逐次最小２乗法に基づく強化学習法の提案

○内部英治，細田耕，浅田稔
大阪大学大学院

Reinforcement Learning Based on Recursive Least Square Method

○Eiji Uchibe, Koh Hosoda and Minoru Asada
Osaka University

Abstract— In this paper, Reinforcement Learning based on Recursive Least Square method is discussed. Based on the RLS-RL, the relation among ρ, λ and γ. The proposed method is implemented by the QR algorithm. As a result, our method can compute the value function stably. Computer simulation is shown and a discussion is given.

Key Words: reinforcement learning, recursive least square method, QR method

1. はじめに

ロボットに自律的に行動を獲得させる手法として強化学習[4]がある．ただし，従来の強化学習法を実ロボットの問題に適用するためには，幾つかの改良を必要とする．課題の一つとして，連続状態・行動空間で表現される価値関数を高速に学習する手法の開発がある．

近年，Boyan[1]は最小２乗法による TD(λ) 法を提案した．この手法では学習率 α を導入する必要がなく，特異値分解という安定な計算アルゴリズムを用いて実装しているため，従来のアルゴリズムと比較して，安定に価値関数を計算できる．ただし，特異値分解を用いているため，計算はオフラインである．

そこで本論文では，Boyan の手法を拡張して，逐次最小２乗法に基づく強化学習法を提案する．逐次計算は QR 法に基づいて実装されているため，数値的に非常に安定に価値関数を推定できる．また，逐次最小２乗法における忘却係数 ρ と強化学習における減衰係数 γ と TD(λ) 学習[4]における λ の関係を明らかにする．提案手法では忘却係数を調節することで，より自然な形で過去のデータを学習に利用できることを示す．簡単な数値実験を通して，提案手法の有効性を示す．

2. 逐次最小２乗法に基づく強化学習

2·1 基本的な考え方

まず，TD (λ) 学習について再考する．状態 $\boldsymbol{x}_t$ から $\boldsymbol{x}_{t+1}$ に遷移したとき，報酬 r_{t+1} を受け取ったとする．このとき，n-step return は

$$R_t^{(n)} = r_{t+1} + \cdots + \gamma^{n-1} r_{t+n} + \gamma^n V_t(\boldsymbol{x}_{t+n}), \quad (1)$$

となる[4]．ここで V_t は価値関数である．

TD(λ) では，1-step return と n-step return の重み付け平均

$$R_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} R_t^{(n)}, \quad (2)$$

を学習に用いる．これを λ-return と呼ぶ．更新量は TD 誤差 $R_t^\lambda - V_t(\boldsymbol{x}_t)$ に学習率 α をかけた

$$\Delta V_t(\boldsymbol{x}_t) = \alpha\left[R_t^\lambda - V_t(\boldsymbol{x}_t)\right], \quad (3)$$

とする．

式 (3) に式 (2) を代入して整理すると，

$$\frac{1}{\alpha}\Delta V_t^\lambda(\boldsymbol{x}_t) = \sum_{k=1}^{t} (\gamma\lambda)^{t-k}\delta_k \quad (4)$$

$$\delta_k = r_k + \gamma V_k(\boldsymbol{x}_{k+1}) - V_k(\boldsymbol{x}_k)$$

となる．これは複数の線形方程式

$$V_k(\boldsymbol{x}_k) = r_k + \gamma V_k(\boldsymbol{x}_{k+1}), \quad k = 1, \cdots$$

が与えられたときの，最小２乗解を求める問題に帰着される．また，式 (4) より，一般の最小２乗法における忘却係数 ρ と強化学習における二つのパラメータには

$$\rho = \sqrt{\gamma\lambda}$$

の関係があることがわかる．

2·2 Q 学習の場合

Q 学習の場合も同様である．状態 $\boldsymbol{x}$，行動 $\boldsymbol{u}$ に対する Q 値を $Q(\boldsymbol{x}, \boldsymbol{u})$ とする．このとき，TD 誤差は

$$\delta = r + \gamma \max_{\boldsymbol{u}'} Q(\boldsymbol{x}', \boldsymbol{u}') - Q(\boldsymbol{x}, \boldsymbol{u}) \quad (5)$$

