論 文

Technical Papers

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—

An Analysis of Actor-Critic Algorithms Using Eligibility Traces:

Reinforcement Learning with Imperfect Value Functions

木村 元
Hajime Kimura
東京工業大学大学院総合理工学研究科
Graduate School of Interdisciplinary Science and Engineering, Tokyo Institute of Technology.
gen@fe.dis.titech.ac.jp

小林 重信
Shigenobu Kobayashi
(同 上)
kobayasi@.dis.titech.ac.jp

Keywords: reinforcement learning, actor-critic, eligibility trace, value function approximation.

Summary

We present an analysis of actor-critic algorithms, in which the actor updates its policy using eligibility traces of the policy parameters. Most of the theoretical results for eligibility traces have been for only critic's value iteration algorithms. This paper investigates what the actor's eligibility trace does. The results show that the algorithm is an extension of Williams' REINFORCE algorithms for infinite horizon reinforcement tasks, and then the critic provides an appropriate reinforcement baseline for the actor. Thanks to the actor's eligibility trace, the actor improves its policy by using a gradient of actual return, not by using a gradient of the estimated return in the critic. It enables the agent to learn a fairly good policy under the condition that the approximated value function in the critic is hopelessly inaccurate for conventional actor-critic algorithms. Also, if an accurate value function is estimated by the critic, the actor's learning is dramatically accelerated in our test cases. The behavior of the algorithm is demonstrated through simulations of a linear quadratic control problem and a pole balancing problem.

1. はじめに

Actor-critic強化学習アルゴリズムは，政策反復法 (policy iteration) の一種であると言われている [Kaelbling 96]．政策反復法は状態評価と政策改善を交互に繰り返す．Actorは，状態から行動への確率分布である“確率的政策 (stochastic policy)”に従って行動を実行する．Criticは，actorの政策下における各状態の“価値 (value)”(利得 (return) の期待値) を推定する．Criticで計算されるTD (temporal difference) またはTD-errorと呼ばれる値を手がかりにしてactorは政策を改善する．Criticによる状態の評価を待っていては時間がかかりすぎるので，actorの政策改善とcriticの政策評価は同時に実行される場合が多い．

Actor-criticアルゴリズムは今まで様々な強化学習タスクへ適用され，有用性が示されてきた．例えばASE/ACE [Barto 83, Gullapalli 92] による倒立振子制御やRFALCON[Lin 96] による倒立振子および梁上を転がるボールの制御，[Doya 96] による倒立振子の振上げ制御などがある．Actor-criticアルゴリズムの解析については [Williams 90] や [Gullapalli 92] などがあるが，Q-learning[Watkins 92] に代表されるvalue-iteration法ほど多くはなされていない．しかしQ-learning等と比較すると以下の実用的利点がある．

- 状態のvalueだけについて推定すればよいため，連続値を含むような行動出力への拡張がQ-learningと比較して非常に容易である [Sutton 98]．
- Actorにおいて確率的政策を用いることからPOMDPの環境 [Jaakkola 94, Singh 94] やマルチプレイヤーゲームの環境 [Littman 94] などへの適用も可能である．
- Actorの部分では状態観測に対する行動出力を学習するため，従来の教師付き学習との組み合わせが容易である．エキスパートの知識は，状態観測に対する行動出力である場合が多い．そのためエキスパートの知識との統合が容易である [Clouse 92]．

適正度の履歴 (eligibility trace) は，報酬の遅れに対処するための基本的メカニズムとして広く用いられて来た [Singh 96]．また，適正度の履歴は非マルコフ性に対処するためにも用いられる [Pendrith 96, Sutton 95]．ASE/ACE [Barto 83] では actor と critic の両方において適正度の履歴を用いている．Critic における適正度の履歴に関する理論は，TD(λ) [Sutton 88] の解析として多くの研究結果が示されてきた．だが actor-critic アルゴリズムにおける actor の適正度の履歴についての解析は皆無であった．本論文では actor が政策パラメータの適正度の履歴を用いて政策を改善する actor-critic アルゴリズムを提案し，その政策改善の方向について考察する．

2. 準備：割引報酬による状態評価

制御対象である“環境”は下記のマルコフ決定過程 (MDP) でモデル化できると仮定する．MDP は，状態集合 $\mathcal{S}$，選択可能な行動の集合 $\mathcal{A}$，状態遷移関数 $T:\mathcal{S}\times\mathcal{A}\to\Pi(\mathcal{S})$ ただし $\Pi(\mathcal{S})$ は状態空間 S において定義される確率分布，報酬関数 $R:\mathcal{S}\times\mathcal{A}\to\mathcal{R}$ より構成される．MDP の状態，行動，報酬は各時刻 $t\in\{0,1,2,\cdots\}$ においてそれぞれ $s_t\in\mathcal{S}$，$a_t\in\mathcal{A}$，$r_t\in\mathcal{R}$ で表される．状態遷移規則は，状態遷移関数 T によって確率分布 $T(s,a,s')=\Pr\{s_{t+1}=s'|s_t=s,a_t=a\}$ で表され，報酬は報酬関数 R によって期待値 $R(s,a)=E\{r_t|s_t=s,a_t=a\}$ で表される．ただし r_t は全時刻において有界であるものとする．

学習の主体“エージェント”について，政策関数 $\pi:\mathcal{S}\to\Pi(\mathcal{A})$ を定義する．これは状態空間から行動空間 $\mathcal{A}$ に定義される確率分布への写像関数である．エージェントの意志決定は，各時刻 t ごとに定義される政策関数 $\pi(t)$ に従った確率 $\pi(a,s,t)=\Pr\{a_t=a|s_t=s,\pi(t)\}$ で行動 a_t を選択する．エージェントと環境は以下のやりとりを行う．

(1) 時刻 t において，エージェントは環境の状態 s_t を観測し，政策関数 $\pi(a,s,t)$ の確率分布に従って行動 a_t を実行する．

(2) 行動 a_t により，環境は $T(s_t,a_t,s')$ の状態遷移確率に従って s_{t+1} へ状態遷移し，期待値 $R(s_t,a_t)$ に従い報酬 r_t をエージェントへ与える．

