

加藤 新吾

松尾 啓志†

名古屋工業大学電気情報工学科‡

1 はじめに

強化学習とは、報酬という特別な入力を手掛かりとして環境に適応する教師無し機械学習の一種である。強化学習の目的は、できるだけ多くの報酬をできるだけ早く獲得することである。

多くの強化学習とその理論的な解析は、環境をマルコフ決定過程 (Markov Decision Processes : MDP) としてモデルしており、状態の観測は完全であることを仮定している。

本研究では、著者らが提案した強化学習法 Profit Sharing with Virtual Queue (PSwVQ) [2] の MDP 環境下での性質について考察を行う。

2 Profit Sharing with Virtual Queue

2.1 Profit Sharing

Profit Sharing (PS) は経験強化型の代表的な強化学習法の一つである。報酬に至るエピソードにおける感覚入力 x と行動 a の対からなるルール系列を記憶しておき、報酬が得られた時点で系列上のルールを次式に従って強化する。

$$w(x_i, a_i) \leftarrow w(x_i, a_i) + f(r, i) \quad (1)$$

ここで、 $w(x_i, a_i)$ はエピソード系列上の i 番目のルールの重み、 r は報酬値、 f は強化関数である。

宮崎らにより、明らかに無駄なルールを強化しない合理的政策を獲得できる強化関数の条件が証明されている [1]。また、PS は MDP 環境下での最適性は保証されていない。 p

2.2 Profit Sharing with Virtual Queue

環境に変化が生じた場合、Profit Sharing では新たな合理的政策を獲得するには多くのエピソードが必要となる。著者らは、環境が変化した場合でも、合理的政策を定数エピソードで獲得する強化学習法 Profit Sharing with Virtual Queue (PSwVQ) を提案した [2]。

PSwVQ では、次式にしたがって系列上のルールを強化する。

$$w(x_i, a_i) \leftarrow w(x_i, a_i) \times \tau + f(r, i) \quad (2)$$

ここで、 τ ($0 < \tau < 1$) は忘却率である。

3 MDP 環境下での性質

本節では、まず強化学習法の一つ n-step SARSA について述べ、次に n-step SARSA と PSwVQ との類似性について着目して、MDP 環境下での PSwVQ の性質について考察を行う。

3.1 n-step SARSA

n-step SARSA は、時刻 t で選択したルール $\overline{x_t a_t}$ を n step 経過後の時刻 $t+n$ において、次式により強化する。

$$Q_{t+1}(x_t, a_t) \leftarrow Q_t(x_t, a_t) + \alpha [R_t^{(n)} - Q_t(x_t, a_t)] \quad (3)$$

$$R_t^{(n)} = r_t + \gamma r_{t+1} + \dots + \gamma^{n-1} r_{t+n-1} + \gamma^n Q_t(x_{t+n}, a_{t+n}) \quad (4)$$

ここで、 $Q(x, a)$ はルール $\overline{x a}$ の評価関数、 r_k は時刻 k に得られる報酬、 α ($0 < \alpha < 1$) は学習率、 γ ($0 < \gamma < 1$) は割引率である。

3.2 PSwVQ と n-step SARSA

強化学習が扱う問題において、報酬は一般的に最終的な目標を達成した場合にのみ得られるので、n-step SARSA の $R_t^{(n)}$ 関数は次式により表される。

$$R_t^{(n)} = \begin{cases} \gamma^n Q_t(s_{t+n}, a_{t+n}) & (\text{after } n \text{ step, reward is not obtained}) \\ \gamma^j \cdot r & (\text{after } j (j < n) \text{ step, reward is obtained}) \end{cases}$$

また、 n が一つのエピソードのルール系列全てを保存できる程十分に大きいと仮定すると、 $R_t^{(n)}$ 関数は次式により表される。

$$R_t^{(n)} = \gamma^j \cdot r \quad (5)$$

* A property of PSwVQ in MDP environment

† Shingo Kato, Hiroshi Matsuo

‡ Department of Electrical and Computer Engineering, Nagoya Institute of Technology

この仮定が成立しないと、PSでの学習が成立しなくなるのでPSwVQと比較する際には妥当な仮定であるといえる。

このとき、n-step SARSAは次式によりルールを更新する。

$$Q_{t+1}(x_t, a_t) \leftarrow Q_t(x_t, a_t) + \alpha [\gamma^j \cdot r - Q_t(x_t, a_t)] \quad (6)$$

