Gauss-Sigmoid ニューラルネットワークを用いた強化学習の安定性

大分大学　○前原 伸一，杉坂 政典，柴田 克成
Stability of Reinforcement Leaning Using a Gauss-Sigmoid Neural Network
Shinichi Maehara, Masanori Sugisaka, Katunari Shibata
Department of Electrical and Electronic Engineering, Oita University, 870-1192, Japan

Abstract ： Boyan has point out that the combination of reiforcement learning and sigmoid based neural network sometimes leads instability of the learning. In this paper, it is proposed that a gauss-sigmoid neural network, in which continuous input signals are put into a sigmoid based neural network through a RBF network, is utilized for reinforcement learning. It is confirmed by a simulation that the learning is faster and more stable when the gauss-sigmoid neural network is used than when the sigmoid based neural network is used.

1. はじめに

近年、自律ロボットや学習機械の開発などにおいて、強化学習の自律学習能力が注目を集めている。従来、強化学習は、行動などのプランニングの学習としてとらえられており、予め設計された状態空間から各行動へのマッピングを学習することが一般的であった。しかし、ニューラルネットと組み合わせることにより、センサからモータまでの、認識等も含めた一連の処理を総合的に学習することが可能となる[1)]。また、ニューラルネットの中間層が連続値状態空間の役割を果たし、これを他のタスクとの間で共有することで効率的に学習することも可能である。

一方、Boyan らは、推力が小さな車が反動をつけて山を登る hill car 問題などを例として、ニューラルネットと強化学習の組み合わせは学習の不安定につながることがあると指摘した[2)]。これに対し、Gordon や Sutton は CMAC などの入力信号を局所化する方法（縮小写像）を用いて on-line 学習させることによって学習の不安定性を回避できることを示した[3)4)]。さらに、逆にシグモイド型のニューラルネットのように、縮小写像でないと学習が発散する場合があることも示している[3)]。Boyan らの取り扱った hill car 問題は強い非線形性が要求されるため、CMAC や RBF(Radial Basis Fuction)などのように連続信号を局所化して、テーブルルックアップに近いかたちで表現することが有効である。しかし、これらの方法では基本的に局所化された信号の線形和という形で出力が表現され、中間層を有していないため、局所化された信号から適応的に大域的な表現を獲得することはできない。したがって、たとえば、ロボットに複数のタスクを学習させる際に、最初に学習させたことを次の学習に利用することはできず、空間認識のようにタスク間で共通に使えるものがあっても、１から学習し直さなければならない。また、GRBF(Generalized RBF)[5)]は RBF の汎化能力を改善したものと言えるが、RBF ユニットが密な領域では汎化能力は改善されない。

そこで、本稿では、筆者らの一部が提案した、Gauss-Sigmoid ニューラルネット[6)]を強化学習に用いることを提案し、 Boyan らの行った hill car 問題のシミュレーションにおいて、学習の安定性を検証する。

2. hill car 問題

本論文では、タスクとして強い非線形関数近似を必要とする hill car 問題を考える。 hill car 問題を Fig.1 に示す。

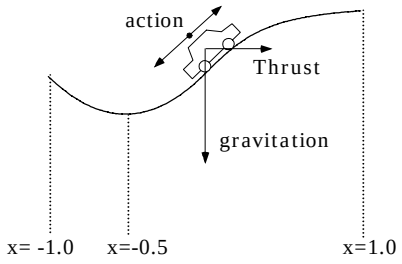


Fig.1　hill car 問題

この図の斜面の式は

$$\begin{cases} f(x) = x(x+1) & x < 0 \\ f(x) = x/\sqrt{1+5x^2} & x \geq 0 \end{cases}$$

