制御工学者のための強化学習入門

佐々木　智丈*・加　嶋　健　司**

＊株式会社富士通研究所 人工知能研究所　神奈川県川崎市中原区上小田中4–1–1
＊＊京都大学　京都府京都市左京区吉田本町 36–1
＊Artificial Intelligence Laboratory, Fujitsu Laboratories Ltd., 4–1–1 Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa, Japan
＊＊Kyoto University, 36–1 Yoshida-Honmachi, Sakyo-ku, Kyoto, Japan
＊E-mail: tomotake.sasaki@fujitsu.com

キーワード：強化学習 (reinforcement learning)，最適制御 (optimal control)，適応制御 (adaptive control).
JL 0003/19/5803–0182

1. はじめに：本稿のねらい

近年の人工知能研究の隆盛の中で，一般的には機械学習の一分野に位置づけられることが多い「強化学習」にも高い関心が集まっている．特に，Atari 2600 のさまざまなゲームを攻略した DQN や囲碁の世界トッププレーヤたちに勝利した AlphaGo をきっかけとして強化学習に興味をもった読者も多いのではないだろうか．

強化学習 (reinforcement learning) あるいは強化 (reinforcement) という用語・概念は，元々は生物のもつ適応能力・学習能力の研究の中で導入されたものである．その後，人工知能分野において人工物に同様の能力をもたせるための研究が行われるようになり，この用語が援用されるようになった．そして現在，工学用語としての強化学習は 3 つの意味をもっている．すなわち，1) ある特徴をもった問題，2) そのような問題に対して有効に働く手法，3) そのような問題を扱う研究分野，である[1]．

制御工学の用語に基づくと，「強化学習問題」とは「適応最適制御問題」であり，「強化学習手法」とは「適応最適制御手法」である．制御工学になじみある読者には，この関係の把握が研究や応用を見通し良く進める上で役立つことと思う．本稿は上記関係の理解を助け，今後の研究・応用への足掛かりを提供することを目標としている．

この目標のため，2 章ではまず，最適制御および適応制御との関係を示しつつ，なるべく一般的な形で強化学習問題を説明し，その後，強化学習分野における標準的な問題（割引付き有限マルコフ決定問題の強化学習版）と離散時間線形二次レギュレーション (LQR) 問題の強化学習版がそのサブクラスとして理解できることを示す．3 章では強化学習手法の概論を述べた後，上記 2 つの問題に対し方策反復法ベースの強化学習手法を示す．4 章では強化学習とモデルベースト制御の融合に関して述べ，最後に本編で触れられなかった話題も含め文献案内を行う．

2. 強化学習問題

本章の目的は，「強化学習問題」が「適応最適制御問題」であることを説明することである．そのために，まずなるべく一般的に最適制御問題を説明し，それとの対比で強化学習問題を説明する．本来，**表 1** の項目のすべての組み合わせに対して最適制御問題と強化学習問題が考えられるが，ここでは制御対象を状態（内部状態）をもつ離散時間システムとして扱い，無限時間地平問題の場合を説明する．

表 1　制御対象を分類する代表的な項目

状態あり（動的システム）／状態なし（静的システム）
連続時間／離散時間
（状態，制御入力等が）連続値／（〃）離散値（特に有限個）
確率的／確定的
時変（非定常）／時不変（定常）
部分観測（状態が直接測定できない）／完全観測（〃できる）

制御対象の変化および制御対象からの出力が以下の離散時間状態方程式，出力方程式で表わされるとする．

$$x_{k+1} = f(x_k, u_k, e^{\mathrm{p}}_k, k) \quad (1)$$

$$y_k = h(x_k, u_k, e^{\mathrm{o}}_k, k) \quad (2)$$

$x_k \in \mathcal{X}$, $u_k \in \mathcal{U}$, $y_k \in \mathcal{Y}$ は時刻 $k \in \mathbb{Z}_{\geq 0} = \{0, 1, \dots\}$ における制御対象の状態，制御入力，出力である．$e^{\mathrm{p}}_k \in \mathcal{E}^{\mathrm{p}}$, $e^{\mathrm{o}}_k \in \mathcal{E}^{\mathrm{o}}$ は状態，制御入力，時刻以外に状態変化／出力に影響する要因を表わす．ここでは状態集合 $\mathcal{X}$，入力集合 $\mathcal{U}$ 等は何らかの集合であるとだけしておく．$f : \mathcal{X} \times \mathcal{U} \times \mathcal{E}^{\mathrm{p}} \times \mathbb{Z}_{\geq 0} \to \mathcal{X}$, $h : \mathcal{X} \times \mathcal{U} \times \mathcal{E}^{\mathrm{o}} \times \mathbb{Z}_{\geq 0} \to \mathcal{Y}$ である．

さらに，上記出力 y_k に加え，$\mathcal{Y} \times \mathcal{U} \times \mathcal{E}^{\mathrm{r}} \times \mathbb{Z}_{\geq 0} \to \mathcal{R} \subseteq \mathbb{R}$ or $\mathcal{X} \times \mathcal{U} \times \mathcal{E}^{\mathrm{r}} \times \mathbb{Z}_{\geq 0} \to \mathcal{R} \subseteq \mathbb{R}$ なる関数 ρ で規定される即時的な評価値（即時報酬あるいは即時コスト）

$$r_{k+1} = \rho(y_k, u_k, e^{\mathrm{r}}_k, k) \text{ or } \rho(x_k, u_k, e^{\mathrm{r}}_k, k) \quad (3)$$

