[計測自動制御学会論文集
Vol.42, No.3, 255/264 (2006)]

# 汎化行為概念の適応的獲得
## —双シェマモデルベースの強化学習—

谷口忠大*・椹木哲夫*

# Incremental Acquisition of Generalized Behavioral Concepts
## —Dual-Schemata Model Based Reinforcement Learning—

Tadahiro TANIGUCHI* and Tetsuo SAWARAGI*

In this paper, we introduce a reinforcement learning architecture method for autonomous robots to obtain generalized behavioral concepts. Reinforcement learning is a well formulated method for autonomous robots to obtain a new behavioral concept by themselves. However, these behavioral concepts cannot be applied to other environments that are different from the place where the robots have learned the concepts. On the contrary, we, human beings, can apply our behavioral concepts to some different environments, objects, and/or situations. We think this ability owes to some memory structure like Schema System that was originally proposed by J.Piaget. We previously proposed a modular-learning method called Dual-Schemata model. In this paper we add a reinforcement learning mechanism to this model. By being provided with this structure, autonomous robots become able to obtain new generalized behavioral concepts by themselves.

**Key Words:** schema, reinforcement learning, modular learning, hierarchical learning, generalized behavioral concept

## 1. はじめに

現在，機械システムの設計論は望みの状態を静的に維持することを目的にする安定性を志向した思想から，自律的に動的なパターンを生成することにより所望の振る舞いを生成したり，自律的に環境に適応することにより新たな機能を形成していくという動的な機能構成を目指した思想へと展開しつつある．そこでは人間に隷属的な道具としての機械の設計という捉え方から，自律適応系としての機械の設計へという発想の転換が本質的である．

このような自律適応系は，動的に変化する環境下でさまざまなタスクをただ一つの身体を通して為していくということを求められる．しかし，そのような動的環境の性質を前もって設計者がすべて想定することは不可能であり，ゆえに，行なうべきタスクもアプリオリには決定できない．このようなニーズを満たすためには，変化する環境や求められる振る舞いに対して適応的に機能生成を行なっていくような発達学習機構が必要である．

このような適応性，発達性を実現する枠組みとして強化学習[21]や模倣学習[22]などが注目されている．そのような中，われわれは自律ロボットが環境との相互作用を続ける中で，環境との相互作用の構造的変化に自律的に気づき行動学習および概念形成を行なうためのモデルとして双シェマモデル（Dual-Schemata model）を提唱してきた[1)~5)]．しかし，過去の研究では行為シェマ（intentional schema）として表象される行為概念は前もって設計者によって与えなければならないという制約があった．これに対し本論文では，双シェマモデルにおける行為シェマに対し強化学習手法を通して行為概念を学習させる方法論を提案する．ここで獲得される行為シェマは通常の強化学習で獲得される方策とは異なり，環境ダイナミクスの変化に対する普遍性をもっている汎化行為概念となっているのが特徴である．本稿では行為概念，および汎化行為概念について議論した後に，汎化行為概念を獲得するための学習機構を提案する．

## 2. 汎化行為概念

本章では本論文の主題でもある，汎化行為概念について議論を行なう．そして，センサ空間のアトラクタとしての汎化行為概念の表現を提案する．

### 2.1 行為とは何か？

一般にロボットの学習理論などで行為という場合，それはモータ出力 $a_t$ を指す場合が多い．しかし，グリッド空間で迷路抜けタスクを行なうようなエージェントではなく，実空

* 京都大学工学研究科機械理工学専攻　京都市左京区吉田
* Department of Mechanical Engineering and Science, Kyoto University, Sakyo-ku, Kyoto
(Received May 25, 2005)
(Revised September 16, 2005)

間で活動するエージェントにとって瞬間的なモータ出力が何らかの意味をもつことは少なく，対象に対して働きかける一連の動作一纏まりで初めて行為として認識される．その点から見れば，ある目的をもった方策関数 $a_t = \pi(s_t)$ のような，センサ入力に対し時々刻々と出力を変化させていく関数関係によって示される制御則のことを行為と呼ぶ方が自然であると考える．また，その制御則によって表出される振る舞いをその制御則が表象する行為概念と呼ぶことにする．たとえば，ある方策関数（制御則）を用いたときに，観察者から見てロボットが歩行したとするならば，その行為概念は「歩行」であると名づけて構わない．

### 2.2 汎化行為概念

ところで，われわれは歩く，投げる，叩くなどのさまざまな行動を表象する言葉，またはそれに相当する行為概念をもっている．しかし，たとえば「投げる」という概念を例にとってみても，投げる距離や，投げるもの，風の具合などによって，実際に筋肉に伝えられる筋指令はまるで変わってくる．つまり用いるべき方策 $a_t = \pi(s_t)$ 自体が変化する．ロボットの言葉に置き換えてみると筋指令はモータ出力であり，姿勢角やボールの軌道がセンサ入力となる．そして，投げるものや風の具合といったものは環境や対象のダイナミクスとして行為の結果に影響を与える．このように，同一でないモータ出力の時系列にもかかわらずわれわれは「投げる」いう一つの言葉でそれらを表象している．これはわれわれが環境や対象のダイナミクスには依存しない形である種の行為概念をもっていることを示唆している．しかし，一般に強化学習の枠組みなどを通じて獲得される行為は固定された環境ダイナミクス下で獲得され，ほかの環境下で応用することができない (Fig. 1)．このような行為概念は環境の変化に対する汎化性がないといえる．ここでいう汎化性とは一般に学習器の汎化性と呼ばれるものとは異なる．学習器の汎化性とは経験していない状態量について正しい応答を返せるかということである．これに対してわれわれの述べている行為の汎化性とは，同じ状態量に対しても環境・対象次第で異なる応答を返さねばならないという点で，いわゆる学習器の汎化性より上位の概念として捉えられる．獲得した技能を新規環境で利用できるかという問題は novelty problem と呼ばれ人間の運動学習の研究においても古くから研究されている[17]．われわれはこのような環境変化に対する普遍性をもつ行為概念を汎化行為概念と呼ぶことにする．本論文はこの汎化行為概念を自律ロボットに適応的に獲得させることを目的としている．

### 2.3 行為主体にとっての行為

汎化行為概念は行為主体にとってどのようなものであろうか．行為主体はおのずからのセンサを通してしかおのずからの行為の結果は観測できないと考えるのが妥当であろう．つまり，行為主体はおのずからの行為をセンサ空間（状態空間）内の軌跡として捉えざるを得ない．そうしたとき，行為主体の行為概念とはセンサ空間内のアトラクタとして捉えることが出来る．目的志向的な行為は点アトラクタであり，歩行などの周期運動的な行為はリミットサイクルとして表現される[18]．このように考えたときのセンサ空間内でのアトラクタを形成する力学場は内部的なダイナミクスであり環境のダイナミクスには依存しない．よって，このようなアトラクタは汎化行為概念として異なった環境ダイナミクス間で再利用可能となる．われわれはこのような視点から後に述べる双シェマモデル内で行為シェマにセンサ空間内でのアトラクタとしての役割を負わせることにより汎化行為概念の実現を行なった．

