目次

第 1 章

強化学習

1.1　強化学習とマルコフ決定過程

1.1.1　強化学習の目的

本章で扱う強化学習 (reinforcement learning, RL) では環境 (environment) と，その中で行動するエージェント (agent) という概念が導入される．環境とは，エージェントが相互作用する対象であり，エージェントの行動によってその状態が変化するものである．一方，エージェントは環境内で行動を選択し，学習を行う主体（例えば生物やロボットなど）を意味する．エージェントは環境内で行動し，状態と行動に応じて報酬 (reward) を得る．強化学習ではエージェントには望ましい行動が教師信号として与えられない代わりに，この報酬が与えられる．強化学習の目的は，エージェントが環境との相互作用を行い，結果として得られる報酬をより多く獲得する（目標を達成する）ために行動の選択を調整することである．

1.1.2　状態と行動

環境とエージェントの状態 (state) を $s \in \mathcal{S}$ とし，エージェントの行動 (action) を $a \in \mathcal{A}$ とする．ここで，$\mathcal{S}$ は環境とエージェントのあらゆる可能な状態の集合であり，$\mathcal{A}$ はエージェントが選択できる行動の集合である．状態や行動は離散的または連続的であり得る．

状態と行動が離散的である例として，グリッド状の迷路の探索課題が挙げられる．この場合，環境は迷路全体を指し，状態集合 $\mathcal{S}$ は迷路内におけるエージェントの位置からなり，行動集合 $\mathcal{A}$ は，{上, 下, 左, 右} の 4 つの移動方向からなる．移動しない（その場で待つ）ことが行動集合に含まれる場合もある．

状態と行動が連続的である例としては，動物の歩行が挙げられる．この場合，環境は動物を取り巻くすべての要素を指し，エージェントは動物（厳密にはその神経系）に相当する．状態集合 $\mathcal{S}$ は環境の状態（地面や大気の状態など）に加え，動物自身の状態（環境内での位置や体の各部位の配置など）が含まれる．一方，行動集合 $\mathcal{A}$ は特定の筋肉の筋緊張の強弱などで表される．

1.1.3 報酬

エージェントは行動の結果として，状態に応じた報酬 $r \in \mathbb{R}$ を得る．この報酬は正にも負にもなり得る．望ましい行動をとった場合には正の報酬が得られ，望ましくない行動をとった場合には負の報酬，すなわち罰 (punishment) が与えられる．報酬は即時に得られることもあれば，長期的な成果としてもたらされることもある．

具体例として，動物の歩行を考えてみよう．正の報酬としては，移動先で得られる水や餌（食料）などがある．一方，負の報酬には，歩行による疲労（エネルギー消費）や痛み（筋肉痛，障害物との接触，外敵の攻撃など）が含まれる．

生物においては，環境や自身の状態からさまざまな要素が報酬として与えられ，その生物（エージェント）がすべての報酬を明示的に設定する必要はない．しかし，強化学習の枠組みでは，エージェントに課題を解かせるために，人間が適切に報酬を定義する必要がある．この過程を報酬設計 (reward design) と呼ぶ．例えば，迷路探索課題では，動物の歩行における報酬を抽象化し，ゴール到達時に正の報酬を与え，移動に伴って一定の負の報酬を課すといった形で報酬を設計することができる．

1.1.4 マルコフ決定過程 (MDP)

これまで説明した状態・行動・報酬の遷移について考えよう．エージェントが状態 s_t において行動 a_t をとると，状態 s_{t+1} に遷移し，報酬 r_{t+1} を受け取る[*1]．状態 s_{t+1} と報酬 r_{t+1} が直前の状態 s_t と行動 a_t のみに依存し，過去の状態や行動の履歴には依存しない場合，この過程はマルコフ性 (Markov property) を持つと言える．このとき，環境とエージェントの状態遷移確率は $p(s_{t+1}, r_{t+1} \mid s_t, a_t)$ で表される．これは「状態 s_t で行動 a_t を選択した際に，次の状態が s_{t+1} になり，報酬 r_{t+1} を得る確率」を示している．このように状態遷移がマルコフ性を持ち，エージェントの行動が次の状態への遷移確率を決定する確率過程をマルコフ決定過程 (Markov decision process; MDP) と呼ぶ．MDP が成立する，すなわち状態遷移がマルコフ性を持つためには，状態 s_t が環境とエージェントの相互作用に関する十分な情報を持つ必要がある．

