適応的探索法を用いた強化学習

非会員 梅 迫 公 輔 (山口大学大学院理工学研究科)

正 員 大 林 正 直 (山 口 大 学 工 学 部)

非会員 小 林 邦 和 (山 口 大 学 工 学 部)

Reinforcement Learning Using Adaptive Search Method

Umesako Kosuke, Non-member (Graduate School of Sci. and Eng, Yamaguchi University), Obayashi Masanao, Member, Kobayashi Kunikazu, Non-member (Faculty of Eng., Yamaguchi University)

We propose an adaptive probability density function (PDF) to select an effective action on reinforcement learning (RL). The uniform distribution function and the normal distribution function of an action are often used to select an action. When these functions are used, however, the information of search direction is not considered. The proposed method utilizing the information of it enables RL to reduce the number of trials, which is needed to real environment learning. Furthermore, the proposed method can be applied easily to various methods of RL, for example, actor-critic, stochastic gradient ascent method. The performance of our proposed method is demonstrated by computer simulations.

キーワード：強化学習，学習の高速化，連続行動空間，actor-critic，確率的傾斜法，正規化ガウス関数ネットワーク

Keywords: reinforcement learning, acceleration of learning, continuous action space, actor-critic, stochastic gradient ascent method, normalized Gaussian network

1. ま え が き

強化学習⁽¹⁾は、試行錯誤により学習を行う学習方式である。エージェントは与えられた環境の中で行動を行い、その結果として得られた報酬を基に学習を行う。この学習方式は環境に対する詳細な知識が不要である為、幅広い対象に対して柔軟に学習を行う事が出来る。又、生体の学習との関連も示唆されている⁽¹⁾。

試行錯誤を実現するため、エージェントはある確率密度関数 (Probability Density Function: PDF) に従って行動を決定する。行動が連続値の場合、これまで正規分布^{(1)~(5)}又は一様分布⁽⁶⁾が用いられて来ている。しかしこれらを用いた場合、試行錯誤による探索は全ての方向に対して常に等確率で行われることになる。しかし、生体が何かを試行錯誤により学習する場合、探索方向を適応的に変化させている。

一方、強化学習において、少ない試行回数で効率良く学習を行う事は重要な課題の一つである。特に実環境でロボットに行動学習を行わせる場合などでは、試行を重ねる事に大きな時間と手間が掛かり、又、失敗を繰り返せばロボットが破損してしまう場合もある。

そこで本研究では、試行錯誤において生体と同様に探索方向を適応的に変化させる手法を強化学習へ導入する。従

来の正規分布を用いた手法^{(1)~(5)}では中心と分散のみを変化させ行動の探索領域を決定しているが、それでは探索領域をきめ細かく調整出来ない。一様分布を用いた場合⁽⁶⁾も同様である。提案法は、更に探索領域の偏り (探索方向) も調整可能である。それにより、探索領域をより細かく調整し、報酬の多く得られる領域を集中的に探索する事が可能となる。これにより少ない試行で目的の達成が可能となる事が期待出来る。

本章でまず枠組みを与え、次に例として、強化学習法の一つである actor-critic⁽¹⁾と Stochastic Gradient Ascent method (SGA)⁽⁷⁾に適用する。シミュレーションにより従来法との比較を行い、提案法を用いる事で試行回数を大きく減少させる事が可能である事を示す。

2. 適応的探索法の枠組み

ランダムサーチにおいて、パラメータ探索の成功・不成功情報を活用し、探索の集中化と多様化を統一した枠組みで取り扱う最適手法 RasID⁽⁸⁾が提案されている。この考えを強化学習での試行錯誤へ適用し、探索を効率的に行う。

〈2・1〉 adaptive PDF エージェントの行動決定において、従来の一様分布や正規分布に代え、RasID で用いられている PDF を改変した関数である adaptive PDF $\pi(a_k)$ を導入する。

図 1 adaptive PDF
Fig. 1. Adaptive PDF.

$$\pi(a_k) = \begin{cases} p_k \beta_k e^{\beta_k(a_k - \mu_k)} & (a_k \leq \mu_k) \\ q_k \beta_k e^{-\beta_k(a_k - \mu_k)} & (a_k > \mu_k) \end{cases} \dots (1)$$
$$p_k + q_k = 1$$

