1D3-01

# 未定義フラグを用いた価値関数の動的初期化による強化学習の学習速度の改善

Improvement of learning speed of reinforcement learning by dynamic initialization of a value function with an undefined flag

福永 哲也
Tetsuya Fukunaga

岐阜工業高等専門学校
Gifu National College of Technology

This paper describes a new learning method of Reinforcement Learning using dynamic initialization. In this method, all Q values are set "Undefined" before learning. Then, the Q values are initialized to observed target value dynamically. Learning experiment with 100×100 Grid-World clearly shows the validity of proposed method. In addition, the experiment with a mobile robot is conducted to confirm the advantage of proposed method in actual robot application. In the experiment, influence of priority of exploration and exploitation in this method is also evaluated. From the result of experiment using mobile robot, it is confirmed that the proposed method accelerate learning-speed.

## 1. はじめに

近年, 二足歩行ロボットやペットロボットのように自律型のロボットが我々の生活に近いところで多く見られるようになってきた. このような自律ロボットはその行動決定をロボット自身が持つ知能に基づいて決定することになる. そこで用いられるのがロボット自身に学習を行わせるという手法である. この学習という手法で知能を形成するならば, 知能を作るプログラマは全ての事象における行動を決定することはせず, ロボット自身に様々な事象を体験させることで求める動作をするように観察するだけでよいということになる. 現在, 様々な学習手法の中で強化学習が注目されている. 強化学習においてプログラマが学習のために用意しなければならないのは, 状態に対する報酬だけでよいため, 全ての事象における行動をプログラミングする場合に比べて格段に容易である. しかし, 与えられた環境が複雑であるほど, その学習にかかる時間は指数関数的に増大してしまう. そのため, いかに学習の速度を向上させるかという点が強化学習の大きな研究テーマであるといえる. そこで本研究では, 代表的な強化学習法の一つである *Q* 学習に対して, 学習の初期段階の速度を飛躍的に向上させる動的初期化を行う強化学習法を提案し, その効果を検証する.

## 2. 提案手法

ここでは, 本研究の基となる *Q-Learning* 及び提案手法について述べる.

### 2.1 Q 学習

強化学習の代表的アルゴリズムである *Q* 学習[Sutton 98] は

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha T \qquad (1)$$

$$T = r + \gamma \max_{a} Q(s',a)$$

のように, 各学習ステップで観測される目標値 *T* に *Q* 値を漸近することで, *Q* 値を最適化する. ここで, *Q*(*s*,*a*)は状態 *s* における行動 *a* の価値, $\alpha$は学習のステップサイズ, $\gamma$は割引率, *r* は報酬を表す. 通常, *Q* の値は任意の初期値に初期化される. *Q* の初期値は0である場合や, オプティミスティック初期値を用いる場合がある[Sutton 98]. 式(1)からも分かるように, *Q* 値の更新はステップサイズ $\alpha$を用いて, 少しずつ行われるため, 正しく近似された *Q* 値の獲得には数多くの経験が必要となる. 過去の行動の記憶を基に, 効率的に数多くの経験を積ませることによって学習を高速に行う手法として, *Dyna-Q* 学習や優先度スイープ等を用いた方法がある[Sutton 98].

### 2.2 動的初期化を行う Q 学習

*Dyna-Q* 学習等の方法は, *Q* 学習において式(1)に示される *Q* 値の更新を, いかにして数多く有効に行うかに注目し, 学習速度を改善している. これに対し, 今回提案する動的初期化を行う *Q* 学習(以後 *DI-Q* 学習:*Dynamic Initialization Q-Learning*)では, *Q* 値の初期化に注目する.

学習初期には式(1)における max*Q*(*s'*,*a*)も, 初期化された *Q* 値となる. しかし, この値には“学習すべき問題の解に対する情報”が含まれていない. また, 報酬 *r* も 0(don't care)であれば, 式(1)における *T* には“学習すべき問題の解に対する情報”がまったく含まれていないことになる. そこで, 提案手法では, *Q* 値は最初 "未定義(UD)" であるとし, *T* に初めて“学習すべき問題の解に対する情報”が含まれた場合に *Q*(*s*,*a*)を *T* で初期化し, その *Q*(*s*,*a*)は "定義済み(D)" となる. 学習手順を図 1 に示す. 状態 *s* から行動 *a* を決定する政策としては, 状態 *s* における *Q* 値が UD でない場合には, $\varepsilon$*-greedy* 行動選択手法を用いる. また, 状態 *s* における *Q* 値が UD の場合には, オプティミスティック初期値における探索と同様の方法で状態行動空間を探索するように行動を選択する. この場合, 未定義な行動の中で経験の少ない行動の経路を優先的に選択することになり, 状態行動空間の効率的な探索が可能となる. 実際には一つの状態において *Q* 値が UD の行動と D の行動が混在するため, $\varepsilon$*-greedy* による知識利用優先, あるいは UD に対する探索優先のいづれかの方法を採用する.