式(2)と式(6)とを比較すると、

$$\begin{cases} \tau & = & 1 - \alpha \\ f(r, i) & = & \alpha \cdot \gamma^j \cdot r \end{cases} \quad (7)$$

を満たすとき、PSwVQとn-step SARSAは等価になることが分かる。また、PSwVQは合理性を保ったまま式(7)を満たすことができる。

n-step SARSA自体の最適性は証明されていないが、基礎となる1-step SARSA(SARSA(0))はMDP環境下で最適解に収束することが証明されている[3]。つまり、n-step SARSAとSARSA(0)の等価性が示せれば、PSwVQのMDP環境下での最適性を保証できることになる。n-step SARSAとSARSA(0)の等価性は証明されていないが、あるエピソードを無限回繰り返したとき、両手法とも同じ評価関数に収束することから、n-step SARSAがMDP環境下で最適解に収束する可能性は非常に高いと考えられる。

4 実験

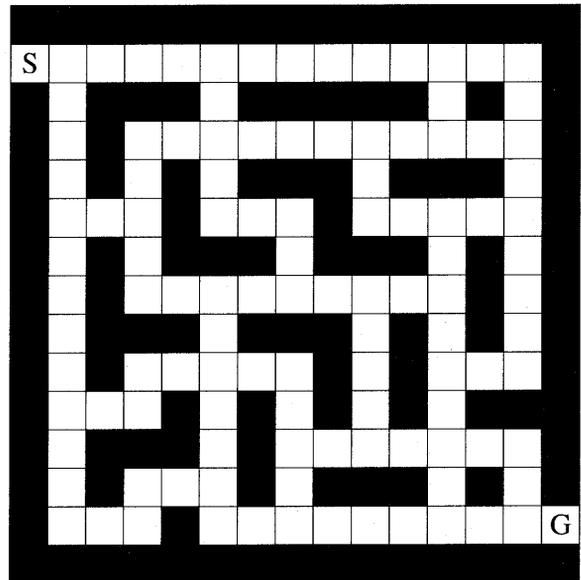
4.1 実験方法

図1に示す15×15の迷路問題を対象に実験を行った。学習器はスタートから出発し、ゴールに到達したら報酬を得る。学習器は上下左右の4方向に1コマ移動することができる。壁があるマスに移動することは許さない。学習器の行動選択として、PSwVQはSARSA(0)の最適性を保証するGLIE政策[3]を利用し、PSではルールレット選択を用いる。

1.0×10⁷エピソードの繰り返しを1試行とし、乱数の種を変えて10試行実験した。

4.2 実験結果および考察

表1に最適政策獲得率を示す。ここで、最適政策獲得率は“最適政策を得た状態数 ÷ 行動選択可能な全状態数 (=111)”を表す。また、ある状態で重みが最大であるルールが最適ルールの時、その状態を最適政策を得た状態と表現する。表1から、PSでは最適政策を習得できない問題に対して、10試行とも全状態で最適政策を習得したことが分かる。



S : Start point G : Goal point

図1: 実験環境

	平均最適政策獲得率
PSwVQ	1.00
PS	0.97

表1: 最適政策獲得率

5 まとめ

本研究では、MDP環境下でのPSwVQの性質について考察し、n-step SARSAとの等価性を示した。また、実験的にPSwVQが最適政策に収束することを確認した。PSwVQがMDP環境下で最適政策に収束することを数学的に証明することが今後の課題である。

参考文献

- [1] 宮崎 和光, 山村 雅幸, 小林 重信: 強化学習における報酬割り当ての理論的考察, 人工知能学会誌, Vol.9, No.4, pp.580-587(1994)
- [2] Singo KATO, Hiroshi MATSUO: "A theory of profit sharing in dynamic environment", 6th Pacific Rim International Conference on Artificial Intelligence (PRICAI2000), Lecture Note in Artificial Intelligence 1886. pp.115-124 (2000)
- [3] Satinder Singh, Tommi Jaakkola, Michael L. Littman, and Csaba Szepesvári. Convergence results for single-step on-policy reinforcement learning algorithms. Machine Learning. To appear. (1998)