3C1-2

模倣学習と強化学習の調和による効率的行動獲得

Efficient Acquisition of Behaviors by Harmonizing Reinforcement Learning with Imitation Learning

田渕 一真[*1] Kazuma Tabuchi
谷口 忠大[*1] Tadahiro Taniguchi
椹木 哲夫[*1] Tetsuo Sawaragi

[*1] 京都大学大学院工学研究科
Graduate School of Engineering, Kyoto University

This paper presents a composite machine learning architecture of imitation learning and reinforcement learning. Humans usually learn tasks through both imitation learning and reinforcement learning. After observing superiors, learners start practicing through trial and error. In this context, imitation learning and reinforcement learning seem harmonized as a smooth series of learning phases. From the viewpoint of machine learning usually requires many trials and errors in an agent's learning phase. However, imitating other people's way of performing the task can reduce the amount of time. Based on this idea, the composition of reinforcement learning and imitation learning is proposed as an integrated machine learning architecture. An additional reward system is introduced, which connects these learning algorithms more naturally.

1. はじめに

人間が技能を獲得するプロセスには様々な形態が考えられる．その中でも，”学ぶ（まなぶ）”という言葉と”まねぶ”が同源である事が示唆するように，模倣を通じて他者から学習することは，人間の技能の獲得において本質的役割を演じている．この模倣学習のプロセスを計算論的に理解する事は，自律適応的なロボットの設計論として重要なだけでなく，人間組織の技能伝承のプロセスを理解する上でも重要である．

自律ロボットの学習方法としての模倣学習研究を見ると，模倣とはしばしば教師あり学習のプロセスのみで定式化される．しかし，我々が他者の行動を真似て学習するプロセスとは，そのような受け身的な学習に終始しない．もし，模倣学習が師匠から弟子，弟子からその弟子への教師あり学習の連鎖で表現されるとすれば，その技能は次第に劣化していくのみである．この劣化プロセスは印刷機でコピーのコピーをとるといった作業に等しい．稲邑らはモーションキャプチャによって人間の動作から教師事例を生成し，HMM を通してその運動を記憶させるミメシスモデルを提案している[稲邑 2004]．この例を含め，多くの模倣学習の研究では模倣される他者を完全な教師と見なし，完全なコピーを取ることを模倣学習としている．しかし，実際の我々の学習プロセスを振り返ると，むしろ，模倣は他者の行動を参考にするための方法にすぎず，参考にした後に自ら試行錯誤を行うことでより洗練された行動へと変化させていくプロセスこそ模倣学習において本質的である．つまり，模倣学習は試行錯誤に基づく強化学習と融合して初めて一連の学習プロセスを形成すると考えられる．

本研究では，他のエージェントの行動を観察し模倣学習を行い，その後に試行錯誤に基づく強化学習により修正していくというプロセスを一つの学習器の上で実現する手法を提案する．さらに，学習者が模倣対象エージェントの行動観察結果を強化学習時に積極的に利用するための補助的な報酬を導入することによって，これら二つの異質な学習則を調和させる．これによって他者の行動を真似つつ，さらに自らの試行錯誤を通じて洗練させていく効率的な学習手法を検討する．

連絡先：
〒606-8253 京都市左京区吉田本町
京都大学大学院工学研究科 機械理工学専攻
機械システム創成学分野
田渕一真
Tel, Fax: 075-753-5044
E-mail: kazuma@102514.mbox.media.kyoto-u.ac.jp

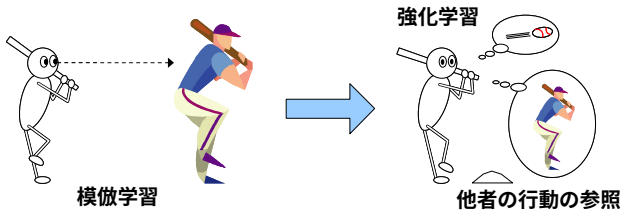

図 1. 模倣学習から強化学習へ

2. 模倣学習と強化学習の融合

本研究で検討するモデルでは，まず観察した他者の行動を真似ることで大まかに望ましい行動に近づき，次いで副報酬を伴った強化学習を行うことで，模倣によって得た方策を洗練していくという順で学習を行う．なお，学習器には線形近似器を用い，最急降下法によって学習する．

