ISSN 2186-5647

非同期強化学習における決定論的方策勾配の導入

日大生産工（学部）　○高萩 悠　　日大生産工　山内　ゆかり

1. まえがき

強化学習[1]と呼ばれる学習手法がある。近年では強化学習はディープラーニングと組み合わせられ深層強化学習と呼ばれるものがある。深層強化学習として有名なものにDeep Q-network[2](DQN)やAlpha碁がある。これらの強化学習手法は計算資源や学習時間に膨大なリソースを消費する。

2016年にMnihらは深層強化学習のための非同期手法として、Asynchronous Advantage Actor-Critic[3]と呼ばれる学習手法を提起した。この手法は複数のエージェントを用いて非同期に探索し、行動選択の方策を評価し、その方策を更新する。複数エージェントで分散的に学習するため、学習時間をDQN等の深層強化学習手法よりも削減することに成功した。Actor-Criticの方策更新は基本的に確率論的に行われるものであり、学習の収束に時間がかかってしまう。

一方、Silverらは Deterministic Policy Gradient algorithm[4] という決定論的な方策勾配の更新方法を提起した。方策勾配を決定論的に更新することで、確率論的に更新するよりも環境を適切に探索することができる。

そこで本研究ではMnihらの非同期手法に決定論的方策勾配法を取り入れることにより、短時間でできる学習をより適切なものにする。

2. 従来研究

2.1. 強化学習

強化学習とはSuttonらが提唱した学習方法である。ある環境内におけるエージェントに行動選択をしてもらい、最終的に獲得する報酬を最大化する学習手法である。有名な手法にQ学習やTD学習がある。Q値の更新や状態価値の更新をする際には次の時間の行動価値や状態価値を使用する。

$$\Delta V(s) = \mathrm{r} + \gamma \mathrm{V}(\mathrm{s}') - \mathrm{V}(\mathrm{s}) \qquad (1)$$

$$Q(s_t, a_t) = \mathrm{Q}(s_t, a_t) + \alpha(r_t + \gamma \max_b \mathrm{Q}(s_{t+1}, b) - \mathrm{Q}(s_t, a_t)) \qquad (2)$$

式(1)はTD学習での状態価値の更新式である。これはTD誤差とも呼ばれる。$V(s)$が状態sでの状態価値であり、rはその状態の報酬であり、γは割引率である。s'は次の時間の状態である。式(2)はQ学習の更新式である。s_tはその時刻での状態であり、a_tはその時刻に実際に取った行動である。r_tはその時刻での報酬であり、αとγは学習率と、割引率である。$\max_b \mathrm{Q}(s_{t+1}, b)$とは次の時刻の状態においての最大のQ値である。

2.2. Asynchronous Advantage Actor-Critic

Asynchronous Advantage Actor-Critic (A3C)とはMnihらが提唱した強化学習の学習手法の一つである。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。

2.2.1. Actor Critic

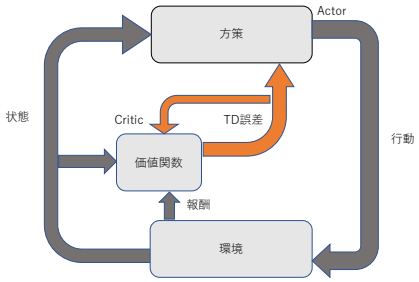

Figure 1　Actor-Critic のモデル

Figure1はActor-Criticのモデルを図にしたものである。Actorは方策によって行動を選択し、Criticは状態価値関数に応じて方策を修正する。一般的なActor-Criticは修正量としてTD誤差を利用する。このTD誤差はActorとCriticの両方を修正する。エージェントはActorによって選択された行動を起こし、環境内で1ステップ行動する。エージェントが行動を起こすと、エージェントの状態が1ステップ分変化する。この状態変化は価値関数と方策に影響を及ぼす。また、エージェントが環境から得た報酬は価値関数に影響を及ぼす。

2.2.2. Advantageな更新

Advantageとは複数ステップ先を考慮して更新をする手法である。強化学習における更新式は式(1)、式(2)のように1ステップ後の時間での行動選択価値や、状態価値を反映させるものが一般的であるが、Advantageは2ステップ、3ステップ後の行動価値や報酬を取り入れて更新をするものである。NステップのAdvantageな状態価値の更新式は式(3)のようになる。

