強化学習による Tray Tilting 問題へのアプローチ
— DYNA-Q アルゴリズムの適用と考察 —

○三輪憲治，三上貞芳，嘉数侑昇（北大）

制御問題の多くは，制御対象が反応を起こすまでに時間遅れが存在し，制御対象の性質を正確に把握することは難しい．その様な問題の例として Tray Tilting 問題を取り上げ，目的を達成させる行動を学習によって獲得することを考える．本研究では，強化学習法の１つである DYNA-Q アルゴリズムを適用させた場合の特徴と問題点について検証し，考察を行う．

１．はじめに

制御問題を強化学習で取り扱う場合，制御対象は様々な動きを取ることが考えられる為，状態を少なく分割した場合，１つの連続した行動列の中に多くのボールの動きを集約することになり，少ない状態の中から選択した行動が最適な行動なのかが問題になる．また状態を細かく分類した場合，ボールの動きを正確に把握できるが，探索空間が広くなり，その広い探索空間を探索しきれるのかが問題になる．

本研究で取り扱う強化学習法であるDYNA-Qアルゴリズムは，強化学習法の研究で多く用いられるQ-Learningにplanningを加え，広い探索空間での収束の遅さの改善を図っている．このDYNA-Qアルゴリズムを利用することで，制御対象を正確に検知するために状態を細かく分割し，それに伴い広い状態空間を探索することになっても，早い学習の収束が期待できる．

制御問題の例としてTray Tilting 問題を取り上げる．Tray Tilting 問題とは，ある平板上にボールを起き，ボールが平板から落ちないように平板を動かしてボールを制御する問題である．この問題の性質として，二次遅れを持った非線形システムであり，制御対象の取りうる行動が非常に多彩なため，制御対象を制御することが大変困難な問題であることである．以上から本研究での制御問題の例として興味深いものである．

本研究では，強化学習法によるTray Tilting 問題へのアプローチの一手法を示す．具体的には，Trayの領域，および行動を離散的に扱い，目的達成の為のTrayの行動を，強化学習法の１つであるDYNA-Qアルゴリズム[1][2]によって獲得していく手法を，計算機を用いた実験を用いて検証し，考察を行う．

２．DYNA-Q アルゴリズム

DYNA-Q アルゴリズムは Sutton によって提案された強化学習法である．その特徴は，Watkins によって提案された強化学習法である Q-Learning[3]の，広い状態空間での学習の収束に時間がかかる欠点を改善するために，agent が実際に経験した事柄を基に内部世界を構築し，それを用いて仮想的な学習を行う機構を組み込むことで学習速度向上を目指している点である．

Q-Learning は状態と行動の対に対する評価値（Q 値）を用いて行動を選択し，その結果によって Q 値を更新していく学習アルゴリズムである．

DYNA-Q アルゴリズムは，Q-Learning が実世界において行動選択と学習が終わった後，その経験の全て，またはその一部を用い内部世界を構築し，その中から planning 候補を選択し，仮想的な学習を行う．この planning は実世界において経験した事柄からのみ選択され，実際の状態変化を伴わないこと以外，実世界における学習となんら変わりはない．

３．提案手法

本研究では，学習者に対してボールがTray上のどの領域に存在しているかという情報しか与えられないという仮定の中で行われるものとする．そのため速度ベクトル・加速度ベクトルを使うことができず，Q-Learningの状態ベクトルとして $x_{t-2}, x_{t-1}, x_t, a_t$ を用い，Q値を，$Q_t(x_{t-2}, x_{t-1}, x_t, a_t)$ としてボールの速度方向・加速度方向を取り入れられる様にする．ボールがTrayから落下するまでを１試行とて，planningは１試行毎に行われるものとする．Trayの大きさをを500mm四方とし，行動はTrayの傾斜の付け方を，x軸回転の傾斜が0を境に± 2 段階の計5通り，y軸回転の傾斜が同じく5通りの合わせて25通り用意し，その選択はボルツマン分布に従って確率的に選択されるようにした．

４．計算機実験

４．１　概要

実験の初期状態はTrayが水平であるとし，学習率 αは0.2，割引率 γは0.9とした．Tray上にボールが乗ることを許す最長時間を120秒と設定した．Trayは16分割，64分割，121分割のものを用意した．評価に関しては，Trayの中心領域ほど高いrewardが与えられるものとした．64分割のTrayはTrayの周辺領域に負のrewardを配置したものと，そうでないものの2種類用意し，121分割のものは周辺領域に負のrewardを配置したもののみを用意した(Fig.1).

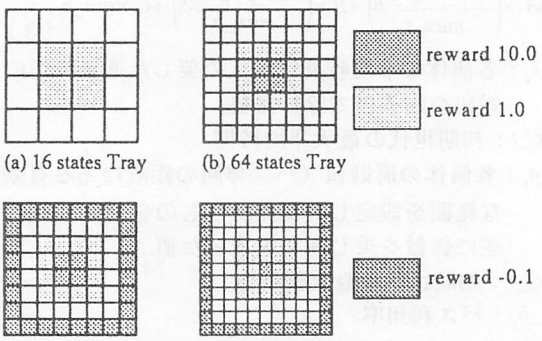

Fig.1 : Tray layout

４．２　実験

実験は次の6通りの実験(Table 1)を，planning回数を100に固定して，planningの参照する過去の系列の長さを換えて行う．グラフは横軸を実験回数，縦軸をボールがTray上に乗った時間とし，それぞれ10回の実験の平均を取っている．スタート領域固定の設定で行った実験に関しては，更に区間5の移動平均をとった．

