多自由度ロボットの実時間学習制御：離散状態表現の自動生成と遷移モデル学習

○木村 元，　小林 重信
東京工業大学 大学院総合理工学研究科

Real time learning control of high D.O.F. robots: Automatic generation of discrete states and learning transition models

○ Hajime Kimura,　Shigenobu Kobayashi
Interdisciplinary Graduate School of Science and Eng., Tokyo Institute of Technology

Abstract: We present a model-based RL approach to cope with continuous space of high D.O.F. robots, combining model learning and an actor-critic method. The model learner generates a discrete state-transition model, that helps improvement of both the policy and state-representation. In general, model-based methods tends to fail in Non-Markovian problems, but the proposed method, using actor-critic, can find good policies in such environments.

1　はじめに

本論文では，４脚ロボットやヒューマノイド等の多自由度ロボットにおいて，学習によって自動的に制御規則を獲得するための新しい接近法を提案する．上記のような多自由度ロボットの学習問題では，以下のような特徴がある：

1. 状態・行動空間は連続で高次元の非常に大きな空間になる．
2. 状態遷移に不確実性が伴う．
3. 特に実機では，動かすためのコストが高いため，試行錯誤回数は制限される．

ここで特に３番目の項目（試行回数の制限）に関連して，下記の要求が挙げられる：

- 限られた試行回数の範囲内で，最適でなくても良いので，それなりに良い制御規則を得たい
- その後試行を追加していくことで制御規則を改善したい
- 新たに試行するよりも，試行データを全て記憶しておくコストのほうが安い
- 人間が知識を持っている場合には簡単に与えることができ，その知識に誤りがあっても自動的に修正できる

状態表現の与え方は，学習速度や得られる制御規則の質に大きな影響を与える．連続で膨大な空間の扱いが求められる場合は特に顕著であり，単純なグリッドによる離散化ではすぐに空間爆発を起こし，学習不能に陥る．高い自由度を持つヒューマノイド型の実機を制御するための状態表現方法としてStateNet[金広，稲葉，井上2002]が提案されている．StateNetは代表的な姿勢を状態として表し，その間の遷移を行動とする単純で優れた表現であり，ロボットに人間の知識を与えることも容易である．しかし状態表現を設計者が全て与える必要があり，また遷移の不確実性をほとんど考慮していない問題がある．

状態遷移に不確実性を伴う環境における意思決定プラニング方法として，環境をマルコフ決定過程としてモデル化し，ダイナミックプログラミング (以下 DP) によって最適な制御規則を得る方法がある [Barto, Bradtke, Singh 95]．これはモデルベース手法と呼ばれ，試行錯誤を通じて環境のモデルを学習・構築していく．この手法は，モデルが獲得できれば別のタスクの学習も容易に行えたり，異なるタスクにおける試行錯誤データも同一モデルの学習へ無駄なく利用できる利点がある．しかし，この手法は正確なモデルを獲得しなければならないという問題がある．特に本論文で取り上げるロボットでは，1) あまり多くのデータが得られない，2) 状態観測に不完全性が存在する箇所がある，という問題があるため，獲得されるモデルは不完全となりがちである．実際，実験で示すが，推定モデルからの DP で得られた政策を実機に適用しても，ほとんど役に立たない．状態観測の不完全性の原因は，センサノイズによる場合もあるが，状態空間を離散化する際の分割が粗いために生じる非マルコフ性の影響が大きい．通常，この問題への対処として状態分割を細かくする方法が取られるが，先に述べとおり多次元空間においては，単純なグリッド分割は状態空間の爆発を起こす．不完全観測が疑われる領域を適応的に分割していく方

法もあるが，不完全観測をゼロにすることは不可能であり，実験でも示すが DP だけに頼った政策獲得には限界がある．

近年，モデルを使わずに試行錯誤によって制御規則を直接探索する強化学習による接近法の有用性が示されている．Q-learning [Watkins 1992] や TD 法 [Sutton & Barto 1998] は，DP の理論に基づいているため非マルコフ性の存在する環境では学習は安定しないが，アルゴリズムが単純な割にそれなりに良い動作を獲得できるためよく用いられる．Actor-critic など政策勾配法は，非マルコフ性のために DP を基礎とする方法 (Q-learning など) が対処できない環境でも適切な政策を獲得できる [木村，小林 2001]．強化学習において連続な空間を扱う方法として，CMAC など連続関数近似を用いて value 関数を表現する方法 [Sutton & Barto 1998] や，空間を適応的に分割・離散化していく方法 [浅田，野田，細田 1997；高橋，浅田 1999；矢入，堀，中須賀 2001] などが提案されている．しかしこれらの強化学習法は，学習に多くの試行回数を要する問題があるため，本研究の問題のように試行回数に制限がある場合には十分な学習が行いにくい．また，異なるタスクを学習する場合は，別のタスクの試行錯誤データや政策が再利用できない．