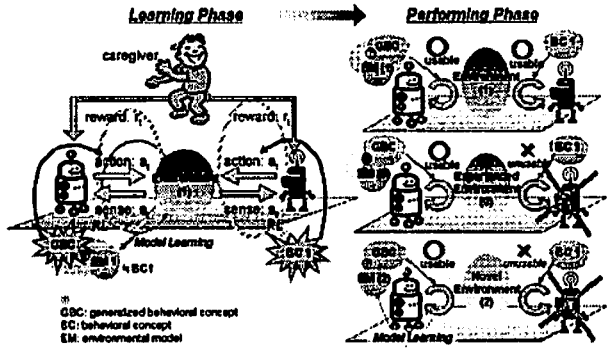


**Fig. 1** an overview of differences between a generalized behavioral concept and a behavioral concept

## 3. 階層的モジュール型学習器

前章で述べたような汎化行為概念を自律ロボットに獲得させるには，通常の強化学習を用いて $a_t = \pi(s_t)$ の形で行為を獲得したのでは不十分である．このような欠点を補うためには，強化学習の研究から一歩踏み出した学習器の構造化についての研究が必要である．現在，モジュール性や階層性を導入した学習器の研究が盛んである．本章ではそれらの研究を紹介し，汎化行為概念の獲得におけるモジュール性，階層性の役割について考える．

### 3.1 モジュール型学習器

近年，ロボット制御の手法や脳のモデルとして，適応的なモジュール型の制御器が注目を集めている[6),7)]．モジュール型の制御器は，将来の予測値や制御入力値を出力するモジュールを分散的に配置し，これらを状況に応じて切り替えることにより離散的に切り替わる環境ダイナミクスや一つの制御器では覆いきれない非線形な系に対応するという特徴をもつ．これに加え，適応的なモジュール型の制御器では個々のモジュール自体が環境との相互作用を通じて学習し変化していく仕組みをもつ．しかし，これらの枠組みではモジュールを関数空間中に設計者が事前に分散的に配置する必須となる．その配置の仕方が獲得される各モジュールの担当するダイナミクスに強い影響を与える．特に MOSAIC[6)]の仕組みでは一度，融合してしまったモジュールは決して二つに分かれることはない．ゆえに初期状態の各モジュールに対して十分な多様性を与えておくことが必要となる．つまり，MOSAIC は初期に与えた多様性を喪失しながら構造化を進める．これ

に対し，われわれは環境がもつダイナミクスの相転移的変化に着目し，その変化を文脈依存的かつ主観的に発見し新たなモジュールを累増的に作成していくモジュール型学習モデルとして双シェマモデル（Dual-Schemata model）を提案してきた[1),3)~5)]．これは，環境のダイナミクスの変化に対応しつつ，おのずからの内部に多様性を創出していくことにより，構造化をすすめる，分化（differentiation）を基本的なダイナミクスとした学習モデルである．これら二つのダイナミクスは対照的に捉えられるが，現段階では双方が長所と短所を持ち合わせており，優劣で考えられるものではなく，双方のダイナミクスの特性についてより深く研究される必要がある．

前者のモジュール型の強化学習機構としては銅谷，片桐らのモジュール型強化学習がある[12)]．しかし，これは報酬とダイナミクスから最適方策を導出する強化学習の中心部分を，リカッチ方程式の代数的解法で求めるために，強化学習機構としては特殊なモデルであり，他手法との比較が難しい．また，一つの線形の予測モデルにつき一つの線形の制御器という考え方は，決して普遍的ではなく適用範囲が制限されてしまう．さらに，線形予測器により分割された状態間での報酬分配が考えられないため，大域的なタスク成功に到るとは限らない．これらを解決する手法は現在，杉本，鮫島らによって提案されている[13)]．後者の分化累増的にモジュールを獲得する強化学習アプローチには高橋らのモジュール型強化学習[14)]がある．

どちらにせよ，モジュールは環境のダイナミクスをその分担により分節化することが出来る．ゆえに，自律ロボットは，環境の変化に対しモジュールを切り替えることで，動的環境下においても適応的に振舞うことが出来る．

### 3.2 階層的強化学習

また，強化学習の研究において，学習を階層化する階層的強化学習の枠組みが議論されている．強化学習を階層化する利点は複数あるが，最も頻繁に強調されるのは上位層に全体のゴールに対してサブゴールを生成させることにより探索を効率化するというものである[8),11)]．しかし，この利点を利用するためには一般的には上位層を下位層より粗くセルで区切り離散化する必要がある．これは階層性そのものがもたらす利点というより，階層的強化学習の上位層と下位層の離散化の粒度の差，つまり分節の仕方の差から生じる利点である．これに対し，われわれは階層性のもつ本質的な利点を自律ロボットの獲得行為概念の汎化性の視点から捉えている．われわれの双シェマモデルでは下位層（知覚シェマ）に環境のダイナミクスを任せることにより，上位層（行為シェマ）においては環境のダイナミクスに依存しない行為表象を獲得することが出来る．しかし，われわれの方法では特に階層ごとの粒度の違いは導入しないために，6章の実験で示すように強化学習の高速化という面においての寄与は期待できない．獲得した行為概念の再利用という面では浅ら[10)]の研究などがあるが，行為概念の再利用において保存される部分の決定があいまいであるといった点や，再利用時に変更部分を忘却してしまうために，再び同じ環境に戻ったときに再学習を要するなどの欠点がある．次章以降でわれわれの提案手法である双シェマモデルについての説明を行なう．

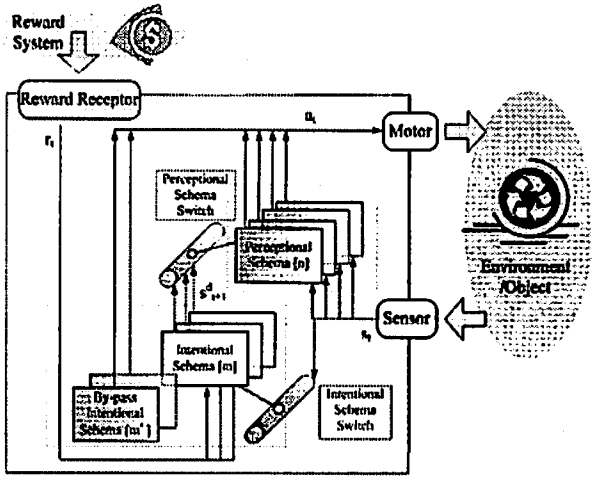


**Fig. 2** Dual-Schemata model

## 4. 双シェマモデル：知覚シェマの適応

本論文ではわれわれが提案してきた双シェマモデルにおいて，その行動学習のために強化学習を用いる手法を提案するが，まず，そのための導入として本章では双シェマモデルの導入を行なう．詳細は文献[1)~3)]を参照されたい．

