1F3-2

Actor-critic アルゴリズムにおける actor の効率的学習のための critic の学習

Learning value function for efficient policy learning in actor-critic algorithm

横山 裕樹*1　浅田 稔*2
Hiroki Yokoyama　Minoru Asada

*1*2大阪大学大学院工学研究科 知能・機能創成工学専攻
Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University

We propose an actor-critic algorithm in which both the actor and the critic use eligibility traces. It is already reported that eligibility traces enable the actor to learn a good policy even though the approximated value function in the critic is inaccurate. Then, in the case where the actor uses eligibility traces, the role of the critic is not to make the actor's learning accurate, but to accelerate or to stabilize. Our proposal is explicitly minimizing the distance between the actor's update rule and true gradient of the expected reward. In this paper, we apply the proposed algorithm to pole balancing problems and show that the algorithm can learn faster than the previous algorithms which use TD method for the critic's learning.

1. はじめに

強化学習の分野で古くから用いられてきた手法に actor-critic アルゴリズムがある [Barto 83]．この手法では，エージェントは actor と critic から構成され，actor は行動を決定し，critic は環境から情報を集めることで状態の価値を推定し，これに基づいて actor の行動を評価する．強化学習の手法には他にも Q 学習や sarsa など，行動の価値に基づくものが知られているが，actor-critic アルゴリズムは行動を選択するための方策を状態価値とは別に学習するため，

- 行動の選択に必要な計算コストが少なくて済む，特に行動が連続値で表現される場合にも実装が容易である，
- 明示的に確率的な方策を学習できる，

という点において優れている．

強化学習が適用される課題においてしばしば問題となるのが報酬の遅延である．これを解決する手法として

- TD(temporal difference) 法による状態および行動価値の推定 [Sutton 90]，
- 適正度の履歴（eligibility trace）を用いた学習 [Singh 96]，

の二つがよく知られている．前者は環境のマルコフ性を仮定することによって得られる Bellman 方程式に基づいて状態や行動がどれだけ望ましいかを推定する手法であり，学習が完了すれば，本来は遅延して得られる報酬を価値関数によって直ちに知ることができる．しかし，マルコフ的でない環境や学習が不完全な場合に最適な方策が学習できないという問題がある [Singh 96]．後者は状態や行動の履歴を短期記憶として保持することで，現在の報酬を，頻繁に経験する過去の状態・行動の価値に反映させる手法で，非マルコフ的な環境にも対応できることが知られている [Pendrith 96]．

適正度の履歴は，はじめはヒューリスティックな手法として用いられたが，その後多くの研究によって理論的な分析が成されてきた．木村ら [Kimura 98, 木村 00] は，actor の学習に適正度の履歴を用いた actor-critic アルゴリズムの特性について分析し，価値関数が不正確な場合でも actor が最適な方策を獲得できることを示した．本研究では，彼らのアルゴリズムにおける critic の学習について着目し，actor の学習をより効率的にする新たなアルゴリズムを提案する．

連絡先: 横山裕樹，yokoyama@ams.eng.osaka-u.ac.jp

2. 適正度の履歴を用いた actor-critic アルゴリズム

本節では，まず一般的な actor-critic アルゴリズムを定式化し，その後木村らのアルゴリズムについて説明する．

Actor は各時刻 t における行動 a_t を確率的方策 $\pi(s,a) = p(s|a;\boldsymbol{w})$ に従って生成する．ここでベクトル $\boldsymbol{w}$ は方策を記述するパラメータである．エージェントの目的は，将来得られる報酬 $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$ の期待値を最大化することである．ただし，r_t は時刻 t において得られる即時報酬，$0 \leq \gamma < 1$ は直近の報酬を重視するように重み付けする割引率である．従って，actor は各時刻 t において

$$E(\boldsymbol{w}_t) = E[R_t|\boldsymbol{w}_t] \tag{1}$$

を増加させるように $\boldsymbol{w}_t$ を更新するべきであるが，一般に $E(\boldsymbol{w})$ を直接知ることはできない．そこで critic が状態価値 $V(s) = E[R_t|s_t = s]$ を学習することで将来の報酬を予測し，より価値の高い状態へ遷移するような行動を強化するように actor に強化信号を送る．Critic の学習には TD 法が用いられることが多い．

