報酬駆動型環状ロボットの実現と評価

○荒牧 岳志, 木村 元, 小林 重信
東京工業大学

Realization and Assessment of Reward Driven Ring Robots

○Takeshi Aramaki, Hajime Kimura, Shigenobu Kobayashi
Tokyo Institute of Technology

Abstract: Reinforcement Learning is a self-adaptive learning framework and obtains control rules through interaction with the environment. In this paper, we treat a control method of a movement of a ring robot that consists of circular five-links. We apply reinforcement learning to the problem and choose Actor-Critic based Stochastic Gradient Ascent as its algorithm. 'Actor' module operates as an action selector using one Gaussian distribution. However, this method requires much time for its learning. This paper proposes actor module that acts in accordance with several Gaussian distribution. And we indicate that it is useful for reducing learning time through control tasks in the ring robot.

1 はじめに

ロボットの制御規則獲得など、教師なしの学習法として「強化学習」[3]が注目されている。だが、連続で膨大な行動を持つ実問題では、実時間で学習するのが困難であるなど、まだまだ問題がある。

強化学習の問題として、５つのリンクを環状に結んだロボット(fig.1) の前進動作を行う制御手法の獲得問題[1][2]を取り上げる。このロボットはでんぐり返し動作を繰り返すことによって前に進む事ができる。人間がこのモデルを設計する事は困難で、従来の制御手法を適用し前進させることが難しい。本研究では強化学習を用いてこのロボットの前進動作を獲得する。

強化学習のアルゴリズムとして代表的なものに、Q-learning[4]等があるが、本研究では容易に連続値の行動を扱う事ができ、状態観測の不完全性にも強いと言われる確率傾斜法に基づく actor-critic[5]を用いる。

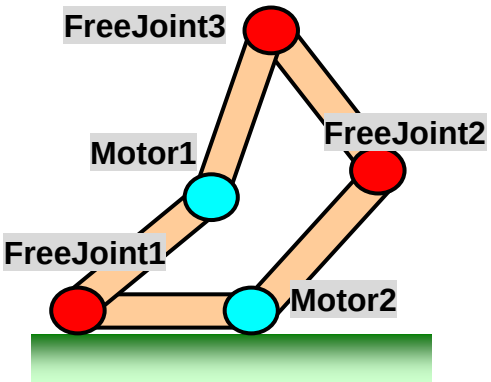

Fig. 1 ５角形環状ロボットのモデル

ほふくロボットや４足歩行ロボットなどの先行研究[6]として行われている正規分布を確率的政策として用いた actor-critic では、RBF を用いることで環境に応じて適応的に学習ができるものの、行動の探索空間が広く、学習時間がかかり過ぎるという問題が生じる。

また、別の先行研究[9]として確率的2分木の行動選択を用いたactor-criticが提案されている。この手法は、探索領域を階層的に縮め、学習時間を短くするのに有効な方法である。しかし、学習が進むにつれて最適と見積もられる政策が変化する場合には、問題が生じ、適応的な学習が困難であることを本研究で示す。

そこで、複数の正規分布基底を重みづけ足し合わせたものを政策の確率密度分布として用いる事で、行動空間中で有望と思われる領域を大まかに探索し、さらにその行動を微調整していく行動選択法および政策表現法を提案する。領域で探索することにより、探索領域を縮めることができ、学習速度を速くすることができる。また、その行動の微調整を行うことで、動的な環境の変化にも追従できるような適応的な学習が行える。

