Summer School 数理物理 2021
機械学習の数理

強化学習

森村 哲郎 *

概要

強化学習は意思決定モデルをデータから学習する機械学習の一分野である．教師あり学習や教師なし学習などの従来の機械学習と異なり，強化学習には報酬という概念が登場し，報酬を最大にするような意思決定モデルを学習する．システムに関する知識が十分でなくても，データと報酬から意思決定モデルを学習するため，多岐にわたる領域での応用が期待されている．近年，囲碁やテレビゲームなどで強化学習と深層学習を用いて人間を超えるパフォーマンスを達成できることが示され、多くの人を驚かせた．しかし，実際に動かすと，学習しない，再現性が乏しいなど，強化学習法をブラックボックスとして用いることは困難であることが多い．そこで，本稿では強化学習の原理の理解を目指し，強化学習の根幹となる基礎的な数理を紹介する．まず，1 節でマルコフ決定過程など強化学習が扱う数理モデルを説明し，2 節では強化学習の基礎となるマルコフ決定過程の解法（プランニング）を説明する．3 節ではプランニング手法をサンプル近似することで Q 学習などの代表的な強化学習法を導出する．

目次

* 株式会社サイバーエージェント．e-mail: morimura_tetsuro@cyberagent.co.jp

1　はじめに

強化学習は逐次的意思決定モデルの学習を扱い，登場するモデル（構成要素）として，大まかに「制御対象のシステム」と「学習対象の方策モデル」がある．本節では，はじめに強化学習の概要を紹介し，1.2 項で制御対象のシステムを記述する標準的な数理モデルであるマルコフ決定過程を導入し，1.3 項で方策モデルと呼ばれる意思決定モデルを規定する関数を説明する．最後，1.4 項で逐次的意思決定問題の分類を簡単に紹介する．

なお，本稿は簡便さを優先して厳密性を犠牲にし，内容も基礎的なものにかなり限定している．そのため，厳密な理論に関心がある場合は [6, 5]，価値関数近似や方策勾配法など幅広い内容に興味ある場合は [29, 34]，深層強化学習やマルチエージェント強化学習など発展的な内容に興味がある場合は [32, 35, 11] などを参照されたい．

1.1　強化学習とは

強化学習（reinforcement learning; RL）は意思決定ルールの最適化を目指すという点ではオペレーションズ・リサーチと同じだが，適用するシステムや環境に関する完全な知識を前提とせず，設計者が「何をすべきか（Goal）」を報酬という形でアルゴリズムに入力して，「どのように実現するか（How）」をデータなどから学習するという特徴がある．そのため，システムに関する知識が十分でなくても，（大量に）データを取得できるのであれば，強化学習によって目的を達成するような意思決定ルールを得られる可能性があり，多岐にわたる領域での応用が期待されている．実際に近年，強化学習が決定的な役割を果たす実問題がビジネスインテリジェンスや医療，金融，文章要約，広告などの領域で次々に見出され，さらなる関心を集めている [2, 20, 21, 8, 33, 1, 7, 10, 27, 23, 17]．また，ゲームの分野においての成功も顕著であり，1992 年の Tesauro によるバックギャモンの成功から [30]，最近では囲碁やビデオゲームで強化学習と深層学習などを用いて人間を超えるパフォーマンスを達成できることが示され [19, 25, 24]，多くの人を驚かせている．

意思決定は状態と呼ばれる現在の状況を表すものに基づき行われ，その結果として報酬や新しい状態を観測し，再び意思決定を行うといったことを繰り返す．例として，次のキャンペーン最適化問題を考えよう．なお，逐次的意思決定ルールは**方策**（policy）と呼ばれ，方策の最適化問題のことを**逐次的意思決定問題**（sequential decision-making problem）という．

例 1.1. 小売における値下げセールなどのキャンペーンの実施について単純化した問題を考える．各販売期でキャンペーンを実施するかどうかを決定し，売上の長期平均を最大にすることが目的である．図 1 のように，各期の売上はキャンペーンの実施の有無と顧客の購買意欲に依存し，購買意欲は「低（`low`）」と「中（`mid`）」，「高（`high`）」の 3 段階あるとする．キャンペーンを実施すれば，実施しない場合よりも大きな売上をあげることができる．しかし，需要の先食いが発生し，次の販売期の購買意欲は `low` になり，次期の売上は落ちてしまう．一方，キャンペーンを実施しなければ，購買意欲が `high` の場合は `high` のまま，それ以外は購買意欲が一段階高くなり，次期の売上は増加する．

典型的には，強化学習では，本例の各期の売上を「報酬」，顧客の購買意欲を「状態」，キャンペーン実施の有無を「行動」として扱う．方策は状態に依存し，決定論的とすれば，次のパターンが考えられる．

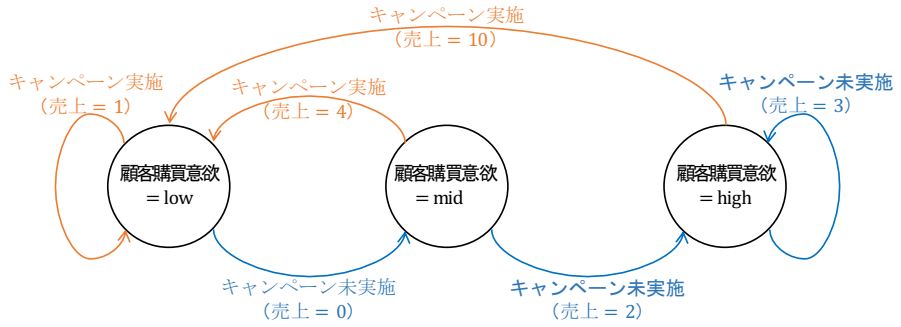

図 1　キャンペーン最適化問題

- 方策 A：キャンペーンを実施し続ける．平均売上は約 1.
- 方策 B：キャンペーンを一切実施しない．平均売上は約 3.
- 方策 C：購買意欲 `mid` でキャンペーンを実施する．平均売上は約 2.
- 方策 D：購買意欲 `high` でキャンペーンを実施する．平均売上は約 4.

平均売上（平均報酬）を最大にする方策は D であり，キャンペーンを乱発せず，顧客の購買意欲が `high` になるまで待って，キャンペーンを実施すべきであることがわかる．一方，近視眼的になり，即時的な売上の良い行動（キャンペーン実施）を取り続ける方策 A は，即時的には最良の選択でも，平均売上という長期評価においては最悪の方策であることがわかる．強化学習は一般に，直近で損をしてでも，トータルで得をするような方策 D を学習することを目指している．　□

例 1.1 は状態数や行動数は少なく，状態の遷移は決定論的で，また各行動に対する応答や報酬を完全に知っているとしていたので，全ての方策を簡単に列挙でき，最適な方策を見つけることをできた．しかし，多くの場合，「制御対象のシステム」の状態行動数は多く，状態遷移は確率的で複雑であり，さらに未知であるから，簡単には解けない．そのような一般の逐次的意思決定問題に対して取り組む数理的枠組みが強化学習である．

ただし，任意の制御対象のシステムに対する学習法を考えることは現実的でなく難しいため，通常はシステムに対して仮定をおく．その典型的な仮定が 1.2 項で紹介する**マルコフ性**である．多くの場合，マルコフ性が成り立つ「状態」よばれる情報を観測できるとする**マルコフ決定過程** に対する学習法を考える．また，本稿では扱わないが，マルコフ決定過程の仮定を緩めて，部分情報しか観測できないとする**部分観測マルコフ決定過程**に対する強化学習法もある [32, 34].

学習対象の方策モデルについても，例 1.1 では，現在の状態のみに依存し，決定論的に行動を選択する簡単なもののみを扱った．しかし，過去の状態や行動などの履歴に依存して行動を選択する方策モデルや，確率的に行動を選択するような複雑なものも考えることができる．では，どこまで複雑な方策モデルを考える必要があるだろうか．1.3 項で，方策モデルを分類し，標準的な問題においては簡単な方策モデルのみを扱えば十分であることを示す．1.4 項では逐次的意思決定の問題設定を整理する．

1.2　マルコフ決定過程

強化学習の数理の基礎になるマルコフ性やマルコフ決定過程を説明する．

1.2.1 確率過程とマルコフ性

サイコロを振ったときの「サイコロの目」のように，ランダム性のある事象の生起しやすさを定量的に示すものが**確率**（probability）であり，とりうる値とその値になる確率が与えられている変数のことを**確率変数**（random variable）という．また，「実際にサイコロを振って出た目」のように，実際に生起した値のことを実現値という．本稿では，確率変数と実現値をアルファベットの大文字 X と小文字 x で区別し，確率変数のとりうる値の集合をカリグラフ体を用いて $\mathcal{X}$ のように書くことにする．

次に**確率過程**（stochastic process）を説明する．サイコロを単発的に振るのではなく，繰り返し振って出てくる目の数列，もしくは目の累積和の数列のように，変数の値が時間とともに確率的に変化するような確率変数の系列のことを確率過程という．そのため，確率過程は時間ステップ t をパラメータとして，$\{X_t, t \in \mathcal{T}\}$ と書くことが多い．ここで，$\mathcal{T}$ は時間ステップ t がとりうる値の集合で，連続時間を扱うため $\mathcal{T}$ を実数集合 $\mathbb{R}$ とする場合もあるが，本稿では次のような離散的な点列からなる集合を考える．

$$\{X_t, t \in \mathbb{N}\} \triangleq X_1, X_2, \ldots \text{（もしくは，} \{X_t, t \in \mathbb{N}_0\} \triangleq X_0, X_1, \ldots\text{）}$$

一般の確率過程では，時間ステップ t の確率変数 X_t が $x \in \mathcal{X}$ をとりうる確率は，

$$\Pr(X_t = x \,|\, X_1 = x_1, \ldots, X_{t-1} = x_{t-1})$$

のように時間ステップ t 以前の全ての実現値に依存する．ここで，Pr(A|B) は事象 B が与えられたときの事象 A の条件付き確率である．一方で，強い制約を課した最も単純な確率過程として，各確率変数 $X_1, X_2, \ldots$ が互いに独立で同一の確率分布に従う場合を考えることも多い．この時，$X_1, X_2, \ldots$ は独立同一分布 (independent and identically distributed; i.i.d.) に従うといい，任意の $x_1, \ldots, x_{t-1}, x \in \mathcal{X}$ に対して，

$$\Pr(X_t = x \,|\, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}) = \Pr(X_k = x), \quad \forall k \in \mathbb{N}$$

が成り立つ．もし手持ちのデータが i.i.d. に従うとみなせるのであれば，データの並びや時系列性の考慮が不要になり，標準的な機械学習や多腕バンディットの方法を利用できるので，一般に扱い易い．しかし，多くの意思決定の問題に対して i.i.d. の仮定を置くことはできず，強化学習では i.i.d. よりも弱い制約である**マルコフ性**（Markov property）を仮定する．

マルコフ性は将来の確率変数の条件付き確率分布が現時間ステップ t の値 x_t のみに依存して，x_t が与えられれば $t-1$ 以前の値 $x_1, \ldots, x_{t-1}$ には依存しない性質のことである．つまり，マルコフ性という特性をもつ確率過程は，任意の $t, k \in \mathbb{N}$ と $x_1, \ldots, x_t, x \in \mathcal{X}$ に対して，

$$\Pr(X_{t+k} = x \,|\, X_1 = x_1, \ldots, X_t = x_t) = \Pr(X_{t+k} = x \,|\, X_t = x_t)$$

を満たす．確率変数 X を状態変数とみなせば，$\Pr(X_{t+1} = x' | X_t = x)$ は状態 x から次ステップで状態 x' に遷移する確率を表すことから，一般に状態遷移確率（state transition probability）とよばれる．また，マルコフ性をもつ確率過程のことをマルコフ過程（Markov process）といい，さらに状態変数のとりうる値が離散的（有限または可算）の場合，**マルコフ連鎖**（Markov chain）という．

このマルコフ性という性質は強化学習法を考えるうえで大切な特徴になる．なぜなら，もしマルコフ性が成り立たないような任意の確率過程を学習の対象にしてしまうと，行動選択の際に考慮す

べき情報が時間ステップ t に対して組合せ的に増大してしまい，一般に扱えなくなるためである．そのため，強化学習を実問題に応用する際は，強化学習法を適用する前に，対象のシステムがマルコフ性を満たすように確率変数を定義するなど確率過程を注意深く設計することが肝要になる．例えば，アタリ（Atari）社のビデオゲームを強化学習で自動操作させる事例では，直近 4 フレームから状態（確率変数 X_t）を定義して，学習の対象となるシステム（確率過程）が概ねマルコフ性を満足するようにして，強化学習法を適用している [19].

1.2.2 マルコフ決定過程

強化学習は行動選択ルールの最適化を扱うため，従来の「状態（state)」のみの確率過程ではなく，行動などを追加した確率制御過程（stochastic control process）と呼ばれる種類の確率過程を考える．

マルコフ連鎖に「行動（action)」と意思決定の良し悪しの基準になる「報酬（reward)」を取り入れた確率制御過程が**マルコフ決定過程**（Markov decision process; MDP）と呼ばれるもので，以下の 5 つ組 $\mathrm{M} \triangleq \{\mathcal{S}, \mathcal{A}, p_{s_0}, p_{\mathrm{T}}, g\}$ で定義される [22].

- 有限状態集合: $\mathcal{S} \triangleq \{1, \dots, |\mathcal{S}|\} \ni s$
- 有限行動集合: $\mathcal{A} \triangleq \{a^1, \dots, a^{|\mathcal{A}|}\} \ni a$
- 初期状態確率関数: $p_{s_0} : \mathcal{S} \to [0, 1]$ ： $p_{s_0}(s) \triangleq \Pr(S_0 = s)$
- 状態遷移確率関数: $p_{\mathrm{T}} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ ：

$$p_{\mathrm{T}}(s'|s, a) \triangleq \Pr(S_{t+1} = s' \,|\, S_t = s, A_t = a), \quad \forall t \in \mathbb{N}_0 \triangleq \{0, 1, \dots\}$$

- 報酬関数: $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

ここで，確率変数 S_t と A_t は時間ステップ $t \in \mathbb{N}_0$ での状態と行動である．また $|\mathcal{X}|$ は，$\mathcal{X}$ が有限集合の場合，$\mathcal{X}$ の要素数を表す．本稿ではこのような有限状態集合，有限行動集合の離散時間マルコフ決定過程を主に扱うが，連続状態空間や連続行動空間，連続時間のマルコフ決定過程に関する強化学習の研究も盛んである．[9, 18, 32, 34]．なお，マルコフ連鎖に報酬のみを追加したマルコフ過程や $|\mathcal{A}| = 1$ のマルコフ決定過程は**マルコフ報酬過程**（Markov reward process）と呼ばれる．

定義から，報酬関数 g は有界関数であり，

$$|g(s, a)| \leq R_{\max}, \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A} \tag{1}$$

を満たす定数 $R_{\max} \in \mathbb{R}$ が存在することを仮定していることになり，報酬の集合 $\mathcal{R}$ を次のように定義する．

$$\mathcal{R} \triangleq \{r \in \mathbb{R} : r = g(s, a), \exists (s, a) \in \mathcal{S} \times \mathcal{A}\}$$

定義上，$\mathcal{R}$ の要素数 $|\mathcal{R}|$ は有限個で，$|\mathcal{R}| \leq |\mathcal{S}||\mathcal{A}|$ を満たす．なお，より一般的な報酬関数として，次状態にも依存するような $\tilde{g}(S_t, A_t, S_{t+1})$ や報酬分布関数 $\Pr(R_t \leq r \,|\, S_t = s, A_t = a)$ などを用いることもあるが，多くの場合，g と同様にして扱うことが可能である．

次に，マルコフ決定過程への入力となる行動の選択ルールを規定する関数を定義しよう．これは**方策**（policy）または政策と呼ばれ，様々な型の方策を考えることができるが，本稿では特に断らない限り，現時間ステップの状態 s のみに依存して確率的に行動を選択する**確率的方策**（stochastic

policy）$\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$:

$$\pi(a|s) \triangleq \Pr(A = a \,|\, S = s) \tag{2}$$

を用いることとする．ここで，方策 π を含めたマルコフ決定過程 M を

$$\mathrm{M}(\pi) \triangleq \{\mathcal{S}, \mathcal{A}, p_{s_0}, p_{\mathrm{T}}, g, \pi\} \tag{3}$$

と表記する．また，任意の確率的方策 π を含む方策集合を

$$\Pi \triangleq \left\{\pi : \mathcal{A} \times \mathcal{S} \to [0, 1] \,:\, \sum_{a \in \mathcal{A}} \pi(a|s) = 1, \forall s \in \mathcal{S}\right\} \tag{4}$$

とおく．他の方策集合や各方策集合の特性や十分性などについては 1.3.2 項で説明する．

ここで，マルコフ決定過程の時間発展 $(s_0, a_0, r_0, \ldots, s_t, a_t, r_t, \ldots)$ の具体的な手順を示す．

マルコフ決定過程 $\mathrm{M}(\pi) = \{\mathcal{S}, \mathcal{A}, p_{s_0}, p_{\mathrm{T}}, g, \pi\}$ の時間発展

0. 時間ステップ t を $t = 0$ と初期化して，初期状態確率 p_{s_0} に従い初期状態 $s_t \sim p_{s_0}$ を観測する
1. 状態 s_t と方策 $\pi(\cdot|s_t)$ から，行動 a_t を選択する
2. 行動 a_t を実行し，その結果として，報酬関数 $g(s_t, a_t)$ により定まる報酬値 r_t と，状態遷移確率 $p_{\mathrm{T}}(\cdot|s_t, a_t)$ により定まる次の状態 s_{t+1} を観測する
3. 時間ステップ t を一つ進め，$t := t + 1$，手順 **1.** に戻る

なお，$\sim$ は左辺の値が右辺の確率分布に従い定まること，もしくは左辺の確率変数が右辺の確率分布に従うことを意味し，$:=$ は右辺から左辺への代入演算子である．

強化学習は以上のようなシステムとの相互作用（図 2）から方策を学習することを考えている．以降は強化学習の一般的な呼び名にあわせて [28]，制御対象のシステムのことを**環境**（environment），制御器や意思決定者を**エージェント**（agent）と呼ぶことにする．

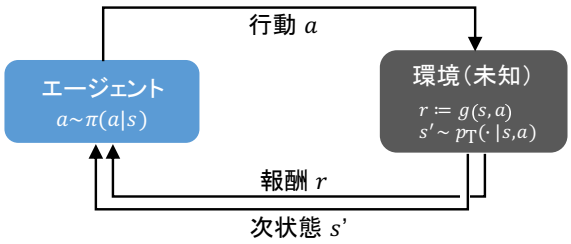

図 2　エージェントと環境の相互作用

1.3　方策

1.3.1 項でマルコフ決定過程における方策の集合を定義し，それらの特徴を 1.3.2 で示す．

1.3.1　方策の分類

式 (2) で定義した確率的方策 π の集合 Π の部分集合として**決定的方策**（deterministic policy）π^{d} の集合 Π^{d} を考えることができる．

$$\Pi^{\mathrm{d}} \triangleq \left\{\pi^{\mathrm{d}} : \mathcal{S} \to \mathcal{A}\right\} \tag{5}$$

なお，π^{d} や Π^{d} の上付き文字 d は deterministic policy の頭文字に由来し，今後も同様の命名法を用いる．なお，$\pi(a|s) := \mathbb{I}_{\{a=\pi^{\mathrm{d}}(s)\}}$ のように π^{d} を確率的方策 π の形式に書きなおすことができるので，Π^{d} は Π に含まれることがわかる．ここで，$\mathbb{I}_{\{\mathrm{B}\}}$ は指示関数であり，B が真なら 1，そうでなければ 0 を出力する．

これまでに導入した方策 π や π^{d} は状態 s のみに依存し，過去の経験とは独立に行動を選択することからマルコフ方策（Markov policy）と呼ばれる分類に属し，また時間ステップ t が進展しても意思決定ルール（方策関数）は変わらないので，マルコフ方策のなかでも**定常なマルコフ方策**（stationary Markov policy）と呼ばれる分類に属す．一方で，一般の**マルコフ方策**として，時間ステップ t の進展に従い方策関数が変化するような非定常な方策系列

$$\boldsymbol{\pi}^{\mathrm{m}} \triangleq \{\pi_0 \in \Pi, \pi_1 \in \Pi, \dots\} \in \boldsymbol{\Pi}^{\mathrm{M}} \tag{6}$$