と表され、 $f(x)$ を用いて

$$\begin{cases} \dfrac{d\dot{x}}{dt} = \{action/m - (g*(\dot{x}/\sqrt{1+\dot{x}^2}))\}/\sqrt{1+\dot{x}^2} \\ \dfrac{dx}{dt} = \dot{x} \end{cases} \quad (1)$$

m:車の質量，g:重力

という2つの微分方程式で車の運動を記述することができる。この問題で重要なことは車が丘を登ろうとする力より重力が強いことである。そのため、ある程度右向きに勢いがあれば右側に一度で登ることができるが、そうではない場合、一度左方向へ登り、その反動を利用しなければ登ることは不可能である。このとき、右方向に登れるか登れないかの境界部分では、理想的な評価関数と出力すべき力は共に不連続となり、その近似には強い非線形性が要求される。

3. Gauss-Sigmoid ニューラルネットワーク

シグモイド関数を出力関数とするニューラルネット（以下 NN と略す）では、大域的な表現が可能である反面、シグモイド関数の非線形性が弱いという特徴のため、ステップ関数などの強い非線形性関数近似には適していない。このことから、RBF など局所的な表現が可能な関数をニューラルネットにそのまま用いることが考えられるが、それだけでは、前述のように大域的な情報を表現できない。そこで、本論文では Fig.2 に示すように RBF の出力をシグモイド関数の中間層への入力とする Gauss-Sigmoid NN を用いる。シグモイド関数は単独では強い非線形性の関数近似が苦手であるが、入力として局所化された信号を用いることによって、容易に強い非線形性を実現することができる。そして、さらに、シグモイド型のニューラルネットの中間層において必要に応じて適応的に空間を再構成し、大域的な表現を獲得することも可能となる。

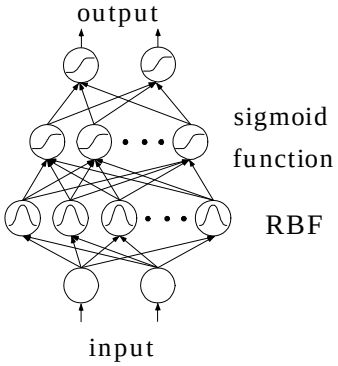


Fig.2 Gauss-sigmoid Neural Network

4. Actor-critic アーキテクチャ

強化学習の主なアルゴリズムとして、Q-learning と Actor-critic アーキテクチャが挙げられる。Q-learning は離散動作しか扱うことができないため、ここでは、Boyan らの論文と同様に、Actor-critic アーキテクチャを用いる。Actor-critic アーキテクチャは Actor（動作生成部）と Critic （状態評価部）から構成され、Critic では過去の経験をもとに現在の状態の評価を行い、Actor ではより高い評価値状態へ移動するための動作信号を学習する。状態評価値は、報酬がもらえるまで時間の経過とともに指数関数的に増大するように学習する TD(Temporal difference learning)学習に基づき、

$$\hat{r} = r_t + \gamma\ P(x_t) - P(x_{t-1}) \quad (2)$$

で表される TD 誤差を減少させるように、critic で1単位時間前の評価値 $P(x_{t-1})$

$$\Delta P(x_{t-1}) = \alpha_p \hat{r}_t \quad (3)$$

を学習していく。ここで、x_t は時刻 t での入力、γ は割引率、$P(x_{t-1})$ は時刻 t の状態評価値、r_t は時刻 t で得られる報酬、α_p は学習係数である。

一方、Actor では、時刻 t での動作 $\alpha(x_t)$ を中心としたガウス分布から決定した $\tilde{a}_t$ にしたがって動作し、その後、状態評価値が大きくなるように動作 $a(x_t)$ を次式にしたがって学習をする。

$$\Delta a(x_t) = \alpha_a(\tilde{a}_t - a(x_t))\hat{r}_t \quad (4)$$

ここで、α_a は学習係数、$\tilde{a}_t - a(x_t)$ は試行錯誤量を表す。

ここでは、Gauss-Sigmoid NN の出力ユニットを 2 つ設け、ひとつを動作、ひとつを評価として取り扱う。動作がベクトルの場合は、その要素数分だけ動作用の出力を用意する。

