Reinforcement Learning

value-based RL (MC, TD)

2024.05.29
안재성

RL and DP

- 학습의 목적
  1. **Q function이 무엇인가**
  2. **Value based RL이 무엇인가**
  3. **Off-policy 와 On-policy가 무슨 차이인가**
  4. **MC와 TD의 차이가 무엇인가**

RL and DP

- **What is different between RL and DP**

RL도 결국 Bellman Equation을 푸는 것이다.

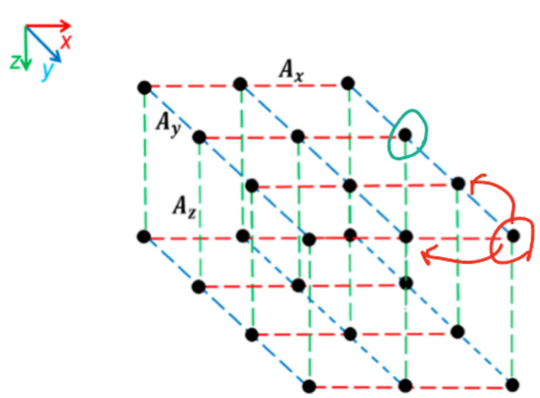
$$V_k^*(x_k) = \min_u \sum_{x_{k+1}} p(x_{k+1} \mid x_k, u)[r + V_{k+1}^*(x_{k+1})]$$

우변 다음 state x_{k+1} **을 구할 때, 모델을 사용** $x_{k+1} = f(x_k, u_k)$

1. State-space model이 없거나

2. State-space model이 잘못 됐거나

3. Curse of dimensionality



용어 정리

- **Control society VS Machine Learning society**

최적제어		강화학습	
State	x	State	S
Control	u	Action	A
Dynamics	$f(x, u)$	Environment	$p(s' \mid s, a)$
Controller	π	Agent	π
Cost	$r(x, u)$	Reward	$r(s, a)$
Batch		Episode	

- **What is important thing in RL**

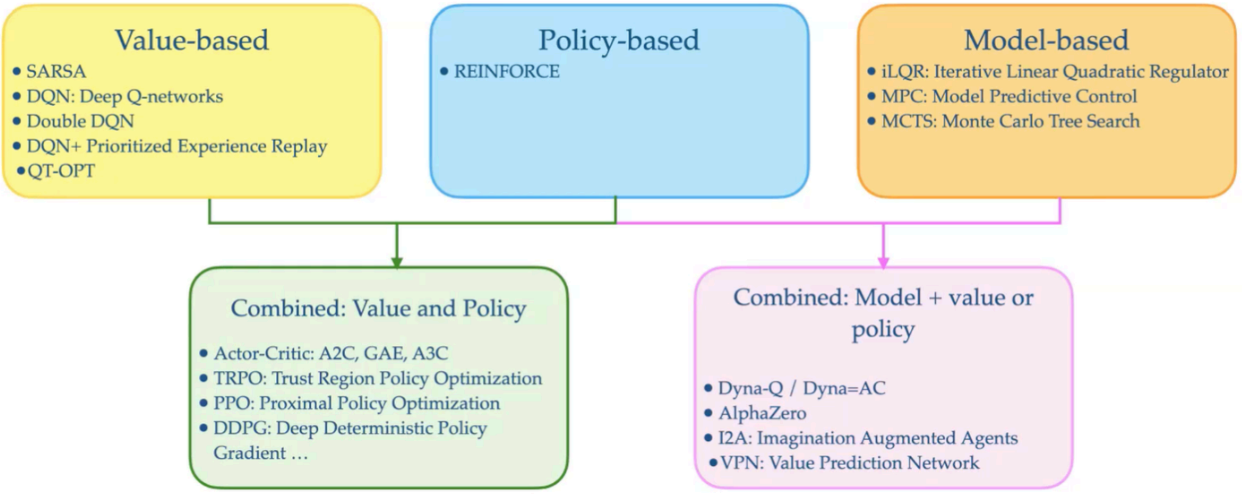
$$G_t = r(x_t, u_t) + \gamma r(x_{t+1}, u_{t+1}) + \gamma^2 r(x_{t+2}, u_{t+2}) + \cdots + \gamma^{T-t} r(x_T, u_T)$$
$$= \sum_{k=t}^{T} \gamma^{k-t} r(x_k, u_k)$$

$$\theta^* = \arg\max J(\theta)$$

$$J(\theta) = \mathbb{E}_{\tau \sim p_{\theta(\tau)}} \left[\sum_{t=0}^{T} \gamma^t r(x_t, u_t) \right]$$

