強化学習による倒立振子の安定化制御について

海 津　宏

A Study on Stabilization Control of Inverted Pendulum using Reinforcement Learning

Hiroshi KAIZU

It is important to study an automatic generation technique of the conventional hierarchical rules. The educational inverted pendulum is a Single-Input Multiple-Output system. In the SIMO system is difficult to construct by conventional controller which has considering multi-parameters by trial and error . This paper studies to discuss the stabilization control of inverted pendulum using reinforcement learning. Some results of computer simulation are shown to prove the effectiveness of this simple model.

Keywords: Reinforcement Learning, Inverted Pendulum, Stabilization Control, Computer Simulation

１．はじめに

人間は、成長に伴い知識や経験を獲得し、これらを柔軟に利用することによって、動的環境下でも適切な行動や判断を行うことができる。実際にはコンピュータを活用しつつ、このような知識や経験の獲得とその適用に関して柔軟で簡便な知的制御の実現が求められている。また、近年多様な自律型のロボットが登場しており、それらのロボットは様々な動的環境下での活動を可能とし、災害現場などその活躍の領域をさらに拡張しつつある。しかし、環境領域の拡張に伴い設計者が予知できなかった状態に陥る可能性も増加する。そのためには、状況に柔軟的対応が可能となりうる制御プログラムを構成するが求められているが、その実現はかなり難しく、機械学習による適応自動生成ルールなどの研究開発が有用である。さらに、初学者のための教育用制御システムの構築は、今後一層重要であると考えられている。比較的古くから用いられている倒立振子自動制御装置以外に、ケペラ（khepera）やBeauto Balancer（図１参照）、自律型移動体など知的制御のための実験教育に有効と思われるものも増加しているが、教育用教材としての課題はなお山積している。

以上より本稿では、強化学習を用いたレール型倒立振子システムの安定化制御に関してシミュレーションによる検証、考察と教育用教材としての適用等について報告する。

平成 22 年 10 月 4 日受理

図１　Bauto Balancer Duo [5)]

２．強化学習

強化学習では、学習者(以下エージェントと呼ぶ)をある環境に置き、外部からの情報を入力していない状態で自ら判断し行動させて、その結果を記憶して次回以降の行動に反映させていく学習制御である。特徴としては、最適な行動を人間がエージェントに直接教えるのでなく、エージェント自身が遷移した環境の状態を観測しつつ、試行錯誤を通して得られたスカラー値の報酬を評価して、その評価の度合いによって次の行動規範に反映させていく。即ち、エージェントは、行動→状態遷移→状態観測→評価を試行錯誤的に繰り返していく。成功した場合には報酬が、失敗した場合には罰が与えられる。これら一連の結果をもとに、指定した行動目的をエージェントが自動的に達成するために、過去に成功した行動の中から報酬を効果的に得るための方策を逐次学習して獲得することになる。ただし、強化学習で獲得した知識は、学習した環境に限定

されることが多くており、汎用性が高くない。そのため新しい環境に順次適応するためには、その環境に適した知識を再び獲得することが求められている。経験や知能化のモジュールを構築し、ロボットの知識として効率よく総合的に（または部分的に）組み合わせて実現することが広く求められている。

強化学習では、一般的な教師付学習と異なり、状態入力に対する正しい行動選択を評価するルールが無く、報酬として与えられる情報を手がかりに学習する。このとき実際上は、報酬にノイズや遅れなどの課題が存在する。従って、行動を実行した直後の報酬だけでは、エージェント自身はその行動が正しかったかどうかを判断できないという実問題を含んでいる。図２は、強化学習のアルゴリズムの概要を表している。

