未来の観測状態を考慮したTD法について

鈴木 豪　小谷 善行
東京農工大学
{go, kotani}@fairy.ei.tuat.ac.jp

概要

TD(Temporal Differential:時間的差分)法は、Samuel(1963)のチェッカープログラムに始まるマルチステップ予言問題に対する学習アルゴリズムである。Sutton(1988)は過去の観測状態を考慮するTD(λ)法を提案し、それが従来行われていた観測状態と最終結果のみからの学習法よりも効率がよいことを示した。

本稿では未来の観測状態のみを考慮したTD(0,μ)法、および過去と未来の両方の観測状態を考慮したTD(λ,μ)法を提案する。TD(λ,μ)法においてμ=0とおくとTD(λ)法と一致し、TD(λ)法の一般化になっている。最後に"4×3の世界"という仮想世界を使って行った学習実験の結果を示す。

TD learning in consideration of the future data

Tsuyoshi SUZUKI　Yoshiyuki KOTANI
Tokyo University of Agriculture and Technology
{go, kotani}@fairy.ei.tuat.ac.jp

Abstract

Temporal Differential (TD) learning is learning algorithm for a multi-step prediction problem. Sutton(1988) proposed TD (λ) learning that past observation condition was taken into consideration. In this paper, we propose TD(0, λ) learning that future observation condition was taken into consideration and TD (λ, μ) learning that past and the future observation condition was taken into consideration. When it is put with μ=0 in the TD (λ, μ) learning, it corresponds to the TD (λ) learning, and it is the generalization of the TD (λ) learning. Finally, we show the result of the learning experiment using of the TD(λ,μ) in the "world of 4×3".

1 はじめに

過去の経験から将来の振舞いの予言を学習する問題を、予言学習問題という。例えば、将棋のある局面が経験から勝ちに結びつくかどうかの予言を学習するなどである。TD(Temporal Differential:時間的差分)法は Samuel(1963)のチェッカープログラムに始まるマルチステップ予言学習問題に対する学習アルゴリズムである。ここで、マルチステップ予言問題とは、予言の正当性が複数ステップを経てからでないと明らかにされないものを言う。将棋の場合には、ある局面が勝ちにどのくらい結びつくかという予言の正当性はゲーム終了時の勝敗が出るまで明らかにならないので、局面から勝敗の予測を学習することはマルチステップ予言学習問題である。

予言学習アルゴリズムとして最も素朴なものは、Widrow and Hoff(1960)により発表された LMS(Least Mean Square)法である。これは観測状態と最終結果のみからの学習を行う。一方、Sutton (1988)は過去の観測状態を考慮する TD(λ)法を提案し、その手法が LMS 法よりも効率がよいことを示した。

TD(λ)法では、過去の観測状態のみを考慮して学習を行っている。本稿では未来の観測状態のみを考慮した TD(0,μ)法、および過去と未来の両方の観測状態を考慮した TD(λ,μ)法を提案する。TD(λ)法は TD(λ,μ)法においてμ=0 とした特別な場合であり、その意味で TD(λ,μ)法は TD(λ)法の一般化になっている。

第 2 節では LMS 法と TD(λ)法について概観する。第 3 節では TD(λ,μ)法を提案し、これが TD(λ)法の一般化になっていることを示す。第 4 節においては“4×3の世界”という仮想世界を導入し、第 5 節でこの世界を使った LMS 法、TD(λ)法、TD(0,μ)法、TD(λ,μ)法の学習実験の結果を示す。最後に、第 6 節においてこれらの実験の考察を行う。

