F-037

正規化ガウス関数ネットワークを用いた群強化学習に関する検討
A Study on Swarm Reinforcement Learning with Normalized Gaussian Network

高橋　朋之[1]　　堀内　匡[1]
Tomoyuki Takahashi　　Tadashi Horiuchi

1. はじめに

人間の学習では，まず自分が試行錯誤を繰り返すことでタスクを達成していると考えられる．コンピュータにおいてもこのような行動学習の能力を実現する枠組みとして「強化学習」に関する研究が盛んになされている．一方，人間は複数の人間でタスクを行うことがある．これは，難しいタスクや，効率良く行う必要のあるタスクを協力して達成しようとする試みである．複数名での協力は，分担作業を可能にし，また，それぞれの個性（特性）の違いにより，行動学習に「多様性」を生む．一般に，他者の個性を知ることで，自身の個性の幅を広げることができる．本研究では，この「多様性」を強化学習に導入し，効果を検討する．なお，対象問題として状態空間が連続な問題を扱う．

2. 強化学習

強化学習は，学習者（以下エージェント）が自ら行動した結果に対し報酬が与えられる環境で，受け取る報酬を最大にするような政策を学習する手法である．本研究では，強化学習を実現する手法の一つで，逐次経験を用いる TD（Temporal Difference）法による学習を行う．

2.1 TD 学習

TD 学習では，実際の行動を通して得られた経験から逐次的に状態価値を推定する．エージェントが，状態 s においてある政策に基づき行動 a を選択した結果，状態 $\acute{s}$ に遷移し，報酬 r を得たとする．このとき TD 学習における状態価値 $V(s)$ は，式 (1) のように更新される．

$$V(s) \leftarrow V(s) + \alpha[r + \gamma V(\acute{s}) - V(s)] \qquad (1)$$

ここで α は学習率で，γ は将来にわたっての割引率である．式 (1) における $r + \gamma V(\acute{s}) - V(s)$ は，価値関数の目標値に対する誤差であり，TD 誤差 δ とよぶ．

$$\delta = r + \gamma V(\acute{s}) - V(s) \qquad (2)$$

この TD 誤差を用いて，評価の推定を行う評価部分（Critic）と政策を推定する行動部分（Actor）を別々に学習する方法を Actor-Critic 法という．Actor-Critic 法は行動価値を直接には学習しないため，連続行動空間の学習に対応できる．本研究では，Actor-Critic 法に基づく学習を行い，Actor および Critic は，連続値の入出力を扱うことができる正規化ガウス関数ネットワーク（Normalized Gaussian network： NGnet）によって関数近似を行う．

2.2 NGnet

NGnet は，状態空間を複数の動径基底関数 (Radial Basis Function： RBF) で分割し，各基底関数の出力を正規化した後に線形和をとる関数近似器である．RBF のパラメータとネットワーク結合重みを適切に調節することにより，任意の関数近似能力を有する．本研究では，入力として環境の状態 s を与え，Actor 出力 a および Critic 出力 V を得る．図 1 にその構造を示す．RBF のパラメータの更新に進化的 recruitment 戦略を，ネットワーク結合重みの更新に TD 誤差を用いる．進化的 recruitment 戦略 [2] は，NGnet の RBF ユニットを必要に応じて生成，それを評価し，淘汰・複製・突然変異による世代交代を行う．

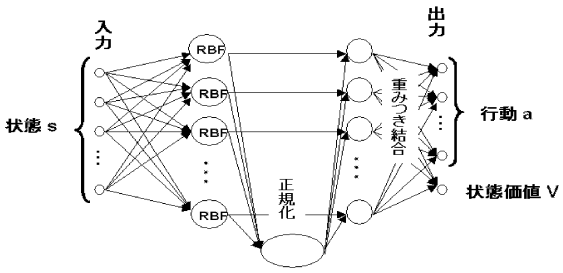

図 1: NGnet の構造

3. 群強化学習

強化学習は，エージェントが教師信号を必要としないが，学習に多くの試行回数が必要であることなどの課題がある．この課題に対し，効率的な学習を実現させる方法の一つとして，群強化学習 [3] が提案されている．群強化学習は，独立した同一のタスクを複数のエージェントで協力（情報交換等）して行う強化学習である．本研究では，連続値入出力を扱うことができる NGnet により群強化学習を実現する手法を提案し，エージェント群に多様性を持たせる方法について検討する．

3.1 NGnet の受け渡しによる群強化学習

群強化学習では，具体的にどのように協力するかを考える必要がある．NGnet による学習では，Actor や Critic の各出力や RBF ユニットのパラメータなど，様々なものが交換できる情報として考えられるが，今回，NGnet そのものを情報として受け渡す群強化学習を提案する（図 2 参照）．これは，ある程度の学習後に優秀だと判定された NGnet を他のエージェントに渡す方法である．渡されたエージェントはその NGnet を用いて，引き続き学習を行う．具体的な方法については，4.2 節で触れる．

3.2 多様性の実現

群強化学習は，エージェント同士が異なる情報を持っている場合に有効に働くと考えられる．エージェントは，試行錯誤による経験によって，それぞれで異なった情報を得ていくが，本研究では，より積極的に異なる情報を