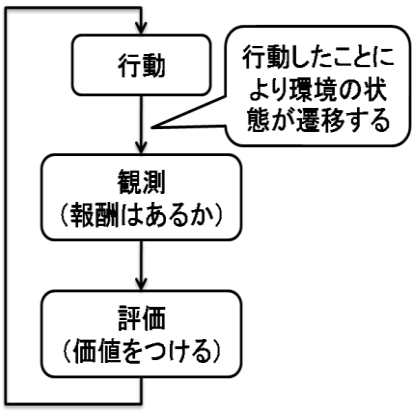

図２　強化学習のアルゴリズム

実問題では連続値の状態入力と同様、連続値の行動出力を求められることも多い。行動空間を離散化するのが普通だが、粗く離散化すると細やかな制御が出来ないという問題が生ずる。逆に、離散化が細かすぎると探索空間が増大し、通常の離散ＭＤＰにおける学習制御方法では、なかなか学習が進まなくなり非実用的となる。実際の応用例としは、携帯通信の周波数帯の動的割り当て、在庫管理・生産ラインの最適化　倒立振子の振り上げ安定化（階層化と actor-critic に基づく連続値行動の組み合わせによる）学習などの例が報告されている。

一般的に強化学習は、方策(policy)、報酬関数(reward function)、価値関数(value function)、環境のモデル(model)の４つで構成されている。強化学習問題においてマルコフ性を満たす環境は、マルコフ決定過程(MDP)と呼ばれ状態 s_{t+1} への遷移が、そのときの状態 s と行動 a のみ依存し、それ以前の状態や行動には関係しない。

MDP は環境のダイナミクスを次のようなモデル化をしたものである。環境のとりうる状態の集合を $\mathbf{S} = \{ s_1, s_2, \ldots, s_n \}$、エージェントがとりうる行動の集合を $\mathbf{A} = \{ a_1, a_2, \ldots, a_n \}$ とすれば、 環境下のある状態 $s \in \mathbf{S}$ において，エージェントのある行動 a を実行すると，環境は確率的に状態 $s' \in \mathbf{S}$ へ遷移する． その遷移確率を $\Pr\{ s_{t+1} = s' \mid s_t = s, a_t = a \} = P^a(s,s')$により表す。このとき環境からエージェントへ報酬 r が確率的に与えられるが、その期待値を $E\{ r_t \mid s_t = s, a_t = a, s_{t+1} = s' \} = R^a(s,s')$ と表現できる。

一方、Q 学習は状態と行動をセットにして、価値関数を振り分けることが出来る。Q 学習のアルゴリズムを次に示す。

step1　Q(s,a)を任意に初期化
step2　各 episode に対して繰り返し
　　1) s を初期化
step3　各 step に対して繰り返し
　　1)　エージェントは環境の時刻 t における状態 s_t を観測する。
　　2)エージェントは任意の行動選択方法に従って行動 a_t を実行する。
　　3)環境から報酬 r_t を受け取る。
　　4)状態遷移後の状態 s_{t+1} を観測する。
　　5)以下の更新式により Q 値を更新する。

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma \max_a Q(s_{t+1}, a)]$$

　　6)時間 t を t+1 へ進めて手順(3.1)に戻る。
　　7)s が終端状態なら手順(2.1)に戻る。

上式において、Q(s,a)は Q 値、s はエージェントが遭遇している環境の状態、a は状態 s においてエージェントが取りうる行動、r は報酬を、α は学習率($0 < \alpha \leqq 1$)、γ は割引率($0 \leqq \gamma < 1$)を示している。$\max_a Q(s_{t+1},a)$は、状態 s_{t+1} で取りうる行動の価値(Q 値)が最も高いものを表している。エージェントが価値関数をどのように行動選択に反映していくかを示すものが、行動価値手法であり、主に greedy 手法と ε -greedy 手法がある。後者の特徴は次の通りである。

ε -greedy 手法：greedy 手法において、小さい確率 ε で価値関数の値とは無関係に、エージェントが取りうる行動 a を一様に選択することで、さらに価値が高い行動を探索することが可能となる。

３．倒立振子とモデル

図３は、倒立振子システムの概念図である。直線のレール上をカートが移動することにより、カート上の振子を倒さずに中心座標点に鉛直にすることを目的としている。なお，レールと台車の間の摩擦と振子が取りつけられている駆動系の摩擦は，無視できるものとする。倒立振子は，代表的な教材であり、多重型、並列型など非線形で不安定な制御対象として。ソフトコンピューティングなど各種制御系設計法の有効性を検証するためによく用いられている。