2.1 模倣学習

模倣学習はその名の通り他者の行動を真似た行動をとることができるようになるための学習であるが，その特徴は，模倣を行う場合には模倣する対象は理想的な行動を行っていると一時的に見なすことにある．模倣学習は模倣される他者がいて初めて成り立つ学習であるから，他のエージェントが居ない場合や，エージェントの集団にとって全く新規の行動を学ぶ時には用いることが出来ない．その点では適用できる状況が限られる学習法である．

模倣学習では他者の行動を観察し，それと同様の行動が出来るように行動形成を行う．しかし，通常，運動学習において学習主体が模倣対象となる他者の運動出力としての出力信号を直接計測することは不可能であり，外部の観察者としてその行動出力を観察できるのみである場合が多い．例えば，人間は模倣を行うときに，他の人間が脳から筋肉に出力した筋電位を計測したり，手が実際に出したトルクを計測したりすることはない．

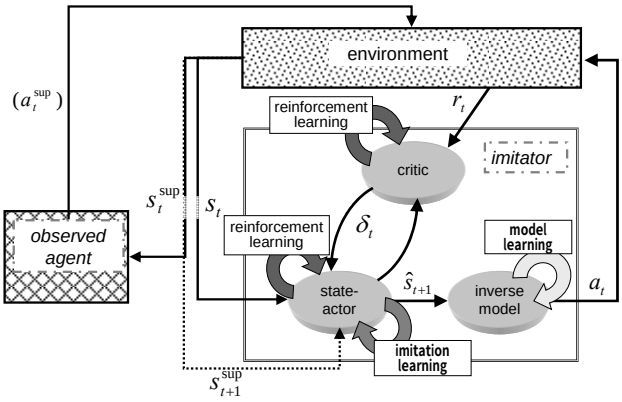

図2. 模倣と強化による統合学習器の概観

また, 仮に計測できたとしても, 自身の身体と他者の身体の相違から同じ運動指令を出力する事が同じ行動結果を生み出すとは限らない.

これに対して谷口らは身体の持つダイナミクスに依存しない行為概念表現としてセンサ空間上でのアトラクタを行動の表現単位として獲得することの有効性を動的環境下における強化学習の文脈から提案し, これを汎化行為概念と呼んでいる[谷口 2006]. また, 中西らは同様に行為を表象する状態空間上のアトラクタを運動プリミティブと呼び, その獲得についての研究を行っている[中西 2004]. 本研究では, この方式をとる事で学習者と被模倣者の身体の違いを吸収し模倣学習をモデル化する. エージェントは他者の行動の結果としての軌跡を観察することで, エージェント内部にセンサ空間上のアトラクタの一部を獲得し, 保持する. 行動時にはこれを駆動し, 目標となるセンサ入力値を逐次, 逆モデルに入力し, 対応する運動出力を出力させることで行動する. 逆モデルはモデル学習を行うことにより獲得することが可能であり, 事前の設計を要さない.

本稿では, タスクと無関係に学習が可能な逆モデル (inverse model) をあらかじめ学習しておいた後, ある状態に対して目標とするセンサ入力値と模倣する対象のデータを逐次比較することによって軌道決定モジュール(state-actor)の学習を行い, 模倣行動を獲得する. すなわち, 現在の状態を $\boldsymbol{s}_t$, 観測した模倣対象の行動後の状態を $\boldsymbol{s}_{t+1}^{\sup}$, 軌道決定モジュールの持つ線型近似器の重みを θ_i, 基底関数を $f_i(\boldsymbol{s}_t)$ とすると, 目標とするセンサ入力値 $\hat{\boldsymbol{s}}_{t+1}$ および 1 回の学習における重みの更新値 $\Delta\theta_i$ は,

$$\hat{\boldsymbol{s}}_{t+1} = \sum_i \theta_i f_i(\boldsymbol{s}_t) \tag{1}$$

$$\Delta\theta_i \propto (\boldsymbol{s}_{t+1}^{\sup} - \hat{\boldsymbol{s}}_{t+1}) f_i(\boldsymbol{s}_t) \tag{2}$$

と表される.