2 LMS 法と TD 法

いま、時刻 1,2,3,… における観測データを $x_1, x_2, \cdots, x_n$ とし、最終的な結果を z とする。また、予言は観測データ x とパラメータ w を用いて $P(x,w)$ と表されるものとする。このとき、LMS 法ではパラメータ w を

$$w \leftarrow w + \sum_{t=1}^{n} \Delta w_t$$

$$\Delta w_t = \alpha (z - P_t) \nabla_w P_t$$

で更新し、学習を行う。ここで正定数αは学習率と呼ばれる。

LMS 法は、予言と実際の結果の平均自乗誤差

$$Err = \frac{1}{n} \sum_{t=1}^{n} \left(z - P(x_t, w) \right)^2$$

を最小にするようにパラメータ w を更新する手続きになっている。

LMS 法の欠点は、観測された各データが独立ではないことを無視していることである。例えば将棋の場合、ある局面はそれに至るまでの局面と深い関係がある。LMS 法ではこれらの制約を無視しているため、一般にはゆっくりと学習が行われる。

LMS 法では予言と実際の結果の差からパラメータの更新を行うのに対し、TD(λ)法では予言間の差からパラメータの更新を行う。TD(λ)法では、過去の観測状態のみを考慮している。この手法では、パラメータ w を

$$w \leftarrow w + \sum_{t=1}^{n} \Delta w_t$$

$$\Delta w_t = \alpha(P_{t+1} - P_t) \sum_{k=1}^{t} \lambda^{t-k} \nabla_w P_k$$

で更新する。ここでλは予言間の依存関係を表す正定数である。とくに、$\lambda = 1$としたTD(1)は、wの更新周期をそろえればLMS法と一致する。

TD法がLMS法よりも効率よく学習するゲームの例を図2.1に示す(Sutton 88)。この例では未知の局面から矢印を通って不利な局面に至り、多くの場合は負けにつながるが、偶然にも勝ってしまったとする。LMS法では未知の局面の予言は勝ちに近いものと学習が行われるのに対し、TD法では近い未来の不利であるということを学習するため効率がよい。

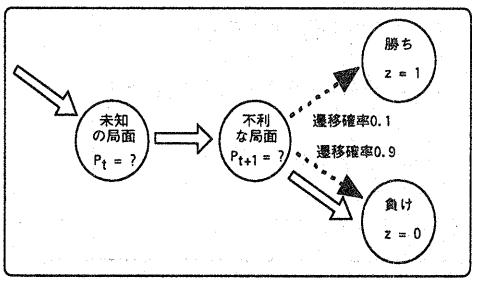

図2.1 TDとLMSの比較

3 未来を考慮した学習

TD(λ)法では、現在の観測状態は過去の観測状態に関係して起こるという仮定に基づいて、図3.1のように現在の観測状態と過去の観測状態との関連を予言の勾配に指数型の重みをつけた形で考慮している。

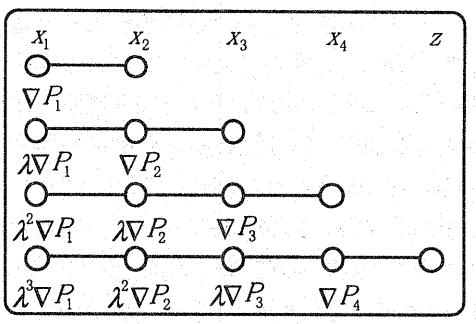

図3.1 TD(λ)法における学習

一方、現在の観測状態は過去の観測状態と関係があるということと同様に、未来の観測状態も現在の観測状態に関連があり、未来から現在状態を推測できると考えることができる。この仮定の下、TD(λ)法の予言の勾配に指数型の重みづけを、未来項に対して行った図3.2のTD(0, μ)法が考えられる。TD(0, μ)法では過去との関係ではなく、未来との関係のみを考慮している。