本研究では、ロボットの前進動作の学習を、提案手法を用いる事により、より少ない時間で獲得できる事を、シミュレーションおよび作成した実機での実験により示す。

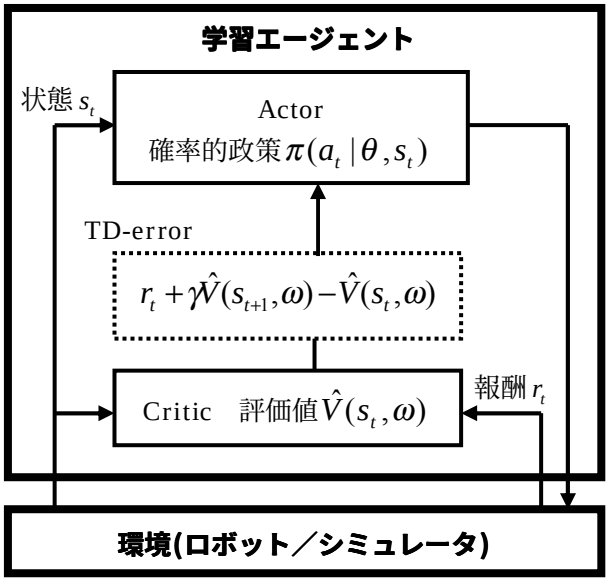

Fig.2　Actor-Critic の一般的な枠組

2　５角形ロボット

５角形ロボットは Fig.1 のようなモデルで５つの関節に２つのモータと３つの自由に回転ができる関節がある。5 角形ロボットは、自由度は２であるため、２つのモータで角度を決めることで形が決まる。

また、各関節を結ぶリンクには離散的な信号の傾斜センサがついており、ある程度の姿勢を知ることができる。しかし、これらのセンサは離散信号で on, off の 2 値しか分からない。よって、本問題は、現在の姿勢角や回転速度が完全には観測できないなど大きな隠れ状態を含んだ問題であり、ヒューリスティックにコントローラを設計するのは難しい。

3　強化学習アルゴリズム

3.1 Actor-Critic

本研究では、強化学習のアルゴリズムとして、容易に状態・行動空間が連続値を用いる事ができ、隠れ状態問題に対して頑健とされる actor-critic アルゴリズムを用いる。

Actor-critic アルゴリズムの一般的な枠組を Fig.3-1 に示す。状態入力に対して行動出力への確率分布 π を決定する actor と、状態に対する評価の推定値の保持・更新する critic の２つのモジュールから構成されている(fig.2)。θ,ω はそれぞれ、actor と critic で使用する内部パラメータである。

Fig.3 は、本ロボットに適用した eligibility trace(適正度の履歴)を用いた actor-critic の詳細である。Critic の評価値の更新は TD(λ) 法で TD($\lambda = \lambda_v$) である。対して actor の更新は TD(λ) とはやや異なる（詳しくは文献[5]参照)。

1. 学習エージェントが環境より時刻 t の状態 s を観測し、確率 $\pi(a_t \mid \theta, s_t)$ により行動 a を決め、実行する。θ は政策パラメータ。
2. 報酬 r_t 及び状態 s_{t+1} を観測する。TD－error を次の式で計算する。

$$\begin{aligned}(\text{TD}-\text{error}) = \\ r_t + \gamma\hat{V}(s_{t+1},\omega) - \hat{V}(s_t,\omega)\end{aligned} \quad (1)$$

$\gamma\ (0 \le \gamma \le 1)$ は割引率、value 値 $\hat{V}(s_t,\omega)$ は critic が出力した割引報酬の期待値を表す。ω は $\hat{V}(s_t,\omega)$ を推定するためのパラメータ。

3. TD 法で critic の $V(s_t,\omega)$ を更新。

$$\begin{aligned} e_v(t) &= \frac{\partial}{\partial\omega}\hat{V}(s_t,\omega) \\ \bar{e}_v(t) &\leftarrow e_v(t) + \bar{e}_t(t) \\ \Delta\omega(t) &= (\text{TD}-\text{error})\bar{e}_t(t) \\ \omega &\leftarrow \omega + \alpha_v\Delta\omega(t) \end{aligned} \quad (2)$$

e_v は ω の eligibility(適正度)を表し、$\bar{e}_v$ はその trace(履歴)を表す。α_v は critic の学習率。

4. TD-error で actor $\pi(a_t \mid s_t,\theta)$ を更新。

$$\begin{aligned} e_\pi(t) &= \frac{\partial}{\partial\theta}\ln\left(\pi(a_t \mid \theta, s_t)\right) \\ \bar{e}_\pi(t) &\leftarrow e_\pi(t) + \bar{e}_\pi(t) \\ \Delta\theta(t) &= (\text{TD}-\text{error})\bar{e}_\pi(t) \\ \theta &\leftarrow \theta + \alpha_\pi\Delta\theta(t) \end{aligned} \quad (3)$$

e_π は θ の eligibility を表し、$\bar{e}_\pi$ はその trace を表す。α_π は Actor の学習率。

5. Eligibility を下の式で割り引く。

$$\begin{aligned} \bar{e}_v(t+1) &\leftarrow \gamma\lambda_v\bar{e}_v(t) \\ \bar{e}_\pi(t+1) &\leftarrow \gamma\lambda_\pi\bar{e}_\pi(t) \end{aligned} \quad (4)$$