連絡先:福永哲也, 岐阜高専電子制御工学科, 〒501-0495 岐阜県本巣市上真桑 2236-2, fukunaga@gifu-nct.ac.jp

全ての $Q(s,a)$を"UD"に初期化し，状態 $s$ を初期化
以下を繰り返し：
(a)政策に従い $s$ に対する行動 $a$ を選択し出力
(b)次状態 $s'$，報酬 $r$ を観測
(c)$r \neq 0$ または $s'$における $Q$ 値に"UD"でないものがある場合に以下を実行：
$Q(s,a)$="UD"の場合は初期化し，それ以外の場合は式(1)で更新
(d)$s \leftarrow s'$

図 1. 学習手順

実際のプログラムに実装する場合には，$Q$ テーブルと同じサイズの未定義フラグを用意し，フラグが UD の場合には状態行動空間の探索を行い，D の場合には $\varepsilon$-*greedy* による行動を選択した．

## 3. 実験結果

本手法の有効性を確認するために，もっとも単純なグリッドワールドを用いて学習実験を行った．また，自律移動ロボット（Khepera）を用いて，一次元視覚センサ上の黒い物体に近づく動作の学習実験を行った．

図 2 はグリッドワールドにおける学習実験の結果を示す．この実験では格子サイズを 100×100 とし，スタートとゴールは対角に設定した．割引率 $\gamma$=0.9 とし，ステップサイズ $\alpha$は $Q$ 学習については 0.1 および 0.9 とし，*DI-Q* については 0.1 とした．また，$Q$ 学習ではオプティミスティック初期値として $Q$ の初期値は 0 とし，全ての行動に-1 の報酬を与え，*DI-Q* ではゴールしたときのみ+10 の報酬を与えた．

この実験では，報酬が確定的であるため，$Q$ 学習では $\alpha$を大きくすると学習が早くなり，$\alpha$=0.9 の場合には学習初期に素早くステップ数が減少しているが，その値はすぐに最適値に収束せず，エピソードが増えるに従って徐々にステップ数は減少を続けている．一方，*DI-Q* では最も初期のステップ数の減少は $\alpha$=0.9 の $Q$ 学習よりも遅れているが，その値は 100 エピソード付近で最適値近くに収束している．最初 $\alpha$=0.9 の $Q$ 学習よりもステップ数の減少が遅いのは，本手法では UD の行動に対する探索を優先しており，UD の行動が含まれている間は，最適な行動をとらないためと考えられる．しかし，エピソードが 100 を越えたあたりからは *DI-Q* の方がステップ数が小さくなっており，そのステップ数は最適値近くに収束している．この結果より，提案手法により学習初期の学習速度の改善が確認できる．

図 3 は自律移動ロボットによる実験結果を示す．実験では黒色の目標物体に一定の距離まで接近することを目標として実験を行った。報酬は、目標を達成したときにのみ与えられる。また、目標を達成するとエピソードは終了したとみなし、状態をランダムに初期化する。学習に際して、視覚センサの情報を基に目標物体までの距離を4種類、方向を5種類に分類し、目標物体が視界にない場合も含めて状態は21状態とした。また、ロボットは5 種類の行動を選択可能である。各方法について 30 分間の学習実験を数回繰り返し、エピソードの平均ステップ数を評価した。また実験では一つの状態に D, UD の行動が混在する場合に知識利用優先（exploitation）と探索優先（exploration）の 2 種類の実験を行い，さらに UD から D への更新を1度に 5 ステップ分遡って行う方法（sweep）について実験を行った．

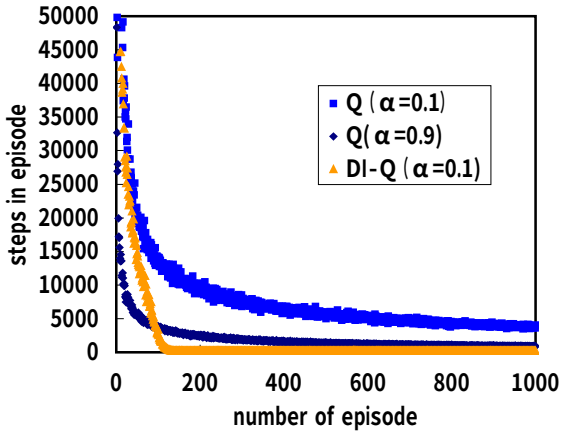


図 2. 100×100 グリッドワールドにおける学習実験結果（100 回の実験の平均値）

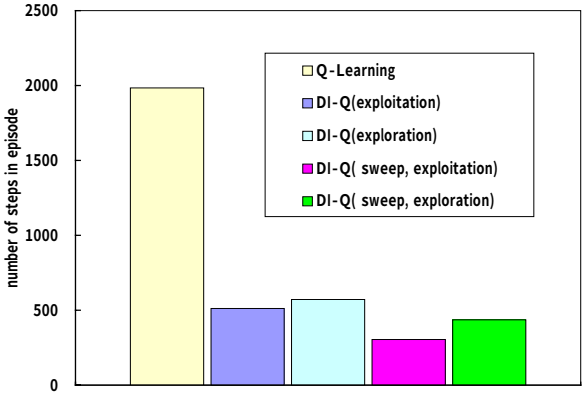


図 3. 自律移動ロボットにおける行動獲得実験

図から明らかなように，提案手法では $Q$ 学習に比べて平均のステップ数が 1/4～1/5 程度に小さくなっており，学習初期の学習高速化の効果が確認できる．

## 4. まとめ

$Q$ 学習において $Q$ 値の初期化を動的に行うことで学習初期の学習速度を高速化する学習手法を提案し，グリッドワールドおよび自律移動ロボットによる学習実験によって，提案手法の有効性を検証し，効果を確認した．本手法は，$Q$ 学習だけでなく，次状態の価値を用いて価値関数を更新していく *TD* 学習などの強化学習法へも適用可能であり，実問題における未知の行動獲得などに強化学習を適用する場合に有効であると考えられる．

## 参考文献

[Sutton 98] 三上貞芳，皆川雅章訳，Richard S. Sutton, Andrew G. Barto: 強化学習，森北出版,1998.