# 強化学習を用いた Ball & Beam 実験装置の制御

2011SE041 日比野綾佳　2011SE230 佐藤いずみ
指導教員：大石泰章

## 1　はじめに

強化学習とは，目標値を設定しそれに近づくと報酬，遠ざかると罰が与えられる環境下で，機械が何度も試行錯誤を繰り返し，報酬を最大化するように学習するという方法である．文献 [1] では状態と入力が連続なシステムの強化学習を扱っており，入力に制限がある場合も考えている．強化学習を適用したシステムは，文献 [1] では振子の振り上げ運動，文献 [4] ではロボットの起き上がり運動である．しかし強化学習による制御の有効性を調べるには，さらに多くのシステムに適用することが必要である．

本研究の目的は，実験キット [3] を使って製作した Ball & Beam という実験機において，強化学習による制御を行うことである．本来ならば，システムが簡単であるため，PID 制御などの制御方法を用いて制御するのが一般的だが，ここで強化学習を用いることによって，強化学習の有効性を確認する．あわせて強化学習を適用する上での注意点を考える．

## 2　制御対象

### 2.1　Ball & Beam 実験装置

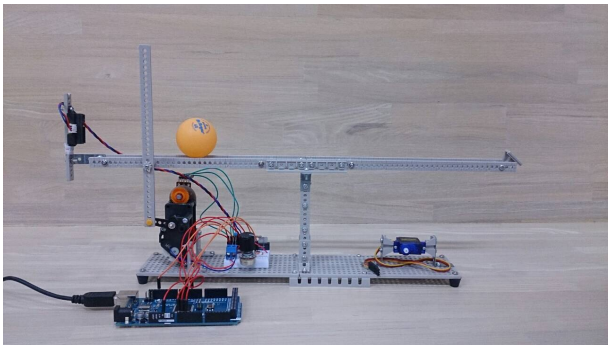

図 1　Ball & Beam 実験装置

図 1 は Ball & Beam 実験装置の写真である．これは，モーターを動かすことでアームと呼ばれる横長のレールの角度を変化させ，ボールを目標の位置に動かす装置である．文献 [2] に基づき，PID 制御等多くの制御が用いられているので，強化学習を用いる対象として適当であると判断した．

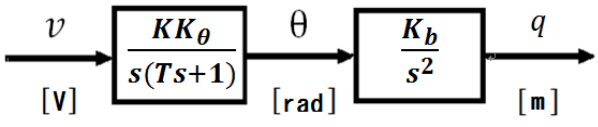


図 2　Ball & Beam 実験装置のブロック線図

図 2 は Ball & Beam 実験装置のブロック線図である．入力として電圧 $v$ をモーターに与え，アームの角度 $\theta$ を変化させ，アーム左端からの距離 $q$ が目標値に等しくなるようにボールを運動させるというものである．

表 1 に基づいて制御対象のモデル化を行う．図 2 のゲイン K と時定数 T は未知定数である．また，$K_b$=$100\times\frac{3}{5}g = 60g$，$K_\theta$=$\frac{\pi}{180}\times(\frac{2.1}{15})$[rad/deg] である．強化学習では，状態変数の次元が高くなると学習が急激に難しくなる．そのため本研究では角度 $\theta$ を入力として直接指定できると考え，制御対象は二次元のシステムであるとして扱う．

表 1　用いる記号

| 意味 | 記号 | 単位 |
|---|---|---|
| 傾いたビームの角度 | $\theta$ | [rad] |
| ボールの質量 | $m$ | [kg] |
| ボールの半径 | $r$ | [m] |
| ボールの慣性モーメント | $J$ | [kgm$^2$] |
| ボールの位置 | $x$ | [m] |
| 重力加速度 | $g$ | [m/s$^2$] |
| 電圧 | $v$ | [V] |

### 2.2　制御対象の状態空間表現

入力を $\theta$ とした場合の状態方程式を導出する．文献 [2] に基づき Ball & Beam 実験装置の運動方程式をたて，これを状態空間表現して，強化学習を適用する．

次の運動方程式が得られる:

$$(mr^2 + J)\ddot{x} = r^2 mg\sin\theta.$$

ピンポン球の中心回りの慣性モーメントは

$$J = \frac{2}{3}mr^2$$

である．更に $\theta \simeq 0$ と仮定すると

$$\ddot{x} = \frac{3}{5}g\theta$$

を得る．以上より状態変数を

$$x = \begin{bmatrix} q \\ \dot{q} \end{bmatrix}$$

と定めたとき, 状態空間表現は

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ \frac{3}{5}g \end{bmatrix} u$$

となる．

理論上 $\theta$ を入力と考えているが，実機においては $v$ を入力としているので，理想的な $\theta$ を満たすような $v$ を間接的に実現しなければならない．したがって，精度の高い制御を行うことは困難であることが予想される．