5. シミュレーション

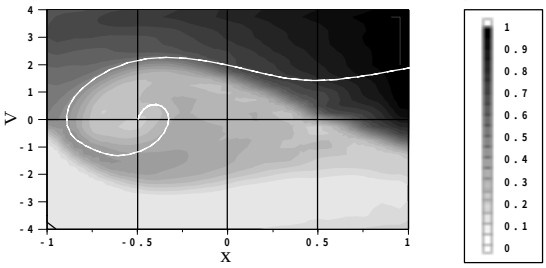
ここでは、Gauss-Sigmoid NN を用い、Actor-critic アーキテクチャで学習を行う。まず、車の位置 x、速度 v の初期状態は $-1 \le x \le +1, -4 \le v \le +4$ の範囲で乱数で決定し、その状態に対応した動作に一様乱数を３乗した値を加えたものを車に推力として与える。次に、その推力により(1)式に基づいてルンゲクッタ法を用いて遷移した状態を求め、それを評価する。強化信号は、車が丘の頂上に到達すれば 1、その他の場合は 0、左側に飛び出した場合、もしくは丘の頂上に達した場合は、(2)式の$P(x_{t-1})$を 0 とし評価と動作を学習する。このとき、推力は-3.0～+3.0 に制限した。このことにより、車は左側の x=-0.74 より高いところに登

った後でないとゴールに到達できず、逆に勢いをつけすぎると左側に飛び出してしまい、評価は下がってしまうという難しさがある。

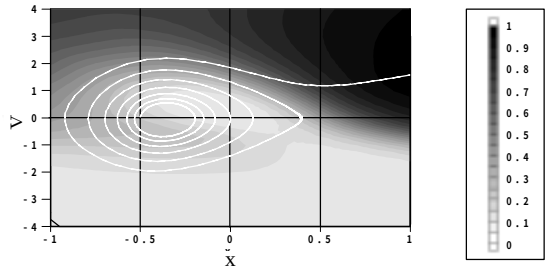
このような手順で初期状態を毎回ランダムに変えて学習を行った。速度は-4〜+4、Gauss-Sigmoid NN の中間層（シグモイド関数）の数は 40、RBF は 320 個とし、シグモイド型 NN の中間層も 40 とした。ただし、シグモイド関数は-0.5 から+0.5 の値域のものを使用している。これは、動作が正負対称であること、学習係数を大きくできることなどを考慮したためである。

