仮想個体と実ロボットとの並列環境強化学習システムにおける諸問題

山口 智浩（大阪大学大学院　基礎工学研究科）
tomo@sys.es.osaka-u.ac.jp　　http://w3.sys.es.osaka-u.ac.jp/tomo/

要旨：実ロボット学習における困難な問題として，学習速度と環境変動への適応の両立がある．実ロボットベースの学習は，多大な時間を要するため，大半の学習システムは，実ロボットの動作する実験環境を模した仮想環境を計算機上に構築し，計算機上でのオフライン学習結果をもとに，実ロボットの行動を獲得してきた．しかしながらオフライン学習では，シミュレーションの高速化によって学習時間は短縮できても，仮想環境と実験環境との差異がそのまま実ロボットでのタスク実行の精度の悪化に反映され，しかも実験環境での環境変動へのオンライン的対応は，原理的に不可能である．

すなわち，学習速度と環境変動への適応の両方を改善するには，実ロボット学習自体の高速化が要求される．そこでこれを実現する手法として，まず両環境でシーケンシャルに強化学習を行う準備洗練強化学習法を紹介し， 次に両環境で仮想個体と実ロボットとが並列に強化学習しながら，互いの学習結果を選択的に伝播させることで，両者の平均化を避けつつ，実ロボットの学習速度を仮想個体並に高速化する実時間適応型模倣学習システムの概要について述べる．

強化学習法は，環境の静的な性質（ex. マルコフ性）を仮定すれば，学習の収束性が数学的に保証されるが，環境変動を伴う場合には，時間遅れを伴いながら変動に対して報酬を確率的に最大化するように追随する風見鳥的モデルとみなすことができる．しかしながら並列環境の場合には，実ロボット環境での変動によって生じる，仮想環境とのずれを如何に補正するかといった問題が生じる。これを解決するため、環境のモデルとして遷移観測マルコフ決定性（TOMDP）を定義し、観測不能の状態遷移によって生じる観測状態での非マルコフ決定性を、観測した状態遷移確率の変動から間接的に推定する方法を提案する。

尚、後半の発表内容（TOMDP,Change-Point Problem [３]）については、OHP資料を参照されたい。

１　研究の背景

ロボットと生物との大きな違いに様々な環境への適応や学習能力の有無がある．従来の産業用ロボットは，工場等の既知環境で特定の作業を効率良く繰り返し実行することを目的として設計されてきた．しかしながら，未知環境，あるいは想定外の状況にロボットが直面しうる場合，例えば惑星の探査ロボットや，社会や家庭等の身近ではあるが千差万別な環境において，用いられる環境や起こりうる状況があらかじめ特定できないため，あらかじめプログラムした行動だけでは，あらゆる状況に対応でき，安定にタスクを遂行する保証はない．したがって，ロボットが環境に適応し，直面した状況に応じて必要な知識をその都度オンラインで学習する，実ロボット学習の能力が将来必要となる．そういった背景から著者は，これまで実ロボットの学習機能を実現するため，強化学習法を中心として，模倣学習 [１] やバイアス知識を用いた学習の高速化および，仮想環境での学習結果を実環境での先見知識として用い，実ロボットでの学習回数の削減を行う準備洗練学習法 [２] 等を提案したが，その原理的限界はロボットの動作速度にあった．

ロボットの動作速度よりも速く学習や試行錯誤を行なうには，どうすれば可能か？実ロボット学習の最大の問題点は，学習時間の長さである．ロボット学習は，大きく分けて，計算機上でのシミュレーション（仮想環境と呼ぶ）での学習結果を実ロボットで用いるオフライン学習と，実験環境中の実ロボットを直接学習に用いるオンライン学習の２つに分類できる．計算機上でのオフライン学習は，人間に例えると，頭の中でのイメージトレーニングに相当し，実ロボットを直接用いる場合と比べるとはるかに高速に学習できるが，仮想環境では表現できない実環境との差異がタスク達成の誤差に反映され，しかも学習後に発生した実環境での新たな状況には原理的に対応できないという欠点がある．したがって学習後の状況変化に対応するには，実ロボットでのオンライン学習を続ける必要がある．しかしながら，実ロボットではロボット自身の機械的な動作速度が力学的に制約されているのに対し，仮想環境では，シミュレーション精度を選べば自由に学習速度を設計でき，しかも用いる計算機の計算能力が年々向上しているため，両者の１行動に要する時間コストは数１０~数１００倍の差がある．

そこでこの限界を超えるアイデアとして，実環境と仮想環境との行動（学習）速度差に着目する．例えば実ロボットの１行動に要する時間コストが仮想個体の１００倍だとすると，実ロボットが１行動する動作時間中に，仮想個体では１００回の行動が仮想環境上で実行可能である．したがって両環境での行動速度差を利用し，実ロボットが１回行動学習する間に空き時間を利用して，仮想個体での高速学習（メンタルシミュレーション）を繰り返し行う，学習の２重ループをもつ並列環境強化学習システムを提案する．

