アルファゼロ型強化学習アルゴリズムを用いた最適制御手法の開発

Development of Optimal Control Using AlphaZero Reinforcement Learning Algorithm

楊　坤[*1], 渡部　雅也[*1], Dinesh Bahadur Malla[*1], 坂本克好[1], 山口浩一[1], 曽我部東馬*[1,2]

[1] 電気通信大学 基盤理工学専攻

[2] 電気通信大学 i-パワードエネルギー・システム研究センター

Deep Learning and Reinforcement Learning are developing rapidly in recent years. A lot of researches which apply deep reinforcement learning to the field such as game and robot control have generated great success. In this paper, we examine the possibility of adopting AlphaZero, an reinforcement learning algorithm demonstrates an unprecedented level of versatility for an game AI, to optimal control problems and gain insight on its ability to control the actions under noisy environment that is difficult to handle by using conventional control mechanism.

1. 序論

制御理論は 1778 年にワットの蒸気機関に適用された P(Proportional, 比例)制御から始まり、1940 年代に比例制御にオフセットを除去する I(Integral, 積分)制御と出力の先行抑制を行う D(Differential, 微分)制御を加えた PID 制御の普及により、制御理論として広く認知された。その後 20 世紀後半に線型代数を制御に応用した線形システム論が誕生し、多入力多出力システムを取扱えるようになり、状態方程式を解くことで解析的に制御の最適解が得られるようになった。

しかし、システムに非線型成分、外乱が存在する場合やシステムの詳細を知ることができないなどの場合は、システムを数式化し解析することが困難になる。そのため現在でも複雑なシステムにおいては PID 制御が用いられ、試行錯誤を通じてパラメータを調整する手法が使われている。

一方、1970 年代から、ニューラルネットワークと強化学習が急速に発展してきた。「任意関数を近似できる」ニューラルネットワークと「環境に対して事前知識を必要とせず、探索に通じて行動を最適化する」強化学習を用いて、非線型成分、外乱が含まれる複雑なシステムでも近似することができ、制御の最適化を行なえるようになった。

そして近年、強化学習分野は幾つのブレイクスルーを遂げている。2015 年に Google DeepMind 社が発表した DQN アルゴリズムがビデオゲームで超人的な高得点を叩き出し、話題となり、その後も最適制御への応用が成果を上げ、制御問題において DQN の有用性が確認された。2016 年、同じく Google DeepMind 社から AlphaGo が発表され、同年にトップ棋士イ・セドルを破り、世界を驚かせた[1]。2017 年に後続バージョンの AlphaZero が発表され、AlphaGo に 100 戦全勝という圧倒的な進化を遂げた。そのため、近年 AlphaZero の強化学習アルゴリズムを用いて、タンパク質のおりたたみなどより難解な「ゲーム」に挑戦し、別分野へ応用する試みが始まった。

本研究は Google DeepMind 社が発表した AlphaZero の論文[1]に基づき、AlphaZero 強化学習アルゴリズムを再現し、最適制御問題に対して AlphaZero 型強化学習アルゴリズムの有効性を検証する。また、適用した結果を現存の DQN アルゴリズムとどのような差があるかを比較する。更に外乱テストにおけるアルゴリズム予測性能のロバスト性を検証する。

連絡先：曽我部 東馬,電気通信大学 i-パワードエネルギー・システム研究センター，sogabe@uec.ac.jp

2. 最適制御と強化学習

2.1 最適制御と強化学習の関係

制御は、制御対象のプラントに思い通りの出力$r(t)$をさせるためにコントローラからの入力$u(t)$を決めることである。そして、最適制御は、図1のように制約条件のもとで、ある評価関数Jを最小化するように制御を行うことである。