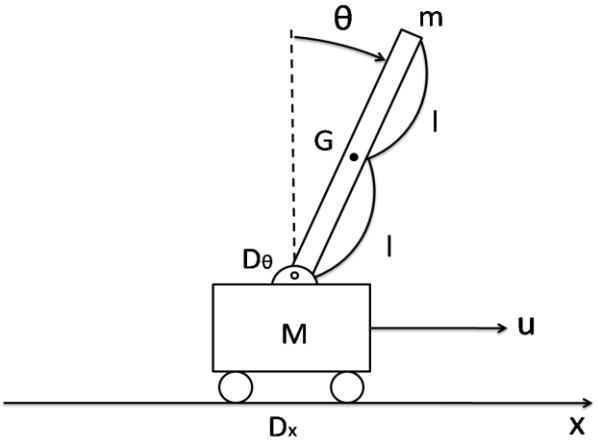

図３　倒立振子の概念図

m:棒の質量，2l:振子長，G:振子の重心，
M:カート質量，x:カートの変位（座標），
θ:振子の偏角，u:外力，
$\mathbf{D}_x$:x の粘性摩擦係数，D_θ:θの粘性摩擦係数

本モデルより、次式のラグランジュの運動方程式において運動、損失、ポテンシャルの各エネルギーの平衡条件より運動方程式が導出できる。ここで $\mathbf{q}(t)$は、一般化座標を表している。

$$\frac{d}{dt}\left(\frac{\partial J}{\partial \dot{q}_j}\right)-\frac{\partial J}{\partial q_j}+\frac{\partial D}{\partial \dot{q}_j}+\frac{\partial U}{\partial q_j}=u_j\ \ (i=1,\dots,p)$$

外力を u_x=u、u_θ=0 とすると、次の非線形運動方程式が導出される。

$$(M+m)\ddot{x}+ml\ddot{\theta}\cos\theta-ml\dot{\theta}^2\sin\theta+D_x\dot{x}=u$$

$$\frac{4}{3}ml^2\ddot{\theta}+ml\ddot{x}\cos\theta+D_\theta\dot{\theta}-mgl\sin\theta=0$$

状態変数を$x,\theta,\dot{x},\dot{\theta}$ とし、平衡点近傍で線形近似して展開すると次式が導出される。

$$(M+m)\ddot{x}+ml\ddot{\theta}+D_x\dot{x}=u$$

$$\frac{4}{3}ml^2\ddot{\theta}+ml\ddot{x}+D_\theta\dot{\theta}-mgl\theta=0$$

制御対象の状態方程式と出力方程式が導出できる。なお、非線形制御でのシミュレーションは次の教材用課題としても重要である。

$$\dot{x}(t)=\begin{bmatrix}0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 0 & -mgl/N & -4lD_x/3N & D_\theta/N\\ 0 & (M+m)g/N & D_x/N & -(M+m)\,D_\theta/Nml\end{bmatrix}x(t)+\begin{bmatrix}0\\0\\4l/3N\\-1/N\end{bmatrix}u(t)$$

ただし、N=(4M+m)l/3

$$y(t)=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\end{bmatrix}x(t)$$

４．シミュレーション

本稿では、Q 学習による倒立振子の安定化制御において、報酬 r および学習率α、割引率γを主な変数として４つの状態変数の挙動から教育システムとしての有効性を含め検証した。プログラミングには c 言語、刻み時間は 0.001 などとした。可視化など教育用展開としては今後改良する必要がある。制御対象の状態 s、各状態におけるエージェントが取りうる行動 a、各状態及び行動から与えられる報酬 r の設定が重要となる。振子モデル本体の設定は次の通りとする。