が存在し[注1]，それに基づいて評価関数が定義されているとする．ここで，$e^{\mathrm{r}}_k \in \mathcal{E}^{\mathrm{r}}$ は出力 or 状態，制御入力，時刻以外に即時評価値に影響を与える要因を表わす．この関数 ρ について，以下の 2 つの場合が考えられる．

(i) 関数 ρ は制御対象に内在する即時評価値の生成機構を表わしており，r_k が出力同様に観測される[注2]

[注1] 文献 1) に従い (3) 式左辺の添字を $k+1$ とした．本稿での強化学習関連の定義・記述についてはこれ以降も本文献をベースにする．
[注2] たとえば r_k としてビデオゲームの得点や，発電機／電気機器の単位時間当たりの発電量／消費電力量を思い浮かべるとよい．

(ii) 関数 ρ は観測可能な量の関数として設計者によって定義され，r_k は制御用の計算機内で計算される

(ii) では関数 ρ の適切な設計が問題となることがあるが，一度 ρ を設計した後は (i) に統合できるのでそうする．

最適制御問題とは，f, h, ρ が完全にわかっているという状況（あるいは仮定）の下で，評価関数を最大化または最小化する制御則を導出する問題ということができる．

一方，本稿の主題である強化学習問題については，以下のようにいうことができる．

【強化学習問題】とは，

- f, h, ρ は未知または部分的／不完全にしかわからない

ただし，

- u_k を決定して制御対象に印加し，その結果として生じる y_k および r_{k+1} は観測できる

という状況の下で，

- データ $(u_0, y_0, r_1, \dots)$ に基づく自己更新（自己調整）によって，評価関数を最大化または最小化する制御則を自ら獲得する制御器を構成する

問題である．

評価関数の最大化や最小化という目標は最適制御問題から引き継がれているが，その他は制御工学分野における適応制御問題と共通している．これが「「強化学習問題」とは「適応最適制御問題」である」と述べた意味である．以上の問題設定をブロック線図形式で**図 1** に示す．この図からも，適応制御問題との関係が見て取れるだろう．強化学習分野では制御対象は環境 (environment)，制御器はエージェント (agent)，制御入力は行動 (action) と呼ばれるので括弧内に付記した．

つぎに，強化学習分野における標準的な問題を以上で説明した問題のサブクラスとして導出する．まず，完全観測 $(y_k = x_k)$ であり，状態集合，入力集合は有限集合であるとする $(\mathcal{X} = \{\mathrm{x}^1, \mathrm{x}^2, \dots, \mathrm{x}^N\}, \mathcal{U} = \{\mathrm{u}^1, \mathrm{u}^2, \dots, \mathrm{u}^M\})$．加えて，報酬集合は非負実数の有限集合であるとし $(\mathcal{R} = \{\mathrm{r}^1, \mathrm{r}^2, \dots, \mathrm{r}^L\} \subset \mathbb{R}_{\geq 0} = [0, \infty))$，(1), (3) 式の特殊形として

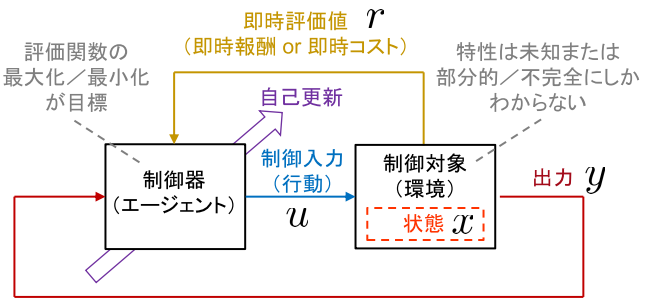

図 1　強化学習問題を表わすブロック線図

強化学習問題は，制御対象の特性がわからない状況で，データに基づく自己更新によって評価関数を最大化または最小化する制御則を自ら獲得する制御器を構成する問題である．即時評価値は制御対象内で生成されるとして図示している．

$$x_{k+1} = f(x_k, u_k, e_k^{\mathrm{p}}) \tag{4}$$

$$r_{k+1} = \rho(x_k, u_k, e_k^{\mathrm{r}}) \tag{5}$$

を考える．$\mathcal{E}^{\mathrm{p}}, \mathcal{E}^{\mathrm{r}}$ も有限集合とし，$e_k^{\mathrm{p}}, e_k^{\mathrm{r}}$ が時刻によらず同一の同時確率分布 $\{p(e^{\mathrm{p}}, e^{\mathrm{r}})\}_{e^{\mathrm{p}} \in \mathcal{E}^{\mathrm{p}}, e^{\mathrm{r}} \in \mathcal{E}^{\mathrm{r}}}$ に従うとすると，制御対象の状態変化と報酬の発生も確率的となり，(4), (5) 式は離散時間定常有限マルコフ決定過程を表わす式となる．確認は容易なので詳細は省略するが，このとき $x_k = x$ で $u_k = u$ を加えたときに $x_{k+1} = x', r_{k+1} = r$ となる確率が時刻によらない条件付き同時確率分布

$$\{p(x', r|x, u)\}_{x, x' \in \mathcal{X}, u \in \mathcal{U}, r \in \mathcal{R}} \tag{6}$$

