1 目標，収益，報酬

- エージェントの目標は累積報酬を最大化すること (報酬仮説)
  - ***報酬仮説*** *Reward Hypothesis*
  - *目標: 期待報酬の最大化*
- 時刻 t における報酬 R_t : **スカラ値**
- 時刻 t におけるエージェント行為の評価

1.1 逐次的意思決定 Sequential Decision Making

- 目標 Goal: 総収益を最大化する行動を選択すること
- 行為，行動 Actions は長期的結果
- 収益は遅延することも有る
- 直近の報酬を選ぶよりも，長期的な報酬を考えた方が良い場合がある

2 収益 Return

- **収益** return G_t: 割引付き収益

$$G_t = R_{t+1} + \gamma R_{t+2} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{1}$$

- 割引率 The discount $\gamma \in [0, 1]$: 現時点から見た将来の報酬を計算するため
  - ***遅延報酬*** *delayed reward の評価*
  - 0 *に近ければ* ***近視眼的*** *評価*
  - 1 *に近ければ* ***将来を見通した*** *評価*

3 価値関数 Value Function

- **状態価値関数** v と **行動価値関数** q
- **状態価値関数** state-value function:

$$v_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s] \tag{2}$$

$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1} | S_t = s)] \tag{3}$$

- **行動価値関数** action-value function:

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a] \tag{4}$$

$$q_\pi(s, a) = \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a] \tag{5}$$

4 最適価値関数 Optimal Value Function

- 最適状態価値関数 :

$$v_*(s) = \max_\pi v_\pi(s) \qquad (6)$$

- 最適行動価値関数 :

$$q_*(s,a) = \max q_\pi(s,a) \qquad (7)$$

- ベルマン方程式 一般に非線形になるので難しい

5 最適価値関数 Optimal Value Functions

- 最大の価値を与える関数

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) = Q^{\pi^*}(s,a) \qquad (8)$$

- 最適価値関数 Q^* が得られれば最適方策 π^* を求めることができる

$$\pi^*(s) = \mathrm{argmax}_a\, Q^*(s,a) \qquad (9)$$

- 全ての意思決定における最適価値:

$$\begin{aligned} Q^*(s,a) &= r_{t+1} + \gamma \max_{a_{t+1}} r_{t+2} + \gamma^2 \max_{a_{t+2}} r_{t+3} + \ldots \\ &= r_{t+1} + \gamma \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1}) \end{aligned} \qquad (10)$$

- **ベルマン方程式** Bellman equation:

$$Q^*(s,a) = \mathbb{E}_{s'}\left[r + \gamma \max_{a'} Q^*\left(s', a'\right) | s, a \right]. \qquad (11)$$