強化学習

1　強化学習問題

学習という意味を考えたとき、恐らく自転車の乗り方を覚えたときのような試行錯誤による学習と、授業を通して知識を得る学習の 2 つを思い浮かべる。確かにどちらも学習と言えるが、教師的な存在の有無という点で両者は異なる。**強化学習**は前者的な学習を指し、後者は教師あり学習に分類される。

強化学習を一言で表すならば、**エージェント**が**環境**から**報酬**信号を受け、それを基に何をすべきか学習する仕組み」である。強化学習ではエージェントや環境、並びに報酬と言う言葉をよく用いる。エージェントとは学習や行動を担うものであり、ロボットや制御器が該当する。環境はそっくりそのままエージェントが置かれた環境のことであるが、実空間の環境だけでなく例えば将棋ゲームといったものも含まれる分、私たちが普段使う環境よりも広い意味を持つ。

エージェントや環境と比べ、報酬は強化学習の中でより特徴的な存在だと言える。まず、強化学習が教師あり学習と最も異なる点は、行動の評価を基に学習することである。つまり、教師あり学習では最適な行動を直接教えて貰うのに対し、強化学習では行動の良し悪しを教えて貰えるものの、最良か否かまでは分からない。この良し悪しに関連する数値が正に報酬である。

もう少し強化学習の枠組みを言うなら、エージェントと環境は離散的な時間ステップ $t = 0, 1, ...$ において相互作用する。各時刻において、エージェントは何かしらの状態 $s_t \in \mathcal{S}$($\mathcal{S}$ は可能な状態の集合) に置かれ、これに対し何か行動 $a_t \in \mathcal{A}(s_t)$($\mathcal{A}(s_t)$ は s_t 状態化で取り得る行動の集合) を取る。その結果、エージェントの状態は s_{t+1} に遷移し、フィードバックとして報酬 r_{t+1} を受け取る。エージェントは一連の報酬 $r_t(t = 1, 2, ...)$ を受け取って、今までの行動 $a_t(t = 0, 1, ...)$ を改める (学習する)。以上の枠組みを図式化したものが図 1.1 である。

強化学習の問題では不確実性が随所に現れる。例えば状態遷移 s_{t+1} や受け取る報酬 r_{t+1} は不確かであることが多い。ちなみに、s_{t+1} に関する確率分布が s_t と a_t のみに依存し、かつ r_{t+1} に関する確率分布が s_t と a_t と s_{t+1} のみに依存する場合、このような強化学習問題を**マルコフ決定過程** (**MDP**) と言う。従って図 1.1 の枠組みは MDP について述べていたのであった。残念なことに、MDP を仮定できない環境も多く存在する (例えばトランプの大富豪。今の持ち札だけでなく、相手の行動履歴からも自分の行動を判断した方がよい)。ただし、そういった問題に取り組むときも、確率過程をシンプルにするために、何とかして MDP に近似できないか考察すべきである。本資料では全ての環境に対し MDP であることを仮定する。

$s_t = s$ かつ $a_t = a$ のとき、$s_{t+1} = s'$ となる確率を $\mathcal{P}^a_{ss'}$ と書くことにする。つまり、

$$\mathcal{P}^a_{ss'} \equiv \Pr\{s_{t+1} = s' | s_t = s, a_t = a\} \tag{1.1}$$

である。このような確率を**遷移確率**と言う。またこのときの r_{t+1} の期待値を $\mathcal{R}^a_{ss'}$ と書く。つまり、

$$\mathcal{R}^a_{ss'} \equiv \mathrm{E}\{r_{t+1} | s_t = s,\ a_t = a,\ s_{t+1} = s'\} \tag{1.2}$$