- **Learnable function in RL**

	Model-free		**Model-based**
	Value-based	Policy-based	
학습 대상	Estimate Q function	policy 자체를 학습	모델을 학습
	SARSA 등등	REINFORCE	LQR, MPC



Value-based RL

- **Value function**

$$V^\pi(x_t) = \mathbb{E}_\pi\left[\sum_{\tau=t}^{\infty}\gamma^\tau r(x_\tau, u_\tau | x_\tau)\right]$$

policy가 stationary policy

$$\begin{aligned} V^\pi &= r_0 + \gamma p_0 r_1 + \gamma^2 p_0 p_1 r_2 + \cdots \\ &= r_0 + \gamma p_0 (r_1 + \gamma p_1 r_2 + \cdots \\ &= r_0 + \gamma p_d V^\pi \end{aligned}$$

Time-invariant

$$V^\pi(x_t) = r(x_t, \pi(x_t)) + \mathbb{E}_\pi\left[V^\pi(x_{t+1})\right]$$

$$\pi(x_t) = \arg\min_u (r(x_t, u) + \mathbb{E}_\pi\left[V^\pi(x_{t+1})\right])$$

Time-varying

$$V_t^\pi(x_t) = r(x_t, \pi(x_t)) + \mathbb{E}_\pi\left[V_{t+1}^\pi(x_{t+1})\right]$$

- **Q function : Action value function**

$$Q^\pi(\underline{x_t, u_t}) = \mathbb{E}_\pi\left[\sum_{\tau=t}^{\infty}\gamma^\tau r(x_\tau, u_\tau | \underline{x_t, u_t})\right]$$

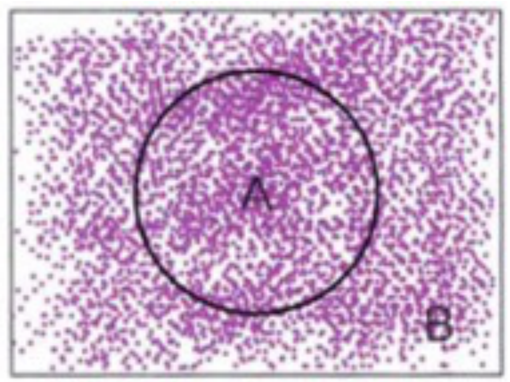
$$V^\pi(x_t) = \mathbb{E}_\pi\left[Q^\pi(x_t, u_t)\right]$$

Relationship between Value func and Action value func : 상태가치는 상태변수 x_t 에서 선택 가능한 모든 행동 u_t에 대한 행동가치의 평균값

$$Q^\pi(x_t, u_t) = r(x_t, u_t) + \mathbb{E}_\pi\left[Q^\pi(x_{t+1}, \pi(x_{t+1})\right]$$

$$\pi^*(x_t) = \arg\min_u Q^*(x_t, u)$$

Monte-Carlo(MC)



if $n \to \infty$, then

$$\frac{1}{n}\sum_{i=1}^{n} I(\text{red\_dot}_i \in A) = \frac{S(A)}{S(B)}$$

- Goal: Learn Q^{π} from episodes of experience under policy π
- Recall that $Q^{\pi}(x_t, u_t) = \mathbb{E}[\sum_{\tau=t}^{\infty} \gamma^{\tau} r(x_{\tau}, u_{\tau}) | x_t, u_t]$
- Return: $G_t = r_t + \gamma r_{t+1} + \cdots + \gamma^{T-t-1} r_T + \cdots$

- Monte-Carlo method: Replace expectation with empirical mean

$$Q^{\pi}(x_t, u_t) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{\tau=t}^{\infty} \gamma^{\tau} r(x_{\tau,i}, u_{\tau,i})$$

Monte-Carlo(MC) and Temporal-difference(TD) policy iteration

- **MC**

For each episode

- Generate an episode π: $x_0, u_0, r_0, \dots, x_T$
- $G \leftarrow 0$
- For each step, $t = T, T-1, \dots 0$:
  - $G \leftarrow \gamma G + r_t$
  - $C(x_t, u_t) \leftarrow C(x_t, u_t) + 1$
  - $Q(x_t, u_t) \leftarrow Q(x_t, u_t) + \frac{1}{C(x_t, u_t)}(G - Q(x_t, u_t))$
  - $u^* \leftarrow \mathrm{argmin}_u Q(x_t, u)$