λ_v と λ_π $(0 \le \lambda_v, \lambda_\pi \le 1)$ は適正度の履歴の割引率。

6. $t \leftarrow t+1$ として 1.に戻る。

Fig.3　Actor-Critic のアルゴリズム

3.2 複数の正規分布基底による政策の提案

3.2.1 従来手法の問題点と提案手法の利点

４足歩行ロボットなどの先行研究[6]で用いられてきた１つの正規分布を政策の確率密度関数として使う従来手法は、学習初期において、行動探索はその正規分布の中心に行われることが多く、また、その正規分布基底の平均値を学習するのに多くの時間がかかる。これは、学習時間がかかる原因になっている。

また、2 分木などの離散的な行動の選択を行うアルゴリズム[9]では、適切な政策の確率密度分布が固定である時、効果を表し、後の実験で示すように５角ロボットのような学習とともに適切な政策の確率密度分布が動的に変わるという問題に対しては、弱いと考えられる。

そこで本研究では、複数の正規分布基底を重み付け足し合わせたものを政策の確率密度関数とすることを提案する。学習初期は各基底の重みを学習することで探索空間を大まかに絞ることができ、その後、残った正規分布基底の平均値および標準偏差の値を学習させる。探索空間を絞ることで学習を速くすることができ、基底の移動で、動的に変わる環境の変化にも追従できると考えられる。

3.2.2 政策表現

N 個あるうちの k 番目の正規分布基底の確率密度関数 $g(\mathbf{a}\,|\,k,\mathbf{x},\theta)$ は、平均値を $\mu_{k,i}$、標準偏差を $\sigma_{k,i}$ とすると次の式で表すことができる。

$$g(\mathbf{a}\,|\,k,\mathbf{x},\theta)=\prod_{i=1}^{d}\frac{1}{\sqrt{2\pi}\sigma_{k,i}}\exp\left\{-\frac{1}{2}\frac{(a_i-\mu_{k,i})^2}{\sigma_{k,i}^2}\right\} \quad (5)$$

政策の確率密度関数 $\pi(\mathbf{a}\,|\,\mathbf{x},\theta)$ を

$$\pi(\mathbf{a}\,|\,\mathbf{x},\theta)=\alpha\sum_{k=1}^{N}m_k g(\mathbf{a}\,|\,k,\mathbf{x},\theta) \quad (6)$$

とする。ここで、α は正規化定数で、

$$\alpha=\frac{1}{\int_{\mathbf{a}\in\mathbf{A}}\sum_{k=1}^{N}m_k g(\mathbf{a}\,|\,k,\mathbf{x},\theta)\,d\mathbf{a}} \quad (7)$$

である。また、ここで、m_k は k 番目の正規分布基底に対する重みで、$\mathbf{x}$ は状態 $\mathbf{s}$ の特徴ベクトルである。各正規分布パラメータは以下の式で計算を行う。

$$\begin{aligned}\mu_{k,i}&=\frac{1-\exp(-\sum_j\theta_{\sigma k,i,j}x_j)}{1+\exp(-\sum_j\theta_{\sigma k,i,j}x_j)}\\ \sigma_{k,i}&=\frac{1}{1+\exp(-\sum_j\theta_{\sigma k,i,j}x_j)}\\ m_k&=\exp(-\sum_j\theta_{\sigma k,i,j}x_j)\end{aligned} \quad (8)$$

3.2.3 行動選択の手順

行動選択の手順は次のようになる。

1. まず、$p(k'|\,\mathbf{x},\theta)=m_{k'}\Big/\sum_{j=1}m_j$ の確率で k' 番目の正規分布基底を選択する。
2. 正規分布 $N(\mu_{k'},\sigma_{k'}^2)$ の確率で行動 $\mathbf{a}$ を選ぶ。
3. 選択した行動 $\mathbf{a}$ が行動の定義域 $\mathbf{A}$ の中でなければ、1.の選択からやりなおす。