Table 1 : Experiments

Experiment	Tray	Start ball point	Trial	Result
Exp.1	(a)	point fixed	250	Fig . 2
Exp.2	(b)	point fixed	500	Fig . 3
Exp.3	(c)	point fixed	500	Fig . 4
Exp.4	(d)	point fixed	500	Fig . 5
Exp.5	(a)	domain fixed	1000	Fig . 6
Exp.6	(d)	domain fixed	1000	Fig . 7

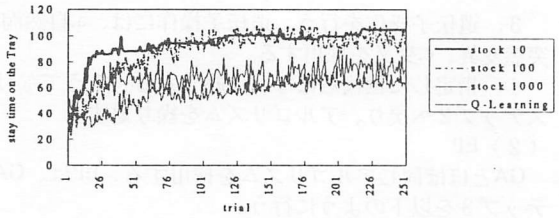

Fig . 2 : Result of Exp.1

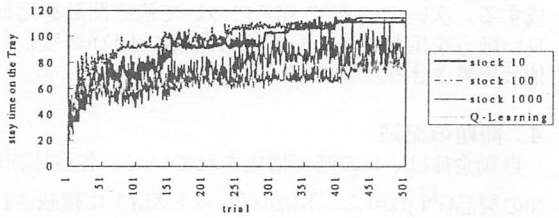

Fig . 3 : Result of Exp.2

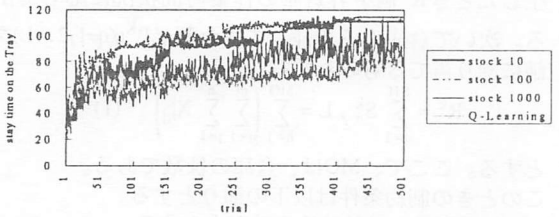

Fig . 4 : Result of Exp.3

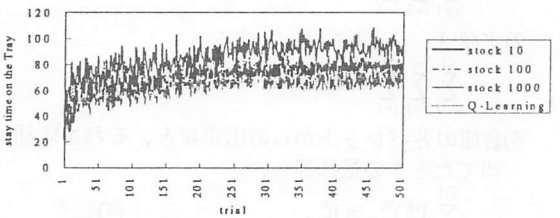

Fig . 5 : Result of Exp.4

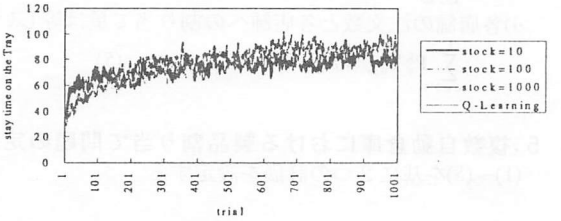

Fig . 6 : Result of Exp.5

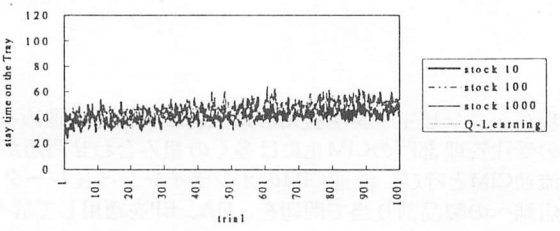

Fig . 7 : Result of Exp.6

4．3　考察

スタート位置を固定した条件で実験を行った場合，Q-Learning よりも DYNA-Q のほうが，Tray 上にボールが乗る時間が長くなり，またその値に早く収束している．これは DYNA-Q の特徴である planning の効果が出ていると思われる．また，16 分割の Tray の場合 stock 10 のものが最も良い結果を示したのに対し，64 分割では stock 100，121 分割では stock 1000 のものが最も良い結果を示した．更に 64 分割のものでは reward の与え方を変えた場合でも stock 100 のものが最も良い結果を示したことから，有効な内部世界を構成するための経験の蓄積量は状態空間の広さに依存すると考えられる．

状態分割の数に注目すると，分割数が増えると Q-Learning に対する DYNA-Q の優位さが減少している．これは分割数が多いと 1 つの状態が小さくなり，ボールの二次遅れを持った複雑な動きの影響により reward の享受が正確に行われなかったことと，Q 値の数が 16 分割のものは約 12 万個なのに対し，121 分割のものになると約 4500 万個にまで増大し，有効な行動の連続を決定することが困難になったことが考えられる．

スタート領域を固定した条件で実験を行った場合，当初状態分割数を多くした方がボールの動きをより正確に分類できるため，良い結果が得られると予想したが，結果は分割数が多い方が目的を達成することが出来ず，また DYNA-Q の有効性も確認できなかった．これは DYNA-Q の行う planning が実世界において経験した内容を使って行われるため，未来において planning した内容と同じ現象が起きなければ planning の効果を期待できないためと考えられる．

以上より，二次遅れを持った系に DYNA-Q アルゴリズムを適用した場合，状態分割の数と planning の参照する過去の系列の長さの設定が重要であると言える．

5．おわりに

本研究では，制御問題の例として Tray Tilting 問題を取り上げ，DYNA-Q アルゴリズムの適用とその有効性を検証した．

参考文献

(1) Richard S. Sutton : *Integrated Architectures for Learning, Planning , and Reacting Based on Approximating Dynamic Programming,* Proceedings of the Seventh International Conference on Machine Learning , pp.216-224 (1990)

(2) Jing Peng and Ronald J. Williams : *Efficient Learning and Planning Within the Dyna Framework,* Proceedings of the Second International Conference on Simulation of Adaptive Behavior, pp.281-290 (1992)

(3) Christopher J.C.H. Watkins, Peter Dyan : *Technical Note Q-Learning,* Machine Learning. 8.,Kluwer Academic Publishers Co., pp.279-292 (1992)