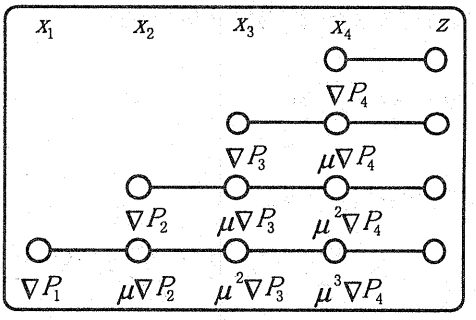

図3.2 TD(0, μ)法における学習

TD(0, μ)法は次式で表される:

$$w \leftarrow w + \sum_{t=1}^{n} \Delta w_t$$

$$\Delta w_t = \alpha(P_{t+1} - P_t) \sum_{k=t}^{n} \mu^{k-t} \nabla_w P_k$$

過去との関連を考慮したTD(λ)法および未来との関連を考慮したTD(0, μ)法と同様

に、すべての時刻における観測状態を考慮した TD(λ, μ)法が考えられる。これは過去との関連に関しては TD(λ)法と同様に、未来の項に関しては TD(0, μ)と同様に指数型の重みをつけたものである(図 3.3)。TD(λ, μ)法は TD(λ)法のように過去との関連だけでなく、未来との関連までも考慮して、より良い学習を行おうという考えに基づいている。

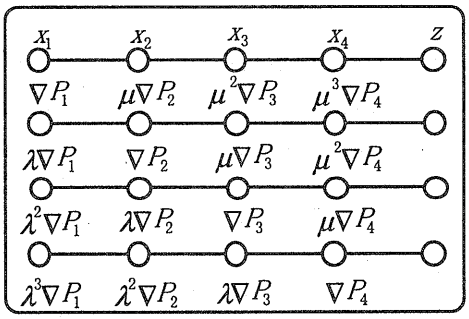

図 3.3 未来と過去を考慮した学習

これを定式化すると次のようになる:

$$w \leftarrow w + \sum_{t=1}^{n} \Delta w_t$$

$$\Delta w_t = \alpha(P_{t+1} - P_t) \times \left(\sum_{k=1}^{t} \lambda^{t-k} \nabla_w P_k + \sum_{k=t+1}^{n} \mu^{k-t} \nabla_w P_k \right)$$

TD(λ, μ)法において、μ=0 とおいて未来項をなくすと、TD(λ,0)は TD(λ)法と一致する。すなわち、TD(λ, μ)は TD(λ)法の一般化になっている。

4 "4×3の世界"

"4×3の世界"とは図 4.1 のように正方形が横4、縦3に並んだ仮想世界で、各升から各升への遷移確率(図中の矢印の隣に書いてある数字)が与えられている。このとき升(1,1)からスタートしてゴールの升(3,4)または(2,4)に向かう。升(3,4)では報酬1が与えられ、升(2,4)では報酬0が与えられる。