ここで、 a_k は時間ステップ t における行動ベクトル $a_t \in R^m$ の k 番目の要素、 p_k, q_k は a_k に対する探索方向を決定するパラメータ、 β_k は探索範囲に関するパラメータである。図 1 の様に、 μ_k を中心として左右非対称の形状となっており、 p_k, q_k により探索方向を、 β_k により探索範囲を制御することが出来る。これらのパラメータを逐次的に更新することにより、適応的な探索が実現される。具体的には、 $a_k > \mu_k$ である行動が有用だと判断された、即ち、正方向探索が成功した（以降、中心 μ_k より大きい方向の探索を正方向、小さい方向の探索を負方向探索と定義する。）場合 q_k を、負方向探索が成功した場合には p_k を増大させる。探索が失敗した場合は探索方向を等方向とし $q_k = p_k = 0.5$ とする。以上の操作により、探索の成功・不成功情報を活用出来、効率的な探索を行う事が可能である。なお、 a_k は次式により計算出来る⁽⁸⁾。

$$0 \leq z \leq p_k \Rightarrow a_k = \frac{1}{\beta_k} \ln\left(\frac{z}{p_k}\right) + \mu_k \dots (2)$$
$$p_k < z \leq 1.0 \Rightarrow a_k = -\frac{1}{\beta_k} \ln\left(\frac{1-z}{q_k}\right) + \mu_k \dots (3)$$

$z : [0,1]$ の一様乱数

3. actor-critic への適用

〈3・1〉 actor-critic actor-critic は Temporal Difference に基づく強化学習法であり、報酬を明示的に予測し学習を行う。本研究でのシステムを図 2 に示す。actor-critic は、環境の状態 $x_t \in R^n$ から行動ベクトル a_t を出力する。そして結果として得られた報酬 r_t を基に学習を行う。具体的には、critic は将来に渡って得られる報酬の総和の予測値（予測報酬） P_t を計算し、その誤差 $\hat{r}_t$ を出力する。そして、 $\hat{r}_t$ の絶対値が小さくなるように学習を行う。actor は行動ベクトル a_t を出力し、 P_t が大きくなるように学習を行う。 $\mu_t \in R^m$ は確率的選択を行う前の行動ベクトル、 $n_t \in R^m$ はノイズである。予測誤差 $\hat{r}_t$ は次式により計算される。

$$\hat{r}_t = r_t + \gamma P_{t+1} - P_t \dots (4)$$

ここで γ は、割引率と呼ばれる $0 \leq \gamma < 1$ の定数である。 $\hat{r}_t$ は TD error と呼ばれ、学習によってこれが減少する様

図 2 actor-critic
Fig. 2. actor-critic.

に critic を構成する。
関数近似器には、本研究では正規化ガウス関数ネットワーク（Normalized Gaussian network: NGnet）⁽⁹⁾を用いる。その学習パラメータは、出力層の結合荷重と、中間層の活性化関数の中心と分散である。学習には Back Propagation を用い、critic では $\hat{r}_t^2/2$ を評価関数として、actor では $-\hat{r}_t \cdot n_t$ を誤差信号とみなして誤差逆伝播を行う。
〈3・2〉 提案法の導入 ノイズ n_t を adaptive PDF により発生させる事で、行動選択部（Action selector）に提案法を導入する。

$$0 \leq z \leq p_k \Rightarrow n_k = \frac{1}{\beta_k} \ln\left(\frac{z}{p_k}\right) \dots (5)$$
$$p_k < z \leq 1.0 \Rightarrow n_k = -\frac{1}{\beta_k} \ln\left(\frac{1-z}{q_k}\right) \dots (6)$$

$z : [0,1]$ の一様乱数

探索方向の更新は下式により行う。

- 探索成功の場合
負方向探索 ($a_k \leq \mu_k$) のとき
$$p_k \leftarrow \alpha_k p_k + (1 - \alpha_k) \dots (7)$$

正方向探索 ($a_k > \mu_k$) のとき
$$q_k \leftarrow \alpha_k q_k + (1 - \alpha_k) \dots (8)$$

- 探索不成功の場合
$$p_k \leftarrow 0.5, q_k \leftarrow 0.5 \dots (9)$$

ここで、 α_k は探索確率の指数平滑定数 ($0 < \alpha_k < 1$) である。なお、上記において p_k を更新した場合は $q_k = 1 - p_k$ とし、 q_k を更新した場合は $p_k = 1 - q_k$ とする。探索の成功・不成功の判別は、 $P_t - P_{t-1} > 0$ の場合を探索成功、それ以外を不成功とする。又、 β_k は次式により決定する。

$$\beta_k = \bar{\beta}_k e^{\eta_k \bar{P}} + \underline{\beta}_k \dots (10)$$
$$\bar{P} = \max(0, P) \dots (11)$$

ここで $\bar{\beta}_k, \underline{\beta}_k, \eta_k$ は定数である。即ち、予測報酬が小さいときにはノイズは大きくなり広範囲に探索が行なわれ、多いときにはノイズは小さくなり安定した制御が行なわれる。

図3 倒立振り子
Fig. 3. A pendulum.

