✱ Model Free Control

- monte-Carlo Control
- Sarsa
- Q-learning

① monte-Carlo Control

monte-Carlo prediction을 통해 Policy Evaluation을 수행했다.

이제 Policy improvement를 해야지! ⟶ greedy Policy 이용!

$$\cdots \pi'(s) = \underset{a}{\operatorname{argmax}}\, q_\pi(s,a)$$

$$= \underset{a}{\operatorname{argmax}}\, R_s^a + \gamma \cdot \sum_{s'} P_{ss'}^a V_\pi(s')$$

→ 환경... 모르지! 다른 방법?...

애초에 Evaluation에서 V_π가 아니라 q_π를 찾자!

- ϵ-Greedy Exploration ⟶ 다른 Action도 작은 확률로 해보겠다!

$$\pi(a|s) = \begin{cases} \frac{\epsilon}{m} + (1-\epsilon) & \text{if } a = \underset{a}{\operatorname{argmax}}\, q_\pi(s,a) \\ \frac{\epsilon}{m} & \text{else} \end{cases}$$

해당 State only

⟷

기존 Greedy Policy

$$\pi'(s) = \underset{a}{\operatorname{argmax}}\, q_\pi(s,a)$$

② Sarsa

→ MC 대신 TD!

$$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha(R_{t+1} + \gamma \cdot Q(S_{t+1},A_{t+1}) - Q(S_t,A_t))$$

③ Q-learning

→ off-Policy learning!

↳ $\mu(a|s)$를 따라 움직이며 $\pi(a|s)$를 평가!

$$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha(R_{t+1} + \gamma Q(S_{t+1},A') - Q(S_t,A_t))$$

다음 Action은 μ를 따른다! / 이어지는 Action은 π를 따라가!

↓ improve

μ와 π를 모두 update!

$\pi(S_{t+1}) = \underset{a'}{\operatorname{argmax}}\, Q(S_{t+1},a')$ greedy!, μ ← ϵ-greedy!

$$\rightarrow R_{t+1} + \gamma \cdot Q(S_{t+1},A')$$

$$= R_{t+1} + \gamma \cdot Q(S_{t+1}, \underset{a'}{\operatorname{argmax}}\, Q(S_{t+1},a'))$$

$$= R_{t+1} + \max_{a'} \gamma \cdot Q(S_{t+1},a')$$

$$\therefore\ Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha(R_{t+1} + \max_{a'} \gamma \cdot Q(S_{t+1},a') - Q(S_t,A_t))$$