## 3　強化学習の適用

### 3.1　報酬と罰

強化学習において，制御の良し悪しの判断基準となる報酬関数を定める．今回は，アーム上においてボールを赤外線センサーから 0.2[m] 離れた位置で静止させることを目標とする．そこで，目標値 $q = 0.2$ に近づく程大きな報酬を与え，それ以外の時は，目標値から離れる程少ない報酬を与えるとする．また，ボールがビームの軌道から外れた時，すなわち q が 0 より小さいまたは q が 0.4 より大きくなった時，10 秒間罰として報酬 $r(x)$ の式に $-1$ を与えるものとする．

報酬は次の形で表せる：

$$r(x) = R(x) - S(u).$$

$R(x)$ は状態 $x$ に対する報酬関数，$S(u)$ は入力 $u$ の絶対値が最大値 $u^{max}$ を越えないようにするためのコスト関数である．今回は，

$$R(x) = -100(q - 0.2)^2 + 0.04 - \dot{q}^2$$

とする．これは，q=0.2，$\dot{q}$=0 の時報酬最大とする関数である．

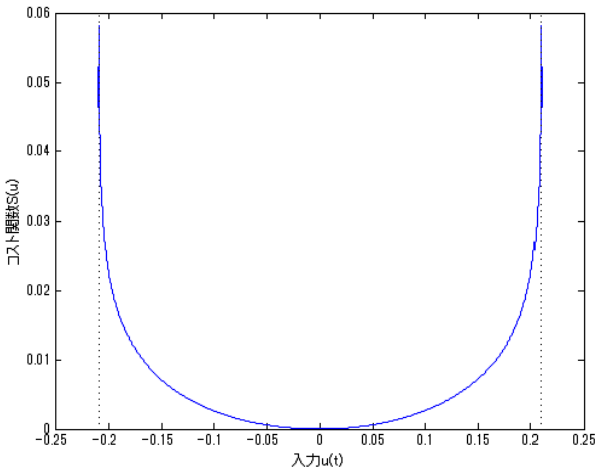


図 3　コスト関数 S(u) のグラフ

またコスト関数 $S(u)$ は

$$S(u) = c\int_0^u s^{-1}(\frac{u}{u^{max}})du$$

と表し，$c = 0.1$，$s(x)$=$\frac{1}{2}\arctan(\frac{\pi}{2}x)$ とする．今回は，入力の最大値は，$u^{max}$=$\frac{\pi}{15}$[rad] と設定する．

図 6 にコスト関数のグラフを示す．入力 $u$ が $u^{max}$ または $-u^{max}$ に近づくほど，コスト関数 $S(u)$ の値は大きくなる．

### 3.2　価値関数

価値関数とは，未来の報酬を考えた上で，現在の状態の価値を評価する関数である．報酬が即時的な意味での良さを示しているのに対し，価値関数は未来を見通した上での良さを示している．価値関数が定まれば，入れるべき入力が計算できる．本研究では次の形の価値関数を考える．

$$V^{\mu}(x) = \int_t^{\infty} e^{-\frac{s-t}{\tau}} r(x(s), u(s))ds. \qquad (1)$$

この時，未来の報酬を割り引いて評価しているので，価値関数の値は発散せずに収束するようになっている．$\tau$ はその時定数である減衰率である．この式をこのまま計算するのは困難である．そこで，

$$\dot{V}(x(t)) = \frac{1}{\tau}V(x(t)) - r(x(t))$$

が成り立つことに着目し，

$$delta(x(t)) := r(x(t)) - \frac{1}{\tau}V(x(t)) + \dot{V}(x(t))$$

とした時，$\delta(x(t))$=0 を満たす $V(x)$ を得ることで学習を行う．そのために，ノーマライズド・ガウシアン・ネットワークを用いる．

### 3.3　ノーマライズド・ガウシアン・ネットワーク

価値関数を推定する為に，ノーマライズド・ガウシアン・ネットワークを使用する．これは $N$ 個の基底関数 $b_1(x), b_2(x), \cdots, b_N(x)$ の重みつき和 $w_1b_1(x) + w_2b_2(x) + \cdots + w_Nb_N(x)$ であり，係数 $w_1, w_2, \cdots, w_N$ を学習して価値関数 $V(x)$ を近似することを考えるものである．パラメータ $w$ は以下のように更新する：

$$\dot{w}_k = \eta\delta(t)e_k(t). \qquad (2)$$

ここで，$e_k(t)$ はエリジビリティートレースと呼ばれるもので，$\delta$ を調節するものである．$\eta$ は学習係数である．基底関数 $b_1(x), b_2(x), \cdots, b_N(x)$ としては状態空間中の点 $c_1, c_2, \cdots, c_N$ をそれぞれ中心とするガウス関数を正規化したものを考える．このためノーマライズド・ガウシアン・ネットワークと呼ばれる．

