LF-002

価値関数族の区間推定を用いた多目的強化学習法

Multi-Criteria Reinforcement Learning based on Interval Estimation of Value Functions

吉田 学 [1]　平岡 和幸 [2]　三島 健稔 [2]
Manabu YOSHIDA　Kazuyuki HIRAOKA　Taketoshi MISHIMA

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．したがって，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学習できる手法が，最近提案された [1]．一括学習により得られる推定最適価値関数は，個別に Q 学習を行なって得られるものとほぼ等しい．しかしその誤差がどの程度におさまるか保証されない点が課題として残されていた．

本研究では，Q 学習を単に近似するのではなく，その上限および下限を逐次計算する方法を提案する．これによって，近似の影響を常時監視しながら学習を進めることが可能となる．

2. 強化学習と荷重報酬和モデル

時刻 $t = 1, 2, 3, \ldots$ 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 γ は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. 一括学習法

$Q^*_{\boldsymbol{\beta}}, Q_{\boldsymbol{\beta}}$ 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 [1]．この手法では，M 次元ベクトルの有限集合 Ω 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

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

により Ω を更新する．ここに $X \sqcup Y$ は併合 $X \cup Y$ から，また $X \boxplus Y$ は Minkowski 和

$$X \oplus Y = \{x + y \,|\, x \in X, y \in Y\} \tag{5}$$

から，それぞれ凸包をとった結果の頂点集合である．両者とも，効率的なアルゴリズムが計算幾何 [3] の分野で開発されている．

頂点集合 Ω^{new} で表される推定価値関数 $Q_{\boldsymbol{\beta}}$ は，各 $\boldsymbol{\beta}$ に対して個別に Q 学習の更新を施した結果と等しい．ただし，Minkowski 和の性質から，Ω の要素数は更新につれ単調に増加する．したがって実用のためには何らかの近似が必要となる．なお，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とを注意しておく．

4. 推定最適価値関数 $Q_{\boldsymbol{\beta}}$ の下限と上限

[1] では，凸包計算時の凹凸判定にマージン $\epsilon > 0$ を導入して前述の難点を回避している．幾何学的にはこれは，図 1 に示した三角形の面積が $\epsilon/2$ 以下のときに白丸の頂点を削除することと等しい．したがって，得られる頂点集合 Ω^L は本来よりも内側になる．すると対応する $Q^L_{\boldsymbol{\beta}}$ は本来よりも小さくなるので，この近似は $Q_{\boldsymbol{\beta}}$ の下限を与える．

一方，図 2 に示した三角形の面積が $\zeta/2$ 以下のときに白丸の頂点を削除し菱形の頂点を追加すれば，得られる頂点集合 Ω^U は本来よりも外側になる．すると対応する $Q^U_{\boldsymbol{\beta}}$ は本来よりも大きくなるので，この近似は $Q_{\boldsymbol{\beta}}$ の上限を与える．

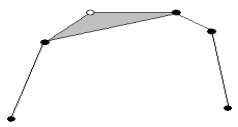

図 1: 内側の近似 ($M = 2$)

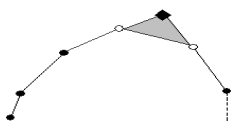

図 2: 外側の近似 ($M = 2$)

両者を組合せることで Q を区間推定するのが提案手法である．すなわち，2 組の頂点集合 Ω^L, Ω^U を保持し，それぞれを (4) で更新する．ただしその際，Ω^L には図 1 の近似を，Ω^U には図 2 の近似を施す．これにより頂点数の爆発は抑制され，しかも $Q^L_{\boldsymbol{\beta}} \le Q_{\boldsymbol{\beta}} \le Q^U_{\boldsymbol{\beta}}$ が保証される．特に，区間幅 $Q^U_{\boldsymbol{\beta}} - Q^L_{\boldsymbol{\beta}}$ が許容誤差より小さくなれば，近似の影響は無視してよい．

5. 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．

5.1 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和が与えられる．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．

[1]埼玉大学大学院理工学研究科博士前期課程
[2]埼玉大学大学院理工学研究科数理電子情報部門

初期状態は $s_0 = S$ とし，行動 a_t は上下左右の 4 つを等確率で常にランダムに選択した．割引率は $\gamma = 1/2$，学習係数は $\alpha = 0.7$ に設定した．

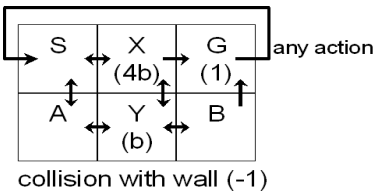

図 3: タスク (括弧内は報酬値)

$\boldsymbol{\beta} = (b, 1)$ 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]．図 2 の削除判定を両端点について行う際には，鉛直な辺を両端へ仮想的に追加する．