を考えることができる．また，時間不変の定常な式 (2) や (5) の方策の系列を

$$\boldsymbol{\pi}^{\mathrm{s}} \triangleq \{\pi, \pi, \dots\} \in \boldsymbol{\Pi}^{\mathrm{S}}, \qquad \pi \in \Pi \tag{7}$$

$$\boldsymbol{\pi}^{\mathrm{sd}} \triangleq \{\pi^{\mathrm{d}}, \pi^{\mathrm{d}}, \dots\} \in \boldsymbol{\Pi}^{\mathrm{SD}}, \quad \pi^{\mathrm{d}} \in \Pi^{\mathrm{d}} \tag{8}$$

と定義するが，定常な方策系列を扱っていることが文脈から明らかな場合，簡単化のため，$\boldsymbol{\Pi}^{\mathrm{S}}$ を Π もしくは $\boldsymbol{\Pi}^{\mathrm{SD}}$ を Π^{d} として表記することがある．なお，$\boldsymbol{\Pi}^{\mathrm{S}}$ や $\boldsymbol{\Pi}^{\mathrm{SD}}$ の頭文字 S は stationary（定常）に由来する．

次に，現在の状態だけではなくそれ以前の経験にも依存して行動選択をする非マルコフ方策を定義する．現在の時間ステップ t までの全ての経験の履歴

$$\{s_0, a_0, \dots, s_{t-1}, a_{t-1}, s_t\} \triangleq h_t \in \mathcal{H}_t \tag{9}$$

に基づいて行動選択確率を決めるような**履歴依存の方策**（history-dependent policy）$\pi_t^{\mathrm{h}} : \mathcal{A} \times \mathcal{H}_t \to [0, 1]$,

$$\pi_t^{\mathrm{h}}(a|h_t) \triangleq \Pr(A = a \,|\, H_t = h_t) \tag{10}$$

である．ここで，時間ステップ t の履歴の確率変数，実現値，集合をそれぞれ H_t, h_t, $\mathcal{H}_t$ と記している．また，任意の π_t^{h} を含む方策集合を

$$\Pi_t^{\mathrm{h}} \triangleq \left\{ \pi_t^{\mathrm{h}} : \mathcal{A} \times \mathcal{H}_t \to [0, 1] \,:\, \sum_{a \in \mathcal{A}} \pi_t^{\mathrm{h}}(a|h_t) = 1 \right\} \tag{11}$$

と表記し，方策 π_t^{d} の系列とその集合を次のように定義する．

$$\boldsymbol{\pi}^{\mathrm{h}} \triangleq \{\pi_0^{\mathrm{h}}, \pi_1^{\mathrm{h}}, \dots\} \in \boldsymbol{\Pi}^{\mathrm{H}} \triangleq (\Pi_t^{\mathrm{h}})_{t \in \mathbb{N}_0} \tag{12}$$

ここで留意すべきは，方策集合 Π_t^{h} は時間ステップ t までに知り得る情報を条件にする任意の方策を含むので，時間ステップ t で考えられる全ての方策を含むということである．

以上をまとめると，各方策系列の集合について，図 3 のような，

$$\boldsymbol{\Pi}^{\mathrm{SD}} \subseteq \boldsymbol{\Pi}^{\mathrm{S}} \subseteq \boldsymbol{\Pi}^{\mathrm{M}} \subseteq \boldsymbol{\Pi}^{\mathrm{H}} \tag{13}$$

という包含関係がある．なお，ここで定義した方策系列以外にも，「履歴依存の決定的方策系列」や「非定常な決定的マルコフ方策系列」という方策系列も考えられる．

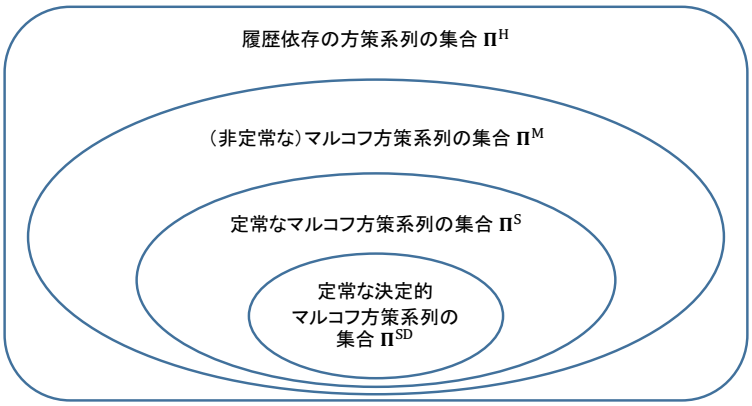

図 3　方策系列の集合の関係性

1.3.2　方策の特徴

式 (13) から，方策系列を引数とする任意の目的関数 f に対して，

$$\max_{\boldsymbol{\pi}\in\mathbf{\Pi}^{\mathrm{SD}}} f(\boldsymbol{\pi}) \leq \max_{\boldsymbol{\pi}\in\mathbf{\Pi}^{\mathrm{S}}} f(\boldsymbol{\pi}) \leq \max_{\boldsymbol{\pi}\in\mathbf{\Pi}^{\mathrm{M}}} f(\boldsymbol{\pi}) \leq \max_{\boldsymbol{\pi}\in\mathbf{\Pi}^{\mathrm{H}}} f(\boldsymbol{\pi}) \tag{14}$$

が成立するので，最適化問題

$$\operatorname*{argmax}_{\boldsymbol{\pi}\in\mathbf{\Pi}^{\mathrm{H}}} f(\boldsymbol{\pi})$$

を解きたくなる．しかし，$\mathbf{\Pi}^{\mathrm{H}}$ のサイズは時間ステップ数に対して組合せ的爆発してしまうため，一般に $\mathbf{\Pi}^{\mathrm{H}}$ についての最適化問題を扱うことは困難である．そして，実は次の命題から，(報酬が履歴全体に依存しない限り) $\mathbf{\Pi}^{\mathrm{H}}$ を扱う必要はなく，マルコフ方策集合 $\mathbf{\Pi}^{\mathrm{M}}$ を考えれば十分であることがわかる．

命題 1. マルコフ方策の妥当性
任意のマルコフ決定過程 M と履歴依存の方策系列 $\boldsymbol{\pi}^{\mathrm{h}} = \{\pi_0^{\mathrm{h}}, \pi_1^{\mathrm{h}}, \dots\} \in \mathbf{\Pi}^{\mathrm{H}}$ に対して，次を満たすようなマルコフ方策の系列 $\boldsymbol{\pi}^{\mathrm{m}} = \{\pi_0^{\mathrm{m}}, \pi_1^{\mathrm{m}}, \dots\} \in \mathbf{\Pi}^{\mathrm{M}}$ が存在する．

$$\Pr(S_t = s, A_t = a \,|\, \mathrm{M}(\boldsymbol{\pi}^{\mathrm{h}})) = \Pr(S_t = s, A_t = a \,|\, \mathrm{M}(\boldsymbol{\pi}^{\mathrm{m}})), \quad \forall (t, s, a) \in \mathbb{N}_0 \times \mathcal{S} \times \mathcal{A} \tag{15}$$

略証：履歴依存の方策系列 $\boldsymbol{\pi}^{\mathrm{h}}$ に従い時間進展するマルコフ決定過程 $\mathrm{M}(\boldsymbol{\pi}^{\mathrm{h}})$ について，各時間ステップ $t \in \mathbb{N}_0$ で到達確率が 0 でない状態

$$s \in \mathcal{S}_t \triangleq \left\{ s \in \mathcal{S} \ : \ \Pr(S_t = s \,|\, \mathrm{M}(\boldsymbol{\pi}^{\mathrm{h}})) > 0 \right\}$$

に対して，

$$\pi_t^{\mathrm{m}\star}(a|s) \triangleq \frac{\Pr(S_t = s, A_t = a \,|\, \mathrm{M}(\boldsymbol{\pi}^{\mathrm{h}}))}{\Pr(S_t = s \,|\, \mathrm{M}(\boldsymbol{\pi}^{\mathrm{h}}))}, \quad \forall a \in \mathcal{A} \tag{16}$$

という行動選択確率をもつマルコフ方策の系列を $\boldsymbol{\pi}^{\mathrm{m}\star} \triangleq \{\pi_0^{\mathrm{m}\star}, \pi_1^{\mathrm{m}\star}, \dots\}$ と定義する．帰納法を用いて，$\boldsymbol{\pi}^{\mathrm{m}\star}$ が式 (15) を満たすことを示す [34]. □

命題 1 から，報酬関数 g の引数である S_t と A_t の同時確率については，どのような履歴依存の方策もマルコフ方策で正確に再現できることがわかる．そして，この同時確率の一致性を達成するマ

ルコフ方策の構成方法が式 (16) ということである．だたし，命題 1 は系列全体 $(S_0, A_0, \ldots, S_t, A_t)$ の同時確率の一致性までは言及しておらず，あくまでも (S_t, A_t) の同時周辺確率

$$\Pr(S_t, A_t) = \sum_{s_0 \in \mathcal{S}} \sum_{a_0 \in \mathcal{A}} \cdots \sum_{s_{t-1} \in \mathcal{S}} \sum_{a_{t-1} \in \mathcal{A}} \Pr(S_0 = s_0, A_0 = a_0, \ldots, S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, S_t, A_t)$$

の一致性に関する結果である．系列全体の生起確率は一致しない反例については [22, 34] などを参照されたい．

命題 1 から，初期状態が s_0 の場合の S_t, A_t の同時周辺確率関数 $\varphi_t^{\boldsymbol{\pi}} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ を

$$\begin{aligned} \varphi_t^{\boldsymbol{\pi}}(s, a | s_0) &\triangleq \Pr(S_t = s, A_t = a \,|\, S_0 = s_0, \mathrm{M}(\boldsymbol{\pi})) \\ &= \mathbb{E}\big[\mathbb{I}_{\{S_t = s\}} \mathbb{I}_{\{A_t = a\}} \,|\, S_0 = s_0, \mathrm{M}(\boldsymbol{\pi})\big] \end{aligned} \tag{17}$$

と定義すれば，次の結果を得る[*1]．

系 2. マルコフ方策の十分性

方策 $\boldsymbol{\pi}$ の目的関数 f を同時周辺確率関数の系列 $\varphi_0^{\boldsymbol{\pi}}, \varphi_1^{\boldsymbol{\pi}}, \ldots$ を引数とする関数 $\tilde{f}$ を用いて，任意の $\boldsymbol{\pi} \in \boldsymbol{\Pi}^{\mathrm{H}}$ に対して，

$$f(\boldsymbol{\pi}) = \tilde{f}(\varphi_0^{\boldsymbol{\pi}}, \varphi_1^{\boldsymbol{\pi}}, \ldots)$$

のように書くことができるとき，次が成り立つ．

$$\max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}^{\mathrm{M}}} f(\boldsymbol{\pi}) = \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}^{\mathrm{H}}} f(\boldsymbol{\pi}) \tag{18}$$

系 2 から，系 2 の条件を満たすタイプの目的関数については履歴依存の非マルコフ方策 $\boldsymbol{\Pi}^{\mathrm{H}}$ までを考える必要性はなく，マルコフ方策 $\boldsymbol{\Pi}^{\mathrm{M}}$ のみを最適化対象にすれば十分であるといえる．

1.4 逐次的意思決定問題の分類

1.4.1 項で逐次的意思決定の問題設定し，1.4.2 項でマルコフ決定過程の分類する．

1.4.1 問題設定

逐次的意思決定問題で調整できるものは方策 π であり，環境モデルであるマルコフ決定過程 $\mathrm{M} \triangleq \{\mathcal{S}, \mathcal{A}, p_{s_0}, p_{\mathrm{T}}, g\}$ は一般に時間不変とされる．もし環境モデルが既知なら，次節で示すように，データが無くても，環境モデルから方策を最適化することが可能である．そのため，データから方策を学習する場合と区別することが多く，環境モデルから最適方策を求めることを，**学習**（learning）といわず，**プランニング**（planning）もしくは**プランニング問題**ということが多い．

一方，環境モデルが未知の場合，プランニングの場合とは異なり，従来の最適化ソルバーをそのまま適用できるような最適化問題として定式化できず，データ（環境との相互作用の結果）からの学習が必要になる．本稿では，環境モデルが未知の場合の方策の学習問題を**強化学習問題**（reinforcement learning problem）と呼ぶことにする．ここで，逐次的意思決定問題の分類を図 4 に示す．

[*1] $\mathrm{M} = \{\mathcal{S}, \mathcal{A}, p_{s_0}, p_{\mathrm{T}}, g\}$ で初期状態確率 p_{s_0} を含むにも関わらず，式 (17) の Pr や $\mathbb{E}$ の条件部は $(S_0 = s, \mathrm{M}(\pi))$ となっていて，初期状態変数 S_0 に関する 2 つの条件 $S_0 = s$ と $S_0 \sim p_{s_0}$ を持つことになるが，本稿では条件部で左側に書いてある条件（$S_0 = s$）を優先する．

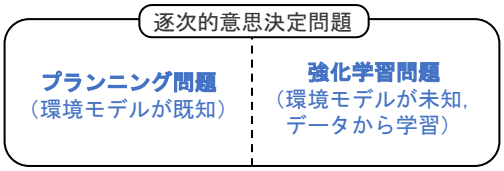

図 4　逐次的意思決定問題の分類

強化学習問題の設定として大きく 2 つある．一つは従来の機械学習と似た設定で，環境との相互作用などから得たデータが大量にあって，そのデータから方策を学習する**バッチ学習**（batch learning）である．バッチ学習はオフライン学習と呼ばれることもある．もう一つは逐次的に環境と相互作用してデータを収集しながら学習する**オンライン学習**（online learning）である．後者の場合，次の 2 つの意思決定戦略があり，それらのバランスを考慮する必要がある．このことを**探索と活用のトレードオフ**（exploration-exploitation trade-off）という．

- データ収集・探索（exploration）：データが十分でないという立場から，環境についての不確実性を減らし，新たな発見をできるように行動を選択する戦略
- データ活用（exploitation）：データは既に十分にあるという立場から，データから最良と判断できる行動を選択する戦略

1.4.2　マルコフ決定過程の単一化

対象とする逐次的意思決定問題の設定によって，次のように終了条件の異なるマルコフ決定過程が考えられる．

(A)　ゴール状態があり，ゴール状態に到達したら終了する
(B)　予め決められた時間ステップになったら終了する
(C)　終了しない（無限時間長のマルコフ決定過程）

以下のように (A) と (B) のマルコフ決定過程のもつ意味を変えずに，表現型を少し変更するだけで，(C) のマルコフ決定過程として再定式化することができる．

移動経路問題をそのまま定式化すれば (A) のマルコフ決定過程になるが，ゴール状態に遷移したらマルコフ決定過程を終了させるのではなく，次時間ステップも同ゴール状態に確率 1 で遷移し，その報酬を 0 とすれば，(C) の無限時間長のマルコフ決定過程に変換することができる．なお，このように他の状態に遷移しない状態のことを**吸収状態**（absorbing state）という．

(B) のマルコフ決定過程としては，四半期など期限が予め決められたもとでのトータルの売上の最大化などがある．もし経過時間の情報を状態に取り込み状態を拡張し，規定の終了時間ステップにれば吸収状態に遷移するマルコフ決定過程を考えれば，それは (C) のマルコフ決定過程に対応する．例えば，もとの状態が $\mathcal{S} \triangleq \{i, j, k\}$ の 3 つあり，終了時間ステップ T が 2 の場合，吸収状態を z として，拡張した状態集合は $\tilde{\mathcal{S}} = \{i_0, j_0, k_0, i_1, j_1, k_1, z\}$ となる．つまり，拡張後の状態サイズは $|\mathcal{S}| \times T + 1$ のように終了時間ステップ T に比例して大きくなる．

以上より，大抵のマルコフ決定過程を (C) の無限時間長のマルコフ決定過程と見なすことできるので，本稿では主に (C) を扱うこととする．

2 プランニング

環境が既知の場合の逐次的意思決定問題であるプランニング問題を扱う．プランニング問題の特徴や解法は，環境が未知である強化学習問題を扱うための基礎になる．例えば，プランニング問題の特徴を調べることで，環境が未知の場合においても，どのクラスの方策までを扱えば良いかを知ることができる．また，次節で示すが，プランニング方法の確率的近似することで TD 法や Q 学習法など強化学習の代表的な方法導出される．

本節では，2.1, 2.2 項でリターンや価値関数を定義し目的関数を与える．2.3 項では，動的計画法（具体的な方法ではなく方法の総称）の数理を紹介し，最適方策の必要十分条件も示す．そして，動的計画法の実装例として， 価値反復法と方策反復法を 2.4 項を説明する．

2.1 リターンと価値関数

2.1.1 項でリターン $C \in \mathbb{R}$ と呼ばれる確率変数を定義し，2.1.2 項で強化学習で大切な役割を果たす価値関数とベルマン方程式を説明する．

2.1.1 リターン

リターン C_t は時間ステップ t から得られる報酬を指数減衰させて累積したもので，

$$\begin{aligned} C_t &\triangleq \lim_{K\to\infty} \sum_{k=0}^{K} \gamma^k g(S_{t+k}, A_{t+k}) \\ &= \lim_{K\to\infty} \sum_{k=0}^{K} \gamma^k R_{t+k} \end{aligned} \tag{19}$$

と定義され，割引累積報酬（discounted cumulative reward）とも呼ばれる．ここで，$\gamma \in [0, 1)$ は**割引率**（discount rate）と呼ばれる**ハイパーパラメータ**（hyper-parameter）である．ハイパーパラメータとは学習によって調整されるものではなく，課題の目的に応じて予め人が設定するパラメータのことである．短期的なリターンを考慮したいのであれば γ を小さく，長期的なリターンを考慮したいのであれば，γ を 1 に近づける．リターン C は状態遷移や方策の確率分布に依存して，確率的に値が決まるので確率変数である．なお，リターンの実現値を c と書くことにする．ここで注目すべきは，式 (19) のリターンの定義から，

$$\begin{aligned} C_t &= R_t + \lim_{K\to\infty} \sum_{k=1}^{K} \gamma^k R_{t+k} \\ &= R_t + \gamma C_{t+1} \end{aligned} \tag{20}$$

のようにリターンは再帰的な構造を持つことである．この構造は後述のベルマン方程式などに利用される．また，報酬関数は定義より，報酬は有界 $|R| \leq R_{\max}$（式 (1)）だから，リターンも次の通り有界である．

$$|C_t| \leq \sum_{k=0}^{\infty} \gamma^k R_{\max} = \frac{R_{\max}}{1-\gamma}, \quad \forall t \in \mathbb{N}_0$$

2.1.2 価値関数とベルマン方程式

引数が状態 $s \in \mathcal{S}$ であり，初期状態 S_0 を s として，方策 $\boldsymbol{\pi} \in \boldsymbol{\Pi}^{\mathrm{H}}$ に従った場合の条件付きリターン C_0 の期待値を返す関数のことを**価値関数**（value function）$V^{\boldsymbol{\pi}}\colon \mathcal{S} \to \mathbb{R}$ という．

$$V^{\boldsymbol{\pi}}(s) \triangleq \mathbb{E}^{\boldsymbol{\pi}}[C_0 \,|\, S_0 = s] \tag{21}$$

ここで，$\mathbb{E}^{\boldsymbol{\pi}}$ は $\mathrm{M}(\boldsymbol{\pi})$ で条件付けされた期待値演算子である．

$$\mathbb{E}^{\boldsymbol{\pi}}[\,\cdot\,] \triangleq \mathbb{E}[\,\cdot \,|\, \mathrm{M}(\boldsymbol{\pi})]$$

なお，$\mathbb{E}$ は期待値演算子であり，$\mathbb{E}[X|Y]$ は条件 Y が与えられた時の確率変数 X の期待値を表す．

同様にして，状態 s で行動 a を選択し，あとは方策 $\boldsymbol{\pi}$ に従った場合の期待リターンを返す関数を

$$Q^{\pi}(s,a) \triangleq \mathbb{E}^{\pi}[C_0 \,|\, S_0 = s, A_0 = a] \tag{22}$$

