

自律移動ロボットにおける強化学習による軌道生成法の検討

Trajectory Generation for a Mobile Robot by Reinforcement Learning

清水 昌樹, 藤田 誠, 宮本 弘之

Masaki Shimizu, Makoto Fujita, Hiroyuki Miyamoto

九州工業大学大学院

Kyushu Institute of Technology, Kitakyushu

shimizu-masaki@edu.brain.kyutech.ac.jp

Abstract

強化学習はロボットに自律的に行動を生成させるための非常に有効な方法である。強化学習の中の Q-learning は比較的簡単な方法であるが、タスクが複雑になるに従い学習時間が長くなる。さらに離散的に状態と行動空間を分割するため、ロボットはスムーズな行動を得ることができない。これら問題を解決するため、我々は二つの方法を提案する。我々の方法の有効性を確かめるため、我々はロボットサッカーのような動的環境における Ball-To-Goal タスクの軌道生成に Q-learning を適用するシミュレーションを行った。

1 緒言

人間の仕事の一部をロボットに分担させたり、人間にとって作業が困難な環境や危険な環境におけるロボットの活躍が望まれている。その分野において近年、自律移動ロボットに関する研究が盛んである。ロボットが直面するタスクに対して人間が設計した行動計画は優れた性能を示す。しかし設計者の能力以上の複雑なタスクでは行動計画の設計者にかかる負担は大きくなる。

設計者の負担を軽減させるために強化学習を使いタスクの最適な行動計画を自律的に生成させることでロボットに行動を獲得させる。行動の指針を与える必要のある教師あり学習とは異なり、強化学習は環境の観測と可能な行動から自律的に学習が可能である。

もし観測可能な状態と行動が多すぎると学習は収束しない。さらに Q-learning での状態や行動は不連続である。これらの問題のため、エージェントと呼ばれる学習者は実環境や連続空間に対して正確に反応することが難しい。最近、この問題を解くため多くの研究がなされている。例えば Q-learning に連続空間を使うためいくつかの研究が

提案されている [3][4]。さらに階層型強化学習によって巨大な状態行動空間を解く研究も提案されている [5][6]。

本論文では上記の問題を解くため 2 つのシンプルな方法を提案する。最初の方法は本来未知な学習の初期値を設計者が設定することである。少ない知識を与えることによってエージェントは設計者の意図を理解し学習を始める。設計者が与えた初期値が最適な行動計画に近ければ、エージェントは全く間違った方向を探索しなくても良い。2 つめの方法は、状態と行動を等間隔で分割しないことである。移動ロボットの場合、従来は設定した数に従って状態と行動を等間隔に分割していた。しかし様々なターゲットへ向かう行動を獲得するタスクでは、このような分割は便利ではない。従って我々はターゲット周辺方向の行動空間を細かく分割し、その他の行動空間を荒く分割した。

本論文では、これら 2 つの方法を使った 1 つの学習モジュールだけで Ball-To-Goal タスクの行動軌道を 2 輪移動ロボットに獲得させることを目的とする。またロボットは単純な Ball-To-Goal タスクだけでなく攻撃や守備のような状況に応じた高いレベルのタスクの複雑な行動軌道も獲得させる。提案した方法の有効性を調べるためシミュレーションにより実験を行った。

2 プルトイ法

Ball-To-Goal タスクはターゲットに向かうという意味から To-Ball タスクと To-Goal タスクに分けることができる。そこでこの節では、まずロボットに Ball-To-Goal タスクを学習させる前に人の手によって設計した理想的な To-Ball タスクでの運動制御法を提案する。従来、我々が用いていた運動制御法では、前進や回転の速度を加えたものがモータに与えられる。この方法だとロボットは常に前進指令を受けていた。しかしこれではロボットの進行方向とボールの進行方向が大きく異なっていたとき、すばやくボールにアプローチできていなかった。これまでの運動制御法はロボットからボールまでの距離と角度が

