# 学習段階に応じた閾値による多目的一括強化学習法の改良

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，確率的なゆらぎを含んだ未知環境において，最適な行動則を自動的に獲得する枠組を与える．現実の問題では，複数の指標を総合的に最適化したいという多目的問題が多く，指標の荷重和をとる便法がしばしば用いられる．この荷重を変えると無限通りの学習タスクが得られるが，各タスクの最適価値関数は荷重に関して線形にはならない．また，各指標の最適価値関数を求めてその荷重和をとった結果は，全体の最適価値関数とは一般に一致しない．

この形のタスク族に対し，それぞれの最適価値関数を一括して区間推定する手法が提案された．推定区間の幅を狭くすることは，保持するベクトルの本数を減らすこととトレードオフの関係にある．本研究ではこのトレードオフを緩和し，両者がより両立できるように手法を改良する．

## 2. 強化学習と荷重報酬和モデル

時刻 $t = 0, 1, 2, \ldots$ ごとに学習エージェントは状態 $s_t \in S$ を観察して行動 $a_t \in A$ を選択し，$(s_t, a_t)$ に基いて報酬 $r_{t+1}$ と次状態 $s_{t+1}$ が確率的に決定される．状態集合 $S$ と行動集合 $A$ は共に有限とする．学習の目標は，将来にわたる期待総報酬

$$R_{t+1} = E\left[\sum_{k=0}^{\infty} \gamma^k r_{t+1+k}\right] \tag{1}$$

を最大化する方策 $\pi : S \to A$ の獲得である．割引率 $\gamma$ は $0 \le \gamma \le 1$ の範囲であらかじめ指定される．代表的な強化学習法として，最適行動価値関数 $Q^*$ の推定値 $Q$ を逐次更新する Q 学習がよく知られる．

本研究では，報酬は部分報酬の荷重和で与えられるとする．

$$r_{t+1}(\boldsymbol{\beta}) = \boldsymbol{\beta} \cdot \boldsymbol{r}_{t+1} \tag{2}$$

$\boldsymbol{r}_{t+1}$ は $M$ 個の部分報酬を並べたベクトルであり，$\boldsymbol{\beta}$ は対応する $M$ 個の荷重を並べたベクトルである．$\boldsymbol{\beta}$ を固定すれば学習タスクが一つ定まるので，その $Q^*, Q$ を各々 $Q^*_{\boldsymbol{\beta}}, Q_{\boldsymbol{\beta}}$ で表す．

## 3. 一括強化学習法と区間推定

$Q^*_{\boldsymbol{\beta}}, Q_{\boldsymbol{\beta}}$ が $\boldsymbol{\beta}$ に関して区分線形かつ凸になることを利用し，全 $\boldsymbol{\beta}$ に対する $Q_{\boldsymbol{\beta}}$ を一括して更新する学習法が提案された．この手法では，$M$ 次元ベクトルの有限集合 $\Omega$ を用いて

$$Q_{\boldsymbol{\beta}}(s, a) = \max_{\boldsymbol{q} \in \Omega(s,a)} \boldsymbol{q} \cdot \boldsymbol{\beta} \tag{3}$$

の形で $Q_{\boldsymbol{\beta}}$ を表現し，学習係数 $\alpha > 0$ に対し

$$\Omega^{\mathrm{new}}(s_t, a_t) = (1-\alpha)\Omega(s_t, a_t) \dot{\boxplus} \alpha\left(\boldsymbol{r}_{t+1} + \gamma \bigsqcup_{a \in A} \Omega(s_{t+1}, a)\right) \tag{4}$$

により $\Omega$ を更新する．ここで $X \sqcup Y$ は併合 $X \cup Y$ から，また $X \dot{\boxplus} Y$ は Minkowski 和 $X \oplus Y = \{x + y \mid x \in X, y \in Y\}$ から，それぞれ凸包をとった結果の頂点集合である．

頂点集合 $\Omega^{\mathrm{new}}$ で表される推定行動価値関数 $Q_{\boldsymbol{\beta}}$ は，各 $\boldsymbol{\beta}$ に対して個別に Q 学習を行った結果と等しい．ただし，Minkowski 和の性質から $\Omega$ の要素数は更新につれ単調に増加する．この難点は手法 (3)(4) に起因するものではなく，2 節で定義した $Q_{\boldsymbol{\beta}}$ 自身の持つ性質である．

この難点を回避するため，次のように $Q$ を区間で推定する．

- 内側近似
  図 1 に示した $\triangle L_1 L_2 L_3$ の面積が $\zeta^L/2$ 以下のときに頂点 $L_2$ を削除する．得られる頂点集合 $\Omega^L$ は本来よりも内側になるので，この近似は $Q_{\boldsymbol{\beta}}$ の下限 $Q^L_{\boldsymbol{\beta}}$ を与える．
- 外側近似
  図 2 に示した $\triangle U_2 U_{cp} U_3$ の面積が $\zeta^U/2$ 以下のときに頂点 $U_2, U_3$ を削除し頂点 $U_{cp}$ を追加する．得られる頂点集合 $\Omega^U$ は本来よりも外側になるので，この近似は $Q_{\boldsymbol{\beta}}$ の上限 $Q^U_{\boldsymbol{\beta}}$ を与える．