価値関数の学習に TD(0) を用いた場合の木村らのアルゴリズムを Alg.1 に示す．このアルゴリズムにおける方策の更新則 $\Delta\boldsymbol{w}$ は次のように変形できる．

$$\sum_{t=0}^{\infty} \Delta\boldsymbol{w}_t = \sum_{t=0}^{\infty} \boldsymbol{e}_t \left(R_t - \hat{V}_{t-1}(\boldsymbol{s}_t)\right) \tag{2}$$

ここで $\boldsymbol{e}_t = \frac{\partial}{\partial \boldsymbol{w}_t} \log \pi(\boldsymbol{a}_t, \boldsymbol{s}_t; \boldsymbol{w}_t)$ は時刻 t における方策の適正度を表す．右辺の $\hat{V}_{t-1}$ は時刻 t においてすでに critic によって決定されており，a_t, V_t と条件付き独立であるため，REINFORCE アルゴリズムの定理 [Williams 92] を適用することにより

$$E\left[\boldsymbol{e}_t(R_t - \hat{V}_{t-1}(s_t))\middle| s_t, \boldsymbol{w}_t\right] = \frac{\partial}{\partial \boldsymbol{w}_t} E[R_t|s_t, \boldsymbol{w}_t] \tag{3}$$

Algorithm 1 [Kimura 98] のアルゴリズム

$\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{w}_t, \overline{\boldsymbol{e}}_0$ を初期化する．
$t \leftarrow 1$
loop
　環境から状態 $\boldsymbol{s}_t$ を観測する．
　方策 $\pi(\boldsymbol{a}_t, \boldsymbol{s}_t; \boldsymbol{w}_t)$ に基づいて行動 $\boldsymbol{a}_t$ を実行する．
　環境から報酬 r_t と次の状態 $\boldsymbol{s}_{t+1}$ を観測する．
　Critic は以下のように actor への強化信号 δ_t を計算する．
　　$\delta_t = r_t + \gamma \hat{V}(\boldsymbol{s}_{t+1}; \boldsymbol{v}_t) - \hat{V}(\boldsymbol{s}_t; \boldsymbol{v}_{t-1})$
　Actor は以下のように方策を更新する．
　　$\boldsymbol{e}_t = \frac{\partial}{\partial \boldsymbol{w}_t} \log \pi(\boldsymbol{a}_t, \boldsymbol{s}_t; \boldsymbol{w}_t)$
　　$\overline{\boldsymbol{e}}_t = \boldsymbol{e}_t + \beta \overline{\boldsymbol{e}}_{t-1}$
　　$\Delta \boldsymbol{w}_t = \delta_t \overline{\boldsymbol{e}}_t$
　　$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \alpha_w \Delta \boldsymbol{w}_t$
　Critic は以下のように価値関数を更新する．
　　$\boldsymbol{v}_{t+1} = \boldsymbol{v}_t + \alpha_v \delta_t \frac{\partial}{\partial \boldsymbol{v}_t} \hat{V}(\boldsymbol{s}_t; \boldsymbol{v}_t)$
　$t \leftarrow t + 1$
end loop

が成立する．これは式 (2) 右辺の各項の期待値が $E(\boldsymbol{w})$ の勾配に等しいことを示している．このことから，$\Delta \boldsymbol{w}_t$ による $\boldsymbol{w}_t$ の更新を繰り返すことで，平均的に $E(\boldsymbol{w})$ が増加していくことがわかる．さらに，式 (3) より，更新則の期待値は $\hat{V}$ に依存しないことがわかる．以上より，価値関数が不正確な場合でも actor が最適な方策を獲得することが確認できる．

3. 提案手法

[木村 00] の結果は，「価値関数の推定値 $\hat{V}$ は方策の更新則の期待値 $E[\Delta \boldsymbol{w}]$ には影響を与えない」と解釈することができる．このため $\hat{V}$ がどのような一定値を取っていても，方策の更新を十分な回数にわたって行うことで，方策 π を最適なものに近づけることができる．これは，一見すると $\hat{V}$ の学習が無意味になったかのように思われる．しかしながら，[木村 00] でも指摘されている通り，$\hat{V}$ は方策更新のステップ幅をコントロールしているという点で，学習全体に尚も一定の役割を果たしている．これを上記の解釈に当てはめると，「$\hat{V}$ は方策の更新則の二次以上の統計量には影響を与えている」と考えられる．方策の更新則 $\Delta \boldsymbol{w}$ は，十分な回数繰り返すことで平均的に真の勾配 $\frac{\partial}{\partial \boldsymbol{w}} E[R|\boldsymbol{w}]$ に近づいていくが，各ステップにおける値は少しずつ異なっている．この誤差は $\hat{V}$ に依存しており，critic が学習することで小さくなると考えられる．

