重み空間の逐次分割にもとづく多目的一括強化学習法

坪田悠吾　吉田学　平岡和幸　三島健稔
埼玉大学大学院理工学研究科

1 序論

強化学習は，未知環境において試行錯誤的に学習することにより最適な方策を獲得する枠組である．強化学習における学習者（エージェント）には行動に応じて環境から報酬が与えられる．エージェントの目標は，最終的に得られる総報酬の期待値を最大化する最適方策の獲得である．そのための手法のひとつとして，行動価値関数 Q を逐次推定する Q 学習が知られている．

従来の強化学習手法の大部分は単一のタスクに対する学習を行うものであり，異なるタスクに対しては個別に学習しなおす必要があった．しかし現実の問題においては複数の評価基準を考慮したい場合も多い．そこでそれらの荷重和を報酬と設定し，荷重を変えて得られる無限通りのタスクに対する強化学習を一括して取り扱う．これに関して次の性質が知られる [1].

1. 最適行動価値関数 Q^* は荷重の区分線形凸関数
2. 各荷重に対する個別の Q 学習で推定される Q も荷重の区分線形凸関数
3. その Q の区分数は学習につれ単調増加

性質 1, 2 を利用して一括 Q 学習を実現するためには，性質 3 への対策が必要となる．本研究は既存手法 [2] と異なるアプローチでこれを行う．

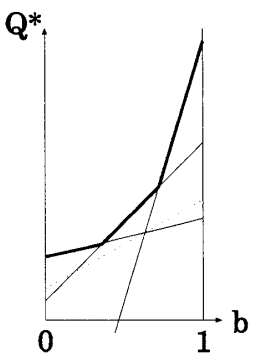

図 1: The optimal expected reward Q^*

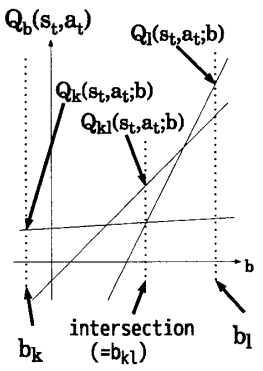

図 2: How to make new elements

Multipurpose Paralell Reinforcement Learning by Adaptive Partitioning of Weight Space
Yugo TSUBOTA , Manabu YOSHIDA ,Kazuyuki HIRAOKA and Taketoshi MISHIMA
Graduate School of Science and Engineering , Saitama University

2 荷重報酬和モデル

状態集合を S，行動集合を A と表す．状態行動対 $(s,a) \in S \times A$ に応じて部分報酬 $r^1,\ldots,r^M$ が与えられる．各部分報酬は観測可能とし，それらの荷重和を報酬 r とする．荷重 $\boldsymbol{\beta}$ を一つ指定するとそれに応じたタスクが一つ定まるから，荷重報酬和モデルは無限通りのタスクを含んでいる．学習ステップ t における行動 a_t に対する報酬は，具体的には

$$r_{t+1}(\boldsymbol{\beta}) = \sum_{i=1}^{M} \beta_i r_{t+1}^i = \boldsymbol{\beta} \cdot \boldsymbol{r}_{t+1} \tag{1}$$

$$\boldsymbol{\beta} \equiv (\beta_1,\ldots\beta_M) \in \mathbb{R}^M \tag{2}$$

$$\boldsymbol{r}_{t+1} \equiv \left(r_{t+1}^1,\ldots r_{t+1}^M\right) \in \mathbb{R}^M \tag{3}$$

で表される．本研究においては $\boldsymbol{\beta} = (b,1)$ の場合を扱う．その場合 Q^* は b の区分一次関数となる（図 1）．この性質をふまえて次節の手法を提案する．

3 逐次分割にもとづく一括強化学習法

3.1 学習点における Q 値の更新

状態行動対 (s,a) ごとに，複数の学習点

$$b_1(s,a),\ldots,b_{n(s,a)}(s,a) \tag{4}$$

を設定する．それらの初期値は $b_1(s,a) = b_{\min}$, $b_2(s,a) = b_{\max}$, $n(s,a) = 2$ とする ($b_{\min}, b_{\max}$ はあらかじめ決めた定数)．各 $b_k(s,a)$ 付近における Q^* の推定値を

$$Q_k(s,a;b) = bQ_k^1(s,a) + Q_k^2(s,a) \tag{5}$$

と表す（図 2）．係数 $Q_k^1(s,a)$ および切片 $Q_k^2(s,a)$ は次式にしたがって更新される $(i=1,2)$.

$$\begin{aligned} Q_k^i(s_t,a_t) \longleftarrow\ & (1-\alpha)Q_k^i(s_t,a_t) \\ & + \alpha\{r_{t+1}^i + \gamma \max_{a\in A} Q^i(s_{t+1},a;b_k(s_t,a_t))\} \end{aligned} \tag{6}$$

ここに

$$\begin{aligned} Q^i(s,a;b) &= Q_{k(s,a;b)}^i(s,a) \\ k(s,a;b) &= \mathrm{argmax}_k Q_k(s,a;b) \end{aligned} \tag{7}$$