1.1.5 部分観測マルコフ決定過程 (POMDP)

動物は感覚器を通して外界を認識しているが，外界のすべてを認識できるわけではない．これと同様に，エージェントは環境およびエージェント自身の状態 s_t を直接観測できるとは限らない．エージェントが環境およびエージェント自身から受け取る情報を観測 (observation) o_t とすると，$o_t = s_t$ の場合は MDP が成立する．

しかし，現実の多くの問題では，エージェントは s_t の一部しか観測できない場合や，観測に不確実

[*1] 状態 s_t において行動 a_t を行った後に受け取る報酬を r_t とする流派もある．

性 (uncertainty) を含む場合がある．この場合，環境は部分観測マルコフ決定過程 (partially observable Markov decision process; POMDP) で記述される．例えば，動物が視覚経路から外部の環境を観測する場合，瞬時的には視野の範囲しか外界を観測できず，また視野の範囲の物体であっても二次元の網膜像からは物体の三次元的形状を正確に得ることはできない（形状は推論する必要があり，その過程には不確実性が含まれる）．このような状況では，エージェントは観測の不確実性を考慮し，状態に対する信念 (beliefs) を持って意思決定を行う必要がある．

1.1.6 方策と軌道

与えられた状態 s に対してエージェントの行動 a を決定する関数を方策 (policy) と呼び，π で表される．ある状態 s に対して常に同じ行動 a を決定する方策を決定論的方策と呼び，$a = \pi(s)$ で表される．一方で行動を確率的に決定する方策を，確率的方策と呼び，$\pi(a \mid s) = p(a \mid s)$ で表される．ここで π のみを使用する場合は方策それ自体を意味し，$\pi(a \mid s)$ は状態 s が与えられた時に a を選択する確率を意味する．

次に 軌道 (trajectory) を定義する．軌道とは，あるエージェントが環境と相互作用する中で得られる状態，行動，報酬の系列全体をまとめたものであり，

$$\tau := \{s_0, a_0, r_1, s_1, a_1, r_2, \ldots, s_T, a_T, r_{T+1}, s_{T+1}\} \tag{1.1}$$

のように表される．ここで T は任意の終端時刻を表し，s_{T+1} は終端状態 (terminal state) と呼ばれる．終端時刻 T が有限であり，目標の達成や失敗などの条件で明確に終了する（終端状態がある）軌道は，特に エピソード (episode) あるいは試行 (trial) と呼ばれる．すなわち，エピソードは終端条件を満たして終了する軌道であり，無限に続く可能性のある軌道（例えば定常方策による継続的な制御）と区別されうる．T が有限の場合，方策 π の下で，軌道（エピソード）τ を取る確率は，マルコフ性より，

$$p(\tau) := p(s_0) \prod_{t=0}^{T} p(s_{t+1}, r_{t+1} \mid s_t, a_t) \pi(a_t \mid s_t) \tag{1.2}$$

と表される．ただし，$p(s_0)$ は初期状態 s_0 を取る確率である．

1.1.7 収益

強化学習は望ましい方策を得ることが目的であるが，そのためには方策の「良さ」を評価する必要がある．単純に瞬時的な報酬 r_t で方策を評価した場合，即時的には報酬が少ないが後に大きな報酬が貰えるような方策を取らなくなってしまうため，これは望ましくない．こうした，行動に対する報酬が即時に得られず，後に得られるような場合の報酬を遅延報酬 (delayed reward) と呼ぶ．方策の評価のためには遅延報酬も含めた報酬を将来全体において累積的に評価することが必要であり，評価した値を収益