この選択法によって得られた行動 $\mathbf{a}$ は式(6)を満たす。

4 実験および結果の考察

５角形環状ロボットの実験は、シミュレーション及び製作した実機で行う。実験の目的は、強化学習を用いて、5 角形ロボットを前進させる動作を実ロボットで実時間に学習させることである。

4.1 ５角形ロボットへのアルゴリズムの実装

５角形ロボットに actor-critic アルゴリズムを実装する上で、次のような設定を行った。

・行動表現

強化学習の行動は２次元ベクトル $\mathbf{a}_t$ で表すことにし、その要素を $\mathbf{a}_t=\{a_1,a_2\}$ で表す。a_1,a_2 は $[-1,+1]$ の連続値の値域を持つ。５角形ロボットはその形からモータの角度に制約条件があり、a_1,a_2 は次の式で角度 θ_1,θ_2 [rad]に写像し、これをサーボモータの角度指令値として用いる。

$$\begin{aligned}\theta_1&=\frac{\pi}{12}a_1a_2+\frac{9\pi}{24}a_1+\frac{\pi}{6}a_2+\frac{15}{24}\pi\\ \theta_2&=-\frac{\pi}{12}a_1a_2+\frac{9\pi}{24}a_1-\frac{\pi}{6}a_2+\frac{15}{24}\pi\end{aligned} \quad (9)$$

・状態表現

状態を７次元ベクトル $\mathbf{s}_t=\{s_1,s_2,s_3,\cdots,s_7\}$ とする。本研究で用いているサーボモータは角度を測ることができないので、s_1,s_2 は時刻 $t-1$ の a_1,a_2 の値をそのまま使うことにする。$s_3\sim s_7$ は現在のセンサの状

態とする。s_1 s_2 は、$[-1,+1]$ の値域を持つ連続値で、$s_3 \sim s_7$ は 0 または 1 の離散値である。

・報酬設定

本研究ではロボットが前進することを目標にしているので、報酬は単位時間にロボットが前進した距離とする。ただし、ロボットには距離を測る装置はないので、センサの状態の遷移から実際に進んだ距離を推定する。ロボットが 1 回転すると 10 の報酬が入るようにする。

また、各アルゴリズムでは RBF を使って状態 $\mathbf{s}$ を特徴ベクトル $\mathbf{x}$ に局所化して用いる。ここで、i 番目の RBF のガウス関数を、

$$f_{gi}(\mathbf{s}) = \exp\left(-\sum_{j=1}^{n} \frac{(s_i - \mu_{gi,j})^2}{2\sigma_{gi,j}^2} \right) \qquad (10)$$

とする。μ_{gi} ,σ_{gi} はガウス関数の平均値、標準偏差である。本研究では RBF の出力は以下のように正規化する。

$$x_i = f_{gi}(\mathbf{s}) \Big/ \sum_{j=1}^{n'} f_{gj}(\mathbf{s}) \qquad (11)$$

本研究では、ガウス関数の平均値 μ_{gi} と標準偏差 σ_{gi} は学習しないことにする。そして、各 RBF ユニットを状態の s_1 s_2 軸に対しては 3 個、$s_3 \sim s_7$ 軸に対しては 2 個、格子状に配置した。計 $3^2 \times 2^5 = 288$ 個の RBF ユニットの標準偏差は一律 $\sigma_{gk,i} = 0.2$ とする。

また、断りのないときは、actor のパラメータ $\boldsymbol{\theta}, \mathbf{e}_\pi$ と critic のパラメータ $\boldsymbol{\omega}, \mathbf{e}_v$ は全て 0 に初期化する。

4.2 実験結果

まず、シミュレーションを用いて、正規分布を使った従来手法[6]と 2 分木を使った従来手法[9]、複数の正規分布基底を用いた提案手法の3つについて実験を行った。

割引率を $\gamma = 0.95$ とし、学習率は $\alpha_\pi = 0.5$、$\alpha_v = 0.1$、$\lambda_\pi = \lambda_v = 1.0$ とした。複数の正規分布を用いた手法では、基底の数を $N = 25$ とし、初期配置は格子状にした。2 分木による手法では、各行動軸を 16 分割し、2 分木構造で確率的に選択するようにした。学習の 1step を 300ms とし、それぞれ 100000step を 5 試行行い、平均を求めた。