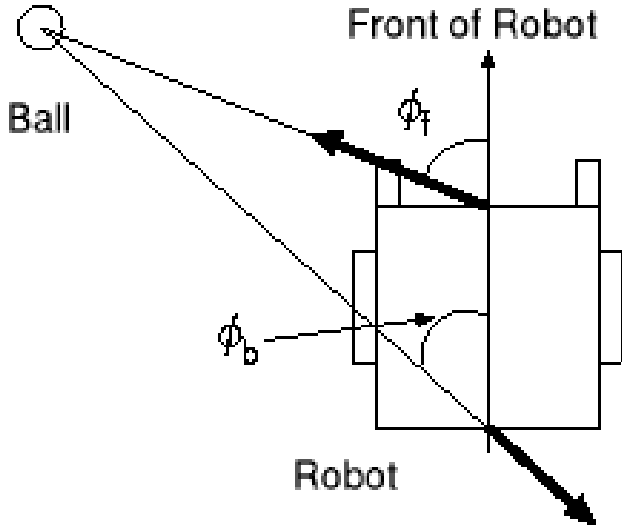


Figure 1: A new approach "Pulltoy"

ら以下のように計算される。

$$\begin{cases} V_r = V_c + A \times L \\ V_l = V_c - A \times L \end{cases} \quad (1)$$

ここで V_r は各ホイールの速度、 A は角速度に関するパラメータ、 L は左右のタイヤの間隔、 V_c はロボットの基準速度である。式(1)による方法を便宜上、スラスタ法と呼ぶ。

式(1)はスムーズに行動を生成するが、ボールがロボットの後方にあるときでさえ、前進に回転指令を加えボールにアプローチしていた。ロボットサッカーにおいては、ボールにどれだけ早くアプローチできるかが非常に重要になる。従って我々は人が行うように状況に応じて前進より回転を優先させる新しい運動制御法を提案する。

人はボールが自身の後ろにある場合、まずボールに向き、その後ボールへアプローチする。このようなアプローチがロボットには有効だと考える。しかし人間の行うアプローチを正確に解析し設計するのは難しい。従って新しいアプローチが必要になる。この論文ではボールにスムーズにアプローチする人の方法を簡略化した方法を提案する(図1)。新しい方法は式の中にロボットの先端からボールへの角度を使う。この角度が大きくなると図中の矢印の方向に引っ張られすばやく回転する。我々はこの新しいアプローチをプルtoy法¹と呼ぶ。

しかしロボットの先端から引っ張られるだけでは最適な行動を設計することはできないので、図1のようにロボットの後方とボールを結んだ角度も使用する。これは引っ張る力ではなく押す力になる。図1の ϕ_b は式を最適化すると省略できる。従ってプルtoy法は式(2)になる。

¹ プルtoyとは子供が小さな木馬や車の先端に結ばれた紐を引っ張って遊ぶおもちゃである。

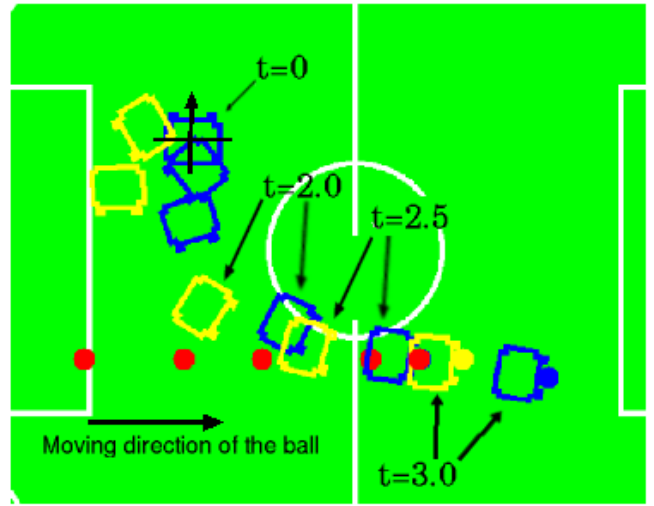


Figure 2: Comparison of Pulltoy and Thruster trajectory. Blue is Pulltoy. Yellow is Thruster. Ball is moving from left below to right. Pulltoy approach is acquiring the ball at $t=2.5$. Thruster is acquiring it at $t=3.0$.

$$\begin{cases} V_r = V_{max} \\ V_l = V_{max}(\cos(\phi_f) - \sin(\phi_f)) \end{cases} \quad (0 \leq \phi_f < \frac{\pi}{2})$$

$$\begin{cases} V_r = -V_{max}(\cos(\phi_f) - \sin(\phi_f)) \\ V_l = -V_{max} \end{cases} \quad (\frac{\pi}{2} \leq \phi_f \leq \pi)$$

(2)

(もし ϕ_f がマイナスの場合、 r と l が逆転する)ここで ϕ_f はロボット先端からボールまでの角度、 V_{max} はロボットの最大速度である。式(2)は非常に単純な方法である。さらに ϕ_f はボールだけでなく他のターゲットに対しても使える。