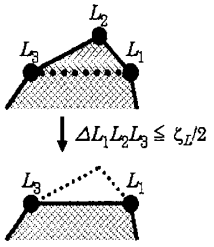


図 1: 内側の近似 ($M = 2$)

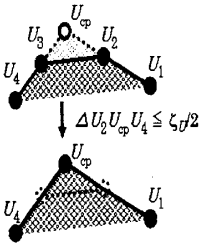


図 2: 外側の近似 ($M = 2$)

$\Omega$ の代わりに $\Omega^L, \Omega^U$ を保持しそれぞれを (4) 式で更新することによって，要素数が抑制され，かつ $Q^L_{\boldsymbol{\beta}} \le Q \le Q^U_{\boldsymbol{\beta}}$ が保証される．特に，区間幅 $Q^U_{\boldsymbol{\beta}} - Q^L_{\boldsymbol{\beta}}$ が許容誤差内に収まれば近似の影響は無視してよい．

## 4. 学習段階に応じた閾値による多目的強化学習法

要素数の抑制と区間幅の縮小はトレードオフの関係にある．一般的に学習の初期段階では要素数が，最終段階では区間幅が問題となる．そこで，学習が進むにつれてマージン $\zeta^X$ を調節することが考えられる ($X$ は $L$ または $U$)．[1] においては，適応的に $\zeta^X$ を調節することで前述の問題解決を図っている．しかし [1] の手法では，パラメータ設定を誤ると $\zeta^X$ が過度に大きくなることや収束の遅さが課題として挙げられ

---

**Improvement of Parallel Reinforcement Learning for Weighted Multi-Criteria Model with Decreasing Margin**

Manabu YOSHIDA†,Kazuyuki HIRAOKA† and Taketoshi MISHIMA†

†Graduate School of Science and Engineering, Saitama University 338-8570, Saitama City, Japan

yoshida@me.ics.saitama-u.ac.jp

た．そこで，本論文においては予め定めた関数に従って $\zeta^X$ を調節する手法を提案する．これにより，$\zeta^X$ の発散による学習の破綻は回避できる．

[1] において妥当な結果が得られた際の $\zeta^L, \zeta^U$ は共に，学習初期段階では急速に，その後さらに学習が進むと緩やかに減少していた．これを踏まえて，本論文では次式の関数 $\zeta^X(t)$ を用いる．

$$\tilde{\zeta}^X(t) = \zeta^X_{\text{init}} \cdot \exp\left(-(Ct)^{\frac{1}{n}}\right)$$
$$\zeta^X(t) = \max(\tilde{\zeta}^X(t), \zeta^X_{min})$$

$\zeta^X_{\text{init}}$ は $\zeta^X$ の初期値，$\zeta^X_{\text{min}}$ は $\zeta^X$ の最小値，$t$ は学習ステップ数，$C, n$ は関数の減少具合を決めるパラメータである．

## 5. 数値誤差

図 2 において，もし $U_1 \sim U_4$ がほぼ同一直線上に並ぶと，数値誤差により $U_{cp}$ が不正な値となる場合があり，$\Omega^U$ の作成が正しく行われない．本論文においては簡便な対策として小さい固定マージン $\zeta^L_{\text{error}}$ を設け，$\Omega^U$ に対しても内側近似を施し，このような状況を回避する．本論文では $\zeta^L_{\text{error}} = 10^{-14}$ とする．

## 6. 実験

荷重報酬和モデルの特性を端的に表す基礎的タスク [1] を用いて，提案手法の挙動を検証した．

### 6.1 タスクと実験設定

実験に用いたタスクを図 3 に示す．各升目が状態 $s$ を表し，上下左右 4 通りの行動 $a$ はそれぞれ矢印に沿った状態遷移を引き起こす．ただし，矢印のない方向へ行動した場合は壁への衝突とみなされ，状態は変化しない．報酬 $r$ としては，図 3 に示した括弧内の報酬値の与えられる．ただし，壁への衝突時は報酬値 $(-1)$ がさらに加えられる．このタスクの最適方策は，$b$ の値に応じて 5 通りに変化する．

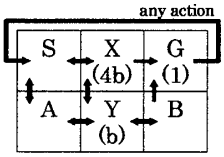


図 3: タスク (括弧内は報酬値)

