訳者まえがき

Google の AlphaGo によるプロ棋士打破は，人工知能がヒトを超えた学習を行った歴史的出来事として認識された．この事例が象徴的に現わしているように，人の手によって作られた正解例をもとに学習する教師有り学習とは異なり，強化学習ではこれまで人が試したこともないような，人を超える手を生み出すことを可能とした．強化学習は，囲碁などのゲームのほかにも自動運転やロボット制御など制御分野への応用でも注目されている．このように，強化学習は，実社会にインパクトを与える応用が生まれつつあり，機械学習の中でも特にホットな分野といえる．

その一方で，日本語で強化学習を体系的に学べる教科書はまだ多くはない．また，代表的な強化学習の教科書である Sutton and Barto (1998) の邦訳書では，新しいアルゴリズムが十分には掲載されていない．実際，本書に掲載されている約 3 分の 1 のアルゴリズムは，Sutton and Barto (1998) では触れられていない．本書の翻訳のきっかけもまさにそこに起因している．今回，特に強化学習を学ぶことに意欲的な若手研究者，実務家，大学院生が集まったことで，本書の勉強会が企画されるのみならず，訳者総勢 12 名による翻訳作業が始まることとなった次第である．

ここで，本書がどういった本であるかを簡単にご紹介したい．本書は出版当時，トップ会議 (AAAI) のチュートリアルで利用されたり，出版以降わずか数年で 400 弱の引用がされたりといった事実から窺えるように，入門書として広く読まれている良書である．本書の内容は動的計画法などの基本的かつ重要なアルゴリズムから始まり，比較的新しい手法の基礎となる部分が網羅されながら，全体の分量はコンパクトに抑えられている．本書の想定読者は大学生・大学院生であり，特別な前提知識なしに自己完結しており，平易な語り口でアルゴリズムのエッセンスが理解しやすいよう工夫されているため，自習にも適している．翻訳では，この原文のニュアンスを保持したまま，日本語として自然に理解できるよう腐心したため，時には直訳からは出てこないような翻訳も敢えて行った．出版から

7年あまり過ぎたことで，カバーされていないアルゴリズムも存在するものの，その多くは本書で掲載されたアルゴリズムをその基礎においている．特に近年の深層学習を利用した強化学習アルゴリズムに，本書で紹介されたアルゴリズムがどのように使われているかを簡単に解説した付録を追加することで，近年の深層学習を利用した強化学習アルゴリズムに対する理解が深まるような工夫を行っている．

本書の出版にあたり，多くの方々にご協力を頂いた．草稿の輪読に協力して頂いた小川義人さん，菊池悠太さん，久米絢佳さん，関谷英爾さん，奥村エルネスト純さん，大録誠広さん，川尻亮真さんには深く感謝申し上げたい．また，共立出版の方々には本書の企画から刊行に至るまで手厚くサポートをして頂いたことをここに記し，厚く御礼申し上げる．

本書が，これから強化学習を学びたいという読者の学習の一助になれば幸いである．

2017年7月

前田新一・小山雅典・小山田創哲

まえがき

強化学習 (reinforcement learning; RL) は機械学習の一つの分野と学習問題の一種の両方を指す言葉であり，学習問題としては，長期的な目標を示す数値を最大化するようシステムを制御する学習を指す．図 1 は，強化学習の典型的な設定を示している．制御器はまず，制御対象となるシステムから現在の状態と直前の状態遷移に伴う報酬を観測し，それに基づいてシステムに対して働きかける行動を計算する．システムがこの制御器の行動に応答し，新しい状態に遷移することで，サイクルが繰り返される．ここで主眼となる問題は，報酬和を最大化するようにシステムを制御する制御方法の学習である．このような学習問題は，データがどのように得られるか，どのように性能が評価されるかといった詳細がそれぞれの問題で異なる．

本書では，我々が制御対象とするシステムが確率的であると仮定する．さらに，ここでは制御器がシステムの状態を推論する必要がないぐらい十分，詳細にシステムの状態を観測できることを仮定する．このような特徴をもった問題は，マルコフ決定過程 (Markov decision process; MDP) の枠組みでうまく記述できる．この MDP を “解く” ための標準的なアプローチは動的計画法である．動的計画法は，良い制御器を見つける問題を良い価値関数を見つける問題に落とし込む手法である．しかしながら，MDP が極めて少ない状態数と行動数しかもたない単純な問題から離れると，動的計画法は実行不可能となる．こ