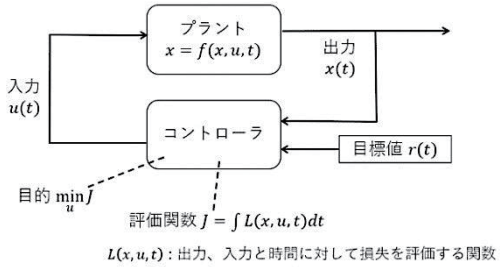

図 1　最適制御の模式図

すなわち、最適制御問題は一般的に評価関数

$$J = \int L(x,u,t)\,dt \qquad (2.1)$$

の最小値を与える$u(t)$を求める問題へと帰着できる。しかし、実際多くの場合では評価関数の最小値を解析的に解くことができない。

強化学習はエージェントが環境に働きかけ、働きかけによって生じる対象の変化の観測のみに通じて、ある目的を目指す最

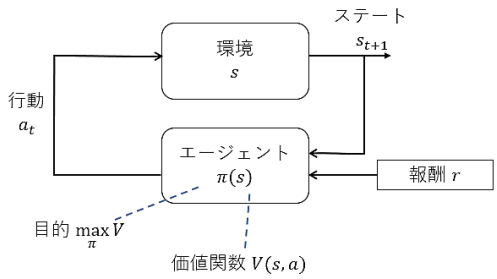

図 2 強化学習の模式図

適な行動を発見する枠組みである。強化学習において、目的は報酬r を用いて表され、ある行動を取ることで将来に渡ってもらえる報酬の期待値は価値v と表す。強化学習の目的は一般的に探索を通じて行動価値を最大化することになる。

ここで、最適制御の目標値と評価関数を強化学習の報酬に変換すれば、最適制御問題に強化学習を応用することができ、従来手法で適用できなかった複雑なプラントに対しても強化学習を用いたアプローチが可能になる。

2.2 アルファゼロの概要

アルファゼロ強化学習アルゴリズムは、DQN などのアルゴリズムと異なり[3]、探索にモンテカルロ木探索(Monte Carlo tree search, MCTS) を用い、価値(Value)と方策(Policy)をすべてニューラルネットワークに予測させ、木探索によるセルフプレイで得られた経験のみで予測を修正する構成になっている。従来のアルファ碁に比べ、価値を予測するバリューネットワークと方策を予測するポリシーネットワークが一つのニューラルネットワークに統合され、マルチタスク学習により予測精度の向上が図られた。

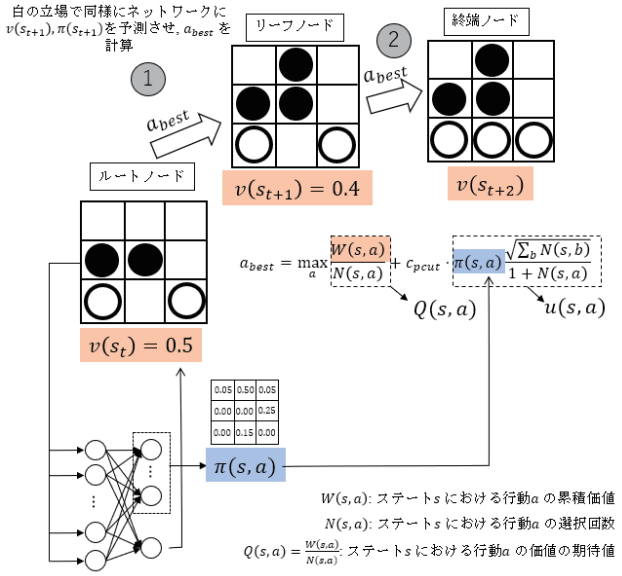

図 3 三目並べにおける MCTS の選択(Selection)

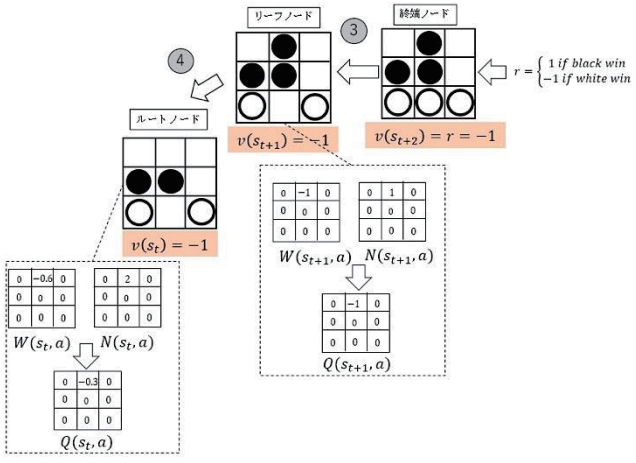

図 4 三目並べにおける MCTS の更新(Backup)処理

また、ニューラルネットワークの性能向上により、木探索でのプレイアウト(報酬を貰うまで探索木を伸ばす)処理が不要となり、より高速に探索を行えるようになった。

アルファゼロのセルフプレイは、今の盤面(ステート)をルートノードとし、図 3, 図 4 のように一定回数の木探索を行う。探索中に行動の価値$Q(s,a)$ が評価、更新され、価値の高い手はa_{best} の式によって、選択される回数も多くなる。探索が終わったあとルートノードでの各行動の選択回数に比例した確率でセルフプレイの次の一手を決め、この過程を繰り返してゲームが終わるまで進める。そして、一定回数のセルフプレイを行い、途中で経験したステート、その時に探索から得た行動方策、ゲーム最後の結果(報酬)を記録し、ニューラルネットワークに学習させる。この手順を繰り返すことで、ニューラルネットワークの予測精度が上がり、その予測結果を使った木探索の質も上がる。それによってセルフプレイでより最適な行動が取られ、蓄積される学習データの精度も上がる。この正帰還によって、アルファゼロが強くなる。