と定義する．これは**行動価値関数**（action value function）もしくは **Q 関数**と呼ばれる．

方策を定常なマルコフ方策 $\pi \in \Pi$ に限定すれば，式 (20) のリターンの再帰性から，

$$\begin{aligned} V^{\pi}(s) &= \mathbb{E}^{\pi}[R_t + \gamma C_{t+1} \,|\, S_0 = s] \\ &= \sum_{a \in \mathcal{A}} \pi(a|s) \left(g(s,a) + \gamma \sum_{s' \in \mathcal{S}} p_{\mathrm{T}}(s'|s,a) \mathbb{E}^{\pi}[C_{t+1} \,|\, S_{t+1} = s'] \right), \quad \forall s \in \mathcal{S} \end{aligned}$$

なので，次の価値関数 V^{π} に関する再帰式を得る．

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(g(s,a) + \gamma \sum_{s' \in \mathcal{S}} p_{\mathrm{T}}(s'|s,a) V^{\pi}(s') \right), \quad \forall s \in \mathcal{S} \tag{23}$$

これは**ベルマン期待方程式**（Bellman expectation equation）もしくは単に**ベルマン方程式**（Bellman equation）と呼ばれ，多くの強化学習法の基礎になる．

2.2 目的関数と最適価値関数

プランニング問題を具体的に定式化するため，目的関数を 2.2.1 項で定義する．2.2.2 項では，目的関数の最適値と価値関数の最適性の関係性を確認し，マルコフ方策のみを扱えば十分なことや，最適価値関数やベルマン最適方程式を紹介する．

2.2.1 目的関数

逐次的意思決定問題は一般にリターン C に関する何かしらの統計量 $\mathcal{F}[C \,|\, \mathrm{M}(\boldsymbol{\pi})]$ を用いて目的関数（objective function）$f : \boldsymbol{\Pi} \to \mathbb{R}$：

$$f(\boldsymbol{\pi}) \triangleq \mathcal{F}[C \,|\, \mathrm{M}(\boldsymbol{\pi})]$$

（や制約条件）を定義して，方策についての最適化問題として定式化される．つまり，制約無しの逐次的意思決定問題は最適方策

$$\boldsymbol{\pi}^* \triangleq \underset{\boldsymbol{\pi} \in \boldsymbol{\Pi}}{\mathrm{argmax}} \{ f(\boldsymbol{\pi}) \} \tag{24}$$

の探索問題である．なお，最適方策の定義上，最適方策は一つとは限らず，複数存在し得る．

目的関数の例として，時間ステップ $t=0$ からのリターン C_0 の期待値

$$\begin{aligned} f_0(\boldsymbol{\pi}) &\triangleq \mathbb{E}^{\boldsymbol{\pi}}[C_0] \\ &= \sum_{s\in\mathcal{S}} p_{s_0}(s)\, V^{\boldsymbol{\pi}}(s) \end{aligned} \tag{25}$$

がある．少し一般化して，重み関数 $w:\mathcal{S}\to\mathbb{R}_{\geq 0}$ による価値関数の重み付き和を

$$f_{\mathrm{w}}(\boldsymbol{\pi}; w) \triangleq \sum_{s\in\mathcal{S}} w(s) V^{\boldsymbol{\pi}}(s) \tag{26}$$

と定義することもある．

本節では目的関数として，ひとまず，価値関数の重み付き和 f_{w}（式 (26)）の特別な場合である，状態 s の期待リターン

$$f_{\mathrm{w}}(\boldsymbol{\pi}\,;\, \boldsymbol{e}_s^{|\mathcal{S}|}) = \mathbb{E}^{\pi}[C_0 \,|\, S_0 = s] = V^{\boldsymbol{\pi}}(s) \tag{27}$$

を考える．ここで，$\boldsymbol{e}_m^n$ は第 m 要素が 1 で他の要素はゼロの n 次元ベクトルである．なお，この目的関数はある特定の状態 s から開始する $\mathrm{M}(\boldsymbol{\pi})$ の期待リターンに対応する．

2.2.2 最適価値関数とベルマン最適方程式

2.1.2 項で，方策 $\boldsymbol{\pi}\in\boldsymbol{\Pi}^{\mathrm{H}}$ の期待リターンとして価値関数 $V^{\boldsymbol{\pi}}(s)\triangleq\mathbb{E}[C_0\,|\,S_0=s,\mathrm{M}(\boldsymbol{\pi})]$ を定義した．ここでは，**最適価値関数**（optimal value function）$V^*:\mathcal{S}\to\mathbb{R}$ を定義する．

$$V^*(s) \triangleq \max_{\boldsymbol{\pi}\in\boldsymbol{\Pi}^{\mathrm{H}}} V^{\boldsymbol{\pi}}(s) \tag{28}$$

最適価値関数はその定義から式 (27) の目的関数の最適値に一致する．

式 (17) の同時周辺確率 $\varphi_t^{\pi}(s,a|s_0)$ を用いれば，価値関数 $V^{\boldsymbol{\pi}}$ を

$$\begin{aligned} V^{\boldsymbol{\pi}}(s_0) &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}^{\boldsymbol{\pi}}[g(S_t,A_t)\,|\,S_0=s_0] \\ &= \sum_{t=0}^{\infty}\gamma^t \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \varphi_t^{\boldsymbol{\pi}}(s,a|s_0) g(s,a), \quad \forall\boldsymbol{\pi}\in\boldsymbol{\Pi}^{\mathrm{H}}, \forall s_0\in\mathcal{S} \end{aligned} \tag{29}$$

のように書くことができ，価値関数はマルコフ方策の十分性に関する系 2 の条件を満たすことがわかる．よって，系 2 より，最適価値関数を

$$V^*(s) \triangleq \max_{\boldsymbol{\pi}\in\boldsymbol{\Pi}^{\mathrm{H}}} V^{\boldsymbol{\pi}}(s) = \max_{\boldsymbol{\pi}\in\boldsymbol{\Pi}^{\mathrm{M}}} V^{\boldsymbol{\pi}}(s), \quad \forall s\in\mathcal{S} \tag{30}$$

と書き直すことができ，最適化対象の方策の集合として非マルコフ方策集合 $\boldsymbol{\Pi}^{\mathrm{H}}$ までを扱う必要はなく，マルコフ方策集合 $\boldsymbol{\Pi}^{\mathrm{M}}$ を考えれば十分である[*2]．

式 (23) のベルマン期待方程式の場合と同様に，式 (20) のリターンの再帰性（$C_t = R_t + \gamma C_{t+1}$）

[*2] 後述の命題 9（最適方策の必要十分条件）で示すが，実は $\boldsymbol{\Pi}^{\mathrm{M}}$ よりも単純な方策である定常な決定的マルコフ方策集合 Π^{d} でも最適価値関数を達成することができる．つまり，$V^*(s)=\max_{\pi\in\Pi^{\mathrm{d}}} V^{\pi}(s)$ が成立する．

を利用すれば，式 (30) は

$$
\begin{aligned}
V^*(s) &= \max_{\boldsymbol{\pi}\in\boldsymbol{\Pi}^{\mathrm{M}}} V^{\boldsymbol{\pi}}(s) = \max_{\boldsymbol{\pi}\in\boldsymbol{\Pi}^{\mathrm{M}}} \mathbb{E}^{\boldsymbol{\pi}}[C_0 \,|\, S_0 = s] \\
&= \max_{\boldsymbol{\pi}\in\boldsymbol{\Pi}^{\mathrm{M}}} \mathbb{E}^{\boldsymbol{\pi}}[g(s, A_0) + \gamma C_1 \,|\, S_0 = s] \\
&= \max_{\pi_0\in\Pi} \mathbb{E}^{\pi_0}\Bigg[g(s, A_0) \\
&\qquad + \gamma \max_{\{\pi_1,\pi_2,\dots\}\in\boldsymbol{\Pi}^{\mathrm{M}}} \mathbb{E}^{\{\pi_1,\pi_2,\dots\}}[C_1 \,|\, S_1 \sim p_{\mathrm{T}}(\cdot|s, A_0)] \,\Big|\, S_0 = s \Bigg] \\
&= \max_{\pi_0\in\Pi} \sum_{a_0\in\mathcal{A}} \pi_0(a_0|s) \Bigg(g(s, a_0) + \gamma \sum_{s_1\in\mathcal{S}} p_{\mathrm{T}}(s_1|s, a_0) V^*(s_1) \Bigg)
\end{aligned}
$$

となる．よって，次の V^* に関する再帰式が成り立つことがわかる．

$$
V^*(s) = \max_{a\in\mathcal{A}} \left\{ g(s, a) + \gamma \sum_{s'\in\mathcal{S}} p_{\mathrm{T}}(s'|s, a) V^*(s') \right\}, \quad \forall s \in \mathcal{S} \tag{31}
$$

これは**ベルマン最適方程式**（Bellman optimality equation）と呼ばれる．式 (31) は以下の**最適性の原理**（principle of optimality）が成り立っていることと等価であり，次項で紹介する動的計画法を適用する上で必要な構造である．

最適性の原理（principle of optimality）[3, 4] [*3]

時間ステップ T までの逐次的意思決定問題の最適方策系列を $\boldsymbol{\pi}^* = \{\pi_0^*, \pi_1^*, \dots, \pi_T^*\}$ とし，$\boldsymbol{\pi}^*$ に従い意思決定をして，時間ステップ t で到達可能な状態集合を $\tilde{\mathcal{S}}_t$ とする．この時，任意の $t \in \{0, \dots, T\}$ と $s \in \tilde{\mathcal{S}}_t$ について，時間ステップ t で状態 s から再開し，時間ステップ T までの元の問題の部分問題を考えた場合，$\boldsymbol{\pi}^*$ の部分系列 $\{\pi_t^*, \dots, \pi_T^*\}$ が最適解（の一つ）であることを最適性の原理という．

本稿では以降，簡便化のため，ベルマン期待方程式（式 (23)) とベルマン最適方程式との区別が特に不要な場合，両者を総称して**ベルマン方程式**（Bellman equation）ということがある．

2.3 動的計画法の数理

動的計画法（dynamic programming; DP）とは帰納的な計算手段（inductive computation）に基づき，逐次的意思決定の最適化問題を解くアプローチのことである [22, 5]．動的計画法は 2.2.2 項で示した**最適性の原理**の性質をもつ逐次的意思決定問題に対するアプローチで，最適性の原理を利用して最適化問題を部分問題に分割して，部分問題を再帰的に繰り返し解くことにより最適解を求める．問題を部分問題に分割して解くことで，最適解を一括で求めようとする場合に比べ，大幅に解の候補を絞ることができ，一般に効率よく最適解を求めることができる．ただし，最適化問題

[*3] 初期の「最適性の原理」の記述として，Bellman (1957, p.83)［3] によるものを以下に引用する．

> “An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.”

（目的関数や制約条件）の設定によっては最適性の原理が成り立たず，その場合は動的計画法で求まる解と真の最適解は一般に一致しない．

本項では，2.3.1 項で動的計画法の各反復の操作の基本となるベルマン作用素を導入し，2.3.2 項で動的計画法を数理的に紐解く上で必要な結果を示し，2.3.3 項でその数理的な性質を議論する．2.3.4 項では，最適方策を定義し，その必要十分条件などを示す．

2.3.1 ベルマン作用素

動的計画法は，一般に，**ベルマン作用素**（Bellman operator）や**動的計画写像**（DP mapping）と呼ばれる処理 B を関数 v に適用し，v の更新

$$v := \mathsf{B}(v) \tag{32}$$

を繰り返し，徐々に v を価値関数や最適価値関数に近づける [5]．ここで，ベルマン作用素として，状態関数 $v : \mathcal{S} \to \mathbb{R}$ に対する 2 つの作用素 B_π と B_* を定義する．

まず，価値関数 V^π についてのベルマン期待方程式（式 (23)）に対応する方策 π で条件付けされた**ベルマン期待作用素**（Bellman expectation operator）$\mathsf{B}_\pi : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ として次がある．

$$(\mathsf{B}_\pi(v))(s) \triangleq \sum_{a\in\mathcal{A}} \pi(a|s) \left(g(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_{\mathrm{T}}(s'|s,a) v(s') \right), \quad \forall s \in \mathcal{S}. \tag{33}$$

以後，簡便化のため，関数についての作用素を $\mathsf{B}_\pi(v) \triangleq \mathsf{B}_\pi v$ のように表記する．なお，$\mathsf{B}_\pi v(s)$ は，初期状態が s のマルコフ決定過程 $\mathrm{M}(\pi)$ が時間ステップ $t = 1$ で終了し，その終了時の状態 s' の価値を $v(s')$ とした際の時間ステップ $t = 0$ の s の価値に対応すると解釈できる．

もう一つのベルマン作用素として，式 (31) のベルマン最適方程式に対応する**ベルマン最適作用素**（Bellman optimality operator）$\mathsf{B}_* : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$

$$(\mathsf{B}_* v)(s) \triangleq \max_{a\in\mathcal{A}} \left\{ g(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_{\mathrm{T}}(s'|s,a) v(s') \right\}, \quad \forall s \in \mathcal{S} \tag{34}$$

がある．なお，ベルマン期待作用素 B_π とベルマン最適作用素 B_* の区別が特に必要ない場合，本稿ではそれらを単にベルマン作用素 B と呼ぶことがある．

次に，ベルマン作用素の繰り返し適用について確認する．まず表記法だが，方策 $\pi_0, \pi_1, \ldots, \pi_{k-1}$ のベルマン期待作用素 $\mathsf{B}_{\pi_0}, \mathsf{B}_{\pi_1}, \ldots, \mathsf{B}_{\pi_{k-1}}$ を $\mathsf{B}_{\pi_{k-1}}$ から逐次的に関数 $v : \mathcal{S} \to \mathbb{R}$ に適用する場合，

$$(\mathsf{B}_{\pi_0}(\mathsf{B}_{\pi_1}(\ldots(\mathsf{B}_{\pi_{k-1}} v)\ldots)) \triangleq \mathsf{B}_{\pi_0}\mathsf{B}_{\pi_1}\ldots\mathsf{B}_{\pi_{k-1}} v$$

もしくは $\boldsymbol{\pi} = \{\pi_0, \pi_1, \ldots, \pi_{k-1}\}$ として，

$$\mathsf{B}_{\pi_0}(\mathsf{B}_{\pi_1}(\ldots(\mathsf{B}_{\pi_{k-1}} v)\ldots)) \triangleq \mathsf{B}_{\boldsymbol{\pi}}^k v$$

と表記する．なお，同一の B_π の繰り返し適用の場合は，単に $\mathsf{B}_\pi(\mathsf{B}_\pi v) \triangleq \mathsf{B}_\pi^2 v$ のように表記する．ベルマン最適作用素についても同様の表記法を用いることとする．

ベルマン作用素の繰り返し適用の具体例として，$\boldsymbol{\pi} = \{\pi_0, \pi_1\}$ のベルマン期待作用素 $\mathsf{B}_{\boldsymbol{\pi}}^2$ の関数

v への適用は

$$
\begin{aligned}
(\mathsf{B}_{\boldsymbol{\pi}}^2 v)(s) &= (\mathsf{B}_{\pi_0}(\mathsf{B}_{\pi_1} v))(s) \\
&= \sum_{a\in\mathcal{A}} \pi_0(a|s) \Bigg(g(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_{\mathrm{T}}(s' \mid s,a) \\
&\qquad \times \sum_{a'\in\mathcal{A}} \pi_1(a'|s') \Big(g(s',a') + \gamma \sum_{s''\in\mathcal{S}} p_{\mathrm{T}}(s'' \mid s',a') v(s'') \Big) \Bigg), \quad \forall s \in \mathcal{S}
\end{aligned}
$$

となる．同様に，ベルマン最適作用素の場合は，

$$
\begin{aligned}
(\mathsf{B}_*^2 v)(s) = \max_{a\in\mathcal{A}} \Bigg\{ & g(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_{\mathrm{T}}(s'|s,a) \\
& \times \max_{a'\in\mathcal{A}} \Bigg\{ g(s',a') + \gamma \sum_{s''\in\mathcal{S}} p_{\mathrm{T}}(s''|s',a') v(s'') \Bigg\} \Bigg\}, \quad \forall s \in \mathcal{S}
\end{aligned}
$$

となり，k 回適用した場合，

$$
\begin{aligned}
(\mathsf{B}_*^k v)(s) = \max_{a_0\in\mathcal{A}} \Bigg\{ & g(s,a_0) + \gamma \sum_{s_1\in\mathcal{S}} p_{\mathrm{T}}(s_1|s,a_0) \\
& \times \max_{a_1\in\mathcal{A}} \Bigg\{ g(s_1,a_1) + \gamma \sum_{s_2\in\mathcal{S}} p_{\mathrm{T}}(s_2|s_1,a_1) \\
& \cdots \\
& \times \max_{a_{k-2}\in\mathcal{A}} \Bigg\{ g(s_{k-2},a_{k-2}) + \gamma \sum_{s_{k-1}\in\mathcal{S}} p_{\mathrm{T}}(s_{k-1}|s_{k-2},a_{k-2}) \\
& \times \max_{a_{k-1}\in\mathcal{A}} \Bigg\{ g(s_{k-1},a_{k-1}) + \gamma \sum_{s_k\in\mathcal{S}} p_{\mathrm{T}}(s_k|s_{k-1},a_{k-1}) v(s_k) \Bigg\} \Bigg\} \cdots \Bigg\} \Bigg\}, \quad \forall s \in \mathcal{S}
\end{aligned}
\tag{35}
$$

となる．よって，関数 $\mathsf{B}_*^k v$ は時間ステップ長が k の有限時間のマルコフ決定過程で，状態 s で終了する価値を $v(s)$ とする場合の $t=0$ の最適価値関数に対応することがわかる．ただし，$k=0$ の場合は $(\mathsf{B}_*^0 v)(s) = v(s)$, $s \in \mathcal{S}$ とする．

最後に，ベルマン作用素と，ベルマン方程式や価値関数との関係性を確認する．ベルマン作用素 B_π や B_* を用いて，式 (23) のベルマン期待方程式を

$$
V^\pi(s) = (\mathsf{B}_\pi V^\pi)(s), \quad \forall s \in \mathcal{S} \tag{36}
$$

と書くことができ，また式 (31) のベルマン最適方程式を

$$
V^*(s) = (\mathsf{B}_* V^*)(s), \quad \forall s \in \mathcal{S} \tag{37}
$$

と書き直すことができる．これは V^π と V^* がそれぞれのベルマン作用素 B_π と B_* の**不動点** (fixed point) であり，$V^\pi = \mathsf{B}_\pi^k V^\pi$, $\forall k \in \mathbb{N}$ などが成立し，B_π を何度 V^π に適用しても V^π のままであることを意味する．不動点とは定義域と値域が同じ $\mathcal{X} \ni x$ であるような関数や作用素 $\mathcal{F} : \mathcal{X} \to \mathcal{X}$ に対して，$\mathcal{F}x = x$ を満たす x のことである．このような x は $\mathcal{F}$ の解と呼ばれることもある．なお，V^π や V^* がベルマン作用素 B_π や B_* の唯一の不動点であるかどうかは自明ではないが，後述の命題 7 で唯一であることを示す．

2.3.2 ベルマン作用素の特徴

動的計画法の解析するうえで役に立つベルマン作用素の特徴を幾つか紹介する.

補題 3. 単調性の補題

任意の関数 $v : \mathcal{S} \to \mathbb{R}$ と $v' : \mathcal{S} \to \mathbb{R}$ が

$$v(s) \leq v'(s), \quad \forall s \in \mathcal{S} \tag{38}$$

を満たすとき,

a. ベルマン最適作用素 B_* について,

$$(\mathsf{B}_*^k v)(s) \leq (\mathsf{B}_*^k v')(s), \quad \forall s \in \mathcal{S},\ \forall k \in \mathbb{N}_0$$

b. 任意のマルコフ方策系列 $\boldsymbol{\pi} \triangleq \{\pi_0, \pi_1, \dots\} \in \boldsymbol{\Pi}_k^{\mathrm{M}}$ のベルマン期待作用素の積 $\mathsf{B}_{\pi_0}\mathsf{B}_{\pi_1}\dots\mathsf{B}_{\pi_{k-1}} \triangleq \mathsf{B}_{\boldsymbol{\pi}}^k$ について,

$$(\mathsf{B}_{\boldsymbol{\pi}}^k v)(s) \leq (\mathsf{B}_{\boldsymbol{\pi}}^k v')(s), \quad \forall s \in \mathcal{S},\ \forall k \in \mathbb{N}_0$$

が成立する.