双シェマモデルはダイナミクスが異なる複数環境下を遷移する自律ロボットがその環境の変化に自律的に気づき，おのずからの内部の記憶構造である知覚シェマを累増的に分化させていく枠組みである．

### 4.1 双シェマ

双シェマとは以下の二種類のモジュールを指す．

$$\text{知覚シェマ } PS_n : s_{t+1} = F_n(s_t^{\dagger}, a_t) \quad (1)$$

$$a_t = I_n(s_t^{\dagger}, s_{t+1}) \quad (2)$$

$$\text{行為シェマ } IS_m : s_{t+1}^d = G_m(s_t^{\dagger}) \quad (3)$$

ここで，$s_t$ はセンサ入力，$a_t$ はモータ出力である．センサ入力以外の短期記憶などの入力を $m_t$ として，$s_t$ との直和をとったものを $s_t^{\dagger}$ とする．知覚シェマは，あるセンサ状態 $s_t$ において，つぎのセンサ状態 $s_{t+1}$ になるためにはどのようなモータ出力 $a_t$ を出せばよいかという関係を表わす関数（逆モデル）をもつ．また，行為シェマは現在の状態に依存しつつ次状態の目標値 $s_{t+1}^d$ を知覚シェマに与えるモジュールである．どの行為シェマが選択されるかによって，どのような振る舞いを自律ロボットが行なおうとするかが決定される．また，実際のモータ出力は $PS_n$,$IS_m$ が選択されているときには下式で決定される．

$$a_t = H_{nm}(s_t) = I_n(s_t^{\dagger}, G_m(s_t^{\dagger})) \quad (4)$$

つまり，行為シェマと知覚シェマを合成 (composite) させることによって一つの行動則が生成される．

以降，本論文では $m_t$ については考慮せず $s_t^! = s_t$ と見なすが，以降の議論は $m_t$ が存在する場合へも自然に拡張可能である．

### 4.2 均衡化・分化

双シェマモデルにおいて重要な知覚シェマの適応プロセスが均衡化（equilibration）と分化（differentiation）である．均衡化は現在の環境と知覚シェマとを対応させる同化（assimilationig3）と，知覚シェマの内部関数を，環境との相互作用を上手く記述できるように更新する調節（accommodation）から成り立っている．これらは情報として $s_t^!, s_{t+1}, a_t$ というセンサ・モータ出力値以外は一切利用しないために，自律ロボットは速い教師あり学習を通して行なうことができる．これは自身以外のいかなる外部他者の存在をも仮定せずに行なうことが出来る適応プロセスである．同化の判断が MOSAIC[6] における責任信号のようにほかのモジュールとの相対的な関係によって決定するのではなく，基本的には主観的誤差を用いた個々のモジュールの自律的な判断に基づいて行なわれるのが双シェマモデルの一つの特徴である．この自律的判断がモジュールの累増性を実現している．その自律的判断のために主観的誤差 $R_n$ およびシェマ活性度 $V_n$ の定義を行なう．まず，以下により誤差 $E_n$ を定義する．（$||*||_{each}$ は各成分について絶対値を取るという意味．）

$$E_n(t) = ||s_{t+1} - F_n(s_t, a_t)||_{each} \quad (5)$$

これは予測誤差と呼ばれる．文献[1]でわれわれは誤差として逆モデル誤差を用いた．しかし，双シェマモデルの特殊化モデルである，LDS（Light Dual-Schemata Model）においては，内部関数を線形関数に制限することにより，順モデル $F_n$ を求めてから逆モデル $I_n$ を求めるために，誤差計測には予測誤差を利用する[2),3)]．$F_n$ は事前に設計せずに同化した経験のみについて確率的勾配法を用いて学習により獲得する（詳しくは文献[1-3]）．また，$I_n$ は $F_n$ から近似的に求めることができる．後に行なう実験ではこの LDS を用いるので以降その説明を行なう．つぎに，誤差予測器 $\hat{E}_n$ を定義する．$\hat{E}_n$ は最も簡単には，過去にそのシェマを用いて得られた誤差の重み付き平均として得られるが，LDS においては以下のように移動量 $ds_t = ||s_{t+1} - s_t||_{each}$ の線形関数として定義する．

$$\hat{E}_n(t) = \hat{E}(ds_t) \quad (6)$$

これも内部関数と同時に学習を続けさせる．これらを用いて主観的誤差 $R_n$ は以下のように定義される．

$$R_n(t) = ||diag(\hat{E}_n(t))^{-1} * E_n(t)||_\infty \quad (7)$$

$diag(*)$ は $*$ を対角成分とした対角行列である．もし $R_n(t) \leq k$ ならば知覚シェマ選択器により選択されていた知覚シェマ $PS_n$ は時刻 t で経験をおのずからに同化する．同化されない経験は廃棄される．ここで $k$ は同化のための閾値であり，同化係数（assimilation threshold）と呼ぶ．上記の同化を繰り返す中で $PS_n$ の内部で定義された順モデルはおのずからが同化した新たな経験に合うように更新し続ける．知覚シェマ活性度 $V_n(t)$ は自律ロボットが現在対面している環境のダイナミクスに対して各知覚シェマがおのずからとそれとがどれだけコヒーレントかを示す変数である．$V_n(t)$ は主観的誤差に従って毎時以下の式によって更新される．

$$V_n(t+1) = p * V_n(t) + (1-p) * exp(-\frac{1}{2}R_n(t)^2) \quad (8)$$

p は定数でどれだけ過去の活性度に固執するか（persistency）を表わす．通常 $V_n = 0$ は「現在の環境は $PS_n$ に対応していない」ことを示す．また，$V_n = 1$ は「現在の環境が $PS_n$ に対応している」ことを示す．知覚シェマ選択器はこの知覚シェマ活性度を参照しながら以下の法則に従い知覚シェマを選択する．

（1） もしすべての知覚シェマの活性度があらかじめ定めた閾値 $V_{turn}$ を上回らなければ，知覚シェマ選択器は新たな知覚シェマを作り出す．新たな知覚シェマは現在の総数が $N$ 個の場合 $PS_{N+1}$ と名づけられる．