(3) 時刻 t を $t+1$ に進めてステップ1へ戻る．

エージェントは一般に，環境の状態遷移規則 $T(s,a,s')$ や報酬 $R(s,a)$ に関する知識をあらかじめ持っていない．

強化学習の目的は，エージェントの利得 (return) を最大化すること，およびそのための政策を見つけることである．利得の評価方法として，次式の割引報酬の合計を用いる．

$$V_t=\sum_{k=0}^{\infty}\gamma^k r_{t+k} \tag{1}$$

ただし，割引率 $0\le\gamma<1$ は時間 t に対する未来の報酬の重み付けを行うパラメータである．

エージェントの政策が時間とともに変化しない場合，つまり $\pi(t)=\pi$ (ただし，$t=0,1,2\cdots$) のとき，π は定常政策と呼ばれる．このときの行動選択確率を $\pi(a,s)=\Pr\{a_t=a|s_t=s,\pi\}$ と表す．MDP では利得の期待値は全状態 s について以下に定義される．

$$\begin{aligned}V^\pi(s)&=E\left\{\sum_{t=0}^{\infty}\gamma^t r_t|s_0=s,\pi\right\}\\&=\sum_{a\in\mathcal{A}}\pi(a,s)\Big(R(s,a)+\gamma\sum_{s'\in\mathcal{S}}T(s,a,s')V^\pi(s')\Big)\end{aligned} \tag{2}$$

$V^\pi(s)$ は value 関数と呼ばれ，状態 s の評価値を示す．全ての状態 $s\in\mathcal{S}$ において，$V^\pi(s)\le V^{\pi'}(s)$ となるとき，定常政策 π は π' より優れているという．MDP では，他のどんな政策よりも優れた，あるいは同等な政策が少なくとも一つ存在し，これが最適政策である [小笠原 67]．

ある政策 π_n において全状態について $V^{\pi_n}(s)$ を求めた後，全状態 s について式 (3) を満たすように政策 π_n を π_{n+1} に改善し，再び同様の value の計算と政策改善を繰り返していく手法は，政策反復法 (Policy Iteration) と呼ばれる [Sutton 98]．

$$\begin{aligned}&\sum_{a\in\mathcal{A}}\pi_{n+1}(a,s)\Big(R(s,a)\\&+\gamma\sum_{s'\in\mathcal{A}}T(s,a,s')V^{\pi_n}(s')\Big)-V^{\pi_n}(s)\\&\le 0\end{aligned} \tag{3}$$

状態遷移規則 $T(s,a,s')$ や報酬 $R(s,a)$ が既知ならば，最適政策への収束が保証されるが，強化学習ではこれらは予め知ることはできないため，次に紹介する actor-critic によって政策改善が行われる．

3. Actor-Critic アルゴリズム

図1と図2に一般的な actor-critic アルゴリズムの概要を示す [Crites 94, Sutton 90]．Actor-critic アルゴリズムは，政策反復法における value の計算を critic による value の推定に置き換え，さらに式 (3) による政策改善の判定を，TD-error という確率変数を用いた判定に置き換えたものと考えられる．その根拠は，critic の $\hat{V}=V$ のとき，TD-error の期待値は式 (3) の左辺に等しいことによる．よって actor-critic アルゴリズムは，critic によって推定された value 関数を用いて，value を増加させる方向へ政策を改善するものと考えられる．

Actor の政策表現と政策改善方法にはバリエーションがあるが，critic のアルゴリズムとしては多くの場合 TD 法が用いられる．Actor-critic は以下の2点に特徴がある：第一に actor は確率的政策を用いる，第二に actor は TD-error を用いて政策を改善する点である．本論文では特に actor のアルゴリズムに注目する．

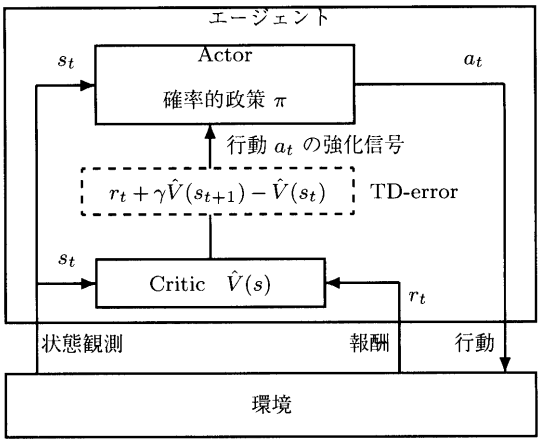


図 1 一般的な actor-critic アルゴリズムの構成

(1) エージェントは環境において状態 s_t を観測する.
Actor は，確率的政策 $\pi(t)$ に従って行動 a_t を実行する.
(2) Critic は報酬 r_t を受け取り，次の状態 s_{t+1} を観測し Actor への強化信号として以下の TD-error を計算する.

$$(\text{TD-error}) = \left[r_t + \gamma \hat{V}(s_{t+1})\right] - \hat{V}(s_t)$$

γ $(0 \le \gamma \le 1)$ は割引率,
$\hat{V}(s)$ は Critic が推定した割引報酬の期待値を表す.
(3) TD-error を用いて actor の行動選択確率を更新する.
(TD-error) > 0 ならば，実行した行動 a_t は比較的好ましいと考えられるので，この選択確率を増やす.
逆に (TD-error) < 0 ならば，実行した行動 a_t は比較的好ましくないと考えられるので，この選択確率を減らす.
(4) TD 法等を用いて critic の value の推定値を更新する.
例えば TD(0) ならば以下のように計算する.
$\hat{V}(s_t) \leftarrow \hat{V}(s_t) + \alpha(\text{TD-error})$, ただし α は学習率である.
(5) ステップ (1) から繰り返す.

図 2 一般的な actor-critic アルゴリズムの処理手続き

4. Actor へ適正度の履歴を付加

4·1 確率的政策のパラメータ表現

パラメータベクトル $W \in \mathcal{R}^\ell$ を定義する．これは各時刻 $t \in \{0,1,2,\cdots\}$ において値 $W(t)$ をとるものとする．本論文では，$W(t)$ を用いて政策関数 $\pi(t)$ を $\pi(a,s,t) = \pi(a,s,W(t))$ のようにパラメータ表現する．エージェントは W を調節することにより政策 π を変えることができる．エージェントの行動選択確率を表す機構が，例えばニューラルネットならば，W はリンクの重み変数に相当し，重み付きのルールベースシステムならば，W はルールの重みに相当する．$\pi(a,s,W)$ の具体的な関数形については，エージェントに実装できる計算資源の制限など，一般に個別の問題ごとに制約が存在する．すなわちエージェントの構造および制約条件は $\pi(a,s,W)$ の関数形で規定される．W をパラメータとした分布関数として政策 π を記述することの利点は，上記のように様々な行動選択機構を持つエージェント全てに対して共通の理論的基礎を与えられることである．また行動 a の集合が連続値の場合は分布関数 $\pi(a,s,W)$ を確率密度関数とすれば，行動が離散の場合と全く同様に扱える．