である。これは (s_t, a_t, s_{t+1}) を入力とした関数と見なせるため、**報酬関数**と呼ぶこともある。

強化学習では行動の選択も不確実性を有するものとして扱う。状態 $s_t = s$ のときに $a_t = a$ を選ぶ確率を $\pi(a_t = a|s_t = s)$ と書き、$\pi(a_t|s_t)$ のことを**方策**と言う。強化学習の目標は良い方策を得ることであり、学習の間方策は頻繁に修正される。

方策に不確実性を与える理由の一つに、探索的な学習の促進がある。前述の通り、強化学習では最良の行動を教えてもらえない。そのため、収益 (後述) を多く得られる行動を自ら見つけなければならない。一方で最適な方策は、これまでの学習結果に限って考えれば最適と思われる行動を選択するようなものでなければならない。つまり、学習中の方策にはこれまでの知識の活用と、更に良いと言える行動がないかの探索が求められる。この活用と探索は進化計算における選択と突然変異に似ている。また、方策に不確実性を与えることは明らかに探索的な作用である。もちろん、学習終盤では探索行為を控えなければならない (本来所望していた方策は必ず最適な行動をとるようなものであるため)。活用と探索については後ほどより深く議論する。

もう一つ方策に不確実性を与える理由として、不確実な方策自体が最適と言える場合もあることを述べておく。例えばポーカーの問題を考えよう。ポーカーでは相手を惑わすために、ときに不合理な行動を取る。このような高度な技法をエージェントに学ばせるには、不確実性のある方策を考えなければならない。

さて、これまでの議論より、行動選択を左右する方策、状態遷移を司る環境、そして報酬が強化学習問題の中核を成すことに気付く。これらの中でエージェントが作用できるのは方策のみである。状態は物理系の状態や将棋の盤面に相当するの

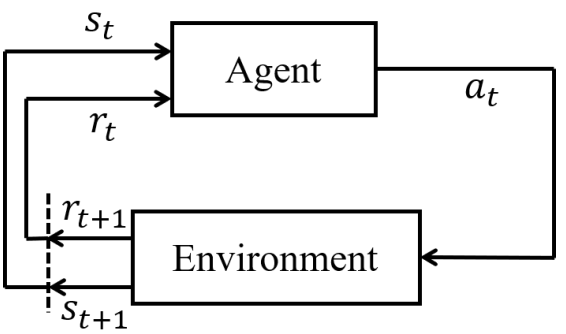

図 1.1　強化学習の概念図。

で、エージェントが決められないのは当然であろう。報酬に関しては、エージェントや環境よりも問題設定者や解析者が決める。エージェントは報酬信号から行動を改めるので、達成したいことに即して私たちが報酬を定義する。

1.1 収益と価値関数

1.1.1 収益

ここまでエージェントは報酬信号を受け取ることは触れてきたが、具体的にこれをどう活用するのかまでは言及してこなかった。ここで報酬信号とは、時刻 t のときにそれ以降で受け取る報酬列 $r_\tau \tau = t+1, t+2, ...$ のことである。当然ながら時刻 t で実施した行動 a_t は報酬信号に影響を与える。

報酬信号が与えられたときに、a_t の良否評価の方法は非常に重要である。仮に即時的な報酬のみを考えたならば、a_t は r_{t+1} の値によってのみ評価される。これは一見良さそうに見えるが、あまり上手くいかないことも多い。

例えば私たちの生活の中で、この日は休むという選択を取ったとする。すると体の疲労は取り除かれ、直感的に良い報酬を得られそうである。ただし、この結果から休むという行動が良いと評価するのは安直かもしれない。もしもその日に大学の必修科目のテストがあれば、来年度に留年という結果が待っているだろう。留年の責任は休むという行動が取るべきで、決して留年が決まった直前の行動ではない。このようにエージェントに高度な目標を達成してほしい場合、エージェントは長期的な報酬も含めて行動の評価をしなければならない。そのため報酬信号を全体的に見ることは重要である。