表１　倒立振子の物理定数

振子の長さ:2l[m]	1.0
振子の質量:m[kg]	0.1
カートの質量:M [kg]	1.0
x の粘性摩擦係数:D_x	1.0
θの粘性摩擦係数:D_θ	0.1
重力加速度:g[m/s^2]	9.8

次に、状態と行動について設計を行う。今回の制御では、$x, \theta, \dot{x}, \dot{\theta}$ のパラメータによって制御入力 u の値を変化させることにする。つまり、エージェントが遭遇する状態は $x, \theta, \dot{x}, \dot{\theta}$ のパラメータによって割り振られ、その状態の Q 値によりエージェントは行動(制御入力 u)を選択していく。

行動 u の大きさは 1.0[N]で固定とし、力の方向(+ or －)だけを制御するものとする。状態 $x, \theta, \dot{x}, \dot{\theta}$ は表２の通り、連続的な区間をある閾値で区分けを行う。報酬の設定例は表３の通りである。その他の設定は次の通りである。

1)エージェントが遭遇する状態は、計 648 通り。
2)エージェントが各状態で取りうる行動は、
　±1.0[N]の 2 通り。
3)x が±1.0[m]以上、または θ が π/10 [rad]を超えた場合は、制御失敗とする
　3.1)マイナス報酬 r=－5
　3.2)状態変数 $x, \theta, \dot{x}, \dot{\theta}$ を初期値に戻し、制御を再び開始する。

このような条件下で各パラメータ(報酬、学習率、割引率)を設定可変し、倒立振子の状態変数の挙動を検証する。

表２　状態変数の区間と分割区間数

x[m]	～－1.0～－0.2～0 ～0.2～1.0～	6 区間
θ[rad]	0～π/180～π/10～	3 区間
ẋ[m/s]	～－0.15～－0.08～0 ～0.08～0.15～	6 区間
θ̇[rad/s]	～－π/6～－π/30～0 ～π/30～π/6～	6 区間

表３　報酬 r の設定例

・｜x｜＜0.2, ・　θ＜π/180, ・｜dx/dt｜＜0.08, ・｜dθ/dt｜＜π/30	報酬 r ＝３
・｜x｜＞1.0 　or ・　θ＞π/10	罰 r ＝－３

５．結果

次表は、結果のグラフ例一覧である。

表４　結果例一覧

報 酬	episode	状態変数	図番号
無し	1	x、θ	4(a)
		dx/dt,　dθ/dt	4(b)
	100	x、θ	4(c)
		dx/dt,　dθ/dt	4(d)
有り	10	x、θ	5(a)
		dx/dt,　dθ/dt	5(b)
	100	x、θ	5(c)
		dx/dt,　dθ/dt	5(d)
学習率α	（γ=0.5　一定）		6(a)
割引率γ	（α=0.5　一定）		6(b)

図４(a)～(d)は、報酬を与えない場合であり、初期行動ｕ＝１[Ｎ]の影響を受け、step とともに発散（不安定）となる様子が確認された。当然ではあるが、step や episode が増加した場合でも、各状態変数が発散となるよう漸増し、安定化制御や学習効果は全く確認できない。一方、多少に関わらず成功報酬を与えた場合は、図５より step および episode の増加に伴い、安定制御への学習効果などが確認された。ただし、シミュレーション上からは、チャタリング現象のような微小雑音振動のような現象が最終定常特性時に生じることが多かった。さらに、報酬と罰の与え方や区間の細分化などの影響も確認することができ、強化学習による学習効果が簡単な本モデルでも検証された。次に、図６より安定化までの平均 step 数による比較検討した場合、学習率や割引率による明確な影響は明らかにはならなかった。適度なαとγの組合せ（α小、γ大）においては、学習効果が表れる場合もあるが設定条件等への一層の配慮、調整（報酬と罰の組合せ度合い等）が必要となっている。以上より、強化学習による学習効果は、線形化倒立振子システムの安定化制御において有効性が確認された。今後、可視化、非線形制御、外乱ロバスト特性、効果的な学習効果の拡張、およびファジィ制御・遺伝的アルゴリズムなどソフトコンピューティングとの最適化の連携が課題である。

