2A1-4

報酬の遅れが大きい状況下での強化学習の学習速度の検討

Investigating the Learning Speed of Reinforcement Learning with a Long Delayed Reward

伊丹英樹 Hideki Itami　板舛尚樹 Naoki Itamasu　岡夏樹 Natsuki Oka

京都工芸繊維大学 大学院工芸科学研究科
Graduate School of Science and Technology, Kyoto Institute of Technology

Abstract: It is known that reinforcement learning with a long delayed reward has a problem of slow learning speed. We propose a learning algorithm, FABL, which is based on a heuristics. The heuristics tells that the final action taken in a state is a right action in the state with high probability. We experimentally demonstrate that FABL is much faster than Q-learning in maze learning tasks, although the shortest path is not always found by FABL.

1. はじめに

近年, お手伝いロボットや介護ロボットに代表される, 我々の日常生活に密接に関わるようなロボットへの期待が高まってきている. 特に, 与えられた環境に自らを適応させ, 柔軟な判断を行う能力の実現には大きな関心が寄せられている.

日常生活環境で動作するロボットの実現には様々なアプローチが考えられる. 例えば, あらかじめ想定される全ての状況に対応した行動や, 応用範囲が広い行動を用意しておくことが考えられる. しかし, 我々の生活の中で想定される状況は無数にあり, その全てに対応することは難しい. そこで, 環境の学習による適応が必要となる. 状況に対しての行動をあらかじめ用意するのではなく, 環境を学習し, ロボット自らが行動を選択することで, 柔軟な判断を行う能力を実現する. この方法であれば, 環境の変化に対しても適応することができる.

ロボットが教師なしで試行錯誤的に学習する手法としては強化学習[Sutton 98]が知られているが, 報酬の遅れが大きい状況下では, 学習速度の点で実用規模の問題に適用できないという課題がある.

本研究では, 報酬が得られるまでの時間の違いに注目して 1 つの episode に現れたすべての状態における行動価値を一気に更新する方法により, 与えられた環境を速やかに学習するアルゴリズムを提案し, その性能をシミュレーション実験により評価する.

2. 最終行動ヒューリスティクス

本研究では, 試行錯誤の中で最終的にとった行動に着目した学習アルゴリズムを提案するが, 本節では, そこで中心的な働きをするヒューリスティクス[伊丹 07]について述べる.

最終行動ヒューリスティクス(Final Action Heuristics, 以降, FAH と略記する):試行錯誤を繰り返して最終的にゴールへ到達した試行において, ある特定の状態に対する行動は複数回試みられた可能性があるが, その中の各状態で最終的にとった行動は正しかった可能性が高い. また, ある状態で最終的にとった行動とは異なる行動は誤っていた可能性が高い.

連絡先:伊丹英樹, 京都工芸繊維大学 大学院工芸科学研究科 情報工学専攻, 〒606-8585 京都市左京区松ヶ崎橋上町, TEL075-724-7125, m7622003@edu. kit. ac. jp
板舛尚樹, 同上, m7622002@edu. kit. ac. jp
岡夏樹, 同上, nat@kit. ac. jp

図 1 に例を示す. ある状態 A において行動 α を実行し, 状態 B になったとする. その後, 状態遷移によって再び状態 A となったときに行動 β を実行し, 状態 C になったとする. その後の状態遷移では再び状態 A は現れず, ゴールに到達したとする. この場合, FAH に基づくと,「行動 β が状態 A での正しい行動である可能性が高い」,「行動 α は状態 A での誤った行動である可能性が高い」と判断できる.

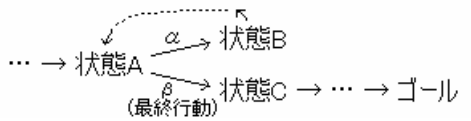

図 1 エージェントの状態遷移と最終行動

FAH は文献[伊丹 07]で実験に用いた迷路の探索タスクにおいて, 平均 90%程度の割合で成立することがわかっており, このヒューリスティクスに基づいて迷路探索タスクにおける教示の意味学習が可能であることが示された[伊丹 07]が, 行動学習においても有効であると予想される.