このような特徴を持つ模倣学習は, 直接的で高速な学習を可能とするが, この方法はいくつかの欠点を持つ. 一つは, 模倣する対象が望ましい状態に素早く到達するような理想的な行動をとっているとは限らないこと. もう一つは仮に理想的な行動をとっていたとしても, 観察できる回数は少なく, エージェント自身が取り得る状態全てについての観察事例を獲得する事は叶わず, 全状態空間中で観察事例が粗にしか存在しないということ. ゆえに, 観察模倣学習自体は完全であったとしても, それによって模倣したエージェントが常に理想的な行動を行い, タスクを達成出来るようになる保証はない. 以上より, 模倣学習のみで可能な学習には限界がある. よって, これを乗り越えるためにはエージェントが自ら試行錯誤を通じて追加の学習を行う必要がある.

2.2 強化学習

上記のような模倣学習の欠点を補う方法のひとつとして, 他者の存在に依存しない自律的で試行錯誤的な強化学習を用いることが考えられる. 強化学習は期待累積報酬の最大化を目指した目的指向的学習で, かつ経験した状態行動対のみをオンラインで評価する経験ベースの学習であるため, タスクに習熟した他者の存在を前提とせず, 自らの行う試行錯誤のみを通じて意味のある行動を獲得することができるという特徴を持つ. しかし, 学習の結果として獲得される挙動に関する規範が陽に与えられる訳ではないので, 環境全体にわたってしらみつぶしに分析する必要がある. そのため, 複雑で多次元な問題における学習では学習速度が急激に減少してしまう欠点がある[Sutton 1998]. 本研究では, 模倣学習の後に強化学習を行うことによって, 模倣学習の持つ欠点を補う事を目指すが, これは強化学習の欠点を補うものでもある.

本稿では強化学習の手法としては, 方策(actor)と価値関数(critic)を独立に持つ Actor-Critic 学習法[木村 2000]を用いる. Actor-Critic 学習法を採用することで一つの学習器で模倣による教師あり学習と強化学習両方を実現させる事が出来る. ただし, 本モデルにおける actor の出力は(1)式で表される目標とするセンサ入力値 $\hat{\boldsymbol{s}}_{t+1}$ とする. 本稿ではこのように次状態を出力としてだす actor を state-actor とよぶ事にする. 環境への出力は逆モデルを通して行われる. すなわち, 現在の状態 $\boldsymbol{s}_t$ について, 報酬を r_t, 割引率をγ, critic の持つ線型近似器の重みを w_i, 基底関数を $\phi_i(\boldsymbol{s}_t)$ とすると, $\boldsymbol{s}_t$ における価値 $V(\boldsymbol{s}_t)$, TD 誤差 δ_t および重みの更新値 Δw_i は,

$$V(\boldsymbol{s}_t) = \sum_i w_i \phi_i(\boldsymbol{s}_t) \tag{3}$$

$$\delta_t = r_{t+1} + \gamma V(\boldsymbol{s}_{t+1}) - V(\boldsymbol{s}_t) \tag{4}$$

$$\Delta w_i \propto \delta_t \phi_i(\boldsymbol{s}_t) \tag{5}$$

と表され, 強化学習時において, 現在の状態が $\boldsymbol{s}_t$ のときの運動出力後の次状態を $\boldsymbol{s}_{t+1}$ とおくと, state-actor の重みの変化量 $\Delta\theta_i$ は,

$$\Delta\theta_i \propto \delta_t(\boldsymbol{s}_{t+1} - \hat{\boldsymbol{s}}_{t+1}) f_i(\boldsymbol{s}_t) \tag{6}$$

と表される.

2.3 副報酬の導入

上のように教師あり学習をベースに, 強化学習を行わせた例としては村田らの人間の歩行データを元にした二足歩行の強化学習による獲得などがある[村田 2005]. しかし, 模倣によって獲得できる行動は状態空間において局所的なものであり, 状態空間全域を網羅したものではない. そのため, 模倣学習を行い, その後で強化学習を別に行う, といったような単純な組み合わせでは, その学習内容を強化学習時には十分に利用できない可能性がある. そこで, 強化学習時において他者の行動の観察データをより積極的に利用するための手段として観察経験に基づいた副報酬の概念の導入する.