[1]松江工業高等専門学校，Matsue College of Technology

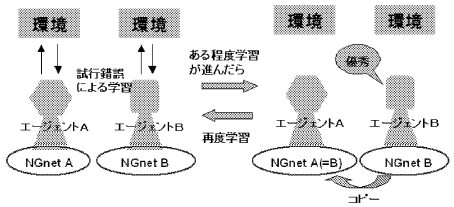

図 2: NGnet の受け渡しによる群強化学習

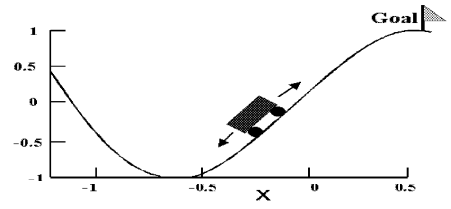

図 3: マウンテンカー問題

持たせるために，エージェントの設計パラメータ値を変えることを提案する．本研究で用いる NGnet は，設計パラメータが多く，多様なエージェントを設計するのに適している．

4. 実験

本研究では，状態空間が連続なマウンテンカー問題（図 3 参照）を対象とした実験を行う．これは急な坂道を登る台車の推進力制御の問題で，台車はゴールに辿り着くために，斜面を行き来して推進力を得る必要がある．台車エージェントは，入力（状態）として連続値の位置 x と速度 v を得て，出力（行動）として連続値の推進力 a を出力する．時間に対する位置 x と速度 v の変化は以下の式に従う．

$$\begin{aligned} x_{t+1} &= x_t + v_{t+1}\Delta t \\ v_{t+1} &= v_t + \left(-9.8mcos\left(3x_t\right) + \frac{a_t}{m} - kv_t\right)\Delta t \end{aligned}$$

$m = 0.2[kg]$ は台車の質量，$k = 0.3$ は摩擦係数，$\Delta t = 0.1[s]$ はシミュレーションステップである．推進力 $a[kgm/s^2]$ は，台車エージェントが $[-0.2, 0.2]$ の範囲で調整できるものとする．報酬は，台車エージェントの行動（ステップ）の度に-1 を，ゴールに到達したときに 1 を与える．

本実験では，スタート位置をランダムに設定し，ゴールに到達，もしくは，1000 ステップ終了で 1 エピソードとした．エピソード終了後は，スタート地点を再びランダムに設定し，次のエピソードに移る．これを繰り返し，100 エピソードで 1 実験終了とする．後述するいずれの実験についても 100 実験の結果で評価している．学習の有効性を示す指標として，タスク達成の速度と精度について考える．

4.1 エージェント設計による違い

設計パラメータ値の異なるエージェントの学習結果の例を表 1 に示す．median of steps は，各実験学習後の NGnet の評価の中央値で，failure rate は，学習の失敗率である．学習後の NGnet の評価は，全てのスタート地点に関して，ゴール到達までにかかったステップ数を平均したものであり，その際にゴールに到達できないスタート地点があった場合，そのエージェントの学習を失敗したものと判定した．表 1 より，各エージェントの設計パラメータの違いが生む多様な結果を確認できる．また，学習過程においても，NGnet の RBF ユニットの追加頻度などの違いを確認できた．

表 1: エージェント設計による学習結果の違い

	median of steps	failure rate[%]
agent_A	37.98	24.2
agent_B	38.90	12.4
agent_C	40.79	2.8

表 2: 群強化学習の学習結果

	median of steps	failure rate[%]
agent_A	37.98	24.2
agent_AA	42.41	7.0
agent_AB	43.59	10.4
agent_AC	47.00	0.8

4.2 NGnet の受け渡しによる群強化学習

群強化学習での実験は，同じ設計パラメータ値を持つエージェントで形成される群（以下同群）と異なる設計パラメータ値を持つエージェントで形成される群（以下異群）で行った．NGnet の受け渡しは，5 エピソード終了後に，その 5 回の平均ステップ回数が少ない NGnet を選択する方法で行った．エージェント群は 4.1 節で実験したエージェントから 2 つ選んで形成した．agent_A に注目した群強化学習の結果を表 2 に示す．群を形成するエージェントによって結果は異なるが，単独で行う学習に比べ，median of steps は増加し，failure rate は改善する傾向が見られる．

5. まとめ

本研究では，NGnet を用いた群強化学習を提案し，情報として NGnet を受け渡す方法について，効果を確認した．今後は，さらに実験を行い，多様性をもつエージェント群で学習を行う有効性について検討する．また，NGnet の受け渡しによる群強化学習とは異なる手法や，別の問題を対象とした実験を行うことを検討する．

参考文献

[1] 八谷大岳，杉山将，強くなるロボティック・ゲームプレイヤーの作り方，毎日コミュニケーションズ，2008

[2] 近藤敏之，伊藤宏司，進化的 recruitment 戦略を用いた強化学習による自律移動ロボットの制御器設計，計測自動制御学会論文集，Vol.39，No.9， pp.857-864，2003

[3] 飯間等，黒江康明，Actor-Critic を用いた群強化学習法，第 35 回知能システムシンポジウム資料，pp.9-14，2008