ISSN 2186-5647

2-52

決定論的な方策勾配を用いた強化学習における探索能力の向上

日大生産工　石井琢磨　　日大生産工　山内　ゆかり

1. まえがき

現在強化学習は多くの発展を遂げ様々な学習アルゴリズムが生まれている。しかしこれらの手法はいずれも計算資源や学習時間に膨大なリソースを消費する。今回は、Asynchronous[1)]の考えを取り入れ複数エージェントを追加し非同期的な分散学習に決定論的な方策勾配[2)]を導入することで計算時間を短縮することが出来るのではないかと考えた。本論文ではGrid Worldの迷路問題実験により提案手法と従来のA3Cを比較し、学習時間及びその精度について報告する。

2. 従来研究

2-1 強化学習

強化学習とはある環境内におけるエージェントに行動選択をして獲得する報酬を最大化する学習方法である。例としてQ学習やTD学習が挙げられる。

2-2 方策

方策とは行動を決定するルールのようなもので環境からのフィードバックに、より良い方策になるよう修正していく。

2-3 決定論的な方策勾配

決定論的方策勾配とは方策勾配法の行動方策を更新する手法である。方策勾配法とはエージェントの行動確率をニューラルネットワークで表現する手法である。方策勾配を更新する際、基本的には確率論的に更新をするものである。しかし、決定論的に方策勾配を更新する手法を導入することで確率論的に方策勾配を更新するよりも早期に学習が収束し、高い精度で学習することが出来る。

$$\nabla_\theta J(\pi_\theta)$$

$$= \int_S p^\pi(s) \int_A \nabla_\theta \pi_\theta(a|s) Q^\pi(s,a) da ds \quad (1)$$

$$= E_{s\sim p^\pi, a\sim\pi_\theta}[\nabla_\theta log\pi_\theta(a|s) Q^\pi(s,a)]$$

(1)の式は確率論的な方策勾配を表したもので状態空間と行動空間の2つに対して期待値をとっている。

$$\nabla_\theta J(\mu_\theta)$$

$$= \int_S p^\mu(s) \nabla_\theta \mu_\theta(s) \nabla_a Q^\mu(s,a) ds \quad (2)$$

$$= E_{s\sim p^\mu}[\nabla_\theta Q^\mu(s, \mu_\theta(s))]$$

(2)の式は決定論的な方策勾配を表したものである。確率的な方策勾配が状態空間と行動空間に期待値を取っているのに対し決定論的な方策勾配は状態空間のみに対して期待値を取っているため、方策勾配を推定する際必要なサンプル数が少なくなり計算量が小さくなる。

2-4 Actor-Critic

Actor-Critic[3)]はTD誤差学習手法の1つである。Actor-Criticのメリットは、行動選択に最小限の計算量しか必要ない点、また確率的な行動選択を学習できる点である。今回は決定論的な方策を用いるため十分な探索を保証するのにオフポリシーActor-Criticアルゴリズムを使用する。

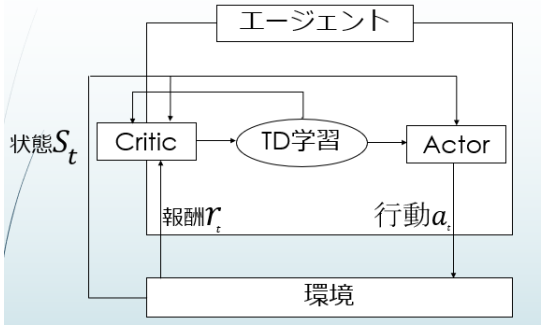

Figure 1　Actor-Critic のモデル図

Figure1はActor-Criticのモデル図を示したものである。Actorは環境から状態を認識し、エージェントはActorに与えられている確率分布に従い行動aをとり1つ先の状態sへ移動する。その際Criticが環境から報酬を得てTD誤差を計算する。Criticは行動価値をTD誤差が0に近づくように更新する。

TD誤差とはエージェントが行動する前に想定していた行動の評価値と実際に行動して得た評価値との差であり、TD誤差が正ならば想定より良い行動、逆に負ならば想定より悪い行動ということになる。