Asynchronous Reinforcement Learning with a Deterministic Policy Gradient

Yu TAKAHAGI and Yukari YAMAUCHI

$$V(s_0) = \sum_{k=0}^{N-1} \gamma^k r_k + \gamma^N V(s_N) \qquad (3)$$

s_kはkステップ後の状態である。ある時刻の状態$V(s_0)$は$N-1$ステップまでは1ステップごとに割引された各ステップの報酬を獲得し、Nステップ目では状態価値を得る。このように1ステップだけでなくその先の数ステップを状態の更新式に取り入れたものがAdvantageの要素である。A3Cは非同期の要素とAdvantageの要素を取り入れたActor-Criticである。

2.3. Deterministic Policy Gradient algorithm

Deterministic Policy Gradient algorithm (DPG) とはSilverらが提案した決定論的方策勾配で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できる。

$$\begin{aligned} &\nabla_\theta J(\pi_\theta) \\ &= \int_S \rho^\pi(s) \int_A \nabla_\theta \pi_\theta(a|s) Q^\pi(s,a) da ds \qquad (4) \\ &= \mathbb{E}_{S\sim\rho^\pi, a\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s)\, Q^\pi(s,a)] \end{aligned}$$

式(4)は確率論的方策勾配法での方策勾配である。この式は方策勾配定理に基づいて計算されている。この定理は状態分布ρ^πがパラメータθに依存しているものの、状態分布の勾配に依存しない特徴を持っている。π_θは確率論的方策であり、Q^πは方策πによって行動をとった時の行動価値関数である。決定論的方策勾配は式(5)のようになる。

$$\begin{aligned} &\nabla_\theta J(\mu^\theta) \\ &= \int_S \rho^\mu(s)\, \nabla_\theta \mu_\theta(s) \nabla_a Q^\mu(s,a)|_{a=\mu_\theta(s)} ds \qquad (5) \\ &= \mathbb{E}_{S\sim\rho^\mu}[\nabla_\theta \mu_\theta(s) \nabla_a Q^\mu(s,a)|_{a=\mu_\theta(s)}] \\ &= \mathbb{E}_{S\sim\rho^\mu}[\nabla_\theta Q^\mu(s, \mu_\theta(s))] \end{aligned}$$

基本的には式(4)と似ているが方策が決定論的になっている。μは決定論的方策であり、状態sに対して行動aを一意に出力する。確率論的方策勾配では状態空間と行動空間の2つに対して期待値を取っているのに対して、決定論的方策勾配では状態空間のみに対して期待値を取っている。

3. 提案手法

本研究ではA3Cをより早期に学習し、環境に対してより適切な行動を選択できる強化学習手法として、A3Cに決定論的方策勾配を導入する。

決定論的方策勾配法はSilverらによって確率論的方策勾配法よりも早期に学習が収束し、環境に対して適切な行動を選択することが分かっている。そのため、A3Cでは確率論的に方策を更新しているActor-CriticにDPGを導入し、決定論的に方策を更新することでより早期に学習を収束させることを目指す。

4. 実験および検討

今回の実験での実験環境は50×50のGrid World迷路問題である。エージェントは1ステップごとに上下左右の4方向のいずれかに行動することができる。

この実験では従来手法である確率論的に方策を学習するA3Cと提案手法である決定論的に方策を学習するA3Cの2種類のA3Cにおいてどちらのほうが短時間で学習が収束するかを比較する。また、学習収束後に学習の完了した方策でエージェントに行動させたときにどちらが適切な行動をするかも比較する。

5. まとめ

本研究ではA3Cに決定論的方策勾配法を導入することでより早期に学習が収束するかを試みる。しかし、現段階では提案手法が実装できていない。

参考文献

1) Richard S. Sutton and Andrew G. Barto, “Reinforcement Learning An Introduction” MIT Press (1998)
2) Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Graves, Alex, Antonoglou, Ioannis, Wierstra, Daan, and Riedmiller, Martin. “Playing atari with deep reinforcement learning” In NIPS Deep Learning Workshop (2013)
3) Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves ,Tim Harley, Timothy P. Lillicrap, David Silver and Koray Kavukvuoglu, “Asynchronous Methods for Deep Reinforcement Learning” Proceeding of the 33rd International Conference on Machine Learning. (2016) vol.48
4) David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, Martin Riedmiller “Deterministic Policy Gradient Algorithm” Proceeding of the 31st International Conference on Machine Learning. (2014) vol.32