証明：

a. 帰納法により証明する. $k = 0$ のときは条件式 (38) より明らか. $k = n$ のとき,

$$(\mathsf{B}_*^n v)(s) \leq (\mathsf{B}_*^n v')(s), \quad \forall s \in \mathcal{S}$$

が正しいと仮定すると, $p_{\mathrm{T}} \geq 0$ より,

$$\sum_{s' \in \mathcal{S}} p_{\mathrm{T}}(s'|s,a)(\mathsf{B}_*^n v)(s') \leq \sum_{s' \in \mathcal{S}} p_{\mathrm{T}}(s'|s,a)(\mathsf{B}_*^n v')(s'), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

だから, 任意の $s \in \mathcal{S}$ に対して,

$$\begin{aligned}(\mathsf{B}_*^{n+1} v)(s) &= \max_{a \in \mathcal{A}} \left\{ g(s,a) + \gamma \sum_{s' \in \mathcal{S}} p_{\mathrm{T}}(s'|s,a)(\mathsf{B}_*^n v)(s') \right\} \\ &\leq \max_{a \in \mathcal{A}} \left\{ g(s,a) + \gamma \sum_{s' \in \mathcal{S}} p_{\mathrm{T}}(s'|s,a)(\mathsf{B}_*^n v')(s') \right\} = (\mathsf{B}_*^{n+1} v')(s) \end{aligned} \tag{39}$$

は明らかに成立する. よって, $k = n + 1$ のときも **a.** が成立するので, 帰納法により証明された.

b. a. と同様にして, 証明される. □

この単調性の補題は, 任意の状態の関数 v の大小関係はベルマン作用素を適用しても変わらない, つまり保存されるということを意味する.

ここで, 関数 $v : \mathcal{S} \to \mathbb{R}$ と $v' : \mathcal{S} \to \mathbb{R}$ の和を

$$(v + v')(s) \triangleq v(s) + v'(s), \quad \forall s \in \mathcal{S}$$

と表記し, また同様に, 関数 v に対して定数 $b \in \mathbb{R}$ を足したものも,

$$(v + b)(s) \triangleq v(s) + b, \quad \forall s \in \mathcal{S}$$

と表記することにする．このとき，ベルマン作用素の定義式 (33) または (34) から，任意の関数 $v : \mathcal{S} \to \mathbb{R}$，スカラー $b \in \mathbb{R}$ に対して，

$$(\mathsf{B}(v+b))(s) = (\mathsf{B}v)(s) + \gamma b, \quad \forall s \in \mathcal{S}$$

を得るので，同様にしてベルマン作用素を繰り返し適用すれば，次の補題が成り立つことがわかる．

補題 4.
任意の $v : \mathcal{S} \to \mathbb{R}$，$b \in \mathbb{R}$ に対して，

$$(\mathsf{B}_*^k(v+b))(s) = (\mathsf{B}_*^k v)(s) + \gamma^k b, \quad \forall s \in \mathcal{S},\ \forall k \in \mathbb{N}_0$$
$$(\mathsf{B}_{\boldsymbol{\pi}}^k(v+b))(s) = (\mathsf{B}_{\boldsymbol{\pi}}^k v)(s) + \gamma^k b, \quad \forall \boldsymbol{\pi} \in \boldsymbol{\Pi}_k^{\mathrm{M}},\ \forall s \in \mathcal{S},\ \forall k \in \mathbb{N}_0$$

が成立する．

最後に，ベルマン作用素の縮小性に関する補題を示す．

補題 5. ベルマン作用素の縮小性
任意の有界関数 $v : \mathcal{S} \to \mathbb{R}$, $v' : \mathcal{S} \to \mathbb{R}$ と $k \in \mathbb{N}_0$ に対して，

a. ベルマン最適作用素 B_*^k について，

$$\max_{s \in \mathcal{S}} \left| (\mathsf{B}_*^k v)(s) - (\mathsf{B}_*^k v')(s) \right| \leq \gamma^k \max_{s \in \mathcal{S}} |v(s) - v'(s)| \tag{40}$$

b. 任意の $\boldsymbol{\pi} \in \boldsymbol{\Pi}_k^{\mathrm{M}}$ のベルマン期待作用素 $\mathsf{B}_{\boldsymbol{\pi}}^k$ について，

$$\max_{s \in \mathcal{S}} \left| (\mathsf{B}_{\boldsymbol{\pi}}^k v)(s) - (\mathsf{B}_{\boldsymbol{\pi}}^k v')(s) \right| \leq \gamma^k \max_{s \in \mathcal{S}} |v(s) - v'(s)|$$

が成立する．

証明：

a. まず，

$$\underline{\varepsilon} \triangleq \min_{s \in \mathcal{S}} \{v(s) - v'(s)\}, \quad \overline{\varepsilon} \triangleq \max_{s \in \mathcal{S}} \{v(s) - v'(s)\}$$

とおけば，

$$v'(s) + \underline{\varepsilon} \leq v(s) \leq v'(s) + \overline{\varepsilon}, \quad \forall s \in \mathcal{S}$$

を得る．上式に対して，B_* を k 回繰り返し適用すれば，補題 3 と補題 4 より，

$$\begin{aligned} & (\mathsf{B}_*^k v')(s) + \gamma^k \underline{\varepsilon} \leq (\mathsf{B}_*^k v)(s) \leq (\mathsf{B}_*^k v')(s) + \gamma^k \overline{\varepsilon}, \quad \forall s \in \mathcal{S} \\ \Leftrightarrow \quad & \gamma^k \underline{\varepsilon} \leq (\mathsf{B}_*^k v)(s) - (\mathsf{B}_*^k v')(s) \leq \gamma^k \overline{\varepsilon}, \quad \forall s \in \mathcal{S} \end{aligned} \tag{41}$$

を得る．そして，ε を

$$\varepsilon \triangleq \max_{s \in \mathcal{S}} |v(s) - v'(s)| = \max\{|\overline{\varepsilon}|, |\underline{\varepsilon}|\}$$

とおけば，式 (41) より，

$$-\gamma^k \varepsilon \leq \gamma^k \underline{\varepsilon} \leq (\mathsf{B}_*^k v)(s) - (\mathsf{B}_*^k v')(s) \leq \gamma^k \overline{\varepsilon} \leq \gamma^k \varepsilon \quad \forall s \in \mathcal{S}$$

を得るので，式 (40) は成立する．

b. **a.** と同様にして，証明できる． □

補題 5 の式 (40) のような縮小性の特性を持つ作用素は一般に**収縮写像**（contraction mapping）と呼ばれる．

2.3.3 動的計画法の特徴

ベルマン作用素の縮小性の補題 5 の式 (40) の v' に V^* を代入すれば，ベルマン最適方程式 (37) より $V^* = \mathsf{B}_* V^*$ なので，任意の関数 $v : \mathcal{S} \to \mathbb{R}$ に対して，

$$\max_{s \in \mathcal{S}} \left| (\mathsf{B}_*^k v)(s) - V^*(s) \right| \leq \gamma^k \max_{s \in \mathcal{S}} |v(s) - V^*(s)| \tag{42}$$

が成立する．よって，V^* と $\mathsf{B}_*^k v$ の誤差の最大絶対誤差（最大値ノルム）は，動的計画法の反復回数 k について指数関数的に減衰することがわかる．またベルマン期待作用素についても同様に，

$$\max_{s \in \mathcal{S}} \left| (\mathsf{B}_{\boldsymbol{\pi}}^k v)(s) - V^{\pi}(s) \right| \leq \gamma^k \max_{s \in \mathcal{S}} |v(s) - V^{\pi}(s)| \tag{43}$$

を得る．以上より，以下の命題が成り立つことがわかる．

命題 6. 動的計画法の収束性

a. 任意の有界の状態関数 $v : \mathcal{S} \to \mathbb{R}$ に対して，ベルマン最適作用素 B_* を k 回繰り返し適用した関数 $(\mathsf{B}_*^k v)$ は最適価値関数 V^* に漸近的に等しくなる．

$$V^*(s) = \lim_{k \to \infty} (\mathsf{B}_*^k v)(s), \quad \forall s \in \mathcal{S} \tag{44}$$

b. 任意の有界の状態関数 $v : \mathcal{S} \to \mathbb{R}$ に対して，マルコフ方策系列 $\tilde{\boldsymbol{\pi}} \triangleq \{\pi_0, \dots, \pi_{k-1}\} \in \boldsymbol{\Pi}_k^{\mathrm{M}}$ のベルマン期待作用素 $\mathsf{B}_{\tilde{\boldsymbol{\pi}}}^k$ を適用した関数 $(\mathsf{B}_{\tilde{\boldsymbol{\pi}}}^k v)$ は $\boldsymbol{\pi} \triangleq \{\tilde{\boldsymbol{\pi}}, \pi_k, \pi_{k+1}, \dots\} \in \boldsymbol{\Pi}^{\mathrm{M}}$ の価値関数 $V^{\boldsymbol{\pi}}$ に漸近的に等しくなる．

$$V^{\boldsymbol{\pi}}(s) = \lim_{k \to \infty} (\mathsf{B}_{\tilde{\boldsymbol{\pi}}}^k v)(s), \quad \forall s \in \mathcal{S} \tag{45}$$

命題 6 は，動的計画法（式 (32)）に従い，ベルマン作用素を用いて状態関数 v を繰り返し更新すれば，最適価値関数や価値関数を求められることを意味する．この考え方をそのまま実装した方法として，2.4.1 項で説明する価値反復法がある．

また命題 6 より，ベルマン方程式の解（ベルマン作用素の不動点）の特徴を次に通り確認できる．

命題 7. ベルマン方程式の解の一意性

a. ベルマン最適方程式 (37) の解になる関数 $v : \mathcal{S} \to \mathbb{R}$ は

$$(\mathsf{B}_* v)(s) = v(s), \quad \forall s \in \mathcal{S} \tag{46}$$

を満たすが，それは最適価値関数 V^* ただ一つである．

b. 定常方策 $\pi \in \Pi$ のベルマン期待方程式 (36) の解になる関数 $v : \mathcal{S} \to \mathbb{R}$ は

$$(\mathsf{B}_\pi v)(s) = v(s), \quad \forall s \in \mathcal{S}$$

を満たすが，それは π の価値関数 V^π ただ一つである．

証明：

a. 背理法を用いて示す．V^* とは異なる式 (46) を満たす関数 $v' : \mathcal{S} \to \mathbb{R}$,

$$\exists s \in \mathcal{S}, \quad V^*(s) \neq v'(s) \tag{47}$$

が存在すると仮定する．v' は式 (46) を満たすので，$v' = \mathsf{B}_* v'$ だから，命題 6 より，

$$v'(s) = \lim_{k \to \infty} (\mathsf{B}_*^k v)(s) = V^*(s), \quad \forall s \in \mathcal{S}$$

となり，式 (47) と矛盾する．

b. **a.** と同様にして，証明できる． □

2.3.4 最適方策

期待リターンについての最適方策を定義し，その存在性や必要十分条件を示す．

定義 8. （理想的な）最適方策

任意の初期状態 $s \in \mathcal{S}$ からの期待リターンを最大化する方策 $\boldsymbol{\pi}^*$ を最適方策と呼ぶ．

$$V^*(s) = V^{\boldsymbol{\pi}^*}(s), \quad \forall s \in \mathcal{S} \tag{48}$$

ある特定の状態 s の期待リターンを最大にする方策は $\mathrm{argmax}_{\boldsymbol{\pi}} V^{\boldsymbol{\pi}}(s)$ であり，常に存在するが，定義 8 の最適方策 $\boldsymbol{\pi}^*$ はあらゆる状態の期待リターンを最大にするものだから，その存在性は自明ではない．もし存在する場合，どの方策集合までを考えれば十分といえるのだろう．次の命題がこれらの問の答えになる．

命題 9. 最適方策の存在性と必要十分条件

最適方策になりうる定常な決定的方策 $\pi^* \in \Pi^{\mathrm{d}}$ $(\subset \boldsymbol{\Pi}^{\mathrm{M}})$ が存在し，π^* が最適方策であることの必要十分条件は，π^* のベルマン期待作用素 B_{π^*} によるベルマン期待方程式 (36) の解 V^{π^*} と最適価値関数 V^*（式 (28)）が一致すること，つまり

$$V^*(s) = (\mathsf{B}_{\pi^*} V^*)(s), \quad \forall s \in \mathcal{S} \tag{49}$$

が成立することである[*4]．

証明：命題 7 より，π^* のベルマン期待作用素 B_{π^*} に対応するベルマン期待方程式 $v = (\mathsf{B}_{\pi^*} v)$ を満たす状態関数 v は価値関数 V^{π^*} のみだから，式 (49) が成り立つのであれば，式 (48) を得る．よって，式 (49) は π^* が最適方策であることの十分条件であることが示された．

反対に，π^* が最適方策であれば，最適方策の定義から $V^* = V^{\pi^*}$ であるから，命題 7 b. から式 (49) を得る．よって，式 (49) は π^* が最適方策であることの必要条件でもある．

最後に，式 (49) を満たす $\pi^{\mathrm{d}*} \in \Pi^{\mathrm{d}}$ の $\mathsf{B}_{\pi^{\mathrm{d}*}}$ が存在することを示すことで，最適方策になる定常な決定的方策 $\pi^{\mathrm{d}*}$ が存在することを示す．ここで，$\mathrm{argmax}_{a \in \mathcal{A}}$ の計算において，同一の最大値を持つ複数の行動 a がある場合，ある決まった順序で扱うものとする．定常な決定的方策 $\pi^{\mathrm{d}*} \in \Pi^{\mathrm{d}}$ を

$$\pi^{\mathrm{d}*}(s) := \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ g(s,a) + \gamma \sum_{s' \in \mathcal{S}} p_{\mathrm{T}}(s'|s,a) V^*(s') \right\} \tag{50}$$

のように設計すれば，ベルマン作用素の定義式 (33) と (34) から，$\mathsf{B}_{\pi^{\mathrm{d}*}} = \mathsf{B}_*$ が成り立つ．また，命題 7 a. より，$\mathsf{B}_{\pi^{\mathrm{d}*}} = \mathsf{B}_*$ であれば，$\mathsf{B}_{\pi^{\mathrm{d}*}}$ は式 (49) を満たす．以上より，式 (49) を満たす $\mathsf{B}_{\pi^{\mathrm{d}*}}$ が存在することが示された． □

命題 9 は, 最適方策の定義 8 と最適価値関数 V^* の定義(式 (28))$V^*(s) \triangleq \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}^{\mathrm{H}}} V^{\boldsymbol{\pi}}(s),\ \forall s \in \mathcal{S}$ から,

$$V^{\pi^{\mathrm{d}*}}(s) \geq V^{\boldsymbol{\pi}}(s), \quad \forall \boldsymbol{\pi} \in \boldsymbol{\Pi}^{\mathrm{H}}, \forall s \in \mathcal{S}$$

を満たす定常な決定的方策 $\pi^{\mathrm{d}*} \in \Pi^{\mathrm{d}}$ が存在することを意味する．よって，最適方策 $\pi^{\mathrm{d}*}$ は式 (26) の目的関数 $f_{\mathrm{w}}(\pi; w) \triangleq \sum_s w(s) V^{\pi}(s)$ について，任意の重み関数 $w \in \mathbb{R}^{\mathcal{S}}_{\geq 0}$ に対して，

$$f_{\mathrm{w}}(\pi^{\mathrm{d}*}; w) = \sum_{s \in \mathcal{S}} w(s)\, V^{\pi^{\mathrm{d}*}}(s) = \sum_{s \in \mathcal{S}} w(s) \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}^{\mathrm{H}}} V^{\pi}(s) \geq \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}^{\mathrm{H}}} f_{\mathrm{w}}(\boldsymbol{\pi}; w)$$

を満たすので,

$$f_{\mathrm{w}}(\pi^{\mathrm{d}*}; w) = \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}^{\mathrm{H}}} f_{\mathrm{w}}(\boldsymbol{\pi}; w), \quad \forall w \in \mathbb{R}^{\mathcal{S}}_{\geq 0}$$

が成立する．つまり，最適方策の学習とは目的関数 $f_{\mathrm{w}}(\pi; w),\ w \in \mathbb{R}^{\mathcal{S}}_{\geq 0}$ を最大にする方策の探索問題であるといえる．ここで，重み w の条件をゼロ以上でなくて，ゼロより大きいとしているのは，もしも重みが非ゼロの何れの状態からも到達できない状態 $s^\star$ が存在してしまうと，$s^\star$ での行動選択は f_{w} に影響を与えず，そのような f_{w} を最大化するような方策を求めても，$s^\star$ での行動選択を最適化できないからである．

命題 9 はアルゴリズムを設計するうえで重要な結果である．第一に，最適化対象の方策集合として時間ステップ依存の非定常な方策や確率的方策などのサイズが大きい方策集合を考える必要はなく，簡単な方策集合である定常な決定的方策の集合 Π^{d} のみを扱っても問題ないことを保証する．第二に，学習している方策のベルマン期待方程式の解が最適価値関数と一致するかどうかで最適方策かどうかを確認できる．また，最適価値関数を求めれば，式 (50) から最適方策が求まる．実際に，次節の動的計画法に基づくアルゴリズムはこれらの結果に基づき導出される．

ところで，定常な決定的方策 Π^{d} は他の目的関数でも真の最適方策になりうるだろうか．次の系で，定常な決定的方策でも最適方策になる目的関数の十分条件を示す．

系 10. 定常な決定的方策の十分性

確率分布を実数に変換する作用素 $\mathbb{F}$ が，任意の確率分布 X, Y について，

$$\begin{aligned} \text{単調性} \quad & X \le Y \text{ ならば，} \mathbb{F}[X] \le \mathbb{F}[Y] \\ \text{平行移動不変性} \quad & \text{任意の実数 } b \in \mathbb{R} \text{ に対して，} \mathbb{F}[X+b] = \mathbb{F}[X] + b \\ \text{正の同時性} \quad & \text{任意の } \lambda \ge 0 \text{ に対して，} \mathbb{F}[\lambda X] = \lambda \mathbb{F}[X] \end{aligned}$$

であるとし，任意の $s \in \mathcal{S}$ について，

$$\begin{aligned} & w(s) \ge 0 \\ & v^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \big(g(s,a) + \gamma \mathbb{F}[v^\pi(S_{t+1}) \,|\, S_t = s, A_t = a, \mathrm{M}] \big) \end{aligned} \tag{51}$$

を満たす重み関数 w と効用関数 v^π を用いて，目的関数を

$$f(\pi) \triangleq \sum_{s \in \mathcal{S}} w(s)\, v^\pi(s)$$

と定義する．このとき，f の最適方策（定義 8）となる定常な決定的方策が存在する．つまり，以下が成立する．

$$\max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}^{\mathrm{H}}} f(\boldsymbol{\pi}) = \max_{\pi \in \Pi^{\mathrm{d}}} f(\pi)$$

略証：価値関数についてのベルマン作用素 B と同様に，式 (51) の効用関数 v^π について $v^\pi = \mathsf{F}_\pi v^\pi$ や $v^* \triangleq \max_\pi v^\pi = \mathsf{F}_* v^*$ を満たすような作用素 F を考えれば，$\mathbb{F}$ が仮定より単調性，平行移動不変性，正の同時性であるため，F についても補題 3, 4 と同様な結果を即座に得る．この結果を用いれば，補題 5 や命題 6, 7, 9 と同様な結果を得るので，目的関数 f の最適方策になりうる定常な決定的方策 $\pi^* \in \Pi^{\mathrm{d}}$ は存在する． □

作用素 $\mathbb{F}$ が $\mathbb{E}$ のとき，式 (51) の効用関数 v^π は価値関数 V^π に対応し，$\mathbb{F}$ が min の場合，v^π は期待最小リターン，v^* は最大最小リターン（maxmin）に対応する [14]．なお，v^π は一般に反復的リスク指標（iterated risk measure）と呼ばれる [13].