  For all $u \in U$
    - $\pi(u|x_t) \leftarrow u^*$ with prob. $1 - \epsilon$ (ϵ-greedy)

$$Q_{n+1} = \frac{1}{n}\sum_{i=1}^{n} G_i$$

$$Q_{n+1} = Q_n + \frac{1}{n}(G_n - Q_n)$$

Q function Look-up table

$Q(x,u)$	u_1	u_2	u_3	...
x_1				
x_2				
x_3				
x_4				
x_5				
x_6				
x_7				
x_8				
$\vdots$				

1
2
3
4

- **TD**

$$Q(x_t, u_t) \leftarrow Q(x_t, u_t) + \frac{1}{C(x_t, u_t)}(G - Q(x_t, u_t)) \quad \rightarrow \quad Q(x_t, u_t) \leftarrow Q(x_t, u_t) + \alpha(r_t + \gamma Q(x_{t+1}, u') - Q(x_t, u_t))$$

$$G_t = r_t + \gamma r_{t+1} + \cdots + \gamma^{T-t-1} r_T + \cdots \approx r_t + \gamma Q^{\pi}(x_{t+1}, u')$$

- **On-policy** **Off-policy**

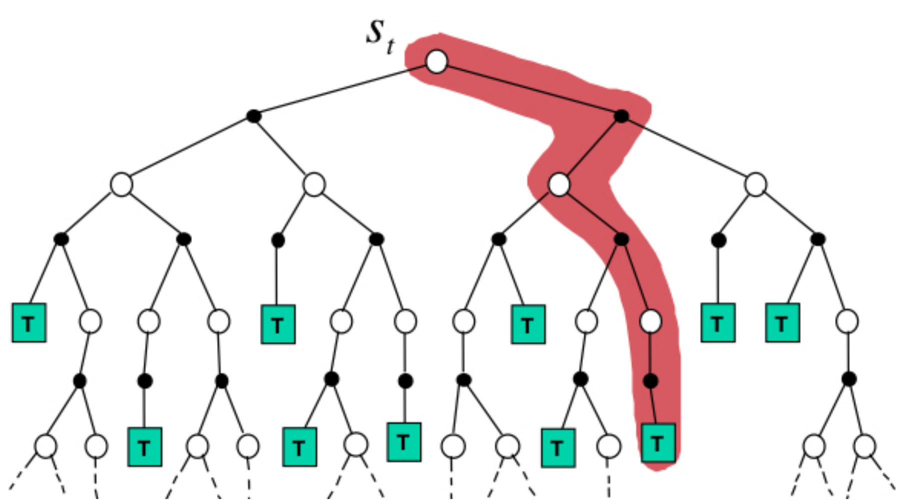
$u' \leftarrow u_{t+1}$

$u' \leftarrow \arg\min_u Q(x_{t+1}, u)$

Monte-Carlo(MC) and Temporal-difference(TD) policy iteration

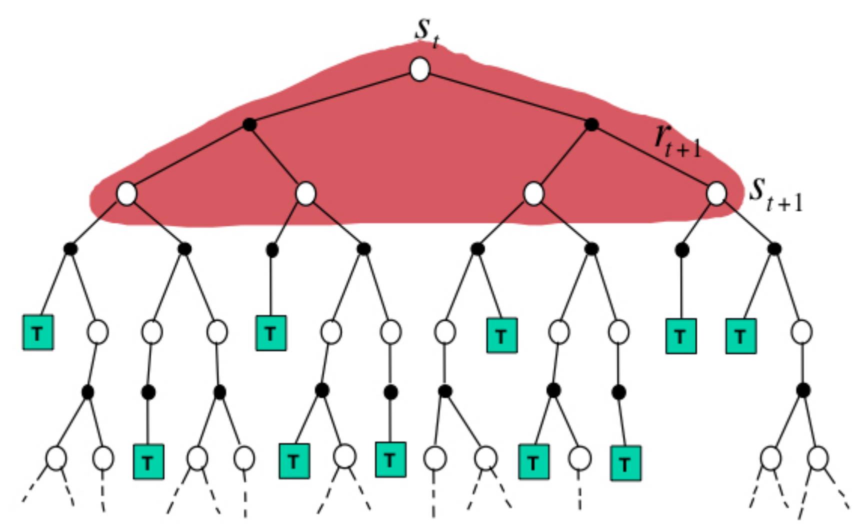
- **MC**

$$V(S_t) \leftarrow V(S_t) + \alpha\,(G_t - V(S_t))$$



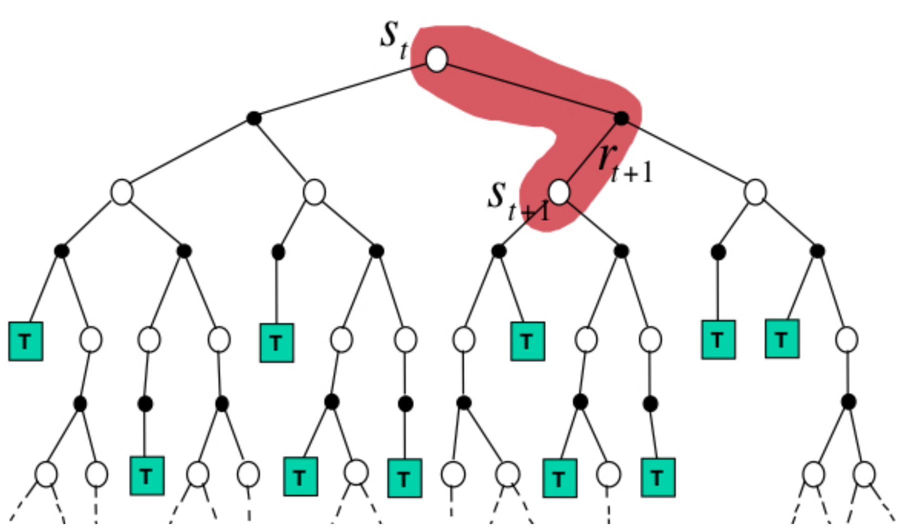
- **DP**