本論文では，多自由度ロボットの制御規則獲得問題に対し，StateNet に類似した状態-行動表現を適用し，環境の遷移モデルの学習と強化学習を融合して制御規則を獲得する方法を提案する．本手法の特徴は，非マルコフ性が生じがちな粗い状態表現のもとで大雑把な遷移モデルを構築し，そのような不完全なモデルを以下の 3 つの事項に利用する点にある：

- モデルから推定される状態評価値を利用して強化学習を加速させる
- モデルを利用して別のタスクの初期政策を生成し，無駄な探索を省く
- データに対するモデルの不完全さの大きな状態を検出することで状態表現を改善する

本手法を仮想的な多足ほふくロボット問題へ適用し，シミュレーションにより有効性を示す．さらに，足先に接触センサを取り付けた 4 足実ロボットへ適用した予備的実験を示す．

2　問題設定および学習目標

本論文では Fig.1 に示す 4 足ロボットを 15 分程度の実時間での試行錯誤で前進動作の制御規則を獲得することを目標とする．しかし実機では，様々なアルゴリズムの動作確認や定量的評価ができるほどの実験データを得ることは物理的に困難である．また一般的に問題の規模が異なる場合に，使用する学習アルゴリズムで正しく学習できるかどうか検証する必要がある．そこで問題規模を変化させることのできる簡易な問題として，仮想的な多自由度ほふくロボットによるほふく動作獲得問題を考え，計算機上にそのシミュレータを作成し，この学習によって定量的評価を行う．

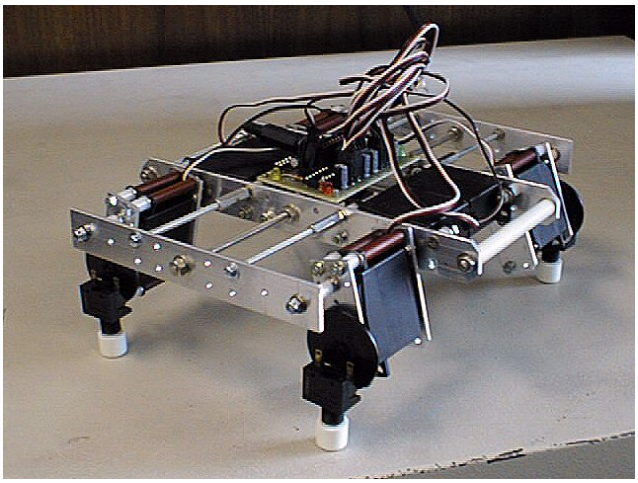

Figure 1: 4 足ロボット

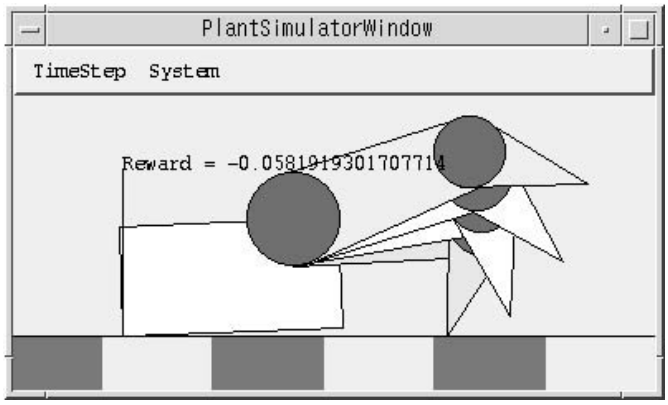

Figure 2: 多自由度ほふくロボットシミュレータ (図は $n = 4$，4 本足の場合)．写真右側が前方．長方形がロボット本体，丸は各足の関節を表す．

これらのロボットを制御するコントローラが学習主体の「エージェント」になり，ロボット本体を含めた外界がエージェントにとっての「環境」になる．エージェントは事前に環境やロボットのダイナミクスを知らされているとは限らないため，試行錯誤を通じて制御規則を形成していくことが求められる．各時間ステップにおいてエージェントは環境の状態を観測し，行動を選択する．行動を実行後，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．報酬はロボットに実行させるべきタスクの達