2.3 最適制御への応用

アルファゼロの入力であるボードゲームの盤面をプラントのステート、出力である次の一手の位置をコントローラの出力に対応させ、評価関数を変換して報酬とすれば、アルファゼロを最適

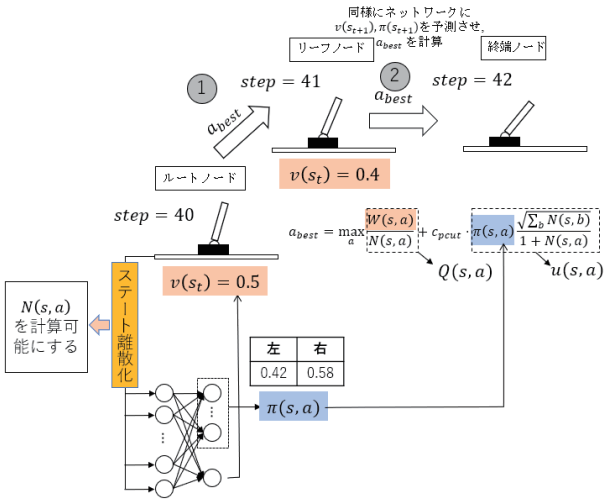

図 5 CartPole における MCTS の選択と展開処理

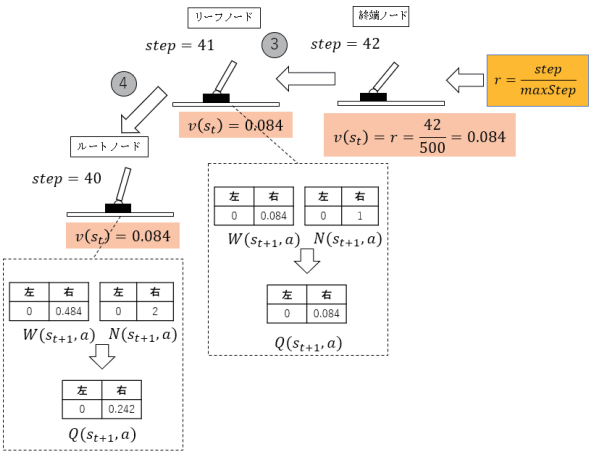

図 6 CartPole における MCTS の更新処理

制御に応用できる。アルファゼロの入力の盤面と出力の次の一手の位置は共に離散値であり、最適制御に応用にあたり、プラントのステートとコントローラの出力が連続値である場合、その値を離散値するか、あるいはアルファゼロの入出力を連続値に対応させる必要がある。制御問題においてプラントの状態は実際センサーなどによって測り、出力されるのでセンサーに精度があり、観測される状態は実質離散値となることを考慮し、本研究はより実装しやすい入出力の離散化を選んだ。

3. 評価実験

OpenAI Gym[2]に公開されている典型な制御問題である CartPole のシミュレーター(図 7)を用いて、アルファゼロの強化学習アルゴリズムを応用し、学習を試み、結果を別論文[3]で公開されている DQN の結果と比較した。また、学習済みのモデルを用いて、50 ステップ毎にランダムに振り子を左あるいは右に 0.1~0.15 rad を倒す外乱(振り子の角度の大きさが 0.21 rad を超えたらゲーム終了となる)を加え、外乱に対するモデルの応答を実験した。

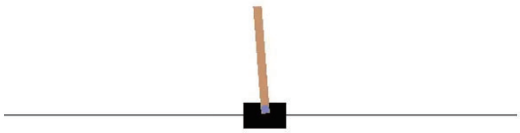

図 7 OpenAI Gym の CartPole シミュレーター

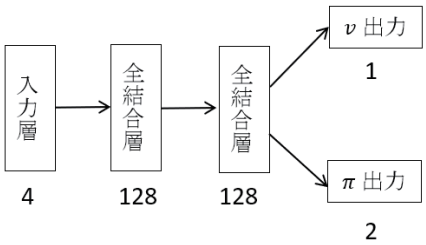

図.8 CartPole 実験におけるアルファゼロのニューラルネットワーク構成

Table 1 CartPole 実験におけるアルファゼロパラメータ設定

項目	条件
最大ステップ数	500
イテレーション回数	100
イテレーションあたりエピソード数	20
ステート丸め桁数	3
MCTS 探索回数	80
C_{puct}	1