$$V(S_t) \leftarrow \mathbb{E}_\pi\,[R_{t+1} + \gamma V(S_{t+1})]$$



- **TD**

$$V(S_t) \leftarrow V(S_t) + \alpha\,(R_{t+1} + \gamma V(S_{t+1}) - V(S_t))$$



SARSA and Q-learning

SARSA: on-policy

- For each episode:

  For each step, $t = 0, 1, \ldots, T$:
  - Given x_t, choose $u_t = \mathrm{argmin}_u Q(x_t, u)$ (+ϵ-greedy)
  - Observe r_t, x_{t+1}
  - Choose $u_{t+1} = \mathrm{argmin}_u Q(x_{t+1}, u)$ (+ϵ-greedy)
  - $Q(x_t, u_t) \leftarrow Q(x_t, u_t) + \alpha\big(r_t + \gamma Q(x_{t+1}, u_{t+1}) - Q(x_t, u_t)\big)$

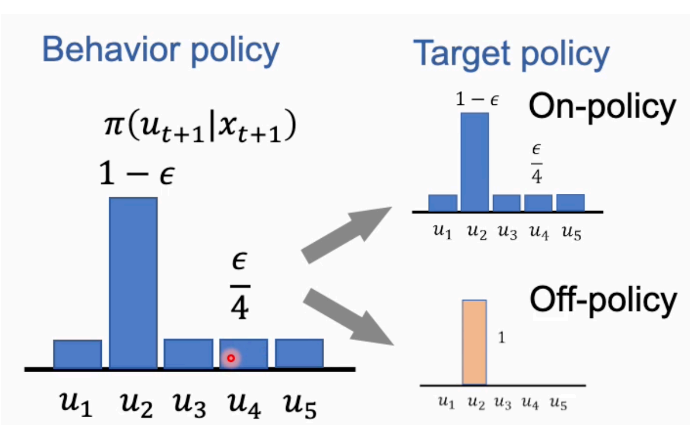
$Q(x,u)$	u_1	u_2	u_3	…
x_1				
x_2				
x_3				
x_4		(x_t, u_t)		
x_5				
x_6			(x_{t+1}, u_{t+1})	
x_7				
x_8				
$\vdots$				

Q-learning: off-policy

- For each episode:

  For each step, $t = 0, 1, \ldots, T$:
  - Given x_t, choose $u_t = \mathrm{argmin}_u Q(x_t, u)$ (+ϵ-greedy)
  - Observe r_t, x_{t+1}
  - $Q(x_t, u_t) \leftarrow Q(x_t, u_t) + \alpha\big(r_t + \gamma \min_u Q(x_{t+1}, u) - Q(x_t, u_t)\big)$