基底関数の中心 $c_1, c_2, \cdots, c_N$ を状態空間中にどのように配置するかを考える．本研究では状態空間の中に，適当な範囲を設定し，これを分割することで格子点を定め，ここに $c_1, c_2, \cdots, c_N$ 等を配置する．分割範囲を狭め，分割数を多くすることで，精度が上がり，良い結果が得られると予想される．

しかし分割範囲の外の運動は評価出来ないので，考えられる運動を含む最小範囲を設定することが必要である．

また，分割数を大きくしすぎると計算量が多くなり好ましくない．運動の範囲を確認した上で，試行錯誤をし，適切な分割数を考えることが必要である．

本研究では，状態変数をそれぞれ 16 分割する．つまり，$0 \leqq q \leqq 0.4$，$-0.3 \leqq \dot{q} \leqq 0.3$ とし，$16^2$ の等間隔の格子を設定して基底関数の中心を定める．

## 4　シミュレーション

### 4.1　Ball & Beam 実験装置

今回のシミュレーションは，サンプル時間 0.02 秒で 1 回の試行を 20 秒間とし，それを 100 回繰り返して学習を行う．

$\tau$=15，$\eta$=0.01 としたときのボールの位置 q と入力 u，価値関数 V の変化を示す．図 4，5，6 は初期値 q=0.0530，$\dot{q}$=0 とした場合のグラフである．図 4 は，多少の誤差は見られるが，ボールの位置 q=0.2[m] で収束している．そして図 5 から分かるように，入力が 0 に収束しているので，目標値に近づいていると考えられる．また図 6 の価値関数 V は，距離が 0.2 以下の時，速度が正の方向に大きくなる程，価値関数の値は大きくなり，距離が 0.2 以上の時，速度が負の方向に大きくなる程，価値関数の値は大きくなる．

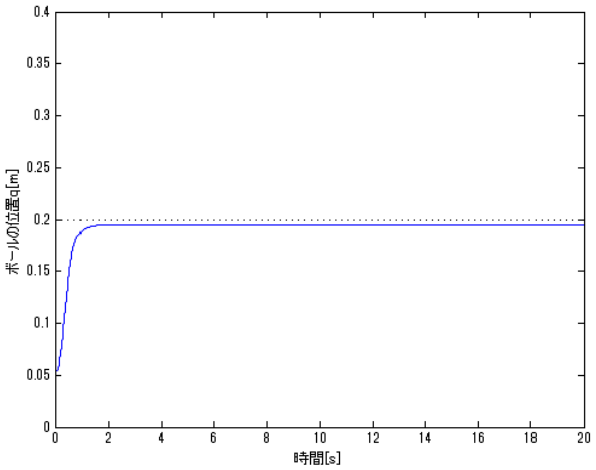


図 4　ボール位置 q の変化

また，図 7，図 8 は初期値が異なる場合でのシミュレーション結果である．初期値がどのような場合でも，多少の誤差は見られるがボールの位置 q は 0.2[m] に収束し，入力 $\theta$ は 0[rad] に収束している．

以上から，学習はほぼ適切に行われていると考える．

### 4.2　減衰率 $\tau$ と学習係数 $\eta$ の影響

本研究では，価値関数 V を大きくすることと，目標値にボールを収束させるため，入力をだんだん小さくすること，つまり V を小さくすることを目標とする．

前者の場合では，(1) 式より $\tau$ の値が小さいほど V の値は大きくなる．また，V は学習パラメータ w と基底関数 b の重み付け和で表されるので，(2) 式より V が大きくなるならば，$\eta$ も大きくなることが望ましいと考えられる．