となる．

$Q(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{\phi}^T(\boldsymbol{x}, \boldsymbol{u})\boldsymbol{\beta}$ と表現できると仮定する．ここで，$\boldsymbol{\beta}$ は推定するパラメータベクトルであり，$\boldsymbol{\phi}$ は特徴器である．いま，$\boldsymbol{\beta}$ を m 次元ベクトルとする．すると，逐次最小２乗法に基づく強化学習では式 (5) において，$\delta = 0$ としたときの連立方程式

$$\boldsymbol{\phi}^T(\boldsymbol{x}, \boldsymbol{u})\boldsymbol{\beta} = r + \gamma \max_{\boldsymbol{u}'} Q(\boldsymbol{x}', \boldsymbol{u}') \quad (6)$$

を解けばよい．式 (6) の右辺の Q 値は，それまでの推定されたパラメータ $\boldsymbol{\beta}$ で計算する．次節で，式 (6) を安定に解く手法を説明する．

3. QR 法を用いた逐次最小２乗法

一般に最小２乗法とは $\min_{\boldsymbol{\beta}} ||\boldsymbol{b} - \boldsymbol{A}\boldsymbol{\beta}||^2$ を最小にする $\boldsymbol{\beta}$ を求める問題である．ここで $\boldsymbol{A}$ は適当なデータ行列であり，$\mathrm{rank}\boldsymbol{A} = r$ とする．逐次的に $\boldsymbol{\beta}$ を計算する方法として，逆行列の補題を用いる方法[3]があるが，ここでは，より数値的に安定な QR 分解を用いた方法[2]を採用する．

3·1 データ行列が非特異 ($r = m$) である場合

QR 分解を用いた方法では，$\boldsymbol{A}$ の QR 分解 $\boldsymbol{A} = \boldsymbol{QR}$ を用いて，

$$\boldsymbol{Q}^T \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{Q}^T \boldsymbol{b}, \tag{7}$$

を解くことを考える．ここで $\boldsymbol{Q}$ は n 次の直交行列 ($\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_n$)，$\boldsymbol{R}$ は m 次の上三角行列である．式 (7) を整理すると，

$$\begin{bmatrix} \boldsymbol{R} \\ \mathbf{o} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} \tag{8}$$

とできる．よって，$\boldsymbol{\beta}_{LS}$ は $\boldsymbol{\beta}_{LS} = \boldsymbol{R}^{-1}\boldsymbol{y}$ と表現できる．実際には $\boldsymbol{R}$ は上三角行列であるから，逆行列を直接計算することなく $\boldsymbol{\beta}_{LS}$ を計算することができる．

3·2 データ行列が特異 ($r < m$) である場合

QR 分解を一般の場合に拡張すると，$\boldsymbol{Q}$ は $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_r$ であるような $n \times r$ 行列 $\boldsymbol{Q}$ と $r_{ij} = 0 (i > j)$ であるような $r \times m$ 行列 $\boldsymbol{R} = (r_{ij})$ を用いて表現できる．ただし，$\boldsymbol{A}$ の第 1 列から第 r 列ベクトルが 1 次独立である必要があるため，事前にそのような変換 (ピボット選択) がしてあると仮定する．

このとき，式 (8) において

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 & \boldsymbol{R}_2 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

と分割する．ここで $\boldsymbol{\beta}_1$ は r 次元，$\boldsymbol{\beta}_2$ は $m - r$ 次元のベクトルであり，$\boldsymbol{R}_1$ は $r \times r$，$\boldsymbol{R}_2$ は $r \times (m - r)$ 行列である．このとき，求める解は

$$\boldsymbol{\beta}_1 = {\boldsymbol{R}_1}^{-1} \left(\boldsymbol{Q}^T\boldsymbol{\beta} - \boldsymbol{R}_2\boldsymbol{\beta}_2 \right), \tag{9}$$

