F-038

POMDPs環境下での知識利用型強化学習法
The Exploitation Reinforcement Learning Method on POMDPs

植村 渉† 上野 敦志† 辰巳 昭治†
Wataru Uemura Atsushi Ueno Shoji Tatsumi

1. はじめに

強化学習法は，試行錯誤により報酬を獲得し，その報酬情報により行動系列を評価する方法である．Profit Sharing法では，報酬を行動系列に分配し，累積することで評価する．従来の分配方法では，POMDPs環境で適切に累積できない場合があることが知られている．本研究では，適切に累積できない理由が，報酬分配時の分配量の差にあることを明らかにする．また，報酬分配時の分配量に差をつけず，行動系列の長さを考慮した分配方法Episode Profit Sharing(EPS)を提案する．EPSの分配方法が，ループ系列の強化を行わないことを証明し，実験により性能を確認する．

2. Profit Sharing

強化学習のモデルでは，エージェントは時刻tにおいて環境から知覚入力として**状態**s_tを知る．その状態に対してエージェントは実行できる行動群の中から一つを選び**行動**a_tとして出力する．状態s_tで行動a_tを実行することを**ルール**(s_t, a_t)と呼び，ルールを選択する判断基準を**政策**と呼ぶ．行動の実行により，エージェントは次状態s_{t+1}に遷移する．次状態s_{t+1}が目標状態であるとき，エージェントは**報酬**r_tを受け取る．目標状態でないときは，報酬を受け取らない$(r_t = 0)$．報酬を獲得するまでの行動系列を**エピソード**と呼ぶ．

Profit Sharingは，各状態のルールに報酬の一部を分配し，それを累積することで強化を行う．報酬を分配する関数を**強化関数**$f(x)$と呼び，目標状態からさかのぼって分配するため，強化関数の引数はエージェントのルール実行の時系列と逆方向の関係にある．強化作業は，次式に従って報酬rを分配する．

$$\omega(s_x, a_x) \leftarrow \omega(s_x, a_x) + r \times f(x). \quad (1)$$

$\omega(s_x, a_x)$はルール(s_x, a_x)に累積された価値である．

Profit Sharingでは，報酬を獲得しないと学習作業を実行しないため，報酬を獲得しないルールを選択し続けることは，エージェントにとって致命的である．このような状況とは，行動系列内のループであり，ループを構成する行動系列を**ループ系列**と呼ぶ．

ある状態においてループ系列への分岐がある場合，ループ系列へのルールを**迂回ルール**，ループ系列から抜け出るルールを**非迂回ルール**と呼ぶ．このとき，ループ系列の学習を防ぐためには，報酬分配時に迂回ルールの強化量が非迂回ルールの強化量よりも少ない必要がある．このような強化量の大小関係が成り立つ場合を，**ルールの強化の抑制**と呼ぶ．従来提案されている報酬分配方法では，抑制を行う時の強化関数は等比減少の形となる[1]．

†大阪市立大学大学院工学研究科, Osaka City University

3. POMDPs環境

問題環境がマルコフ性を持つが，エージェントの知覚能力に制限があり問題環境を正しく認識できない場合がある．このような不完全知覚問題の問題クラスを**部分観測可能マルコフ決定過程**(POMDPs)と呼ぶ．

Profit Sharingにとって問題が生じる場合は，強化すべきルールと強化してはいけないルールを混同する場合である．このとき，どちらのルールも均一に報酬を分配する必要がある．

定理 1 *POMDPs*環境におけるエピソードに対する報酬分配の必要条件は，

$$f_x = \begin{cases} \alpha_{s_x} & \text{ルール } x \text{ の強化が初めての場合} \\ 0 & \text{それ以外の場合} \end{cases} \quad (2)$$

である．ここで，ルールxはf_xにて強化されるルールである．α_{s_x}はf_xにて強化される状態s_xに分配する強化量であり，状態ごとに一定である必要がある．

4. EPS: 新しい報酬分配方法

本研究では，POMDPs環境における報酬分配の必要条件を満たし，報酬量と行動系列の長さを考慮した報酬分配方法として，次の強化関数を提案する．

$$f_x = \begin{cases} 1/L^W & \text{ルール } x \text{ の強化が初めての場合} \\ 0 & \text{それ以外の場合} \end{cases} \quad (3)$$

である．ここで，Lは状態における行動数，Wは行動系列の長さである．式(3)の報酬分配方法を，エピソードを一括して強化することよりEpisode Profit Sharing(EPS)と呼ぶ．以下，EPSがループ系列を強化しないことを確認する．