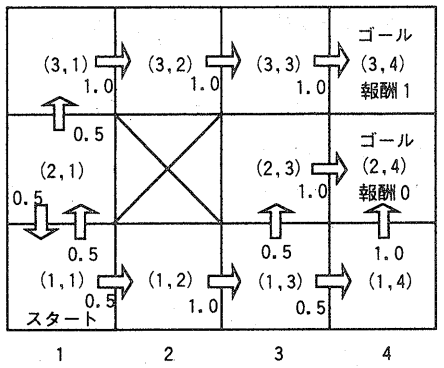

図 4.1 4×3の世界

升(i,j)にいるときに期待できる報酬の正確な期待値 $e_{(i,j)}$ は

$$e_{(2,4)} = e_{(2,3)} = e_{(1.4)} = e_{(1,3)} = e_{(1,2)} = 0$$

$$e_{(1,1)} = \frac{1}{2} e_{(1,2)} + \frac{1}{2} e_{(2,1)}$$

$$e_{(3,4)} = e_{(3,3)} = e_{(3,2)} = e_{(3,1)} = 1$$

$$e_{(2,1)} = \frac{1}{2} e_{(1,1)} + \frac{1}{2} e_{(3,1)}$$

より、図 4.3 のようになる。

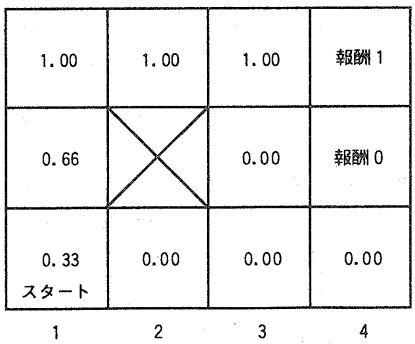

図 4.3 正確な期待値

5 学習実験

LMS法、TD(λ)法、TD(0, μ)法、TD(λ, μ)法を使って"4×3の世界"の各升に期待できる報酬の期待値を学習させた。学習を行うための問題設定を次のようにする。スタート地点(1,1)から出発して時刻tに升(i,j)にいる時の状態を観測ベクトル

$$x_t = (\delta_{(1,1)}, \delta_{(1,2)}, \delta_{(1,3)}, \delta_{(1,4)}, \delta_{(2,1)}, \delta_{(2,3)}, \delta_{(3,1)}, \delta_{(3,2)}, \delta_{(3,3)})$$

$$\delta_{(i,j)} = \begin{cases} 1 & (\text{升}(i,j)\text{にいるとき}) \\ 0 & (\text{それ以外}) \end{cases}$$

で表す。また、各升の報酬の期待値を並べた期待値ベクトルを

$$E_t = (e_{(1,1)}, e_{(1,2)}, e_{(1,3)}, e_{(1,4)}, e_{(2,1)}, e_{(2,3)}, e_{(3,1)}, e_{(3,2)}, e_{(3,3)})$$

$e_{(i,j)}$: 升(i, j)の期待値

とする。このとき、予言を

$$P_t = E^T x_t$$

と定義する。これは時刻tにおいて升(i,j)にいたときに期待される報酬の期待値を表す。

このとき、学習を行う期待値ベクトルの更新式は予言 P_t が線型なので

$$E \leftarrow E + \sum_{t=1}^{m} \Delta E_t$$

$$\Delta E_t = \alpha(z - E^T x_t) x_t \quad (for\ LMS)$$

$$\Delta E_t = \alpha(E^T x_{t+1} - E^T x_t) \sum_{k=1}^{t} \lambda^{t-k} x_k \quad (for\ TD(\lambda))$$

$$\Delta E_t = \alpha(E^T x_{t+1} - E^T x_t) \sum_{k=t}^{n} \mu^{k-t} x_k \quad (for\ TD(0,\lambda))$$

$$\Delta E_t = \alpha(E^T x_{t+1} - E^T x_t) \times \left(\sum_{k=1}^{t} \lambda^{t-k} x_k + \sum_{k=t+1}^{n} \mu^{k-t} x_k \right) \quad (for\ TD(\lambda,\mu))$$

となる。

また、以下の実験で指標とする誤差は、学習を行う升の集合を

$$S = \{(1,1),(1,2),(1,3),(1,4),(2,1),(2,3),(3,1),(3,2),(3,3)\}$$

学習された升 (i, j) の期待値を $e_{(i,j)}$、計算で求めた正確な期待値を $t_{(i,j)}$ とするとき

$$Err = \frac{1}{\|S\|} \sum_{(i,j)\in S} \left(t_{(i,j)} - e_{(i,j)} \right)^2$$

で定義する(平均自乗誤差)。ここで、$\|S\|$ は S の要素数である。

それぞれの手法において、各升に与えた期待値の初期値および与えた観測データ列は同一である。また、α =0.05、データ数100、反復回数300回で固定している。

(1) LMS法による学習結果

LMS法の学習結果を図5.1に示す(括弧内は正確な期待値)。升(2,1)の誤差は0.10と大きいが、それ以外の升についてはほぼ正確な値が学習できている。最終的な平均自乗誤差は0.04となった。