（2） 知覚シェマ選択器は知覚シェマ活性度が $V_{turn}$ 以上のものの中から，最も古いものを優先して選択する．

このようなルールに従うことにより知覚シェマは文脈に寄り添いながら，選択，作成されることが出来る．$V_{turn}$ は $V$ についての閾値である．本稿では $V_{turn} \equiv exp(-\frac{1}{2}k^2)$ に設定する．ここで固執率 $p$ と同化係数 $k$ は共にシェマの同化，分化を支配する重要なパラメータである．一般に k が大きいほどシェマは安定化するが分化は起こさないために環境変化の識別能力は低下する．逆に小さすぎるとシェマは E の小さい経験以外をほとんど同化しなくなるために $\hat{E}$ は 0 に漸近していきシェマが消れる（i.e. 何も同化せず常に活性度が 0 になる）という現象が起きる．予測誤差が正規分布すると仮定した場合，$2 < k < 5$ 程度の値が適当である．また，$p = 0$ のときにはその瞬間の経験のみでシェマ活性度を決定するために環境の変化には鋭敏であるが，耐ノイズ性は低くなる．逆に $p$ が 1 に近づくと耐ノイズ性は高まるが，大きすぎると環境が変化してからシェマ活性度が十分に変化するまでに時間がかかりすぎるために環境の変化に鈍感になり実時間での適応性が損なわれる．これらの設計論はシェマの適応ダイナミクスを考える上で興味深いが，紙面の都合上後の稿に譲る．

## 5. 行為シェマの強化学習による獲得

本章では双シェマもデルにおいて二つ目のモジュール群である行為シェマの適応ダイナミクスを強化学習の手法を用いて導入する．

### 5.1 センサ空間のアトラクタとしての行為

双シェマモデルの枠組みにおいて，行為シェマは $(s_t, s_{t+1})$ の二項関係を規定する．これにより，順次決定されるセンサ値が自律主体にとっての行為となる．これは自律主体にとって一纏まりのチャンク化された行為とは，受容される一連のセンサ値の系列で決められる状態空間内のアトラクタとして描かれることを前提とした定義である．行為シェマは自律ロ

ボットが他者の助けなしに得ることの出来るセンサ空間の中のベクトル場として張られ，実際のモータ出力 $a_t$ とはまったく無縁に定義される．自律主体の行為のおのずからにとっての意味とは，おのずからのセンサ入力のみから認識されるため，モータ出力 $a_t$ がどのように出力されようとも，その影響がおのずからのセンサ越しに現れない場合，有意義な行為としては理解されえない．つまり，行為シェマ $IS_m$ は自律主体にとって理解されうる一纏まりの行為を表象している．行為シェマは知覚シェマと同様に並列に複数存在し，行為シェマ選択器によって選択されるが，本章では行為シェマに外部教示者による報酬をあてがうことにより，外部教示者の意図する所望の振る舞いを強化学習により学習させる方法を提案する．説明は簡単のために離散時間で記す．

### 5.2 行為シェマの強化学習

行為シェマの適応プロセスを実装するにあたり，強化学習の理論としては，銅谷らが提案した連続時空間上でのActor-Critic法を用いる[15]．これはActor-Critic法を連続空間に拡張したもので，価値関数（Critic）$V$ と方策関数（Actor）$G$ を基底関数の線形和として表現することにより，連続空間上へ拡張したものである．

$$V(s_t) = \sum_i \omega_i \mathrm{b}_i(s_t) \tag{9}$$

$$G(s_t) = \sum_j \theta_j \mathrm{a}_j(s_t) \tag{10}$$

また，方策関数における連続な粒度での探索を表現するためにGullapalliの確率的強化学習の手法[20]を用いることにより，行動空間，状態空間共にセル状に離散化する必要をなくしている．さらに，連続時間上でメタパラメータを設定することによりロボットの行動周期の時間巾の設定とは独立にすべてのメタパラメータを調節できるようにしている．ただし，一定の行動周期に固定する場合においては，離散時間における立式と違いはない．これらの定式化は連続な実空間，および状態行動空間の適切な分節化が出来ないような環境で活動するロボットの強化学習を行なう場合に非常に有用である．また，Actorの学習においてprofit sharingを積極的に行なうために，木村らのActorにおける適格度トレースの枠組み[16]を導入した．これらの方法を用いた強化学習では通常，方策の出力としてモータ出力 $a_t$ を割り当てるが，行為シェマの出力であるセンサ入力の目標値 $s_t^d$ を割り当てることによりこれらの強化学習手法を行為シェマの適応ダイナミクスとして利用した．しかし，行為シェマにおける学習の場合，最終的に現れるセンサの次入力としての $s_{t+1}$ は知覚シェマのモデル化における精度が悪い場合は，知覚シェマに出力される $s_{t+1}^d$ とは異なる．よって，われわれは次入力として得られる $s_{t+1}$ を実際に行為シェマがとった行為と見なし，actorの学習に利用した．また，局所的にしか与えられない報酬に対しても対処できるようにするために，適格度トレース（eligibility trace）をcriticだけでなく，actorにも導入した．具体的な更新式は以下のようになる．

#### 5.2.1 価値関数の更新

Actor-Critic法における価値関数とは時刻 $t$ での状態 $s_t$ を始点にして現在の方策 $\pi$ に従った場合に将来にわたって獲得することの出来る報酬の重みつき期待値として定義される．

$$V_t^\pi = E^\pi[\Sigma_{k=0}^{\infty} \gamma^k r_{t+k}] \tag{11}$$

上式において，通常の強化学習では $r_t$ の条件として決定論的な場合には $r_t = r(s_t, a_t)$ であること，非決定論的な場合には $s_t, a_t$ を変数にもつ確率密度関数から生成されることを求めるが，時刻 $t$ における行為を次状態そのもの（$s_{t+1}$）とするので，$r_t = r(s_t, s_{t+1})$ で定められるものとする．

推定された価値関数は将来に得られる報酬の重み付き期待値という役割と同時に報酬の予測モデルとしての役割も果たす．その報酬についての予測誤差はTD誤差 $\delta_t$ と呼ばれて以下で表わされる．

$$\delta_t = r_t - \hat{r}_t = r_t - (V(s_t) - \gamma V(s_{t+1})) \tag{12}$$

基本的には学習はこのTD誤差の二乗誤差を最小化するように慣性項つきの勾配法を用いて行なうが，報酬の分配（profit sharing）を考えるために以下で更新される適格度トレース（eligibility trace）$\mathrm{e}^{\omega_i}$ を導入する．

$$\mathrm{e}_t^{\omega_i} = \gamma\lambda \mathrm{e}_{t-1}^{\omega_i} + \frac{\partial V}{\partial \omega_i} \tag{13}$$

ここで，$\lambda$ は適格度トレースの減衰率パラメータである．この $\mathrm{e}^{\omega_i}$ を用いて毎時以下のように基底関数の係数を更新する．

$$\omega_i \leftarrow \omega_i + \alpha\delta_t \mathrm{e}_t^{\omega_i} \tag{14}$$