図4 学習曲線
Fig. 4. Learning curves.

〈3・3〉実験法 まず基本的な例題として、振り子の倒立制御を行う（図3参照）。そのダイナミクスは

$$m g \ddot{\theta} = m g l \sin \theta - \mu_v \dot{\theta} + T \quad \cdots \cdots (12)$$

で表される。ここで θ は倒立位置からの角度、 T は制御トルクである。本実験では質量 $m = 1$ [kg]、重力加速度 $g = 9.8$ [m/s²]、リンク長 $l = 1$ [m]、粘性係数 $\mu_v = 0.01$ とした。振り子のシミュレーションは、Runge-Kutta 法（サンプリングタイム 0.02 秒）により電子計算機上で行なった。時間ステップ t において強化学習システムが観測する状態は $\mathbf{x}_t \equiv (\theta_t, \dot{\theta}_t)$ 、行動は $\mathbf{a} \equiv T$ とする。（本例題では行動がスカラの為、これ以降 (1) 式における各変数の添え字 k は省略する。）又、報酬 r_t は下式⁽¹⁰⁾に従って試行の間常に与えられる。

$$r_t = \exp \left(-\frac{(\theta_t)^2}{2(\alpha_{R1})^2} - \frac{(\dot{\theta}_t)^2}{2(\alpha_{R2})^2} \right) \quad \cdots \cdots (13)$$

即ち、振り子が倒立位置で静止した状態で報酬は最大となる。ここで、 $\alpha_{R1} = 0.2$ 、 $\alpha_{R2} = 1$ とする。

試行は $(\theta_t, \dot{\theta}_t) = (\pi/18[\text{rad}], 0[\text{rad/s}])$ （10° で静止した状態）から始まり、20 秒間制御と学習を行う。もし $|\theta_t| \geq \pi/6$ となった場合は失敗とみなし、その試行は打ち切る。以上を 1 試行とし、この試行を繰り返す事により学習が進行する。なお、試行の最後の 10 秒間 $\pm \pi/360$ [rad] 以内を維持できた場合を倒立成功とみなす。

本実験では $\gamma = 0.95$ とする。各 NGnet のノード数は (3-64-1) であり、出力層ノードから入力層ノードへのリカレント結合を持つ。adaptive PDF に関するパラメータは、 $\alpha = 0.9$ 、 $\beta = -0.4$ 、 $\bar{\beta} = 1$ 、 $\eta = 0.05$ である。

表 1 実験結果

Table 1. Experimental results.

	The number of success	The average number of trials until success	Computational time per sampling [μ s]
Uniform distribution	99 / 100	1954.8	429
Normal distribution	98 / 100	1966.1	433
Adaptive PDF	100 / 100	831.1	431

図5 試行の代表例（提案法）
Fig. 5. An example of trials (proposed).

又、PDF として一様分布と正規分布を用い、adaptive PDF との比較を行う。一様分布を用いた場合のノイズ n は次式により計算される。

$$n = z \cdot \bar{n} \cdot \exp \left(-\alpha_U \dot{P} \right) \quad \cdots \cdots (14)$$

但し、 z は $[-1, 1]$ の一様乱数、 $\bar{n}$ はノイズの大きさの上限、 α_U は定数である。 $\dot{P}$ は (11) 式により計算される。正規分布を用いる場合の PDF $g(n)$ は次のように表される。

$$g(n) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{n^2}{2\sigma^2} \right) \quad \cdots \cdots (15)$$

$$\sigma = \bar{\sigma} \cdot \exp \left(-\alpha_N \dot{P} \right) \quad \cdots \cdots (16)$$