$0<\alpha<1$ は学習係数，$0<\gamma<1$ は割引率である．ただし，argmax の対象は更新を一定回数終えた学習点に限定する．そのような学習点を「大人」と呼び，他は「子供」と呼ぶ．初期学習点 $b_{\min}, b_{\max}$ は特例として最初から大人とみなす．

以上により Q^* の推定値

$$Q(s,a;b) = bQ^1(s,a;b) + Q^2(s,a;b) = Q_{k(s,a;b)}(s,a;b) \tag{8}$$

を得る．

3.2 学習点の追加と削除

(s_t, a_t) に対する「大人」$b_k(s_t, a_t), b_l(s_t, a_t)$ の各ペアについて，$Q_k(s_t, a_t; b_{kl}) = Q_l(s_t, a_t; b_{kl})$ となる b_{kl} を求め，新たな学習点として追加する．この時点の b_{kl} は「子供」である．

一方，以下のいずれかに当てはまる「大人」$b_k(s_t, a_t) \neq b_{\min}, b_{\max}$ は削除する．

- $Q_k(s_t, a_t; b_k(s_t, a_t)) < Q(s_t, a_t; b_k(s_t, a_t))$
- 同じ b に別の「大人」が存在する．
- 同じ切片，傾きをとる他の「大人」が存在する．

4 挙動の検証

検証は図 **3** に示すタスク [1] により行った．状態は S, A, B, X, Y, G の 6 通り，行動は UP, DOWN, LEFT, RIGHT の 4 通りであり，状態は，行動にしたがって図で示される隣り合った枡目に遷移する．また，枡目の無い方向に行動した場合は，状態は変化しないものとする．ただし，特例として状態 G においてはどの行動でも状態 S に遷移するものとする．各状態で得られる報酬は，括弧内に示されている．G 以外の状態から隣接する枡目が存在しない方向に行動した場合には，-1 の報酬がさらに与えられる．このタスクの $Q^*(s, a)$ は，b に関して 5 ないし 6 個の区分数をもつ区分一次関数となる．

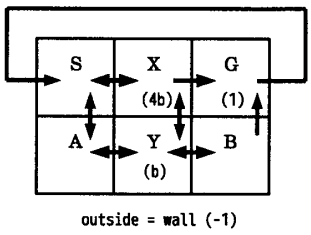

図 3: Task for experiments

学習係数は $\alpha = 0.7$，割引率は $\gamma = 0.8$，大人になるまでに必要な更新回数は $T = 20000$，初期学習点は，$b_{\min} = -2.0$，$b_{\max} = 2.0$ と設定した．また、報酬 r_2 に -0.3〜0.3 の範囲の一様乱数をノイズとして付加した．実験の結果得られた行動価値関数の例を図 **4** に示す．学習点の個数は，Q^* の区分数の 2 倍以下におさえられている（図 **5**）．

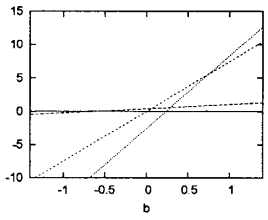

図 4: Q(S,DOWN;b) at $t = 3 \times 10^6$

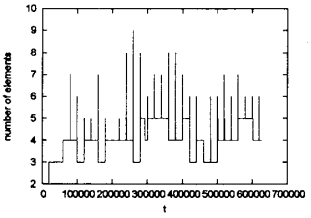

図 5: n(S, DOWN)

Q から求められた方策を**表 1** に示す．方策が切り替わる b の値を Q^* のそれと比較したところ，誤差は平均 0.020，最大で 0.068 であった．

5 結論

本研究では，荷重報酬和モデルの一括強化学習問題に対して，重み空間の逐次分割にもとづくアプローチを提案した．また、基礎的な例題についてその妥当性を確認した．

学習点の追加と削除には，現段階では素朴な方法しか用いていない．今後は，これらを工夫することにより学習速度・精度の改善を試みる予定である．また，[2] との比較も含めたさらなる実験的検証も課題として残されている．

参考文献

[1] 平岡和幸，三島健稔（2006）：荷重報酬和モデルで表されるタスク族に対する一括強化学習法，日本神経回路学会誌 Vol. 13, No. 4, pp. 137–145.

[2] K Hiraoka, M Yoshida, and T Mishima (2007) : Parallel Reinforcement Learning for Weighted Multi-Criteria Model with Adaptive Margin, Proc. ICONIP 2007 , to appear.

表 1: 獲得された方策（括弧内は真の最適方策）

状態	行動			
	UP	LEFT	DOWN	RIGHT
S			$b \leq -0.089107$ ($b \leq -0.091563$)	$-0.089107 < b$ ($-0.091563 < b$)
A	$b \leq -0.707686$ ($b \leq -0.640000$)			$-0.707686 < b$ ($-0.640000 < b$)
B	$b \leq 0.323146$ ($b \leq 0.319149$)	$0.323146 < b$ ($0.319149 < b$)		
X	$0.739265 < b$ ($0.750000 < b$)	$0.337511 < b \leq 0.739265$ ($0.319149 < b \leq 0.750000$)	$b \leq 0.337511$ ($b \leq 0.319149$)	
Y	$0.011326 < b$ ($0.000000 < b$)			$b \leq 0.011326$ ($b \leq 0.000000$)