で表わされることになる．表記単純化のため，これ以降は (6) 式を $\{p(x', r|x, u)\}$ と表わす（他の場合も同様）．この状況における典型的な評価関数は割引因子 γ $(0 \leq \gamma < 1)$ を用いて定義される割引付き累積報酬の期待値

$$\mathbb{E}[\textstyle\sum_{k=0}^{\infty} \gamma^k r_{k+1}] \tag{7}$$

である．$\{p(x', r|x, u)\}$ が既知の状況でこの評価関数を最大化する制御則を導出する問題を本稿では（色々と形容詞を省略して）割引付き有限マルコフ決定問題と呼ぶ．$\{p(x', r|x, u)\}$ が未知または部分的／不完全にしかわからない状況で評価関数 (7) を最大化するような制御則を制御器に獲得させるというのが，強化学習分野における最も標準的な問題設定である．強化学習分野では，通常は (6) のような式から説明が始まるが，(4), (5) 式から出発したことで標準的な強化学習問題を少し身近に感じ，そしてそれが「適応最適制御問題」であると納得してもらえたら本稿の目標の一端は達成されたことになる．

強化学習問題に対する理解を深めるため，もう 1 つ例をあげる．やはり完全観測 $(y_k = x_k)$ で，$\mathcal{X} = \mathbb{R}^n$, $\mathcal{U} = \mathbb{R}^m$ の場合を考える．f, ρ が $A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}$ および半正定対称行列 $Q \in \mathbb{R}^{n \times n}$，正定対称行列 $R \in \mathbb{R}^{m \times m}$ を用いて $f(x, u) = Ax + Bu, \rho(x, u) = x^\top Qx + u^\top Ru$ と表わされるとき，(1)，(3) 式の特殊形として

$$x_{k+1} = Ax_k + Bu_k \tag{8}$$

$$r_{k+1} = x_k^\top Qx_k + u_k^\top Ru_k \tag{9}$$

を得る（(A, B) は可到達とする）．評価関数が累積コスト

$$\sum_{k=0}^{\infty} r_{k+1} = \sum_{k=0}^{\infty} \left(x_k^\top Qx_k + u_k^\top Ru_k \right) \tag{10}$$

で，これを最小化する制御則を導くとすると，われわれがよく知る離散時間 LQR 問題となる．A, B, Q, R が未知または部分的／不完全にしかわからない状況で評価関数 (10) を最小化する制御則を制御器に獲得させるとすると，先ほどと同様に強化学習問題（適応最適制御問題）となる．

補足 1：ここまでの説明では，強化学習問題と適応制御問

題の共通性を強調してきた．一方で，後者における標準的な設定では，制御器の目標は規範モデルの出力への追従とされる．この違いのため，強化学習手法といわゆる適応制御手法では制御則・更新則の具体形は異なったものとなる．最適制御問題を適応的に解く手法については，強化学習以外に適応動的計画法 (adaptive dynamic programming) や近似動的計画法 (approximate〜) といった名称で研究されていることも多い．このような類義語については文献 2) の §1.3.1 を参照いただきたい．

補足 2：本稿冒頭で，強化学習は一般的には機械学習の一分野に位置づけられることが多いと述べた．機械学習の立場から見た場合，強化学習問題の大きな特徴は，報酬という手掛かりは与えられるものの，ある状態で取るべき行動の正解（教師データ）が与えられないことである．また，環境が状態を有する場合，エージェントの行動がその瞬間だけではなく環境の将来にも影響を及ぼし，それに伴って後々獲得するデータを変えるという点も他の多くの機械学習問題にはない特徴である．この点で，強化学習は機械学習の中で以前からダイナミクスと制御を扱ってきた分野ということができる．

3. 強化学習手法

強化学習問題に対して有効に働く制御器を本稿では強化学習制御器と呼ぶことにする．強化学習制御器の構造は一般に**図 2** のようになっている．すなわち，制御器内には可調整パラメータを含んだ制御則（強化学習分野では方策；policy と呼ばれる）とそれを変更する更新則が存在する．制御則は常に制御入力を決定し続け，その一方で更新則は出力，更新則が決定した制御入力，即時評価値に基づいて制御則内の可調整パラメータを適切なタイミングで変更する．前章で強化学習問題と適応制御問題との関係をみたので，制御器がこのような構造を取ることは自然に理解されるであろう．この共通構造のもと，当然ながら制御則と更新則を規定するさまざまな手法が存在する．強化学習手法とはそれらの総称である．