但し、 $\bar{\sigma}$ は標準偏差の上限、 α_N は定数である。パラメータの値は $\bar{n} = 2$ 、 $\alpha_U = 0.1$ 、 $\bar{\sigma} = 2$ 、 $\alpha_N = 0.1$ である。本研究では、各パラメータは試行錯誤により決定している。

〈3・4〉実験結果 100 種の乱数系列を用いて 100 回のシミュレーションを行った。その結果、提案法の学習が最も高速であった。学習曲線を図4に示す。この図は 100 回のシミュレーションに対して平均化したデータである。縦軸は平均報酬であり、1 試行の間に得られた報酬の平均値である。adaptive PDF を用いた場合（提案法）、少ない試行回数で多くの報酬が得られている。次に、結果を表1に示す。提案法では全ての乱数系列に対して学習が成功したが、従来法では 10000 回試行を重ねても倒立が成功しない場合があった。又、提案法は従来法の約半分の試行回数で

1. 環境の観測 x_t を受け取る。
2. $\pi(a_t, w, x_t)$ の確率で行動 a_t を実行する。
3. 環境から報酬 r_t を受け取る。
4. 内部変数 w の全ての要素 w_i について以下の $e_i(t)$ と $\bar{D}_i(t)$ を求める。但し γ は割引率 ($0 \leq \gamma < 1$) である。

$$e_i(t) = \frac{\partial}{\partial w_i} \ln \{ \pi(a_t, w, x_t) \}$$

$$\bar{D}_i(t) = e_i(t) + \gamma \bar{D}_i(t-1)$$
5. 以下の式を用いて $\Delta w_i(t)$ を求める。

$$\Delta w_i(t) = (r_t - b) \bar{D}_i(t),$$
 但し b は定数である。
6. 政策の改善: 以下の式で w を更新

$$\Delta w(t) = (\Delta w_1(t), \Delta w_2(t), \dots, \Delta w_i(t), \dots),$$

$$w \leftarrow w + \alpha \Delta w(t),$$
 但し α は非負の学習定数である。
7. 時間ステップ t を $t+1$ へ進めて, 1 へ戻る。

図6 SGA アルゴリズム

Fig. 6. Algorithm of the SGA.

成功している。計算量に差は殆どみられない。

提案法による試行の代表例を図5に示す。振子が負の方向に振れている状況 (およそ 3.5[s] から 5[s] のまで部分) では $q > 0.5$, つまりトルクが大きな方向を重点的に探索しており, 振子を静止させる段階 (およそ 6[s] から 9[s] までの部分) では振子を倒立状態で静止させる方向 (振子を倒す方向) を重点的に探索している。以上の様に, 探索方向を適応的に変化させる事により学習が効率的に行われている。

4. SGA への適用

〈4・1〉 SGA SGA はメモリレス型強化学習法であり, 計算コストが低く, 報酬の明示的な推定が不要であるという特徴を持つ。その為, 成功・不成功が判断出来ず, actor-critic と同様な手法では探索方向を制御出来ない。その SGA アルゴリズムを図6に示す⁽⁷⁾。

〈4・2〉 提案法の導入 行動 (本例題では制御トルク) a を決定する PDF である $\pi(a, w, x_t)$ として adaptive PDF を用いることにより, 提案法を SGA に導入する。

$$\pi(a, w, x_t) = \begin{cases} p\beta e^{\beta(a-\mu)} & (a \leq \mu) \\ q\beta e^{-\beta(a-\mu)} & (a > \mu) \end{cases} \dots (17)$$

$$p + q = 1$$

但し, 中心 μ は次式により適切に変化させる。

$$\mu = \begin{cases} \bar{\mu} & \text{if } \mu_{tmp} > \bar{\mu} \\ -\bar{\mu} & \text{if } \mu_{tmp} < -\bar{\mu} \\ \mu_{tmp} & \text{otherwise} \end{cases} \dots (18)$$

$$\mu_{tmp} = \alpha_1 \theta_t \cdot w_1 + \alpha_2 \dot{\theta}_t \cdot w_2 + \alpha_3 \cdot w_3 + \alpha_4 a_{t-1} \cdot w_4 \dots (19)$$

図7 学習曲線

Fig. 7. Learning curves.