本研究では，以上のような方策更新のステップ幅のコントロールを明示的に実現する critic の学習則を提案する．具体的には，方策の更新則と真の勾配との距離の期待値

$$J = \frac{1}{2} E\left[\left\| \sum_{t=0}^{\infty} \Delta \boldsymbol{w}_t - \sum_{t=0}^{\infty} \frac{\partial}{\partial \boldsymbol{w}_t} E[V_t|\boldsymbol{w}_t] \right\|^2 \right]$$
$$= \frac{1}{2} E\left[\left\| \sum_{t=0}^{\infty} \Delta \boldsymbol{w}_t \right\|^2 \right] - \frac{1}{2} \left\| \sum_{t=0}^{\infty} \frac{\partial}{\partial \boldsymbol{w}_t} E[V_t|\boldsymbol{w}_t] \right\|^2 \quad (4)$$

を最小化するように $\hat{V}$ を学習する．[木村 00] の結果からわかる通り，TD(0) 等の手法を用いて critic の学習を行っても，結果的にステップ幅をコントロールすることができる．しかし，TD 学習は Bellman 方程式に基づく状態価値の推定手法であり [Sutton 90]，期待報酬を最大化するという目的に対して必ずしも最適とは限らない．

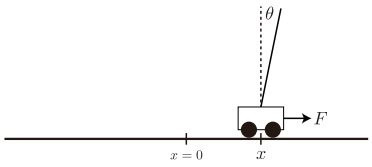

図 1: 倒立振子制御問題．

目的関数 J を最小化するために，$\boldsymbol{v}_t$ で微分する．式 (4) の第 2 項は木村らの結果から $\boldsymbol{v}_t$ に依存しないことと，$\Delta \boldsymbol{w}_{t+1}$ に $\hat{V}(\boldsymbol{s}_{t+1}; \boldsymbol{v}_t)$ が含まれていることに注意すると，

$$\Delta \boldsymbol{v}_t = \frac{\partial J}{\partial \boldsymbol{v}_t} = E\left[C_t \sum_{\tau=0}^{\infty} \Delta \boldsymbol{w}_\tau \right], \quad (5)$$
$$C_t = \frac{\partial}{\partial \boldsymbol{v}_t} \hat{V}(\boldsymbol{s}_{t+1}; \boldsymbol{v}_t) \boldsymbol{e}_{t+1}^{\mathrm{T}}$$

が得られる．この式は時刻 t における $\boldsymbol{v}_t$ の更新則に未来 $\tau > t$ の情報が必要であることを示しており，そのままでは計算できない．そこで，次式のように変形する．

$$\sum_{t=0}^{\infty} C_t \sum_{\tau=0}^{\infty} \Delta \boldsymbol{w}_\tau = \sum_{t=0}^{\infty} C_t \sum_{\tau=0}^{t} \Delta \boldsymbol{w}_\tau + \sum_{t=0}^{\infty} C_t \sum_{\tau=t+1}^{\infty} \Delta \boldsymbol{w}_\tau$$
$$= \sum_{t=0}^{\infty} C_t \left(\sum_{\tau=0}^{t} \Delta \boldsymbol{w}_\tau \right) + \sum_{t=0}^{\infty} \left(\sum_{\tau=0}^{t-1} C_\tau \right) \Delta \boldsymbol{w}_t \quad (6)$$

第 2 行の二つの括弧は $\Delta \boldsymbol{w}_t$ と C_t の累積和を表している．従って，$\boldsymbol{e}_t$ と同様に累積和による履歴を用いることで，目的関数 J の勾配に近づいていく更新則 $\Delta \boldsymbol{v}$ を構成できる．

式 (6) には，それぞれ t, τ の様々な組み合わせによる項が含まれている．従って，更新則には $\Delta \boldsymbol{w}_t$ と C_t との時間を大きく隔てた二つの組み合わせが情報として含まれ得る．これは，$\boldsymbol{w}, \boldsymbol{v}$ が固定されている場合は問題にはならないと考えられる．しかし，実際には時間経過に従って学習が進行し，$\Delta \boldsymbol{w}_t, C_t$ はそれぞれ変化するため，両者の間に矛盾が生じる可能性がある．そこで，$|t - \tau|$ が大きい組み合わせを割り引くための定数 $0 \leq \beta \leq 1$ を新たに導入し，

$$\Delta \boldsymbol{v}_t = E\left[C_t \sum_{\tau=0}^{\infty} \beta^{|t-\tau|} \Delta \boldsymbol{w}_\tau \right] \quad (7)$$