#### 5.2.2 方策関数の更新

Actor-Critic法においては探索のための項を方策内に陽に設計する必要がある．よって，実際の方策は標準偏差1のノイズ $n_t$ を用いて以下のように書ける．

$$s_{t+1}^d = \mathrm{s}(\mathrm{s}(G(s_t)) + \sigma_t n_t) \tag{15}$$

また，実際の状態空間は各次元についての上界，下界をもつ場合が多い．そのために飽和量を超えたものを定義域内に納めるための関数 s を用いている．探索率 $\sigma_t$ については Willson，木村らのように勾配法を用いて制御する方法[16]，森本，銅谷らのように価値関数で制御する方法[11),15]，または時間の進行に対して単調に減少させていくフィードフォワード的な制御などが考えられるが，本研究では価値関数による制御を採用する．このどれを用いるかは本論文の主旨ではないので議論を避ける．

Actorの学習は基本的にはノイズによる探索を通して実際に起こした行動が期待報酬よりも多くの報酬を得た場合にはその行動を強化するという仕組みで行なわれる．具体的な更新式は価値関数と同様に適格度トレースを用いて以下のように表わされる．

$$\mathrm{e}_t^{\theta_i} = \gamma\lambda \mathrm{e}_{t-1}^{\theta_i} + \frac{\partial G}{\partial \theta_i}(s_{t+1} - G(s_t)) \tag{16}$$

$$\theta_i \leftarrow \theta_i + \alpha\delta_t \mathrm{e}_t^{\theta_i} \tag{17}$$

以上は係数ベクトル $\omega, \theta$ の更新に本質的には勾配法を用い

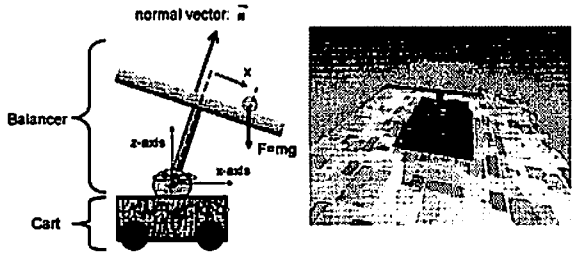


Fig. 3 an overview of the agents used in the experiments

たものであるが，学習のなだらかさを確保するために実際には慣性項を用いた更新を行なう．

## 6. 中心への位置制御の獲得

本章では前述のモデルを用いて行なった実験について記す．

### 6.1 実験環境

本研究では振動を起こさずに加速や移動を行なう台車に乗り，その上で転がるボールを平板の角度を変化させることにより制御するというタスクを行なうエージェントを考える．論点を明確にするために，課題自体は非常にシンプルなものを用いている．具体的には Fig. 3 に示すような系を考える．ここで，平板を制御するエージェント（Balancer）と台車を制御するエージェント（Cart）は別であり，お互いの制御指令などは観測することが出来ないとする．Fig. 3 の Balancer 部と Cart 部に「目」の記号をつけているのはこの重要点を強調するためである．つまり，Balancer は制御しようとする球にかかる慣性力という形で Cart の行動の影響を受ける．それにもかかわらず，Cart の行動を観測できない環境で制御を行なわなければならない．この系の y 方向についての物理量を Fig. 3 に示す．x 方向についてもほぼ同様である．Fig. 3 における平板の規格化された法線ベクトルを $\vec{n} = (n_x, n_y, n_z)$ とする．Balancer が動かしうる法線ベクトルの範囲には $-\frac{1}{2} \leq n_x, n_y \leq \frac{1}{2}$ の制限があるものとする．Balancer は二次元の行動出力 $a_t$ を出す．$a_t = (a_t^x, a_t^y)$ により法線ベクトルは以下の式で直接的に制御される．

$$(n_x, n_y, n_z) = \frac{1}{2}(a_t^x, a_t^y, \sqrt{1-(a_t^x)^2-(a_t^y)^2}) \quad (18)$$

平板は一辺 2[m] の正方形で，その中心を原点とし x, y 座標を設定する．このとき，平板の上のボールの運動は近似的に以下のように表わされる．

$$(\ddot{x}, \ddot{y}) = (gn_x - \alpha_x, gn_y - \alpha_y) \quad (19)$$

$\alpha = (\alpha_x, \alpha_y)$ は Cart の運動や姿勢によって現れる外因的な項である．Cart の運動については詳述しないが，基本的には y 方向に加減速と左右方向への操舵が可能である．これに伴いボールには慣性力，遠心力がかかる．また，路面の傾きによっても $\alpha$ は変化する．ボールの位置と速度に微小なセンサノイズを乗せたものを Balancer のセンサ入力 $s_t$ とし，Balancer は毎ステップこれを獲得することが出来る．

$$s_t = (x, \dot{x}, y, \dot{y}) + \epsilon_t \quad (20)$$

Table 1 Intentional Schemata

| Index | Name | Policy |
|---|---|---|
| $IS_0^{\times}$ | move random | $a_t = G_0^{\times} \equiv n_t$ ( $n_t$ is a noise term) |
| $IS_1$ | shake a ball | $s_{t+1}^d = G_2(s_t) \equiv n_t$ |
| $IS_2$ | (for reinforcement learning) | $s_{t+1}^d = G_1(s_t)$ |

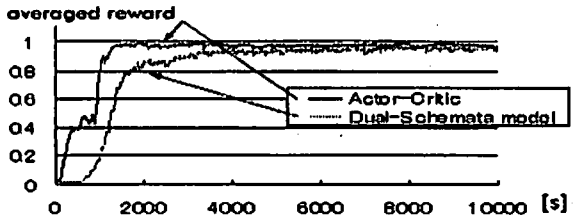


Fig. 4 transition of the weighted averaged rewards in simple reinforcement learning task

ボールが平板から落ちた場合には，つぎのステップで平板の上のランダムな位置にボールを置きなおす．このとき，特にペナルティは与えない．今回のタスクでは 3 章で述べたセンサ以外の入力 $m_t$ は用いない．

行為シェマに内における Actor の基底関数としては Gauss 型の RBF を各次元 4 個（計 $4 \times 4 \times 4 \times 4 = 256$ 個）ずつ，Critic には各次元 5 個ずつ（計 $5 \times 5 \times 5 \times 5 = 625$ 個）を等間隔に配置した．これらの配置はヒューリスティックによる．本実験では actor と critic の基底関数の粒度を変えているが特にその必要はない．また，実験を通して強化学習のメタパラメータはそれぞれ $\gamma = 0.6, \lambda = 1.0, \alpha = 0.1$ に設定した．また，ロボットの行動周期は 2[Hz] とした．