4·2 アルゴリズムの詳細

(1) エージェントは環境において状態 s_t を観測する.
Actor は確率的政策 $\pi(a_t,s_t,W(t))$ により行動 a_t を実行
(2) Critic は報酬 r_t を受け取り，次の状態 s_{t+1} を観測し actor への強化信号として以下の TD-error を計算する.

$$(\text{TD-error}) = \left[r_t + \gamma \hat{V}_t(s_{t+1})\right] - \hat{V}_{t-1}(s_t) \quad (4)$$

γ $(0 \le \gamma \le 1)$ は割引率,
$\hat{V}_t(s)$ は Critic が推定した割引報酬の期待値を表す.
(3) TD-error を用いて actor の行動選択確率を更新する.

$$\begin{aligned} e_i(t) &= \frac{\partial}{\partial w_i(t)} \ln\Big(\pi(a_t,s_t,W(t))\Big) \\ D_i(t) &= e_i(t) + \beta D_i(t-1)\ , \\ \Delta w_i(t) &= (\text{TD-error})\, D_i(t) \\ W(t+1) &\leftarrow W(t) + \alpha_p \Delta W(t) \end{aligned} \quad (5)$$

ただし，$e_i(t)$ は適正度，$D_i(t)$ は適正度の履歴,
$W(t) = (w_1(t), w_2(t), \cdots, w_\ell(t))$ は政策パラメータ,
β $(0 \le \beta < 1)$ は適正度の履歴の割引率,
α_p は actor の学習定数を表す.
(4) TD 法等を用いて critic の value の推定値を更新する.
例えば TD(0) ならば以下のように計算する.
$\hat{V}_{t+1}(s_t) \leftarrow \hat{V}_t(s_t) + \alpha(\text{TD-error})$, ただし α は学習率.
$\hat{V}_{t+1}(s) \leftarrow \hat{V}_t(s)$, $\forall s \ne s_t$
(5) 時刻 t を 1 つ進めてステップ (1) から繰り返す.

図 3 Actor に適正度の履歴を用いた actor-critic アルゴリズム

図 3 に本論文が対象とする actor-critic の詳細を示す．ASE/ACE [Barto 83] はこのアルゴリズムの具体例の一つと考えられる．処理ステップ (3) に示されている actor の適正度は，REINFORCE アルゴリズム [Williams 92] で定義されている物と同一である．適正度 $e_i(t)$ は政策関数 π のパラメータ $w_i(t)$ と実行した行動 a_t の相関を表している．$D_i(t)$ は適正度の履歴 (eligibility trace) を表し，適正度を割引率 β で割引きながら足し合わせることにより，今まで実行してきた行動の履歴に関する情報を圧縮して保持する．Actor は正の TD-error を受け取ると，図 3 の手順 (3) の処理によって，今まで実行してきた全行動の選択確率を高めるように W を更新する．ただし過去に実行した行動ほど強化の大きさは割引率 β によって割引かれる．従来の actor-critic(図 2) では，時間 t の TD-error は行動 a_t だけの強化に用いられていたのに対し，図 3 の手法では，時間 t の TD-error が行動 a_t の強化だけでなく，今まで実行してきた行動全て $a_{t-1}, a_{t-2}, \cdots$ の強化に用いられている．このことは，図 2 の actor-critic を式 (3) の代わりに TD-error を用いた政策反復法の一種とする考え方からは無意味に思えるが，非常に興味深い特徴を持つことを後に示す．

適正度の履歴の割引率 $\beta = 0$ の場合，本手法は完全に図 2 の actor-critic の一種になる．また，$\beta = \gamma$ で，かつ

criticが出力する割引報酬の推定値 $\hat{V}(s)$ が全ての状態 s について任意の定数 b のとき，図3のアルゴリズムは確率的傾斜法 [木村 96, Kimura 97] と同一になる．

Actorは，政策を表現するための実数パラメータベクトル W および適正度の履歴を保持するための D_i に相当するメモリ容量を必要とする．D_i に要するメモリ量は W のそれと同一である．

4·3 提案手法の解釈

Actorの割引率 $\beta=\gamma$ すなわちvalue関数の割引率に等しい場合，図3における $\Delta w_i(t)$ の合計は以下のように計算できる．

$$\begin{aligned}
&\sum_{t=0}^{\infty}\Delta w_i(t)\\
&=\sum_{t=0}^{\infty}\left(r_t+\gamma\hat{V}_t(s_{t+1})-\hat{V}_{t-1}(s_t)\right)D_i(t)\\
&=\sum_{t=0}^{\infty}\left(r_t+\gamma\hat{V}_t(s_{t+1})-\hat{V}_{t-1}(s_t)\right)\sum_{\tau=0}^{t}\gamma^{t-\tau}e_i(\tau)\\
&=\sum_{t=0}^{\infty}e_i(t)\sum_{\tau=t}^{\infty}\gamma^{\tau-t}(r_\tau\\
&\quad+\gamma\hat{V}_\tau(s_{\tau+1})-\hat{V}_{\tau-1}(s_\tau)\Big)\\
&=\sum_{t=0}^{\infty}e_i(t)\left(\left(\sum_{\tau=t}^{\infty}\gamma^{\tau-t}r_\tau\right)-\hat{V}_{t-1}(s_t)\right) \qquad (6)\\
&=\sum_{t=0}^{\infty}e_i(t)\left(V_t-\hat{V}_{t-1}(s_t)\right) \qquad (7)
\end{aligned}$$

ただし $t<0$ のとき $D_i(t)=0$，かつ $\hat{V}_t$ は任意の有界な定数とする．$\hat{V}_t(s)$ は全ての t,s において有界とする．式(7)は式(6)へ式(1)を適用して得る．式(7)中の $e_i(t)(V_t-\hat{V}_{t-1}(s_t))$ は，時刻 t に実行した行動 a_t をどのように強化するのかを示している．

REINFORCEアルゴリズムの定理 [Williams 92] によれば，行動 a_t の選択確率が w をパラメータとした関数 g により $\Pr\{a_t|w\}=g(a_t,w)$ のように表されるシステムにおいて，評価値 r が w によって $E\{r|w\}$ のように与えられる場合，以下が成り立つ．

$$E\{(r-b)\frac{\partial}{\partial w}\ln g(a_t,w)\}=\frac{\partial}{\partial w}E\{r|w\} \qquad (8)$$

