論 文 Technical Papers

重み付けされた複数の正規分布を用いた政策表現

最適行動変化に追従できる実時間強化学習と環状ロボットへの適用

A Policy Representation Using Weighted Multiple Normal Distribution

Real-time Reinforcement Learning Feasible for Varying Optimal Actions

木村 元 Hajime Kimura
東京工業大学 Tokyo Institute of Technology
gen@fe.dis.titech.ac.jp, http://www.fe.dis.titech.ac.jp/~gen/

荒牧 岳志 Takeshi Aramaki
東京工業大学 Tokyo Institute of Technology
aramaki@fe.dis.titech.ac.jp

小林 重信 Shigenobu Kobayashi
東京工業大学 Tokyo Institute of Technology
kobayasi@dis.titech.ac.jp, http://www.fe.dis.titech.ac.jp/

keywords: reinforcement learning, hierarchical representation, actor-critic, robotics

Summary

In this paper, we challenge to solve a reinforcement learning problem for a 5-linked ring robot within a real-time so that the real-robot can stand up to the trial and error. On this robot, incomplete perception problems are caused from noisy sensors and cheap position-control motor systems. This incomplete perception also causes varying optimum actions with the progress of the learning. To cope with this problem, we adopt an actor-critic method, and we propose a new hierarchical policy representation scheme, that consists of discrete action selection on the top level and continuous action selection on the low level of the hierarchy. The proposed hierarchical scheme accelerates learning on continuous action space, and it can pursue the optimum actions varying with the progress of learning on our robotics problem. This paper compares and discusses several learning algorithms through simulations, and demonstrates the proposed method showing application for the real robot.

1. はじめに

ロボットの制御規則をロボット自身が獲得していく手段として，強化学習は有望なアプローチの一つであり，代表的な強化学習法として Q-learning [Watkins92] が知られている．しかし実ロボットの連続な状態・行動空間において，試行錯誤を通じて適切な行動を獲得する場合，Q-leaning をそのまま用いるのでは膨大な学習時間を要するため，様々な手法が提案されている．Q-learning の状態-行動評価関数 (Q 関数) を連続空間へ拡張する手法として，CMAC 等の線形アーキテクチャを用いる方法 [Sutton98] やメモリーベースな方法 [Santamaria98][深尾 98]，Fuzzy-Q [堀内 98]，正則化理論を用いて補完する Q-learning [深尾 98] などが提案されている．しかし，これらの手法では行動選択において行動空間をグリッドに離散化して扱っており，高次元の行動空間においてきめ細かな行動の学習を試みると，行動選択の計算量が膨大になるという問題がある．

連続な行動空間での行動選択処理において有利な手法として actor-critic 法 [Sutton98] がある．actor-critic 法は，状態（あるいは状態-行動対）を評価するための critic と呼ばれる部分と，ある政策に従って行動を選択し critic の評価値を利用して政策を改善する actor という部分より構成される．政策表現に任意性があるために連続値の行動出力への対応が容易である．Actor-critic アルゴリズムを用いた neuro-fuzzy コントローラ [Lin96] [Lin99] や，政策を EM 法によって獲得する方法 [石井 2000][Yoshimoto2000] などが提案されている．この actor-critic アルゴリズムを下位層とし，上位層に Q-learning を用いた階層化によって効率良く学習する手法 [森本 2001] や，さらに環境の遷移モデルを学習して政策学習に利用する方法 [鮫島 2001] なども提案されている．

しかし，上記の方法は状態（あるいは状態-行動対）の評価値を得た後に政策を得る，あるいは改善することを基本としており，状態観測の不完全性などにより正確な評価値が得られない場合に学習がうまく行かないという問題がある．著者らは actor に適正度の履歴を用いることで，状態観測に不完全性が存在する環境における学習

が可能な actor-critic 法 [木村 2000] を提案し，actor の行動選択に正規分布を用いて４足ロボットの歩行動作獲得 [木村 2002] へ適用したり，actor の政策を確率的 2 分木で表現する階層的な actor-critic 法を提案した．

本論文では，５リンク環状ロボットの移動動作獲得問題を取り上げ，実機による試行錯誤が許容できる実時間で強化学習することを目指す．本学習問題では，１）学習の実時間性，２）状態観測にノイズが存在する環境の扱い，３）最適な行動が学習の進行に伴って変化していく環境の扱い，以上の３点をクリアすることが求められる．そこで本論文では，前述のように連続状態-行動空間を持ち，状態観測にノイズが存在する問題に対し有望な接近法である actor-critic 法を適用する．しかし，従来の actor-critic による実装では，上記の３つの問題点を同時にクリアすることが困難だった．そこで actor の政策表現を工夫し，上位が離散的，下位が連続的行動選択を行う階層的な構造とする方法を提案する．

正規分布を actor の確率的政策とする先行研究の actor-critic では，有望と思われる探索領域を絞り込んでいくまでの過程に時間がかかり過ぎる．別の先行研究で提案された確率的 2 分木による階層的な actor-critic 法では，階層化によって有望と思われる探索領域をすみやかに絞り込んで効率的な学習が行えるが，学習が進むにつれて最適と見積もられる政策や行動が変化する場合には，問題が生じる．本論文の提案手法は，行動空間中で有望と思われる領域を上位層の行動選択で大まかに探索し，さらにその行動を下位層で微調整していくことにより，効率的な学習を行えると同時に動的な環境の変化にも追従することが期待できる．

2. Actor-Critic における政策表現

2･1 Actor-Critic アルゴリズム

Actor-critic アルゴリズム [Sutton98] は連続空間における強化学習法として実績がある [石井 2000][森本 2001][鮫島 2001]．Actor は，状態から行動への確率分布である”**確率的政策** (**stochastic policy**)” に従って行動を選択する．Critic は，actor の政策のもとでの各状態の評価値 (**value**) すなわち割引報酬の期待値を推定する．Actor は，critic で計算される *temporal difference (TD)* エラーを用いて政策を改善する．本論文のロボットへ適用した actor-critic では，actor と critic の両方に適正度の履歴 (eligibility trace)[Sutton98] を用いる．そのため，従来のアルゴリズムに比べて隠れ状態問題に対してロバストである [木村 2000]．隠れ状態問題は，MDP の状態観測にノイズなどの不完全性や，状態変数の一部しか観測できないという部分観測性が加わるだけで発生するため，ロボットの学習においてたびたび直面する．隠れ状態問題では環境のマルコフ性が保障されないため，MDP のマルコフ性に依存する従来のアルゴリズムでは対処が困難である．また，actor-critic は連続な行動空間における学習に適しているが，本研究では政策表現を工夫することで，性能改善を試みる．