のように更新則を変更する．これは式 (6) と同様に

$$\sum_{t=0}^{\infty} C_t \left(\sum_{\tau=0}^{t} \beta^{t-\tau} \Delta \boldsymbol{w}_\tau \right) + \sum_{t=0}^{\infty} \left(\sum_{\tau=0}^{t-1} \beta^{t-\tau} C_\tau \right) \Delta \boldsymbol{w}_t \quad (8)$$

のように計算できる．

以上をアルゴリズムとしてまとめたものを，Alg.2 に示す．$\overline{\boldsymbol{e}}, \overline{\Delta \boldsymbol{w}}$ はそれぞれ $\boldsymbol{e}, \Delta \boldsymbol{w}$ の履歴を表す．また $\alpha_w, \alpha_v > 0$ はそれぞれ actor と critic の学習率である．

4. 数値実験

4.1 倒立振子制御問題

提案手法と既存手法を比較するため，倒立振子制御問題（図 1）に適用した．この問題は，台車に左右方向の力 F を加え

Algorithm 2 提案手法

$\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{w}_t, \overline{\boldsymbol{e}}_0, \overline{\Delta \boldsymbol{w}}_0$ を初期化する．
$t \leftarrow 1$
loop
　環境から状態 $\boldsymbol{s}_t$ を観測する．
　方策 $\pi(\boldsymbol{a}_t, \boldsymbol{s}_t; \boldsymbol{w}_t)$ に基づいて行動 $\boldsymbol{a}_t$ を実行する．
　環境から報酬 r_t と次の状態 $\boldsymbol{s}_{t+1}$ を観測する．
　Critic は以下のように actor への強化信号 δ_t を計算する．
　　$\delta_t = r_t + \gamma \hat{V}(\boldsymbol{s}_{t+1}; \boldsymbol{v}_t) - \hat{V}(\boldsymbol{s}_t; \boldsymbol{v}_t)$
　Actor は以下のように方策を更新する．
　　$\boldsymbol{e}_t = \frac{\partial}{\partial \boldsymbol{w}_t} \log \pi(\boldsymbol{a}_t, \boldsymbol{s}_t; \boldsymbol{w}_t)$
　　$\overline{\boldsymbol{e}}_t = \boldsymbol{e}_t + \beta \overline{\boldsymbol{e}}_{t-1}$
　　$\Delta \boldsymbol{w}_t = \delta_t \overline{\boldsymbol{e}}_t$
　　$\overline{\Delta \boldsymbol{w}}_t = \beta \overline{\Delta \boldsymbol{w}}_{t-1} + \Delta \boldsymbol{w}_t$
　　$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \alpha_w \Delta \boldsymbol{w}_t$
　Critic は以下のように価値関数を更新する．
　　$C_t = \frac{\partial}{\partial \boldsymbol{v}_t} \hat{V}(\boldsymbol{s}_{t+1}; \boldsymbol{v}_t) \boldsymbol{e}_{t+1}^{\mathrm{T}}$
　　$\overline{C}_t = \beta \overline{C}_{t-1} + \overline{C}_t$
　　$\Delta \boldsymbol{v}_t = \beta \overline{C}_t \overline{\Delta \boldsymbol{w}}_{t-1} + \overline{C}_t \Delta \boldsymbol{w}_t$
　　$\boldsymbol{v}_{t+1} = \boldsymbol{v}_t + \alpha_v \Delta \boldsymbol{v}_t$
　$t \leftarrow t + 1$
end loop

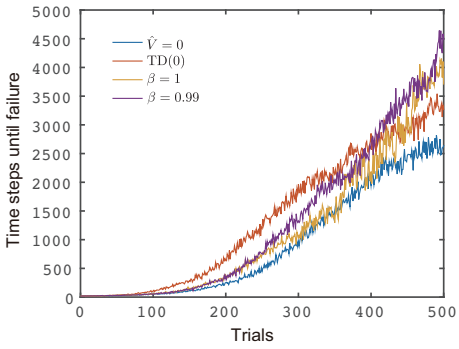

図 2: 倒立を維持したステップ数の変化（それぞれ 100 試行平均）．

ることでヒンジで固定されたポールの角度を間接的に制御し，倒立状態を保つという問題であり，[Barto 83] の実験設定を元に，[Kimura 98] において行動を連続値とするために修正されたものである．