本論文では迷路探索タスクを例題として取り上げるが, この FAH は,「オペレータによる状態空間中の状態遷移」として定式化できる問題に対して広く有効であると考えられる.

3. 提案アルゴリズム

本節では, FAH を基にした学習アルゴリズム(Final Action Based Learning, 以降, FABL)について述べる.

FABL は, 強化学習の一つである Q 学習等と同様に, 特定の状態に対する行動の価値を示す Q 値の更新によって学習を行うアルゴリズムである.

Q 学習等と同様, FABL においても, エージェントは状態 s で可能な行動 a_n の Q 値, $Q(s,a_n)$の値を基に行動を選択する. 本研究では, 行動学習の際の行動選択法として、ε -グリーディ法を用いた. グリーディ法では現在の状態において最も Q 値の高い行動をとり, Q 値が最適な値に収束していれば最適な行動を生成する. ε -グリーディ法は $1-\varepsilon$ の確率でグリーディな行動選択をし, 確率 ε でランダムに行動を選択する手法である.

Q 値の更新は, 状態 s でのエージェントの最後の行動が行動 a であったとき, 以下のように更新する.

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r - Q(s,a)] \quad (1)$$

また, 最後の行動以外の行動 $\bar{a}$ に関しては, 以下のように更新する.

$$Q(s,\bar{a}) \leftarrow Q(s,\bar{a}) - \alpha Q(s,\bar{a}) \quad (2)$$

ここで，α は学習率であり，0<α<1 を満たす値である．r はエージェントが環境から得る報酬である．エージェントが目標状態 s_{goal} に到達したときに得られる報酬を r(>0)とし，その他の状態で得られる報酬を 0 とする．

Singh らは，replacing eligibility trace を提案し，学習速度が改善されることを示した[Singh 96]．replacing eligibility trace は，ある状態の再訪問による trace の増加を抑えることにより，複数回繰返された望ましくない行動の価値の上昇を防ぐ．この考え方は，我々が提案する FABL の設計思想と共通するものである．

4. アルゴリズムの性能評価

本節では，FABL を，強化学習の一種である Q 学習[Watkins 92]およびモンテカルロ法[Michie 68]と比較することでその性能を評価する．

4.1 比較するアルゴリズムの概要

本研究で比較対象とした Q 学習およびモンテカルロ法のアルゴリズムについて，その概要を示す．

(1) Q 学習

状態 s_t において，エージェントが行動 a を選択し，状態 s_{t+1} に遷移したときの Q 学習における Q 値の更新式を以下に示す．

$$Q(s_t,a) \leftarrow Q(s_t,a) + \alpha[r_{t+1} + \gamma \max_p Q(s_{t+1},p) - Q(s_t,a)] \quad (3)$$

ここで，γ は割引率であり，0 以上 1 以下の定数である．r_{t+1} はエージェントが状態 s_{t+1} に遷移したときに得た報酬である．行動学習の際の行動選択は，ε-グリーディ法で行う．

Q 学習では，学習率 α が以下の条件を満たすとき，与えられた環境における最適解が得られることが証明されている．

$$\sum_{t=0}^{\infty} \alpha(t) \to \infty \quad (4)$$

$$\sum_{t=0}^{\infty} \alpha(t)^2 < \infty \quad (5)$$

(2) モンテカルロ法

状態 s において，エージェントが行動 a を選択したときのモンテカルロ法（every-visit MC methods）における Q 値の更新式を以下に示す．

$$Q(s,a) \leftarrow Q(s,a) + \alpha[R_t - Q(s,a)] \quad (6)$$

ここで，R_t は以下に示される値である．

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+1+k} \quad (7)$$

これは，エージェントが得られる報酬の総和を未来に得られる分，割り引いたものである．行動学習の際の行動選択は，ε-グリーディ法で行う．

4.2 性能の比較実験

(1) 実験の仕方

FABL の他アルゴリズムとの比較はエージェントによる迷路の探索タスクを用い，その実験は PC 上のシミュレーションプログラムによる行動学習によって行った．

(2) 実験の環境

実験に用いた迷路は図 2，図 3 である．

迷路は格子状に座標が設定されており，スタート地点，ゴール地点および迷路の通路の座標をエージェントの状態とする．迷路上で隣の升目への移動を 1 step とし，実験の開始からエージェントがゴール地点に到達するまでの状態の遷移を 1 episode とする．