(return) と呼ぶ．最も単純な収益 G_t としては，時刻 $t+1$ 以降の報酬を加算した累積報酬 (cumulative reward) があり，時刻 T に得られる報酬までを考慮する場合は次式で表される．

$$G_t := r_{t+1} + r_{t+2} + r_{t+3} + \cdots + r_T = \sum_{k=t+1}^{T} r_k \tag{1.3}$$

累積報酬は平易であるが，T が大きい場合には G_t が無限大に発散してしまう恐れがある．そこで，G_t の発散を防ぐために割引率 (discount factor) γ $(0 \leq \gamma \leq 1)$ と呼ばれる係数で将来の報酬が減衰するようにする．

$$G_t := r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots + \gamma^{T-t-1} r_T = \sum_{k=t+1}^{T} \gamma^{k-t-1} r_k \tag{1.4}$$

これを割引報酬和 (discounted total reward) と呼ぶ．$T \to \infty$ の場合は $\gamma^{T-t-1} r_T \to 0$ となるため G_t が発散することは防がれる．γ が 0 に近い場合は短期的な報酬を重視し，1 に近い場合は累積報酬のように長期的な報酬も重視して行動選択を行うこととなる．以降では，$T \to \infty$ とし，無限の未来の報酬までを考慮した $G_t := \sum_{k=t+1}^{\infty} \gamma^{k-t-1} r_k$ を収益として考えることとする．この場合，

$$\begin{aligned} G_t &= r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots && (1.5)\\ &= r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \cdots) && (1.6)\\ &= r_{t+1} + \gamma G_{t+1} && (1.7) \end{aligned}$$

が成立する．

1.1.8 価値

方策は状態に応じて変化するため，方策 π の収益は状態ごとに評価する必要がある．状態 s から，方策 π に従って行動を選択した場合の収益の期待値を，状態 s の価値 (value) あるいは状態価値 (state value) と呼び，$v_\pi(s)$ で表す．MDP の場合，$v_\pi(s)$ は以下で定義される．

$$v_\pi(s) := \mathbb{E}_\pi \left[G_t \mid s_t = s \right] = \mathbb{E}_\pi \left[\sum_{k=t+1}^{\infty} \gamma^{k-t-1} r_k \,\middle|\, s_t = s \right] \tag{1.8}$$

ここで，$\mathbb{E}_\pi[\cdot]$ は方策 π に従う場合の $[\cdot]$ 内の確率変数の期待値を取ることを意味する．また，$v_\pi(\cdot)$ を状態価値関数 (state value function) と呼ぶ．

状態価値と同様の発想で，状態 s において行動 a を選択した場合の価値を行動価値 (action value) と呼ぶ．行動価値は，方策 π に従う条件下で，状態 s において行動 a を選択した場合の収益の期待値として計算され，$q_\pi(s,a)$ で表される．

$$q_\pi(s,a) := \mathbb{E}_\pi\left[G_t \mid s_t = s, a_t = a\right] = \mathbb{E}_\pi\left[\sum_{k=t+1}^{\infty} \gamma^{k-t-1} r_k \;\middle|\; s_t = s, a_t = a\right] \tag{1.9}$$

この $q_\pi(\cdot)$ を行動価値関数 (action value function) と呼ぶ．状態 s における価値 $v_\pi(s)$ は，状態 s において取る可能性のあるすべての行動 a の価値 $q_\pi(s,a)$ の期待値として次式のように表すことができる．

$$v_\pi(s) = \sum_a \pi(a \mid s) q_\pi(s,a) \tag{1.10}$$

すなわち，状態 s の価値 $v_\pi(s)$ は，その状態 s での各行動 a の価値 $q_\pi(s,a)$ に方策，つまり行動 a が取られる確率 $\pi(a \mid s)$ の重みをつけた加重平均として計算できる．