本実験では台車の質量を $M = 1.0$(kg)，ポールの質量 $m = 0.1$(kg)，ポールの長さ $l = 1$(m)，台車の摩擦係数 $\mu_c = 0.0005$，ヒンジの摩擦係数 $\mu_p = 0.000002$ とし，$\Delta t = 0.02$(sec) の時間ステップで近似計算した．エージェントは台車の位置 x，速度 $\dot{x}$，ポールの角度 θ，角速度 $\dot{\theta}$ を状態として観測し，台車に加える力 F を行動として出力する．ただし $-20 \leq F \leq 20$ の範囲を超えた分は無視される．エピソード毎に，θ は ± 1(deg) の範囲から一様分布で初期化され，他の変数は 0 で初期化される．$-1 \leq \theta \leq 1$(deg) の範囲または $-2.4 \leq x \leq 2.4$(m) の範囲を逸脱すると報酬 -1 が与えられてエピソードが終了する．

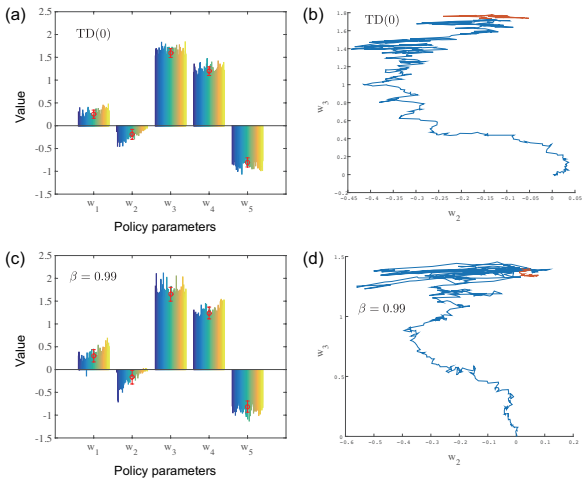

図 3: (a,c) 獲得された方策のパラメータ．それぞれの棒グラフは 100 試行を最終エピソードのステップ数で左から昇順に並べたものである．赤い丸とエラーバーはそれぞれ 100 試行の平均と標準偏差．(b,d) パラメータ空間の部分空間上での学習の軌跡（それぞれ一例）．赤い部分は最後の 50 エピソードに対応する．

4.2 エージェントの実装

Actor の行動を生成する確率的方策 π は，次式のように正規分布を用いて設定した．

$$\pi(F, \boldsymbol{s}; \boldsymbol{w}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(F - \boldsymbol{w}_\mu^{\mathrm{T}} \boldsymbol{s}\right)^2\right) \qquad (9)$$
$$\sigma = (1 + \exp(-w_\sigma))^{-1},$$

ここで，上付きの T は行列の転置を表す．また $\boldsymbol{w} = \{\boldsymbol{w}_\mu, w_\sigma\}$ は学習によって更新されるパラメータである．

Critic は状態の空間を格子状に離散化し，それぞれの格子に対する状態価値を学習する．本実験では $x, \dot{x}, \theta, \dot{\theta}$ をそれぞれ ± 2.4(m), ± 0.2(m/s), ± 1(deg), ± 0.5(rad/s) の範囲で 5 分割した．$\boldsymbol{s} = (x, \dot{x}, \theta, \dot{\theta})^{\mathrm{T}}$ が各軸でそれぞれ i, j, k, l 番目の格子上に存在するときに 1，それ以外のときに 0 となる $\boldsymbol{s}$ の関数を $u_{ijkl}(\boldsymbol{s})$ とすると，価値関数は次式のように表せる．

$$\hat{V}(\boldsymbol{s}; \boldsymbol{v}) = \sum_{i,j,k,l=1}^{5} v_{ijkl} u_{ijkl}(\boldsymbol{s}) \qquad (10)$$

ここで，$\boldsymbol{v} = \{v_{ijkl}\}_{i,j,k,l=1}^{5}$ は学習によって更新されるパラメータで，各格子に対応する状態の価値を表す．以上のように構成したエージェントで Alg.2 を実行した．

