

尤度情報を利用した強化学習法（温度学習の提案）

小堀 訓成

橋本 周司

早稲田大学 理工学部

応用物理学科

1. はじめに

強化学習とは報酬という特別な入力を手掛かりとして環境に適応する機械学習の一つである。問題を報酬という形でエージェントに指示するだけで、自動的に問題が解けるので設計者の負荷が少ない手法といえる。しかし、報酬の設計方法は未だに明確ではなく、問題ごとにエージェントのおかれた状態に基づいて設計者が経験的に決めていく。複雑な問題においては報酬の設定が困難であり設計自身に試行錯誤を要してしまうため、強化学習本来の良さが損なわれてしまう。そこで本論文では、従来の報酬（設計者が設けるタスクの途中における達成度合）は必要としない枠組みでの強化学習を提案する。言い換えれば、最終的にタスクが達成されたか否かという情報だけを用いて学習する機構である。設計者はタスクの達成・失敗のみをエージェントに与えれば良い。具体的には、各状態にボルツマン選択における温度パラメータを設け、タスク達成・失敗の尤度を基に各状態の温度パラメータを学習させること（温度学習）でこれを可能とした。

2. 問題設定・実験環境

本論文の趣旨を明確にするため先に実験環境である“条件付き迷路”（10×10マス）（図1参照）について説明する。エージェントはスタート地点から始まり、上下左右の4方向に行動できる（各マスの座標を状態とした）。経路点を通してゴールしたときを初めてタスクが達成できたものとする。経路点が2つの場合は、順序も満たしてゴールしたときにタスク達成となる（V1 V2の順）。経路点1個の場合、300ステップ以内、2個の場合1000ステップ以内に経路点を通してゴール状態に行けない場合をタスクが失敗したとし、スタート地点に戻す。経路点を通過しないでゴールした場合はタスクが達成するまで、ステップの上限値を超えない限り試行を続けるものとする。エージェントにはどこが経路点であるかという情報は与えない。経路点に到達しても報酬は入らないものとする。従来の強化学習ではタスクが達成するまでの過程において、問題を解きやすくするために報酬を与えていたが、本論文はタスクの達成できたか否かだけの情報のみで学習を行なう。また、このような条件付きの問題は強化学習の枠組みにおいて非常に重要である。強化学習の代表的な例題 Mountain-Car task[1]も条件付きの問題の一つであるように実世界では条件付きの問題がほとんどである。Mountain-Car task



図1 実験環境

ではマルコフ性が満たされる点、本論文の問題より簡単な問題といえる。例えば、経路点が2個の場合、タスクが達成できたケースと経路点1個目のV1を通してゴールに行き失敗したケースが存在し競合するためV1での学習が困難である。よって環境同定型 Q-learning では解けない問題の一つである。

3. 温度学習

多くの強化学習エージェントはその行動則に次式に示す、ボルツマン選択を用いる。

$$P(a_i | s) = \frac{\exp(q_{is} / T)}{\sum_{j=1}^n \exp(q_{js} / T)} \quad (1)$$

ここで s は状態、 a は行動、 $P(a_i | s)$ は状態 s において行動 a_i をとる確率、 n は選択できる行動の数、 q_{is} 状態 s において行動 a_i を選択する行動優先度、 T は q_{is} の変化に対する $P(a_i | s)$ の変化の敏感さを決める温度と呼ばれるパラメータである。 q_{is} の値同様に T は学習に大きな影響を及ぼす。そこで本手法は各状態ごとに温度 T_s を与え、タスク達成の尤度を最大化するように T_s を更新させる。タスク達成・失敗の尤度 $P(g)$ 、 $P(u)$ は

$$P(g) = \sum_{s \in S} \sum_{j=1}^n P(g | a_j, s) P(a_j | s) P(s) \quad (2)$$

$$P(u) = \sum_{s \in S} \sum_{j=1}^n P(u | a_j, s) P(a_j | s) P(s) \quad (3)$$

$P(s)$ は状態 s の出現確率である。また $P(g|a_j, s)$ と $P(u|a_j, s)$ は状態 s で行動 a を選択した後、タスクが達成できる確率と失敗する確率である。学習の評価関数は、次式に示す L を用いる。

