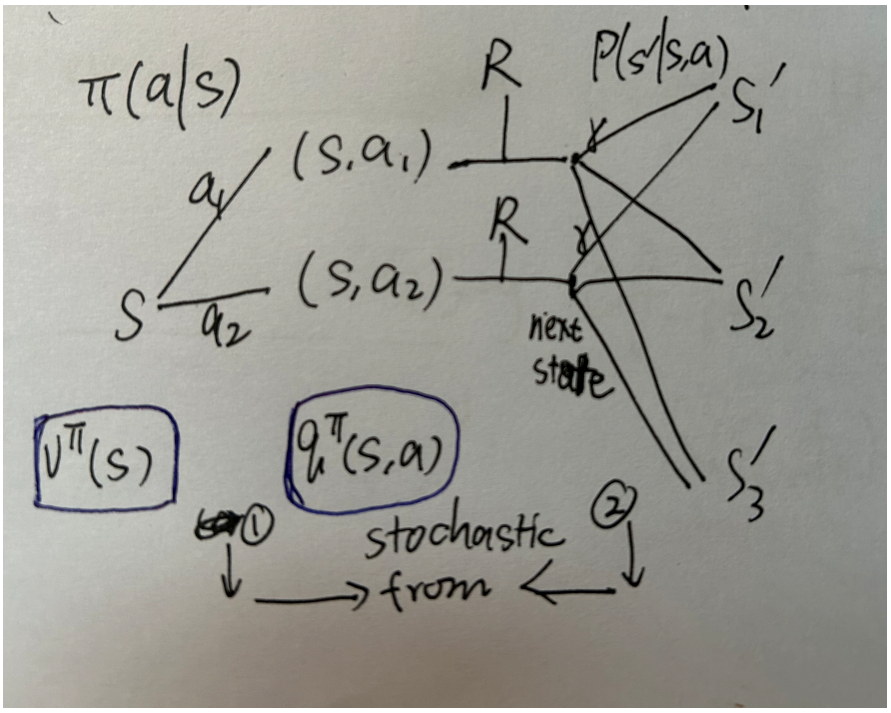

Figure 2: Enter Caption

$$v^{\pi}(s) = E_{\pi}[G_t|s_t = s] \tag{5}$$

$$q^{\pi}(s,a) = E_{\pi}[G_t|s_t = s, A_t = a] \tag{6}$$

$$\tag{7}$$

Relation

$$v^{\pi}(s) = \sum_{a\in A} \pi(a|s) q^{\pi}(s,a) \tag{8}$$

$$q^{\pi}(s,a) = R(s,a) + \gamma \sum_{s'\in S} P(s'|s,a) V(s') \tag{9}$$

$$v^{\pi}(s) = E_{\pi}[R_{t+1} + \gamma v^{\pi}(s_{t+1})|s_t = s] \tag{10}$$

$$q^{\pi}(s,a) = E_{\pi}[R_{t+1} + \gamma q^{\pi}(s_{t+1}, a_{t+1})|s_t = s, A_t = a] \tag{11}$$

$$v^{\pi}(s) = \sum_{a\in A} \pi(a|s)\big(R(s,a) + \gamma \sum_{s'\in S} P(s'|s,a) v^{\pi}(s')\big) \tag{12}$$

$$q^{\pi}(s,a) = R(s,a) + \gamma \sum_{s'\in S} P(s'|s,a) \sum_{a'\in A} \pi(a'|s') q^{\pi}(s',a') \tag{13}$$