2･2 Actor-Critic における従来の政策表現

§1 確率的政策に正規分布を用いる方法

Actor-critic の確率的政策表現方法として正規分布を用いる方法は，連続な行動空間を扱うには最も単純であり，４足ロボットなどの学習に用いられてきた [木村 2000]．ここでは行動空間に上下界が存在する場合に有効であることが示されている reject-and-resample 法を比較手法として用いた [木村 2002]．

しかし確率的政策に正規分布を用いる方法では，行動１次元あたりたった１つの正規分布で探索を行う．そのため学習初期において，行動探索はその正規分布の中心に行われることが多く，最適な行動が行動空間の端に存在するような場合には，学習によって正規分布の平均値 μ を中心からそこまで動かしていかなければならないため，多くの学習時間ステップを要する．これは，実時間で実ロボットの学習を行う上で大きな障害となる．

§2 確率的政策に２分木構造を用いる方法

本論文で扱う問題のように，実問題の多くは行動空間に位相構造が存在するため，物理量的に近接･類似する行動をグループ化して階層的することにより，学習や意思決定を効率良く行うことができる．Actor-critic において，数十以上の多数の類似する行動を効率良く扱うための階層的行動選択方法の１つとして，２分木構造を持つパラメータ化された確率的政策表現がある [木村 2001]．この方法では，まず上位層がおおまかな行動選択をすみやかに学習し，次々と下位層の学習が進行していくことで，有望と思われる行動が絞り込まれていく．このアルゴリズムでは離散的な行動選択しか扱えないため，本論文で扱うロボットの学習では，行動空間をグリッド状に区切り，代表点を選択する．２分木構造の政策では，各階層において確率的に２者選択を行うが，その確率分布関数は状態変数 s と政策パラメータベクトル θ の微分可能な非線形関数として表される．

２分木構造の階層的な行動選択を行う方法では，行動空間を階層的にグループ化して上位層から学習していくため，ランダム選択を行う初期政策から学習する場合には効率が良いが，獲得した行動を再学習によって修正することは，特に上位層においては困難である．従って，最適な行動が時間とともに変化していく動的な環境では，特に上位層でグループ化した行動領域をまたぐように変化していく場合，学習によって適応していくことは難しいと考えられる．

2·3 複数の正規分布による階層的政策表現を用いた Actor-Critic アルゴリズムの提案

§1 基本的なアイデアと特徴

従来の政策表現の問題点を解決するため，重み付けされた複数の正規分布を用いて政策表現を行う方法を提案する．これは上位層において，複数の正規分布のうち 1 つを重み変数の比率に基づいて確率的に選択し，下位層において従来の正規分布に基づく手法によって連続値の行動を選択する．つまり上位層では行動空間を大まかに離散化して行動選択を行い，下位層では連続的な行動選択を行うことで探索空間を効率良く絞り込み，学習を加速する効果が得られるものである．上位層で選択する離散化された行動の領域は，下位層の正規分布の学習によって適応的に変化するため，動的に変化する環境にも適応できるのも本手法の大きな特徴である．

本手法と非常に類似した手法として，NGnet と呼ばれる混合正規分布を政策表現に用いて，状態-行動評価関数（Q 関数）のボルツマン分布に近づくよう EM 法で政策を更新する actor-critic 法がある [Yoshimoto2000]．NGnet による政策表現は，連続空間における任意の分布を表現することを志向しているが，本提案手法は上位層では離散的行動選択を行い，下位層では正規分布のような連続空間での行動選択を行うことで効率良く学習することを志向している．ここで挙げた上位層の行動選択はフラット選択なので結果的に混合正規分布や NGnet に等しいが，上位層の離散的行動選択に 2 分木行動選択などを取り入れることも可能な枠組みである．

§2 提案手法の具体的な処理手順

本論文では，actor に適正度の履歴を用いる actor-critic アルゴリズム [木村 2000] を使用し，提案する政策表現を以下に示す方法にて実現する．N コあるうちの k 番目の正規分布を以下の式で表す：

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

このとき政策の確率密度関数 $\pi(\mathbf{a}|\mathbf{s},\theta)$ は，

$$\pi(\mathbf{a}|\mathbf{s},\theta) = \alpha \sum_{k=1}^{N} m_k g_k(\mathbf{a}|k,\mathbf{s},\theta) \quad (2)$$

ただし m_k は k 番目の正規分布 g_k に対する重みで，政策パラメータ θ および状態変数 $\mathbf{s}$ の関数である．また α は形式的な正規化定数で，$\alpha = 1/\int_{\mathbf{a}\in\mathcal{A}} \sum_{k=1}^{N} m_k g_k(\mathbf{a}|\mathbf{s},\theta)d\mathbf{a}$ である．状態変数 s および政策パラメータ θ を用いて $\mu_{k,i}$, $\sigma_{k,i}$ および m_k を計算する方法は実験設定に示す．適正度の具体的計算については文献 [木村 2000] に従って微分するだけなので省略する．

具体的な行動選択の手順は，まず重み変数 $m_{k'}$ に比例した確率 $p(k'|\mathbf{s},\theta) = m_{k'}/\sum_{j=1}^{N} m_j$ に従って k' 番目の多変数正規分布を選択し，その分布 $N(\mu_{k'},\sigma_{k'}^2)$ に従って行動 $\mathbf{a}$ を選択する．ただし選択した行動 $\mathbf{a}$ が行動の定義域 $\mathcal{A}$ の中でなければ，行動選択は正規分布を選択する部分からやり直す．従って，最終的に k' 番目の正規分布が選択され実行される確率は $p(k'|\mathbf{s},\theta)$ とは一致せず，定義域 $\mathcal{A}$ の端に存在する正規分布は $p(k'|\mathbf{s},\theta)$ で示される選択確率よりも選択されにくい傾向がある．しかしこの行動選択方法によると行動 $\mathbf{a}$ の確率は式 (2) の分布になる．