エージェントは与えられた迷路のスタート地点から探索を始め，ゴールまでの道筋を学習する．エージェントの学習開始時点での Q 値はいずれの状態-行動に関しても 0 であり，学習の経過と共にその値を更新していく．また，報酬は，ゴールに到達したときに 1 とし，それ以外では 0 とする．

本実験では，迷路探索タスクとしてスタートからゴールまでの経路が複数ある場合と単一である場合の 2 つの環境を用いる．図 2，3 にそれぞれの環境を示す．

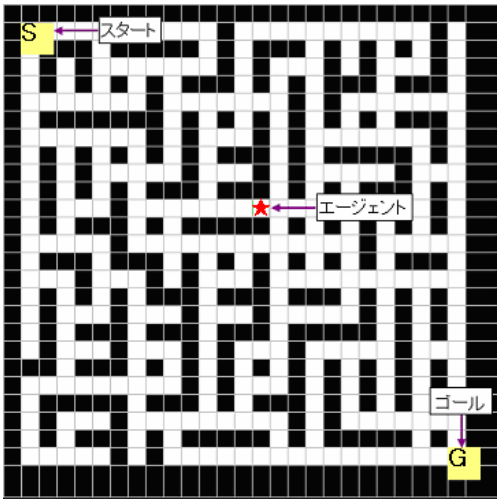

図 2 スタートからゴールまでの経路が複数ある迷路

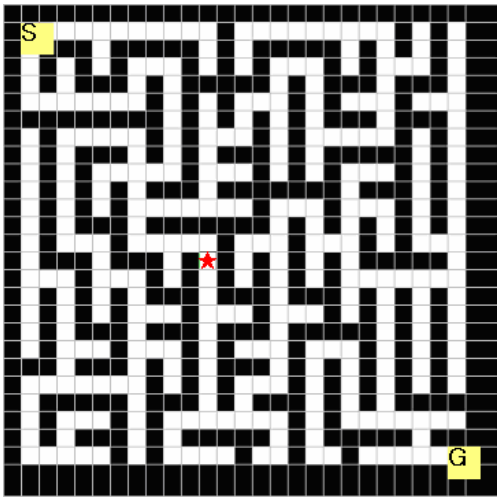

図 3 スタートからゴールまでの経路が単一の迷路

図 2 の迷路の最短経路の step 数は 48 であり，図 3 の迷路の最短経路の step 数は 80 である．

(3) 実験の結果

図 4，5 に図 2，3 に示される環境に対し，Q 学習，モンテカルロ法および FABL を用いて学習を行ったときの step 数と episode 数の関係をそれぞれ示す．図 4 において，グラフ右端の時点での step 数はそれぞれのアルゴリズムで，Q 学習は 48 step，モンテカルロ法は 51.2 step，FABL は 70.8 step となった．図 5 では，全てのアルゴリズムで 80 step に収束した．