副報酬導入の問題点として, 副報酬を加えて学習した場合と加えずに学習した場合において報酬が最大となる状態が異なる可能性があることが挙げられる. 副報酬を導入したことでタスクの課題自体が変化し, 異なる解に収束することとなる. ゆえに, 副報酬は報酬の最大化に関して整合性を持たせた形で導入しなければならない. この問題に対して Ng, Harada, Russell は, 状態のみに依存する関数 $\Phi(s_t)$ を用いる場合, $\gamma\Phi(\boldsymbol{s}_{t+1}) - \Phi(\boldsymbol{s}_t)$ を副報酬として主報酬と足しあわせる, つまり新たな報酬関数 $\bar{r}_t$ を

$$\bar{r}_{t+1} = r_{t+1} + \gamma\Phi(\boldsymbol{s}_{t+1}) - \Phi(\boldsymbol{s}_t) \quad (7)$$

として主報酬に代えて与えれば, 最適行動に影響を与えることなく収束速度を変化させることが可能であることを示した[Ng 1999]. ただし, (7)式に従うことによって最適行動が変化しないことは保証されるが, 学習が効率化するとは限らない. すなわち, $\Phi(s_t)$ の設計によって学習速度は高速化することも低速化することもあり得るものである.

本研究では, 模倣する対象を観察して得られたデータの量を用いて副報酬を作成する. 具体的には, 状態空間を分割し, その領域に属する観察データの数 $D(\boldsymbol{s}_t)$ と, 最も観察データの属する数が多い領域に属するデータ数 $\max D(\boldsymbol{s})$ との比から得られる

$$\Phi(\boldsymbol{s}_t) = \zeta \frac{D(\boldsymbol{s}_t)}{\max D(\boldsymbol{s})} + \xi \quad (8)$$

(ξ, ζは任意の定数)を用いて副報酬を生成する. この副報酬は, 観察された模倣によって学習した行動の系列へ引き込まれるような行動が強化学習によって生まれることを意図したものである. この引き込みによって, 主報酬のもとでの最適行動を変化させることなく, 強化学習における探索対象となる領域について模倣対象が与えられていない領域から与えられている領域への遷移を促す. なお, この副報酬は外部から与えられるのではなく, エージェント内部で生成されるものである.

2.4 学習機構と手順

以上から, 本研究で検討する学習機構は, ある状態に対して目標とする次状態のセンサ入力値を出力する state-actor, 現在の状態と目標とする次状態から行動出力を決定する inverse model, Actor-Critic 学習法における critic の 3 つのモジュールで構成される. そして学習は,

1. タスクに関しての学習を始める前に, 逆モデルについて学習を行う.
2. 観測した模倣対象に関するデータと state-actor の出力を比較することで模倣学習を行う.
3. 観測した模倣対象に関するデータの状態空間における分布から $\Phi(s_t)$ を求め, 副報酬を伴った強化学習を行う.

という手順で行う.

3. 実験

上記のモデルを, シミュレーション上における倒立振子の振り上げタスクに適用し, その有効性を検証した.

3.1 実験内容と評価方法

(1) タスク環境と報酬設計

本実験でタスクとする倒立振子は, 移動方向が水平線上に拘束された台車に一様断面の物理振子を取り付け, 水平方向の力を加えることで振子を振り上げる方式のものである.

このモデルにおける運動方程式は, 振子の長さを ℓ, 台車および振子の質量, 速度減衰係数をそれぞれ M, m, c_s, c_p, 重力加速度を g とすると以下のように表される.

$$(M+m)\ddot{x} + \frac{1}{2}m\ell\ddot{\theta}\cos\theta - \dot{\theta}^2\sin\theta + (c_s + c_p)\dot{x} + \frac{1}{2}c_p\ell\dot{x}\cos\theta = u$$

$$\frac{3}{2}m\ell\ddot{\theta} + m\ddot{x}\cos\theta + mg\sin\theta + c_p\dot{x}\cos\theta + \frac{1}{2}c_p\ell\dot{\theta} = 0.$$

エージェントが試行を開始するときには台車の位置 x, 速度 $\dot{x}$, 振子の角度 θ, 角速度 $\dot{\theta}$ の初期値を無作為に与えた. また一定時間試行を行うと, 再び無作為な初期値を与えてそこから学習させるようにした.