実験結果を fig.5 に示す。また、見やすくするために、1000step の移動平均をとっているものをグラフにした。縦軸の reward/step は、ロボットの前進速度を表す。

正規分布を用いた従来手法と提案手法では、きれいに学習が行え、しっかりしたロボットの前進動作が得られている。そして、提案手法は従来手法の約 2 倍の学習速度が得られた。

しかし、2 分木を使った Actor-Critic では、途中、ロボットの回転動作が止まることがあり、学習がうまく行えなかった。

Fig.4 は、正規分布を使った従来手法と提案手法におけるある状態における確率密度分布を表し、×印は正規分布基底の平均値を表し、円は標準偏差を表す。

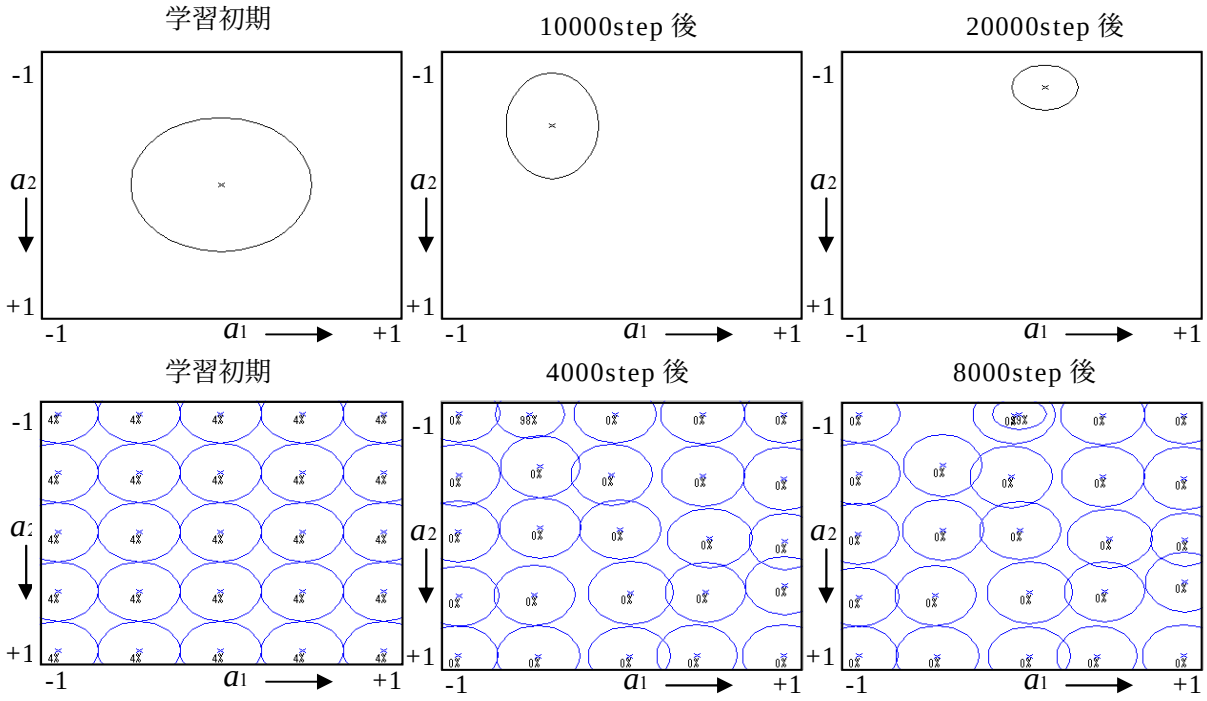

Fig.4 各学習ステップ数における基底の動き。従来手法 $N = 1$(上段)　提案手法 $N = 25$(下段)

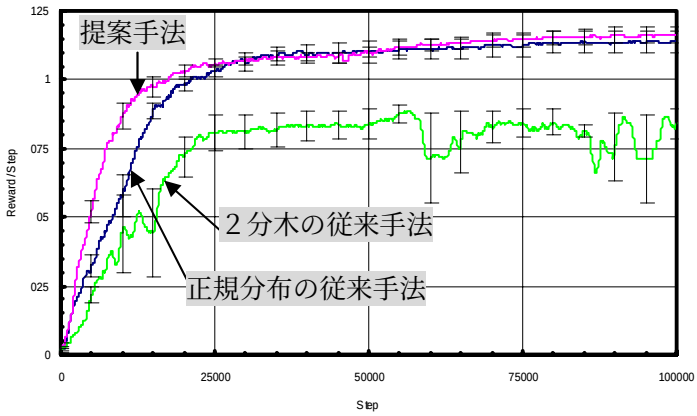

Fig.5 各手法によるシミュレーション結果

提案手法は正規分布を用いた従来手法に比べ学習速度が速くなっている。また、2 分木を用いた actor-critic では、たまに動作が固まることがあり、分散が大きくなっている。

