1. Randomness in return：折扣回報.

Discounted Return.

- $U_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots\cdots$

U_t 的隨機性由 Action，state 決定

策略函數.

1. Action can be randomnesss. $P(A=a \mid S=a) = \pi(a \mid s)$
2. New state can be randomness. $P(S'=s' \mid S=s, A=a) = P(s' \mid s, a)$

狀態狀移函數.

2. Action Value Function. 動作價值函數

對未來折扣獎勵求期望，因是連續型隨機變數

積分去除了折扣獎勵之隨機性.

$$Q_\pi(s_t, a_t) = E[U_t \mid S_t = s_t, A_t = a_t].$$

給動作打分

$$Q^*(s_t, a_t) = \max_\pi Q_\pi(s_t, a_t)$$

3. State value Function：狀態價值函數

評價當前狀態好壞.

$$V_\pi(s_t) = E_A[Q_\pi(s_t, A)] = \sum_a \pi(a \mid s_t) \cdot Q_\pi(s_t, a).$$

$$= \int \pi(a \mid s_t) \cdot Q_\pi(s_t, a)\, da$$

一、價值學習：

- 動作價值函數，會依當下情況給動作打分

$$Q_\pi = E[U_t \mid S_t = s_t, A_t = a_t]$$

$$Q^*(s_t, a_t) = \max_\pi Q_\pi(S_t = s_t, A_t = a_t)$$

上 0.7
右 0.2
左 0.1

$$a_t = \operatorname{argmax} Q^*(s_t, a_t)$$

- 利用神經網路近似Q函數 DQN

"先知"

↳ $Q(s_t, a_t; w)$ to approximate $Q^*(s, a)$.