我々は次にスラスタ法とプルtoy法の軌道を比較した(図2)。

図のようにロボットの背後にボールがある状況では特に効果的であることがわかった。さらにロボットからボールへの初期角度、距離を変え比較実験を行った(図3)。縦軸はボールを得るまでの時間差でスラスタ法からプルtoy法を引いている。横軸はロボットからボールまでの下図が初期角度、上図が初期距離である。これらの図からほとんどの条件においてスラスタ法よりプルtoy法が効果的だといえる。もし初期角度が大きく、初期距離が小さい場合、その傾向は顕著になる。

3 強化学習

強化学習において学習エージェントは環境から状態を観測した後、ポリシーから行動を選択する。その後エージェントは変化した環境を観測することで報酬を得る。最終的にエージェントは受け取った報酬に従いポリシーを修正する。エージェントはこれを繰り返し学習する。我々は

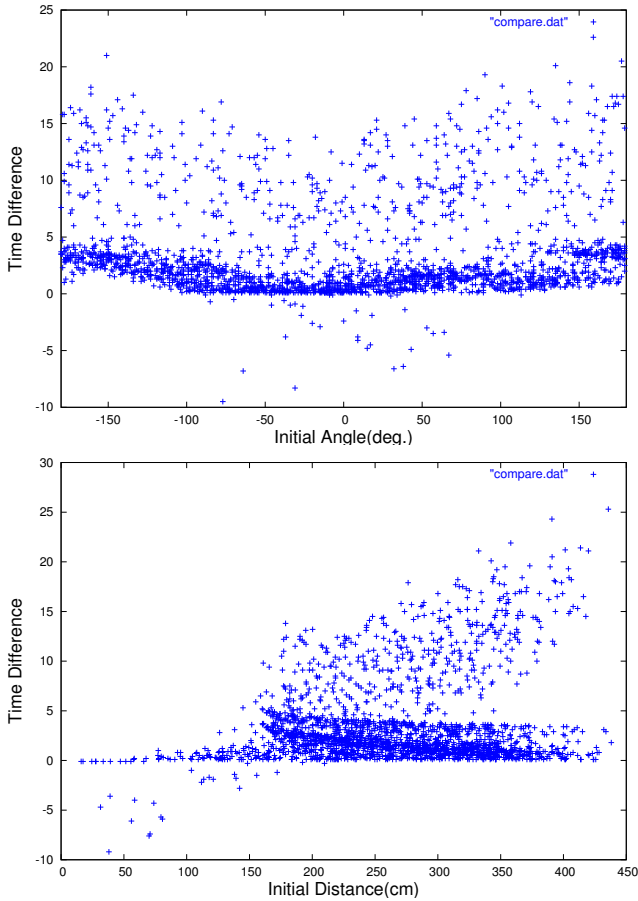


Figure 3: Graph of varying initial angle and distance. Upside in each graph represent that the Pulltoy approach could get the ball before the Thruster approach.

強化学習の1つである Watkins et al. によって提案された Q-learning を使う。Q-learning は行動価値関数 $Q(s,a)$ と呼ばれる関数を持ち、ポリシー における状態 s での行動 a を得るときの価値を計算する。この関数は式 (3) で定義され更新される。

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha [r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \quad (3)$$

ここで r は報酬、 α は学習率、 s_{t+1} は $t+1$ での状態、 γ は一定のステップサイズパラメータである。

学習開始時において正しい行動価値関数が未知なためエージェントは探索行動をしなければならない。エージェントは手本を使わず無知な状態から探索を開始する。そのため報酬を学習初期の段階で獲得することは難しく、良い行動を獲得するには時間がかかる。我々はこの問題を解くため行動を選択するためのポリシーに最適な初期値を前もって設定する。人の手によって比較的設計が簡単なタスクでは初期値を設定できる。設定された初期値から探索を開始することによって効率的な学習が期待できる。この考え方は人にも言えることで、あるタスクを人に教える場合、まず手本を見せることが重要である。本論文