### 6.2 中心への位置制御の獲得

本節では提案手法を用いて，ボールの平板上での中心位置制御が可能であることを示す．報酬系としては以下を用意した．これは単純にボールが平板の中央に来たときにエージェントに報酬を与える．

$$r_t = exp(-\frac{1}{2}((\frac{x}{0.5})^2 + (\frac{y}{0.5})^2)) \quad (21)$$

行為シェマは初期状態では Table 1 に示す 3 つのものを用意する．双シェマモデルでは行為シェマの強化学習のための探索のみならず，知覚シェマのモデル学習のための探索も行なう必要があり，複数の行為シェマを切り替えながら学習することが重要である．よって，タスクの目的である $IS_2$ 以外についても適宜選択し利用する．ただし，$IS_2$ を選択するときのみ $IS_2$ は強化学習を行なうものとする．10000[s] の間学習を行なった結果を Fig. 4 に示す．比較対象として，連続空間の Actor-Critic 手法で基底に同様の RBF を用いたものを示している．この実験の間は Cart は静止しているために，Balancer にとって環境のダイナミクスは常に一定である．両者ともに制御は成功へと収束していくことがわかるが，提案手法のほうが収束は遅くなっている．これは両者で行動空間の定義が異なり通常の Actor-Critic 手法では 2 次元の行動空間であったのが，双シェマモデルにおける行為シェマの行動空間は状態空間に等しくなるので 4 次元になり，方策の探索

空間が広がったためと考えられる．ただし，タスクや基底の取り方によっては双シェマモデルのほうが速く収束する例も報告されている[4]．ただし，本研究の主旨は学習の高速化にはないので，それを除けば学習能力としてはこの時点では両者大きな差はない．このような静的環境下ではなく，動的環境下において行なわれる次章以下のタスクにおいて，この二つのモデルの差は明確になっていく．

## 7. 汎化行為概念の獲得

さて，はじめにも述べたように，われわれ人間は異なる環境ダイナミクス下でも，同様の行為概念を用いた行動を行なう．たとえば，歩くという場合，坂を歩く場合でも，平地を歩く場合でも，砂地を歩く場合でも，その歩くという汎化された行為概念は変わらない．行為シェマにより獲得された行動はこのようなものに近い．そのことを示すために，獲得された行為概念の環境の変化に対する再利用可能性を調べる実験を行なう．

### 7.1 実験環境

実験のパラメータや系は前章と同じものを用いる．環境の変化としてはCartの路面の変化やコーナーでの遠心力を考える．加速する電車の中でバランスを保つためには，同じ立つという行動であっても，停車時や等速運動時とは違う力の入れ方をしないといけないという経験はだれしもがもっているであろう．実験のシナリオとしては以下のものを考える．

（1） 初め10000[s]の間，Cartは停止し，Balancerはボール制御の強化学習を行なう（$\alpha=(0,0)$）．（状況A）

（2） その後5000[s]の間，Cartが等速で約10°の坂を上り始める（$\alpha=(0,-2)$）．Balancerは強化学習を続ける．（状況B）

（3） 獲得した行為概念の汎化性を調べるため，ここでBalancerの強化学習を止め，15000[s]から20000[s]の間は，Cartは200[s]ごとに状況Aと等価な等速運動，状況Bと同じ登坂を繰り返す．Balancerは獲得した方策を実行し続ける．（状況C）

（4） その後にCartは半径8[m]の円周上を秒速4[m/s]（時速14.4[km/h]）で周回を始める（$\alpha=(2,0)$）（状況D）．

まとめると状況Aと状況Bの間に二種類の環境下で強化学習を行なう．状況Cでは一度学習経験のある環境での獲得された行動概念の再利用性を調べ，状況Dでは中心位置制御というやるべきタスクは同じにもかかわらず，今まで経験していない環境に入っていった場合での獲得行為概念の再利用可能性，つまり汎化性を調べる．比較するのは以下の4つである．

エージェント1. 従来のActor-Critic手法
エージェント2. 従来のActor-Critic手法（強化学習を止めない）
エージェント3. 双シェマモデル（知覚シェマの分化を考慮しない）
エージェント4. 双シェマモデル

ここでエージェント2は15000[s]における強化学習の停止を行なわず，強化学習を続ける場合の学習と比較検討する．エージェント3は前章で示した知覚シェマの分化の機構をはずしたものである．これによって，過去に経験した環境についての記憶を分散的に保持することが出来なくなる．知覚シェマの選択則については知覚シェマが分化しないために常に同じ知覚シェマを選択する．実験のパラメータについては前章と同様のものを用い同化係数については$k=2.3$に設定した．

### 7.2 実験結果

以上の4つの手法のもと得られた結果について，報酬の重み付時間平均の推移をFig. 5に示す．それぞれ6回同様の実験を行ない，その平均を取っている．状況A,Bではどのエージェントも十分な時間をかけてやると同様に環境への適応を果たす．平均報酬の変化だけを見ると大きな違いはみられないが，実際におのおののエージェント内で行なわれる学習は異なる．エージェント1，2では環境が変化したために，今までの方策が利用できなくなり修正を求められる．この修正を強化学習ベースで行なっている．よって状況Aにおける方策は忘却される．グラフでは報酬の時間平均を取っているため，大きな差は出ていないが，エージェント3は強化学習より学習速度の速い教師ありのモデル学習を行なうために，強化学習はほとんど用いず，ここまでで獲得した行為シェマを維持したまま，知覚シェマの均衡化を通して環境に適応する．ただし，状況Aにおいて学習した知覚シェマは忘れ去られてしまう．最後にエージェント4は環境の変化に伴い知覚シェマの分化を行ない，この新たな知覚シェマをもって状況Bへと適応する．このとき状況Aで獲得された行為シェマ，知覚シェマは忘却されることなく保持される．

劇的な違いがあらわになるのは状況C以降である．エージェント1は状況Aでの方策をすでに忘却しているので，状況Bに相当する坂道に遭遇したときには上手くボールを制御できるが，状況Aに相当する状況C内でCartが等速で平地をうごいているときには上手くボールを制御することが出来ない．よって，この二つのダイナミクスの切り替わりの中ではエージェント1は一貫した制御を行なうことが出来ない．これに対しエージェント3は知覚シェマの均衡化のダイナミクスを通して随時適応していくことができるので，エージェント1に比べるとダイナミクスの切り替わりにもかかわらず，高い報酬を得続けることが出来る．しかし，知覚シェマの均衡化も時間を要するので，200[s]ごとの切り替わりには追従することが出来ない．これに対し，エージェント4はFig. 7のように0～15000[s]を通じて獲得，保持された状況A用の知覚シェマと状況B用の知覚シェマをスイッチングするだけで新たな環境に適応できるため，状況Cでも高い報酬をとり続けることができる．その後，状況Dになるとエージェント3は知覚シェマの均衡化，エージェント4は知覚シェマの分化および均衡化を通じて，強化学習つまり行為概念の新たな学習を行なわなくても適応することが出来る．これに比べて