これを踏まえたとき、評価指標としてまず初めに思い浮かぶのが、報酬の総和

$$R_t = r_{t+1} + r_{t+2} + ... \tag{1.3}$$

であろう。これを**収益**と言う。ただし、報酬の設定によっては R_t は容易に発散する。特に恒常的に動く機械の制御のような問題では、報酬信号は無限に続くので発散しやすい。

そこで、**割引率** $\gamma (0 \leq \gamma \leq 1)$ を導入し、式 (1.3) を

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} ... = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{1.4}$$

のように修正する。これを**割引収益**と言う。$\gamma = 1$ のとき、割引収益は収益と等しくなる。一方で $\gamma = 0$ のとき $R_t = r_{t+1}$ となるので、エージェントは即時的な報酬からのみ学習するようになる。γ がこれらの間を取る場合、割引収益は時刻 t に近いときに得た報酬に重みを持たせつつ、報酬信号全体を加味するようになる。

1.1.2 エピソードタスクと連続タスク

ところで、エージェントが達成しようとする問題をタスクと呼ぶことがあるが、強化学習のタスクには**エピソードタスク**と**連続タスク**の 2 種類がある。

エピソードタスクは終端状態という特別な状態を持つ。これは例えばゲームオーバーやゲームクリアによる終了や、組み立てロボットが作業を完了した状態などである。エピソードタスクの場合、式 (1.4) のように報酬信号が無限に続くことは稀である。

一方で連続タスクは終端状態を持たないタスクのことで、例えばエアコンの温度管理タスクなどが挙げられる。連続タスクの割引収益は式 (1.4) の通りである。また、式 (1.3) の収益は連続タスクの場合特に発散しやすい。

このように、一見するとエピソードタスクと連続タスクで収益の定義を微修正しなければならない気もするが、少しだけエピソードタスクの解釈を拡張すれば、エピソードタスクでも式 (1.4) を使うことができる。例えば $t = T$ で終端状態に達したとき、本来なら $T+1$ 以降の報酬はないはずだが、そこを $r_\tau = 0 (\tau = T+1, T+2, ...)$ と拡張する。そうすればエピソードタスクでも式 (1.4) を使えるし、割引収益の値も変わらない。本資料では特に断りがない限りこれを前提にして、エピソードタスクと連続タスクを統一的に記述する。

1.1.3 価値関数

ところで状態遷移や報酬には不確実性があるので、ある状態のときにある行動を取ったとしても、その後に得られる収益にもバラつきがある。従って、収益を評価指標と述べたが、厳密には「強化学習は期待される収益が最大となるような方策を求める」と言った方が良さそうである。

価値関数とは状態 (もしくは状態行動対) の関数であり、そこから後に得られる収益の期待値を返す。入力である状態 (もしくは状態行動対) から以後の状態遷移と行動選択は方策に依存するため、価値関数も方策に依存する。従って価値関数は方策毎に定義される。

状態のみを入力とする価値関数のことを**状態価値関数**と言い、$V^\pi(s)$ と書く。定義より、$V^\pi(s)$ は

$$V^\pi(s) = \mathrm{E}_\pi \{R_t | s_t = s\} = \mathrm{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s \right\} \tag{1.5}$$

となる。ここで E_π は方策 π に従ったときの期待値を表す。定義より、終端状態の状態価値関数はゼロとなる。次に、状態と行動を入力とする価値関数のことを**行動価値関数**と言い、$Q^\pi(s,a)$ と書く。定義より、$Q^\pi(s,a)$ は

$$Q^\pi(s,a) = \mathrm{E}_\pi \{R_t | s_t = s, a_t = a\} = \mathrm{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a \right\} \tag{1.6}$$