2.4 動的計画法の実装

動的計画法の実装として，価値反復法（2.4.1 項）と方策反復法（2.4.2 項）の 2 つの代表的な方法を紹介する．

2.4.1 価値反復法

「動的計画法の収束性」の命題 6 a. から，ベルマン最適作用素を状態関数に繰り返し適用することで最適価値関数を求めることができ，「最適方策の存在性と必要十分条件」の命題 9 の式 (50) から，最適価値関数から最適方策を計算できることがわかる．これらを直接的に実装した方法として，アルゴリズム 1 の Bellman (1957) [3] による**価値反復法**（value iteration algorithm）がある．

なお，B_* の代わりに B_π を用いて，関数 $v : \mathcal{S} \to \mathbb{R}$ の更新

$$v(s) := \mathsf{B}_\pi v(s), \quad \forall s \in \mathcal{S} \tag{52}$$

を反復すれば，命題 6 b. から，価値関数 V^π が求まる．

アルゴリズム 1. 価値反復法 [3]

[入力] $\mathcal{S}$, $\mathcal{A}$, p_{T}, g, γ, 終了閾値 $\epsilon \in (0, \infty)$
[出力] （推定の）最適方策 $\pi^{\mathrm{d}}_{v'} : \mathcal{S} \to \mathcal{A}$, 最適価値関数 $v' : \mathcal{S} \to \mathbb{R}$

1. 初期化
   価値関数 $v : \mathcal{S} \to \mathbb{R}$ を任意に初期化
2. 価値関数の更新

$$v'(s) := \max_{a \in \mathcal{A}} \left\{ g(s,a) + \gamma \sum_{s' \in \mathcal{S}} p_{\mathrm{T}}(s' \mid s, a) v(s') \right\}, \quad \forall s \in \mathcal{S}$$

3. 収束判定
   もし $\max_{s \in \mathcal{S}} |v(s) - v'(s)| < \epsilon$ ならば，決定的方策

$$\pi^{\mathrm{d}}_{v'}(s) := \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ g(s,a) + \gamma \sum_{s' \in \mathcal{S}} p_{\mathrm{T}}(s' \mid s, a) v'(s') \right\}, \quad \forall s \in \mathcal{S}$$

   を求め終了;
   それ以外は，$v'(s) := v(s), \forall s \in \mathcal{S}$ として，2. から繰り返す

2.4.2 方策反復法

「最適方策の存在性と必要十分条件」の命題 9 から，単純な有限の方策集合である定常な決定的方策集合 Π^{d} が最適方策を含むので，定常な決定的方策を逐次的に更新して最適化する方法としてアルゴリズム 2 の**方策反復法**（policy iteration algorithm）が提案された．なお，前節で示した価値反復法（アルゴリズム 1）は，方策ではなくて，価値関数を逐次的に更新する方法であった．

方策反復法の収束性を確認する．まずその基礎となる方策改善の単調性の命題を示す．

命題 11.（方策改善の単調性）
アルゴリズム 2 の方策反復法の任意の繰り返し回数 $k \in \mathbb{N}_0$ の方策 π^{d}_k と π^{d}_{k+1} について，

$$V^{\pi^{\mathrm{d}}_k}(s) \leq V^{\pi^{\mathrm{d}}_{k+1}}(s), \quad \forall s \in \mathcal{S}$$

が成立する．また，

$$\forall s \in \mathcal{S},\ V^{\pi^{\mathrm{d}}_k}(s) = V^*(s) \quad \Leftrightarrow \quad \forall s \in \mathcal{S},\ V^{\pi^{\mathrm{d}}_k}(s) = V^{\pi^{\mathrm{d}}_{k+1}}(s) \tag{54}$$

$$\exists s \in \mathcal{S},\ V^{\pi^{\mathrm{d}}_k}(s) \neq V^*(s) \quad \Leftrightarrow \quad \exists s \in \mathcal{S},\ V^{\pi^{\mathrm{d}}_k}(s) < V^{\pi^{\mathrm{d}}_{k+1}}(s) \tag{55}$$

が成立する．

証明は [34] を参照されたい．命題 11 の式 (54) は，方策の更新により価値関数 V^{π} が変化しないのであれば，方策は既に最適方策に収束していることを意味している．また式 (55) は，方策が最適方策に収束していないのであれば，方策の更新により何れかの状態で必ず価値関数が改善されることを意味している．つまり，アルゴリズム 2 の方策反復法は繰り返しの度に，現在の方策が最適方

アルゴリズム 2. 方策反復法（policy iteration algorithm）[15]

[**入力**] 状態集合 $\mathcal{S}$, 行動集合 $\mathcal{A}$, 遷移確率関数 p_{T}, 報酬関数 g, 割引率 γ

[**出力**] 最適方策 $\pi^{\mathrm{d}} : \mathcal{S} \to \mathcal{A}$, （最適価値関数 $V : \mathcal{S} \to \mathbb{R}$）

1. 初期化

   決定的方策 $\pi^{\mathrm{d}} : \mathcal{S} \to \mathcal{A}$ を任意に初期化

2. 方策評価

   方策 π^{d} のベルマン方程式（$V^{\pi^{\mathrm{d}}}$ に関する連立一次方程式）

$$V^{\pi^{\mathrm{d}}}(s) = (\mathsf{B}_{\pi^{\mathrm{d}}} V^{\pi^{\mathrm{d}}})(s), \quad \forall s \in \mathcal{S} \tag{53}$$

   を解いて，π^{d} の価値関数 $V^{\pi^{\mathrm{d}}}$ を求める

3. 方策改善

   改善方策 $\pi^{\mathrm{d}\prime} : \mathcal{S} \to \mathcal{A}$ を求める

$$\pi^{\mathrm{d}\prime}(s) := \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ g(s,a) + \gamma \sum_{s' \in \mathcal{S}} p_{\mathrm{T}}(s' \mid s, a) V^{\pi^{\mathrm{d}}}(s') \right\}, \ \forall s \in \mathcal{S}$$

4. 収束判定

   もし $\pi^{\mathrm{d}}(s) = \pi^{\mathrm{d}\prime}(s), \forall s \in \mathcal{S}$ ならば，終了；

   それ以外は，$\pi^{\mathrm{d}}(s) := \pi^{\mathrm{d}\prime}(s), \forall s \in \mathcal{S}$ として，2. から繰り返す

策でない限り，必ず方策を改善する．また，$\mathcal{S}$ と $\mathcal{A}$ は有限集合なので決定的方策の集合 $\Pi^{\mathrm{d}} \ni \pi^{\mathrm{d}}$ の要素数は有限であるため，方策反復法は常に有限の繰り返し回数で最適方策に収束することがわかる．

3　強化学習法

2 節では環境（マルコフ決定過程）が既知として方策の最適化（プランニング問題）を考えたが，ここでは，環境は未知であり，環境とエージェントの相互作用などによって得られたデータから方策を学習することを考える．具体的には，2 節の動的計画法を確率的近似の考え方に従い標本近似して，TD 学習（3.2 項）や Q 学習（3.3 項）などの代表的な強化学習アルゴリズムを導出する．

なお，本節で紹介する強化学習法はモデルフリー（環境非同定）型の強化学習とも呼ばれ，環境を陽に推定せずに学習するアプローチである．本稿では紹介しないが，環境を明示的に推定するモデルベース（環境同定）型の強化学習法もある [32, 29, 34].

3.1　データ

1.4.1 項で示した通り，データに基づく意思決定の問題設定を大きく分類すれば，データが既に手元にあって，データに含まれる全ての標本から方策などを学習するバッチ学習（batch learning）と，データを収集しながら逐次的に学習するオンライン学習（online learning）がある．また，データとはエージェントと環境が相互作用した履歴を記録したものであり，単一の意思決定系列

$$h_T \triangleq \{s_0, a_0, r_0, \ldots, s_{T-1}, a_{T-1}, r_{T-1}, s_T\} \tag{56}$$

の場合と，複数系列の場合がある．系列の最小構成は，ある状態 s で行動 a を実行し，その結果として観測される報酬 r と次状態 s' の 4 つ組 $\{s, a, r, s'\}$ であり，（強化学習における）標本，もしくは**経験データ**（experience data），もしくは単に**経験**という．例えば，最小構成の系列（標本）が N 個ある場合は，

$$\{h_1^{(1)}, \ldots, h_1^{(N)}\} = \{(s_0^{(1)}, a_0^{(1)}, r_0^{(1)}, s_1^{(1)}), \ldots, (s_0^{(N)}, a_0^{(N)}, r_0^{(N)}, s_1^{(N)})\}$$

となる．また，このような経験データの集合のことを**履歴データ**（historical data）と呼ぶことにする．

3.2　価値関数の推定

本項では，方策 π を固定して，π に従い行動選択する場合の期待リターン（価値関数）$V^\pi(s) \triangleq \mathbb{E}^\pi[C_0 \,|\, S_0 = s]$ （式 (21)）をデータから推定することを考える．なお，固定された方策 π の価値関数 V^π の推定は方策反復法における方策評価に対応し，多くの強化学習法の部分問題になる基礎的な問題設定であり，この部分だけを切り出して研究している論文も数多く存在する．方策の学習については 3.3 項で扱う．

ここでは，通常の強化学習の問題設定に従い，状態遷移確率や報酬関数の環境情報は未知とする．ただし，状態数や行動数については既知とし，価値関数の推定器 $\hat{V}$ は価値関数と同じ自由度をもつ関数 $\hat{V} : \mathcal{S} \to \mathbb{R}$ を用いることを想定する[*5]．なお，このような関数は各状態などの各要素に対して値を設定することから，ルックアップテーブル（lookup table）もしくはテーブル形式（tabular

[*5] データ（標本）から推定する量や推定方式を推定器もしくは推定量（estimator）と呼び，$\hat{p}_{\mathrm{T}}$ のように推定対象にハット記号（ˆ）をつけて表記する．データは一般に確率的に観測されるので，推定量は確率変数であり，推定量の実現値のことを推定値（estimate）と呼ぶ.

form）の関数と呼ばれる．価値関数より低い自由度を持つような関数近似器や深層ニューラルネットワークモデルなどを用いた価値関数の推定については [29, 34, 11] などを参照されたい．

履歴データ $\{s_0, a_0, r_0, \ldots, s_T, a_T, r_T\}$ からの最も素朴な価値関数（期待リターン）の推定方法として，

$$\hat{V}(s) := \frac{\sum_{t=0}^{T'} \mathbb{I}_{\{s=s_t\}}\, c_t}{\sum_{t=0}^{T'} \mathbb{I}_{\{s=s_t\}}}, \quad \forall s \in \left\{ s \in \mathcal{S} \,:\, \textstyle\sum_{t=0}^{T'} \mathbb{I}_{\{s=s_t\}} > 0 \right\} \tag{57}$$

のように**モンテカルロ推定**するアプローチが考えられる．ここで，c_t は時間ステップ t からの実績リターン $c_t \triangleq \sum_{k=t}^{T} \gamma^{k-t} r_k$ であり，$T' \in \mathbb{N}$ はハイパーパラメータで $T' \leq T$ である．T' を用いるのは，大きな偏りを持つ可能性がある終端時間ステップ T に近い時間ステップ $t' \in \{T'+1, \ldots, T\}$ のリターン $c_{t'}$ を除外するためである．例えば，$t' := T-1$ のリターンは $c_{t'} := r_{T-1} + \gamma r_T$ と計算され，2 時間ステップ以上先の報酬は全てゼロと仮定した偏りのある推定になっている．また，特に割引率 γ が 1 に近い場合，リターンを正確に計算するには，T' を十分に小さくする必要性があり，モンテカルロ推定に利用できる標本数が少なくなり，一般に推定の効率は良くない [28]．そのため，ベルマン作用素に基づく動的計画法の特徴を利用して価値関数を推定する以降で示すアプローチがよく用いられる．

環境が未知でベルマン作用素を直接的に計算できないので，まず 3.2.1 項でその標本近似を示し，価値関数の推定に必要な道具を準備する．そして，3.2.2 項でバッチ学習の場合，3.2.3 項ではオンライン学習の場合の価値関数推定を考える．オンラインの価値関数推定法として **TD 学習**（temporal difference learning; 時間的差分学習）と呼ばれる最も代表的な強化学習法を紹介する．

3.2.1 ベルマン作用素の標本近似

命題 6（動的計画法の収束性）で，状態関数にベルマン作用素 B_π や B_* を繰り返し適用することで，関数 v が価値関数 V^π や最適価値関数 V^* に収束することをみた．しかし，環境が未知なので，ベルマン作用素を計算できない．そこで，$\mathrm{M}(\pi)$ に従い収集した履歴データ

$$h_t^\pi \triangleq \{s_0, a_0, r_0, \ldots, s_{t-1}, a_{t-1}, r_{t-1}, s_t \,|\, \mathrm{M}(\pi)\} \in \mathcal{H}_t$$

から，ベルマン期待作用素 B_π を標本近似することを考える．ここで，$\mathcal{H}_t$ は h_t^π の集合（標本空間）であり，h_t^π は確率変数 $H_t^\pi \triangleq \{S_0, A_0, R_0, \ldots, S_{t-1}, A_{t-1}, R_{t-1}, S_t \,|\, \mathrm{M}(\pi)\}$ の実現値に対応する．つまり，$h_t \triangleq \{s_0, a_0, r_0, \ldots, s_{t-1}, a_{t-1}, r_{t-1}, s_t\}$，$H_t \triangleq \{S_0, A_0, R_0, \ldots, S_{t-1}, A_{t-1}, R_{t-1}, S_t\}$ とおけば，履歴データ h_t^π は

$$h_t^\pi \sim \Pr(H_t = h_t \,|\, \mathrm{M}(\pi))$$

に従い観測される標本である．簡単化のため，上式を単に $h_t^\pi \sim \mathrm{M}(\pi)$ と書いたり，どのような方策 π でデータを収集したかを考慮する必要がない場合，h_t^π を h_t と書いたり，$h_t \sim \mathrm{M}$ と表記することもある．

式 (33) の方策 π のベルマン期待作用素 B_π は関数 $v : \mathcal{S} \to \mathbb{R}$ に対して，

$$\mathsf{B}_\pi v(s) = \mathbb{E}^\pi[R_t + \gamma v(S_{t+1}) \,|\, S_t = s], \quad \forall s \in \mathcal{S} \tag{58}$$

と書けるので，B_π の直接的な近似アプローチとして，時間ステップ $T \in \mathbb{N}$ までの履歴データ h_T^π

を用いて，任意の $s \in \mathcal{S}$ について，

$$\hat{\mathsf{B}}(v; h_T^\pi)(s) \triangleq \begin{cases} \dfrac{1}{\sum_{t=0}^{T-1} \mathbb{I}_{\{s_t=s\}}} \displaystyle\sum_{t=0}^{T-1} \mathbb{I}_{\{s_t=s\}} \left(r_t + \gamma v(s_{t+1})\right) & (\sum_{t=0}^{T-1} \mathbb{I}_{\{s_t=s\}} > 0) \\ v(s) & (\text{それ以外}) \end{cases} \tag{59}$$

のように標本近似することが考えられる．以降，$\hat{\mathsf{B}}$ を**近似ベルマン期待作用素**（approximated Bellman expectation operator）もしくは**近似ベルマン作用素**（approximated Bellman operator）と呼ぶことにする．以下，$\hat{\mathsf{B}}$ の性質を示す．

はじめに，$\hat{\mathsf{B}}$ は環境を最尤推定し，推定した環境モデルから計算されるベルマン作用素に対応することを確認する．行動 a について周辺化した報酬関数 $\bar{g}(s) \triangleq \sum_a \pi(a|s) g(s,a)$ を最小二乗法で推定もしくは最尤推定すれば，任意の $s \in \mathcal{S}$ について，

$$\hat{g}(s; h_T^\pi) = \begin{cases} \dfrac{1}{\sum_{t=0}^{T-1} \mathbb{I}_{\{s_t=s\}}} \displaystyle\sum_{t=0}^{T-1} \mathbb{I}_{\{s_t=s\}} r_t & (\sum_{t=0}^{T-1} \mathbb{I}_{\{s_t=s\}} > 0) \\ 0 & (\text{それ以外}) \end{cases} \tag{60}$$

が求まり，また（周辺化）状態遷移確率 $\bar{p}_{\mathrm{T}}(s'|s) \triangleq \sum_a p_{\mathrm{T}}(s'|s,a)$ を多項分布を用いて最尤推定すれば，次の遷移確率を得る[*6]．

$$\hat{p}_{\mathrm{T}}(s'|s; h_T^\pi) = \begin{cases} \dfrac{1}{\sum_{t=0}^{T-1} \mathbb{I}_{\{s_t=s\}}} \displaystyle\sum_{t=0}^{T-1} \mathbb{I}_{\{s_t=s\}} \mathbb{I}_{\{s_{t+1}=s'\}} & (\sum_{t=0}^{T-1} \mathbb{I}_{\{s_t=s\}} > 0) \\ \mathbb{I}_{\{s'=s\}} & (\text{それ以外}) \end{cases} \tag{61}$$

以上より，$\hat{g}$ と $\hat{p}_{\mathrm{T}}$ を用いて式 (59) の近似ベルマン作用素 $\hat{\mathsf{B}}$ を

$$\hat{\mathsf{B}}(v; h_T^\pi)(s) = \hat{g}(s; h_T^\pi) + \gamma \sum_{s' \in \mathcal{S}} \hat{p}_{\mathrm{T}}(s'|s; h_T^\pi) v(s'), \quad \forall s \in \mathcal{S}$$

と書くことができ，$\hat{\mathsf{B}}$ は最尤推定した環境のベルマン作用素と同一であることがわかる．また，上式から，$\hat{\mathsf{B}}$ は報酬関数と状態遷移確率がそれぞれ $\hat{g}$ と $\hat{p}_{\mathrm{T}}$ であるマルコフ報酬過程に対するベルマン期待作用素とみなせるので，縮小性（補題 5）など 2 節で示したベルマン期待作用素の性質を $\hat{\mathsf{B}}$ はもつ．

次に，近似ベルマン作用素 $\hat{\mathsf{B}}$ の不偏性を示す．履歴データの確率変数 H_T から，時間ステップ $T-1$ までに状態 s に訪問した回数を求めた確率変数を

$$K \triangleq \sum_{t=0}^{T-1} \mathbb{I}_{\{S_t=s\}}$$

[*6] 履歴データ h_t^π で一度も観測されていない状態 s について，報酬関数 $\hat{g}(s)$ や状態遷移確率 $\hat{p}_{\mathrm{T}}(\cdot|s)$ の決め方は任意で，ここでは $\hat{g}(s) = 0$，$\hat{p}_{\mathrm{T}}(s'|s) = \mathbb{I}_{\{s'=s\}}$ としている．また，式 (59) でも観測のない状態 s の扱いは同様に任意であり，このように観測の無い状態については学習することはできない．ただし，状態が連続空間にあるなど状態空間に何かしら構造があるのであれば，観測の無い状態も観測のある状態との類似性から何かしら学習することは可能である [28, 32].