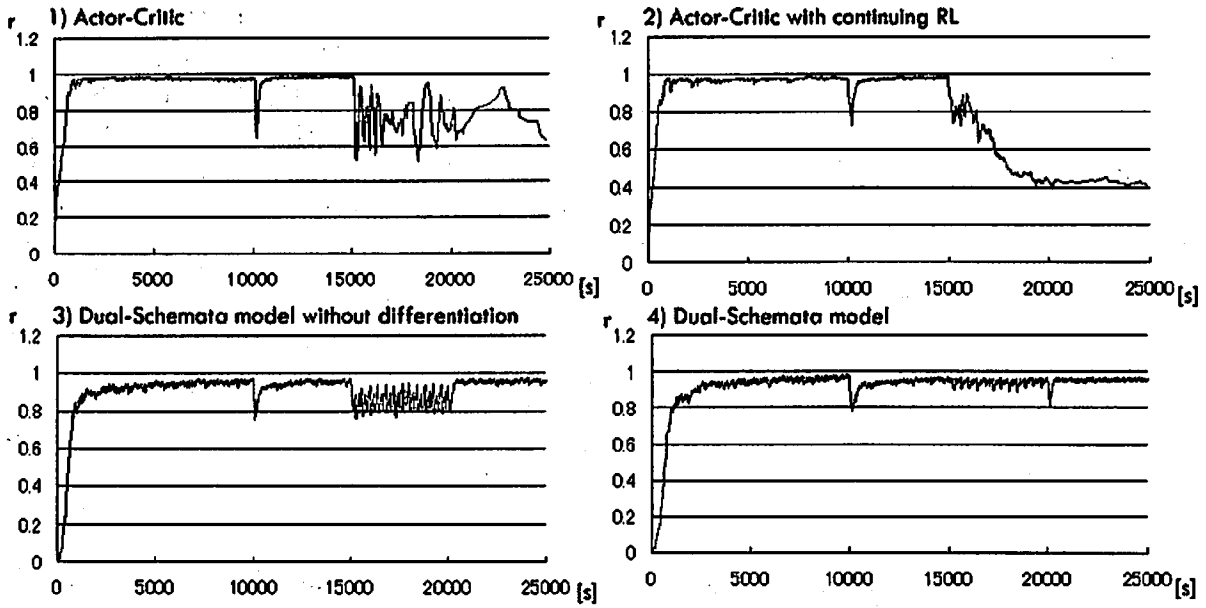


**Fig. 5** transition of weighted averaged rewards which were obtained by the 4 kinds of learning architectures

エージェント1は状況B用の方策を採り続けなければならないために，良い結果を出すことができない．

最も悪いのは強化学習をダイナミクスがシフトしていくな環境下で行ない続けた場合である．通常環境のダイナミクスが変化した場合，HJB方程式の解としての強化学習の最適解は変わってしまう[15]．よって，ダイナミクスが変化する中での強化学習はしばしば，片一方のダイナミクス下だけで学習した結果を両方のダイナミクスに用いた場合より悪い結果を引き起こす[14]．今回の結果はそのような場合である．一定でないダイナミクスに支配された状況Cを通して崩壊させられた価値・方策は状況Dにおいて新たな環境において強化学習を始めることにすら悪影響を与えているように見える．ただし，エージェント1と2の獲得報酬の優劣はどのようなダイナミクスの変化を与えるかと，タスクに依存し，優劣は逆転する場合もある[4]．重要なのは通常の強化学習手法では環境が変わるごとに外部から与えられるたった1次元の報酬という情報を頼りに再学習を行なう必要があるが，これは過去の行為概念を破壊し，再構築するプロセスにほかならない．これに対し，双シェマモデルでは行為シェマは状況Aで獲得された後は，破壊されることはない．状況A～Dを通して一つの行為シェマが用いられる．つまり，これは環境変化に依存しない行為概念が獲得できたといえる．

### 7.3 行為シェマとセンサ空間上でのアトラクタ

主体にとっての一纏まりの行為とは，おのずからのセンサ空間内に描かれるアトラクタとして捉えられるのではないかという提案を2章において行なった．アトラクタとは主に力学系のダイナミクスで一点に収斂していくもの（点アトラクタ）や，定常的な軌道へと収斂していくもの（リミットサイクル）を指すが，それぞれがリーチングなどのゴールのあるタスクと，歩行などの周期的な運動に対応している．リミットサイクルを形成する非線形振動子の結合によって，歩行を実現しようという研究は近年非常に盛んである[19]．われわれはこのセンサ空間内の力学系を行為シェマの内部関数において実現することにより，外部ダイナミクスに依存しない行為概念の内部表現を試みたわけである．

ところで，本稿の枠組みを用いて得られる汎化行為概念は，前節で述べたようなボールの中心への位置制御のように，ある固定値に安定化させるようなものだけではない．例として平面の中央で約半径0.5[m]の円軌道に沿って周期運動を行なわせるような平板の制御に関する学習に関する実験を行なった．このための報酬系を以下のように定義する．基本的には平板の中心周りの角速度 $\omega$ を $\pm 1$ で飽和させた項と半径0.5の円軌道上で最大となる項の積で構成されている．

$$r_t = max(-1.0, min(1.0, \omega)) * exp(-\frac{1}{2} * (\frac{R-0.5}{0.3})^2) \quad (22)$$

学習結果の詳細については，紙面の制約上，別項に譲るが，結果的に，エージェントは前節同様に複数環境下で円運動を起こすことのできる行為シェマを獲得した．その結果，板上の適当な点に置いたボールは，獲得された行為シェマによりロボットの視点から見たセンサ空間内のアトラクタに引き込まれていくことが確認された (Fig. 6)．このように，本手法を用いることで，さまざまなタスクに対して環境ダイナミクスの変化に依存しない汎化行為概念を個別に獲得することが出来る．また，この結果に例証されるように，本稿で導入した汎化行為概念とは異なるタスクに対する複数の行為を包含するような，上位の一般化概念ではなく，タスクごとに個別でありながら，環境の変化に対して汎化性を有した行為概念を指している．さらに，2.3節で述べたように，本稿の定義

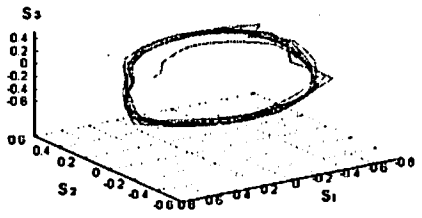

Fig. 6 an obtained attractor in the agent's sensor space

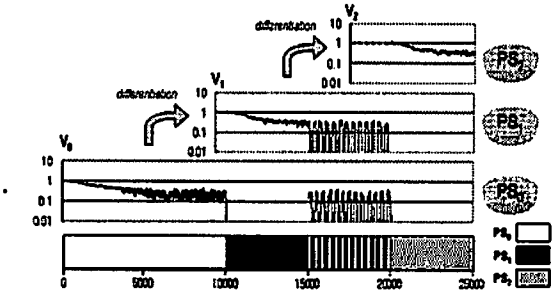

Fig. 7 (top) time course of schemata's activity and differentiation process, (bottom) selected schema at each time