成を反映する．設計者はロボットにさせたいタスクを報酬として表現しなければならない．本論文では前進動作をさせるので，各ステップでのボディの移動速度を与える．

これらのロボットは，値域が既知の連続値および離散値の状態変数を持つ．関節の角度を連続状態変数で表し，各足のタッチセンサの状態を離散状態変数で表す．学習主体であるエージェントは，これら多次元の状態ベクトルを環境の状態として観測する．エージェントからの行動出力は，関節の角度の目標値を与える．よって行動は連続状態変数と同次元の連続値ベクトルである．エージェントが行動を出力すると，ロボットは関節の角度を目標値へ近づける方向へ動かして状態遷移を行う．関節が目標値になるか，あるいは途中でセンサの値が変化するとイベントが発生し，そこまでの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．よって，途中でセンサの出力が変化した場合には，次のステップでの連続値状態は出力した行動と同じになるとは限らず，遷移先についても不確実性が存在する．

ほふくロボット (Fig.2) は n 本の足 (n は任意に設定可能) を持つ移動ロボットであり，問題規模としては $2n$ 自由度となる．各足には関節が 2 つあり，それぞれの足を地面と接した状態で動かすことによって本体が前後に移動する．ここで注意すべきことは，複数の足で同時に地面を漕ぐことはできず，常に本体を支持している 1 本の足のみで漕いだ分だけロボット本体が移動するよう設定してある点である．つまり 2 本以上の足で同時に漕ぐことに優位性は無いのである．この設定により，足の本数が増えていったときに効率良く前進するためには，地面を漕ぐ作業を各足で分担しなければならない．状態入力 s_t，行動出力 a_t は 4 足歩行ロボットと同じで，それぞれ現在の関節角度，次の時刻に取るべき関節角度の目標位置とした．関節はロボット全体で $2n$ 個となるため，s_t, a_t は $2n$ 次元のベクトルとなる．状態・行動空間は 0 から 1 までの区間に正規化される．特に足の本数 $n = 4$ のときは 4 本足ロボットと同一の状態-行動空間を持つ同規模の学習問題となるが，解空間は異なるため，獲得される解が一致する訳ではない．またここでは簡単のため，エージェントの観測する状態入力に外乱はないものとした．

４足ロボット後方には，ボディの移動を検出して報酬信号を生成するための車輪が２個付いており，ロボットの前進速度と旋回速度を計測できる．ロボットの学習目標はまっすぐに前進することなので，各時間ステップでの報酬は，前進速度から旋回速度を差し引いて与えた．各足先にはタッチセンサが付いており，値が変化するたびに意思決定のイベントが発生する点はほふくロボットと同じである．

解空間の特徴として，ほふくロボットの場合，最適ではないがそれなりに前進できる解が多数存在する問題であるのに対し，４足ロボットは前進できる解は少なく，へたに動くとボディが旋回して負の報酬が入るため「動かない」という強力な局所解が存在する問題である．

3　モデル学習と強化学習の融合

本論文ではモデル学習と強化学習を融合した新しいアルゴリズムを提案する．以下に概要を述べる．

3.1　状態・行動表現方法

【代表点による状態空間離散化】
学習器は連続な空間中にいくつかの点を定義し，距離が閾値以下の領域を代表させて離散状態とする．
【生成的な離散状態表現】
意思決定のイベント発生時の状態入力に対して，最も近い代表点の距離が閾値を超える場合，その位置に新たに代表点を置いて離散状態を生成する．
【状態間の遷移を行動として定義】
行動は目標とする状態として表現される．しかし必ずしも目標状態へ遷移できるとは限らず，別の状態へ遷移することもある．本研究では，状態の増加に伴って類似の行動がむやみに増加しないよう，目標とする状態は数を固定する．