ここで, $\bar{\mu}$ は中心の大きさの制限値, a_{t-1} は1ステップ前の行動である。荷重 $w_1 \sim w_4$ を SGA アルゴリズムにて学習更新することにより, 振子の制御に適した政策を獲得する。

ここで, SGA は報酬を明示的には推定しないため, 試行の成功・不成功を逐次的に判別することが出来ない。従って actor-critic の場合と同様なパラメータ更新法 ((7)~(9) 式) は使用出来ない。そこでパラメータ p, q, β についても SGA を用いて更新する。各パラメータの計算式を以下に示す。

$$p = \alpha_P w_P \dots (20)$$

$$q = 1 - p \dots (21)$$

$$\beta = \frac{1}{1 - e^{-\alpha_{\beta 1} |w_{\beta}|}} + \alpha_{\beta 2} \dots (22)$$

$\alpha_P, \alpha_{\beta 1}, \alpha_{\beta 2}$ は定数である。 w_P, w_{β} は学習パラメータであり, 探索範囲 β が負になるのを防ぐ為, (22) 式では w_{β} の絶対値を用いている。又, $0 \leq p \leq 1$ であるので, w_P は $\underline{p}/\alpha_P \leq w_P \leq \bar{p}/\alpha_P$ に制限する。ここで $\underline{p}$ は p の下限, $\bar{p}$ は p の上限を表し, $0 \leq \underline{p} \leq \bar{p} \leq 1$ である。

〈4・3〉 実験法 ここでも〈3・3〉と同様に振子の倒立を行う。各パラメータは, $\gamma = 0.95, b = 0.6, \bar{\mu} = 5, \alpha = 0.00002, \alpha_1 = \alpha_2 = 500, \alpha_3 = 0.1, \alpha_4 = 0.01, \alpha_P = 2, \alpha_{\beta 1} = 1, \alpha_{\beta 2} = -0.5, \underline{p} = 0.1, \bar{p} = 0.9$ である。学習パラメータの初期値は, $w_i = 0 (i = 1, 2, \dots, 4), w_P = 0.25, w_{\beta} = 0.5$ である。

又, $\pi(a, w, x_t)$ に正規分布を用いた手法との比較を行う。この場合では,

$$\pi(a, w, x_t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right) \dots (23)$$

となる。 μ の計算には (18), (19) 式を用いる。 σ の計算は (22) 式と同様に,

$$\sigma = \frac{1}{1 - e^{-\alpha_{\beta 1} |w_{\beta}|}} + \alpha_{\beta 2} \dots (24)$$

により計算する。パラメータは学習定数 α のみ異なり $\alpha = 0.005$ である。

表 2 実験結果
Table 2. Experimental results.

	The number of success	The average number of trials until success	Computational time per sampling [μs]
Normal distribution	100 / 100	55.5	258
Adaptive PDF	100 / 100	8.9	261

図 8 Khepera ロボット
Fig. 8. Khepera robot.

〈4・4〉 実験結果 乱数系列を変え 100 回のシミュレーションを行った。その結果、提案法の学習が最も高速であった。学習曲線を図 9 に示す。提案法は従来法よりも少ない試行回数で多くの報酬が得られている。次に、結果を表 2 に示す。提案法は従来法に比べて約 7 分の 1 の試行回数で倒立が成功しており、actor-critic の場合と同様、学習を高速に行えている。計算量はほぼ同じであった。

5. 行動要素が複数の場合

これまでは行動がスカラの場合を扱ってきたが、本節では行動がベクトルの問題へ提案法を適用する。即ち Khepera ロボット⁽¹¹⁾の 2 つの駆動輪を制御し、行動学習を行う。

〈5・1〉 Khepera ロボットシミュレータ Khepera ロボットは 8 つの距離センサと 2 つの駆動輪を持つ (図 8 参照)。Khepera ロボットシミュレータ⁽¹¹⁾は、このロボットをコンピュータ上でシミュレートする。

各センサは障害物との距離の近さに応じた値を出力する。シミュレータではこの値と車輪の駆動速度に ±10% のノイズが加えられる。又、ロボットの方向には ±5% のノイズが加えられる。この様に環境は決定的では無い。

〈5・2〉 実験法 実験環境として、図 9 の通路を用いる。ロボットの初期状態は図の位置であり、制限時間 t_F 内に出来るだけ早くゴール (斜線部) に辿り着く事が目的である。

強化学習システムが観測するのはセンサ 0, 1, 2 の値を平均し正規化した値 $s_t^{(L)}$ とセンサ 3, 4, 5 の値を平均し正規化した値 $s_t^{(R)}$ 、即ち $\mathbf{x}_t \equiv (s_t^{(L)}, s_t^{(R)})$ であり、行動は左右の車輪の駆動速度 (それぞれ $v_t^{(L)}, v_t^{(R)}$)、即ち $\mathbf{a}_t \equiv (v_t^{(L)}, v_t^{(R)})$ である。報酬は常に与えられ、タイムステップ t での報酬は以下の式で計算される。

図 9 実験環境
Fig. 9. Experimental environment.