による汎化行為概念とはモータ出力とは独立した，センサ空間内のアトラクタとして描かれるが，前節で獲得された盤上の中心への制御が点アトラクタとして表現されるのに対し，このような周期運動へ引き込んでいく行為はセンサ空間上のリミットサイクルとして描かれていることがわかる．

## 8. まとめ

本論文では双シェマモデルにおける行為シェマの学習に強化学習を導入し，階層的モジュール型学習モデルとして提案した．この学習モデルでは上位層に相当する行為シェマが環境のダイナミクスとは独立に求められる行為をセンサ空間内でのアトラクタとして獲得されるために，複数環境ダイナミクス間で利用可能な行為概念を獲得することが出来る．また，双シェマモデルの基本的特徴である，知覚シェマの分化の仕組みと組み合わせることにより，すでに過去の経験からシェマを形成済みの環境であれば，知覚シェマ選択器が自律的に知覚シェマを選択するだけで，追加の学習なしに新たな環境に再適応できることを示した．また本稿では特に実験は示さなかったが，獲得された知覚シェマは中心への制御を行なう行為シェマだけではなく，ほかの行為シェマによって再利用されることが出来る．双シェマモデルにおいて重要なのは，複数環境において行為シェマ（汎化行為概念）が再利用可能なだけでなく，複数行為間で知覚シェマ（環境との相互作用のモデル）が再利用可能だということである．動的環境下で活動する自律適応系にとって，このような二重の再利用可能性が非常に重要だとわれわれは考えている．

今後の双シェマモデルの展望としては，累増的な行為シェマの獲得や，より非線形な系へのモデルの拡張，獲得した表象群のエージェント間コミュニケーションへの利用といったものがある．

謝辞 本研究は21世紀COEプロジェクト「動的機能機械システムの数理モデルと設計論」の一部として遂行いたしました．また，重要な示唆を与えてくれた京都自律適応系研究会ならびに京都大学機械システム創成学研究室のメンバーに感謝します．

### 参考文献

1) T. Taniguchi and T. Sawaragi: 双シェマモデル:自己組織化機械学習手法の提案，日本人工知能学会論文誌，19-6, 493/501 (2004)
2) 谷口，楾木：身体と環境の相互作用を通した記号創発：表象生成の身体依存性についての構成論，システム・制御・情報学会誌，49-12, 440/449 (2005)
3) T. Taniguchi and T. Sawaragi: Design and Performance of Symbols Self-Organized within an Autonomous Agent Interacting with Varied Environments, *IEEE International Workshop on RO-MAN* (2004)
4) T. Taniguchi and T. Sawaragi: Adaptive organization of generalized behavioral concepts for autonomous robots: Schema-based modular reinforcement learning, IEEE International Symposium on Computational Intelligence in Robotics and Automation, in CD-ROM (2005)
5) 谷口，楾木：顔ロボットの移動物追跡のための運動記憶の動的構成による環境適応，SICEシステム・情報部門学術講演会2004講演論文集，185/190 (2004)
6) D. M. Wolpert and M. Kawato: Multiple paired forward and inverse models for motor control, Neural Networks, 11, 1317/1329 (1998)
7) J. Tani and S. Nolfi: Learning to perceive the world as articulated: an approach for hierarchical learning in sensory-motor systems, Neural Networks, 12, 131/1141 (1999)
8) 高橋，浅田：階層型学習機構における状態行動空間の構成，日本ロボット学会誌，21-2, 164/171 (2003)
9) J. Morimoto and K. Doya, Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning, Robotics and Autonoumous Systems, 36, 37/51 (2001)
10) 港，浅田：環境の変化に適応する移動ロボットの行動獲得，日本ロボット学会誌，18-5, 706/712 (2000)
11) J. Morimoto and K. Doya: Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning, Robotics and Autonoumous Systems, 36, 37/51 (2001)
12) K. Doya et al: Multiple Model-based Reinforcement Learning, Neural Computation, 14, 1347/1369 (2000)
13) 杉本，鮫島，銅谷，川人：ダイナミクスの線形性に基づいて状態空間を分割する階層型強化学習，電子情報通信学会ニューラルコンピューティング研究会6月報告，NC2003-16 (2003)
14) 枝澤，高橋，浅田：複数学習器を用いたマルチエージェント環境における行動獲得；第22回日本ロボット学会学術講演会，in CD-ROM (2004)
15) K. Doya: Reinforcement Learning In Continous Time and Space, Neural Computation, 12-1, 219/245 (2000)
16) H. Kimura and S. Kobayashi, An Analysis of Actor/Critic Algorithms using Eligibility Traces: Reinforcement Learning with Imperfect Value Function, 15th International Conference on Machine Learning, 278/286 (1998)
17) R.A. Schmidt: A Schema Theory of Discrete Motor Skill Learning, Psychological Review, 82-4 (1975)
18) 小林，野村，Pakdaman，佐藤：二足歩行運動の動的安定性，信学技報MBE97-48, 1997-7 (1997)

19) 富田，矢野：大脳基底核ー脳幹系のモデル化による二足歩行運動の創発的リアルタイム制御，第16回自律分散システムシンポジウム資料，5/10 (2004)
20) V. Gullapalli: A Stochastic Reinforcement Learning Algorithm for Learning Real-Valued Functions, Neural Networks, 3, 671/692 (1990)
21) R.S. Sutton and A.G. Barto, Reinforcement Learning: An Introduction, MIT Press (1998)
22) S. Schaal, A. Ijspeert and A. Billard: Computational Approaches to Motor Learning by Imitation, Philosophical Transaction of the Royal Society of London: Series B, Biological Sciences 358: 537/547 (2003)

## [著 者 紹 介]

**谷 口 忠 大（学生会員）**

2003年京都大学工学研究科精密工学専攻修士課程修了，現在同博士課程に在籍．2003年度から21世紀COEプロジェクト「動的機械システムの数理モデルと設計論」若手研究助成対象者．2005年度から日本学術振興会特別研究員 (DC2)．生体の認知的発達における，ボトムアップな記号組織化，意味生成に注目し，自律適応系の設計論を研究している．システム情報制御学会，日本神経回路学会，日本人工知能学会，などの会員．

**槇 木 哲 夫（正会員）**

1986年京都大学大学院工学研究科博士課程指導認定退学．同工学部助手．94年同助教授，2002年同教授．2005年改組により同研究科機械理工学専攻教授，その間，91～92年米国スタンフォード大学客員研究員．京都大学工学博士．現在，人間ー機械共存環境下での協調システムの設計・解析と知的支援などに関する研究に従事．ヒューマンインターフェース学会論文賞，計測自動制御学会学術奨励賞，論文賞，著述賞，等受賞．日本機械学会，システム情報学会，IEEE等の会員．