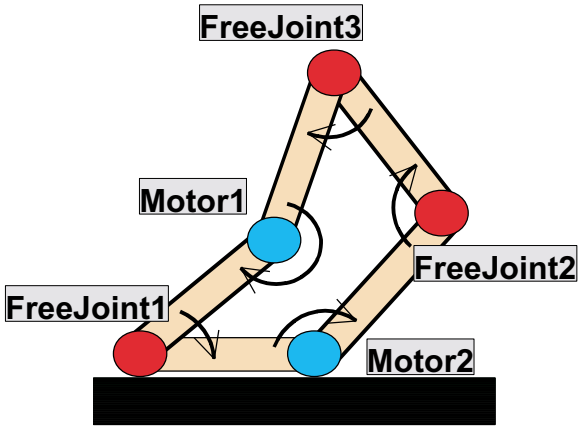

図 1 5 リンク環状 2 自由度ロボット．位置制御サーボモータ 2 個によって形を変える．

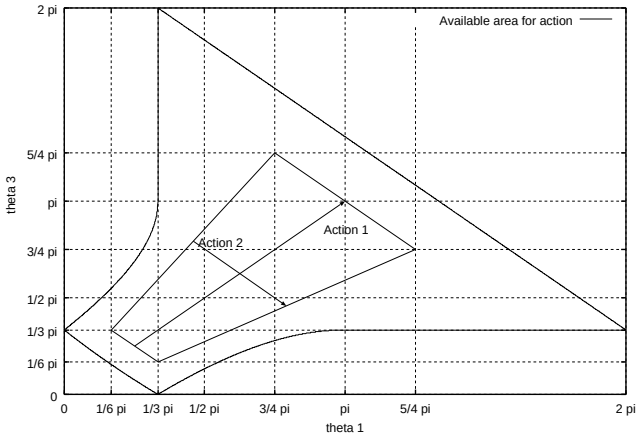

図 2 ロボットのモータを取り付けた関節角度の制約条件と，エージェントの選択する行動の出力範囲．

3. 実　　験

3·1 5 リンク環状 2 自由度ロボット

本論文では図 1 に示す 5 リンク環状 2 自由度ロボットを扱う．各リンクは 1 辺が約 8.5cm のフレームで構成され，それぞれ 5 つの関節によって接続される．そのうち 2 つの関節は位置制御モーターによって駆動され，残りの 3 つの関節は自由に回転できる．2 つのモータによって環状リンクの形は一意に決まる．モータが取り付けられている関節の角度 (θ_1,θ_3) は，機械的な制約より図 2 に示す範囲しか動けない．この範囲の導出は付録に示す．

本ロボットでは，行動出力の領域および状態領域を，こ

の物理的制約を満たす範囲のうち，図中に示す台形領域に限定する．この台形は $(\theta_1,\theta_3)=(\pi/3,\pi/6),(5/4\pi,3/4\pi),(3/4\pi,5/4\pi),(\pi/6,\pi/3)$ の４点で囲まれている．コントローラが選択する行動は，(a_1,a_2) の２次元空間で $-1\leq a_1,a_2\leq 1$ の矩形領域中の点をとり，この座標をモータの関節角度の台形領域へ以下の式を用いて変換する：

$$\begin{aligned}\theta_1 &= \frac{1}{3}\pi a_1 a_2+\frac{7}{12}\pi a_1+\times\frac{1}{6}\pi a_2+\frac{1}{6}\pi\\ \theta_2 &= \frac{1}{3}\pi a_1 a_2+\times\frac{11}{12}\pi a_1-\times\frac{1}{6}\pi a_2+\frac{1}{3}\pi \quad (3)\end{aligned}$$

各関節を結ぶリンクには傾斜センサが付いており，リンクの外側が上方を向いているかどうかを on/off の離散信号で示すので，これらの状態観測を使えば，ある程度の姿勢を知ることができる．また，関節を動かすのは位置制御のサーボモータなので，現在の関節角度についての情報は，１ステップ前に出力した行動値で代用する．ただし，モータが追従できない場合には実際の角度と食違う場合がある．モータの回転速度は観測できない．

各時間ステップにおいてエージェント（コントローラ）は現在の関節角度についての情報および傾斜センサの値（離散値）を観測して現在の状態とし，政策関数に従って行動選択する．よって状態空間は連続２次元空間×離散 5bit 空間，行動は２次元連続空間である．学習目標は，なるべく速く前進する制御規則を獲得することである．

エージェントが行動を選択後，約 0.3 秒後に状態遷移結果として報酬が与えられ，次の時間ステップへと進む．報酬は学習目標である「前進」を反映させる．各リンクに取り付けられた５つの離散値センサは，ロボットが前または後へ転がればそれを反映した遷移をするので，これを利用して報酬を計算する．前進する方向のリンクが地面へ完全に接地する度に 2 ずつ報酬が与えられ，ロボットが 1 回転分前進すれば合計 10 の報酬が与えられる．しかしロボットはその時々の報酬を得るように動けば良いというわけではなく，滑らかに前進するため，将来に渡り多くの報酬が得られる状態へ遷移することが求められる．

本ロボットにおいて，転がって前進する制御規則を考えることは，設計者にとって困難である．なぜなら，５角形のうち２箇所にモーターがあり，左右非対称になっているため，転がりだす姿勢はどのリンクが底にあるかや回転速度などによって異なる．離散センサは 5bit なので，最大 $2^5=32$ 通り，実験的な観測でも少なくとも 10 通りの離散センサのパターンへの対応が必要である．各パターンにおいて，それぞれ２次元連続状態空間が存在し，そこで２次元連続行動空間での意思決定を行うのは容易ではない．