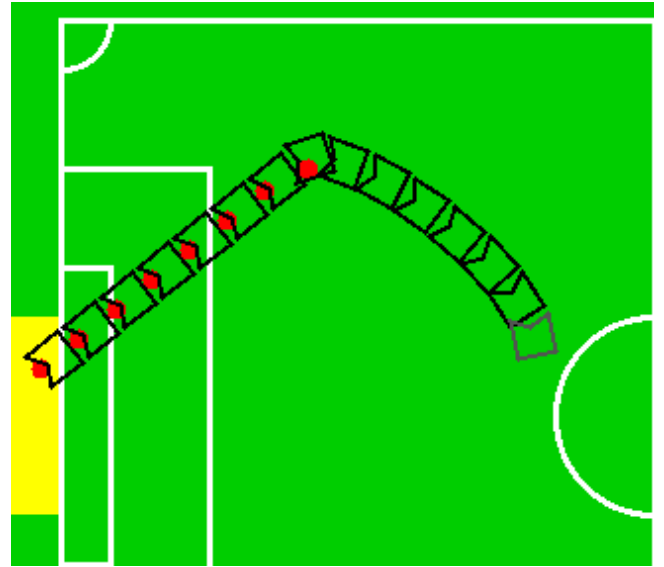


Figure 4: Trajectory of the robot for Ball-To-Goal task that generated by RL setting initial value

では前節で提案した To-Target タスクでのプルトイ法を Ball-To-Goal タスクでの強化学習法の初期値として使用した。

4 状況に応じた戦略の獲得

我々はエージェントに単なる Ball-To-Goal タスクだけでなく状況に応じたハイレベルなタスクでの軌道も獲得させる。ハイレベルなタスクではエージェントとボール、ゴールだけでなく複雑な環境の下でのタスクを考えた場合、通常の直接的なアプローチより、守備や攻撃などロボットの役割やターゲットの種類、敵ロボットの有無によってターゲットへのアプローチの仕方を変える方がより効果的であるといえる。例えば相手に攻められ守備をしなければならない場合、エージェントは自分のゴールとボールの間に入りながらアプローチする方が得点の危険性を考慮すると効果的である。

5 シミュレーション実験

Q-learning は一般的に観測可能な状態とエージェントの取りうる行動を空間的に分割する。本実験においても状態と行動の分割を使用した。状態にはロボットからボールおよびゴールへの角度と距離、ボールの速度を使用した。ボールへの角度の分割数は12、距離は4、ゴールへの角度は12、距離は4、ボール速度は5とした。移動ロボットの場合、一般的にエージェントの行動はロボットの周囲または取りうる行動をほぼ等間隔に分割する。しかしロボットサッカーでの Ball-To-Goal タスクの場合、行動空間は等間隔に分割すると効率が悪い。よって我々はターゲット周辺の行動空間を密に分割し、他の行動空間を荒く分割した。

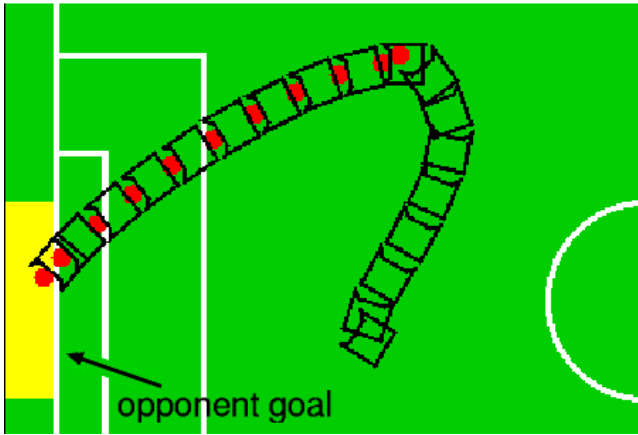


Figure 5: Generated trajectory for offense task

本論文ではシミュレータを使い Ball-To-Goal タスクにおいて軌道を生成させる以下の実験を行った。

5.1 良い初期軌道を与えたときの学習

ロボットの軌道は強化学習でのポリシーの初期値に2節でのプルツイ法を使い生成した。この結果を図4に示す。この時の報酬はボールをゴールへ入れた場合、1を与えている。

プルツイ法やスラスタ法などのように設計が簡単な一般的な方法では To-Ball や To-Goal としてターゲットを変えなければならない。しかし本論文で提案した方法を使った強化学習を使用することで初期位置からボールを獲得し、得点するまでを1つの軌道として生成することができる。また状態にボールの速度を含んでいるためボールの速度に対応した適切な行動を選択し軌道を生成できる。