後者の場合では，(1) 式より $\tau$ の値が大きいほど V の値は大きくなり，同様に $\eta$ の値は小さくなることが望ましいと考えられる．

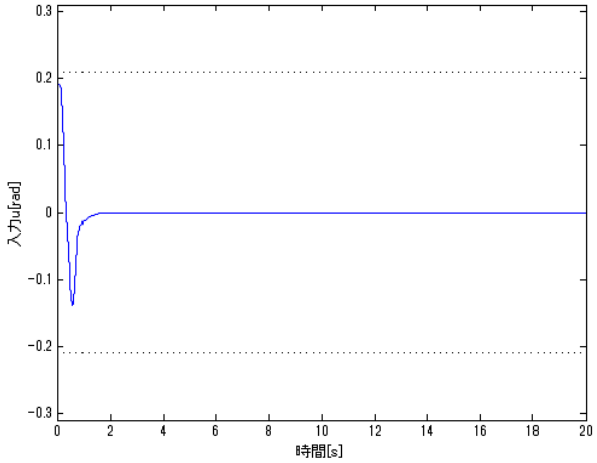


図 5　入力 u の変化

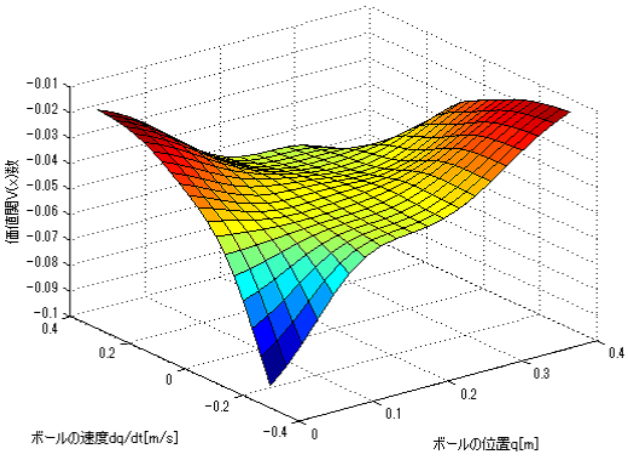


図 6　価値関数

そのため，試行錯誤によって，その値のときの適切な $\tau$ と $\eta$ を決定することが大切であると考えられる．

## 5　おわりに

本研究では，強化学習の有効性を確かめる為に，Ball & Beam 実験装置を用いて検証を行った．また，減衰率 $\tau$ や学習係数 $\eta$，分割数を変更することで，学習の精度への影響を比較した．

研究の成果として，強化学習による入力をビームの角度 $\theta$ とした場合の Ball&Beam 実験装置のシミュレーションに成功した．同時に，制御対象が変化するとパラメータ $\tau$，$\eta$ もそれぞれ変えなければならず，パラメータ選択が重要であることがわかった．

しかし，今回 100 回のシミュレーションを行ったが，実際に実験を 100 回行うことは非現実的である．そのため，パラメータの選択によって試行回数を減らすことが必要だと考える．

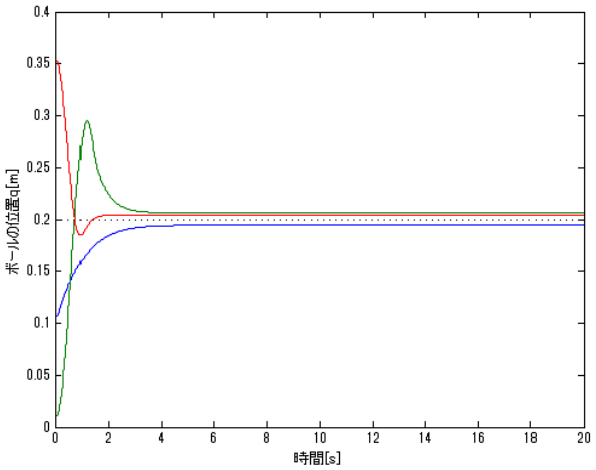


図 7　複数の初期値によるボール位置 q の変化

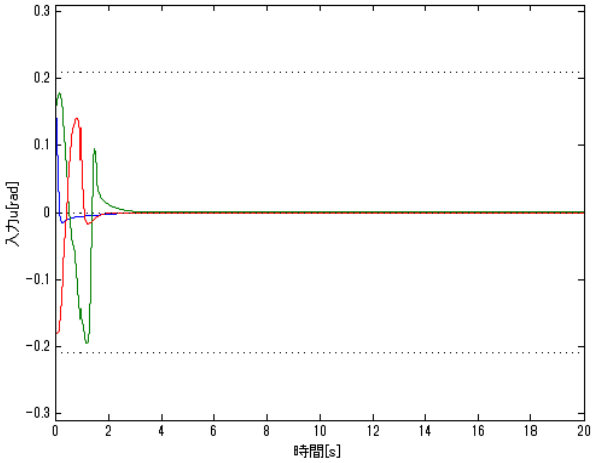


図 8　複数の初期値による入力 u の変化

今後の課題として，ボールの位置 q を目標値である 0.2[m] に誤差なく収束させるため，パラメータの調節を行う．また，入力を電圧 v とする場合のシミュレーションを行い，実機への実装をしたいと考える．

## 6　参考文献

[1] Kenji Doya: Reinforcement learning in continuous time and space, *Neural Computation*, vol. 12, no. 1, pp. 219–245, 2000.

[2] 平田光男：『Arduino と MATLAB で制御系設計をはじめよう！』．TechShare，東京，2012.

[3]『Arduino と MATLAB で制御系設計をはじめよう！Ball&Beam 実験装置 実験キット』．TechShare，東京，2012.

[4] 森本淳・銅屋賢治: 強化学習を用いた高次元連続状態空間における系列運動学習–起き上がり運動の獲得- **電子情報通信学会論文誌**, vol. J82-D-2, no. 11, pp. 2118–2131, 1999.