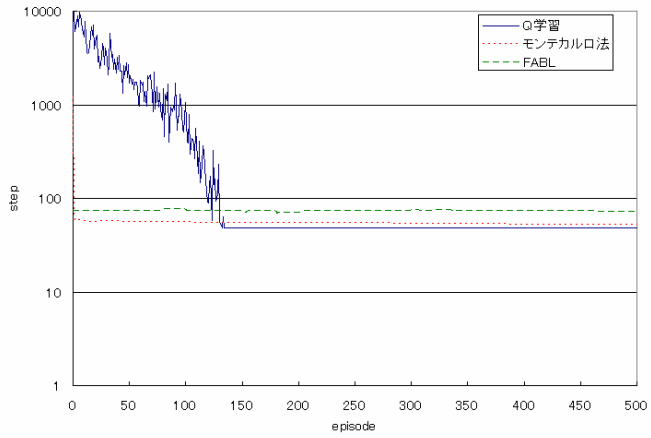

図 4 スタートからゴールまでの経路が複数ある場合の学習速度の比較

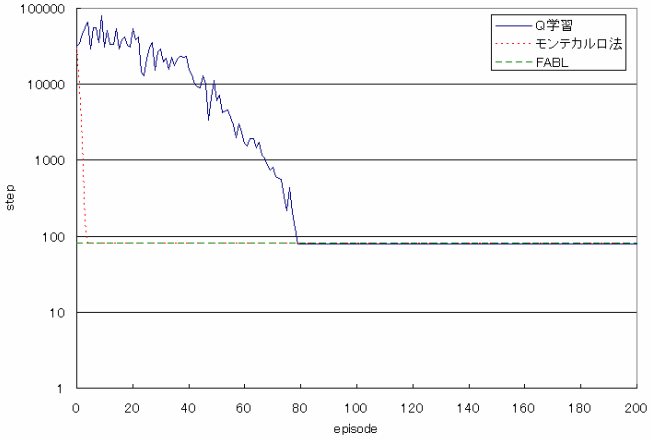

図 5 スタートからゴールまでの経路が単一である場合の学習速度の比較

図 4，5 で示した結果は，1000 episode の学習を 1 試行として，10 試行の平均をプロットしたものである．

学習に用いたパラメータである学習率 α，割引率 γ，および，探査率 ε の値は，試行錯誤によって求めた最適と思われる値を用いた．

5. 考察

5.1 学習速度

図 4，5 に示されるように，学習の収束（正確には、ゴールまでに要するステップ数が大きくは変化しなくなるまで）に要した episode 数は，FABL，モンテカルロ法，Q 学習の順に小さかった．このような結果となった理由について考察する．

Q 学習は，ゴールまでの経路が単一である場合は，学習完了までに，ゴールまでの最短 step 数と同程度の回数の episode が必要となる．これは，1 つ後の状態の行動価値に基づいて学習を進めていく時間差分学習を用いているためであり，与えられた環境が大きくなると学習に多くの時間が必要となる．ゴールまでの経路が複数存在する場合には，さらに多くの時間が必要となる．

モンテカルロ法および FABL では，Q 学習に比べ非常に高速に学習を行っている．これは，ゴールに到達して報酬が得られたときに，エージェントが遷移してきた状態の全ての Q 値が一気に更新されるためである．このような方法で Q 値の更新を行うことによって，少ない episode 数でゴールまでの step 数を大幅に減らすことができる．ただし，5.3 節でも考察するが、モンテカルロ法および FABL では，急速に学習が進むが、ゴールまでの経路が複数ある場合には，急速に最適解が求まっているわけではないことに注意が必要である．

Q 学習は，ゴール地点に近い状態から学習を行っていくという学習の戦略を持つアルゴリズムである．一方，モンテカルロ法および FABL は，ゴールまで到達することのできる経路を見つけ，その経路を中心として他の経路の学習を行っていくという戦略のアルゴリズムであるといえる．このような学習戦略の違いによって，図 4，5 のように学習速度に差が出たと考えられる．

5.2 時間計算量

図 4，5 に示されるグラフ上には表れないが，学習過程における Q 値の更新の時間計算量について考察する．

Q 学習では，1 episode 分の学習のための計算量は，その episode におけるスタートからゴールまでの step 数に比例する．各 step で(3)式にしたがって Q 値を更新するが，(3)式の計算は全体の step 数には依存しない一定の時間で実行できる．

モンテカルロ法においても，1 episode 分の学習のための計算量は，その episode におけるスタートからゴールまでの step 数に比例する．ゴールから遡って(7)式の値をインクリメンタルに計算していくことができ，(6)式の計算は，R_t が求まっていれば，全体の step 数には依存しない一定の時間で実行できるからである．