5.2 実験結果

内側および外側のマージン ϵ, ζ をいずれも 1.0×10^{-10} に設定し，5000 ステップの学習を行った．$b = 0.20$ における上限と下限の区間幅 $Q^U_{\boldsymbol{\beta}} - Q^L_{\boldsymbol{\beta}}$ を図 4 に，また，誤差 $|Q^U_{\boldsymbol{\beta}} - Q_{\boldsymbol{\beta}}|$ および $|Q_{\boldsymbol{\beta}} - Q^L_{\boldsymbol{\beta}}|$ を図 5 に，それぞれを示す．

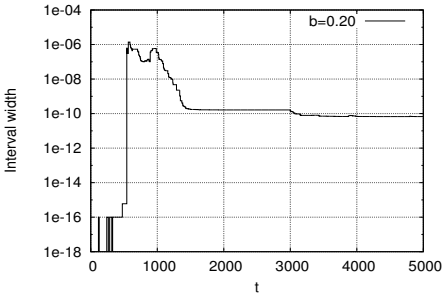

図 4: Q 値の推定区間幅

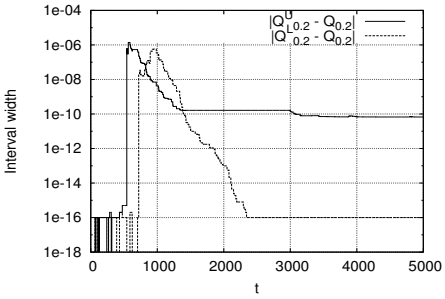

図 5: Q 学習による Q 値との差

図 4 では，数値誤差による区間幅の発生が $t = 542$ 以前に，図 5 では数値誤差による符号の逆転が $t = 2358$ 以降に生じたが，その絶対値はそれぞれ 6.0×10^{-16}，1.0×10^{-16} 程度であり十分小さい．

図 4 において，途中で一時的に区間幅が広がるものの，ステップが進むにつれ区間幅は狭くなっていることが観察できる．3500 ステップ以降は多少のゆらぎを除いて区間幅がほぼ一定となり，その値は 7.79×10^{-11} 以下となった．図 5 からわかるように，最終的な区間幅の大部分は上限の誤差に起因する．現実の強化学習タスクではこれよりはるかに大きなオーダーのノイズが見込まれるので，この程度の区間幅は多くの場合無視できるであろう．

さらに，$\Omega^L(s, a)$ および $\Omega^U(s, a)$ の要素数を図 6 および 7 にそれぞれ示す．要素数は一時的に 90 程度まで増加したが，2500 ステップ以降は 5 ないし 6 まで削減された．これは区分線形関数である真値 $Q^*_{\boldsymbol{\beta}}(s, a)$ の区分数に等しいので，適切な要素数が学習によって得られたと言える．

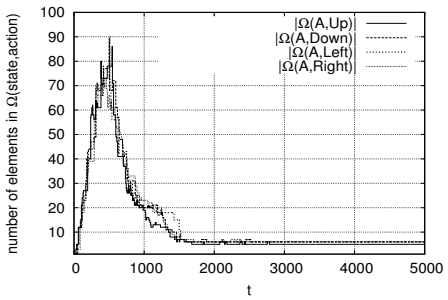

図 6: Ω^L の要素数

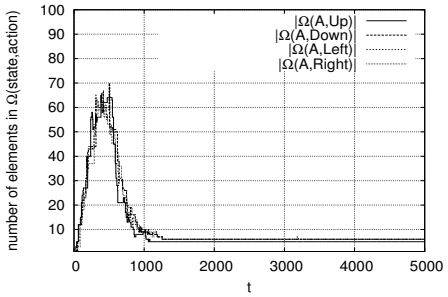

図 7: Ω^U の要素数

6 ． 結論

本研究では，[1] で提案された荷重報酬和モデルで表されたタスク族に対する最適価値関数の一括学習法を対象とした．この手法は，厳密に計算すると個別に Q 学習を行って得られるものと等しい推定最適価値関数を得られる．しかし，関数の形状が際限なく複雑化することが問題となっていた．

Q 学習の単なる近似 [1] では関数の収束性が保証されない．そこで，本研究では上限と下限を逐次更新することで，関数を区間で推定する手法を提案した．これにより，近似の影響を常に監視しながら学習を進めることが可能となった．収束性が証明されている Q 学習との誤差を学習エージェント自身が観察できる点が，提案手法の特長である．

実験の結果，個別 Q 学習との誤差は実用上問題ない程度に十分小さく，提案手法が有効であると考えられる．今後の課題として，マージンと区間幅などとの関係をより広範に検証することが挙げられる．

参考文献

[1] 平岡和幸・三島健稔:「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, 2006.

[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.

[3] F.P. プレパラータ，M.I. シェーモス (浅野孝夫・浅野哲夫訳):『計算幾何学入門』．総研出版，1992.