図 10 学習曲線
Fig. 10. Learning curves.

$$r_t = \begin{cases} (t_F - t)/(2t_F) + 0.5 & \text{if 現在位置がゴール} \\ 0 & \text{if } s_L > T_C \text{ or } s_R > T_C \\ 0.5 & \text{otherwise} \end{cases} \quad (25)$$

但し、 T_C は衝突の判別に用いる閾値である。ロボットがゴールした場合、要したタイムステップの少なさに応じた報酬を得、壁に衝突したとみなされると 0 の報酬を得る。試行はロボットがゴールに辿り着くか $t = t_F$ となるまで行い、その間の全てのステップに対し、制御と学習を行う。これを 1 試行とし、試行が終了するとロボットを初期状態に戻し、次の試行に移る。壁に衝突しても試行は中断されない。

各パラメータは、 $t_F = 500$ 、 $T_C = 600/1023$ である。強化学習システムには、3 章で用いた actor-critic を用いる。NGnet のノード数は actor で (4-625-2)、critic で (3-125-1) である。adaptive PDF に関するパラメータは、全ての k に対して $\mu_k = 0$ 、 $\alpha_k = 0.7$ 、 $\beta_k = -0.9$ 、 $\bar{\beta}_k = 1$ である。

又〈3・3〉と同様、行動選択部のノイズに一樣分布と正規分布も使い、それぞれの比較を行なう。各パラメータの値は、 $\bar{n} = 3$ 、 $\alpha_U = 0.01$ 、 $\sigma = 2$ 、 $\alpha_N = 0.005$ である。その他のパラメータの値は〈3・3〉と同様である。

〈5・3〉 実験結果 試行回数を 100 回とし、100 種の乱数系列に対してシミュレーションを行った。平均化した学習曲線を図 10 に示す。縦軸はゴール迄に掛ったステッ

表 3 実験結果

Table 3. Experimental results.

	The average number of success	The average number of trials until success
Uniform distribution	54.8	15.8
Normal distribution	60.7	11.0
Adaptive PDF	75.9	5.4

ブ数（ゴール出来なかった試行は $t_F + 1$ として計算）であり、横軸は試行回数である。提案法は少ない試行回数でより早くゴールに到達している。次に、結果を表 3 に示す。提案法はゴール到達回数が最も多く、しかもより少ない試行回数でゴールへ到達している。このように、行動がベクトルであるこの問題に対しても、学習を効率良く行える。

6. 検 討

以上の結果から、適応的探索法を用いる事により試行回数を大きく減少させる事が可能である事が確認された。その際、汎化能力に対する影響はあるのだろうか？

その調査の為、 $\langle 3 \cdot 3 \rangle$ の実験において 10° からの倒立を学習したシステムを用いて、 5° 、 15° からの倒立を行った。結果を表 4 に示す。提案法は、 15° からの倒立では従来法と同等であったが、 5° からの倒立では僅かに劣っていた。

その原因は、提案法では振子の角速度が大きい傾向がある為だと考えられる。あるシミュレーションでの学習が終了するまでの振子の状態の軌跡を図 11 に示す。提案法では振子を立たせる方向のトルクを重点的に探索する為、振子の状態は角速度が大きな方向に分布している。その結果、 5° で静止した状態付近の学習があまり行われていない為に、 5° からの成功率が低い結果になったと考えられる。 15° からの倒立については、振子が 10° より外側に振れるのはまだ学習が進んでいない段階であるから、その時点では探索方向もランダムに近く、従来法との差が付かなかったからだと考えられる。

以上の様に、提案法は行動がより良くなると期待される方向を重点的に探索する為、汎化能力に悪影響が出る場合がある。これを防ぐには、ある程度学習速度を犠牲にして探索範囲を大きく取る必要がある。その為には、探索範囲を決定する β_k の初期値や下限を大きくする事や、環境の多様な初期状態に対して学習を行う事などが考えられる。根本的な解決として、探索確率の指数平滑定数 α_k を大きく設定し探索の偏りを弱める事も考えられるが、何れにしても、本論文の特徴である少ない試行回数という効果は減少すると考えられる。

7. ま と め

本研究では強化学習における試行錯誤に関して探索方向を適応的に変化させる事を提案し、その枠組みを与えた。そ

表 4 実験結果

Table 4. Experimental results.