以下では，前章で説明した 2 つの問題設定を題材に強化学習手法の具体例を説明する．説明の仕方としては，まず制御対象の特性（$\{p(x', r|x, u)\}/A, B, Q, R$）が既知の場合の最適制御手法（方策反復法；policy iteration）を説明し，それを制御対象の特性がわからない状況に徐々に適合させていく．これにより強化学習の難しさがどこにあるかが見て取れると思う．また，以下で紹介するのはごく限られた手法であるが，そこに現れる考え方や用語は他の手法について学ぶ際にも役立つはずである．

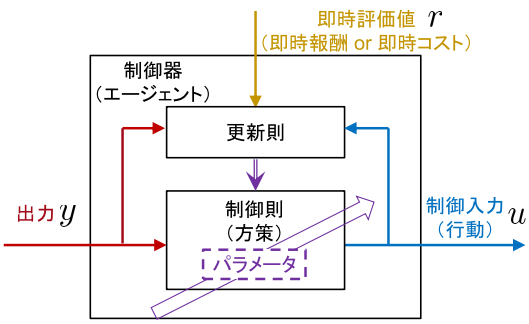

図 2　強化学習制御器の構造
強化学習制御器は，可調整パラメータを含んだ制御則と，可調整パラメータを変更する更新則によって構成される．更新則も制御則も一般には動的であるが，上図では制御則が可調整パラメータを含むこと以上の詳細は省略した．

3.1 割引付き有限マルコフ決定問題の場合

方策反復法：関数 $\pi : \mathcal{X} \to \mathcal{U}$ による状態フィードバック (FB) 制御則

$$u_k = \pi(x_k) \tag{11}$$

を考える．有限集合 $\{\pi(x)\}$ が関数 π 自体を表わすことに注意しよう．次式で定義される $\mathsf{V}_\pi : \mathcal{X} \to \mathbb{R}_{\geq 0}$ を制御則 π の下での状態価値関数 (state value function) という．

$$\mathsf{V}_\pi(x) := \mathbb{E}[\sum_{s=0}^{\infty} \gamma^s r_{k+1+s} | x_k = x, u_{k+s} = \pi(x_{k+s}), s \geq 0]$$

$\mathsf{V}_\pi(x)$ は状態 x から制御則 π に従うことで得られる割引付き累積報酬の期待値であり，最適制御則 π^* は任意の制御則 π，任意の x に対し $\mathsf{V}_{\pi^*}(x) \geq \mathsf{V}_\pi(x)$ を満たすものとして定義される．また，無限和であることを利用して右辺を変形するとつぎの方程式（ベルマン方程式）を得る．

$$\mathsf{V}_\pi(x) = \sum_{(x', r) \in \mathcal{X} \times \mathcal{R}} p(x', r|x, \pi(x))[r + \gamma \mathsf{V}_\pi(x')] \tag{12}$$

条件付き同時確率分布 $\{p(x', r|x, u)\}$ が既知であれば，これは $N(= |\mathcal{X}|)$ 個の未知数 $\{\mathsf{V}_\pi(x)\}$ に対する N 本の連立線形方程式であることに注意しよう．

一方，次式で定義される $\mathsf{Q}_\pi : \mathcal{X} \times \mathcal{U} \to \mathbb{R}_{\geq 0}$ を，制御則 π の下での状態–行動価値関数 (state-action value function) または行動価値関数 (action〜) という．

$$\mathsf{Q}_\pi(x, u) := \mathbb{E}[\sum_{s=0}^{\infty} \gamma^s r_{k+1+s} | x_k = x, u_k = u, u_{k+s} = \pi(x_{k+s}), s \geq 1] \tag{13}$$

これは状態 x で入力 u を使用した後，制御則 π に従うことで得られる割引付き累積報酬の期待値である．この Q_π の使い道であるが，新しい FB 制御則 $\pi' : \mathcal{X} \to \mathcal{U}$ を

$$\pi'(x) := \arg\max_{u \in \mathcal{U}} \mathsf{Q}_\pi(x, u) \tag{14}$$

で定義すると[注3]，任意の x に対して $\mathsf{V}_{\pi'}(x) \geq \mathsf{V}_\pi(x)$ となること，すなわち，制御則を改善できることが示せ

[注3] ここでは，複数の u が最大値を与える場合はそのうちの 1 つを選ぶという意味で記号 arg max を使用している．

る[3)~5)]．同時に，(14) 式の操作のあとで $\pi' = \pi$ であったときにはそれが最適制御則 π^* であることも示せる．また，(13) 式を変形すると以下の式が得られる．

$$\mathsf{Q}_\pi(x,u) = \sum_{(x',r)\in\mathcal{X}\times\mathcal{R}} p(x',r|x,u)[r+\gamma\mathsf{Q}_\pi(x',\pi(x'))] \tag{15}$$

$$= \sum_{(x',r)\in\mathcal{X}\times\mathcal{R}} p(x',r|x,u)[r+\gamma\mathsf{V}_\pi(x')] \tag{16}$$

(15) 式は状態–行動価値関数に対するベルマン方程式である．(16) 式からは $\{p(x',r|x,u)\}$ と $\{\mathsf{V}_\pi(x)\}$ から $\{\mathsf{Q}_\pi(x,u)\}$ が計算できることがわかる．