この状態-行動表現は StateNet[金広，稲葉，井上 2002] に類似しており，連続で膨大な空間を高々数十コのごく少ない離散状態で表現することで，状態空間の爆発を防いでいるが，必要に応じて状態が適応的に増えていく点が異なる．この表現は，タスク達成のための動作が代表的な「姿勢」とその間の遷移という人間にとってわかり易い表現で示されるため，知識の組込みが容易で，また学習結果を自然言語で伝達しやすい．また，代表点の座標を動かすだけで簡単に状態の境界を調節できる利点がある．しかし，行動表現が，目標とする状態であるため，ロボットに与えるべき行動指令が別の表現，例えばモータのトルクなどの場合には，ロボットに目標状態を目指すように動作させるコントローラを下位層にデザインする必要がある．本研究で扱うロボットでは，行動出力は連続状態変数を目標値へ近づけるような仕様なので，このような下位層のコントローラがすでにできていると考えられる．ただし，離散状態変数については下位層のコントローラだけでは制御しようがないため，StateNet 表現のレベルで遷移を制御する必要がある．

3.2 遷移モデル学習

エージェントは環境における連続な状態観測・行動観測を上記の離散状態・行動表現に変換し，状態遷移回数をカウントして遷移確率および報酬の期待値を最尤推定し，マルコフ決定過程モデルを生成する．ただし，1) 状態空間離散化が粗いため，状態観測が不完全 2) データが少ない, という条件下であるため，推定されるモデルは不正確になりがちである．本手法は，このように不完全なモデルを積極的に利用する点に特徴がある．

3.3 タスク達成のための制御規則獲得

1. Observe state s_t, choose action a_t with probability $\pi(a_t|\theta, s_t)$, and perform it.
2. Observe immediate reward r_t, resulting state s_{t+1}, and calculate the TD-error according to
$$(\text{TD-error}) = r_t + \gamma \hat{V}(s_{t+1}) - \hat{V}(s_t) \ , \quad (1)$$
where $0 \leq \gamma \leq 1$ is the discount factor, $\hat{V}(s)$ is an estimated value function by the critic.
3. Update the estimating value function $\hat{V}(s)$ in the critic according to DP in the estimated model.
4. Update the actor's stochastic policy by
$$\begin{aligned} e_\pi(t) &= \frac{\partial}{\partial\theta} \ln\Big(\pi(a_t|\theta, s_t)\Big) \ , \quad (2) \\ \overline{e_\pi}(t) &\leftarrow e_\pi(t) + \overline{e_\pi}(t) \ , \\ \Delta\theta(t) &= (\text{TD-error})\, \overline{e_\pi}(t) \ , \\ \theta &\leftarrow \theta + \alpha_\pi \, \Delta\theta(t) \ , \end{aligned}$$
where e_π is the eligibility of the policy parameter θ, $\overline{e_\pi}$ is its trace, and α_π is a learning rate.
5. Discount the eligibility traces as follows:
$$\overline{e_\pi}(t+1) \leftarrow \gamma \, \lambda_\pi \, \overline{e_\pi}(t) \ , \quad (3)$$
where λ_π $(0 \leq \lambda_\pi \leq 1)$ is a discount factor in the actor.
6. Let $t \leftarrow t+1$, and go to step 1.

Figure 3: Actor に適正度の履歴，critic に環境モデル+DP を用いた actor-critic アルゴリズム

状態観測の不完全性などにより非マルコフ性の存在する環境における有効な強化学習法として actor に適正度の履歴を用いた actor-aritic アルゴリズムがある [木村，小林 2000]．この手法では，critic は状態評価値の推定学習を担当し，actor は行動選択を担当しながら critic の推定した評価値と環境からの報酬を手がかりにして政策を改善する．注目すべき点は，critic において学習に失敗し，正確な行動評価の推定値が得られなくても，学習に時間を要するものの actor は環境からの報酬を用いて政策を改善できることである．観測の不完全性が存在する状態は，全状態の中でも一部であることが多いので，たいていの場合は critic により学習を加速させる効果が得られる．

本研究ではこの点に注目し，従来 TD 法を用いて状態評価値の推定学習を行っていた部分をモデル学習と DP 処理に置き換える．試行データを無駄なく利用したモデルと DP によって，より正確な状態評価値を actor へ提示することにより，政策改善をさらに加速する．Actor-critic アルゴリズムの特徴を引継いでいるので，モデル学習と政策改善を同時に行うことが可能であり，任意の学習ステップで学習を打ち切っても，それなりに改善された確率的政策を得られる特徴を有する．