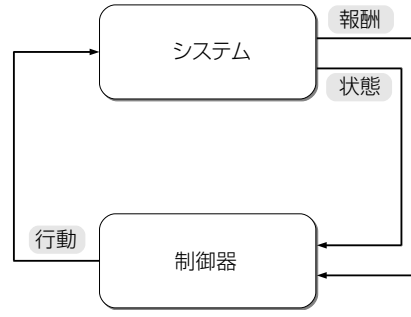

図 1　基本的な強化学習のシナリオ

こで我々が議論する強化学習アルゴリズムは，大規模な問題では実行不可能となる動的計画法を実践的なアルゴリズムに落とし込み，大規模な問題に対して適用可能にするための方法といえる．

強化学習アルゴリズムにこの目標を達成させるにあたって，重要となるアイデアは二つある．一つ目は，制御問題のダイナミクスをコンパクトに表現するためにサンプルを使うことである．このアイデアが重要である理由は二つある．一つは，サンプルを使うと，ダイナミクスが未知の場合でも学習を扱うことが可能になること，もう一つは，たとえダイナミクスが既知であっても，そのダイナミクスに基づいた正確な推論が困難であるかもしれないこと，にある．強化学習アルゴリズムの背後にある二つ目の重要なアイデアは，価値関数をコンパクトに表現するために関数近似法という強力な道具を用いることである．これには，大規模で高次元な状態・行動空間を扱うことを可能にする意義がある．都合の良いことに，この二つのアイデアは相性が良い．なぜなら，サンプルがそれらの属する空間のうちの小さな部分領域に集中していれば，賢い関数近似手法はその性質を利用できるからである．強化学習アルゴリズムを設計・分析するにせよ，適用するにせよ，その核心には動的計画法，サンプルおよび関数近似の間の相互作用の理解が必要になる．

本書のゴールは読者に対してこの美しい分野を垣間見る機会を提供することである．もちろん，このゴールを目指したのは我々が初めてではない．1996 年には Kaelbling らが当時のアプローチとアルゴリズムについて，簡潔で素晴らしいサーベイを書いているし (Kaelbling et al., 1996)，続いて理論的基礎について詳細に記述した本も出版された (Bertsekas and Tsitsiklis, 1996)．数年後，強化学習の “父” である Sutton と Barto が著書を出版し，強化学習における彼らの考えを明快かつわかりやすい形で提示した (Sutton and Barto, 1998)．より新しく包括的な動的計画法/最適制御のツールやテクニックに関する概説は Bertsekas (2007a,b) の上下巻の著書で与えられており，その中で一章分が強化学習の手法に充てられている[1)]．時に，分野が急速に発展する状況では，書籍はすぐに時代遅れになってしまう．実際，増え続ける新しい結果に対応するために，Bertsekas は著書のオンライン版において下巻の第 6 章の整備を続けており，その分量は本書の執筆時点で 160 ページにも及ぶ (Bertsekas, 2010)．他に関連する近年の書籍には，60 ページを第 9 章の強化学習アルゴリズムに充て，平均コスト問題に集中した Gosavi (2003) や，方策勾配法に焦点を当てた Cao (2007) が挙げられる．Powell (2007) はオペレーションズ・リサーチの視点からアルゴリズムとアイデアについて言及したうえで，大規模な制御空間を扱うことのできる手法を強調し，Chang et al. (2007b) は適応サンプリング（つまり，シミュレーションベースの性能最適化）に焦点を当てている．そうした一方，近年出版された Busoniu et al. (2010) は関数近似をその中心に据えている．

[1)] この本では，強化学習はニューロ動的計画法または近似動的計画法と呼ばれている．このニューロ動的計画法という単語は，強化学習アルゴリズムが多くの場合，ニューラルネットワークとともに使われるという事実に由来する．

このように，強化学習の研究者らは良い書籍に事欠いているわけでは決してない．しかしながら Kaelbling et al. (1996) のように，自己完結しつつも比較的短くまとまっている本で，既存の研究者が分野の視座を広げられるに留まらず，初学者が最先端の感覚を養うことができ，しかも最新のコンテンツを備えているものはないように思われる．この穴を埋めることがまさしく本書の目的である．

紙面を短くするため，いくらかの（願わくば深刻ではないと思われる）妥協をする必要があった．一つ目の妥協点は，議論を累積割引報酬和の期待値を指標とする結果のみに絞ることである．この理由は，広くこの指標が使われていることと数学的な扱いが最も容易であることである．次の妥協点は，MDP と動的計画法の背景を極めて簡潔にまとめたことである（補足として，これらの基本的な結果を説明する付録を用意している）．本書は，これらの妥協をしたうえで，読者が本書で示されたアルゴリズムを実装できるだけでなく，強化学習の様々な側面の，何がどのようになっているのかを理解できるレベルで広く浅く取り扱うことを目指す．当然ながら，何について説明するかは選択する必要があった．そこで，何よりも基本的なアルゴリズムやアイデア，そして有効性の実証されている理論を重視することと決めた．さらに，読者に対して複数の選択肢を紹介するだけでなく，その選択に伴うトレードオフを理解してもらうことにも特に注意を払った．可能な限り公平な記述を心がけたが，ご多分に漏れず，個人的なバイアスがいくらか乗ってしまっていることには留意してほしい．また，実用に重きをおいた読者にとって，ここで記述されたアルゴリズムの実装が容易になることを期待して，本書には 20 近くのアルゴリズムの擬似コードを掲載した．