	The number of success	
	5°	15°
Uniform distribution	98 / 100	14 / 100
Normal distribution	93 / 100	28 / 100
Adaptive PDF	81 / 100	27 / 100

図 11 振子の軌跡 (a)：正規分布 (b)：adaptive PDF

Fig. 11. Locus of the pendulum. (a): normal distribution, (b): adaptive PDF.

して提案法を強化学習法の一つである actor-critic と SGA へ適用し、その効果を実験的に検証した。提案法は、振子の倒立制御や自律走行ロボットの行動学習に対して、又、両学習法において、その学習効率の高さを示した。本手法を用いることにより、行動の探索領域をきめ細かく調整する事が可能となり、計算コストを増大させることなく試行回数を大きく減少させることが可能となる。又、行動が連続値の強化学習法に対して容易に適用可能なため、その応用範囲は広いと考えられる。提案法の更なる解析、種々の強化学習法・問題への適用と検証は今後の課題である。
(平成 13 年 7 月 23 日受付、同 13 年 10 月 19 日再受付)

文 献

- (1) A. G. Barto, R. S. Sutton, and C. W. Anderson: Neuron-like adaptive elements that can solve difficult learning control problems, *IEEE Transactions on Systems, Man, and Cybernetics*, 13, pp. 834-846 (1983)
- (2) J. Lee, S. Oh, D. Choi: TD Based Reinforcement Learning Using Neural Networks in Control Problems with Continuous Action Space, *Proc. International Joint Conference on Neural Networks*, pp. 2028-2033 (1998)
- (3) 森本淳, 銅谷賢治: 強化学習を用いた高次元連続状態空間における系列運動学習——起き上がり運動の獲得——, *電子情報通信学会誌*, J82-D-II, pp. 2118-2131 (1999)
- (4) K. Doya: Reinforcement learning in continuous time and space, *Neural Computation*, 12, pp. 243-269 (1999)
- (5) 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 のもとの強化学習—, *人工知能学会誌*, 16 巻, 12 号, pp. 267-275 (2000)
- (6) 落合真吾, 佐野雅己, 沢田康次: 強化学習を用いた球技動作の予測と制御, *電子情報通信学会*, NC97-10, pp. 69-75 (1997)
- (7) 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）
- （8）平澤宏太郎，東郷和幸，胡敬焄，大林正直，邵寧，村田純一：ニューラルネットワークの適応的ランダム探索最適化手法-RasID-，計測自動制御学会論文集，Vol. 34，No. 8，pp. 1088-1096（1998）
- （9）J. Moody, C.J. Darken: Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2), pp. 281-294 (1989)
- （10）石井信，佐藤雅昭：オンライン EM アルゴリズムを用いた強化学習法，電子情報通信学会技術研究報告，NC98-83，pp. 41-48（1999）
- （11）<http://diwww.epfl.ch/lami/team/michel/khep-sim/index.html>

梅 迫 公 輔（非会員）1976年生。1998年3月広島工業大学工学部電子工学科卒業。2000年3月山口大学大学院理工学研究科博士前期課程知能情報システム工学専攻修了。2000年4月同大学院理工学研究科博士後期課程システム工学専攻入学。現在に至る。計測自動制御学会会員。

大 林 正 直（正員）1951年生。1975年3月九州工業大学工学部電気工学科卒業。1977年3月九州大学大学院工学研究科修士課程電気工学専攻修了。1977年4月（株）日立製作所入社大みか工場勤務。1989年4月九州大学助手。1997年7月同助教授。1998年4月山口大学助教授。現在同教授。計測自動制御学会，システム制御情報学会，米国電気電子学会，国際神経回路学会の各会員。博士（工学）。

小 林 邦 和（非会員）1969年生。1991年3月山口大学工学部電子工学科卒業。1993年3月同大学院工学研究科博士前期課程知能情報システム工学専攻修了。大学院博士後期課程システム工学専攻退学。1994年4月同大工学部知能情報システム工学科助手。現在に至る。米国電気電子学会，電子情報通信学会，日本神経回路学会の各会員。