以上の考察から，$\{p(x',r|x,u)\}$ が既知の状況で最適制御則を導出する以下のアルゴリズムが得られる．

[割引付き有限マルコフ決定問題に対する方策反復法]

1. 最初の FB 制御則 π を定める．
2. $\{p(x',r|x,u)\}$ と π を用い，ベルマン方程式 (12) に基づいて $\{\mathsf{V}_\pi(x)\}$ を計算する．
3. $\{p(x',r|x,u)\}$ と $\{\mathsf{V}_\pi(x)\}$ を用い，(16) 式に基づいて $\{\mathsf{Q}_\pi(x,u)\}$ を計算する．
4. (14) 式によって新しい FB 制御則 π' を定義する．
5. $\pi' = \pi$ であれば終了．そうでなければ π' を新しい π として 2. に戻る．

強化学習手法の構成：$\{p(x',r|x,u)\}$ が未知または部分的／不完全にしかわからない状況で制御器に最適制御則を獲得させるのが強化学習問題だが，大きな流れとしては上記の方策反復法を踏襲することで強化学習手法を構成できる．今は $\{\pi(x)\}$ が有限集合なので，これ全体を可調整パラメータとできる．最終的に制御則の改善に使ったのが $\{\mathsf{Q}_\pi(x,u)\}$ であることに着目すると，入力の決定と状態・即時報酬の観測を行いつつ，得られたデータに基づいて推定値 $\{\widehat{\mathsf{Q}}_\pi(x,u)\}$ を求め，これを (14) 式右辺に代入した

$$\pi'(x) := \arg\max_{u\in\mathcal{U}} \widehat{\mathsf{Q}}_\pi(x,u) \tag{17}$$

で制御則を更新するという基本方針を立てられる．これを実現する 1 つの方向として，$\{p(x',r|x,u)\}$ の推定値 $\{\widehat{p}(x',r|x,u)\}$ を計算し，上記ステップ 2., 3. で使用することが考えられる．制御対象のモデルの構築（システム同定）を経由して制御則の変更を行うこのような方式をモデルベースト強化学習という（適応制御における間接法に対応する）．一方，$\{\widehat{p}(x',r|x,u)\}$ を経由せずに直接 $\{\widehat{\mathsf{Q}}_\pi(x,u)\}$ を計算する方向も考えられる．このような方式をモデルフリー強化学習という（適応制御における直接法に対応する）．以下ではテーブル型 SARSA(0) 法に基づいたモデルフリー手法の構成を説明する．

まず，この状況での推定値 $\{\widehat{\mathsf{Q}}_\pi(x,u)\}$ の直接計算は

$$\sum_{(x,u)\in\mathcal{X}\times\mathcal{U}} \Big(\mathsf{Q}_\pi(x,u) - \widehat{\mathsf{Q}}_\pi(x,u)\Big)^2 \tag{18}$$

を評価関数とし，$\{\widehat{\mathsf{Q}}_\pi(x,u)\}$ を決定変数とする最小化問題と考えることができる．この最小化問題に対し，確率的勾配降下法を（形式的に）適用すると，時刻 $k+1$ に実行する推定値の更新式として以下を得る．

$$\widehat{\mathsf{Q}}_\pi(x_k,u_k) \leftarrow \widehat{\mathsf{Q}}_\pi(x_k,u_k) + \alpha_k\Big(\mathsf{Q}_\pi(x_k,u_k) - \widehat{\mathsf{Q}}_\pi(x_k,u_k)\Big)$$

α_k は学習係数やステップサイズと呼ばれる更新の大きさを調整する正の実数値であり，矢印は値の書き換えを表わす．$\mathsf{Q}_\pi(x_k,u_k)$ を実際に観測することはできないので，$\mathsf{Q}_\pi(x_k,u_k) = \mathbb{E}[r_{k+1}+\gamma\mathsf{Q}_\pi(x_{k+1},\pi(x_{k+1}))]$（(15) 式）を手掛かりに，観測された即時報酬と現在の推定値から計算できる $r_{k+1}+\gamma\widehat{\mathsf{Q}}_\pi(x_{k+1},\pi(x_{k+1}))$ に置き換えると以下を得る（テーブル型 SARSA(0) 法）．

$$\widehat{\mathsf{Q}}_\pi(x_k,u_k) \leftarrow \widehat{\mathsf{Q}}_\pi(x_k,u_k) + \alpha_k\delta_k \tag{19}$$

$$\delta_k := r_{k+1} + \gamma\widehat{\mathsf{Q}}_\pi(x_{k+1},\pi(x_{k+1})) - \widehat{\mathsf{Q}}_\pi(x_k,u_k) \tag{20}$$

δ_k は TD 誤差 (temporal difference error) と呼ばれるもので，場合によって定義式に差異はあるが強化学習手法において頻出する．(19), (20) 式はテーブル型 TD(0) 法と呼ばれるアルゴリズムの状態–行動価値関数版となっているので，学習係数が $\sum_{k=0}^{\infty}\alpha_k = \infty, \sum_{k=0}^{\infty}\alpha_k^2 < \infty$（Robbins-Monro 条件）を満たし，すべての $\widehat{\mathsf{Q}}_\pi(x,u)$ が無限回更新されるならば，$\{\widehat{\mathsf{Q}}_\pi(x,u)\}$ は真値に概収束する．