と書き表される。行動価値関数でも、終端状態での値はゼロとなる。

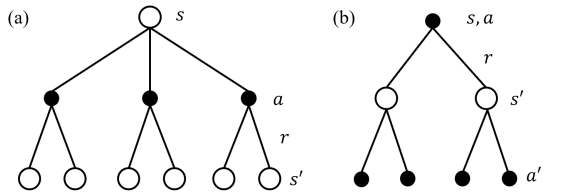

図 1.2　(a) 状態価値関数のバックアップ線図 (b) 行動価値関数のバックアップ線図。

価値関数は経験に基づいて推定することができる。例えば状態価値関数の場合、ある状態 s から方策に従って行動したときに収益 R_t を得たとする。これを何度か経験し、アンサンブル平均から $V^\pi(s)$ を推定できる。行動価値関数に関しても同様である。後述するモンテカルロ法などは価値関数の推定にこのアプローチを取る。一方で遷移確率と報酬関数が既知である場合は、実際に行動せずとも方策毎に価値関数を推定できる。後述する動的計画法はこのアプローチを取る。

1.2　Bellman 方程式

本節では価値関数の再帰的構造に着目して、価値関数に関する方程式を導出する。

まずは状態価値関数について議論しよう。そこで**バックアップ線図**というものを新たに導入する。図 1.2(a) は状態価値関数に関するバックアップ線図である。白いノードは状態を表しており、根の部分の状態 s が今の状態を表している。s と繋がっている黒いノードはそれぞれ一つの行動 a を表している。方策は確率的性質を持つため、s の後に複数の行動 a が繋がっている次第である。s のときに a の行動を取ったとき、エージェントは状態 s' に遷移する。式 (1.1) より、状態遷移も不確実性を有するため、a の後にいくつかの白いノードが繋がる。s' に遷移したとき、環境から報酬 r を貰う。図 1.2 のように行動から状態へと繋がるエッジ上に報酬を記載する。

ここで状態価値関数の方程式を導出する。状態価値関数は定義より、

$$\begin{aligned} V^\pi(s) &= \mathrm{E}_\pi\left\{\sum_{k=0}^\infty \gamma^k r_{t+k+1} | s_t = s\right\} \\ &= \mathrm{E}_\pi\left\{r_{t+1} + \gamma\sum_{k=0}^\infty \gamma^k r_{t+k+2} | s_t = s\right\} \\ &= \sum_{a\in\mathcal{A}(s)} \pi(a|s) \sum_{s'} \mathcal{P}^a_{ss'}\left[\mathcal{R}^a_{ss'} + \gamma \mathrm{E}_\pi\left\{\sum_{k=0}^\infty \gamma^k r_{t+k+2} | s_{t+1} = s'\right\}\right] \\ &= \sum_{a\in\mathcal{A}(s)} \pi(a|s) \sum_{s'} \mathcal{P}^a_{ss'}\left[\mathcal{R}^a_{ss'} + \gamma V^\pi(s')\right] \end{aligned}$$

が成立する。この式は重要なので、始めと最後だけを残した方程式を再度記載する。

$$V^\pi(s) = \sum_{a\in\mathcal{A}(s)} \pi(a|s) \sum_{s'} \mathcal{P}^a_{ss'}\left[\mathcal{R}^a_{ss'} + \gamma V^\pi(s')\right]. \tag{1.7}$$

これを **Bellman 方程式**と言う。Bellman 方程式の意味は図 1.2(a) からよく読み取れる。Bellman 方程式は状態の数だけ存在するため、式 (1.7) から $V^\pi(s)$ に関する連立一次方程式を作ることができる。従って、一般的な連立一次方程式の数値解法で状態価値関数は計算できる。

次に行動価値関数について議論する。図 1.2(b) は行動価値関数に関するバックアップ線図であり、基本的に図 1.2(a) と見方は同じである。ただし、根の部分だけ状態と行動の両方を意味している。これは行動価値関数の定義を考えれば当然であろう。初めに取る行動が確定しているため、s と a のノードを纏めたと考えてもよい。ただし、以後の行動は方策に従って選択されるので、図 1.2(a) のように分岐する。