本ロボットの学習問題は，on/off の離散信号を出力する簡易なセンサと関節の角度を用いて姿勢を観測しており，回転速度が観測できないなどの影響により，最適と思われる行動が政策の学習に伴って変化していくように見える現象が起きる．さらに実機の場合，学習が進んで高速に回転するようになるほど衝撃などによってセンサにノイズが入り，状態観測の不完全性が増していく問題を抱えている．

3･2　強化学習問題の設定

エージェントが選択する行動は２次元ベクトル $\mathbf{a}=(a_1,a_2)$ で表される．この値はロボットのモータ駆動関節の角度を指示し，それぞれ $[-1,1]$ の有界領域だが，これらは直接モータの角度になっているわけではなく，図 2 および式 (3) に示す変換を行った上で２つのサーボモータの角度としている．状態 $\mathbf{s_t}$ は７次元のベクトル $\mathbf{s_t}=(s_1,s_2,\ldots s_7)$ で表され，s_1 と s_2 はロボットのモータ駆動関節の角度指示値で，１ステップ前に出力した行動の値が入る．s_3〜s_7 は姿勢センサの値（0 or 1）が入る．

実機による実験だけでは，学習に多くの時間がかかってしまうため，アルゴリズムの検証や，学習に適切と思われるパラメータの選定，行動学習に適切と思われる状態特徴ベクトルの生成方法についての検討などができない．そこで，短い時間でこれらの検証を行うため計算機シミュレータを作成した．これは実機についての大まかなダイナミクスを模擬できるが，モータのトルク制限が無かったり，全てのリンクが均質で同じ重さであるなど，必ずしも実機を忠実に再現してはいない．実験はシミュレータ上および実機の両方で行った．

3･3　ロボットへの学習アルゴリズム実装

§1　状態の特徴ベクトルの生成

状態 $\mathbf{s_t}$ は７次元のベクトル $\mathbf{s_t}=(s_1,s_2,\ldots s_7)$ で表されるが，状態空間中に十分に細かく局所化された複雑な政策や状態価値関数を表現するため，状態ベクトル $\mathbf{s_t}$ と RBF(radial basis function) を用いて n' 次元の特徴ベクトル $\mathbf{x}=(x_1,\ldots,x_{n'})$ を生成する．特徴ベクトル $\mathbf{x}$ の生成方法の設計は重要な問題であるが，本論文では研究の対象外とし，シミュレータの実験によって十分な表現能力を有すると思われる RBF の数と配置を試行錯誤的に以下のように与えた．各 RBF ユニットは状態変数 s_1,s_2 軸に対しては３個ずつ，s_3〜s_7 軸に対しては２個ずつ格子状に配置した．よって計 $3^2\times 2^5=288$ 個の RBF ユニットを用いて，288 次元の特徴ベクトル $\mathbf{x}$ を生成する．RBF としてはガウス関数を用いた．i 番目の RBF のガウス関数 $f_{gi}(s)$ は，以下の式で表される：

$$f_{gi}(s)=\exp\left(-\sum_{j=1}^{n}\frac{(s_i-\mu_{gi,j})^2}{2\sigma_{gi,j}^2}\right) \quad (4)$$

ただし $\mu_{gi,j}$ は中心値，$\sigma_{gi,j}$ は標準偏差に相当するパラメータで，これらは前述のように設計時にあらかじめ与えられて固定され，学習しない．全ての実験において一律 $\sigma_{gi,j}=0.2$ とした．この関数 $f_{gi}(s)$ を用いて 288 次元の特徴ベクトル $\mathbf{x}$ の i 番目の要素は，以下のように正

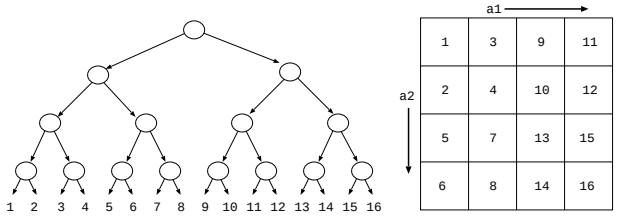

図 3 2 分木表現における行動空間のラベル付け.

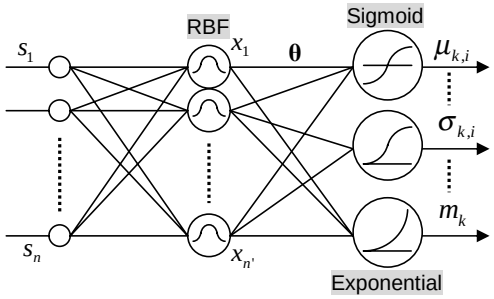

図 4 複数正規分布を用いた階層的政策表現と計算方法.

規化される：

$$x_i = \frac{f_{gi}(\mathbf{s})}{\sum_{j=1}^{n'} f_{gj}(\mathbf{s})} \tag{5}$$

§2 確率的政策の計算方法

確率的政策として正規分布を用いる方法では，行動ベクトル $\mathbf{a}$ は，その i 番目の要素に関するパラメータ μ_i と σ_i を用いて計算される．μ_i と σ_i は状態変数 $\mathbf{s}$ と政策パラメータ θ の関数として以下のシグモイド関数を用いた式で表す：

$$\mu_i = \frac{2}{1+\exp\left(-\sum_{j=1}^{n'}\theta_{\mu i,j}x_j\right)} - 1 \tag{6}$$

$$\sigma_i = \frac{1}{1+\exp\left(-\sum_{j=1}^{n'}\theta_{\sigma i,j}x_j\right)} \tag{7}$$

ただし x_j は式 (5) で示す特徴ベクトル $\mathbf{x}$ の j 番目の要素，$\theta_{\mu i,j}$ および $\theta_{\sigma i,j}$ は政策パラメータである．

また，確率的 2 分木を用いた手法でも，同様に特徴ベクトル $\mathbf{x}$ を用いて政策を表現する．行動空間は 4×4, 16×16 および 64×64 に区切った場合について実装を行った．離散表現と連続な行動空間との対応付けの例を図 3 に示す．実装方法の詳細は文献 [木村 2001] を参照.