ここまでの説明に基づくと，FB 制御則 (11) による入力の決定と状態・即時報酬の観測を行いながら，(19), (20) 式と (17) 式を使って $\{\pi(x)\}$ を更新していけば良いように思えるが，実際にはモデルフリー手法を構成する上で考えるべき課題が 2 つある．1 つは，確定的な FB 制御則にしたがっていると 1 つの状態 x に対して常に同じ制御入力が生成され，$\{\widehat{\mathsf{Q}}_\pi(x,u)\}$ の中に更新されない要素が出てくるという課題である．これについては，たとえば

$$u_k = \begin{cases} \text{choose } u\in\mathcal{U} \text{ randomly} & \text{with prob. } \epsilon_k \\ \pi(x_k) & \text{with prob. } 1-\epsilon_k \end{cases}$$

とするなどして，π によって計算される（良いとされる）制御入力以外も選択されるようにして解決が図られる．探索や試行錯誤と表現されるこの要素は，システム同定における PE 性の確保に対応している．

もう 1 つは，強化学習問題では実際に制御対象に入力を印加しなければならないため，最適でない π に対する $\{\widehat{\mathsf{Q}}_\pi(x,u)\}$ の更新を繰り返すことはある種の損失を生むという課題である．これに対しては，推定値 $\{\widehat{\mathsf{Q}}_\pi(x,u)\}$ が収束していないと思われる段階でも (17) 式によって制

御則を更新することで解決が図られる．最も極端な場合には (19), (20) 式の演算を一回行っただけで制御則更新が行われるが，このようなケースでも，適切な条件の下では最適制御則への概収束が示せる[6]．

3.2 離散時間 LQR 問題の場合

先ほど同様，方策反復法から説明する．われわれはこの問題の解が線形状態 FB 制御則であることを知っているので，

$$u_k = Fx_k \tag{21}$$

を考える（$F \in \mathbb{R}^{m\times n}$ は $A+BF$ が Schur 安定行列となるよう選ばれているものとする）．このとき，F の下での状態価値関数 $\mathsf{V}_F : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ は $\mathsf{V}_F(x) := \sum_{s=0}^{\infty} x^\top [(A+BF)^\top]^s (Q + F^\top RF)(A+BF)^s x$ で定義されるが，これはリアプノフ方程式

$$P_F = (A+BF)^\top P_F (A+BF) + Q + F^\top RF$$

の正定対称解 $P_F \in \mathbb{R}^{n\times n}$ を用いて $\mathsf{V}_F(x) = x^\top P_F x$ と書ける．状態–行動価値関数 $\mathsf{Q}_F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ も同様に定義されるが，結果として以下がなりたつ．

$$\mathsf{Q}_F(x,u) = \rho(x,u) + \mathsf{Q}_F(Ax+Bu, F(Ax+Bu)) \tag{22}$$

$$= x^\top Qx + u^\top Ru + \mathsf{V}_F(Ax+Bu) \tag{23}$$

$$= \begin{bmatrix} x \\ u \end{bmatrix}^\top \begin{bmatrix} S_F^{11} & S_F^{12} \\ (S_F^{12})^\top & S_F^{22} \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} \tag{24}$$

(22) 式が $\mathsf{Q}_F(x,u)$ に対するベルマン方程式であり，$S_F^{11} = Q + A^\top P_F A$, $S_F^{12} = A^\top P_F B$, $S_F^{22} = R + B^\top P_F B$ である．(14) 式は $F'x := \arg\min_{u\in\mathbb{R}^m} \mathsf{Q}_F(x,u)$ となるが，Q_F が二次形式なので右辺を満たす u は $\nabla_u \mathsf{Q}_F(x,u) = 0$ を解くことで得られ，その結果は以下の FB 係数行列の更新式となる．

$$F' = -(S_F^{22})^{-1}(S_F^{12})^\top \tag{25}$$

以上より，A, B, Q, R を用いて P_F, S_F, F' の計算を繰り返す離散時間 LQR 問題に対する方策反復法（＝離散時間代数リカッチ方程式に対するニュートン法[7]）を得る．

やはり先ほど同様，A, B, Q, R が未知な場合でも，FB 係数行列 F を可調整パラメータとし，S_F の推定値 $\widehat{S}_F$ を求めて $F' = -(\widehat{S}_F^{22})^{-1}(\widehat{S}_F^{12})^\top$ とすることを基本方針として方策反復法ベースの強化学習手法が構成できる．A, B, Q, R の推定値 $\widehat{A}, \widehat{B}, \widehat{Q}, \widehat{R}$ を求めて使用すればモデルベースト手法となる．以下では，逐次最小二乗法に基づいたモデルフリー手法[8)~10)] を紹介する．対称行列 $S \in \mathbb{R}^{(m+n)\times(m+n)}$ に対し，$\mathsf{Q}(x,u;S) := [x^\top, u^\top] S [x^\top, u^\top]^\top$ と定義する．ベルマン方程式 (22) より，この状況での推定値 $\widehat{S}_F$ の直接計算は

$$\sum_{(x,u)\in\mathcal{S}} \Big(r + \mathsf{Q}(x', Fx'; \widehat{S}_F) - \mathsf{Q}(x,u;\widehat{S}_F) \Big)^2$$