行動価値関数は以下のように書き換えられる。

$$\begin{aligned} Q^\pi(s,a) &= \mathrm{E}_\pi\left\{\sum_{k=0}^\infty \gamma^k r_{t+k+1} | s_t = s, a_t = a\right\} \\ &= \mathrm{E}_\pi\left\{r_{t+1} + \gamma\sum_{k=0}^\infty \gamma^k r_{t+k+2} | s_t = s, a_t = a\right\} \\ &= \sum_{s'} \mathcal{P}^a_{ss'}\left[\mathcal{R}^a_{ss'} + \gamma \mathrm{E}_\pi\left\{\sum_{k=0}^\infty \gamma^k r_{t+k+2} | s_{t+1} = s'\right\}\right] \\ &= \sum_{s'} \mathcal{P}^a_{ss'}\left[\mathcal{R}^a_{ss'} + \gamma V^\pi(s')\right] \end{aligned}.$$

ここで、状態価値関数と行動価値関数には

$$V^\pi(s) = \sum_{a\in\mathcal{A}(s)} \pi(a|s) Q^\pi(s,a) \tag{1.8}$$

の関係があるため、最終的に

$$Q^\pi(s,a) = \sum_{s'} \mathcal{P}^a_{ss'}\left[\mathcal{R}^a_{ss'} + \gamma \sum_{a'\in\mathcal{A}(s')} \pi(a'|s') Q^\pi(s',a')\right] \tag{1.9}$$

の方程式を得る。これが行動価値関数に対する Bellman 方程式である。

1.2.1 Bellman 最適方程式

これまで強化学習問題の定義と定式化を議論した。特に強化学習は期待収益が最大となるような方策を求めるという視点は重要である。また、収益の期待値として価値関数があり、価値関数は Bellman 方程式に従うことも紹介した。

全ての状態 S に対して $V^{\pi}(s) \geq V^{\pi'}(s)$ であるならば、π は π' よりも優れていると言える (行動価値関数に関しても同様)。実際に強化学習では方策をこのように比較する。他のどれよりも同じか優れている方策を**最適方策**という。最適方策は少なくとも一つ存在することが分かっており、複数存在することもある。方策比較の定義より、最適方策群 π^*(全ての最適方策を同じ π^* で表している) は同じ価値関数を共有する。最適方策における状態価値関数 $V^*(s)$ のことを**最適状態価値関数**と言う。同様に最適方策における行動価値関数のことを**最適行動価値関数**と言う。両者は

$$V^*(s) = \max_{\pi} V^{\pi}(s), \quad Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$$

の等式を満たす。

最適方策の場合、Bellman 方程式を違った形で書くことができる。まず、式 (1.8) より最適方策についても $V^*(s) = \sum \pi^*(a|s) Q^*(s,a)$ が成立する。$V^*(s)$ が最大であるということは、方策 $\pi^*(a|s)$ は s が固定された中で $Q^*(s,a)$ が最大となる a を確実に選択するようなものでなければならない。つまり、

$$\pi^*(a|s) = \begin{cases} 1 & (a = \mathrm{argmax}_a Q^*(s,a)) \\ 0 & (\text{otherwise}) \end{cases} \tag{1.10}$$

だと分かる。このとき $V^*(s) = \max_a Q^*(s,a)$ と書ける訳で、ここから更に

$$V^*(s) = \max_a \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma V^*(s') \right] \tag{1.11}$$

なる関係を導出することができる。これを最適状態価値関数の **Bellman 最適方程式**と言う。最適行動価値関数に関しても同様で、

$$Q^*(s,a) = \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma \max_{a'} Q^*(s',a') \right] \tag{1.12}$$

なる関係が得られる。

状態の数が N であるとき、Bellman 最適方程式も N 個存在する。従って、適当な非線型連立方程式の数値解法を用いれば最適価値関数は求まる。また、式 (1.12) より最適方策も同時に求まることになる。そのため強化学習問題は Bellman 最適方程式の数値解析に帰着させることも可能であろう。