セミパラメトリック統計学の立場からの方策評価法の解析

前田 新一
京都大学大学院情報学研究科　システム科学専攻

強化学習は、動物がより多くの報酬が得られるよう自身の行動を強化していく学習行動を説明するモデルとして心理学の分野で提案されたが、その後、未知の環境において経験を通じて、期待累積報酬を最大化する最適な行動系列を学習するための一般的な枠組み (ここでは行動系列学習問題とよぶ) として定式化され、体系化されてきた。期待累積報酬 R は、各時刻で与えられる報酬 $r_k(k = 1, \cdots, T)$ の時刻に関する和で表現され、報酬は、その時の状態 (あるいは状態遷移) とその時に選択した行動に応じて与えられる。

$$R = E\left[\sum_{k=1}^{T} \gamma^{1-k} r_k(s_k, s_{k+1}, u_k)\right] \tag{1}$$

ここで、γ は割引率とよばれる $0 < \gamma =< 1$ を満たす定数であり、T は終端時刻を表す。T が無限大である問題を infinite horizon と呼ぶ。$E[\cdot]$ は状態と行動の系列に関する期待値演算を表す。s_k,u_k がそれぞれ時刻 k における状態とそこで選択した行動を表す。

本講演では、この行動系列学習問題に対してどのようなアプローチがあるのかの概略について説明した後、そのアプローチの一つである方策評価法が統計の立場からがどのように定式化、解析され、それがどのように役立つかについて話す。ここで、その講演の要旨を論じる。

行動系列学習問題は主に、環境のダイナミクスの種類やダイナミクスに関する知識の有無に応じてとりえるアプローチが異なってくるが、強化学習では、通常、環境のダイナミクスは未知とされ、確率的に選択した行動を通じて得られる状態遷移・報酬に関するサンプルのみから方策を学習していく必要がある。ここで方策とは、各状態でどのような (確率的) 行動をとるかを表す行動則 $\pi(u_t|s_t)$ である。最大化すべき期待累積報酬を、教師信号と解釈すれば教師あり学習とみなすことも可能であり、実際に教師あり学習のアプローチをとる研究も存在する。しかし、通常の教師あり学習が扱う問題と行動系列学習問題とは2点、異なる。行動系列学習問題の「行動」と「系列」の二つのキーワードがその違いを特徴付ける。

一つは、サンプル取得が受動的では無い点である。通常の教師あり学習の問題では、サンプルは受動的に得られるが、行動系列学習問題では各時刻、状態で選択する行動 (つまり方策) は、自分で選択できるという自由度をもつ。また、どの行動系列でどのような累積報酬が得られるか、という行動系列と累積報酬の関係全体を知りたいのではなく、累積報酬が最大となる最適な行動にのみ興味があるという特徴をもつ。

もう一つは、最適化すべき対象が「系列」である点である。とりえる行動系列は、各時刻でとりえる行動の組み合わせになるため、非常に大きな数となり、すべての行動系列を全探索することが不可能となる。また、一連の行動系列が訪れる可能性のある状態数は一般に非常に大きくなるため、たとえ環境のダイナミクスが既知であったとしても動的計画法のような手法もとれない。そこで、現在の方策より、より良い行動を選択できるよう方

策を漸次、改善していく手法がとられる。この方策改善は、方策の勾配を直接、推定して方策を改善する方策勾配法 [16] [2] [7] [9] と、状態価値関数や行動価値関数と呼ばれる方策の評価をもとに方策を改善する価値関数法 [11] [1] [3] [8] [4] とに分けられる。

前者の方策勾配法は、コスト関数に対して方策のパラメータの勾配をとって更新する、という合理的なものである。ただし、コスト関数自身に未知のダイナミクスが含まれるため、勾配の推定はサンプルから推定される必要がある。

後者の価値関数法は、方策評価を行って、その評価に基づき方策を更新する。状態価値関数 $V_\pi^t(s_t)$ は、時刻 t の状態 s_t から始まって方策 π に従って行動をおこなったときに得られる期待累積報酬として以下のように定義される。

$$V_\pi^t(s_t) = E\left[\sum_{k=t}^{T} \gamma^{1-k} r_k | s_t\right] \tag{2}$$

価値関数法では、方策評価の精度が方策の更新の良し悪しに直結するため重要な問題となる。この方策評価の問題は、方策が固定されているときの方策評価を行うため、サンプルが特定の分布から得られることとなり、統計推定の問題として定式化しやすい。ただし、こちらも方策勾配法と同様、環境のダイナミクスを未知としたまま推定することになるので、最尤推定などのパラメトリックな分布推定とは異なり、セミパラメトリック統計における推定問題として捉える必要がある。