$$L = p(g)(1 - p(u)) \quad (4)$$

そこで(4)式対数尤度を最大にするように山登り法によって温度 T_s を更新させる。

$$\ell = \ln p(g) + \ln(1 - p(u)) \quad (5)$$

$$T_s = T_s + \eta \frac{\partial \ell}{\partial T_s} \quad (6)$$

$$\frac{\partial \ell}{\partial T_s} = \sum_{i=1}^n \left[\left\{ \frac{P(a_i, s | g)}{P(a_i | s)} - \frac{P(a_i, s | u)}{P(a_i | s)} \times \frac{P(u)}{1 - P(u)} \right\} \times \left\{ \left(\frac{1}{T_s^2} \left(\frac{\sum_{j=1}^n q_{js} \exp(q_{js} / T_s)}{\sum_{j=1}^n \exp(q_{js} / T_s)} - q_{is} \right) P(a_i | s) \right) \right\} \right] \quad (7)$$

ここで $P(a_i, s | g)$ 、 $P(a_i, s | u)$ はタスク達成または失敗に至るまでに状態 s で行動 a を選択する確率である。(7)式の $P(a_i, s | g)$ 、 $P(a_i, s | u)$ 、 $P(u)$ は、エージェントに M 回のエピソードが試行させて、その行動履歴から推定する。

$$P(a_i, s | g) = \frac{\sum_{j=1}^M N_j(a_i, s, g)}{\sum_{j=1}^M N_j(g)} \quad (8) \quad P(f) = \frac{\sum_{j=1}^M N_j(u)}{\sum_{j=1}^M N_j} \quad (9)$$

$P(a_i, s | u)$ は $P(a_i, s | g)$ と同様に求まる。ここで、 $N_j(a_i, s, g)$ 、 $N_j(u)$ 、 N_j はタスク達成までに状態 s で行動 a_i を選択する回数、タスク失敗までのステップの回数、全ステップの回数である。山口らは尤度を用いて q_{is} を学習させた[2]が、報酬設計の困難さは無い一方で、尤度の推定に多くの試行を要し、学習時間が遅いという問題がある。これはエージェントがランダムに動く程その重みの更新量が小さくなるためにおこる。本手法は各状態に割り当

Reinforcement Learning with temperature control using likelihood function

Norimasa Kobori and Shuji Hashimoto

Department of Applied Physics, Waseda University

てた温度 T_s を更新する。温度 T_s は微小な変化量にも敏感に反応する。よって少ない更新量でもエージェントの行動に影響を与えることができるため学習が早く進むことが期待できる。また(6)式の学習係数は(10)式のように設定した。同式の分母は、 T_s を変化させた時における確率の変異の上限を意味する。温度の変化によって行動選択確率の変化が少ない所は温度を急激に、大きい所は緩やかに調整することができる。

$$\eta = \frac{1}{\sum_{j=1}^n \left| \frac{\partial P(a_j | s)}{\partial T_s} \right|} \quad (10)$$

行動優先度 q_{is} の更新は Profit Sharing を用いた。50 エピソードに一回、タスク達成までのステップ数の最も小さい行動系列に対して以下の条件式に従い報酬 f_i を与えた。ここで、 W はエピソードの長さである。

$$\begin{aligned} & \text{if}(P(a_i | s) > 0.7 \wedge T_s > 1.0) \quad f_n = 0 \\ & \text{else} \quad f_n = R \times f_{n-1} \\ & \quad q_{is} = q_{is} + f_n \quad \text{但し} \quad R < 1 \quad (n=1,2,\dots,W-1) \end{aligned} \quad (11)$$

4. 実験結果・考察

2節で記した“条件付き迷路”にて、経路点が1つと2つの場合においてシミュレーション実験を行なった。Profit Sharing だけで学習する場合と Profit Sharing に温度学習を用いた場合の二つを、(11)式の $R=0.7, 0.9, 1.0$ の3つケースで行なった。また温度学習において、式(8)(10)の確率の推定は $M=5$ ごとに行い、温度 T_s を更新した (T_s の初期値は 0.3)。図2,3は20回実験した結果の平均である。図2のタスク達成までの平均ステップ数を見ると温度を学習した場合はいずれも早い段階(10000step)でステップ数が減少している。これは学習初期においては達成回数の割合が失敗回数に比べて比較的大きいため、温度が下がりゴール方向に行動選択確率が高められるからである。また $R=0.7$ のように、なかなか報酬が過去の行動系列にまで伝播しない場合、ランダムウォークの回数が増え、タスクが失敗するとランダムウォークした回数分だけ $P(a_i, s|u), P(u)$ が増加し、温度が上がってしまう。温度が上がるとまたより一層ランダムウォークを行い、報酬が入らない限りうまく行かなくなってしまう。よって温度学習を用いても温度学習なしの場合と変わらなくなってしまう。ゆえに Profit Sharing によって、長く過去の系列にまで遡った形で報酬を与える方が温度学習には向いていると考えられる。図2,3のタスク達成回数を見ると $R=0.9, 1.0$ の場合、温度を学習させることによってかなりの改善が見られた。これは温度を学習しない場合は、強化関数の R 値が大きいため迂回系列が発生してしまうが、温度を学習させること温度があがり迂回系列から脱することが可能になるためである。これは Profit Sharing の強化関数を設計する上での問題[3]を $R=1.0$ にとり温度を制御することで同時に解決していると考えられる。経路点2個の場合の学習過程における温度の変化を図4に示した ($f_0=0.1, R=1.0$ 温度学習有)。これは図1に示した実験環境の等温度面である。400000ステップ後、S V1 V2 G のルートは温度が下がっていることが確認できた。