初期状態は $s_0 = S$ とした．行動 $a_t$ は確率 4/5 で $\arg\max_a Q^L_{0.2}(s_t, a)$ とし，確率 1/5 でランダムに行動を選択した．割引率は $\gamma = 4/5$，学習係数は $\alpha = 0.7$ に設定した．$\beta = (b, 1)$ と置けば，このタスクを $M = 2$ の荷重報酬和モデルで表現できる．その場合，一括学習における凸包は上部凸包に置きかえてよい [1]．ただし，両端について図 2 の削除判定を行う際には，鉛直な辺を両端へ仮想的に追加する．

### 6.2 実験結果

$n = 6, \zeta^L_{\text{init}} = \zeta^U_{\text{init}} = 1, \zeta^L_{\text{min}} = \zeta^U_{\text{min}} = 10^{-8}$ に設定し，$C$ は $\zeta^X(60000) = \zeta^X_{\text{min}}$ となるよう定めた．本実験では，100000 ステップの学習を 1000 試行行い平均を求めた (実験 1)．$\Omega^L, \Omega^U$ の全状態行動対の総要素数を図 4 に，区間幅 $Q^U_{-0.4} - Q^L_{-0.4}$，$\zeta^L = \zeta^U$ を図 5 にそれぞれ示す．図 4,5 において，$\Omega^L, \Omega^U$ 共に要素数が最大でも 350 程度に抑えられ，$t = 100000$ の段階で $Q^U_{-0.4} - Q^L_{-0.4}$ は $10^{-6}$ まで狭まり，良好な結果が得られた．また，$C, n, \zeta^X_{\text{init}}$ を様々に変えても結果は大きく変わらず，これらのパラメータ設定に対しては頑健であった．一方，$\zeta^X_{\text{min}}$ を変えると，$\zeta^U$ が $\zeta^U_{\text{min}}$ に到達する付近で再び要素数が増加するなどの現象が観察された．この点は更なる検証が必要である．

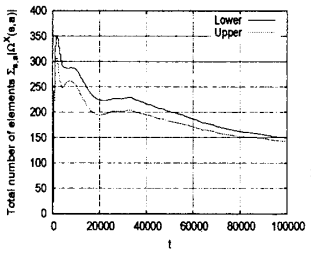


図 4: 総要素数 (実験 1)

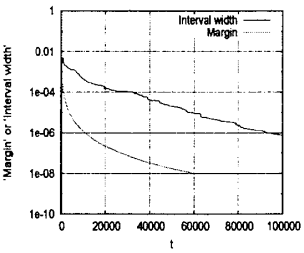


図 5: 推定区間幅 (実験 1)

次に，即時報酬 $r_t$ に範囲 $[-0.1, 0.1]$ の一様乱数をノイズとして加え，100000 ステップの学習を 100 試行行い平均を求めた (実験 2)．実験 1 と同じ設定に対する結果を図 6,7 に示す．学習が進むにつれ要素数が増大し，最終的に $\Omega^L$ は 7000，$\Omega^U$ は 5000 付近で振動した．その時の区間幅は $10^{-5}$ 程度であった．これは $\zeta^X(t) = 10^{-8}$ に固定した場合と同程度の精度であり，$\zeta^X(t)$ を減衰させることの有効性は示せなかった．適切な $\zeta^X(t)$ の関数形が問題依存である可能性も含め，学習パラメータの変更や他の問題による検証が必要である．また，実験 1,2 どちらにおいても，収束速度には改善は見られなかった．これらの検証と改善が今後の課題である．

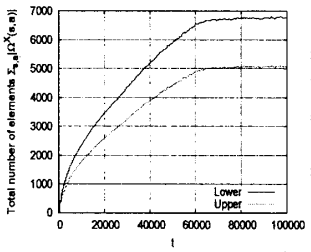


図 6: 総要素数 (実験 2)

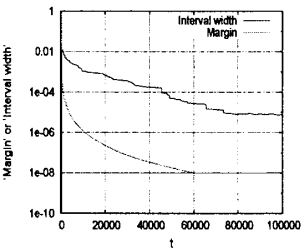


図 7: 推定区間幅 (実験 2)

## 7．結論

荷重報酬和モデルで表されたタスク族に対する一括学習法について改良版を提案した．実験の結果，ノイズがない場合において，学習初期段階では要素数を抑制し，最終段階の区間幅も抑えられた．また，パラメータ設定に対して頑健さが増した．ただし，頑健さに対して更なる検証は必要である．一方，ノイズを含んだタスクでの実験では固定マージンの従来法と同程度の精度しか得られず，適切な $\zeta^X(t)$ の関数形が問題に依存している可能性も考えられる．また，収束速度には依然改善が見られなかったので，これらの検証と改善が今後の課題として挙げられる．

### 参考文献


[1] K.Hiraoka, M.Yoshida, T.Mishima : "Parallel Reinforcement Learning for Weighted Multi-Criteria Model with Adaptive Margin". Proceedings of ICONIP 2007, Nov. 13-16, 2007.

[2] R.S. サットン，A.G. バート (三上貞芳・皆川雅章 訳)：『強化学習』．森北出版，2000.