を評価関数とし，$\widehat{S}_F$（正確には，その中に含まれる $(m+n+1)(m+n)/2$ 個の独立な実変数）を決定変数とする最小化問題と考えることができる．ただし，$r = \rho(x,u), x' = Ax+Bu$ であり，集合 $\mathcal{S} \subset \mathbb{R}^n \times \mathbb{R}^m$ は $r + \mathsf{Q}(x', Fx'; \widehat{S}_F) - \mathsf{Q}(x,u;\widehat{S}_F) = 0$ が $(m+n+1)(m+n)/2$ 本の 1 次独立な方程式となるような集合である．この最小化問題に対して逐次最小二乗法を適用すると，TD 誤差

$$\delta_k := r_{k+1} + \mathsf{Q}(x_{k+1}, Fx_{k+1}; \widehat{S}_F) - \mathsf{Q}(x_k, u_k; \widehat{S}_F)$$

を用いた $\widehat{S}_F$ の更新式が導ける．この場合も (21) 式そのままでは PE 性が確保できないため，$u_k = Fx_k + \varepsilon_k$ という摂動を加えた制御則を使うことで解決が図られる（$\varepsilon_k \in \mathbb{R}^m$）．ある F に対して $\widehat{S}_F$ の更新をどれだけ繰り返すか適切に決定しなければならないことも同様である．

補足：方策反復法をもう少し抽象的に見てみると，ステップ 2., 3. で制御則改善のための手がかりを計算し，ステップ 4. で制御則の改善を実行している．制御則改善のための手がかりとして，制御則の可調整パラメータに関する評価関数の勾配等も考えられ，それをベースに強化学習手法を構成することもできる[1),4),11)~14)]．この場合も，制御器の構造としてはやはり図 2 で理解することができる．

4. モデルベースト制御との融合

2 章の最後で強化学習は機械学習の中でダイナミクスを考慮してきた分野であると述べたが，強化学習分野の知見と制御工学分野の知見を何らかの形で融合することでより良い制御器の構成が期待でき，実際そのような研究も行われている．ここではモデルフリー強化学習手法とモデルベースト制御手法の融合について，いくつかの方向を述べる．

まず，モデルベースト制御手法の知見を活かしたモデルフリー強化学習手法という方向がある．3 章の後半で紹介した離散時間 LQR 問題に対する強化学習手法では，最適な制御則が線形状態フィードバックであることや，状態–行動価値関数が二次形式となること等のモデルベースト制御手法の知見が活用されている．より一般の場合[2),14)~19)] でも，モデルベースト制御手法の知見を制御則および更新則の設計に生かすことができる．

また，不完全ながらも事前に知りえた制御対象のモデル（ノミナルモデル）を使って，モデルベースト制御手法と強化学習手法を併用するという方向もある[20)~22)]．強化学習手法においては一般に，学習（パラメータ調整）

に時間がかかったり，学習中に不適切な制御入力が生成されうるという課題が存在する．強化学習側から見ると，モデルベースト制御手法の利用により，これらの改善が期待できる．一方，制御対象のモデルが誤差を含んでいる場合，モデルベースト制御手法は実際の制御対象に対して期待した性能を発揮できない．モデルベースト制御側から見ると，強化学習手法を用いることで実制御対象に対する最終的な性能の向上が期待できる．

このほか，事前に設計された複数のモデルベースト制御器の切り替えに強化学習制御器を利用する方法[23]も提案され，効果を上げている．強化学習と制御工学の深い結びつきを考えると，ここにあげられなかった新しい融合の仕方も数多く存在するものと思われる．本稿がそのような発展に役立てば幸いである．

5. おわりに：文献案内

本稿では制御工学との関係についてはやや詳しく説明を行ったが，強化学習の観点から見ると基本的な内容のごく一部しか紹介できなかった．最後に，今後の研究・応用に向けた参考のため，執筆にあたり参照した文献と本編で触れられなかった話題に関する文献を紹介する．

強化学習の標準的内容については文献 1) に加え，文献 4), 5), 24)，本誌ならびにシステム制御情報学会誌『システム／制御／情報』に過去掲載された解説記事を主に参照し，2 章では 3), 25) も参考にした．離散時間 LQR 問題に対する強化学習手法に関しては文献 8)〜 10) に加え 16), 17) も参考にした．テーブル型 SARSA(0) 法については，文献 24), 26) の TD(0) 法の説明を参考にしつつ，逐次最小二乗法との類似性を強調する説明を行った．テーブル型 TD(0) 法のより詳細な説明については文献 4), 24), 27) を参照いただきたい．本稿で触れられなかった発展的な内容（階層型強化学習，逆強化学習，脳神経科学との関係，具体的な応用例等）については，上記二誌の各種解説記事や，文献 13), 14), 18) および文献 1) の対応する各章を参照いただきたい．深層強化学習については，文献 4) 邦訳版付録 C と文献 28) が参考になる．