4.3 結果

エージェントが倒立を維持できたステップ数を図 2 に示す．$\hat{V}$ を 0 に固定した場合に比べて TD(0) によって $\hat{V}$ を学習し

た場合のほうが改善がみられるが，提案手法の場合，400 エピソード以降でさらなる改善がみられる．一方，400 エピソード以前では TD(0) のほうが長時間にわたって倒立を維持できていることがわかる．これは，学習初期においては 1 エピソードが短いため，提案手法で多用されているトレースが十分蓄積する前に初期化されることで，提案手法の利点が生かされていないためと考えられる．学習が進み，一つのエピソードが長くなるにつれて，トレースが十分蓄積し，学習効率が改善するという正帰還が働いていることが示唆される．

図 3(a,c) は学習によって獲得された方策のパラメータを示している．ここで，$(w_1, w_2, w_3, w_4)^{\mathrm{T}} = \boldsymbol{w}_\mu$，$w_5 = w_\sigma$ である．各棒グラフでは，最終エピソードにおいて倒立振子を長時間維持できた試行を右に配置している．これを踏まえると，例えば w_2 の最適値は，少なくとも本実験でいえる範囲では，0 付近であることがわかる．また図 3(b,d) は学習中のパラメータの変化の一例を示している．提案手法 (d) では，学習の後半において大きく振動しているが，最後の 50 エピソードにおいて最適値と思われる $w_2 = 0$ に近づくと比較的収束する傾向がみられた．一方，TD(0) の場合 (b) では，より小さい値を中心とする揺動がみられた．これは，提案手法による actor の更新則が他の手法よりも真の勾配に近いため，極大値付近に収束することができたためと考えられる．しかしながら，図 3(d) にみられる後半の大きな振動や試行間のばらつきなどに関しては，さらなる分析が必要である．

5. おわりに

Actor-critic アルゴリズムにおいて，actor の学習に適正度の履歴を用いることで，critic の推定する価値関数が不正確であっても actor が方策を獲得できることが示されている．本研究では，actor による方策の更新則と actor が学習すべき期待報酬の真の勾配との差を明示的に最小化するための critic の学習則を提案した．また，提案手法を倒立振子制御問題に適応し，最適方策を獲得する速度が向上することを示した．

冒頭で述べた通り，遅延報酬を解決する手法として TD 法と適正度の履歴がよく知られており，木村らの手法は actor の学習に非マルコフ的な環境に対して頑健な後者を用いたものである．本研究ではさらに，critic の学習にも累積和による履歴を用い，環境のマルコフ性に依存しない手法を提案した．今後は前節で述べた学習の安定性の問題について改善するとともに，非マルコフ的な環境に適用することで提案手法の特性について検証する．

謝辞

本研究の遂行にあたり，科学研究補助金 (課題番号 24000012) の補助を受けた．

参考文献

[Barto 83] Barto, A., Sutton, R., and Anderson, C.: Neuronlike adaptive elements that can solve difficult learning control problems, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-13, No. 5, pp. 835–846 (1983)

[Kimura 98] Kimura, H. and Kobayashi, S.: An analysis of actor/critic algorithms using eligibility traces: reinforcement learning with imperfect value function, in *Proceedings of the 15th International Conference on Machine Learning*, pp. 278–286 (1998)

[Pendrith 96] Pendrith, M. D. and Ryan, M. R. K.: Actual return reinforcement learning versus Temporal Differences: Some theoretical and experimental results, in *Proceedings of the 13th International Conference on Machine Learning*, pp. 373–381 (1996)

[Singh 96] Singh, S. P. and Sutton, R. S.: Reinforcement Learning with Replacing Eligibility Traces, *Machine Learning*, Vol. 22, pp. 123–158 (1996)

[Sutton 90] Sutton, R. S. and Barto, A. G.: Time-Derivative Models of Pavlovian Reinforcement, in Moore, J. W. and Gabriel, M. eds., *Learning and Computational Neuroscience*, pp. 497–537, MIT Press, Cambridge, MA (1990)

[Williams 92] Williams, R. J.: Simple Statistical Gradient Following Algorithms for Connectionist Reinforcement Learning, *Machine Learning*, Vol. 8, pp. 229–256 (1992)

[木村 00] 木村 元, 小林 重信：Actor に適正度の履歴を用いた Actor-Critic アルゴリズム—不完全な Value-Function のもとでの強化学習—, 人工知能学会誌, Vol. 15, No. 2, pp. 267–275 (2000)