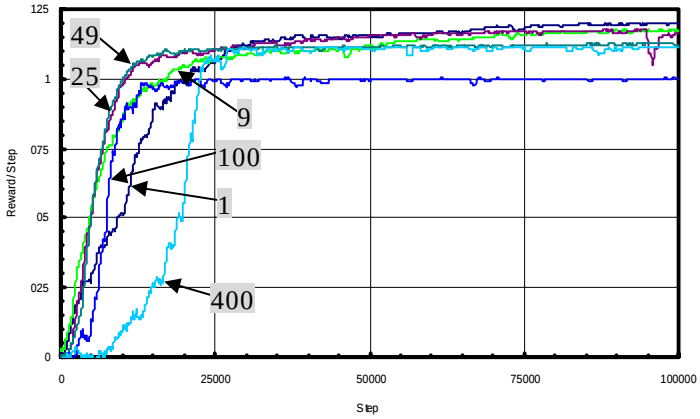

Fig.6 複数基底を用いた手法の学習曲線

グラフ中の数値は正規分布基底の数 N を表す。基底数 1 の時は正規分布を用いた従来手法と等しくなる。基底数が 25,49 といったあたりで学習速度が最も速くなっている。

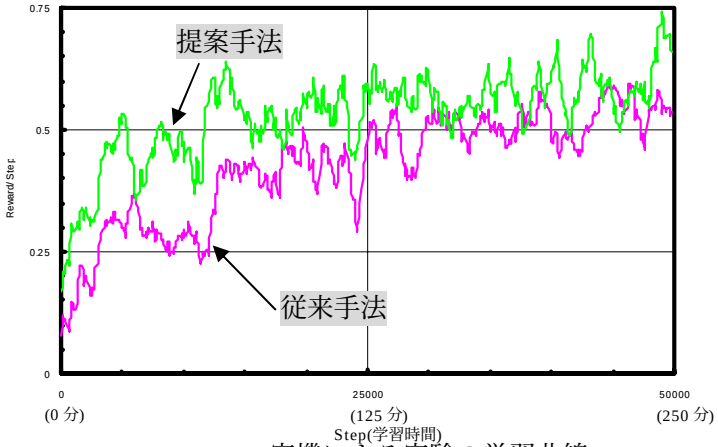

Fig.7 実機による実験の学習曲線

実機ではノイズが多くのり、シミュレーションほどの性能はでなかったが、提案手法は正規分布を用いた従来手法に比べ、学習を速くすることに成功した。

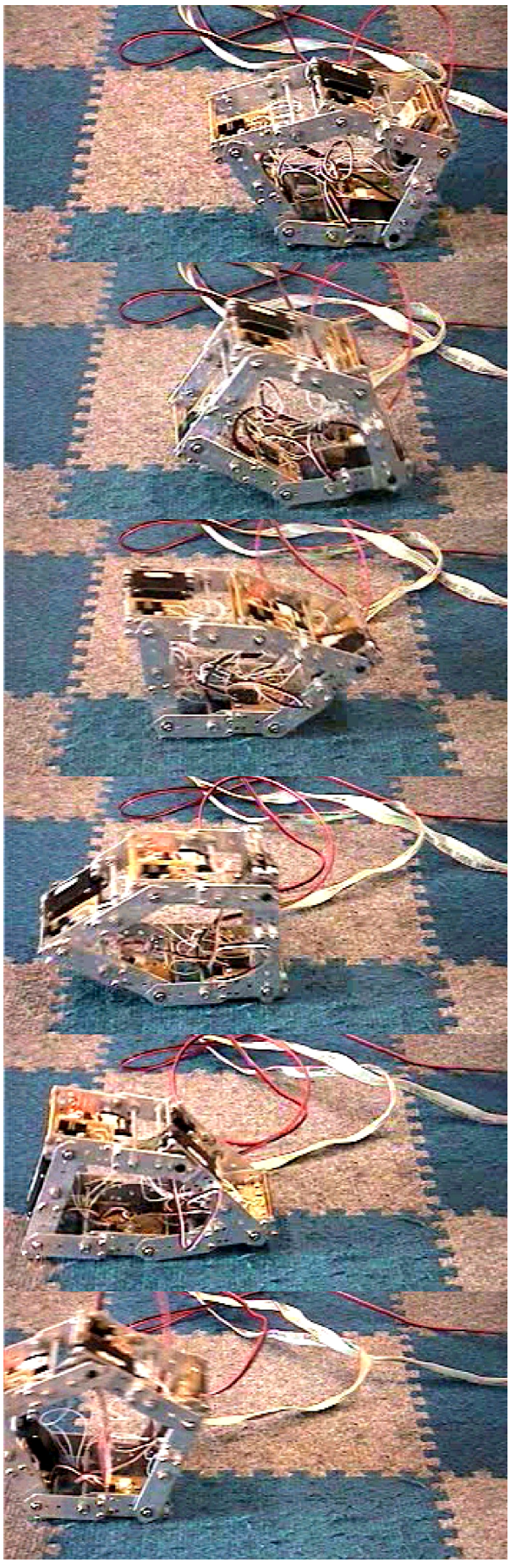

Fig.8 ロボットが回転している様子

Fig.4(下段)に示すように 4000step で最適と思われる確率密度分布が 8000step では移動をしている。これは、学習とともに最適と思われる解が移動していることを表す。そのため、２分木による手法では、解の移動がおきると、２分木の上位層を更新する必要が出てくるため、学習が不安定になってしまうと考えられる。

Fig.6 は、提案手法において、正規分布基底の数 N を変えてシミュレーション実験したものである。$N=1$ のとき、正規分布を用いた従来手法と同じになる。

基底数がある程度多いと学習は速くなるが、多すぎると逆に学習は遅くなる。これは、この提案手法の学習時間が(基底の重みの学習時間)+(基底の平均・標準偏差の学習時間)となっているためであり、基底数が小さいと(基底の重みの学習時間)は短くできるが、(基底の平均・標準偏差の学習時間)が長くなってしまう。逆に規定数が大きいと、(基底の平均・標準偏差の学習時間)を短くできる代わりに、(基底の重みの学習時間)は長くなってしまう。適切な基底数を設定する必要がある。

Fig.7 は、製作した実機を用いて、正規分布を用いた従来手法と提案手法において、50000step(250 分)、１試行の実験を行った学習結果である。従来手法では学習に約４時間かかっていたが、提案手法では約２時間で学習ができている。