また，新しいタスクを学習する場合において，試行前にモデルが学習されている場合は，強化学習の初期政策構築に DP の計算結果を利用することで無駄な試行錯誤を極力排除することができる．不完全観測の存在する特定の状態では，DP による計算で示された最適行動は誤りである可能性が高いのだが，それ以外の状態ではだいたい正しい行動を示していることが多い．よって強化学習によって誤った行動を修正していくだけなので，完全にゼロから政策を探索するよりも効率が良いと考えられる．

Fig.3 は提案する政策獲得手法の詳細である．Actor では確率的政策がパラメータ関数表現され，価値関数の勾配を用いて更新する．エージェントが政策 π のもとで観測 s で行動 a を選択する確率を関数 $\pi(a|\theta, s)$ で表す．エージェントは内部変数 θ を調節することにより確率的政策 π を変える．パラメータ λ_π は actor の適正度の履歴について以下の性質を特徴付ける：λ_π が 0 に近い場合，critic で推定された価値関数 $\hat{V}$ の勾配方向へ政策が更新される．λ_π が 1 に近い場合，実際の報酬のサンプル合計による評価値の勾配方向へ政策が更新される．環境の非マルコフ性や critic の不完全性に対処するためには λ_π を 1 に近づけることが望ましい．

Critic では，遷移モデルより DP を用いて政策 π のもとでの状態評価値 $\hat{V}(s)$ を以下のように計算する：

1. 全ての s について $\hat{V}(s)$ に以下を代入：
$\sum_{s'} \sum_a \Pr(s'|s,a)\pi(a|\theta,s) \Big[R^a(s,s') + \gamma \hat{V}(s')\Big]$
2. 処理手順 1 を収束するまで繰り返す

ここで状態 s にて行動 a をとった場合の s' への遷移確率 $Pr(s'|s,a)$ および報酬期待値 $R^a(s,s')$ は遷移モデルより推定値として得られ，行動選択確率 (政策)$\pi(a|\theta,s)$ は actor より得られる．

3.4 モデルと履歴データによる状態表現の改善

【状態代表点をデータの重心へ移動】
本手法では状態入力に最も近い代表点の距離が閾値を超える場合に，その位置に新たに代表点を置いて離散状態を生成するため，モデルの良し悪しが偶然に作用されやすい．状態の代表点をイベント発生座標の重心へ調節することで，改善が期待できる．

【推定モデルの尤度に応じた状態分割】
履歴データに対する推定モデルの尤度に着目して状態分割を行うことにより状態表現を改善し，粗い状態離散化による非マルコフ性を軽減する．分割が粗くて不完全観測を起こしている状態は，その状態内の領域に応じて遷移確率が大きく異なっている．よって，領域を分割して別々の２状態だとして履歴データから遷移確率分布を作り直し，該当する状態からの遷移の尤度が分割後のほうが有意に増加していれば，その状態は分割すべきであると判断できる．このように遷移データに対する確率モデルの尤度に着目して状態分割する既存方法として，状態判別木を用いた方法 [矢入，堀，中須賀 2001]が提案されている．彼らは判別木を用いるため状態分割が超矩形で行われているのに対し，我々の手法は状態表現が代表点であるので状態空間はボロノイ分割されている点が異なる．また，判別木による分割は，データ全てに対するモデルの尤度最大化を行うのだが，本手法は任意の１状態領域からの遷移の尤度が最大になるよう分割する．よって，他状態から注目している状態領域へ遷移してきた場合，状態が増加しているためデータ全体に対するモデルの尤度は減少してしまうこともあるが，粗い状態分割に起因する非マルコフ性を軽減する意味においては妥当である．分割処理の詳細については別の機会に譲る．