と計算できる．ここで $\boldsymbol{\beta}_2$ は任意である．

3·3 逐次計算の方法への拡張

新しいデータ $\boldsymbol{\phi}_k = \boldsymbol{\phi}(\boldsymbol{x}_k, \boldsymbol{u}_k)$，$\boldsymbol{b}_k = r + \gamma \max Q$ が与えられたとき，逐次的に $\boldsymbol{\beta}$ を計算するには，次のようにすればよい．

$$\boldsymbol{Q}_k^T \begin{bmatrix} \rho\boldsymbol{R}_{k-1} \\ \boldsymbol{\phi}_k^T \end{bmatrix} \boldsymbol{\beta} = \boldsymbol{Q}_k^T \begin{bmatrix} \rho\boldsymbol{y}_{k-1} \\ \boldsymbol{b}_k \end{bmatrix}$$

と表現できる．これより

$$\begin{bmatrix} \boldsymbol{R}_k \\ \mathbf{o} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{y}_k \\ \boldsymbol{\zeta}_k \end{bmatrix}$$

となり $\boldsymbol{\beta}_{LS,k} = \boldsymbol{R}_k^{-1}\boldsymbol{y}_k$ と計算できる．この場合，$\boldsymbol{Q}_k$ は $n+1$ 次の正方行列でサイズが一定なので，計算可能である．実際には，ハウスホルダー変換や Givens 変換 [2] などを用いて，明示的に $\boldsymbol{Q}_k$ 計算することなく，最小 2 乗解を計算できる．

4. 数値実験

提案手法を **Fig.1** のような単純な環境 (1 次元) に適用し，正しく学習できるかを確認した．ロボット (黒丸) は左右に動くことが可能である．報酬は $x > 1$ で $r = 1$，$x < 0$ で $r = 0$ として与え，それ以外は 0 とした．また，特徴器として $\boldsymbol{\phi}$ は 2 次関数を用いた．

Fig.2 に価値関数の推定誤差の推移を示す．ここで，学習中の最適行動を選択する確率を $p = 0.8, 0.2$ の 2 種類の結果について示す．価値関数を表現する基底関数として 2 次関数は適していないが，学習が進むにつれ，推定誤差は小さくなっている．実験結果の詳細については，当日報告する．

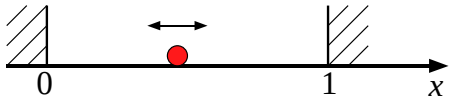
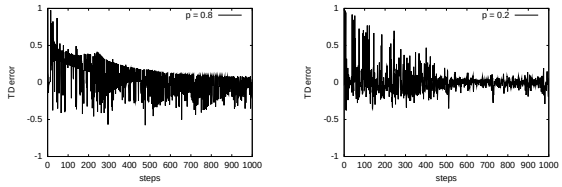


Fig.1 Simple environment



(a) $p = 0.8$ (b) $p = 0.2$

Fig.2 TD error

5. おわりに

本研究では，最小 2 乗法に基づく強化学習法を提案した．提案手法は QR 法を用いて実装されているため，安定に，かつ逐次的に数値計算が可能である．現在，実ロボットの協調行動の学習の問題に対して，本手法を実装中である．

今後の課題として，報酬関数の多次元化や状態ベクトル推定法との統合が考えられる．特に，オンラインで状態ベクトルを推定する手法との統合に取り組んでいる．

謝辞

本研究は，日本学術振興会 未来開拓学術研究推進事業「分散協調視覚による動的 3 次元状況理解」プロジェクト (課題番号 JSPS-RFTF96P00501) の補助を受けた．

参考文献

1) J. A. Boyan. Least-Squares Temporal Difference Learning. In *Proc. of the Sixteenth International Conference on Machine Learning*, 1999.
2) G. H. Golub and C. F. Van Loan. *Matrix computations*. North Oxford academic Publishing Co., Johns Hopkins Press, 1983.
3) 片山. 応用カルマンフィルタ, 第 8 章, pp. 133–154. 朝倉書店, 1983.
4) R. S. Sutton and A. G. Barto. *Reinforcement Learning*. MIT Press/Bradford Books, March 1998.