3 Policy Iteration

4 Mean Estimation

5 on/off- line learning

6 MC

first visit first, every visit

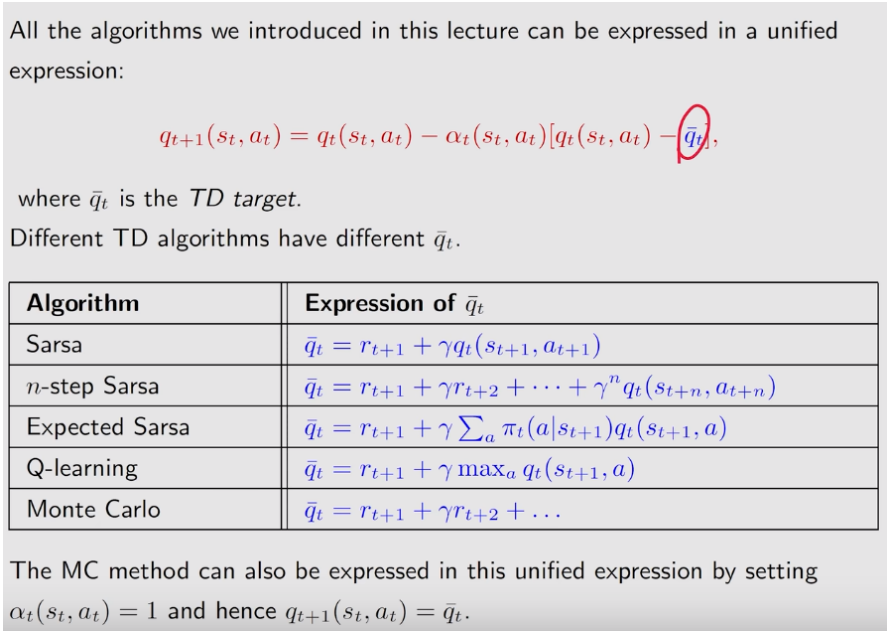

Figure 3: Enter Caption

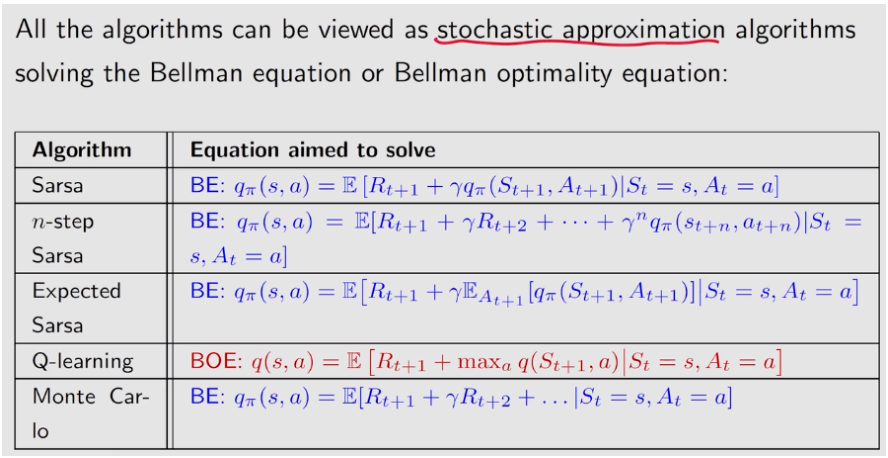

Figure 4: Enter Caption

7 TD

7.1 on/off-policy

The Behavior Policy is used to generate experience samples.

The Target Policy is toward our optimal policy.

On-policy learning: the above two are the same.

Off-policy learning: the above two are NOT the same. (Learn Experience from Others.)

eg. Set behavior policy to be exploration policy, and we use the experience to do the target policy.

7.2 examples

Sarsa is on-policy. MC is on-policy

Q-learning is on/off-policy.