なお、実習課題としては、両手足による振子倒立実演検証や多重・並列・カオスおよび車輪型振子などでの実証、Matlab/Simulink, Scilab/Scicos などとの設計ツール連携などが今後も有効に展開できる。

６．おわりに

制御工学演習、実機制御においても教育効果は重要な考察題目であり、学習効果を強化学習で展開することが一層求められている。振子モデルおよび報酬・罰の設定基準、エージェントの行動の拡張、さらに制御設計法の連携など改善する内容は少なくない。本手法を単独で展開する際には、状態変数の閾値設定、連続－離散空間の数値化など微調整や経験による知能化などの難しさも認められた。しかし、今後より一層初学者のための安価で柔軟簡便な教育用汎用制御システムが開発されていくことを大いに期待している。

参考文献

１）三上貞芳，皆川雅章　訳, Richard S.Sutton ，Andrew G.Barto 著：強化学習，森北出版, 2000

２）池田，斎藤，北村：多層ネットワークによる倒立振子の安定化学習制御，システム制御情報学会論文誌，vol. 3，no. 12，pp. 405. 413，1990

３）亀井，高木：ファジィクラスタリングを用いたファジィ ID3 と制御ルール獲得への応用，日本ファジィ学会誌，vol. 11，no. 1，pp132. 139，1999

４）木村元, 宮崎和光, 小林重信：強化学習システムの設計指針, 計測自動制御学会誌, vol. 38，no. 10，pp. 618-623，1999

５）ヴイストン株式会社, http://vstone.co.jp/top/access/index.html

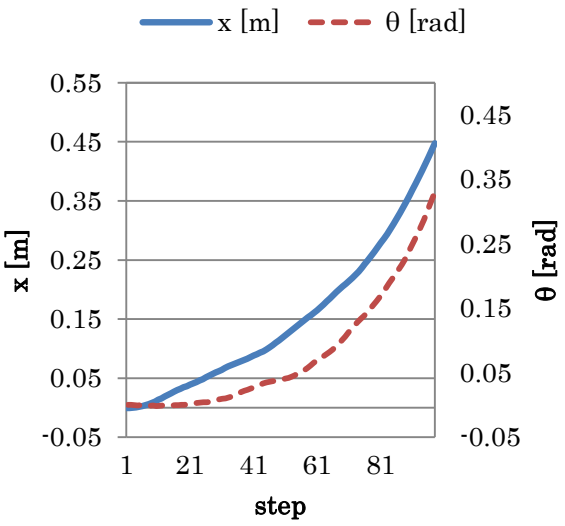

図４(a)報酬無し　x，θ (episode=1)

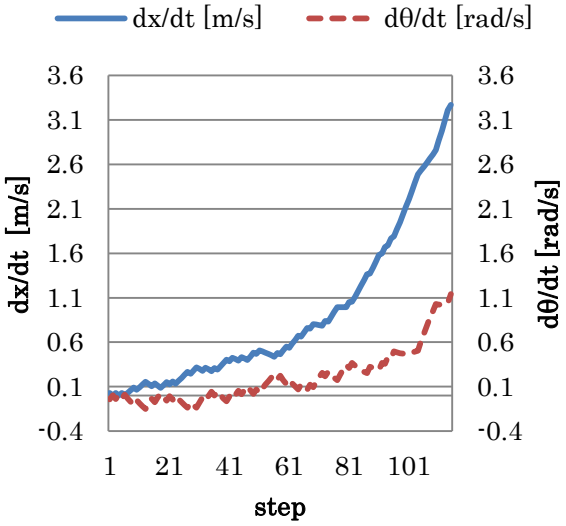

図４(b)報酬無し ｄx/dt，ｄθ/dt(episode=1)