ただし b は a_t および r とは条件付き独立な値なら(定数でも確率変数でも)なんでも良く，評価値 r は直接報酬以外の割引報酬や平均報酬でも成り立つ．すると式(7)の $e_i(t)(V_t-\hat{V}_{t-1}(s_t))$ は，行動 a_t の適正度 $e_i(t)$ を式(8)の $(\partial/\partial w)\ln g(a_t,w)$ に，割引報酬合計 V_t を式(8)の評価値 r に，criticの推定値 $\hat{V}_{t-1}(s_t)$ を式(8)の b に対応させることができる．なぜならcriticの推定値 $\hat{V}_{t-1}(s_t)$ はアルゴリズムの処理手順上，行動 a_t や割引報酬合計 V_t を求めるより前に確定しているので，明らかに a_t や V_t とは条件付独立だからである．よって図3のアルゴリズムは $\beta=\gamma$ のとき実際の利得 V_t の増加が最大となる方向へ，時刻 t に実行した行動 a_t を確率的に強化するように $W(t)$ を更新していると考えられる．注目すべき点は，criticが明示的に推定したvalue関数 $\hat{V}$ は更新の期待値に対して影響を与えていないが，更新の分散の大きさに対して影響を与えている点である．そのため，criticがvalue関数を学習できるかどうかに依存せずに，actorは政策を改善できると考えられる．Criticがactorの政策更新の分散に影響を与える根拠を以下の例で説明する．状態 s_t で行動 a_1 または a_2 を選択でき，a_1 を実行すると割引報酬10，a_2 を実行すると割引報酬8である場合を考える．$\pi(a_1,s_t)=\pi(a_2,s_t)=0.5$ のとき，$V^\pi(s_t)=9$ となる．もしcriticが $\hat{V}(s_t)=9$ を獲得していれば，actorが a_1 をとるなら式(8)中の $V_t-\hat{V}(s_t)$ すなわち $r-b$ は $10-9=1$ で必ずプラス，a_2 をとるなら $V_t-\hat{V}(s_t)=8-9=-1$ で必ずマイナスの値をとり，どちらの行動をとっても確実に正しい方の行動を強化できる．ところがcriticがvalueの推定に失敗し，例えば $\hat{V}(s_t)=0$ の場合，actorが a_1，a_2 どちらの行動をとっても $V_t-\hat{V}(s_t)$ は必ずプラスの値になるので，正しい行動を強化する確率は0.5になる．このように $\hat{V}$ の値によって重み更新の分散は大きく異なり，学習速度に影響するが，重み更新の期待値は同じなのでどちらの場合も学習は可能である．また政策が最適に近付き，かつ $\hat{V}$ も最適valueに近付くと，式(8)の $V_t-\hat{V}(s_t)$ も期待値がゼロに近付く．以上よりcriticはactorの政策更新のステップ幅を適応的に制御する役割を果たしていると考えられる．

上記の考察は，割引率が $\beta=\gamma$ の場合だけだが，β を別の値 $(0\le\beta\le\gamma)$ とすることにより，政策の更新をする際にcriticの推定したvalue関数の勾配を登るか，それともトレーニング系列の割引報酬の勾配を登るのかの度合を調節できる．$\beta=0$ のときは図2に示した一般的なactor-criticと同じなので，criticの推定したvalue関数の勾配を登る．$0<\beta<\gamma$ のときは，おおよそ中間的な登りかたをすると考えられる．

5. 実　　験

本章では，解析から予想された適正度の履歴の効果を確認するため，提案手法を単純な線形制御問題に適用した計算機シミュレーションを示す．

5·1 実験設定: 線形2次形式制御問題

ベンチマークとして以下の線形2次形式制御問題を考える [Baird 94]．ある時間ステップ t において，環境の状態はある連続値の実数 x_t にある．エージェントは同じく連続値の実数 a_t で表される行動を選択する．環境の状態遷移は以下で与える．

$$x_{t+1}=x_t+a_t+noise \qquad (9)$$

ただし，$noise$ は標準偏差 0.5，平均 0 の正規分布で与える．直接報酬は以下のように与える．

$$r_t = -x_t^2 - a_t^2 \tag{10}$$

エージェントの学習目標は，初期状態より計算される以下の割引報酬合計を最大化することである．

$$\sum_{t=0}^{\infty} \gamma^t r_t \tag{11}$$

本問題は線形 2 次形式制御問題であることより，最適な制御規則を解析的に求めることができる．Riccati 方程式より，最適レギュレータは以下の状態フィードバックで与えられる．

$$a_t = -k_1 x_t, \quad \text{ただし}$$
$$k_1 = 1 - \frac{2}{1+2\gamma+\sqrt{4\gamma^2+1}} \tag{12}$$

最適な value 関数は $V^*(x_t) = -k_2 x_t^2$ ただし k_2 は何らかの正の定数の形式で与えられる．本実験では，遷移可能な状態空間は $[-4,4]$ に制限する．式 (9) で示される状態遷移が上記の範囲を超える場合は，制限範囲までしか移動できないものとする．エージェントの行動についても同様に $[-4,4]$ の範囲外の行動を実行しても，環境では制限の範囲でしか実行されないとした．

5·2 エージェントの実装

§1 Actor の実装

行動が連続値の場合，政策 $\pi(a,s,W)$ は確率密度関数となることは **4·1 節**にてすでに述べた．正規分布は平均値 μ と標準偏差 σ の二つのパラメータを持つ．政策 π が式 (13) に示す正規分布で与えられた場合，パラメータ μ と σ に関する適正度は以下のように計算される．

$$\pi(a,s,W) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(a-\mu)^2}{2\sigma^2}\right) \tag{13}$$

$$e_\mu = \frac{a_t - \mu}{\sigma^2} \tag{14}$$

$$e_\sigma = \frac{(a_t-\mu)^2 - \sigma^2}{\sigma^3} \tag{15}$$

このような行動選択メカニズムは Gaussian unit と呼ばれ，ランダムな探査の度合いを自ら制御する特徴を有する [Williams 92]．ここで式 (14),(15) においてパラメータ σ が分母となっていることより，σ が 0 へ近付くと適正度が発散することに注意しなければならない．適正度の発散はアルゴリズムの動作に悪い影響を及ぼす．この問題に対処するための一つの方法として，σ の値に応じて政策パラメータの更新のステップ幅を制御することが考えられる．更新のステップ幅を σ^2 に比例させると，適正度は以下のように計算される．

$$e_\mu = a_t - \mu, \quad e_\sigma = \frac{(a_t-\mu)^2 - \sigma^2}{\sigma} \tag{16}$$