提案手法の階層的な政策表現では，式 (1), (2) で示される k 番目の正規分布 g_k に対する重み m_k，中心値 $\mu_{k,i}$，標準偏差 $\sigma_{k,i}$ はそれぞれ図 4 に示すように特徴ベクトル $\mathbf{x}$ と政策パラメータ θ を用いて以下のシグモイド関数および指数関数で表される：

$$\mu_{k,i} = \frac{2}{1+\exp\left(-\sum_{j\in x}\theta_{\mu k,i,j}x_j\right)} - 1$$

$$\sigma_{k,i} = \frac{1}{1+\exp\left(-\sum_{j\in x}\theta_{\sigma k,i,j}x_j\right)}$$

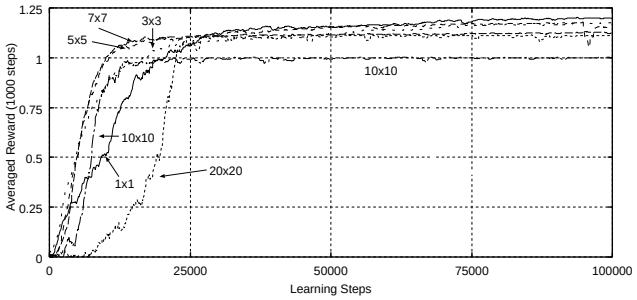

図 5 提案手法の政策表現をシミュレータへ適用した場合の学習曲線.

$$m_k = \exp\left(-\sum_{j\in x}\theta_{mk,i,j}x_j\right)$$

§3 Critic における状態評価値の関数近似方法

本論文で用いる手法の critic は全て式 (5) の特徴ベクトルを用いて以下の線形アーキテクチャによって状態評価値を計算する：

$$\hat{V}(\mathbf{s}) = \sum_{j=1}^{n'} x_j\omega_j \tag{8}$$

ただし ω_j は学習すべき重みパラメータで，TD(0) 法によって更新される．線形アーキテクチャを用いた TD 法の詳しい更新方法は文献 [Sutton98] を参照.

3·4 シミュレーションにおける実験結果

図 5 は提案手法の政策表現を用いた actor-critic によるシミュレータ上での学習曲線を示す．本手法では，下位層として行動空間に複数の正規分布を配置できるため，初期配置として 2 次元の行動空間へ $N=1\times1$, 3×3, 5×5, 7×7, 10×10, 20×20 の 6 通りで格子状に配置してそれぞれ実験を行った．$N=1\times1$, 3×3, 5×5, 7×7 については 5 試行の平均をとり，それ以外は都合により 1 試行のみのデータである．$N=1$ については，政策表現に正規分布のみを用いる従来手法と等価である．下位層の正規分布数 N を 1 より増やしていくことで学習が加速されていくことが分かる．$N=25$ や 49 のとき最大で，$N=1$ の場合に比べると単位ステップあたりの報酬が 1 に到達するまでの時間が約半分まで短縮できた．だが N をさらに増やして $N=100$ や 400 では学習速度が遅くなる．これは上位層での離散的行動数が多くなりすぎて階層化の効果が薄れるためと考えられる．学習パラメータの設定は，これまでと同じで割引率 0.95，actor の学習率 $\alpha_\pi=0.5$，actor の適正度の履歴の割引率 $\lambda_\pi=1$，critic の学習率 $\alpha_v=0.1$，critic の適正度の履歴の割引率 $\lambda_v=1$ と設定した.