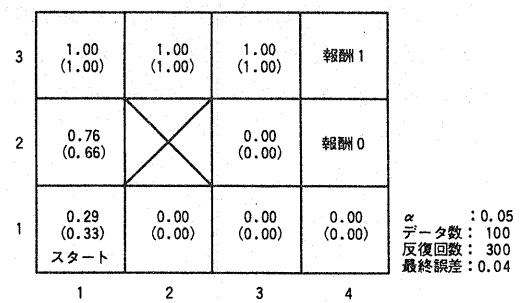

図5.1 LMS法による学習結果

図 5.2 に升(1,1),(1,2),(2,1),(3,1)の学習曲線を示す。値は反復が 160 回程度で収束しているが、升(1,1),(2,1)においては収束後も比較的大きな振動をしている。図 5.3 に誤差収束の様子を示す。

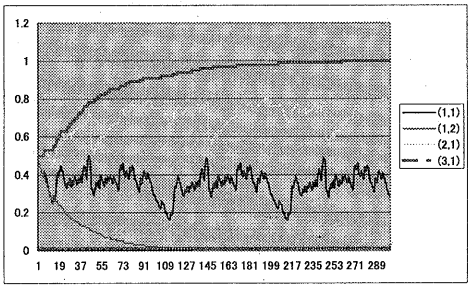

図 5.2　LMS 法における学習曲線

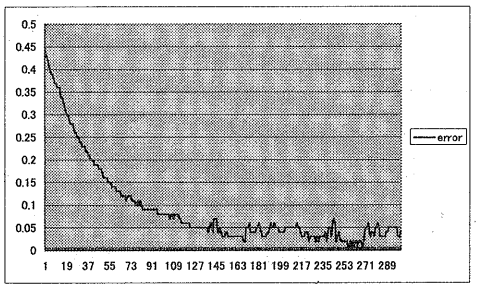

図 5.3　LMS 法における誤差収束

(2) TD(λ)法による学習結果

λ=0.80 とした TD(λ)法の学習結果を図 5.4 に示す。TD(λ)法でも LMS 法と同様にほぼ正確な期待値が学習できている。最終的な結果は升(1,1)の誤差が 0.14 と比較的大きくなったが、それ以外の升については正確な値に近い。最終的な平均自乗誤差は 0.05 であった。

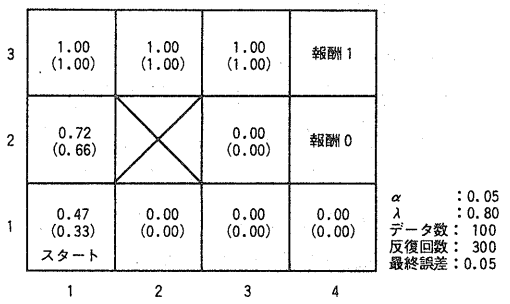

図 5.4 TD(λ)法による学習結果

図 5.5 に升(1,1),(1,2),(2,1),(3,1)の学習曲線を示す。TD 法でもLMS法と同様に 160 回程度の反復でほぼ収束している。図 5.6 に誤差の収束の様子を示す。

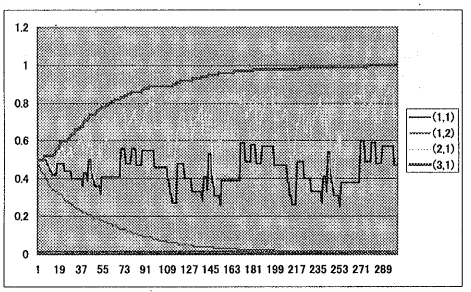

図 5.5　TD(λ)法における学習曲線

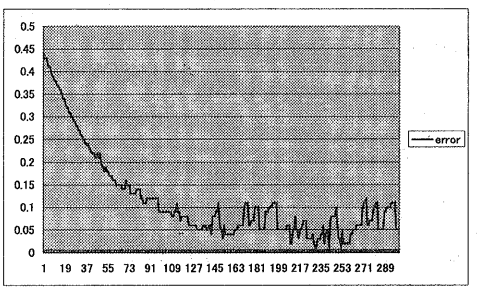

図 5.6　TD(λ)法における誤差収束

(3) TD(0,μ)法による学習結果

未来のみを考慮した TD(0,μ)法の学習結果を図 5.7 に示す。ここではμ=0.20 としている。最終的に得られた値は、LMS 法、TD(λ)法と比較して全体的に誤差が大きく

