1M-01

Differentiable programmingによる強化学習の最適化

黄 伊莎　Tristan Hascoet　高島 遼一　滝口 哲也　有木 康雄
神戸大学

1 はじめに

制御などの問題における一般的な設定は，環境とエージェントが与えられ，エージェントの取る行動によって環境が変化し，それに応じた報酬がもらえることである．強化学習の目的は，報酬を最大化するための行動を決定することである．実応用として，碁や atari ゲームなどで有効性が示されている．しかし，強化学習は膨大な試行錯誤を必要とする，という問題点があった．

近年，Zygot により，物理シミュレーションの環境において，物理的法則が微分可能であることを示された [1]．これは differentiable reinforcement learning (DiffRL) と呼ばれ，強化学習を教師あり学習に置き換えることで，学習時間の大幅な短縮を実現した．しかし，この要因については明らかにされておらず，本稿ではこれについて調査する．

2 DiffRL と強化学習

図 1，2 に，強化学習の手法である deep Q network (DQN) と DiffRL の計算グラフを示す．ここで，○は変数，□はオペレーションを表す．また，時刻 t に取る行動を $a(t)$，時刻 $t+1$ の状態を $S(t+1)$ とし，W はモデルのパラメータである．そして $Exp.reward$ はモデルを通して出力される行動価値の期待値である．$I(t,n) = \frac{\partial S(t+n)}{\partial a(t)}$ は，行動 $a(t)$ が時刻 $t+n$ の状態に与える影響の大きさを表す．

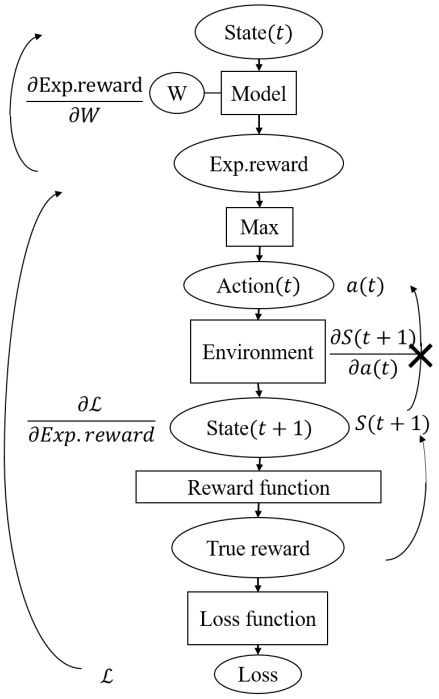

図 1　DQN の計算グラフ

DQN とは Q 学習とニューラルネットワークを組み合わせたものであり，状態を入力し，行動価値を出力する．最大の行動価値が得られる行動によって環境を更新し，次の状態に遷移する．$I(t,n)$ を直接計算することは難しいため，状態更新後の真の報酬と期待報酬の差から損失関数を設計し，モデルの学習を行う．

DiffRL は，DQN と異なり，状態の更新に物理法則を利用することで微分可能な $I(t,n)$ を計算する．そのため，目的の報酬を直接最適化するようにパラメータの更新を行うことができる．これにより，強化学習で必要としていた膨大な試行が削減できるため，学習を高速化することができる．

Optimization of Reinforcement Learning using Differentiable programming
Isa Ko, Tristan Hascoet, Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki, Kobe University

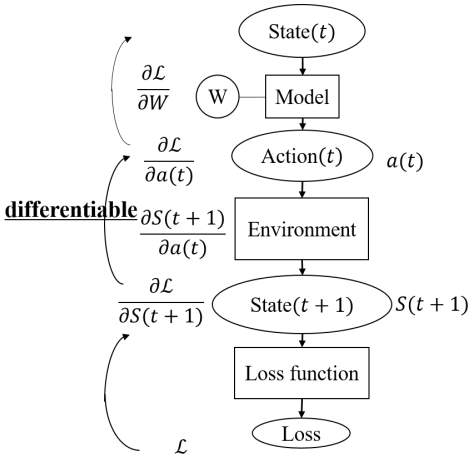

図 2　DiffRL の計算グラフ

3　実験

本実験では，図 3 に示す倒立振子 (cartpole) 問題を扱う．状態 S は，カートの位置 x，速度 v，ポールの角度 θ，ポールの角速度 ω の 4 変数からなる．行動は，左右どちらに動かすかである．x の閾値は 2.4，θ の閾値は 12 度とした．

モデルの最適化には Adam を使用した．損失関数は平均二乗誤差とした．また，DiffRL の教師信号は，x の閾値と θ の閾値の積とした．

図 4 に，DQN と DiffRL の学習前の $I(t,n)$ を示す．図 5 に，DiffRL の学習後の $I(t,n)$ を示す．DQN の学習後 $I(t,n)$ は，学習前から大きく変化しなかったため割愛する．t を 1 から 8 まで変化させ，n は 2 から 9 とした．ここで，勾配の計算には位置変数 x を用いた ($I(t,n) \approx \frac{\partial x(t+n)}{\partial a(t)}$)．学習前と学習後で DQN は $I(t,n)$ を変化させなかったが，DiffRL では同じ t に対し，n が大きくなるにつれて値が小さくなる傾向が得られた．このことから，2 つの手法では異なる方策が取られていると考えられる．しかし，時刻 t に取る行動によって離れた時刻の状態に与える影響が大きくなると学習が不安定であるため，DiffRL の方が好ましいと考えられる．

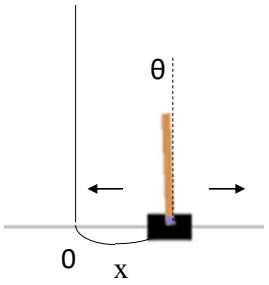

図 3　Cartpole の問題設定

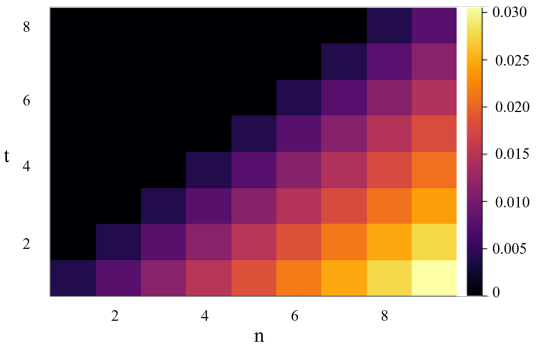

図 4　DQN と DiffRL の学習前の $I(t,n)$

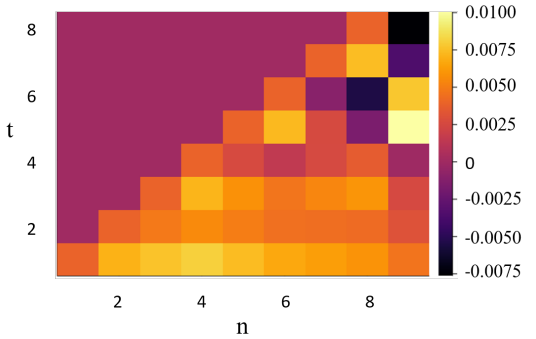

図 5　DiffRL の学習後の $I(t,n)$

4　おわりに

本研究では，DiffRL の有効性について，エージェントの取る行動が将来の状態に与える影響について，位置変数に着目して調査を行った．今後は，速度などその他の変数についても検討する．

参考文献

[1] M. Innes *et al.*, “A differentiable programming system to bridge machine learning and scientific computing,” arXiv preprint arXiv:1907.07587, 2019.