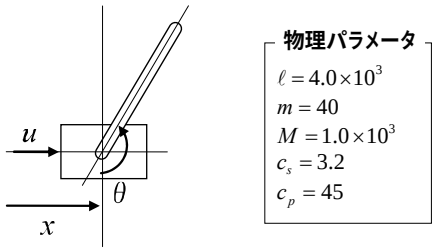

図 3. タスクモデル

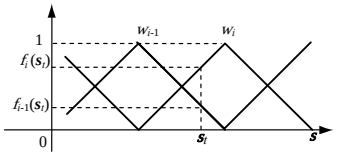

図 4. 基底関数

表 1. 基底配置, 状態分割のパラメータ設計

変数	最小値	最大値	基底数	状態分割
x	-2.0×10^6	2.0×10^6	31	15
$\dot{x}$	-5.0×10^5	5.0×10^5	31	9
θ	-2π	2π	15	7
$\dot{\theta}$	-1.0×10^2	1.0×10^2	31	7

報酬関数は原点において振り上げた状態で最大となるよう

$$r_t = \alpha[\exp\{-\beta(\theta_t - \pi)^2\} + \exp\{-\beta(\theta_t + \pi)^2\} - \cos\theta_t] + \varepsilon\exp(-\phi x_t^2)$$

($\alpha, \beta, \varepsilon, \phi$は設計者の決定する任意の定数)と与えた.

(2) エージェントの設計

エージェントの測定できる変数は $x, \dot{x}, \theta, \dot{\theta}$ とした. 基底関数には図 3 のような三角形状のものを用いた. また, 基底の配置は倒立振子の振り上げが可能だと考えられる状態空間の範囲と基底数を設計者が選び, 各変数に関して等間隔に配置した.

(3) 模倣される他者

模倣対象のデータを得るために, 本実験を行う前に Actor-Critic 学習法を用いてある程度学習させたエージェントを準備する. このエージェントは決して最適な行動を取ることの出来るエージェントではない. これに振り上げを実行させ, その振り上げの状態遷移の過程から模倣を行うエージェントの行動決定時間周期と同じ間隔で, 模倣するエージェントに $x, \dot{x}, \theta, \dot{\theta}$ をサンプリングさせ, 模倣学習の為の観測事例として獲得させた. また, 基底配置を行った状態空間の範囲と同様の範囲を各変数に関して等分割し, 分割された各範囲について観察データから $D(\boldsymbol{s}_t)$ を求めた.

(4) 評価方法

あらかじめ逆モデルの学習, 模倣学習を行わせたのち, 強化学習を行い, 一定時間ごとの獲得報酬をサンプリングした. また比較対象として, $\Phi(s_t) = 0$ としたこと以外は上記のモデルと同じ条件とした模倣学習は行うが副報酬は利用しないエージェントと, Actor-Critic 学習法のみで学習するエージェントについて同様の実験を行った. なお, 強化学習と模倣学習は目的の異なる学習方法であり比較できないため, 強化学習時のデータのみをサンプリングし, 比較した.

3.2 結果と考察

α= 0.5, β= 50, ε= 0.3, ϕ= 2.5×10^{-5}, γ= 0.9, ζ= 0.5, ξ= -0.05, 線形近似器の学習率を 0.1, 力学モデルの更新幅を 0.01, その

他のパラメータを図 2, 表 1 のように設計し, 各モデルについて 4 回ずつ行った実験結果の平均が図 4 である.

この結果より, 模倣学習を行ったエージェントの学習速度は, 行っていないエージェントに比べ高速な立ち上がりを見せた. また副報酬を用いていないエージェントのグラフは, 本研究のモデルを用いたエージェントのものと比べ波形の上下が激しかった. これは, 副報酬を導入していない場合, エージェントは模倣が可能な状態に移行する･模倣状態を維持するための機構を持っていないため, 模倣対象を真似る行動をとった場合には高い報酬が得られるものの, 模倣状態に素早く到達･維持できるとは限らないため学習が不安定化するという現象が起きている可能性が示唆された. さらに, 本研究のモデルを用いたエージェントのグラフの波形は, Actor-Critic 学習法のみで学習したものと比べて最終的な学習結果が安定しなかった. これは, 模倣対象のエージェントの行動が Actor-Critic 学習法により獲得されたものであるために, このエージェントの行動が最適方策に基づいたものであるとは言い難く, Ng, Harada, Russell らの定理が最適方策の不変性は示しているものの, この未熟な模倣対象への依存が主報酬の最大化を阻害したのではないかと考えられる. しかし, 学習者の模倣対象への依存性や, 副報酬の設計の学習プロセスへの影響についてはさらなる研究が必要である.