なっている。平均自乗誤差は 0.08 であった。

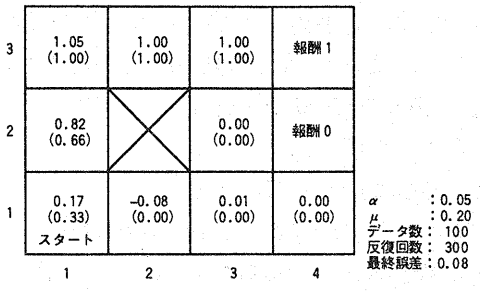

図 5.7 TD(0, μ)法による学習結果

TD(0, μ)法の学習曲線を図 5.8 に誤差収束の様子を図 5.9 に示す。反復が 160 回程度で収束していると考えられるが、それ以降も振動の幅が大きい。

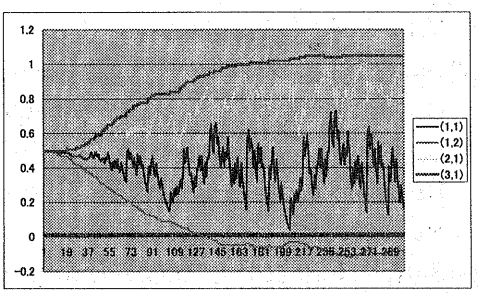

図 5.8 TD(0, μ)の学習曲線

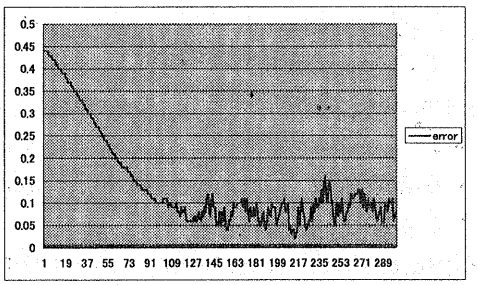

図 5.9 TD(0, μ)の誤差収束

(4) TD(λ, μ)法による学習結果

過去と未来の両方を考慮した TD(λ, μ)法の学習結果を図 5.10 に示す。ここでは、λ=0.80、μ=0.20 としている。升(1,1)で得られた値の誤差は若干大きくなっているものの、それ以外の升についてはほぼ正確な値が学習できている。最終的な平均自乗誤差は 0.05 であった。