図 6 は本手法について特徴ベクトル $\mathbf{x}$ を

$$x_i = \begin{cases} 1 & i=27 \\ 0 & \text{otherwise} \end{cases}$$

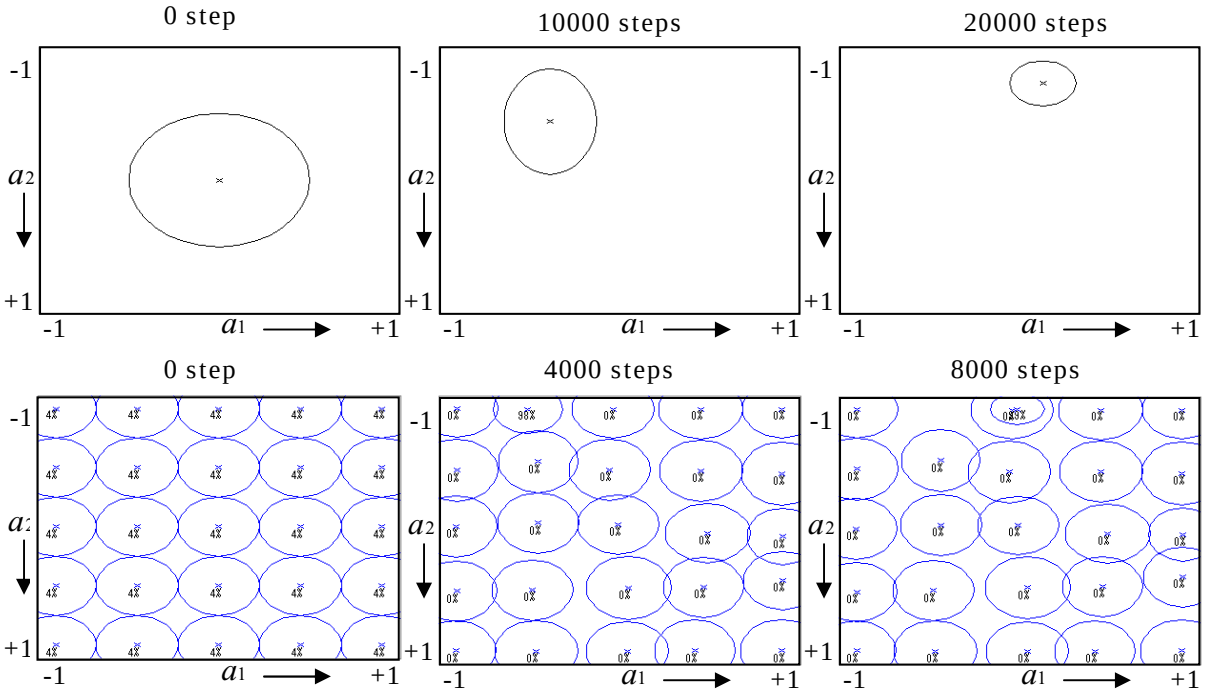

図 6 学習途中における行動空間の正規分布の様子．上 $N = 1 \times 1$，下 $N = 5 \times 5$．

上記に設定したとき，行動空間に配置した正規分布の数 $N = 1 \times 1$ と $N = 5 \times 5$ の各学習ステップにおける正規分布の変化の一例である．図中×印は正規分布の中心値 μ_k を表し，円は標準偏差 σ_k を表す．また下の図で示される×印の下の数字 (%) はその正規分布が上位層で選択される確率 $p(k'|\mathbf{x}, \theta)$ を表す．この図が示すように，たった 1 つの正規分布を用いて学習する場合は，徐々に正規分布の中心値を動かしていくことだけで学習を行っているが，正規分布が複数の場合は，まず有望と思われる行動領域に配置されている正規分布の重みが大きくなり，そのあとにその中心値を移動させる学習を行っている．

注目すべきは図 6 において 4000 step で選択される確率が 98 %となった正規分布の中心が 8000 step では他の正規分布が配置されていた領域へ移動していることである．観察した全ての試行において同様の現象が見られており，学習の進行に伴って政策が変化すると，最適と思われる行動も変化する問題の構造になっていることが分かる．

図 7 は従来の 2 分木構造の政策表現を用いた actor-critic によるシミュレータ上での学習曲線を示す．2 次元の行動空間を 4×4, 16×16, 64×64 の等分割タイル状に離散化したものをそれぞれ 1 試行行った．グラフが示すとおり，この学習方法でも，約 $25,000$ step で収束に向かうが，さらに学習が進行するとしばしば前に進めなくなる状況が生じ，学習は不安定である．学習パラメータの設定は，割引率 0.95，actor の学習率 $\alpha_\pi = 0.5$，actor の適正度の履歴の割引率 $\lambda_\pi = 1$，critic の学習率 $\alpha_v = 0.1$，critic の適正度の履歴の割引率 $\lambda_v = 1$ と設定しており，従来の正規分布のみを用いた場合と同一である．

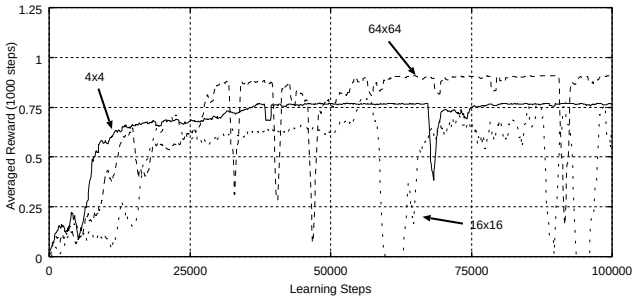

図 7 従来の 2 分木構造の政策表現をシミュレータへ適用した場合の学習曲線．

3·5 実機における実験結果

図 8 は政策表現に正規分布のみを用いる従来手法と提案手法を実ロボット上で学習させて比較したものである．提案法は従来法に比べてすみやかに学習しており，単位ステップあたりの報酬が 0.5 を超えるまでの時間が約

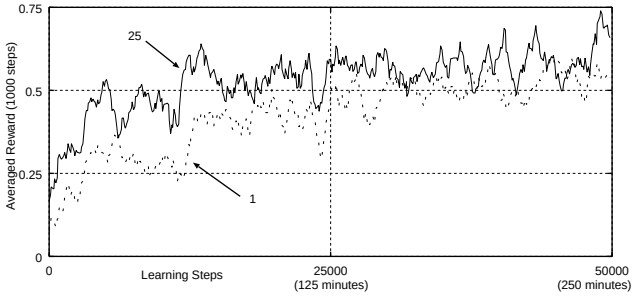

図 8 従来法および提案手法の政策表現を実機に適用した場合の学習曲線．

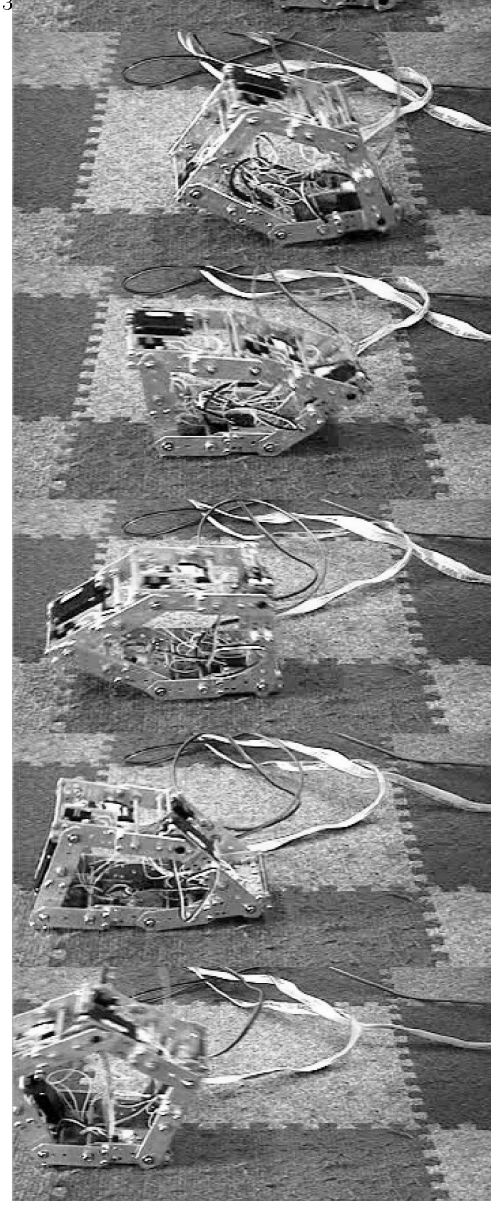

図 9　ロボットの獲得した動作例.