とおく．さらに，状態 s に訪問した時間ステップを $T_1 < T_2 < \cdots < T_K$ と書けば，

$$\begin{aligned}\mathbb{E}^\pi[\hat{\mathsf{B}}(v;H_T)(s) \,\big|\, K>0] &= \mathbb{E}^\pi\left[\frac{\sum_{t=0}^{T-1}\mathbb{I}_{\{S_t=s\}}(R_t+\gamma v(S_{t+1}))}{K}\,\middle|\, K>0\right]\\ &= \mathbb{E}^\pi\left[\frac{\sum_{k=1}^{K}(R_{T_k}+\gamma v(S_{T_k+1}))}{K}\,\middle|\, K>0\right]\end{aligned}$$

と書け，マルコフ性より $R_{T_k}+\gamma v(S_{T_k+1})$ の期待値は k によらず $\mathbb{E}^\pi[R_t+\gamma v(S_{t+1}) \,|\, S_t=s]$ に等しいので，

$$\begin{aligned}\mathbb{E}^\pi[\hat{\mathsf{B}}(v;H_T)(s) \,\big|\, K>0] &= \mathbb{E}^\pi\left[\frac{K}{K}\,\middle|\, K>0\right]\mathbb{E}^\pi[R_t+\gamma v(S_{t+1}) \,|\, S_t=s]\\ &= \mathsf{B}_\pi v(s), \qquad \forall (T,s)\in\mathbb{N}_0\times\mathcal{S}\end{aligned} \tag{62}$$

を得る．以上より，$\hat{\mathsf{B}}$ の条件付き期待値は B_π に等しいこと（$\hat{\mathsf{B}}$ の不偏性）を確認できた．

最後に，マルコフ決定過程 $\mathrm{M}(\pi)$ がエルゴード性を満たすなら，近似ベルマン作用素 $\hat{\mathsf{B}}$ が真のベルマン期待作用素 B_π に収束することを示す．エルゴード性より各状態への滞在確率の極限は初期状態に依存せず，定常分布 p_∞^π に一致し，非ゼロである．

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{I}_{\{s_t=s\}} = p_\infty^\pi(s) > 0, \quad \forall s_0, s\in\mathcal{S}$$

よって，任意の状態関数 $v\in\mathbb{R}^{\mathcal{S}}$ と状態 $s\in\mathcal{S}$ に対しての近似ベルマン作用素 $\hat{\mathsf{B}}(\cdot;h_T^\pi)$ は極限 $T\to\infty$ で，初期状態 $s_0\in\mathcal{S}$ に依存せず，

$$\begin{aligned}\lim_{T\to\infty}\hat{\mathsf{B}}(v;h_T^\pi)(s) &= \lim_{T\to\infty}\frac{\dfrac{1}{T}\displaystyle\sum_{t=0}^{T-1}\mathbb{I}_{\{s_t=s\}}(r_t+\gamma v(s_{t+1}))}{\dfrac{1}{T}\displaystyle\sum_{t=0}^{T-1}\mathbb{I}_{\{s_t=s\}}}\\ &= \frac{p_\infty^\pi(s)\,\mathbb{E}^\pi[R_t+\gamma v(S_{t+1}) \,|\, S_t=s]}{p_\infty^\pi(s)}\\ &= \mathsf{B}_\pi v(s), \qquad \forall s\in\mathcal{S}\end{aligned} \tag{63}$$

となり，ベルマン期待作用素 B_π に収束することを確認できた．

3.2.2 バッチ学習の場合

ある方策 π に従い行動し収集した履歴データ h_T^π が既にあり，そのデータから π の価値関数 V^π を推定するバッチ学習（オフライン学習）を説明する．動的計画法による $\hat{V}$ の更新式 (52) の近似として，真のベルマン期待作用素 B_π の代わりに，単純に式 (59) の近似作用素 $\hat{\mathsf{B}}$ を用いて，

$$\hat{V}(s) := \hat{\mathsf{B}}(\hat{V};h_T^\pi)(s), \quad \forall s\in\mathcal{S} \tag{64}$$

のように $\hat{V}$ を更新するアプローチが考えられる．式 (64) を繰り返し実施すれば，命題 l6 b と命題 7 b より，$\hat{V}$ は次を満たす唯一の不動点

$$\hat{V}_\infty(s) = \hat{\mathsf{B}}(\hat{V}_\infty;h_T^\pi)(s), \quad \forall s\in\mathcal{S} \tag{65}$$

に単調に収束することがわかる．

また，履歴データ h_T^π の系列長 T の極限 $T\to\infty$ では，式 (63) より $\hat{\mathsf{B}}$ は真の B_π に収束するので，推定価値関数 $\hat{V}_\infty$ は真の価値関数 V^π に収束する．

3.2.3 オンライン学習の場合

データが逐次的に追加され，それに従い推定価値関数 $\hat{V}$ を逐次的に更新するオンライン学習問題を扱い，**TD 法**もしくは **TD(0) 法**と呼ばれる TD 学習の原始的な方法を紹介する．

バッチ学習の場合の更新式 (64) をそのままオンライン学習に適用して，各時間ステップ t で $\{a_{t-1}, r_{t-1}, s_t\}$ を経験をする度に履歴データを $h_t^\pi := \{h_{t-1}^\pi, a_{t-1}, r_{t-1}, s_t\}$ と更新して，推定価値関数 $\hat{V}$ を

$$\hat{V}(s) := \hat{\mathsf{B}}(\hat{V}; h_t^\pi)(s), \quad \forall s \in \mathcal{S} \tag{66}$$

のように更新するアプローチが考えられる．しかし，このような更新則を実現するには履歴データを全て記憶しておく必要があり，また全ての状態 $s \in \mathcal{S}$ それぞれに対して $\hat{\mathsf{B}}$ を計算する必要があり，計算量も大きく，効率的ではない．そこで，更新式 (66) を簡単化して，現時間ステップ $t+1$ の観測 $\{s_t, r_t, s_{t+1}\}$ のみを用いて，$\hat{V}(s_t)$ を微小に更新することを考えれば，更新則は

$$\hat{V}(s_t) := (1-\alpha_t)\hat{V}(s_t) + \alpha_t \hat{\mathsf{B}}(\hat{V}; \{s_t, r_t, s_{t+1}\})(s_t) \tag{67}$$

となり，履歴データを記憶しておく必要がなくなる．ここで，$\alpha_t \geq 0$ は**学習率**やステップサイズと呼ばれるハイパーパラメータで，$\alpha_t = 1/(t+1)$ や十分小さい定数などを用いる．ただし，3.4 項で示すように，収束性を保証するには**ロビンス・モンローの条件**（**Robbins-Monro condition**）

$$\alpha_t \geq 0 \ \ (\forall t \in \mathbb{N}_0), \quad \sum_{t=0}^{\infty} \alpha_t = \infty, \quad \sum_{t=0}^{\infty} \alpha_t^2 < \infty \tag{68}$$

を満たす必要がある[*7]．

更新式 (67) の収束性を簡単に示す．詳細な数理や証明は 3.4 項で示す．今，時間ステップ t で状態 s_t にいて，これから行動を選択 $A_t \sim \pi(\cdot|s_t)$ し，報酬を観測 $R_t := g(s_t, A_t)$ して，そして次状態を観測 $S_{t+1} \sim p_{\mathrm{T}}(\cdot|s_t, A_t)$ するという状況にいるとしよう．このとき，式 (67) の右辺第二項 $\hat{\mathsf{B}}(\hat{V}; \{s_t, R_t, S_{t+1}\})(s_t)$ は確率変数であり，その期待値は

$$\mathbb{E}^\pi[\hat{\mathsf{B}}(\hat{V}; \{S_t, R_t, S_{t+1}\})(s_t) \,|\, S_t = s_t] = \mathsf{B}_\pi \hat{V}(s_t)$$

となり，真のベルマン期待作用素 B_π による演算と一致する（式 (62) 参照）．ここで，真のベルマン期待作用素との誤差を

$$X_t \triangleq \hat{\mathsf{B}}(\hat{V}; \{s_t, R_t, S_{t+1}\})(s_t) - \mathsf{B}_\pi \hat{V}(s_t)$$

とおけば，その期待値 $\mathbb{E}^\pi[X_t \,|\, S_t = s_t]$ はゼロであり，また報酬が有界なので，明らかに $\mathbb{E}^\pi[X_t^2 \,|\, S_t = s_t] < \infty$ である．誤差 X_t を用いて更新式 (67) を書き直せば，

$$\hat{V}(s_t) := (1-\alpha_t)\hat{V}(s_t) + \alpha_t\big(\mathsf{B}_\pi \hat{V}(s_t) + X_t\big) \tag{69}$$

となり，真の B_π を用いた更新則 $\hat{V}(s_t) := (1-\alpha_t)\hat{V}(s_t) + \alpha_t \mathsf{B}_\pi \hat{V}(s_t)$ にノイズ X_t が乗っているものと解釈できる．この形式は**確率的近似** (stochastic approximation)，特に**ロビンス・モンロー**

[*7] ロビンス・モンローの条件を満たすには $c > 0$ や $b > 0$ を用いて $\alpha_t = c/(t+b)$ など $\lim_{t\to\infty} \alpha_t = 0$ となるようにする必要があるが，c や b の設定が適切でなく学習初期にもかかわらず α_t が非常に小さくなり，現実的な繰り返し回数で学習が終わらないことがよくある．そのため，収束性を保証できなくなるが，$\alpha_t = c$ のように単に定数を用いることも多い．

のアルゴリズム（Robbins-Monro algorithm）として知られている [6]．つまり，更新式 (67) は B_π による動的計画法（式 (52)）の確率的近似に対応する．確率的近似の数理解析の結果から，更新式 (67) の学習率 α_t がロビンス・モンローの条件（式 (68)）を満たしていれば，極限 $t \to \infty$ で $\hat{V}$ は

$$\hat{V}(s) = \mathsf{B}_\pi \hat{V}(s), \quad \forall s \in \mathcal{S}$$

を満たす不動点に収束することを示すことができる．さらに，ベルマン方程式の一意性（命題 7）より，上式を満たす $\hat{V}$ は唯一 V^π だから，$\hat{V}$ が真の価値関数 V^π に収束することがわかる．

次に更新式 (67) を解釈するため，式 (67) を少し書き換える．

$$\hat{V}(s_t) := \hat{V}(s_t) + \alpha_t \delta_t \tag{70}$$

ここで，δ_t は（α_t を除いた）$\hat{V}$ の更新量であり，

$$\delta_t \triangleq r_t + \gamma \hat{V}(s_{t+1}) - \hat{V}(s_t) \tag{71}$$

である．図 5 のように，δ_t は次時間ステップまでの情報を利用した s_t の価値の推定値 $r_t + \gamma \hat{V}(s_{t+1})$ と，時間ステップ t 時点での推定価値 $\hat{V}(s_t)$ の差分値と解釈できる．このように δ_t は $t+1$ と t の異なる時間ステップでの s_t の予測価値の差異と解釈できることから，**時間的差分誤差**（temporal difference error）もしくは **TD 誤差**（TD error）や TD と呼ばれる．また，式 (70) による価値関数の学習方法は **TD 法**（TD method）と呼ばれ，このような TD 誤差を利用する学習法を総称して **TD 学習**（TD learning）という．TD 法の実装例をアルゴリズム 3 に示す．

図 5　TD 誤差 δ_t（式 (71)）の解釈

3.3　方策と行動価値関数の学習

前項では履歴データからベルマン期待作用素 B_π を近似して，価値関数を推定することを考えたが，ここでは主にベルマン最適作用素 B_* に基づく価値反復法（アルゴリズム 1）を近似的に実行して，最適方策 π^* を学習することを考える．ただし，前項のように単純に B_* を標本近似できないので，まず 3.3.1 項で，ベルマン作用素に行動空間を追加して，ベルマン**行動**作用素と**行動**価値関数を定義する．3.3.2 項でベルマン行動作用素の標本近似方法を示し，3.3.3 項ではバッチ学習，3.3.4 項ではオンライン学習法として **Q 学習法**と **SARSA 法**を導出する．

3.3.1　ベルマン行動作用素と最適行動価値関数

価値反復法（アルゴリズム 1）はベルマン最適作用素 B_*（式 (34)）を用いて状態関数 $v : \mathcal{S} \to \mathbb{R}$ を繰り返し更新して最適価値関数を求める方法だった．第 n 繰り返し目の推定価値関数を $\hat{V}_n$ と書けば，アルゴリズム 1 の価値関数の更新則は

$$\begin{aligned} \hat{V}_{n+1}(s) &= \mathsf{B}_* \hat{V}_n(s) \\ &= \max_{a \in \mathcal{A}} \mathbb{E}\Big\{ g(S_t, A_t) + \gamma \hat{V}_n(S_{t+1}) \,|\, S_t = s, A_t = a \Big\}, \ \forall s \in \mathcal{S} \end{aligned} \tag{72}$$

アルゴリズム 3. TD 法（TD method）

[**入力**] 環境（状態集合 $\mathcal{S}$ と行動集合 $\mathcal{A}$ のみ既知．行動が入力されると，報酬と次状態を出力するブラックボックスなモデル)，方策 π，割引率 γ，学習率 α_t，終了条件（最大時間ステップ数など）

[**出力**] 方策 π の推定価値関数 $\hat{V} : \mathcal{S} \to \mathbb{R}$

1. 初期化
   · 推定価値関数 $\hat{V} : \mathcal{S} \to \mathbb{R}$ を任意に初期化
   · 時間ステップ t を初期化：$t := 0$
   · 初期状態 s_0 を環境から観測
2. 環境との相互作用
   · 方策 $\pi(a|s_t)$ に従い行動 a_t を選択，a_t を環境に入力
   · 環境から報酬 r_t と次状態 s_{t+1} を観測
3. 学習
   · TD 誤差を計算
   $$\delta := r_t + \gamma \hat{V}(s_{t+1}) - \hat{V}(s_t)$$
   · 推定価値関数 $\hat{V}$ を更新
   $$\hat{V}(s_t) := \hat{V}(s_t) + \alpha_t \delta$$
4. 終了判定
   もし終了条件を満たしているならば，終了；
   それ以外は，$t := t + 1$ として，2. から繰り返す

となる．このとき，式 (58) の B_π と異なり，B_* は期待値演算子 $\mathbb{E}$ の外側に max 演算子をもつ．そのため，式 (59) で B_π を標本近似したように，B_* を単純に標本近似することはできない．そこで，$\hat{V}_n$ に補助変数として行動 $a \in \mathcal{A}$ を導入した $\hat{Q}_n : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ を用いて，更新式 (72) を書き直す．関数 $\hat{Q}_n$ を推定行動価値関数として，

$$\hat{Q}_n(s,a) \triangleq \mathbb{E}\Big\{ g(S_t, A_t) + \gamma \hat{V}_n(S_{t+1}) \,|\, S_t = s, A_t = a \Big\}, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

のように定義すれば，式 (72) の更新則を

$$\hat{V}_{n+1}(s) = \max_{a \in \mathcal{A}} \hat{Q}_n(s,a), \quad \forall s \in \mathcal{S} \tag{73}$$

と書ける．よって，

$$\hat{Q}_{n+1}(s,a) := \mathbb{E}\Big\{ g(S_t, A_t) + \gamma \max_{a' \in \mathcal{A}} \hat{Q}_n(S_{t+1}, a') \,|\, S_t = s, A_t = a \Big\}, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \tag{74}$$

のように，状態の関数 $\hat{V}(s)$ についての更新式 (72) を状態行動対の関数 $\hat{Q}(s,a)$ についての更新式に拡張できる．ここで，関数 $q \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ についての作用素 $\Upsilon_* : \mathbb{R}^{\mathcal{S} \times \mathcal{A}} \to \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ を

$$\Upsilon_* q(s,a) \triangleq \mathbb{E}\Big\{ g(s,a) + \gamma \max_{a' \in \mathcal{A}} q(S_{t+1}, a') \,|\, S_t = s, A_t = a \Big\}, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \tag{75}$$

と定義して，**行動価値のベルマン最適作用素**（Bellman optimality mapping for action values）もしくは単に**ベルマン行動最適作用素**と呼ぶことにする．なお，Υ_* を用いれば，式 (74) を $\hat{Q}_{n+1} := \Upsilon_* \hat{Q}_n$ と書ける．

式 (73) のように $\hat{Q}_n$ は $\hat{V}_n$ と関係するので，価値反復法（アルゴリズム 1）の収束判定と同等にするには，各繰り返し n で適当な閾値 $\epsilon > 0$ を用いて，

$$\max_{s \in \mathcal{S}} \left\{ \left| \max_{a \in \mathcal{A}} \hat{Q}_n(s,a) - \max_{a \in \mathcal{A}} \hat{Q}_{n-1}(s,a) \right| \right\} < \epsilon$$

と収束判定すればよい．収束していれば，最適価値 $V^*(s)$ の推定値を $\max_{a \in \mathcal{A}} \hat{Q}_n(s,a)$ と求め，最適方策を

$$\hat{\pi}^*(a|s) := \begin{cases} 1 & (a = \operatorname{argmax}_{b \in \mathcal{A}} \hat{Q}_n(s,b)) \\ 0 & (\text{それ以外}) \end{cases} \tag{76}$$

と推定すればよい．

以下，状態行動の最適価値関数として**最適行動価値関数**（optimal action value function）を導入し，ベルマン行動最適作用素 Υ_* の特徴を整理する．最適行動価値関数は，式 (28) の最適価値関数 $V^*(s) \triangleq \max_\pi V^\pi(s)$ と同様にして，

$$Q^*(s,a) \triangleq \max_{\pi \in \Pi} Q^\pi(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \tag{77}$$

と定義される．このとき，行動価値関数の定義（式 (22)）から，

$$Q^\pi(s,a) = \mathbb{E}\Big\{ g(S_t, A_t) + \gamma V^\pi(S_{t+1}) \,|\, S_t = s, A_t = a \Big\} \tag{78}$$

なので，

$$\begin{aligned} Q^*(s,a) &= \max_{\pi \in \Pi} \Big[\mathbb{E}\Big\{ g(S_t, A_t) + \gamma V^\pi(S_{t+1}) \,|\, S_t = s, A_t = a \Big\} \Big] \\ &= \mathbb{E}\Big\{ g(S_t, A_t) + \gamma V^*(S_{t+1}) \,|\, S_t = s, A_t = a \Big\}, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \end{aligned} \tag{79}$$

を得る．よって，ベルマン方程式の解の一意性（命題 7）を用いて，Q^* や Q^π から次のように V^* や V^π を簡単に求めることが可能である．

$$\begin{aligned} &\max_{a \in \mathcal{A}} Q^*(s,a) = (\mathsf{B}_* V^*)(s) = V^*(s), && \forall s \in \mathcal{S} \\ &\sum_{a \in \mathcal{A}} \pi(a|s) Q^\pi(s,a) = (\mathsf{B}_\pi V^\pi)(s) = V^\pi(s), && \forall s \in \mathcal{S},\, \pi \in \Pi \end{aligned} \tag{80}$$

次に，ベルマン行動最適作用素 Υ_* による動的計画法は，ベルマン最適作用素 B_* の場合と同様，収束することを示す．価値反復法の収束性（命題 6 a.）より，更新式 (72) の $\hat{V}_n$ は任意の初期関数 $\hat{V}_0 : \mathcal{S} \to \mathbb{R}$ に対して，

$$\lim_{n \to \infty} \hat{V}_n(s) = V^*(s), \quad \forall s \in \mathcal{S}$$

だから，式 (74) と (73) より，任意の関数 $\hat{Q}_0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ に対して，

$$\begin{aligned} \lim_{n \to \infty} (\Upsilon_*^n \hat{Q}_0)(s,a) &= \lim_{n \to \infty} \hat{Q}_n(s,a) \\ &= \mathbb{E}\Big\{ g(S_t, A_t) + \gamma \lim_{n \to \infty} \hat{V}_n(S_{t+1}) \,|\, S_t = s, A_t = a \Big\} \\ &= \mathbb{E}\Big\{ g(S_t, A_t) + \gamma V^*(S_{t+1}) \,|\, S_t = s, A_t = a \Big\}, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \end{aligned}$$

が成立する．上式と式 (79) より，

$$\lim_{n\to\infty}(\Upsilon_*^n\,\hat{Q}_0)(s,a)=Q^*(s,a),\quad s\in\mathcal{S},a\in\mathcal{A} \tag{81}$$

を得るので，ベルマン行動最適作用素 Υ_* を繰り返し適用することで，初期関数 $\hat{Q}_0$ によらず，最適行動価値関数 Q^* を求められることがわかる．さらに，「ベルマン方程式の解の一意性」の命題 7 の証明と同様にして，Q^* は Υ_* の唯一の不動点であること，つまり Q^* が

$$Q^*(s,a)=\Upsilon_*Q^*(s,a),\quad \forall(s,a)\in\mathcal{S}\times\mathcal{A}$$

を満たす唯一の解であることを示すことができる．上式は**行動価値のベルマン最適方程式** (Bellman optimality equation for action values) もしくは**ベルマン行動最適方程式**と呼ばれる．また，「ベルマン作用素の縮小性」の補題 5 と同様にして，Υ_* が縮小写像

$$\max_{s,a}|\Upsilon_*q(s,a)-\Upsilon_*q'(s,a)|\le\gamma\max_{s,a}|q(s,a)-q'(s,a)|,\quad \forall q,q'\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$$