FABL においても，1 episode 分の学習のための計算量は，その episode におけるスタートからゴールまでの step 数に比例する．すべての状態における最終行動が何であったかは全ステップ数に比例した手間で求めることができ，(1)式や(2)式の計算は，全体の step 数には依存しない一定の時間で実行できるからである．

5.3 解の最適性の保証

(1) ゴールまでの経路が複数ある場合

ゴールまでの経路が複数ある場合にも Q 学習では最短経路を求めることができた．モンテカルロ法では，最短経路が必ずしも求められなかった．モンテカルロ法は式(6)，(7)に示されるように Q 値の更新においてゴールまでの step 数を考慮したアルゴリズムであり，十分な時間をかければ最適解へ収束するものと思われるが，図4に示す実験の範囲内では，まだ最短経路の長さにまではステップ数が下がりきっていない．不適切な行動がたまたま繰返されたことにより，Q 値が高くなってしまうという現象が生じた可能性がある．

FABL はモンテカルロ法とは異なり, 式(1), (2)に示されるように Q 値の更新に際して, エージェントがゴールするまでに必要とした step 数を考慮していない. そのことによって, 経路にループが生じなかった場合においては, スタートからゴールまでの経路の長さの違いによる Q 値の更新の値に差が出ず, このために FABL では最適解への収束がさらに遅くなったものと考えられる.

ただし, モンテカルロ法および FABL では, 数 episode の試行だけで最適解に近い経路を得ることはできており, 厳密に最短の経路の導出が必要とされない場合には, これらの方法が向いていると考えられる.

(2) ゴールまでの経路が単一の場合

ゴールまでの経路が単一の場合は, モンテカルロ法は数 episode で最短経路を学習することができる. FABL では, 最初の 1 episode のみで最短経路の学習を行うことができる. これは, ゴールまでの経路が単一である場合, 最初に見つかった経路が最短経路と一致するためである.

6. おわりに

本論文では, FAH を基にした FABL を提案し, その学習速度を Q 学習およびモンテカルロ法と比較した. その結果, FABL は厳密な最適解を必ずしも求めることはできないが, 1 episode の学習だけで, 最適解に近い経路の学習を行えることを示した. ゴールまでの経路が単一である場合には, 1 episode の学習で最短経路を求めることができる.

FABL では厳密な最短経路を学習することができるとは限らないが, これはエージェントがゴールに到達するまでに遷移した経路について以下のいずれかを考慮したアルゴリズムにすることで改善がなされると推測される.

- エージェントがゴールに到達するまでに遷移した状態のそれぞれのゴールからの距離
- エージェントがゴールに到達するまでに必要とした step 数

今後の課題としては, 以上を踏まえ, 最適解に近い解が高速に求まるだけでなく, 最適解も高速に求めることができるように, アルゴリズムを改良することが挙げられる. また, 本論文では, 特定の迷路タスクにおけるアルゴリズムの性能を評価しただけであり, 今後, 提案アルゴリズムが, どのような性質を持つ問題に対して, どのような性能を示すかを明らかにしていく必要がある.

参考文献

[伊丹 07] 伊丹 英樹, 川上 茂雄, 野川 博司, 岡 夏樹: 最終行動ヒューリスティクスに基づく教示意味の学習, 平成 19 年度情報処理学会関西支部支部大会講演論文集, pp.83-86, 2007.

[Michie 68] Michie, D., and Chambers, R. A.: BOXES: An Experiment in Adaptive Control, in E. Dale and D. Michie (eds.), *Machine Intelligence 2*, pp. 137-152, Oliver and Boyd, Edinburgh, 1968.

[Singh 96] Singh, S. P., and Sutton, R. S., Reinforcement Learning with Replacing Eligibility Traces, Machine Learning, vol. 22, No. 1-3, pp. 123-158, 1996.

[Sutton 98] Sutton, R. S., and Bart, A. G.: "Reinforcement Learning: An Introduction", MIT Press, 1998. (三上 貞芳, 皆川 雅章 共訳: 強化学習, 森北出版, 2000.)

[Watkins 92] Watkins, C. J. C. H., and Dayan, P.: Q-Learning, *Machine Learning*, Vol. 8, pp. 279-292, 1992.