Actor はまず μ と σ を計算し，正規分布に従って行動出力を決定する．エージェントは二つの内部変数 w_1, w_2 を持ち，これを用いて μ と σ を以下のように計算する．

$$\mu = w_1 x_t, \quad \sigma = \frac{1}{1+\exp(-w_2)} \tag{17}$$

このとき，w_1 はフィードバックゲインとして見ることができる．σ を上記のように計算するのは，σ の値が負になるのを防ぐためである．エージェント内部変数 w_1 と w_2 に対応する適正度をそれぞれ e_1, e_2 と表す．式 (16) より，適正度 e_1, e_2 は以下に与えられる．

$$e_1 = e_\mu \frac{\partial \mu}{\partial w_1} = (a_t - \mu) x_t$$
$$e_2 = e_\sigma \frac{\partial \sigma}{\partial w_2} = ((a_t-\mu)^2 - \sigma^2)(1-\sigma)$$

エージェントのパラメータは，学習係数 $\alpha = 0.001$, w_1 の初期値は 0.35 ± 0.15 の範囲内でランダムに初期化，w_2 は 0 すなわち $\sigma = 0.5$ に初期化した．

§2 Critic の実装

Critic では状態観測の空間を格子状に分割離散化し，それぞれのグリッドに対して value を学習する．本実験では $-4 \le x \le 4$ の状態空間を 3 等分した場合と 10 等分した場合について示す．Critic は TD(0) 法を用いて value 関数を推定する．Critic では観測入力 x_t に対する $V(x_t)$ を推定する．TD(0) の学習率は 0.2 とした．

§3 実験結果

図 4, 図 5, 図 6, 図 7, 図 8 は LQR 問題において割引率 $\gamma = 0.9$ の場合における各アルゴリズムで得られたフィードバックゲインの 100 試行の結果を示す．

図 4 は状態空間を 3 等分した critic と $\beta = 0$ つまり適正度の履歴を用いない actor による学習の様子を示す．図 6 は状態空間を 10 等分した critic と $\beta = 0$ つまり適正度の履歴を用いない actor による学習の様子を示す．図 6 のアルゴリズムは最適なフィードバックゲイン付近に収束している．ところが，図 4 では全く学習できない．これより，状態空間を 3 分割して value を表現する critic では，適正度の履歴なしで政策を学習するには関数近似能力が不十分であるためであることは明らかである．

図 5 は状態空間を 3 等分した critic と $\beta = \gamma = 0.9$ つまり適正度の履歴を使った actor による学習の様子を示す．学習の速さと収束した解の質の両方において，図 4 や図 5 より良い結果を示している．この性能向上には actor の適正度の履歴が関与していることは明らかである．学習が速くなった理由は，適正度の履歴が情報の伝搬を加速しているためと考えられる．解の質が向上した理由は，**4·3 節**の解析で示したように actor が実際の利得の勾配を用いて政策を改善しているためだと考えられる．そのため，得られた政策の平均値に関して，actor に適正度の履歴を用いた場合は critic の性能には依存しない．この特徴は解の分散が大きい図 8 からも観察できる．図 9 は式 (10), で定義される value 関数を μ と σ で張られるパ