であることも容易に確認できる．ここで $\gamma\in[0,1)$ はリターンの割引率である．

最後に，状態価値のベルマン**最適**作用素 B_* に対するベルマン**期待**作用素 B_π のように，行動価値のベルマン最適作用素 Υ_* に対応する**行動価値のベルマン期待作用素**もしくは**ベルマン行動期待作用素**と呼ばれる Υ_π を，任意の定常方策 π と関数 $q:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ について，

$$\Upsilon_\pi q(s,a)\triangleq g(s,a)+\gamma\sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}}p_{\mathrm{T}}(s'|s,a)\pi(a'|s')q(s',a'),\ \ \forall(s,a)\in\mathcal{S}\times\mathcal{A}$$

と定義する．B_π の場合と同様にして，Υ_π は縮小写像であり，行動価値のベルマン期待方程式

$$Q^\pi(s,a)=\Upsilon_\pi Q^\pi(s,a),\quad \forall(s,a)\in\mathcal{S}\times\mathcal{A}$$

を満たす唯一の不動点（解）として，行動価値関数 Q^π（式 (22)）を持つことを示すことができる．よって，式 (74) の Υ_* の代わりに Υ_π を用いて，$\hat{Q}_{n+1}:=\Upsilon_\pi\hat{Q}_n$ のように $\hat{Q}$ を繰り返し更新すれば，$\hat{Q}$ はいずれ Q^π に収束する．

以降，Υ_* と Υ_π の区別が特に必要のない場合，簡単化のためそれらを単に**ベルマン行動作用素**と呼び，Υ と表記することにする．

3.3.2 ベルマン行動作用素の標本近似

3.3.1 項で，ベルマン行動作用素 Υ を状態行動関数 $q:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ に繰り返し適用することで Q^* もしくは Q^π が求まり，Q^* から最適方策 π^* や最適価値関数 V^* を計算できることを示した．ただし，Υ の計算には p_{T} など未知の環境情報が必要なため，通常は Υ を計算できない．そこで，これまで同様，履歴データ $h_T\triangleq\{s_0,a_0,r_0,\dots,s_{T-1},a_{T-1},r_{T-1},s_T\}$ から Υ を近似することを考える．

ベルマン期待作用素 B_π の標本近似 $\hat{\mathsf{B}}$（式 (59)）と同様に，ベルマン行動期待作用素 Υ_π を標本近似すれば，関数 $q:S\times A\to\mathbb{R}$ に対して，

$$\hat{\Upsilon}(q;h_T)(s,a)\triangleq\begin{cases}\dfrac{\displaystyle\sum_{t=0}^{T-1}\mathbb{I}_{\{s_t=s\}}\mathbb{I}_{\{a_t=a\}}\big(r_t+\gamma q(s_{t+1},a_{t+1})\big)}{\displaystyle\sum_{t=0}^{T-1}\mathbb{I}_{\{s_t=s\}}\mathbb{I}_{\{a_t=a\}}} & \big(\textstyle\sum_{t=0}^{T-1}\mathbb{I}_{\{s_t=s\}}\mathbb{I}_{\{a_t=a\}}>0\big)\\ q(s,a) & (\text{それ以外})\end{cases} \tag{82}$$

となる．$\hat{\Upsilon}$ を近似ベルマン行動期待作用素と呼ぶことにする．同様に，ベルマン行動最適作用素 Υ_* は

$$\hat{\Upsilon}_*(q;h_T)(s,a) \triangleq \begin{cases} \dfrac{\sum_{t=0}^{T-1} \mathbb{I}_{\{s_t=s\}}\mathbb{I}_{\{a_t=a\}}\Big(r_t + \gamma \max_{a'\in\mathcal{A}} q(s_{t+1},a')\Big)}{\sum_{t=0}^{T-1} \mathbb{I}_{\{s_t=s\}}\mathbb{I}_{\{a_t=a\}}} & (\textstyle\sum_{t=0}^{T-1} \mathbb{I}_{\{s_t=s\}}\mathbb{I}_{\{a_t=a\}} > 0) \\ q(s,a) & (\text{それ以外}) \end{cases} \tag{83}$$

のように標本近似でき，$\hat{\Upsilon}_*$ を近似ベルマン行動最適作用素と呼ぶことにする．

近似ベルマン行動作用素 $\hat{\Upsilon}$, $\hat{\Upsilon}_*$ は，近似ベルマン作用素 $\hat{\mathsf{B}}$ と同様，データから推定された何かしらのマルコフ決定過程 $\hat{\mathrm{M}}$（環境）における真のベルマン行動作用素とみなせるので，縮小性や不動点の唯一性などベルマン行動作用素と同じ特徴をもつ．ただし，有限標本では一般に $\mathrm{M} \neq \hat{\mathrm{M}}$ なので，不動点は異なる．履歴データを収集する方策系列 $\boldsymbol{\pi}$（もしくは定常方策 π）が真のマルコフ決定過程 M に対して，

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \Pr(S_t{=}s, A_t{=}a \,|\, \mathrm{M}(\boldsymbol{\pi})) > 0, \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A} \tag{84}$$

を満たすのであれば，極限 $T\to\infty$ では，全ての状態行動対を無限回観測するので，$\mathrm{M} = \hat{\mathrm{M}}$ となり，式 (63) と同様にして $\hat{\Upsilon}$ と $\hat{\Upsilon}_*$ はそれぞれ Υ_π と Υ_* に収束する．ここで，データを収集する際の行動選択に用いる方策と，式 (76) のように $\hat{Q}$ から最終的に計算される方策が異なる場合があることに注意されたい．前者は**行動方策**（**behavior policy**）もしくは挙動方策（control policy）と呼ばれ，後者は**目的方策**（**target policy**）もしくは推定方策（estimation policy）と呼ばれます．また，ある特定の方策 π の価値関数を求めるような問題を考えている場合の π も目的方策と呼ばれる．

他に注目すべきは，$\hat{\Upsilon}_*$ は式 (84) を満たすような行動方策 $\boldsymbol{\pi}$ であればどのような $\boldsymbol{\pi}$ でも Υ_* に収束して，（漸近的には）行動方策に非依存であることである．そのため，3.3.4 項で紹介する Q 学習法など $\hat{\Upsilon}_*$ に基づく方法は**方策オフ型の学習**（**off-policy learning**）に分類される．一方，$\hat{\Upsilon}$ は行動方策 π のベルマン行動期待作用素 Υ_π に収束して，行動方策に依存するので，SARSA 法（3.3.4 項）など $\hat{\Upsilon}$ に基づく方法は**方策オン型の学習**（**on-policy learning**）に分類される．なお，TD 法など $\hat{\mathsf{B}}$ に基づく方法も方策オン型の学習である．つまり，方策オン型の学習では目的方策と行動方策が常に同じであり，方策オフ型の学習ではそれらは異なりうる．

3.3.3 バッチ学習の場合

近似ベルマン行動作用素を用いて行動価値関数や最適方策を推定することを考える．3.2.2 項の（状態）価値関数のバッチ学習の場合と同様に，価値反復法に従い，関数 $\hat{Q}: \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ を適当に初期化して，近似ベルマン行動最適作用素 $\hat{\Upsilon}_*$ を用いて，

$$\hat{Q}(s,a) := \hat{\Upsilon}_*(\hat{Q};h_T)(s,a), \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A} \tag{85}$$

のように $\hat{Q}$ を繰り返し更新すれば，$\hat{Q}$ は最適行動価値関数 Q^* の推定器になる．このとき最適方策を決定的方策として，

$$\hat{\pi}^{\mathrm{d}*}(s) := \operatorname*{argmax}_{a\in\mathcal{A}} \hat{Q}(s,a)$$

と推定する．また, もし履歴データ h_T がある定常方策 π に従い収集されていて(つまり $h_T \sim \mathrm{M}(\pi)$ の場合)，$\hat{\Upsilon}_*$ の代わりに $\hat{\Upsilon}$ を用いて式 (85) の更新を実施すれば，$\hat{Q}$ で π の行動価値関数 Q^π を推定していることになる．

なお，$\hat{\Upsilon}_*$ や $\hat{\Upsilon}$ は縮小写像だから，推定器 $\hat{Q}$ は次の近似ベルマン行動方程式を満たす唯一の不動点 $\hat{Q}^*_\infty$ もしくは $\hat{Q}_\infty$ に収束する．

$$\hat{Q}^*_\infty(s,a) = \hat{\Upsilon}_*(\hat{Q}^*_\infty; h_T)(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \tag{86}$$

$$\hat{Q}_\infty(s,a) = \hat{\Upsilon}(\hat{Q}_\infty; h_T)(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \tag{87}$$

ここで，式 (82) の $\hat{\Upsilon}$ の定義から，式 (87) は不動点 $\hat{Q}_\infty$ についての連立一次方程式になっていることがわかる．よって，上記のように $\hat{Q}$ に繰り返し $\hat{\Upsilon}$ を適用しなくても，連立方程式を解くことで解析的に $\hat{Q}_\infty$ を求めることが可能である．一方，$\hat{\Upsilon}$ と異なり $\hat{\Upsilon}_*$（式 (83)）は max 演算子を持つため，一般に式 (86) は $\hat{Q}^*_\infty$ についての連立一次方程式にはならず，解析的に $\hat{Q}^*_\infty$ を求めることはできない．これは式 (85) に従い $\hat{Q}$ を更新する度に，$\hat{\Upsilon}_*$ で選択する 1 ステップ先（$t+1$）の貪欲行動 $a' = \mathrm{argmax}_b \hat{Q}(s_{t+1}, b)$ が変わる可能性があるためである．なお，価値反復法（アルゴリズム 1）も同じ理由により，解析的に最適価値関数を求めることができず，逐次的に状態関数 $\hat{V}$ を更新する必要があった．

3.3.4 オンライン学習の場合

3.2.3 項の TD 法（価値関数のオンライン学習）の導出と同様にして，近似ベルマン行動最適作用素 $\hat{\Upsilon}_*$ によるオンライン学習法を導出する．これは **Q 学習法** [31] と呼ばれる代表的な強化学習法である．また，近似ベルマン行動期待作用素 $\hat{\Upsilon}$ を用いる **SARSA 法**も紹介する．

Q 学習法

各時間ステップ t で $\{a_t, r_t, s_{t+1}\}$ を経験する度に，最適行動価値関数の推定器 $\hat{Q}$ を更新することを考える．TD 法の場合（3.2.3 項）と同様に確率的近似に従い，バッチ型学習（式 (85)）をオンライン学習へ拡張すれば，**Q 学習法**（Q learning method）

$$\hat{Q}(s_t, a_t) := (1-\alpha_t)\hat{Q}(s_t, a_t) + \alpha_t \hat{\Upsilon}_*(\hat{Q}; \{s_t, a_t, r_t, s_{t+1}\})(s_t, a_t) \tag{88}$$

を得る．ここで，α_t は学習率である．

収束性を簡単に確認する．行動 a_t は与えられていて，S_{t+1} はこれから観測されるとして確率変数として扱い，ノイズ項を

$$X_t \triangleq \hat{\Upsilon}_*(\hat{Q}; \{s_t, a_t, r_t, S_{t+1}\})(s_t, a_t) - \Upsilon_* \hat{Q}(s_t, a_t)$$

とおけば，上式 (88) を

$$\hat{Q}(s_t, a_t) := (1-\alpha_t)\hat{Q}(s_t, a_t) + \alpha_t \big(\Upsilon_* \hat{Q}(s_t, a_t) + X_t\big)$$

と書き直すことができる．このとき，Υ_* は縮小写像であり，X_t は平均ゼロで有限分散なので，確率的近似の結果から，式 (84) を満たすような行動方策であれば，適当な条件下で $\hat{Q}$ は真の最適行動価値関数 Q^* に収束する．詳細は 3.4 項に示す．

更新式 (88) を少し整理すれば，次のように TD 学習の形式で更新式を書くことができる．

$$\hat{Q}(s_t, a_t) := \hat{Q}(s_t, a_t) + \alpha_t \delta_t^{(\mathrm{q})} \tag{89}$$

アルゴリズム 4. Q 学習法（Q learning method）[31]

[**入力**] 環境（状態集合 $\mathcal{S}$ と行動集合 $\mathcal{A}$ のみ既知．行動が入力されると，報酬と次状態を出力するブラックボックスなモデル），方策モデル $\pi_t(a|s;\hat{Q})$，割引率 γ，学習率 α_t，終了条件（最大時間ステップ数など）

[**出力**] 最適行動価値の推定値 $\hat{Q}:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$

1. 初期化
   - 推定値 $\hat{Q}:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ を任意に初期化
   - 時間ステップ t を初期化：$t=0$
   - 初期状態 s_0 を環境から観測
2. 環境との相互作用
   - $\pi_t(a|s_t;\hat{Q})$ に従い行動 a_t を選択，a_t を環境に入力
   - 環境から報酬 r_t と次状態 s_{t+1} を観測
3. 学習
   - TD 誤差 δ を計算
   $$\delta_t := r_t + \gamma \max_{a'\in\mathcal{A}} \hat{Q}(s_{t+1},a') - \hat{Q}(s_t,a_t)$$
   - $\hat{Q}$ を更新
   $$\hat{Q}(s_t,a_t) := \hat{Q}(s_t,a_t) + \alpha_t\delta_t$$
3. 終了判定
   - もし終了条件を満たしているならば，終了;
   - それ以外は，$t := t+1$ として，2. から繰り返す

ここで，$\delta^{(\mathrm{q})}$ は Q 学習法の TD 誤差と呼ばれるもので，

$$\delta_t^{(\mathrm{q})} \triangleq r_t + \gamma \max_{a'\in\mathcal{A}} \hat{Q}(s_{t+1},a') - \hat{Q}(s_t,a_t) \tag{90}$$

である．以降，TD 法の TD 誤差（式 (71)）との区別が文脈から明らかな場合，単純化のため $\delta^{(\mathrm{q})}$ を δ と書く．アルゴリズム 4 に Q 学習法を示す．

探索と活用のトレードオフの考慮が必要になる場合，各行動の選択確率は非ゼロであり，また推定価値 $\hat{Q}(s,a)$ が最も大きい行動（貪欲行動）を他の行動より高い確率で選択するような $\hat{Q}$ に依存する方策モデル $\pi_t(a|s;\hat{Q})$ を用いることが多い．なぜなら，学習の進行に応じて $\hat{Q}$ の精度が改善され，行動選択のパフォーマンスも良くなることが期待できるからである．

具体的な方策モデルとして，ε **貪欲方策**（**ε-greedy policy**）

$$\pi_e(a|s;\hat{Q},\varepsilon) = \begin{cases} 1-\varepsilon+\dfrac{\varepsilon}{|\mathcal{A}|} & (a=\arg\max_b \hat{Q}(s,b)) \\ \dfrac{\varepsilon}{|\mathcal{A}|} & (\text{それ以外}) \end{cases} \tag{91}$$

や**ソフトマックス方策**（softmax policy）

$$\pi_s(a|s;\hat{Q},\beta) = \frac{\exp(\beta\hat{Q}(s,a))}{\sum_{b\in\mathcal{A}}\exp(\beta\hat{Q}(s,b))}$$

がよく用いられる．これらにはランダム性（貪欲行動以外を選択する程度）を制御するハイパーパラメターとして，ε もしくは β があるが，それらを時間ステップ t に応じてランダム性が小さくなるように調整して，学習初期はデータ探索を優先的に行い，徐々にデータ活用の比重を大きくするようにして，探索と活用のトレーオフ（1.4.1 項）を考慮することがある．また，学習初期の探索を促進するため，$\hat{Q}$ を単にゼロに初期化するのではなく，行動価値の上限値 $R_{\max}/(1-\gamma)$ などの大きい値に初期化することがある．これは**楽観的な初期化**（optimistic initialization）と呼ばれ，経験の少ない状態行動対の行動価値は大きい値のままなので，そのような行動は選択され易くなる．これは探索と活用のトレードオフの対処に有効であると知られるヒューリスティック，**不確かなときは楽観的に**（*optimism in the face of uncertainty*）[16] を実践しているといえる．ただし，楽観的な初期化はよく用いられるが，学習初期にしか効果がなく，環境が変化してしまうような非定常な問題には機能しないことを注意する必要がある [28].

SARSA 法

Q 学習法は方策オフ型の学習法であったが，方策オン型の学習法として，**SARSA 法** [28] が有名である．3.3.3 項のベルマン行動期待作用素の標本近似 $\hat{\Upsilon}$（式 (82)）によるバッチ学習を，TD 法や Q 学習法と同様にして，オンライン学習に拡張した方法であり，推定器 $\hat{Q}$ を次のように更新する．

$$\hat{Q}(s_t,a_t) := (1-\alpha_t)\hat{Q}(s_t,a_t) + \alpha_t \hat{\Upsilon}(\hat{Q};\{s_t,a_t,r_t,s_{t+1},a_{t+1}\})(s_t,a_t)$$

上式を少し書き換えれば，TD 誤差

$$\delta_t^{(\mathrm{sarsa})} \triangleq r_t + \gamma \hat{Q}(s_{t+1},a_{t+1}) - \hat{Q}(s_t,a_t)$$

に基づく更新式

$$\hat{Q}(s_t,a_t) := \hat{Q}(s_t,a_t) + \alpha_t \delta_t^{(\mathrm{sarsa})} \tag{92}$$

を得る．これは Q 学習法によく似ているが，TD 誤差が異なっていて，Q 学習法の TD 誤差 $\delta^{(\mathrm{q})}$（式 (90)）では次状態 s_{t+1} の行動価値に $\max_{a'\in\mathcal{A}} \hat{Q}(s_{t+1},a')$ を用いているのに対し，SARSA 法では次状態で実際に選択された行動 a_{t+1} の行動価値 $\hat{Q}(s_{t+1},a_{t+1})$ を用いている．なお，SARSA という名前は更新式 (92) を規定する 5 つ組 $\{s_t,a_t,r_t,s_{t+1},a_{t+1}\}$ の頭文字に由来する．

最後に SARSA 法で用いる行動方策について議論する．方策評価のために行動価値の推定問題を考え，評価したい特定の定常方策 π を行動方策とすることもあるが，多くの場合，Q 学習法と同様，ε 貪欲方策モデル（式 (91)）など推定価値 $\hat{Q}$ に依存する方策モデル $\pi_t(a|s;\hat{Q})$ を利用する．このとき，$\hat{Q}$ を更新する度に，結果的に方策も更新され，また更新された方策に従いデータを取得して $\hat{Q}$ を更新するため，「方策評価」と「方策改善」を繰り返し実施して，学習しているといえる．よって，SARSA 法は方策評価と方策改善を繰り返し行う方策反復法（2.4.2 項）と同様の構造を持つことがわかる．

3.4 収束性

はじめに，オンライン型の強化学習法の収束性を証明する基礎的な道具として確率的近似の結果を紹介する．なお，i.i.d. の問題設定で最急勾配法に対して逐次的にサンプリングしてパラメータを更新する方法として確率的勾配法があるが，**確率的近似**（stochastic approximation）はその一

般化（確率過程版）に対応し，TD 法や Q 学習法なども含む．ここでは，確率的近似として次の状態関数 $v : \mathcal{S} \to \mathbb{R}$ の更新則を考える．

$$v_{t+1}(s) := (1 - \alpha_t(s))v_t(s) + \alpha_t(s)\big\{\mathsf{B}_t v_t(s) + X_t(s) + Y_t(s)\big\}, \quad \forall s \in \mathcal{S} \tag{93}$$

ここで，$\alpha_t \in \mathbb{R}^{\mathcal{S}}_{\geq 0}$ は学習率，$\mathsf{B}_t : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ は状態関数の作用素，$X_t, Y_t \in \mathbb{R}^{\mathcal{S}}$ はノイズ（確率変数）である．また，時間ステップ t までの全履歴を

$$\xi_t \triangleq \{v_0, x_0, y_0, \ldots, v_{t-1}, x_{t-1}, y_{t-1}, v_t\}$$

と定義する．以上の準備のもとで，式 (93) の確率的近似の収束性に関する補題を次に示す．

補題 12. 確率的近似の収束性 [6]
式 (93) の状態関数 $v : \mathcal{S} \to \mathbb{R}$ の更新則が次を満たすとする．

(1) 学習率 $\alpha_t(s)$ は全ての $s \in \mathcal{S}$ でロビンス・モンローの条件を満たす．

$$\sum_{t=0}^{\infty} \alpha_t(s) = \infty, \quad \sum_{t=0}^{\infty} \alpha_t^2(s) < \infty, \quad \forall s \in \mathcal{S}$$