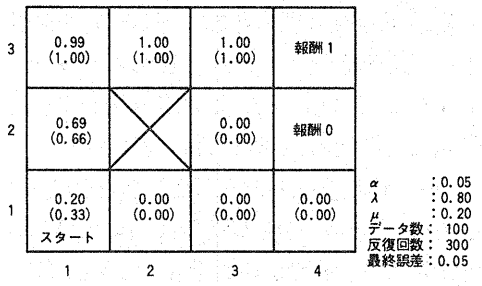

図 5.10 TD(λ, μ)法による学習結果

TD(λ, μ)法の学習曲線を図 5.11 に誤差収束の様子を図 5.12 に示す。

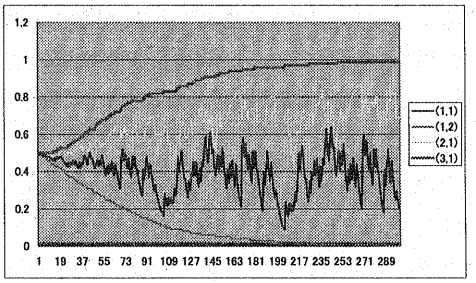

図 5.11 TD(λ, μ)法の学習曲線

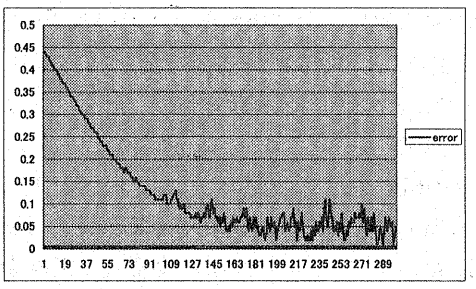

図 5.12 TD(λ, μ)法の誤差収束

6 考察

LMS 法と、その他の学習法では収束の速さ、および最終的に得られた誤差に大きな違いはみられなかった。これは“4×3の世界”が、図 2.1 で示したような、LMS 法がうまく学習できない状態がほとんど発生しない

単純な環境であるためと考えられる。LMS法よりも他の学習方法が誤差が大きくなる傾向にあるのは、LMS 法を除いては、重みの更新量が複数の項の和になっており、1回当たりの更新量が多いため振動の幅が大きくなってしまうことによる。この意味で、同じ学習率αでの比較は単純にはできない。

未来の観測状態のみを考慮した TD(0,μ)法では、学習の初期の段階においては、小さな変動で、学習が少し進んでからは高速に真値に向かうという、望ましい性質が見られた。しかし、ほぼ収束したと考えられるところからの学習曲線の振動は大きく、最終的な誤差は他の学習法と比較して大きくなった。TD(0,μ)法においては、例えば通常の学習で行われるように、αを学習回数と共に小さくしていくなど、振動を抑えるための工夫が必要である。

未来と過去の両方を考慮した TD(λ,μ)法では、最終的な誤差は升(1,1)は幾分大きいものの、その他の升については、他の学習法よりも小さく、より正確な値が得られた。升(1,1)の誤差が大きいのは、TD(λ,μ)法が、学習曲線において比較的大きな振動をしてしまい、大きくずれたときに学習が打ち切られたことによる。これらの振動を抑えれば、より良い結果が得られると期待できる。α、λ、μの値の決定は TD(λ)法と同様に難しい。本稿では結果は示さないが、TD(λ,μ)法において、λ、μを様々に変更して実験を行った結果、過去と未来を均等に扱うのではなく、過去を重視し、未来については若干考慮した方がより良い学習を行う傾向がみられた。

8 まとめと今後の課題

未来の観測状態を考慮した TD(λ,μ)法を提案し、"4×3の世界"を使って学習実験を行った。TD(λ,μ)法は、TD(λ)法を特別な場合として含む、TD(λ)法の一般化になっている。過去のみでなく、未来との関連も考慮することにより、より効率的な学習ができると期待できると考えられる。

今後の課題として

(1) 理論的な裏付けの構築

(2) α、λ、μの設定・更新方法の確立

(3) 将棋など、より複雑なゲームへの適用

などが挙げられる。

参考文献

[1] Beal, D.F. and Smith, M.C. (1997). Learning Piece Value Using Temporal Differences. ICCA Journal, Vol. 19, No. 3. Pp.147-151.

[2] Baxter, J. Tridgell, A. and Weaver, (1998). Experiments in Parameter Learning Using Temporal Differences. ICCA Journal, Vol. 20, No. 3. Pp.147-161.

[3] Sutton, R.S. (1987). Implementation Details of the TD(λ) Procedure for the Case of Vector Predictions and Backpropagation. GTE Laboratories Technical Note TN87-509.1.

[4] Sutton, R.S. (1988). Learning to Predict by the Methods of Temporal Differences. Machine Learning, Vol. 3, pp.9-44.

[5] Sutton, R.S. and Barto, A.G. (1988) Reinforcement Learning:An Introduction. MIT Press.

[6] Tesauro, G. (1995). Temporal Difference Learning and TD-Gammon. Communications of the ACM, Vol. 38, No. 3.