4.1 1つの状態でループが構成される場合

ある状態sにおいて，L個のルールが存在する場合を考える．迂回ルールの強化の抑制が最も困難な場合は，$L-1$個の非迂回ルールと，1個の迂回ルールという組み合わせであり，強化量の期待値をΔとすると，

$$\Delta(s, \text{非迂回ルール}) > \Delta(s, \text{迂回ルール}) \quad (4)$$

が成立する(証明は付録A参照)．

4.2 複数の状態でループが構成される場合

複数の状態$s_l (l = 1, 2, \ldots, M)$で構成されるループ系列を考える．ルールの期待値を計算すると，

$$\Delta(s_l, \text{非迂回ルール}) > \Delta(s_l, \text{迂回ルール}) \quad (5)$$

である(証明は付録B参照)．以上により，EPSはどのような状態でもループ系列の強化を抑制することが保証される．

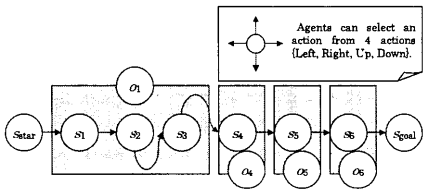

図 1: 実験環境

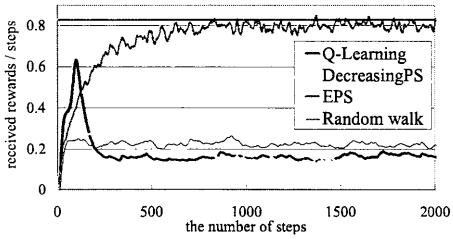

図 2: POMDPs 環境の性能

5	11	14	20	26	31	37		G
4	10		19	25	30	36		45
S	9		18	24	29	35		44
3	8		17	23	28	34	40	43
2	7	13	16	22		33	39	42
1	6	12	15	21	27	32	38	41

図 3: Sutton の迷路

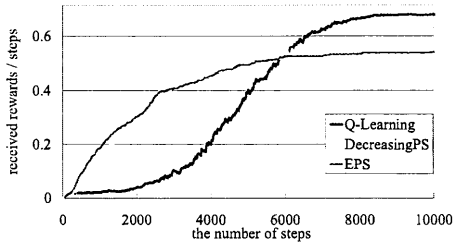

図 4: MDPs 環境の性能

5. 実験とまとめ

EPS の効果を実験にて確認する．全状態のうち不完全知覚が半分生じる場合 (図 1) と，まったく生じない場合 (図 3) の実験を行う．図 1 では，エージェントは状態 s_1，s_2，そして s_3 を同一観測結果 o_1 として知覚する．それぞれ遷移に必要な行動は，右，下，そして上であるため，エージェントは o_1 を観測する時は，左以外の三方向をランダムに選択する必要がある．状態 s_4，s_5，そして s_6 では，観測結果と状態が一致するため，それぞれ遷移する行動を学習する必要がある．評価基準である性能は獲得報酬量を行動数で割ったものであり，実験 1 の性能理想値は $10/12 \simeq 0.833$，実験 2 の性能理想値は $10/14 \simeq 0.714$ となる．結果が図 2 と図 4 である．POMDPs 環境下では，EPS のみが不完全知覚の混同に影響されず，学習を進めていることがわかる．MDPs 環境下では，EPS の学習性能は，従来の等比減少関数を用いた Profit Sharing(DecreasingPS) とほぼ同等の性能であるといえる．

以上，本研究により，EPS が POMDPs 環境でも，MDPs 環境でも，適切に報酬を累積できることが確認できた．

A 1 つの状態でのループ系列の学習の抑制

$L-1$ 個の非迂回ルールと，1 個の迂回ルールという組み合わせが最も抑制が困難である．状態 s 前後の行動選択回数を n_1，n_2 とし，$N = n_1 + n_2$ とする．迂回ルールを選ぶ確率を p として，それぞれのルールの強化量の期待値 Δ を計算する．期待値の差は，

$$\Delta(s, \text{非迂回ルール}) - \Delta(s, \text{迂回ルール})$$
$$= \sum_{i=0}^{\infty} p^i (1-p) \frac{r}{L^{N+i}} \left(\frac{L(1-p)+p}{L(L-1)} \right) > 0 \quad (6)$$

なぜなら，$1 < L, 0 < p < 1$ だからである．よって，EPS の局所的な合理性が証明された．

B 複数の状態でのループ系列の学習の抑制

ある状態 s_l における迂回ルールと非迂回ルールの構成は局所的な合理性と同様に，$L-1$ 個の非迂回ルールと 1 個の迂回ルールの組み合わせの場合が最も抑制が困難である．また迂回ルールを選ぶ確率は，それぞれの状態において偏りがない場合が最も抑制が困難であるため，どの状態でも同じ確率 p とする．期待値の差は，

$$\Delta(s_l, \text{非迂回ルール}) - \Delta(s_l, \text{迂回ルール})$$
$$= \sum_{j=0}^{M-1} \sum_{i=0}^{\infty} (p^M)^i p^j (1-p) \frac{r}{L^{N+Mi+j}} \times \left(\frac{1}{L-1} - \sum_{k=0}^{M-1} \frac{p^{k+1}}{L^{k+1}} \right)$$
$$> \sum_{j=0}^{M-1} \sum_{i=0}^{\infty} (p^M)^i p^j (1-p) \frac{r}{L^{N+Mi+j}} \times \frac{1-p}{L(L-1)(1-\frac{p}{L})} > 0 \quad (7)$$

なぜなら，$1 < L, 0 < p < 1$ だからである．よって，EPS の大局的な合理性が証明された．

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当ての理論的考察，人工知能誌，Vol.9, No.4, pp.580–587 (1994).