Fig.8 は学習した実機が転がっている様子である。

5　まとめ

・強化学習を使うことにより、パラメータの測定なしに５角形ロボットの前進動作が得られた。
・５角形ロボットの前進動作の学習問題は、最適と見積もる行動の確率密度分布が学習とともに動くということが分かった。２分木を用いた actor-critic では、うまく学習ができないことが示した。
・複数の正規分布基底を組み合わせる政策表現を新たに提案した。実機実験では学習時間を約半分にでき、本問題の様な動的に変化する環境でも追従できた。

6　今後の課題

実機のロボットではもっと速い学習が望まれる。提案手法の基底の選択に２分木などの選択法を組み合わせることで、より学習速度を上げることが期待できる。

また、学習の 1step を一律 300ms としてきたが、モデルをセミマルコフ化するなどして、この学習のステップ時間をも学習することで、学習速度を速くすることも可能ではないかと考えられる。

References

[1] 森 治、小峰 晋一郎、村木 誠一郎、小俣 透：劣駆動結合によるパラレルメカニズムへの形態変化：垂直面内での劣駆動シリアルリンク系の終端制御, 第 19 回日本ロボット学会学術講演会(CD-ROM)

[2] Vunthichai Ampornaramveth：On Motion Generation of Autonomous Decentralized Mechanical Systems Using Genetic Methods, ph. D. thesis, Tokyo Institute of Technology (1999)

[3] R.S. Sutton and A.G. Barto：Reinforcement Learning , An Introduction. A Bradford Book.　The MIT Press (1998)