5.1　シミュレーション結果

学習後の各状態に対する状態評価値、状態遷移の様子を Fig.3、学習により得た動作の様子を Fig. 4 に示す。

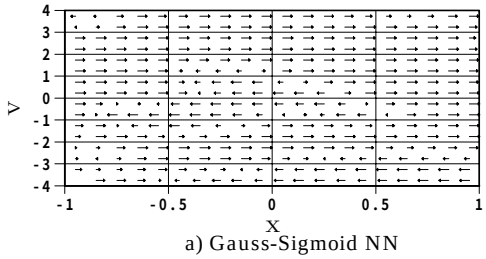


a) Gauss-Sigmoid NN

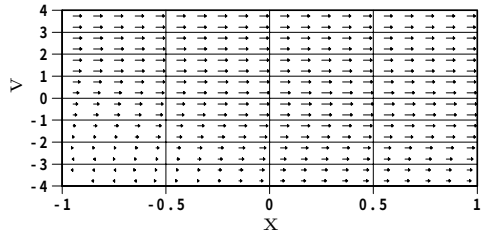


b) Sigmoid based NN

Fig.3　学習後の各状態に対する状態評価値と状態遷移の様子



a) Gauss-Sigmoid NN



b) Sigmoid based NN

Fig.4　学習により得られた動作

このタスクは、車がくぼみの中にいる場合は x＜-0.74,v=0 の状態を通過しなければ、ゴールに到達できない。Fig.3 からわかるようにシグモイド型のニューラルネットでは、何回もいったりきたりしないと上れない。一方、Gauss-Sigmoid NN では右-左-右で登ることができる。状態評価値を見ると、まず、両方とも右上の領域は黒色となっているが、これはそのまま右へ力を出せば登れる部分である。そして、Gauss-Sigmoid NN では評価値の尾根が x＝0.8、v=0 の部分を通って、ら線上になっており、車の軌道がその尾根に沿っていることがわかる。ところが、シグモイド型 NN では x＝0.8、ｖ＝0 のところに明確な尾根はなく、さらに、ら線状の尾根も観察できない。また、Fig.4 からわかるように、シグモイド型 NN で得られた動作はほとんど右方向なのに対し、Gauss-sigmoid NN では x＝-0.5、v=0 あたりでは右向きに動作が加わり、車が少し移動すると左向き、その後、左側に飛び出さないように右向きに動作が得られている。これらは、前述のシグモイド関数の非線形近似能力が不十分なことが原因と考えられる。

5.2 学習の安定性と学習速度

ここでは、シグモイド型 NN、RBF、Gauss-Sigmoid NN の３つの方法で、位置、速度の初期値をそれぞれ 0.25、1.33 間隔の格子状上の点から物理的に登ることが不可能な点を除いた 37 点について、2000 回の学習毎にゴール到達できるまでの平均所要ステップ数を比較する。ただし、100 ステップを過ぎても山を登れない場合、もしくは左側に飛び出してしまう場合は所要ステップ数を 100 とした。なお、シグモイド型 NN は 80、RBF は 10、Gauss-Sigmoid NN は 100 をひとつ前の層のユニット数の平方根で割ったものを学習係数として用いた。

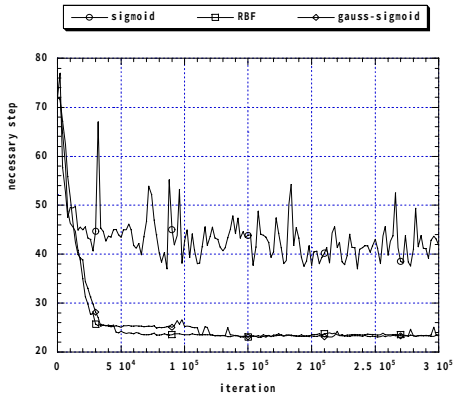


Fig.5 平均所要ステップ回数の比較

Fig. 5 からわかるように、シグモイド型 NN のみの場合、学習が安定せず、収束も遅い。RBF と Gauss-Sigmoid NN とを比較すると多少 RBF の方が収束が早いが、ほとんど同等と言える。

6. あとがき

本稿では、非線形性の強い hill car 問題をタスクとし、強化学習における Gauss-Sigmoid NN による学習の安定性を検証した。その結果、シグモイド型の NN と比較して、安定性の高い学習が可能であることがわかった。今後は、Gauss-sigmoid NN は RBF ユニットの中心と分散を変化させることもできるため、これによって、少ない RBF ユニットで効率的に学習できることを確認していきたい。

参考文献

1) shibata, K., Ito, K. & Okabe, Y. : Direct-Vision-Based Reinforcement Learning in "Going to an Target" Task with an Obstacle and with a Variety of Target Sizes, Proc. of Inter. Conf. On Neural Networks and Their Applications '98 PP.95-102(1998)

2) J.A. Boyan & A.W. Moore : Generalization in Reinforcement Learning:Safely Approximating the Value Function, Advances in Neural Infomation Processing Systems, MIT Press, 7, pp.369-376(1995)

3) Gordon, G. J. : Stable Function Approximation in Dynamic Programing, Proc.of the 12-th ICML, pp.261-268 (1995)

4) Sutton, R. S. : Generalization in Reinforcement Learning:Successful Examples Using Space Coarse Coding, In Advanced in Neural Information Processing System,vol8,pp.1038-1044(1996)

5) 森本淳, 銅谷賢治 : 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, J82-D- Ⅱ ,No.11, pp. 2118-2131(1999)

6) Katunari Sibata and Koji Ito : Gauss-Sigmoid Neural network, Proc. Of IJCNN'99, #747(1999)