速習 強化学習

1　マルコフ決定過程

完備な確率空間を $(\Omega, \mathcal{F}, \mathbb{P})$ とする.

Definition 1.1. (可算 MDP)
可算な状態集合を $\mathcal{X}$, 可算な行動集合を $\mathcal{A}$ とする. この時遷移確率カーネル $P_0 : \mathcal{B}(\mathcal{X} \times \mathbb{R}) \times \mathcal{X} \times \mathcal{A} \to [0, 1]$ を

$$P_0(U, x, a) = P_0(U \mid x, a)$$

と定める. ここで P_0 をは $\mathcal{X} \times \mathbb{R}$ 上の確率測度である. この時 $(\mathcal{X}, \mathcal{A}, P_0)$ を可算 MDP という.

Definition 1.2. (状態遷移確率カーネル)
$(\mathcal{X}, \mathcal{A}, P_0)$ を可算 MDP とする. 状態遷移確率確率カーネル $P : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to [0, 1]$ を

$$P(x, a, y) = P_0(\{y\} \times \mathbb{R} \mid x, a)$$

と定める.

Definition 1.3. (即時報酬関数, 即時報酬)
$(\mathcal{X}, \mathcal{A}, P_0)$ を可算 MDP とする. 即時報酬と呼ばれる確率変数 $R_{(x,a)} : \Omega \to \mathcal{X}$ を考える. ただし $(x, a) \in \mathcal{X} \times \mathcal{A}$ であり $R_{(x,a)}$ は (x, a) に依存する. 即時報酬が $(Y_{(x,a)}, R_{(x,s)}) \sim P_0(\cdot|x, a)$ である時,(ただし $Y_{(r,a)} : \Omega \to \mathcal{A}$ は確率変数) 即時報酬関数 $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ を

$$r(x, a) = \mathbb{E}[R_{(x,a)}]$$

と定める.

Proposition 1.4. 任意の $x \in \mathcal{X}, a \in \mathcal{A}$ について, 確率 1 で $|R_{(x,a)}| < M$ であるならば

$$\|r\|_\infty = \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} |r(x, a)| < M$$

が成立する.

Proof.

$$r(x,a) = \int_\Omega r(x,a) < \int_\Omega M d\mathbb{P} = M \times \mathbb{P}(\Omega) = M$$

これが任意の $(x,a) \in \mathcal{X} \times \mathcal{A}$ について成立するので

$$\|r\|_\infty = \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} |r(x,a)| < M$$

が成り立つ. □

マルコフ決定過程は逐次的な意思決定問題を記述する空間である. $(\mathcal{X}, \mathcal{A}, P_0)$ を可算 MDP とする. $t \in \mathbb{N}$ とし $X_t : \Omega \to \mathcal{X}, A_t : \Omega \to \mathcal{A}$ を確率変数とする. この選択された行動が実行されると, システムにそれを反映した行動が生じる.

$$(X_{t+1}, R_{t+1}) \sim P_0(\cdot | X_t, A_t)$$

ここで, 任意の $x, y \in \mathcal{X}$, $a \in \mathcal{A}$ について $\mathbb{P}(X_{t+1} = y \ | X_t = x, A_t = a) = P(x,a,y)$ である. また, 任意の $w \in \Omega$ に対して $\mathbb{E}[R_{t+1} | X_t(w), A_t(w)] = r(X_t(w), A_t(w))$ である.

Definition 1.5. (行動則)
行動を選択するルールのことを行動則という.

行動則と確率的に決定される初期状態 $X_0 : \Omega \to \mathcal{X}$ によって, 状態-行動-報酬の系列 $((X_t, A_t, R_{t+1}; t \geq 0))$ が確率的に決定される. (X_{t+1}, R_{t+1}) は $(X_{t+1}, R_{t+1}) \sim P_0(\cdot | X_t, A_t)$ より, (X_t, A_t) に依存している.

Definition 1.6. (報酬)
$0 \leq \gamma \leq 1$ を割引率とする. 報酬 R を以下のように定義する.

$$R = \sum_{t=0}^{\infty} \gamma^t R_{t+1}$$

割引率 γ が 0 以上 1 未満で上で定義される報酬を持つ MDP を割引報酬 MDP と呼び, 割引率が 1 の時の報酬を持つ MDP は割引なしと呼ばれる.