[4] C.J.C.H Watkins and P.Dayan：Technical note :　Q-Learning, Machine Learning, Vol. 8, pp. 279-292 (1992)

[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 のもとでの強化学習, 人工知能学会, Vol. 15, No.2, pp.267-275(2000)

[6] 山下 透, 木村 元, 小林 重信：強化学習による多足歩行ロボットの実現, 計測自動制御学会, 第 13 回自律分散システムシンポジウム資料 pp 111-116(2001)

[7] Hajime Kimura and Shigenobu Kobayashi : An Analysis of Actor/Critic Algorithms using Eligibility Traces: Reinforcement Learning with Imperfect Value Functions. In 15th International Conference on Machine Learning , pp. 278-286, 1998.

[8] 柴田克成, 前原伸一, 杉坂政典, 伊藤宏司：Gauss-Sigmoid ニューラルネットワーク, 第 12 回自律分散システムシンポジウム資料, pp.133-138(2001)

[9] 木村 元, 小林 重信：確率的 2 分木の行動選択を用いた Actor-Critic アルゴリズム :多数の行動を扱う強化学習 , 計測自動制御学会誌, Vol.37, no.12, pp.1147--1155 (2001).

Appendix

提案手法の actor の eligibility は以下のように計算できる。

$$e_\pi(t)_{\mu k,i,j} = x_j \frac{(1+\mu_{k,i})(1-\mu_{k,i})}{2}\frac{1}{\sigma_{k,i}^2}\left\{\frac{m_k g(\mathbf{a}\mid k,\mathbf{x},\boldsymbol{\theta})\cdot(a_i-\mu_{k,i})}{\sum_{k'} m_{k'} g(\mathbf{a}\mid k',\mathbf{x},\boldsymbol{\theta})} - \frac{\int_{\mathbf{a}'\in\mathbf{A}} m_k g(\mathbf{a}'\mid k,\mathbf{x},\boldsymbol{\theta})\cdot(a_i'-\mu_{k,i})d\mathbf{a}'}{\sum_{k'}\int_{\mathbf{a}'\in\mathbf{A}} m_{k'} g(\mathbf{a}'\mid k',\mathbf{x},\boldsymbol{\theta})d\mathbf{a}'}\right\}$$

$$e_\pi(t)_{\sigma k,i,j} = -x_j \frac{(1-\sigma_{k,i})}{\sigma_{k,i}^2}\left\{\frac{m_k g(\mathbf{a}\mid k,\mathbf{x},\boldsymbol{\theta})\cdot\left(\sigma_{k,i}^2-(a_i-\mu_{k,i})^2\right)}{\sum_{k'} m_{k'} g(\mathbf{a}\mid k',\mathbf{x},\boldsymbol{\theta})} - \frac{\int_{\mathbf{a}'\in\mathbf{A}} m_k g(\mathbf{a}'\mid k,\mathbf{x},\boldsymbol{\theta})\cdot\left(\sigma_{k,i}^2-(a_i'-\mu_{k,i})^2\right)d\mathbf{a}'}{\sum_{k'}\int_{\mathbf{a}'\in\mathbf{A}} m_{k'} g(\mathbf{a}'\mid k',\mathbf{x},\boldsymbol{\theta})d\mathbf{a}'}\right\}$$

$$e_t(t)_{mk,j} = x_j m_k\left\{\frac{g(\mathbf{a}\mid k,\mathbf{x},\boldsymbol{\theta})}{\sum_{k'} m_{k'} g(\mathbf{a}\mid k',\mathbf{x},\boldsymbol{\theta})} - \frac{\int_{\mathbf{a}\in\mathbf{A}} g(\mathbf{a}'\mid k,\mathbf{x},\boldsymbol{\theta})d\mathbf{a}'}{\sum_{k'}\int_{\mathbf{a}\in\mathbf{A}} m_{k'} g(\mathbf{a}'\mid k',\mathbf{x},\boldsymbol{\theta})d\mathbf{a}'}\right\}$$

しかし、$\sigma_{k,i}\to 0$のとき動作が不安定になる[7]。そこで、本研究では$e_\pi(t)_{\mu k,i,j}, e_\pi(t)_{\sigma k,j}$に$\sigma_{k,i}{}^2$をかけたものを用いた。