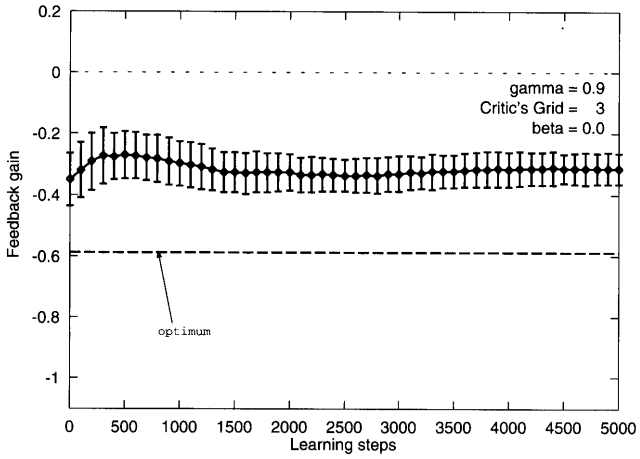


図 4 $\gamma = 0.9$, Critic は 3 分割グリッドの関数近似．Actor に適正度の履歴を用いない場合の 100 試行の平均と分散

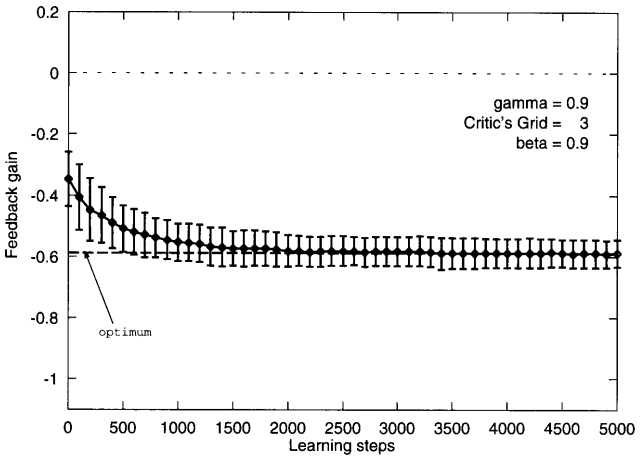


図 5 $\gamma = 0.9$, Critic は 3 分割グリッドの関数近似．Actor に適正度の履歴を使用した場合の 100 試行の平均と分散

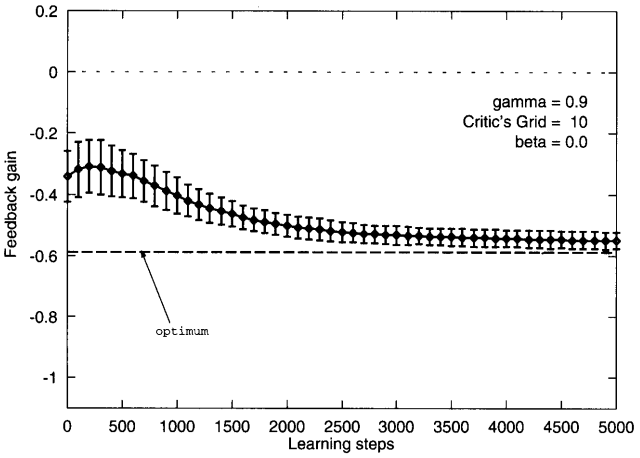


図 6 $\gamma = 0.9$, Critic は 10 分割グリッドの関数近似．Actor に適正度の履歴を用いない場合の 100 試行の平均と分散

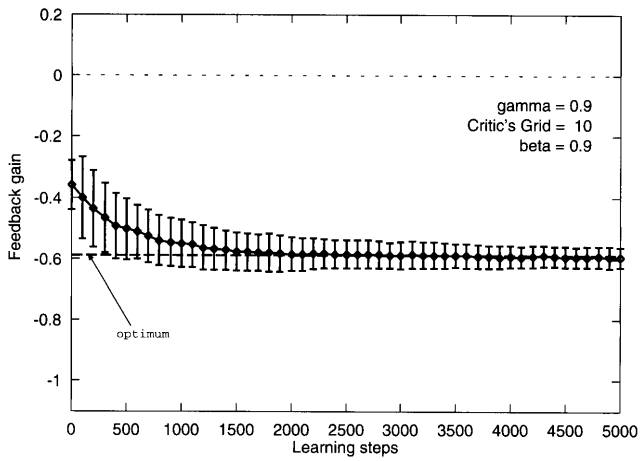


図 7 $\gamma = 0.9$, Critic は 10 分割グリッドの関数近似．Actor に適正度の履歴を用いた場合の 100 試行の平均と分散

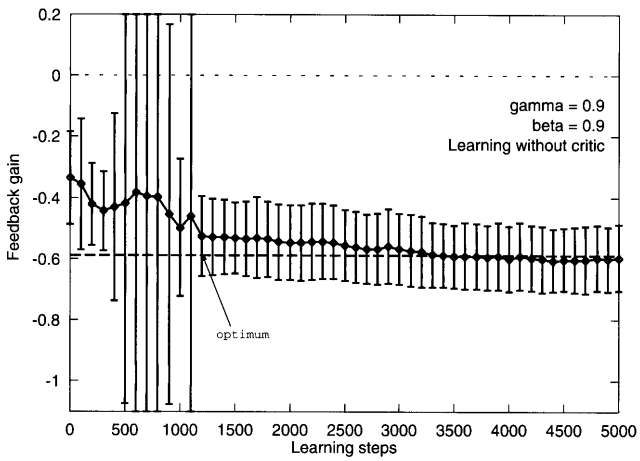


図 8 $\gamma = 0.9$, Critic を用いないで，つまり $\hat{V}(x) = 0$ for all x として Actor の適正度の履歴だけで学習した場合の 100 試行の平均と分散．確率的傾斜法で学習することと等価

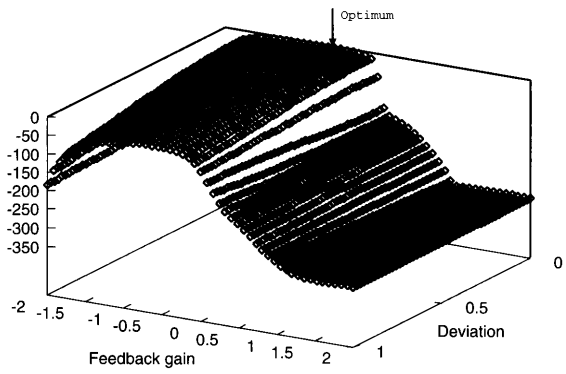


図 9 割引率 γ=0.9 の場合の value function の形状．$\mu = -0.5884$, $\sigma = 0$

ラメータ空間で表示したものである．最適解周辺ではほとんど平らになっている．よって，図8で得た政策の分散が大きい理由は，政策更新のステップ幅のコントロールを行っていないために最適解周辺で政策が収束できずに浮遊しているためであることが分かる．これより，critic が局所解周辺において actor の政策更新ステップ幅を小さくするようにコントロールしていることが分かる．

図7のアルゴリズムは政策の平均値と分散において最も良い解を得ている．これは critic が少ない誤差で value 関数を学習したためだと考えられる．

この実験では，同じ性能を持つ critic を用いて actor-critic を比較した場合，適正度の履歴を使用する actor の方が，使用しない actor よりも良い性能を示した．

6. 倒立振子制御問題への適用

提案手法の有用性について示すため，多次元の状態観測を必要とする非線形・非 2 次形式の倒立振子制御問題 (図 10) へ適用する．[Barto 83] の実験設定を参考にしたが，彼らは行動を離散値としていたので，いくつか修正を加え，行動を連続値として計算機実験を行った．

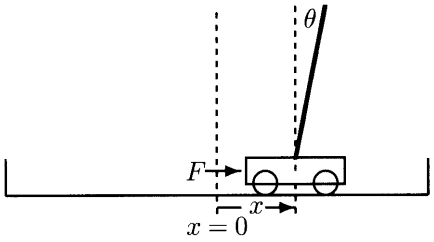


図 10 倒立振子制御問題

6·1 倒立振子問題の定式化

台車の質量 $M = 1.0(\mathrm{kg})$，ポールの質量 $m = 0.1(\mathrm{kg})$，ポールの長さ $2\ell = 1(\mathrm{m})$，重力加速度 $g = 9.8(\mathrm{m/sec^2})$，台車に加えられる力 $F(\mathrm{N})$，台車の摩擦係数 $\mu_c = 0.0005$，ポールの摩擦係数 $\mu_p = 0.000002$ とすると，図 10 のダイナミクスは以下で表される．

$$\ddot{\theta} = \frac{g\sin\theta + \frac{\cos\theta(\mu_c sgn(\dot{x}) - F - m\ell\dot{\theta}^2\sin\theta)}{M+m} - \frac{\mu_p\dot{\theta}}{m\ell}}{\ell\left(\frac{4}{3} - \frac{m\cos^2\theta}{M+m}\right)}$$

$$\ddot{x} = \frac{F + m\ell\left(\dot{\theta}^2\sin\theta - \ddot{\theta}\cos\theta\right) - \mu_c sgn(\dot{x})}{M+m}$$

本実験では $\Delta t = 0.02\mathrm{sec}$ の離散時間システムとして近似した．エージェントは $(x, \dot{x}, \theta, \dot{\theta})$ を観測し，行動として台車に加える力 F を出力する．エージェントは行動出力として任意の連続値をとることができるが，$F = \pm 20(\mathrm{N})$ の範囲を超える場合には，環境はこの制限を超える分を無視して行動を実行する．環境は $(x, \dot{x}, \theta, \dot{\theta}) = (0, 0, 0, 0)$ の初期状態から始まる．ポールの角度 θ が ± 12 度を超えるか，または台車の中心 x が ± 2.4 の範囲からはみ出すと，環境からエージェントへ -1 の報酬が与えられ，環境は初期状態へ戻る．