5. むすび

本論文ではタスク達成、失敗の情報のみを用いた強化

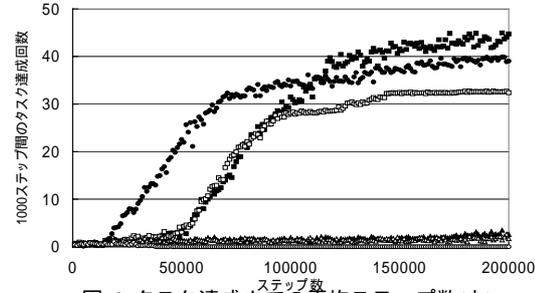
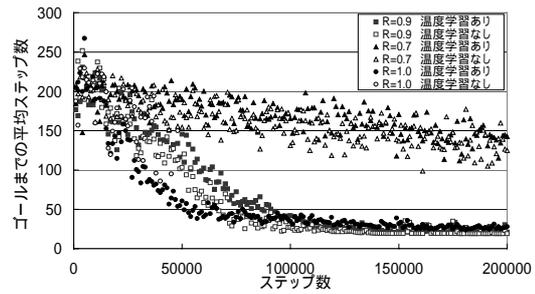


図2 タスク達成までの平均ステップ数(上)・タスク達成回数(下) (経路点 V1)

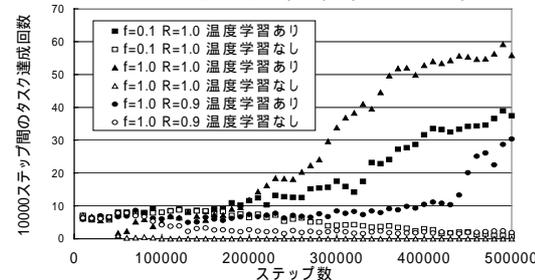


図3 タスク達成回数 (経路点 V1, V2)

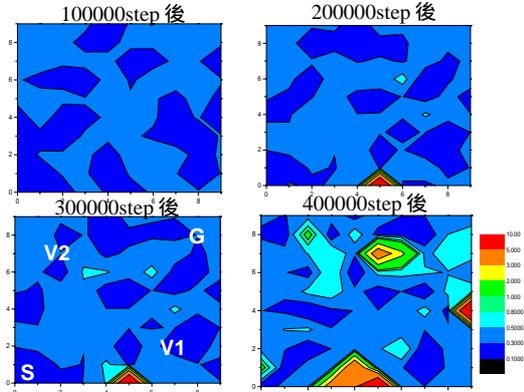


図4 等温度面 (経路点 V1, V2)

学習法として、従来の Profit Sharing の枠組みに各状態に温度パラメータを設け、タスク達成、失敗の尤度関数を基に温度を学習させる手法を提案した。今後は、他の条件付きの問題においても有効性を確かめたいと考えている。また本手法は環境が変化した時への対応や学習の効率化(温度パラメータのアニーリング)を図る上でも有望な手法と考えられる。

6. 参考文献

[1] Richard S. Sutton and Andrew G. Barto: Reinforcement Learning: An Introduction, A Bradford Book, The MIT Press (1998).
 [2] 山口智, 板倉秀清: 尤度最大化を目的としたエージェントの学習アルゴリズム, 電子情報通信学会論文誌 Vol.121-C, No.10, pp1612-1619(2001)
 [3] 宮崎和光, 山村雅幸, 小林重信: 強化学習における報酬割当ての理論的考察, 人工知能学会誌 Vol.9, No.4, pp580-587(1994)