4. 考察と展望

図 9 から、アルファゼロが 500 エピソードで振り子を最大の 500 ステップ倒立できるようになり、それ以後は最大ステップを維持できた。DQN と DQN with Target Network に比べて収束時の性能と安定性が優れている。図 10 から、最大 500 ステップの倒立の学習に対し、外乱実験で 2000 ステップを維持でき、さらに 50 ステップ毎に加えた外乱に対して、Cart の速度と Pole の角度が 0 になるように制御できた。従って最適制御においても、アルファゼロ強化学習アルゴリズムの有用性が確認できた。

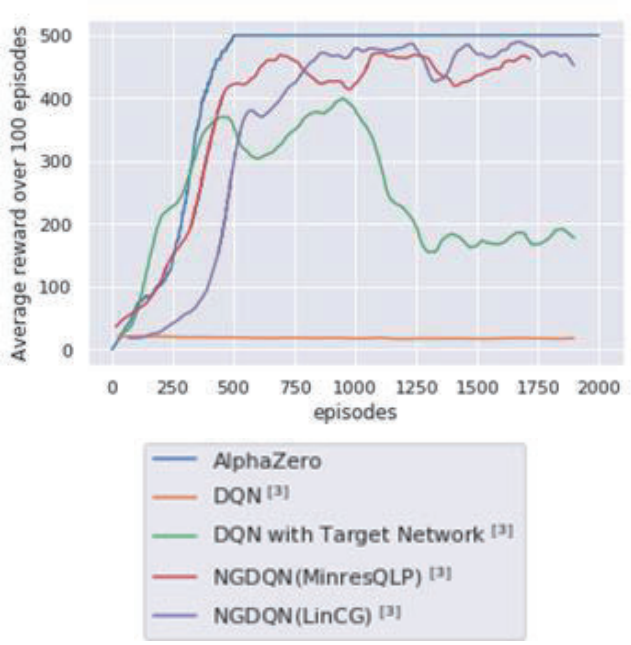

図.9 CartPole 学習結果の比較

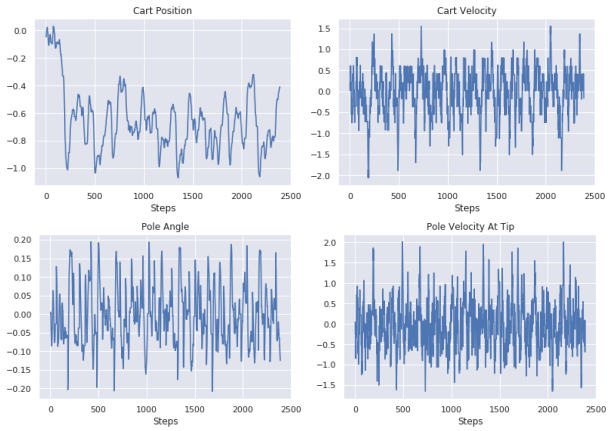

図.10 CartPole 外乱実験の結果

本研究の展望として、今回はプラントのステップ入力を離散化しアルファゼロを適用したが、連続ステート、連続出力を必要とする制御問題では、アルファゼロの MCTS を連続入力、連続出力に対応させる研究が期待される。

5. まとめ

本研究では、最適制御と強化学習の関係、強化学習アルゴリズムのアルファゼロの概要を説明し、アルファゼロを最適制御問題に応用できることを提案した。そして、アルファゼロを最適制御問題の一つである CartPole に応用し、既存の DNQ アルゴリズムと比較した。また外乱実験でアルファゼロが振り子を倒立することができ、外乱のあるプラント制御においても有用であることを確認できた。

参考文献

[1] D Silver, J Schrittwieser, K Simonyan, I Antonoglou, A Huang, A Guez, T Hubert, L Baker, M Lai, A Bolton, Y Chen, T Lillicrap, Fan Hui, L Sifre, G van den Driessche, T Graepel,

D Hassabis, “Mastering the game of Go without human knowledge,” Nature, 2016.
[2] CartPole-v1, https://gym.openai.com/envs/CartPole-v1/
[3] E Knight, O Lerner, “Natural Gradient Deep Q-learning”, arXiv:1803.07482, 2018