25,000 step で，約 2 時間である．学習パラメータの設定はシミュレータと同一とした．図 9 は本手法で獲得した動作の一例を示すが，どちらの手法でも得られた動きは同等であった．図 9 の上のグラフは本手法で獲得した動作による 100 step 分のモータ角度の遷移の一例である．状態観測のノイズの影響でしばしば遷移パターンが乱れるが，ほぼ同じような場所を遷移する傾向が見られる．また図 9 の下のグラフは上のグラフからある 10 step 分を抜き出したものである．およそ 10 step でグラフに対し時計回りに 2 回転の遷移を行うと，ロボットはちょうど 1 回転する．

4. 考　　察

4·1　最適行動が学習とともに変化する現象について

本ロボットの問題は，環境が動的に変化しているわけではない．環境が変化しない限り，最適な「政策」は変化

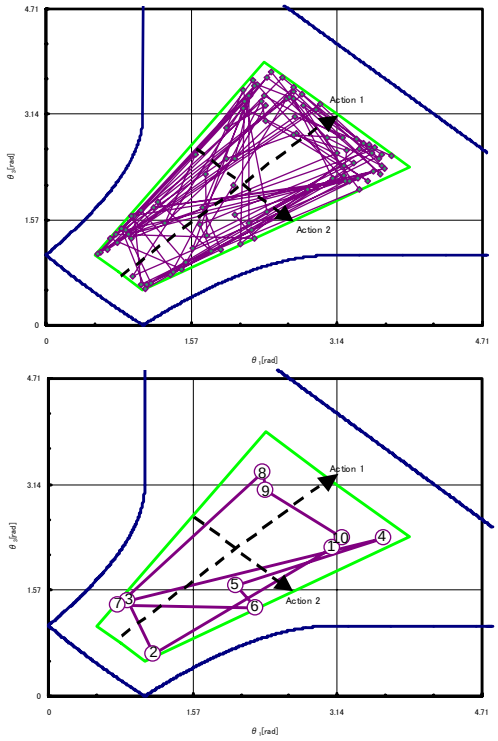

図 10　実ロボット学習後のモータ角度遷移の様子.

しないが，ある状態における最適な「行動」は，エージェントの政策に依存して変化する場合がある．本ロボットの学習初期では，まだ転がることがままならない政策なので，各状態では着実に一歩ずつ転がる行動を目指すが，うまく転がるようになると，なめらかに回転する行動を目指すようになり，これは学習初期に収束しそうに見えた行動領域からずれてくる．図 6 は，実際に変化していく様子を裏付けている．従来の離散行動の学習アルゴリズムだけでは対処が難しく，特に 2 分木を用いた actor-critic では図 7 に示されるようにしばしば学習が不安定になることを示した．

2 分木を用いた方法において学習が進行すると突然進めなくなってしまう状況に陥るのは，次のような理由が考えられる．本ロボットの場合，状態観測に簡易なセンサと関節角度のみを用いており，ロボットの回転速度が観測できないことにより，状態観測に不完全性が生じる．このため，エージェントの状態空間中で同じ状態だと知覚する場所が，学習の進行による政策の変化で実際の位置は変化する．すると各状態においてエージェントが最適とする行動が見かけ上変化してしまうという現象が起きる．最適な行動が変化していく環境において，離散的な 2 分木構造の政策表現でその変化に対応するとき，特に上位層にて分割された領域をまたぐように最適行動が変化していくのに追従していくには，上位層から学習し直す必要があり，このときエージェントはその状態に関して一旦ランダム政策まで戻ってから再学習せざるをえない．よって図 7 のようにしばしば性能が学習初期に戻るという現象が生じたと考えられる．

4·2 階層的政策により学習が加速

上位層で離散的かつ大域的な行動選択，下位層で連続的かつ局所的な行動選択を行う政策表現を用いた actor-critic 法により，単に正規分布だけを用いる学習法より学習すべきパラメータ数が大きいにもかかわらず学習速度を加速する効果があることが実験により示された．また，本手法は最適と思われる行動が学習の進行とともに変化していく場合にも追従できた．

本実験で用いた actor-critic アルゴリズムは，毎ステップにおいて政策パラメータの更新が行われ，その計算コストは政策パラメータ数に比例する．提案手法は正規分布を１つだけ用いる方法に比べるとおよそ正規分布の個数× 1.5 倍の政策パラメータ数になるので，毎ステップにおける政策パラメータ更新の計算コストもそれだけ余計にかかる．しかし，この更新はローカルに並列処理可能であり，またロボットの意思決定時間間隔の時間スケールと対比すると十分な計算時間が与えられているので問題にはならない．

4·3 RBF による特徴ベクトル生成について

本研究のロボットの制御では，状態観測を RBF へ入力して特徴ベクトルを得る方法がそれを用いない手法よりも安定的でほぼ確実に制御規則を獲得できた．よって，RBF による局所的な状態表現を用いると十分な状態表現ができることが分かる．しかし，本論文では RBF による状態の特徴ベクトル獲得は学習の対象外であり，予め与えている．この状態表現獲得は今後の課題である．

4·4 上位層の離散的行動選択方法について

本論文で示した提案手法では，単一正規分布による政策表現と比べた場合の優位性は上位層の設計に左右される．この上位層のデザインは今のところ設計者に依存している．行動の空間がもっと高次元になると離散的行動選択方法にも工夫が必要になってくる．この部分に２分木構造の行動選択方法を用いれば，設計者に依存しにくく，かつ効率の良い学習が期待できる．

5. お わ り に

本論文では，実ロボットにおける強化学習による実時間学習を目標として５リンク環状ロボットの移動動作獲得問題を取り上げた．本問題は状態観測にノイズが存在し，最適な行動が学習の進行に伴って変化していく環境であり，actor-critic における従来の政策表現法では扱いが困難であることを実験的に示した．本論文では上位が離散的，下位が連続的行動選択を行う階層的な政策表現方法を提案し，シミュレーションおよび実機による実験を通じて有効性を示した．今後の課題として，本手法をもっと高次元の自由度を持つ全く別の形態を持つロボットへ適用し，学習システムの汎用性と拡張性を示すことを目指して，ヘビ型移動ロボットで実験中である．

最後に，５リンク環状ロボットに関して貴重な助言をお寄せいただきました本学総合理工学研究科小俣透助教授に謝意を表します．

♢ 参 考 文 献 ♢

[深尾 98] 深尾 隆則，稲山 典克，足立 紀彦：正則化理論を用いた連続的状態と行動を扱う強化学習，システム制御情報学会論文誌，Vol.11, No.11, pp.593–599 (1998).