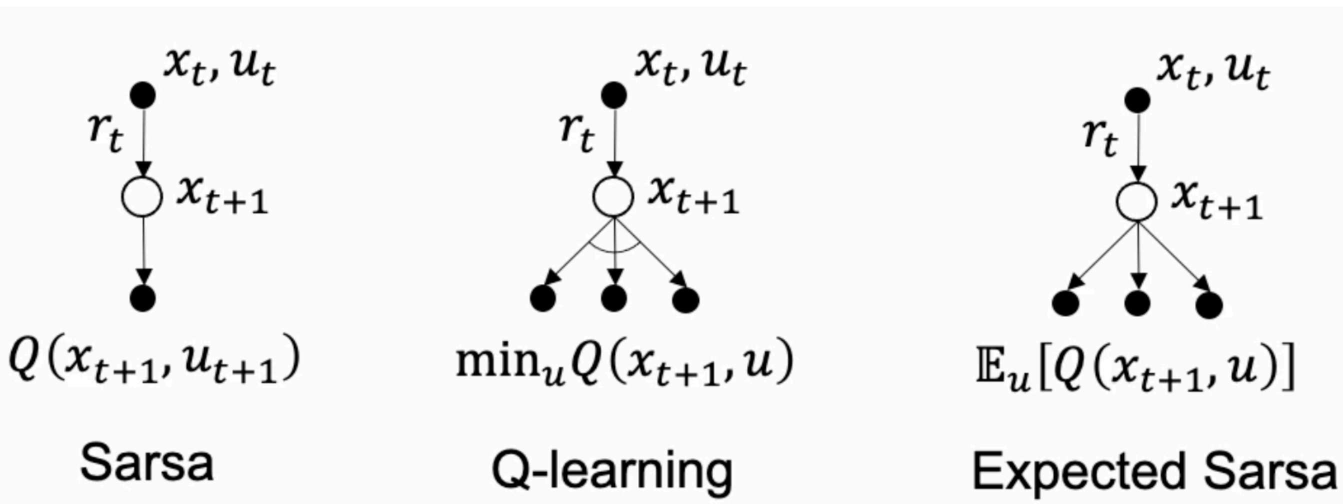
$Q(x,u)$	u_1	u_2	u_3	…
x_1				
x_2				
x_3				
x_4		(x_t, u_t)		
x_5				
x_6	(x_{t+1}, u)	(x_{t+1}, u)	(x_{t+1}, u)	
x_7				
x_8				
$\vdots$				



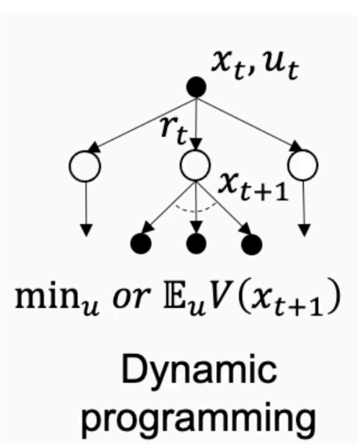
- **MC**


x_t, u_t
x_{t+1}
x_{t+2}
$\cdots$
x_T
r_t
r_{t+1}
Monte Carlo

- **TD**


x_t, u_t
r_t
x_{t+1}
$Q(x_{t+1}, u_{t+1})$
Sarsa
x_t, u_t
r_t
x_{t+1}
$\min_u Q(x_{t+1}, u)$
Q-learning
x_t, u_t
r_t
x_{t+1}
$\mathbb{E}_u[Q(x_{t+1}, u)]$
Expected Sarsa

- **DP**


x_t, u_t
r_t
x_{t+1}
$\min_u \; or \; \mathbb{E}_u V(x_{t+1})$
Dynamic programming

결론

- **MC**

  $G_t = r_t + \gamma r_{t+1} + \cdots + \gamma^{T-t-1} r_T + \cdots$

  - Episode 끝날 때까지 학습 못함, 끝나야 학습
  - High variance
  - Unbiased estimate of Q
  - 초기값에 민감하지 않음

- **TD**

  $G_t \approx r_t + \gamma Q^{\pi}(x_{t+1}, u')$

  - 매 step 마다 학습 가능
  - Low variance
  - Biased estimate of Q
  - 처음 어떤 Q를 쓰냐에 따라 잘 수렴하냐 안하냐 차이