◎ 強化学習の数理モデル

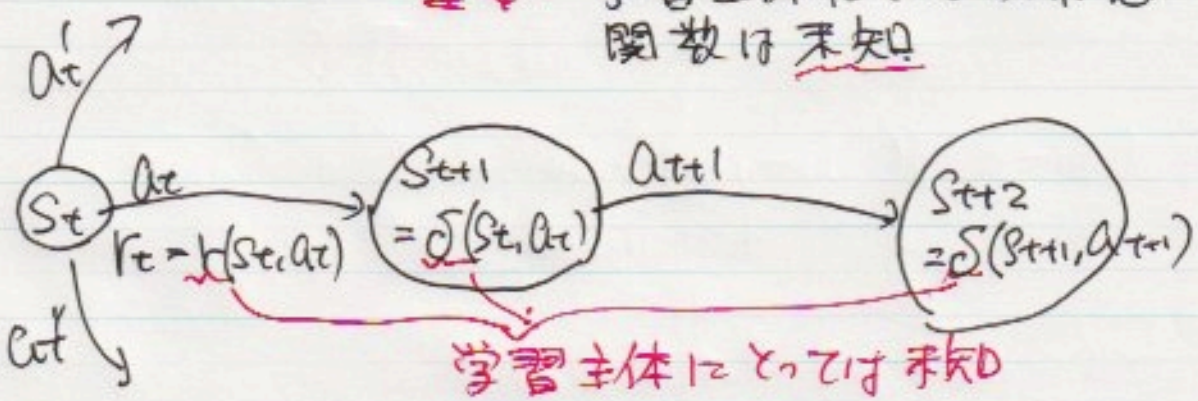
⇒ 離散時間マルコフ意志決定過程
(DiscreteMDP)

- 状態 (state): $s_t \in S$ ← 全状態集合
  時間, $t = 0, 1, 2, \cdots$
- 行為 (action): $a_t \in A$ ← 全行為集合
- 報酬 (reward): $r_t = r(s_t, a_t)$
  ↳ 報酬関数

重要：学習主体にとっては報酬関数は未知

次状態 $s_{t+1} = \delta(s_t, a_t)$
↳ 次状態関数

重要：学習主体にとって次状態関数は未知



マルコフ性：$\delta(s_t, a_t)$, $r(s_t, a_t)$ は s_t と a_t だけに依存している.

◎ 強化学習の目標：報酬を最大化する行動ポリシーを学習する
(制御側)

$\pi(s_t) = a_t$

制御ポリシーを $\pi(s_t) = a_t$ としたとき

累積報酬 $V^{\pi}(s_t)$

$$V^{\pi}(s_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots$$

$$= \sum_{i=0}^{\infty} \gamma^i r_{t+i}$$

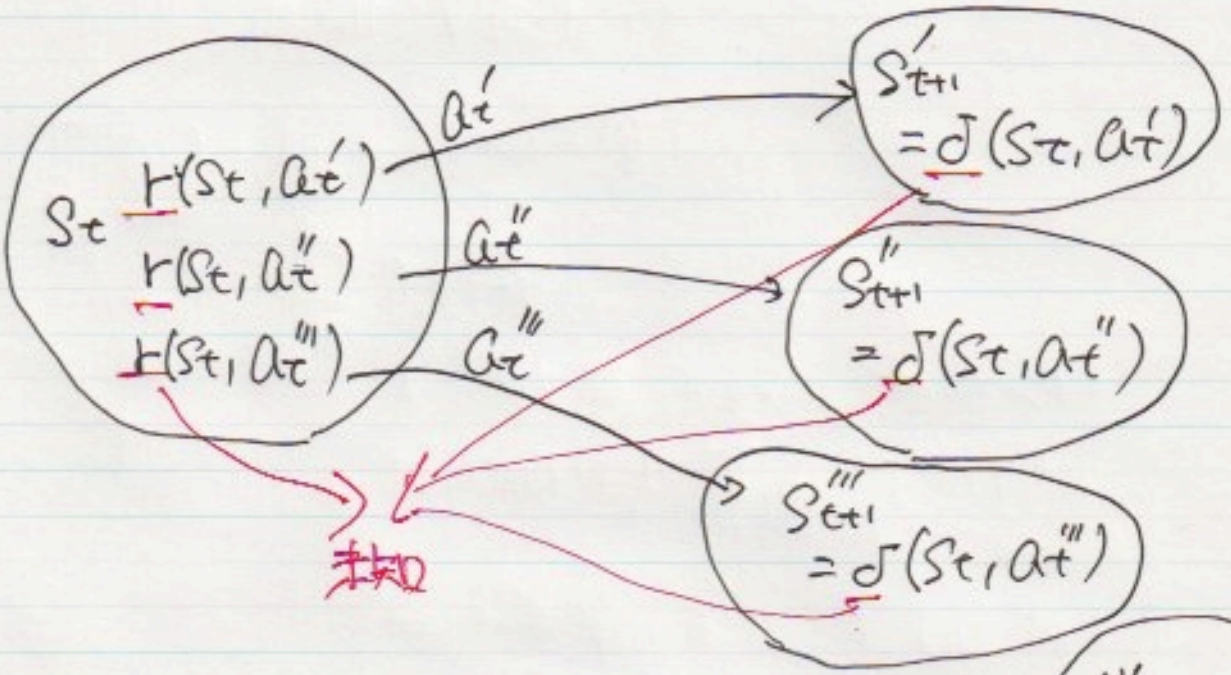
γ：減衰係数（$0 \leq \gamma < 1$）

技術的理由：$V^{\pi}(s_t)$ を有限の値にするために導入する．

（⇒社会学における未来係数と関係がある．）

→ 強化学習の目的は $V^{\pi}(s_t)$ を最大化すること

⊙ Q学習：$V^{\pi}(s_t)$ の最大値を推定



A^* 最適なヒューリスティック探索

$$r(s,a) + \gamma V^{\pi}(\delta(s,a))$$

未知なので推定する

↓ 推定値 Q を導入

$$Q(s,a) \stackrel{\text{def}}{=} r(s,a) + \gamma V^{*}(\delta(s,a))$$

↳ 最適値

"定義より" $V^{*}(\delta(s,a)) = \max_{a'} Q(\delta(s,a), a')$

つまり $$\boxed{Q(s,a) = r(s,a) + \gamma \max_{a'} Q(\delta(s,a), a')}$$