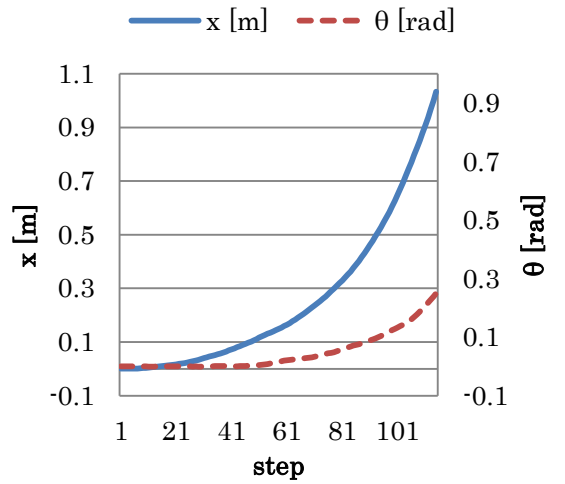

図４(c)報酬無し　x，θ (episode=100)

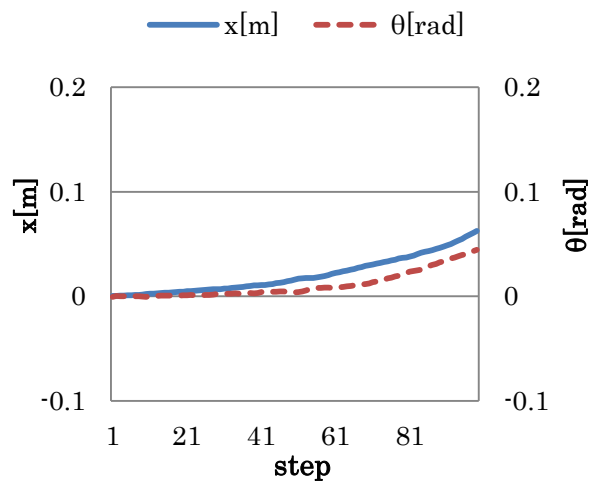

図４(d)報酬無し ｄx/dt，ｄθ/dt (episode=100)

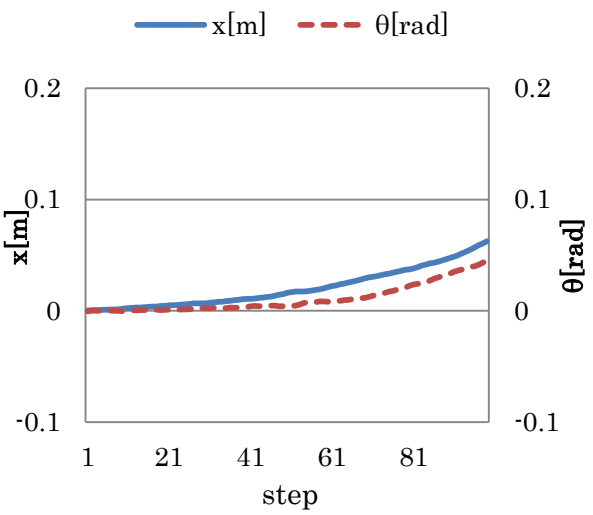

図５(a)報酬有り　x，θ(episode=10)

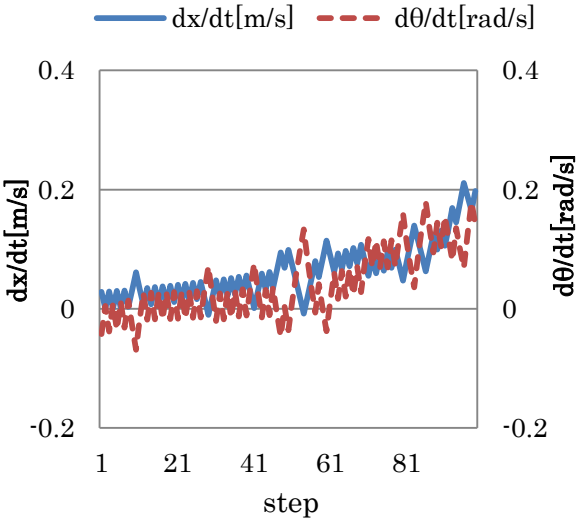

図５(b)報酬有り　dx/dt，dθ/dt(episode=10)

