連載

5分で分かる!? 有名論文ナナメ読み

小林泰介（奈良先端科学技術大学院大学）

Sergey Levine : Reinforcement Learning and Control as Probabilistic Inference : Tutorial and Review

https://arxiv.org/abs/1805.00909 (2018)

強化学習とは

強化学習は Sutton らが 1980 年代に提唱して以来，目立った注目は集めないながらも着実に基礎を構築してきた．そして近年の深層学習との組合せにより，2016 年には囲碁の世界プロに勝利したことは記憶に新しい．ゲームだけでなく，最近では柔軟物体の取り扱いといったダイナミクスを同定することが困難な複雑な環境におけるロボットの制御にも応用され始めている．技術的にも，世界的な注目を集めてからの研究加速が目覚ましく，GAFA や UC Berkley を中心に新たな手法が日夜登場し，それらの体系化も少しずつだが進められている．本稿で紹介する論文もその体系化を狙ったものの 1 つである．

基本的な強化学習はマルコフ決定過程（Markov Decision Process ; MDP）と呼ばれる環境設定下における最適化手法に分類できる．MDP では，世界にはエージェント（学習者）と環境が存在し，互いに状態と行動をやり取りしている．エージェントの行動によって更新される環境の状態はマルコフ性にしたがっており，現状態と行動にのみ依存して確率的なダイナミクスで与えられる．よって，行動も現状態に基づいて決定すれば十分であり，現状態を条件とした行動をサンプルする確率分布を方策と呼ぶ．強化学習ではこの方策を最適化したいのだが，その規範となる情報として，環境から状態と行動に基づく報酬が与えられる．この報酬の将来に渡っての総和を最大化することを目的と定め，方策を最適化することを目指す．

強化学習の分かりづらさの一端には状態・行動という入出力関係の直接的な推論問題になっておらず，報酬（しかも将来に渡っての総和）という間接的な情報を基に方策を学習しなければならない点にある．そのため，他の機械学習手法では一般的なグラフィカルモデルで問題を表現できず，推論問題で発展してきた手法も活用できない．

最適性変数

本稿で紹介する論文は，この問題に対して最適性を明示的に示すためにバイナリ型の最適性変数を導入した．それと同時に最適性変数の発生確率を現状態と行動を条件とした確率分布で定義するとともに，報酬関数はその確率分布のパラメータと解釈した．すなわち，報酬が高いほど現状が最適である確率が高く，逆もまた然りということである．この新たな最適性変数の導入により，図 -1 に示すようにグラフィカルモデル上に現状が最適かどうかを示す情報が加わることとなる．

このとき，状態・行動の相互作用で生まれる軌道が将来に渡って常に最適性を満たす，いわば最適軌道である確率を考えると，そこには将来に渡っての報酬の総和が含まれてくる．すなわち，従来の強化学習では最適軌道である確率を最大化しようと軌道

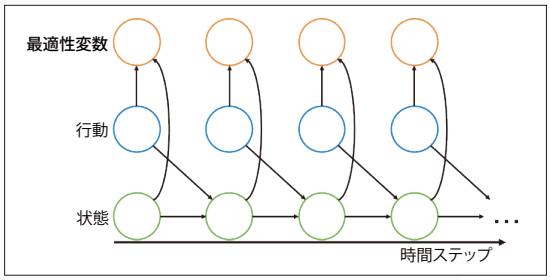

図 -1　最適性変数を加えた強化学習におけるグラフィカルモデル

を方策により修正していたと解釈できる．

このように，紹介する論文は最適性変数を導入したことで新たな確率推論問題としての強化学習を導出していき，従来のものとの共通点や解釈を辿っていくものである．本稿ではページ数にも数式なしでの説明にも限界があるので，概念的に重要となる2章までを紹介する．

推論問題としての最適方策

最適性変数を活用して強化学習の目的を考えてみると，現状態と将来に渡って常に最適であることを条件とした確率的方策の推論問題となる．これは，ベイズの定理とマルコフ性を利用することで，現状態と行動を条件として将来に渡って常に最適である確率とその条件に含まれている行動を周辺化した確率の比で表せる．

また，これらの確率に対して対数を取ったものを従来の強化学習における状態価値関数と行動価値関数☆1と類するものと見なしてみると，従来の価値関数の更新時に登場してきたベルマン方程式に類似した関係式が導出できる．これと従来のベルマン方程式は，環境のダイナミクスが決定論的である場合は完全に一致し，確率的である場合にはイェンゼンの不等式を用いた変分推論を施すことで近似的に一致させられる．特に後者では，状態価値関数と行動価値関数の関係がソフトマックスで与えられ，近年のエントロピー最大化を考慮したSoft Actor-Criticに代表される強化学習手法と一致することが分かる．

2章の最後では，将来に渡って常に最適性を得る最適方策が制御問題としては何を目的として得られるのか解析している．具体的には，前述した最適軌道の確率分布と適当な確率的な方策で得られる軌道の確率分布とのKullback-Leiblerダイバージェンスの最小化問題を考えている．すると，軌道生成に関与する環境のダイナミクスは打ち消されるか最小化問題から除外でき，残るは報酬の総和の期待値と軌道中の各状態で条件付けられた方策のエントロピー項となる．この目的はやはり，すでに価値関数の関係で一致が見られたSoft Actor-Criticに代表されるエントロピー最大化を考慮した強化学習問題と一致する．

☆1 現時刻から将来に渡っての報酬の累積和の期待値．状態価値関数は現時刻の状態のみを条件とし，行動価値関数は状態・行動を条件とする．

将来展望

2章以降では，変分推論を用いた近似計算と，得られた結果と近年提案されている強化学習手法との関係性について述べている．前述しているように最適性変数を導入したことで得られた最適方策の推論問題は，奇しくも結果としてエントロピー最大化項を組み込んだ最新手法と一致した形式で解が与えられた．この事実は次世代の強化学習手法の導出に向けた示唆に富んでいると考えられる．また，制御問題が推論問題と転換する利点として，環境のダイナミクスの学習といった元々推論問題で扱われていたものとの自然な統合が期待できる．報酬関数の設計に関しても，これまではヒューリスティックに与えるだけであったが，最適性変数の確率モデル用パラメータという解釈から，報酬の設計論や整形技術の開発への発展が考えられる．

個人的な見解だが，制御問題としての強化学習を推論問題として捉えるために新たな変数の導入が求められるなど，まだまだアイディアレベルの段階に見える．より適した解釈や洗練された導出を経ることで，今回のような体系化にとどまらず新たな手法の開発につながるのではないかと思う．拙い紹介であるが興味が湧いたなら，ぜひとも自ら精読し，新技術の着想になれば幸いである．

(2020年10月14日受付)

小林 泰介 kobayashi@is.naist.jp

2016年名古屋大学工学研究科博士課程後期課程短縮修了．博士（工学）．同年奈良先端科学技術大学院大学に助教として着任．2018年より約1年間ミュンヘン工科大学にて滞在研究員．機械学習のロボティクス応用に関する研究に従事．