(2) 作用素 B_t は任意の $t \in \mathbb{N}_0$ で（同一の）不動点 v^* を持つ縮小写像である．つまり，次を満たす $\tau \in [0, 1)$ が存在する．

$$\|\mathsf{B}_t v_t - v^*\|_\infty \leq \tau \|v_t - v^*\|_\infty, \quad \forall t \in \mathbb{N}_0$$

(3) ノイズ X_t の期待値はゼロであり，

$$\mathbb{E}[X_t(s) \,|\, \xi_t] = 0, \quad \forall t \in \mathbb{N}_0,\ s \in \mathcal{S}$$

また，与えられた任意のノルム $\|\cdot\|$ に対して，

$$\mathbb{E}[X_t^2(s) \,|\, \xi_t] \leq c + d\|v_t\|^2, \quad \forall t \in \mathbb{N}_0,\ s \in \mathcal{S}$$

を満たす $c, d \in \mathbb{R}_{\geq 0}$ が存在する．

(4) ノイズ Y_t に対して，次式を満たすような 0 に収束する系列 $\{\beta_t \in \mathbb{R}_{\geq 0}\}$ が存在する．

$$|Y_t(s)| \leq \beta_t(\|v_t\|_\infty + 1), \quad \forall t \in \mathbb{N}_0,\ s \in \mathcal{S}$$

このとき，v_t は v^* に収束する．

補題 12 は多くのオンライン型の強化学習法の収束性の証明に利用できる便利な命題である．以下，補題 12 を用いて Q 学習法や SARSA 法の収束性を確認する．

命題 13. Q 学習法の収束性

マルコフ決定過程 $\mathrm{M}(\pi)$ の各時間ステップ $t+1 \in \mathbb{N}$ で関数 $\hat{Q} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ を Q 学習法（式 (89)）

$$\hat{Q}(s_t, a_t) := \hat{Q}(s_t, a_t) + \alpha_t \Big(r_t + \gamma \max_{a' \in \mathcal{A}} \hat{Q}(s_{t+1}, a') - \hat{Q}(s_t, a_t) \Big)$$

に従い更新するとき，$\gamma \in [0, 1)$ であり，

- $\hat{Q}$ の初期化条件：$\|\hat{Q}\|_\infty \leq \infty$
- 累積学習率の条件：

$$\begin{cases} \displaystyle\sum_{t=0}^{\infty} \alpha_t \mathbb{I}_{\{s=s_t\}} \mathbb{I}_{\{a=a_t\}} = \infty, & \forall (s, a) \in \mathcal{S} \times \mathcal{A} \\ \displaystyle\sum_{t=0}^{\infty} \alpha_t^2 \mathbb{I}_{\{s=s_t\}} \mathbb{I}_{\{a=a_t\}} < \infty, & \forall (s, a) \in \mathcal{S} \times \mathcal{A} \end{cases} \tag{94}$$

を満たすなら，$\hat{Q}$ は最適行動価値関数 Q^* に収束する．

証明：補題 12 の 4 つの条件を満たすことを示すことで証明する．時間ステップ t で更新後の $\hat{Q}$ を q_t と書き，学習率を $\tilde{\alpha}_t(s, a) \triangleq \alpha_t \mathbb{I}_{\{s=s_t\}} \mathbb{I}_{\{a=a_t\}}$ とする．さらに，時間ステップ $t+1$ の状態はこれから観測されるとして確率変数 S_{t+1} として扱い，ノイズ X_t を

$$X_t(s, a) \triangleq \begin{cases} g(s, a) + \gamma \max_{a' \in \mathcal{A}} q_t(S_{t+1}, a') - \Upsilon_* q_t(s, a) & (s = s_t, a = a_t) \\ 0 & (\text{それ以外}) \end{cases} \tag{95}$$

とおけば，時間ステップ $t+1$ の Q 学習法の更新則を

$$q_{t+1}(s, a) := (1 - \tilde{\alpha}_t(s, a)) q_t(s, a) + \tilde{\alpha}_t(s, a)(\Upsilon_* q_t(s, a) + X_t(s, a)), \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A} \tag{96}$$

と書き直せる．上式をノイズ Y_t が常にゼロである式 (93) の確率的近似の更新則に対応させれば，補題 12 の条件 (4) を明らかに満足する．また累積学習率の条件（式 (94)）より補題 12 の条件 (1) を満たしており，Υ_* は縮小写像（3.3.1 項）なので条件 (2) も満たす．よって，あとは式 (96) の X_t が補題 12 の条件 (3) を満たすことを示せばよい．

まず，X_t の条件付き期待値は定義（式 (95)）と次式から常にゼロである．

$$\begin{aligned} &\mathbb{E}[X_t(s_t, a_t) \,|\, H_t = h_t, \mathrm{M}] \\ &\quad = \mathbb{E}\Big[g(s_t, a_t) + \gamma \max_{a' \in \mathcal{A}} q_t(S_{t+1}, a') \,|\, H_t = h_t, \mathrm{M} \Big] - \Upsilon_* q_t(s_t, a_t) \\ &\quad = \mathbb{E}\Big[g(s_t, a_t) + \gamma \max_{a' \in \mathcal{A}} q_t(S_{t+1}, a') \,|\, S_t = s_t, A_t = a_t, \mathrm{M} \Big] - \Upsilon_* q_t(s_t, a_t) \\ &\quad = 0 \end{aligned}$$

ここで，h_t は履歴（式 (9)）であり，最後の等式は Υ_* の定義（式 (75)）から成立する．さらに，g は有界なので，

$$\mathbb{E}[X_t^2(s, a) \,|\, H_t = h_t, \mathrm{M}] \leq c + d\|q_t\|^2, \quad \forall (t, s, a) \in \mathbb{N}_0 \times \mathcal{S} \times \mathcal{A}$$

を満たすような $c, d \in \mathbb{R}_{\geq 0}$ が存在する．以上より，X_t は補題 12 の条件 (3) を満足する． □

命題 13 の累積学習率の条件（式 (94)）であるが，任意の状態間を有限の時間ステップで遷移できるマルコフ決定過程については，各状態で各行動の選択確率が $\epsilon \in \mathbb{R}_{>0}$ 以上であり，学習率が α_t がロビンス・モンローの条件を満たすのであれば，式 (94) は成立する（文献 [26] の Lemma 4）．本稿では省略するが，TD 法も Q 学習法と同様にして収束性を示すことができる．

最後に，SARSA 法の収束性を確認する．Q 学習法はベルマン**最適**作用素に基づく方策**オフ**型の学習であったのに対して，SARSA 法は方策に依存するベルマン**期待**作用素に基づく方策**オン**型の学習ため，時間ステップの進展に従い，方策が更新され，ベルマン期待作用素も変化する．そのため，Q 学習法の収束性の場合よりも証明に工夫が必要になる．また，**GLIE**（Greedy in the Limit with Infinite Exploration）方策という追加の条件が必要になる．GLIE 方策とは，全ての状態行動対を無限回観測でき，かつ極限で貪欲方策になるような方策のことである．

命題 14. SARSA 法の収束性

マルコフ決定過程 M の各時間ステップ $t+1 \in \mathbb{N}$ で行動 a_{t+1} を関数 $\hat{Q} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ に基づく方策 $\pi_t(\cdot|s_{t+1}; \hat{Q})$ に従い選択し，$\hat{Q}$ を SARSA 法（式 (92)）

$$\hat{Q}(s_t, a_t) := \hat{Q}(s_t, a_t) + \alpha_t \Big(r_t + \gamma \hat{Q}(s_{t+1}, a_{t+1}) - \hat{Q}(s_t, a_t) \Big)$$

に従い更新するとき，$\gamma \in [0, 1)$ であり，

- $\hat{Q}$ の初期化条件：$\|\hat{Q}\|_\infty \leq \infty$
- 累積学習率の条件：

$$\begin{cases} \displaystyle\sum_{t=0}^{\infty} \alpha_t \mathbb{I}_{\{s=s_t\}} \mathbb{I}_{\{a=a_t\}} = \infty, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \\ \displaystyle\sum_{t=0}^{\infty} \alpha_t^2 \mathbb{I}_{\{s=s_t\}} \mathbb{I}_{\{a=a_t\}} < \infty, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \end{cases}$$

- 方策 π_t は GLIE 方策

を満たすなら，$\hat{Q}$ は最適行動価値関数 Q^* に収束する．

証明：命題 13 と同様，補題 12 の 4 つの条件を満たすことを示すことで証明する．時間ステップ t で更新後の $\hat{Q}$ を q_t と書き，学習率を $\tilde{\alpha}_t(s,a) \triangleq \alpha_t \mathbb{I}_{\{s=s_t\}} \mathbb{I}_{\{a=a_t\}}$ と再定義する．時間ステップ $t+1$ の状態と行動はこれから決定される確率変数として扱い，ノイズ X_t

$$X_t(s,a) \triangleq \begin{cases} g(s,a) + \gamma \max_{a' \in \mathcal{A}} q_t(S_{t+1}, a') - \Upsilon_* q_t(s,a) \\ \qquad\qquad + \gamma \big(v_t^{\pi_t}(S_{t+1}) - q_t(S_{t+1}, A_{t+1}) \big) & (s = s_t, a = a_t) \\ 0 & (\text{それ以外}) \end{cases}$$

とおく．ここで，$v_t^\pi : \mathcal{S} \to \mathbb{R}$ は

$$v_t^\pi(s) \triangleq \sum_{a \in \mathcal{A}} \pi(a|s) q_t(s,a) \tag{97}$$

である．さらに，

$$Y_t(s,a) \triangleq \begin{cases} \gamma\big(\max_{a' \in \mathcal{A}} q_t(S_{t+1}, a') - v_t^{\pi_t}(S_{t+1})\big) & (s = s_t, a = a_t) \\ 0 & (\text{それ以外}) \end{cases} \tag{98}$$

とおけば，時間ステップ $t+1$ での SARSA 法の更新則を次のように書くことができる．

$$\begin{aligned} q_{t+1}(s,a) :=& (1 - \tilde{\alpha}_t(s,a)) q_t(s,a) \\ &+ \tilde{\alpha}_t(s,a)(\Upsilon_* q_t(s,a) + X_t(s,a) + Y_t(s,a)),\ \forall (s,a) \in \mathcal{S} \times \mathcal{A} \end{aligned}$$

上式を式 (93) の確率的近似の更新則に対応させ，Q 学習法の収束性（命題 13）の証明と同様にして，上式が補題 12 の条件 (1), (2), (3) を満足することを示すことができる．よって，あとは式 (98) の Y_t が補題 12 の条件 (4) を満たすこと，つまり，

$$|Y_t(s,a)| \leq \beta_t(\|q_t\|_\infty + 1), \quad \forall (t,s,a) \in \mathbb{N}_0 \times \mathcal{S} \times \mathcal{A} \tag{99}$$

を満たすゼロに収束する系列 $\{\beta_t \in \mathbb{R}_{\geq 0}\}$ が存在することを示せればよい．

まず，Y_t と $v_t^{\pi_t}$ の定義から，

$$0 \leq Y_t(s,a) \leq \gamma\big(\max_{a' \in \mathcal{A}} q_t(S_{t+1}, a') - \min_{a' \in \mathcal{A}} q_t(S_{t+1}, a')\big), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

だから，

$$|Y_t(s,a)| \leq 2\gamma \|q_t\|_\infty, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

と書け，例えば $\beta_t := 2\gamma$ とすることで，有限の時間ステップ t については式 (99) を満たす β_t が存在することを確認できる．よって，あとは $t \to \infty$ でゼロに収束する式 (99) を満たす β_t が存在することを示せばよい．

方策 π_t は GLIE 方策だから，任意の $\varepsilon > 0$, $s \in \mathcal{S}$ に対して，

$$\max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left| \mathbb{I}_{\{a = \pi_{\text{greedy}}(s; q_t)\}} - \pi_t(a|s; q_t) \right| < \varepsilon, \quad \forall t \in \{t^*, t^*+1, \dots\} \tag{100}$$

を満たすような t^* が存在する．ここで，π_{greedy} は貪欲方策

$$\pi_{\text{greedy}}(s; q) \triangleq \operatorname*{argmax}_{a \in A} q(s,a)$$

である．式 (100) から，任意の $s \in \mathcal{S}$ で

$$\Pr\Big(\operatorname*{argmax}_{a' \in \mathcal{A}} q_t(s,a') = A \,|\, A \sim \pi_t(\cdot|s, q_t)\Big) \geq 1 - \varepsilon, \quad \forall t \in \{t^*, t^*+1, \dots\}$$

が成立するので，v_t^π の定義（式 (97)）より，任意の $t \in \{t^*, t^*+1, \dots\}$ で，

$$v_t^{\pi_t}(s) \geq (1 - \varepsilon) \max_{a' \in \mathcal{A}} q_t(s,a') - \varepsilon \min_{a' \in \mathcal{A}} q_t(s,a'), \quad \forall s \in \mathcal{S}$$

は成立する．以上より，Y_t の定義（式 (98)）より，任意の $\epsilon > 0$ に対して，

$$\begin{aligned} |Y_t(s,a)| &\leq \gamma\varepsilon\big(\max_{a' \in \mathcal{A}} q_t(S_{t+1}, a') - \min_{a' \in \mathcal{A}} q_t(S_{t+1}, a')\big) \\ &\leq 2\gamma\varepsilon \|q_t\|_\infty, \qquad \forall t \in \{t^*, t^*+1, \dots\},\ (s,a) \in \mathcal{S} \times \mathcal{A} \end{aligned}$$

を満たすような t^* は存在する．よって，例えば $\beta_t := 2\gamma\varepsilon$ とすれば，β_t は式 (99) を満たし，$t \to \infty$ でゼロに収束する． □

3.5　アクター・クリティック法

SARSA 法は方策反復法（2.4.2 項) に類似するが，より直接的に方策反復法を実装する強化学習のアプローチとして，**アクター・クリティック法**（actor critic method; AC 法）がある．図 6 に示すように，方策である**アクター**（**actor**）と方策評価を行う**クリティック**（**critic**）の 2 つのモジュールからなり，クリティックが報酬などの観測から方策改善のための信号を計算し，それをアクターに与えて方策を更新することを繰り返す．なお，アクター・クリティック法は特定の方法の実装を指すのではなく，方法の総称であり，本稿では割愛するが実装方法は様々ある [12]．SARSA 法は推定器 $\hat{Q}$ が方策の評価結果であると共に方策を定めているので，アクターとクリティックが一つのモジュールに統合されているアクター・クリティック法の特殊型であると解釈できる．

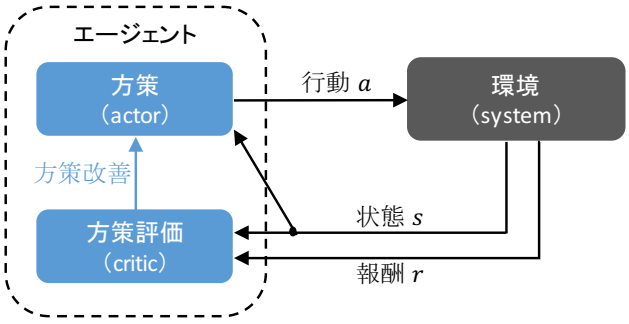

図 6　アクター・クリティック法の構造

参考文献

[1] N. Abe, P. Melville, C. Pendus, C. K. Reddy, D. L. Jensen, V. P. Thomas, J. J. Bennett, G. F. Anderson, B. R. Cooley, M. Kowalczyk, M. Domick, and T. Gardinier. Optimizing debt collections using constrained reinforcement learning. In *International Conference on Knowledge Discovery and Data Mining*, pages 75–84, 2010.

[2] N. Abe, N. K. Verma, C. Apté, and R. Schroko. Cross channel optimized marketing by reinforcement learning. In *International Conference on Knowledge Discovery and Data Mining*, pages 767–772, 2004.

[3] R. E. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[4] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume I. Athena Scientific, 2nd edition, 1995.

[5] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, 4th edition, 2012.

[6] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[7] M. K. Bothe, L. Dickens, K. Reichel, A. Tellmann, B. Ellger, M. Westphal, and A. A. Faisal. The use of reinforcement learning algorithms to meet the challenges of an artificial

pancreas. *Expert Review of Medical Devices*, 10(5):661—-673, 2013.

[8] S. R. K. Branavan, H. Chen, L. S. Zettlemoyer, and R. Barzilay. Reinforcement learning for mapping instructions to actions. In *Annual Meeting of the Association for Computational Linguistics*, 2009.

[9] K. Doya. Reinforcement learning in continuous time and space. *Neural Computation*, 12(1):219–245, 2000.

[10] P. Escandell-Montero, M. Chermisi, J. M. Martínez-Martínez, J. Gómez-Sanchis, C. Barbieri, E. Soria-Olivas, F. Mari, J. Vila-Francés, A. Stopper, E. Gatti, and J. D. Martín-Guerrero. Optimization of anemia treatment in hemodialysis patients via reinforcement learning. *Artificial Intelligence in Medicine*, 62(1):47–60, 2014.

[11] V. Francois-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau. *An Introduction to Deep Reinforcement Learning*. Now Publishers Inc, 2019. (松原崇充/監訳 井尻善久/訳 浜屋政志/訳：深層強化学習入門, 共立出版, 2021).

[12] I. Grondman, L.Busoniu, G. Lopes, and R. Babuska. A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 42(6):1291–1307, 2012.

[13] M. R. Hardy and J. L. Wirch. The iterated cte: A dynamic risk measure. *North American Actuarial Journal*, 8:62–75, 2004.

[14] M. Heger. Consideration of risk in reinforcement learning. In *International Conference on Machine Learning*, pages 105–111, 1994.

[15] R. A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, 1960.

[16] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of AI Research*, 4:237—-285, 1996.

[17] H. Kamigaito, P. Zhang, H. Takamura, and M. Okumura. An empirical study of generating texts for search engine advertising. In *Conference of the North American Chapter of the Association for Computational Linguistics: Industry Papers*, pages 255–262, 2021.

[18] A. Lazaric, M. Restelli, and A. Bonarini. Reinforcement learning in continuous action spaces through sequential Monte Carlo methods. In *Advances in Neural Information Processing Systems*, pages 833–840, 2007.

[19] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostro-vski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[20] Y. Nevmyvaka, Y. Feng, and M. Kearns. Reinforcement learning for optimized trade execution. In *International Conference on Machine Learning*, pages 673–680, 2006.

[21] J. Pineau, A. Guez, R. Vincent, G. Panuccio, and M. Avoli. Treating epilepsy via adaptive neurostimulation: A reinforcement learning approach. *International Journal of Neural Systems*, 19(4):227–240, 2009.

[22] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, 1994.

[23] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence

training for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7008–7024, 2017.

[24] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, and D. Silver. Mastering Atari, go, chess and shogi by planning with a learned model. In *Nature*, volume 588, 2020.

[25] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[26] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38:287—-308, 2000.

[27] T. Spooner, J. Fearnley, R. Savani, and A. Koukorinis. Market making via reinforcement learning. In *International Conference on Autonomous Agents and MultiAgent Systems*, pages 434–442, 2018.

[28] R. S. Sutton and A. G. Barto. *Reinforcement Learning*. MIT Press, 1998. (三上貞芳, 皆川雅章 訳：強化学習, 森北出版, 2000).

[29] R. S. Sutton and A. G. Barto. *Reinforcement Learning*. MIT Press, 2nd edition, 2018.

[30] G. J. Tesauro. Temporal difference learning and TD-gammon. *Communications of the ACM*, 38(3):58–68, 1995.

[31] C. J. C. H. Watkins and P. Dayan. Technical note: Q-learning. *Machine Learning*, 8:279–292, 1992.

[32] M. Wiering and M. V. Otterlo, editors. *Reinforcement Learning: State of the Art*. Springer, 2012.

[33] S. Young, M. Gašić, F. Mairesse S. Keizer, J. Schatzmann, B. Thomsona, and K. Yu. The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech and Language*, 24(2):150–174, 2009.

[34] 森村哲郎. **強化学習**. 講談社, 2019.

[35] 牧野貴樹, 澁谷長史, and 白川真一, editors. **これからの強化学習**. 森北出版, 2016.