4 実験

提案手法のうち，モデル+DP の actor-critic と，状態分割の効果を確認するため，比較対象としてモデルから DP で求めた最適政策，および通常の actor-critic[木村，小林 2000] の２つを用いた．割引率 $\gamma = 0.9$，提案手法および標準的 actor-critic の actor の学習率は 0.02，critic の学習率は 0.4 に設定した．行動としてとる状態点は以下の 8 点である：(1 1 1 1 1 1 1 1), (0 0 0 0 0 0 0 0), (1 1 1 1 0 0 0 0), (0 0 0 0 1 1 1 1), (1 1 0 0 1 1 0 0), (0 0 1 1 0 0 1 1), (1 0 1 0 1 0 1 0), (0 1 0 1 0 1 0 1)．Fig.4 は距離閾値を 2 に設定しランダムに動作させて状態を 13 コ生成し，そのときの同一状態表現で各手法を適用した場合の 1000 step 毎の試行数と獲得した政策の性能を 5 試行分示したものである．Fig.5，6 も同様にそれぞれ状態を 24，46 コ生成した場合を示す．モデル＋ DP に基づく手法で得た政策では，ほとんど前に進めない．13 状態では有意な差は見られないが，状態数が増加すると提案手法は質の高い解へ収束する傾向が見られる． Fig.7 は 13 状態の設定において，状態表現改善の効果を示す．モデル＋ DP によって安定して前進動作が得られるかどうかによって状態表現の良し悪しを判断した．状態座標をデータ重心へ移動した場合，モデルの尤度は全く変化しないのだが，DP だけでは前進する政策が全く得られなかったのが改善され，安定はしないものの頻繁に解を見つけられる

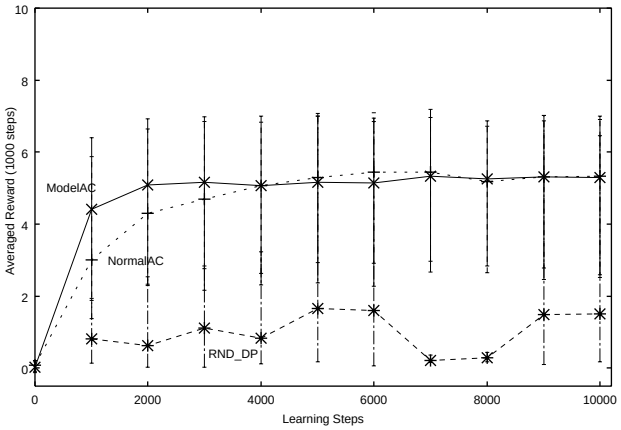

Figure 4: 状態数 13 における試行数と獲得政策の性能 (5 試行)．プロットは平均，誤差棒は上限・下限値．ModelAC は提案手法，NormalAC は通常の actor-critic，RND-DP はランダム試行＋ DP を表す．

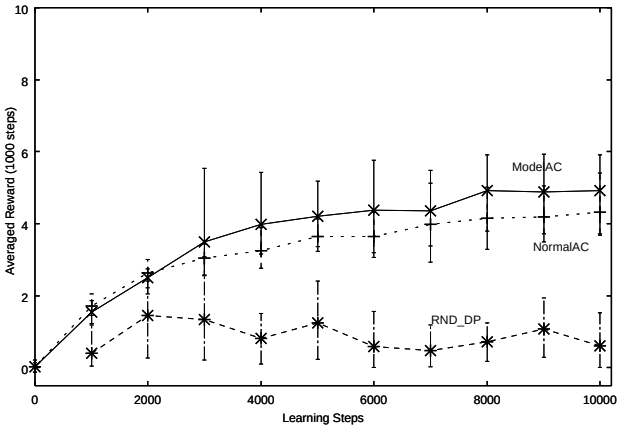

Figure 5: 状態数 24 における試行数と獲得政策の性能 (5 試行)．ModelAC は提案手法，NormalAC は通常の actor-critic，RND-DP はランダム試行＋ DP．

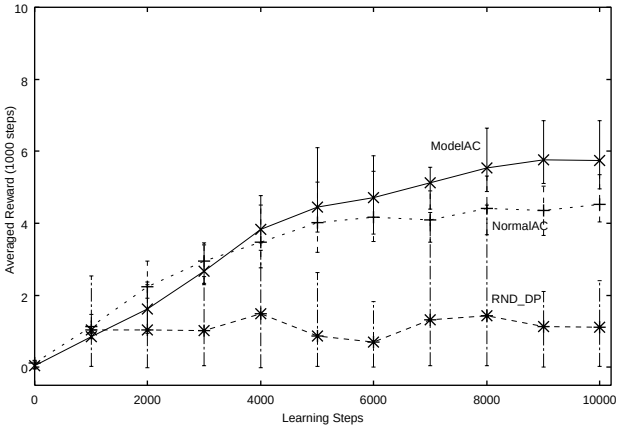

Figure 6: 状態数 46 における試行数と獲得政策の性能 (5 試行). ModelAC は提案手法，NormalAC は通常の actor-critic，RND-DP はランダム試行＋ DP.