2　仮想個体と実ロボットとの並列環境強化学習システム

2.1　仮想個体と実ロボット間での適応型模倣による実時間準備洗練強化学習法

並列環境強化学習システムにおいて仮想個体と実ロボット間で学習結果のやり取りと行いながら，実ロボットの学習を高速化する手法を，実時間準備洗練強化学習法（Rt-SARA）と呼び，以下に手順を示す．

１）感覚データ入力：実ロボットが得たセンサー情報を学習システム（仮想個体）に送る．

２）実ロボットでのオンライン学習：実ロボットでの強化学習を１回行う．

３）メンタルシミュレーションによるオフライン学習：送られたセンサー情報を元に，仮想個体の状態を実ロボットの状態に近くなるよう初期化した後，動作の空き時間を利用して，ロボットが行動可能となるまで強化学習を繰り返し行い，最適な行動を試行錯誤から学習する．

４）仮想個体と実ロボットとの適応型模倣学習による学習結果の伝播：仮想個体の学習結果と実ロボットのもつ学習結果を比較し，パフォーマンスのよい方を選んで，実ロボットの行動として実行し，併せて学習結果を伝播する．

2.2　実時間洗練強化学習法による過度の経験強化の抑制

準備洗練強化学習法では，仮想環境での過度の経験強化が，実ロボットでの洗練学習を困難にする，という問題があった．これに対し，本手法では経験の回数自体を減少させて過度の経験強化を抑制し，しかも高速に学習を収束させることが可能である．

核となるアイデアは，適応型模倣 [１] による仮想個体から実ロボットへの学習結果の逐次的伝播である．ここで適応型模倣とは，模倣する相手と自己のパフォーマンスを比較し，相手の方が良ければ，その差に比例して確率的に相手の学習結果から必要な行動を自己の行動ルールとして模倣する手法である．但し模倣により伝播されるのは，行動ルールのうち強化値を除いた状況-行為ルール部のみである．模倣した行動ルールの強化値は初期値とし，その環境で実行された結果によって強化されていく．

まず，仮想個体と実ロボットの強化学習を同時に開始する．仮想個体は，実ロボットの１００倍の速さで学習するため，初期学習時において，仮想個体の方が確率的によい行動を学習する．よって実ロボットが直面した状況に対して，仮想個体が既に良い行動を学習していれば，その行動を実ロボットが選択的に模倣して実環境で実行し評価することにより，シミュレーション並の速度で実ロボット実環境下での高速学習が可能となる．模倣は両個体のパフォーマンスの差に比例して確率的に行うので，両者の差が小さい場合や実ロボットの方がパフォーマンスが良い場合には，実ロボットは仮想個体とは独立に強化学習し，実環境での洗練学習を行う．本手法の特徴は，仮想個体の学習結果を実ロボットの行動選択のガイドラインとして利用し，模倣したルールの強化学習は実環境で行うことである．適応型模倣により，学習結果の伝播を必要な部分に限定して選択的，逐次的に行い，ルールの強化値は伝播しないため，仮想環境の学習結果への過度の特殊化を実環境で起こすことなく，見かけ上，実時間での実ロボットでの強化学習が実現できる．

2.3　並列環境での洗練強化学習法の利点と問題点

従来のシーケンシャルな準備洗練学習手法 [２] では，学習回数の大半を高速なオフライン学習で実行することで全体の学習時間を短縮したが，全体での学習回数を減らすことは困難であり，しかも実環境が変化した場合には仮想環境での学習結果が無駄，あるいは逆効果になる，という問題点があった．これに対して提案する本手法は，オフライン / オンライン学習の並列実行と，両者でパフォーマンスの良い方の学習結果を伝播させる適応型模倣学習手法との組合せにより，実ロボットでの試行錯誤回数そのものを減少させることができ，実ロボットが，見掛け上シミュレーション並みの速さで学習することが可能となる．しかしながら，並列環境において両者の状態の同期 / 同一化をどうとるか，という問題が発生する．現状では，実ロボットの１行動ごとに実ロボットのセンサー情報に基づいて，仮想環境での過去のセンサー情報と状態との履歴から仮想環境を初期化することによって，仮想環境での学習を実環境に常に近づける解決手法を検討している．

謝辞　Change-Point Problem [3] に関して教示頂き、素朴なアイデアに適切なコメントをして下さった、大阪大学大学院基礎工学研究科数理科学分野統計数理講座の安芸重雄助教授に感謝いたします。

参考文献

[１] 山口，三浦，谷内田：適応型模倣による複数個体の強化学習，人工知能学会誌，Vol.12, No.2, pp.323-331,1997.3

[２] 山口，増渕，田中，谷内田：経験型強化学習における仮想個体から実ロボットへの学習行動の伝播，人工知能学会誌，Vol.12, No.4, pp.570-581, 1997

[３] Prakasa Rao, B.L.S., Asymptotic Theory of Statistical Inference, pp.275-289, John Wiley & Sons Inc., 1987