6·2 エージェントの実装

本実験の actor は，式 (13)，(16) で示したのと同様の実装である．状態空間は $(x, \dot{x}, \theta, \dot{\theta}) = (\pm 2.4\ \mathrm{m}, \pm 2\ \mathrm{m/sec}, \pm\pi \times 12/180\ \mathrm{rad}, \pm 1.5\ \mathrm{rad/sec})$ の範囲に限定した．エージェントは五つの内部変数 $w_1 \cdots w_5$ を持ち，これを用いて μ と σ を以下のように計算する．

$$\mu = w_1\frac{x_t}{2.4} + w_2\frac{\dot{x}_t}{2} + w_3\frac{\theta_t}{12\pi/180} + w_4\frac{\dot{\theta}_t}{1.5}$$

$$\sigma = 0.1 + \frac{1}{1+\exp(-w_5)} \qquad (18)$$

LQR の例題の場合と同様にして，適正度 $e_1, \cdots, e_5$ は以下に与えられる．

$$e_1 = (a_t - \mu)x_t, \quad e_2 = (a_t - \mu)\dot{x}_t$$
$$e_3 = (a_t - \mu)\theta_t, \quad e_4 = (a_t - \mu)\dot{\theta}_t$$
$$e_5 = ((a_t - \mu)^2 - \sigma^2)(1 + 0.1 - \sigma)$$

Critic は状態観測の空間を格子状に分割して離散化し，それぞれのグリッドに対して TD(0) 法を用いて value を推定する．本実験では正規化された状態空間を各軸に対して 3 等分，つまり $3^4 = 81$ 個の矩形に分割した．割引率 $\gamma = 0.95$，$\alpha = 0.5$, $\alpha_p = 0.001$ に設定した．

6·3 実験結果

図 11 は異なる 3 種類のアルゴリズムによる学習の様子を示す．1 trial は初期状態からポールや台車が許容範囲をはみ出すまでを表す．Actor に適正度の履歴を用いた actor-critic が最も良い結果を得たのに対し，適正度の履歴を用いていない actor-critic では全く学習できなかった．Critic を用いずに actor と適正度の履歴だけを用いた確率的傾斜法では学習できた．全てのアルゴリズムにおいて，政策を保持するための関数近似が全く同じであるにもかかわらず，actor に適正度の履歴を用いるかどうかで大きな差が生じた．

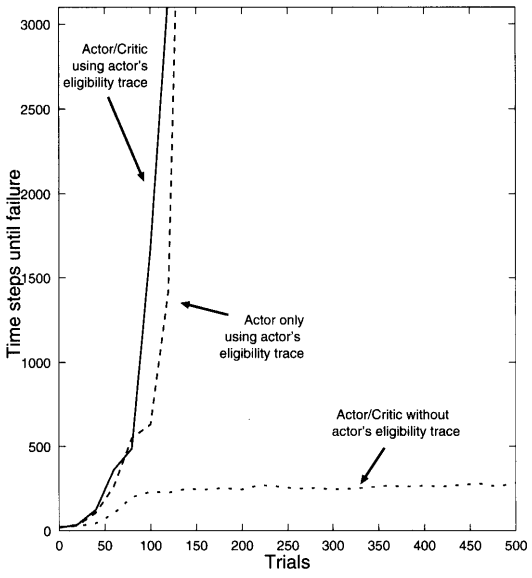


図 11 連続値行動の倒立振子制御問題への適用結果．$\gamma = 0.95$, Critic は $3 \times 3 \times 3 \times 3$ 分割グリッドの関数近似，Actor は線形関数を用いた．それぞれ 100 試行の平均

7. 考 察

政策の表現： Actor-critic アルゴリズムでは，まず政策の関数近似能力が十分であることが要求される．その上で，十分な学習を行うためには critic の関数近似能力を上げるよりも actor に適正度の履歴を用いた方が，本実験では安いコストで学習させることができた．

政策更新ステップ幅のコントロール： 4·3節において，critic は actor の政策更新ステップ幅をコントロールし，局所解の頂上付近では幅を小さくすることを示した．実験を通して critic が学習途中の効率の改善と，収束時の政策パラメータのドリフトを抑える効果のあることを観察した．

非マルコフ環境における頑健性： Critic へ適用するアルゴリズムには任意性がある．TD(λ) は非マルコフ環境に対して頑健であると言われている [Peng 94, Sutton 95]．また actor に適正度の履歴を用いた場合も非マルコフ環境に対して頑健である [Kimura 97]．Critic に TD(λ) を用いれば，本アルゴリズムは非マルコフ環境においてさらに頑健となることが期待できる．

DP ベースの手法と確率的傾斜法の並列学習： Actor における政策学習アルゴリズムの基礎となる確率的傾斜法は，モンテカルロ法により value 関数の勾配を求めて政策改善を行っており，非 DP 手法である．そのため確率的傾斜法単体では，DP を基礎とする強化学習手法に比較し，マルコフ決定過程の環境では学習の効率は悪いが，非マルコフの環境では頑健である．TD(0) を始めとした DP を基礎とする手法は，逆にマルコフ決定過程の環境では学習の効率は良いが，非マルコフの環境では頑健性に欠ける．本手法は critic における value 関数の学習に DP を基礎とする TD(0) 法を用い，actor における政策の学習には非 DP 手法である確率的傾斜法を用いて，高い独立性を保ちながらそれぞれ並列に進行することにより，マルコフ決定過程の環境における学習の効率性と非マルコフ環境における頑健性を両立させることができる．

8. おわりに

本論文では actor に適正度の履歴を用いた actor-critic アルゴリズムを示し，その動作について考察した．Actor の適正度の履歴の割引率と value 関数の割引率が等しい場合，actor の政策更新は critic の推定した value 関数の勾配の方向ではなく，実際のトレーニング系列の割引報酬 (actual return) の勾配の方向へ行われる．そのため critic での value の推定が致命的に不正確であっても actor は政策を学習可能である．Critic が推定する value は actor の政策更新のステップ幅をコントロールする役割を持つ．Critic での value 関数の推定が正確ならば，学習途中の性能が改善され，最適政策付近ではステップ幅が自動的に小さくなる．1 次元の線形 2 次形式制御問題および倒立振子制御問題を取り上げ，actor に適正度の履歴を用いない従来の actor-critic と比較を行い，解析結果より予想された性質について実験的に確認した．非マルコフな環境における本手法の挙動についての解析は今後の課題である．

◇参 考 文 献◇

[Baird 94] Baird, L.C.: Reinforcement Learning in Continuous Time: Advantage Updating, *Proceedings of IEEE International Conference on Neural Networks*, Vol.IV, pp.2448–2453 (1994).