4. おわりに

本研究では, 機械学習の文脈で, これまでは独立に考えられがちであった強化学習と模倣学習が抱える問題について, 両者を融合させることによって相補的に改善する学習手法を検討した. 未だ検討の余地は多分に残されているものの, 模倣学習を行動学習に役立てる際に最適方策の学習効率の向上が可能であること, 副報酬を導入することで模倣学習と強化学習とを調和させながら接続することが可能であることを簡単な例題によって検証した. しかし, 今回提案した手法でも, 模倣される対象が必ずタスクに関して習熟していることを前提し, それを盲目的に学習者が真似るという従来からの模倣学習が持つ構造は変化していない. これに対し, 人間の学習時の行動を考えると, 模倣する価値のある対象を選び, 選択的に模倣するといった一段メタな学習機構が存在すると考えられる. このようなメタな模倣学習の機構が, 人間の模倣学習を通した組織内での技能伝播といった社会的な現象にも強く結びついていると考えられる. 人間の模倣学習の構成論的研究という視点からすれば, 今後はそのようなメタな選択も含めた模倣学習のモデル化を目指していきたいと筆者らは考えている.

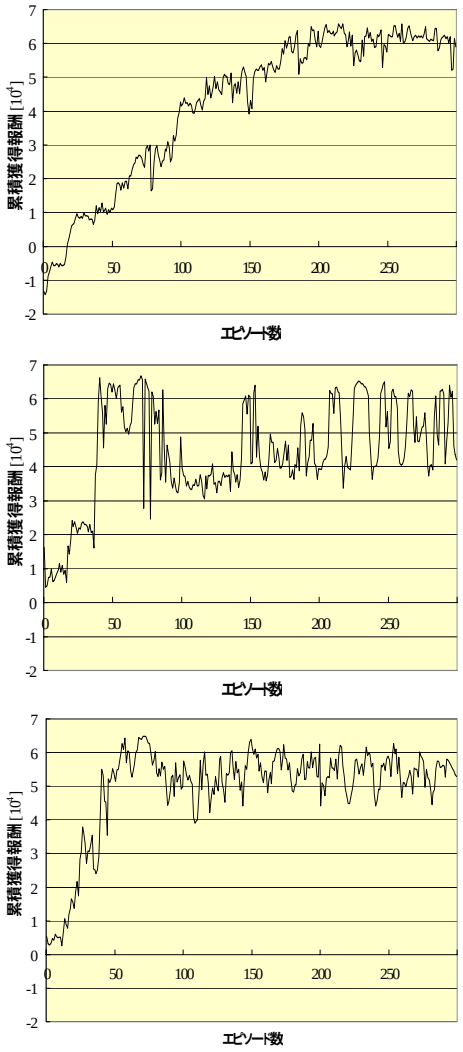

図 5. Actor-Critic 学習法のみで学習したモデル(上), $\Phi(s_t)=0$ としたモデル(中), 提案するモデル(下)の獲得報酬の推移

参考文献

[稲邑 2004] 稲邑哲也, 他:“ミメシス理論に基づく見まね学習とシンボル創発の統合モデル”, 日本ロボット学会誌, vol.22, no.2, pp.256—263,2004.

[谷口 2006] 谷口忠大, 椹木哲夫: “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―”, 計測制御学会論文集, vol.44, no.3, pp.255—264, 2006.

[中西 2004] 中西淳, 他, “運動学習プリミティブを用いたロボットの見まね学習”, 日本ロボット学会誌, vol.22, no.2. pp. 165—170, 2004.

[Sutton 1998] R. Sutton, A.G. Barto, “Reinforcement Learning : An Introduction”, The MIT Press, 1998.

[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”, 人工知能学会誌, vol. 15, no. 2, pp.267--275, 2000,

[Ng 1999] A.Y.Ng, D.Harada and S.Russell: "Policy invariance under reward transformations:Theory and application to reward shaping", In Proceedings of the Sixteenth International Conference on Machine Learning, 1999.

[村田 2005] 村田栄理, 他, “強化学習による多自由度 2 足歩行ロボットの制御“, 第 19 回日本人工知能学会全国大会予稿集, in CD-ROM, 2004.