謝辞　本稿に関わる多くの話題についてご教示・ご議論いただきました沖縄科学技術大学院大学 銅谷賢治教授，国際電気通信基礎技術研究所 内部英治主幹研究員に感謝いたします．（2018 年 12 月 28 日受付）

参 考 文 献

1) R. S. Sutton and A. G. Barto: *Reinforcement Learning: An Introduction*, MIT Press, 2nd edition (2018)
2) D. Liu et al.: *Adaptive Dynamic Programming with Applications in Optimal Control*, Springer (2017)
3) M. L. Puterman: *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, paperback edition (2005)
4) C. Szepesvári: *Algorithms for Reinforcement Learning*, Morgan & Claypool (2010). 邦訳=小山田創哲（訳者代表・編集），前田新一・小山雅典（監訳）：速習　強化学習―基礎理論とアルゴリズム―，共立出版 (2017). 註：本稿では邦訳および原著者 web サイト掲載の 2018 年 6 月 25 日改訂版原稿を合わせて参照した．
5) D. P. Bertsekas: *Dynamic Programming and Optimal Control Vol. I & II*, Athena Scientific, 4th edition (2017, 2012)
6) S. Singh et al.: Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms, *Machine Learning*, **38**–3, 287/308 (2000)
7) 西村，狩野：制御のためのマトリクス・リカッチ方程式，朝倉書店 (1996)
8) S. J. Bradtke: Reinforcement Learning Applied to Linear Quadratic Regulation, In *Advances in Neural Information Processing Systems*, 295/302 (1993)
9) S. J. Bradtke et al.: Adaptive Linear Quadratic Control Using Policy Iteration, In *Proc. Amer. Contr. Conf.*, 3475/3479 (1994)
10) S. J. Bradtke: Incremental Dynamic Programming for On-Line Adaptive Optimal Control, CMPSCI Technical Report 94-62, University of Massachusetts Amherst (1994)
11) 白川，森村：方策勾配に基づくアルゴリズム，文献 18), 42/55 (2016)
12) M. P. Deisenroth et al.: A Survey on Policy Search for Robotics, *Foundations and Trends in Robotics*, **2**–1-2, 1/142 (2013)
13) M. Wiering and M. van Otterlo, eds.: *Reinforcement Learning: State-of-the-Art*, Springer (2012)
14) F. L. Lewis and D. Liu, eds.: *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, IEEE Press (2013)
15) 銅谷，森本，鮫島：強化学習と最適制御，システム/制御/情報，**45**–4, 186/196 (2001)
16) F. L. Lewis et al.: Reinforcement Learning and Feedback Control: Using Natural Decision Methods to Design Optimal Adaptive Controllers, *IEEE Contr. Syst. Mag.*, **32**–6, 76/105 (2012)
17) D. Vrabie et al.: *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*, IET (2013)
18) 牧野，澁谷，白川（編著）：これからの強化学習，森北出版 (2016)
19) R. Kamalapurkar et al.: *Reinforcement Learning for Optimal Feedback Control: A Lyapunov-Based Approach*, Springer (2018)
20) S. Levine and V. Koltun: Guided Policy Search, In *Proc. Int. Conf. on Machine Learning*, 1/9 (2013)
21) J. Si et al.: *Toward Design of Nonlinear ADP Learning Controllers with Performance Assurance*, 文献 14), 182/202 (2013)
22) 大川，佐々木，岩根：強化学習とモデルベースト制御を並列した制御アプローチ，第 61 回自動制御連合講演会予稿集，152/159 (2018)
23) 吉本，銅谷，石井: 強化学習の基礎理論と応用，計測と制御，**44**–5, 313/318 (2005)
24) D. P. Bertsekas and J. N. Tsitsiklis: *Neuro-Dynamic Programming*, Athena Scientific (1996)
25) P. J. Werbos: *Reinforcement Learning and Approximate Dynamic Programming (RLADP) - Foundations, Common Misconceptions, and the Challenges Ahead*, 文献 14), 3/30 (2013)
26) C. Dann et al.: Policy Evaluation with Temporal Differences: A Survey and Comparison, *The Journal of Machine Learning Research*, **15**, 809/883 (2014)
27) 前田新一：統計学習の観点から見た TD 学習，文献 18), 72/111 (2016)
28) K. Arulkumaran et al.: A Brief Survey of Deep Reinforcement Learning, arXiv preprint, arXiv:1708.05866 (2017)

［著　者　紹　介］

佐々木　智丈 君（正会員）

2010 年 3 月東京大学大学院情報理工学系研究科システム情報学専攻博士課程修了．博士（情報理工学）．同専攻システム情報第 5 研究室学術支援専門職員を経て 2010 年 10 月（株）富士通研究所入社，2018 年 10 月より同社シニアリサーチャーとなり現在に至る．マサチューセッツ工科大学および Center for Brains, Minds and Machines リサーチアフィリエイト．制御工学，強化学習，深層学習等の研究に従事．システム制御情報学会，電気学会などの会員．

加嶋　健司 君（正会員）

（本号 p.155 参照）