[Barto 83] Barto, A.G., Sutton, R.S. & Anderson, C.W.: Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol.SMC-13, No.5, pp.834–846, September/October 1983.

[Clouse 92] Clouse, J.A. & Utogoff, P.E.: A Teaching Method for Reinforcement Learning, *Proc. of the 9th International Conference on Machine Learning*, pp.93–101 (1992).

[Crites 94] Crites, R.H. & Barto, A.G.: An Actor/Critic Algorithm that is Equivalent to Q-Learning, *Advances in Neural Information Processing Systems 7*, pp.401–408 (1994).

[Doya 96] Doya, K. : Efficient Nonlinear Control with Actor-Tutor Architecture, *Advances in Neural Information Processing Systems 9*, pp.1012–1018 (1996).

[Gullapalli 92] Gullapalli, V.: Reinforcement Learning and Its Application to Control, *PhD Thesis*, University of Massachusetts, Amherst, COINS Technical Report 92–10 (1992).

[Jaakkola 94] Jaakkola, T., Singh, S.P. & Jordan, M.I.: Reinforcement Learning Algorithm for Partially Observable Markov Decision Problems, *Advances in Neural Information Processing Systems 7*, pp.345–352 (1994).

[Kaelbling 96] Kaelbling, L. P., Littman, M.L., & Moore, A.W.: Reinforcement Learning: A Survey, *Journal of Artificial Intelligence Research*, Vol.4, pp.237–277 (1996).

[Kimura 95] Kimura, H., Yamamura, M. & Kobayashi, S.: Reinforcement Learning by Stochastic Hill Climbing on Discounted Reward, *Proceedings of the 12th International Conference on Machine Learning*, pp.295–303 (1995).

[木村 96] 木村 元，山村雅幸，小林重信: 部分観測マルコフ決定過程下での強化学習：確率的傾斜法による接近，人工知能学会誌, Vol.11, No.5, pp.761–768 (1996).

[Kimura 97] Kimura, H., Miyazaki, K. & Kobayashi, S.: Reinforcement Learning in POMDPs with Function Approximation, *Proceedings of the 14th International Conference on Machine Learning*, pp.152–160 (1997).

[Kimura 98] Kimura, H. & Kobayashi, S.: An Analysis of Actor/Critic Algorithms using Eligibility Traces: Reinforcement Learning with Imperfect Value Function, *15th International Conference on Machine Learning*, pp.278–286 (1998).

[Lin 96] Lin, C.J. & Lin, C.T.: Reinforcement Learning for An ART-Based Fuzzy Adaptive Learning Control Network, *IEEE Transactions on Neural Networks*, Vol.7, No.3, pp.709–731 (1996).

[Littman 94] Littman, M.L.: Markov games as a framework for multi-agent reinforcement learning, *Proceedings of the 11th International Conference on Machine Learning*, pp.157–163 (1994).

[小笠原 67] 小笠原正巳，坂本武司 著，北川敏男 編: 情報科学講座 (全 62 巻)A・5・1 マルコフ過程，共立出版 (1967).

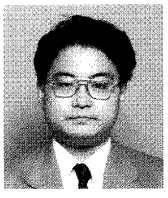
[Pendrith 96] Pendrith, M.D. & Ryan, M.R.K.: Actual return reinforcement learning versus Temporal Differences: Some theoretical and experimental results, *Proceedings of the 13th International Conference on Machine Learning*,

pp.373–381 (1996).
[Peng 94] Peng, J. & Williams, R.J.: Incremental Multi-Step Q-Learning, *Proceedings of the 11th International Conference on Machine Learning*, pp.226–232 (1994).
[Singh 94] Singh, S.P., Jaakkola, T. & Jordan, M.I.: Learning Without State-Estimation in Partially Observable Markovian Decision Processes, *Proceedings of the 11th International Conference on Machine Learning*, pp.284–292 (1994).
[Singh 96] Singh, S.P. & Sutton, R.S.: Reinforcement Learning with Replacing Eligibility Traces, *Machine Learning 22*, pp.123–158 (1996).
[Sutton 88] Sutton, R.S.: Learning to Predict by the Methods of Temporal Differences, *Machine Learning 3*, pp.9–44 (1988).
[Sutton 90] Sutton, R.S.: Reinforcement Learning Architectures for Animates, *Proceedings of the 1st International Conference on Simulation of Adaptive Behavior*, pp.288–295 (1990).
[Sutton 95] Sutton, R.S.: TD Models: Modeling the world at a Mixture of Time Scales, *Proceedings of the 12th International Conference on Machine Learning*, pp.531–539 (1995).
[Sutton 98] Sutton, R.S. & Barto, A.: Reinforcement Learning: An Introduction, *A Bradford Book*, The MIT Press (1998).
[Watkins 92] Watkins, C.J.C.H. & Dayan, P.: Technical Note: *Q*-Learning, *Machine Learning 8*, pp.55–68 (1992).
[Williams 90] Williams, R. J. & Baird, L. C.: A Mathematical Analysis of Actor-Critic Architectures for Learning Optimal Controls through Incremental Dynamic Programming, *Proceedings of the Sixth Yale Workshop on Adaptive and Learning Systems*, pp.96–101, Center for Systems Science, Dunham Laboratory, Yale University, New Haven (1990).
[Williams 92] Williams, R. J.: Simple Statistical Gradient Following Algorithms for Connectionist Reinforcement Learning, *Machine Learning 8*, pp.229–256 (1992).

〔担当委員：安倍直樹〕

1998 年 10 月 22 日　受理

著　者　紹　介



木村　元（正会員）

1992 年東京工業大学工学部制御工学科卒業．1994 年同大学大学院総合理工学研究科知能科学専攻修士課程修了．1997 年同大学大学院博士課程修了，同年 4 月日本学術振興会 PD 研究員，1998 年 4 月，東京工業大学大学院総合理工学研究科助手，現在に至る．人工知能，特に強化学習に関する研究を行っている．計測自動制御学会，日本ロボット学会各会員．

小林　重信（正会員）

1974 年東京工業大学大学院博士課程経営工学専攻修了．同年 4 月，同大学工学部制御工学科助手．1981 年 8 月，同大学大学院総合理工学研究科助教授．1990 年 8 月，教授．現在に至る．問題解決と推論制御，知識獲得と学習などの研究に従事．計測自動制御学会，情報処理学会各会員．