しかし実験のような比較的簡単なタスクの場合、良い初期軌道を与えると学習後に生成される軌道は与えた軌道とほとんど変わらない。つまり今回与えた初期軌道であるプルツイ法は学習の手を借りずとも良好な手法であるといえる。

5.2 高度なタスクにおける軌道生成

図5,6は単にボールにアプローチするだけでなく役割を与えられた戦略下での高度なタスクでの軌道生成の結果を示す。報酬は役割に応じた方向を向きボールをキャッチした場合、1を与えた。さらにその状態からボールをゴールへ入れた場合、さらに1を与えた。

図5で示した攻撃行動は両チームのロボット共ボールに触れていない場合に使用される。ロボットがボールを獲得したときすでにゴールの方向を向いているポリシーを獲得している。相手ロボットがボールを持ちゴールに向かっていている場合、守備の戦略が最適である。この戦略ではゴールとボール間の軌道に自身を移動させ、その後ボールへアプローチする。図6は守備戦略の場合、生成

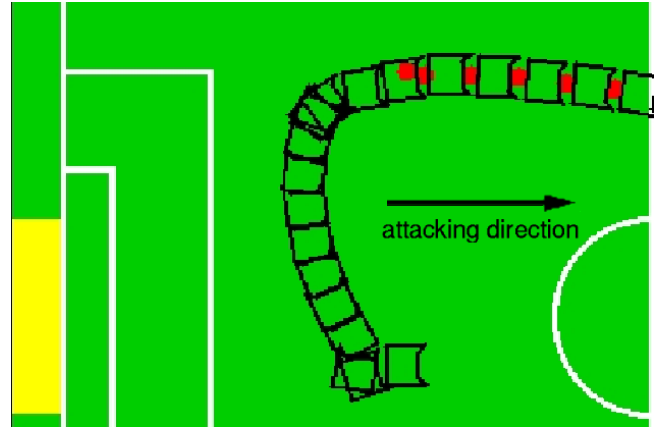


Figure 6: Generated trajectory for defense task

した軌道である。

攻撃や守備等の役割を与えられた場合、プルツイ法や他の運動制御法では満足な結果が得られない。報酬関数は複雑になるが人の手では設計が難しいタスクでは強化学習法が有効である。

6 結言

我々は実験結果から To-Ball タスクや To-Goal タスクのような To-Target タスクにおいてスラスタ法よりプルツイ法が有効なことを示した。もしロボットが小さいサイズ、または強化学習や他の方法ではメモリが足りないような場合、プルツイ法は簡単なため非常に有効だといえる。ロボットは強化学習によって Ball-To-Goal タスクでの軌道を自律的に生成することができた。強化学習を実ロボットへの適用を考慮した場合、学習の効率化、高速化が求められる。この問題は初期値の設定により解決できた。しかし学習後の軌道は人の手による良い初期軌道とほとんど同様の軌道が得られた。このことから単純なタスクに限れば両手法とも有効であることがわかった。

強化学習法において状態空間が大きくなりすぎると学習が収束しなくなる。さらに状況に応じた戦略軌道を設計することは人の手においても難しい。良い初期軌道として用いたプルツイ法であっても戦略に応じた軌道を生成することはできない。これら問題にも我々の比較的単純な方法で対応できた。また攻撃や守備のような難しい戦略に従ってボールにアプローチする軌道も生成できた。今後は複雑になった報酬を簡素化する方法の検討および実ロボットへの適用を目指す。

参考文献

- [1] R.S.Sutton, A.Barto, "Reinforcement Learning: An Introduction", MIT Press, 1998.

- [2] J.Peng, R.J.Williams, "Incremental Multi-Step Q-Learning", Machine Learning, Vol.22,pp283-290,1996.
- [3] A.Sherstov, P.Stone, "On Continuous-Action Q-Learning via Tile Coding Function Approximation", In Under Review., June 2004.
- [4] S.Hagen, B.Kröse, "Neural Q-learning", In Neural Computing & Applications, 12(2), pages 81-88, November 2003.
- [5] A.Barto, S.Mahadevan, "Recent Advances in Hierarchical Reinforcement Learning", Discrete Event Dynamic Systems:Theory and Applications, 13,41-77,2003.
- [6] Y.Takahashi, M.Asada, "Multi-Layered Learning Systems for Vision-Based Behavior Acquisition of A Real Mobile Robot", Proceedings of SICE Annual Conference 2003 in Fukui, Vol.CD-ROM, pp.2937-2942, 2003.