Improvement of search ability in reinforcement learning using deterministic policy gradient

Takuma ISHII and Yukari YAMAUCHI

2-5 Asynchronous Advantage Actor-Critic

Asynchronous Advantage Actor-Critic(A3C)とは強化学習の学習手法の一つで、この手法は複数のエージェントが同一の環境で非同期に学習する手法である。

Asynchronous→非同期という意味でエージェントをそれぞれ複数で動かし分散学習を行うという考え。

Advantage→通常強化学習における更新では1ステップ先の行動と更新しか行わないがAdvantageは２ステップ、3ステップ先の行動価値や報酬を参照するという考え方である。

3. 提案手法

本論文では、Actor-Criticにおける決定論的方策勾配アルゴリズムにAsynchronousの考えを導入し非同期的な分散学習を考察する。決定論的な方策勾配は行動価値関数の勾配が期待でき、このアルゴリズムは通常の確率論的な方策勾配推定と違い一意に方策勾配を求めることができるため効率的に非常に効率的だと考えられる。

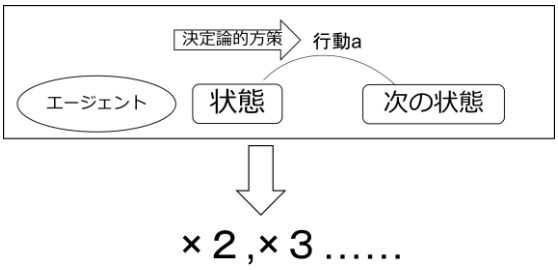

Figure 2　提案手法の概要

Figure2 は提案手法の考えを示したもので、エージェントは現在の状態から決定論的方策により一意に出力された行動に従い次の状態へ進む。このエージェントを複数生成し分散的に学習を行うことで計算時間の短縮を図る。

4. 実験および検討

実験では従来手法のA3CとActor-Criticにおける決定論的方策勾配にAsynchronousの考えを取り入れた提案手法を学習時間と精度の２つで比較する。環境は20×20のGrid World迷路問題を想定して行う。エージェントは1ステップ毎に左右上下のいずれかに必ず動き、次の行動で障害物にあたる場合はその場に留まるように設定する。また、実験に使用するパラメータはTable1に示す。

Table 1　実験パラメータ

マップサイズ	20×20
1エピソード行動回数上限	1000
エピソード上限	500
学習率α	0.30
割引率γ	0.95
ゴール報酬	10
学習調整パラメータc	0.255

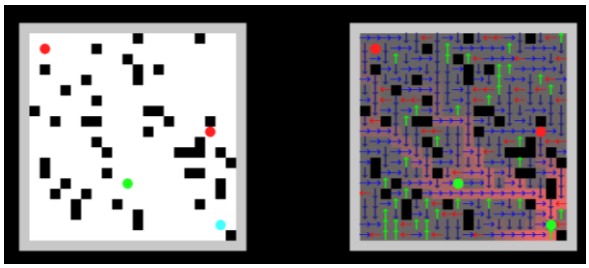

Figure 3　Grid World 迷路問題における実験環境

Figure3はVisual Studioで実装した環境であり、赤い点はスタート、緑の点はゴール、青い点が行動主体のエージェントを表している。右側の矢印はそれぞれの状態において最も行動選択確率が高い行動を示している。また、各状態には行動価値を視覚的に表現するため行動価値の高さに応じて色の濃さが変化するようにしている。

5. まとめ

学習時間、精度の2点において従来手法より提案手法の方が優れていると言える。これは提案手法の決定論的方策勾配によって状態空間にのみ期待値を取っているため、行動空間を参照しない分計算コストが少なくなり効率の良い手法であるということが分かる。また、Asynchronousの導入により非同期的に分散学習を行うため早期に収束し学習時間短縮と精度向上につながる。

参考文献

1) 上野史, 坂本充生 “マルチエージェントシステムにおける協調行動の抽象度と深層強化学習器の関係性の考察”

2) David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, Martin Riedmiller. “Deterministic Policy Gradient Algorithm”,

3) 木下直人. “Actor-Criticによるロボットの強化学習”(2005)