この本が想定する読者は，大学院生や意欲的な学部生のほか，最先端の強化学習について短期間で概観を掴みたい研究者や実務家である．すでに強化学習を用いた研究を行っている研究者であっても，自分がまだあまり馴染み深くない箇所を読んで，強化学習の見識を広げる楽しみ方もできるように心がけた．読者には線形代数，微積分学，確率論の基本的な知識があることを想定している．特に確率変数，条件付き期待値，マルコフ連鎖についての知識を有することを想定している．必要に応じて本質的な概念は説明されるので必須というほどではないが，統計的学習理論について知っていると役に立つだろう．また，本書のいくつかの箇所では，機械学習の回帰手法に関する知識も役立つであろう．

本書は三つのパートから成っている．最初のパートの第 1 章で，必要となる前提知識について説明する．この章において記法の導入を行い，マルコフ決定過程の理論の短い概説と動的計画法の基本的なアルゴリズムに関する説明を行う．MDP と動的計画法についてすでによく知っている読者も，本書で使われる記法に慣れるために，このパートに目を通しておいた方が良いだろう．この章で説明する結果とアイデアは，本書のそれ以降のパートのベースになるので，MDP にあまり馴染みのない読者は読み進める前にこの章の理解に十分な時間を取って頂きたい．

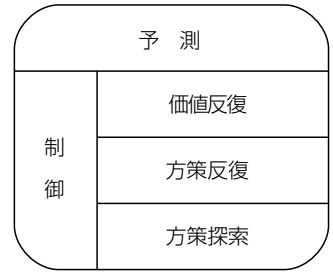

図 2　強化学習における問題とアプローチのタイプ

残りの二つのパートは，強化学習の基本的な二つの問題（図 2）にそれぞれ充てられている．一つ目のパートである第 2 章では，状態に紐付いた価値を予測する学習問題について学ぶ．まず MDP が十分に小さく，状態ごとの価値を配列としてコンピュータのメインメモリに載せられるテーブル形式の場合について，基本的なアイデアを説明する．最初に説明するアルゴリズムは TD(λ) 法と呼ばれる，動的計画法を使った価値反復の学習バージョンともいえる手法である．その後，テーブル形式の場合と異なり，コンピュータのメモリに載せきれないほどの状態数をもつ，より難しくやりがいのある状況について考える．この場合はもちろん，価値を表すテーブルを圧縮する必要が生じるわけであるが，これは，大雑把にいうと，適切な関数近似法によって達成することになる．まず初めに，こうした状況でどのように TD(λ) 法を利用できるかについて述べ，さらに新しい勾配ベースの手法（GTD2 と TDC）について説明を行う．これらは，TD(λ) 法が直面するいくつかの収束性に関する問題を回避できるという点で，TD(λ) 法の改善版ともいえるものである．その後，最小二乗法（特に LSTD(λ) と λ-LSPE）について議論し，先に説明する逐次的な方法と比較する．最後に，関数近似を実装する場合の可能な選択肢と，それぞれの選択に伴うトレードオフについて説明する．

二つ目のパート（第 3 章）は，制御の仕方を学習するために開発されたアルゴリズムの説明に充てている．まず，オンライン性能を最適化することを目標とした方法の説明を行う．特に，“不確かなときは楽観的に” の原則について説明し，この原則に基づいて自らが置かれた環境を探索していく手法について説明する．さらに，バンディット問題と MDP の双方に対する最先端のアルゴリズムを説明する．この章でのメッセージは，巧みに探索することで大きな効果を得ることができるが，それを大規模な問題に対して適用できるようスケールアップするにはさらなる努力が必要になるということである．この章の残りは，大規模な問題に利用できるよう開発された手法について割かれる．大規模な MDP における学習は，小規模な MDP に比べて極めて難しくなるため，その規模が極端に大きくなる場合においては学習の目標は最善の方策を学習することではなく，漸近的に十分良くなる方策を学習することへと緩和される．まず，最適行動価値を直接推定することを目標とした直接法について説明する．この直接法も，動的計画法を使った価値反復の学習バー

ジョンとみなすことができる．その後，動的計画法を使った方策反復の学習バージョンのアルゴリズムと考えられる actor-critic 法について説明する．これについては，直接的な方策改善ベースのものと方策勾配ベースのもの（つまりパラメトリックな方策のクラスを用いるもの）の双方を説明する．第 4 章で，さらに深く探求したい人向けにいくつかのトピックを挙げて締めくくりとする．