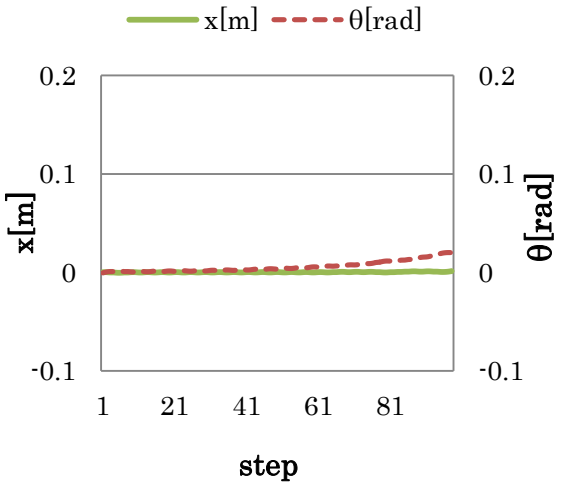

図５(c)報酬有り　x，θ(episode=100)

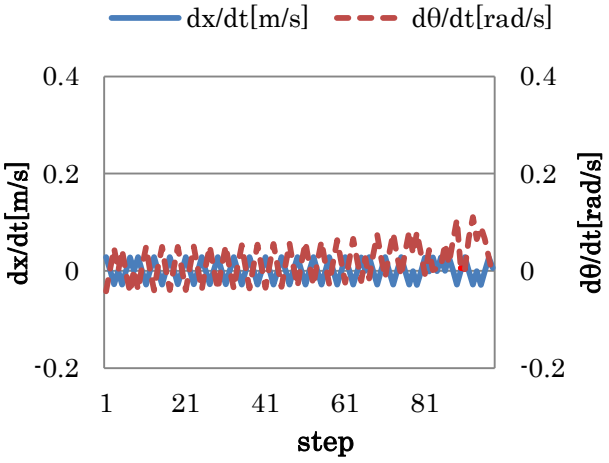

図５(d)報酬有り　dx/dt，dθ/dt(episode=100)

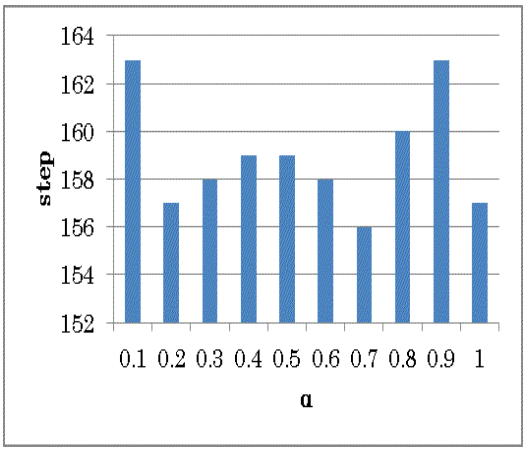

図６(a)　αによる特性（γ=0.5）

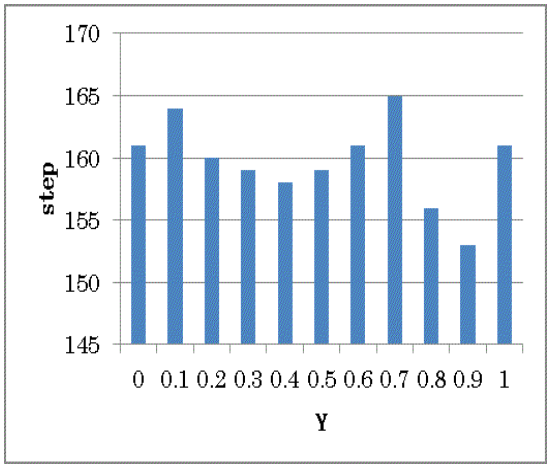

図６(b)　γによる特性（α=0.5）