[堀内 98] 堀内 匡，藤野 昭典，片井 修，椹木 哲夫：連続値入出力を扱うファジィ内挿型 Q-learning の提案，計測自動制御学会論文集, Vol.35, No.2, pp.271–279 (1999).

[石井 2000] 石井 信，佐藤 雅昭：統計的手法にもとづく強化学習と制御ルールの獲得，計測と制御，Vol.39, No.12, pp.763–768 (2000).

[木村 99] 木村 元，宮崎 和光，小林 重信：強化学習システムの設計指針，計測と制御, Vol.38, No.10, pp.618–623 (1999).

[木村 2000] 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，人工知能学会論文集, Vol.15, No.2, pp.267–275 (2000).

[木村 2001] 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，計測自動制御学論文集, Vol.37, no.12, pp.1147–1155 (2001).

[木村 2002] 木村 元，山下 透，小林 重信：強化学習による４足ロボットの歩行動作獲得，電気学会 電子情報システム部門誌 Vol.122-C, No.3, pp.330–337 (2002).

[Lin96] Lin, C. J. and 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).

[Lin99] Lin, C. T. and Chung, I. F.: A Reinforcement Neuro-Fuzzy Combiner for Multiobjective Control, IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetcs, Vol.29, No.6, pp.726–744, 1999.

[森本 2001] 森本 淳，銅谷 賢治：階層型強化学習を用いた３リンク２関節ロボットによる起立運動の獲得，日本ロボット学会誌 Vol.19, No.5, pp.574–579, 2001.

[鮫島 2001] 鮫島 和行，銅谷 賢治，川人 光男： 強化学習 MOSAIC: 予測性によるシンボル化と見まね学習，日本ロボット学会誌 Vol.19, No.5, pp.551–556, 2001.

[Santamaria98] Santamaria, J.C., Sutton, R.S. & Ram, A.: Experiments with Reinforcement Learning in Problems with Continuous State and Action Spaces, Adaptive Behavior 6 (2), pp.163–218 (1998).

[Sutton98] Sutton, R. S. & Barto, A.: Reinforcement Learning: An Introduction, A Bradford Book, The MIT Press (1998).

[Sutton2000] Sutton, R. S., McAllester, D., Singh, S. & Mansour, Y.: Policy Gradient Methods for Reinforcement Learning with Function Approximation, Advances in Neural Information Processing Systems 12 (NIPS12), pp. 1057–1063 (2000).

[Yoshimoto2000] Yoshimoto, J., Ishii, S. and Sato, M.: Online EM reinforcement learning, IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN 2000), III, pp. 163-168, (2000).

[Watkins92] Watkins, C. J. C. H. & Dayan, P.: Technical Note: Q-Learning, Machine Learning 8, pp.279–292 (1992).

[Williams92] Williams, R. J.: Simple Statistical Gradient Following Algorithms for Connectionist Reinforcement Learning, Machine Learning 8, pp. 229–256 (1992).

〔担当委員：山川 宏〕

2003 年 4 月 10 日　受理

◇ 付　　録 ◇

•1　関節角度の制約条件と可動領域の導出

図 1 に示す環状のリンクの角度 $\theta_1,\theta_2,\ldots\theta_5$ は，物理的制約より次の条件式を満たす：

$$0 \leq \theta_1,\theta_2,\theta_3,\theta_4,\theta_5 \leq 2\pi \tag{A.1}$$

$$\theta_1+\theta_2+\theta_3+\theta_4+\theta_5 \;=\; 3\pi \tag{A.2}$$

$$\sum_{i=1}^{5}\exp\left(j\sum_{k=1}^{i}(\pi-\theta_k)\right) \;=\; 0 \tag{A.3}$$

ただし j は虚数単位である．式 (A.1) は各関節ヒンジの可動領域による制限，式 (A.2) は 5 角形の内角の和の条件，式 (A.3) はリンク 1 の根元の座標がリンク 5 の先端座標と一致する条件式である．これより 2 つのモータ関節角度 θ_1,θ_3 が既知のとき，他の角 $\theta_2,\theta_4,\theta_5$ を次の式で求めることができる：

$$\theta_4 \;=\; -\theta_3+\arctan\frac{1}{\tan\theta_3}+\arccos\frac{\frac{1}{2}+\cos\theta_1-\cos\theta_3}{\sqrt{2(1-\cos\theta_3)}}$$

$$\theta_2 \;=\; -\arctan\frac{1}{\tan\theta_1}-\arctan\frac{\sin\theta_3-\sin(\theta_3+\theta_4)}{\cos\theta_3-\cos(\theta_3+\theta_4)-1)}$$

$$\theta_5 \;=\; 3\pi-\theta 1-\theta_2-\theta_3-\theta_4$$

ここで arccos の定義域は $[-1,1]$ なので，

$$-1 \leq \frac{\frac{1}{2}+\cos\theta_1-\cos\theta_3}{\sqrt{2(1-\cos\theta_3)}} \leq 1 \tag{A.4}$$

これらの条件を考慮すると，モータの角度 θ_1,θ_3 の可動領域は図 2 のようになる．

著　者　紹　介

木村　　元（正会員）

1997 年東京工業大学大学院知能科学専攻博士課程修了，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に従事．

荒牧　岳志

2002 年東京工業大学工学部情報工学科卒，同年 4 月（株）スクウェアに所属．

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．