価値関数は、状態遷移のダイナミクスのマルコフ性が仮定される時、ベルマン方程式と呼ばれる自己無撞着方程式を満たす。終端状態や終端時刻が定義されない infinite horizon の問題で定常分布が存在するとき、価値関数は時刻に依存しない状態の関数で表され、ベルマン方程式は、現在の状態の価値関数と、現在の状態から行動を選択して遷移した先の次状態の価値関数との間で成り立つべき方程式となる。

$$V_\pi(s_t) = r_t + \gamma E\left[V_\pi(s_{t+1}) | s_t\right] \tag{3}$$

ここで、$V_\pi(s_t) = E\left[\sum_{k=1}^{\infty} \gamma^{1-k} r_k | s_t\right]$ である。

このベルマン方程式の利用は、サンプルから推定する際の推定量の分散を減らす利点をもつ。すなわち、式 (1) のように状態・行動系列という長時間の系列に関して期待値を取る必要のあった期待累積報酬が、式 (3) のように 1 時刻間という短時間の状態遷移で定義づけられるベルマン方程式の解として得られる。ただし、式 (3) は任意の状態 s_t で成り立つ必要があることに注意する。これまで、ベルマン方程式を成り立たせる価値関数を求めるために、数多くの推定アルゴリズムが提案されてきた [11] [1] [3] [8] [4]。いずれもベルマン方程式が成り立つように推定量を少しずつ更新する方法で、サンプル平均によって推定されたベルマン方程式の残差である TD 誤差を最小化する形となっている。

これまでこれらの推定アルゴリズムの相対的な良さは、シミュレーションによって主に確かめられてきた。理論的には、一致性に関しては確かめられていても、収束速度の解析や推定分散の大きさに関しては一部のアルゴリズム [13] [5] を除いて評価されておらず、どのアルゴリズムが良いかについての理論的な議論は行われなかった。

統計の立場からは、このベルマン方程式はセミパラメトリック推定において M 推定量を求めるための推定方程式の一種と考えることができるが、このような M 推定量一般に対して漸近解析を行うことで、既存の TD 最小化のアルゴリズムが対象としていた M 推定量を含んだ一般的な推定量を求めることができる。この一般的な推定量の漸近二乗誤

差を評価することでどのような推定量を用いるべきかの議論を統一的に行うことができる [15] [14]。

上記では、強化学習の主流といえるアプローチについて概観した。一方、近年、確率最適制御の研究に進展があった。Linear-Quadratic Regulator(LQR) と呼ばれるダイナミクスを表す微分方程式が行動に対して線形関数、報酬が行動に対して二次関数である問題において、リカッチ方程式を解くことで最適な方策が得られることが知られていたが、ダイナミクスを確率微分方程式に一般化した際にも特定の条件を付与することで、ファインマン-カッツの経路積分の定式化に持ち込むことができ、最適な方策が求められる、といったことが新たに示された [6] [17]。これらの研究では、環境のダイナミクスを既知として推定に利用する必要があったが、この研究に触発されて環境のダイナミクスを未知とした場合の推定アルゴリズムが新しい強化学習のアルゴリズムとして提案されている [12]。これは、従来の強化学習アルゴリズムに比べて高次元の状態をもつ問題を解くことに成功しており、注目を集めている [12] [10]。しかし、もともとの確率最適制御の研究 [6] [17] とは大きく異なるものとなっており、このアルゴリズムはもはや最適な方策を求めるものではなく、方策を改善する方策改善法の形をとる。これらの研究の位置付けについても議論する。

参考文献

[1] L. Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Machine Learning*, pp. 30–37, 1995.

[2] J. Baxter and P. L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, Vol. 15, No. 4, pp. 319–350, 2001.

[3] S. J. Bradtke and A. G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, Vol. 22, No. 1, pp. 33–57, 1996.

[4] A. Geramifard, M. Bowling, M. Zinkevich, and R. S. Sutton. iLSTD: Eligibility traces and convergence analysis. In *Advances in Neural Information Processing Systems 19*, pp. 441–448, 2007.

[5] S. Grunëwälder and K. Obermayer. Optimality of LSTD and its relation to TD and MC. Technical report, Berlin University of Technology, 2006.

[6] H. J. Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics: Theory and Experiment*, p. P11011, 2005.

[7] V. R. Konda and J. N. Tsitsiklis. On actor-critic algorithms. *SIAM Journal on Control and Optimization*, Vol. 42, No. 4, pp. 1143–1166, 2003.

[8] A. Nedić and D. P. Bertsekas. Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dynamic Systems*, Vol. 13, No. 1, pp. 79–110, 2003.

[9] J. Peters and S. Schaal. Natural actor-critic. *Neurocomputing*, Vol. 71, No. 7-9, pp. 1180–1190, 2008.

[10] Norikazu Sugimoto and Jun Morimoto. Phase-dependent trajectory optimization for cpg-based biped walking using path integral reinforcement learning. In *EEE/RAS International Conference on Humanoid Robots (Humanoids2011)*, 2011.

[11] R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, Vol. 3, No. 1, pp. 9–44, 1988.

[12] E. Theodorou, J. Buchli, and S. Schaal. A generalized path integral control approach to reinforcement learning. *Journal of Machine Learning Research*, Vol. 11, pp. 3137–3181, 2010.

[13] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, Vol. 42, No. 5, pp. 674–690, 1997.

[14] T. Ueno, S. Maeda, and S. Ishii. Asymptotic analysis of value prediction by well-specified and misspecified models. *Neural Networks*, Vol. 31, pp. 88–92, 2012.

[15] T. Ueno, S. Maeda, M. Kawanabe, and S. Ishii. Generalized TD learning. *Journal of Machine Learning Research*, Vol. 12, pp. 1977–2020, 2011.

[16] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, Vol. 8, No. 3, pp. 229–256, 1992.

[17] 佐藤訓志, H. J. Kappen, 佐伯正美. 軌道積分に基づく非線形確率最適制御の解法. 第12回計測自動制御学会制御部門大会予稿集, p. P0194, 2012.