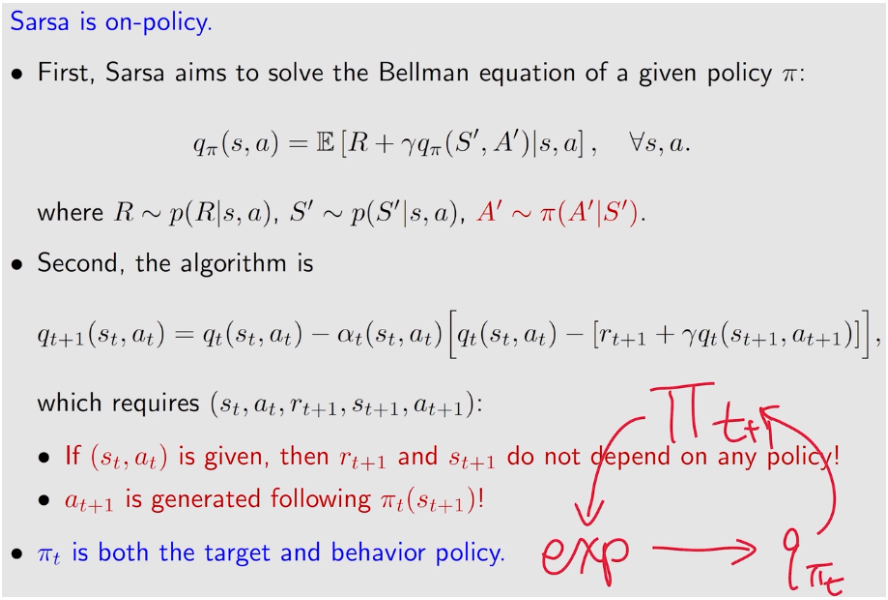

Figure 5: Enter Caption

Q-learning tabular version

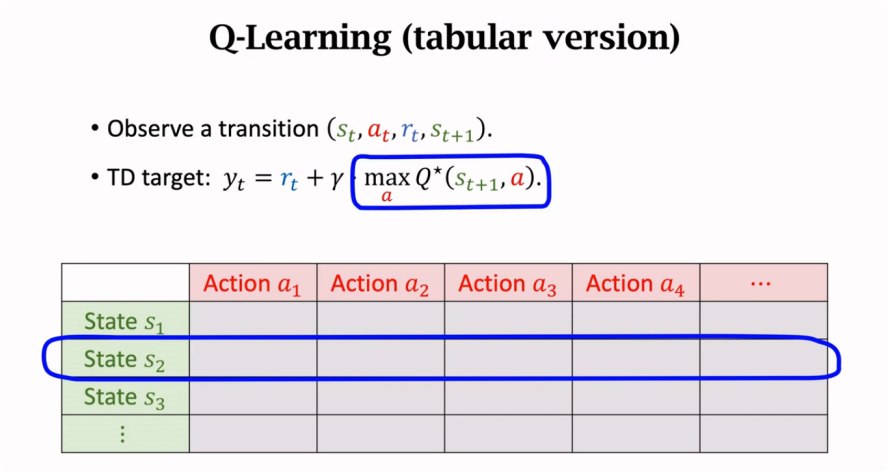

Figure 6: Enter Caption

functional approximation: Linear

Deep learning (DQN)

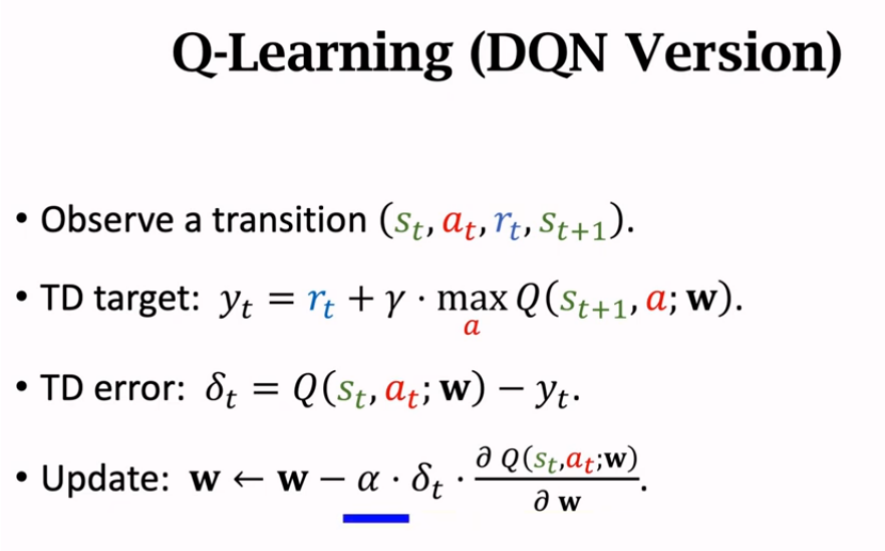

Figure 7: Enter Caption

8 Value Function Approximation

9 Policy Function Approximation (Policy Gradient)

10 ref

ZHAO shiyu course

le yuan blog

Liang tengyuan, course

SDSC8006, Clint Chin Pang Ho

References