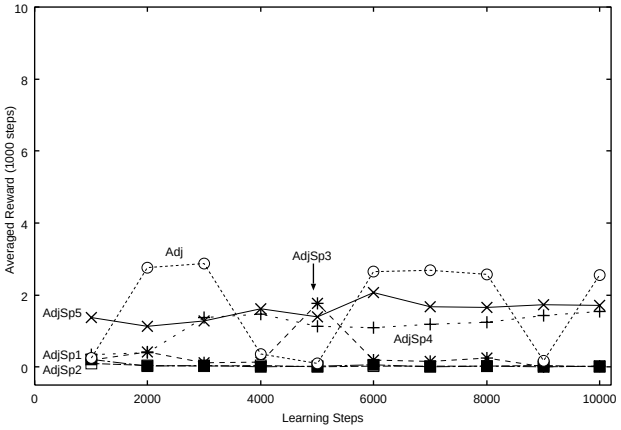

Figure 7: 状態数 13 における状態表現改善の効果．ランダム動作で 1000 step 試行毎に DP を実行して得た政策の性能を示す．Adj は状態座標をデータ重心へ移動した場合，AdjSp1 は Adj の状態表現を分割により 14 状態へ，AdjSp2 は 15 状態，…AdjSp5 は 18 状態である．

ようになっている．分割を増やしていくと，安定してそれなりの解を見つけられるようになる．しかし，本手法ではモデルの尤度は分割を増やすとわずかながら低下していく．

Fig.1 の実機において，上記の 3 手法およびモデルによる初期政策の設定などを適用した結果，残念ながらどの手法も前進動作を獲得できず，ボディを動かさないという局所解から抜け出せなかった．

5　考察とまとめ

【重要な状態への訪問の不足】
環境探査において，行動選択方法は任意であるので，ランダム選択でも良いが，タスク達成に必要な状態領域を十分に探索できないことがある．本論文では紙面の都合で示していないが，ほふくロボットで後進動作を獲得する場合はランダム動作でモデルを構築しても後進動作に必要な行動をとることがほとんど無いため，モデルから政策を得ることは期待できない．実ロボットにおける学習失敗も状況が類似していると考えられる．

【状態表現の偶然性】
本手法では，必要に応じて離散状態を生成するが，その位置は実行される行動に依存する．状態表現の良し悪しは生成される位置に大きく左右される．提案手法のような状態分割では試行毎に性能が大きく変ってしまう．よってこれ以外の方法で良好な状態表現を安定的に得る別な方法が望まれる．

【政策の局所探索の限界】
提案手法は actor-critic の拡張であり，政策を局所探索により改善していく方法であるため，局所解へ陥るとそこから抜け出せない問題を抱えている．今後の展開としては，モデルにおける DP 操作において最適政策を計算し，その政策へ確率的にスイッチするような探索が有望だろう．

参考文献

浅田 稔, 野田 彰一, 細田 耕：ロボットの行動獲得のための状態空間の自律的構成，日本ロボット学会誌 Vol.15, No.6, pp.886–892, 1997.

Barto, A. G., Bradtke, S. J. & Singh, S. P.: Learning to act using real-time dynamic programming, *Artificial Intelligence 72*, pp.81–138. (1995).

金広文男，稲葉雅幸，井上博允：StateNet: 障害回復機能を内蔵する行動空間の状態遷移図表現，日本ロボット学会誌 Vol.20, No.8, pp.835–843 (2002).

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，人工知能学会誌, Vol.15, No.2, pp.267–275 (2000).

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).

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

高橋 泰岳，浅田 稔：実ロボットによる行動学習のための状態空間の漸次的構成，日本ロボット学会誌 Vol.17, No.1, pp.118–124, 1999.

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

矢入 健久，堀 浩一，中須賀 真一: 複数行動結果を考慮した最尤推定に基づく状態一般化法，人工知能学会誌, Vol.16, No.1, pp.128–138 (2001).