ISSN 2186-5647

P-28

制約更新を用いたZeroth-Order Actor-Critic

日大生産工　○片山 泰一　　日大生産工　山内　ゆかり

1. まえがき

強化学習はゲームやロボット制御、自動運転など幅広い領域で活用されている。強化学習の最も一般的な手法の一つにActor Critic法がある。この手法は方策を元にActorが行動を決め実行し、その行動によって得られた状態や報酬をCriticが環境から観測する。それをもとにActorが方策を更新するという作業を繰り返し行う。その他にもRL問題をブラックボックス最適化とみなし、一次の勾配情報を用いず、ゼロ次的な方法で最適政策を直接探索する方法があり、最近の研究ではこの手法を取り入れるものが多く存在している。

Yuheng Lei らは Zeroth Order Acter Critic(ZOAC)[1]を提案した。ZOACは一次政策評価とゼロ次政策改善を組み合わせることで、サンプル効率、最終性能、および学習されたポリシーのロバスト性において上回ったという報告がされている。

本研究では，ポリシー改善のため、RL側での制約付き更新を提案する。実験としては、迷路問題における計算機実験により提案手法とZOACを比較し、制約をかけて更新することがもたらす学習の向上、安定性について報告する。

2. 従来研究

2.1 Actor-Critic

Actor-Criticは環境に対する行動(Actor)とその行動から得られた状態と報酬から評価(Critic)し、それを元に行動の方策を更新するアルゴリズムだ。大まかな流れを図１に示す。

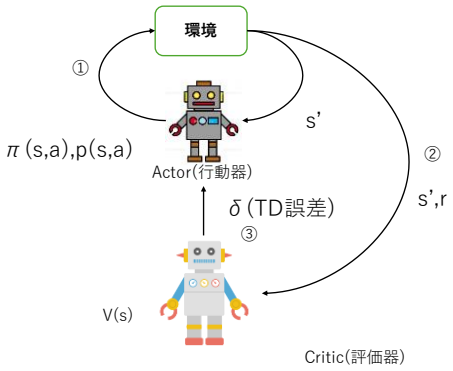

図１　Actor Critic

この手法ではTD誤差を用いる。TD誤差とは現在の状況から、次時点における状態価値を推定し、そこを目標とみなして行動し、推定値との差のことだ。この誤差を０に近づけていくことが求められる。TD誤差δ及び状態価値更新の計算式は次のように書ける。

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \quad (1)$$

$$V(s_t) = V(s_t) + \alpha\delta_t \quad (2)$$

$\alpha(0 < \alpha \leq 1)$は学習率と呼ばれ、どれだけ誤差を反映して次の状態の評価値に近づけるのかを調節する。また行動価値の更新でも TD 誤差を用いる。式は次のように書ける。

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha\delta_t \quad (3)$$

2.2 Evolution Strategy (ES)

Q 学習や方策勾配法などの強化学習の代わるものとして ES[2]がある。この手法はブラックボックス関数の最適化手法の一つで問題の内部構造や性質に関する詳細な知識がなくても最適化が可能だ。

ES は方策のパラメータ θ を直接更新、改善する。そのため行動回数に影響されず、Back Propagation が不要である。次に方策ポリシー θ の更新に必要な手順と式を説明する。

まず始めに摂動ε_iをランダムに初期化する。この時パラメータの次元と同じ次元数を持たせる。 ε はポリシーのパラメータにノイズを加えるために使用され$N(0,1)$に従う乱数である。

次にポリシー θ に摂動を加え候補パラメータを生成する、その時の計算は次のように書ける。

$$\theta_i = \theta + \sigma * \varepsilon_i \quad (4)$$

σ はノイズをどれだけ反映するかの調整をしている。その後生成した各パラメータθ_iに基づいてエピソードを実行し、確率的な累積報酬F_iを得て θ を更新する。式は次のように書ける。

$$F_i = F\theta_i \quad (5)$$

$$\theta_{t+1} = \alpha \frac{1}{n\sigma} \sum_{i=1}^{n} F_i \varepsilon_i \quad (6)$$

ここで n はサンプル数である。

2.3 Zeroth-Order Actor-Critic (ZOAC)

このアルゴリズムはタイムステップ毎に摂動を伴うロールアウト収集、一次政策評価(PEV)とゼロ次政策改善(PIM)を各反復で交互に行う。

Zeroth-Order Actor-Critic with Constrained Update

Taichi KATAYAMA and Yukari YAMAUCHI

行動政策$V^{\beta}(s)$の状態値関数は目標値と出力結果の誤差を最小化することを目標とする。目標値は各反復において計算され、計算式は以下のように書ける。

$$\hat{G}_t = V_w(S_t) + \sum_{k=0}^{T-t-1} (\gamma\lambda)^k r_{t+k} + \gamma V_w(S_{t+k+1}) - V_w(S_{t+k}) \quad (7)$$

この時0<λ<1は、目標値のバイアスと分散のトレードオフを制御するハイパーパラメータである。PEVの目的関数は次のように書ける。

$$J_{critic}(W) = E_{(s,\hat{G})}[\frac{1}{2}(V_w(s) - \hat{G})^2] \quad (8)$$

次にPIMを行う。ここではアドバンテージ関数を用いて、行動が現在のポリシーに対してどれほど優れているかを測定し、行動選択の評価を補助する。アドバンテージ関数は次のように書くことができる。

$$\hat{A}_N^{\pi_{\theta+\sigma\varepsilon_{i,j}}} = \sum_{k=0}^{N-1} (\gamma\lambda)^k [r_{i,jN+k} + \gamma V_w(s_{i,jN+k+1}) - V_w(s_{i,jN+k})] \quad (9)$$

ゼロ次勾配は、サンプリングされたランダム方向の加重和として推定することができ次のように書ける。

$$\nabla_\theta J_{actor}(\theta) \approx \frac{1}{nH\sigma} \sum_{i=1}^{n} \sum_{j=0}^{H-1} \hat{A}_N^{\pi_{\theta+\sigma\varepsilon_{i,j}}} \varepsilon_{i,j} \quad (10)$$

Actor側では摂動を用いるがCritic側では摂動を用いずに計算している。

2.4 Proximal Policy Optimization (PPO)

方策勾配法方策パラメータの更新方向はわかっても、更新幅がわからないという問題があった。この問題を解決するために更新幅を制約して方策ネットワークをより安定的に学習することが求められた。PPO[3]はKL制約付きパラメータ更新を使用して更新幅を制約する手法である。

このアルゴリズムでは代理目的関数(Surrogate Objective)にクリッピング処理を行う。代理目的関数の計算では更新前の方策と更新後の方策の変化比が含まれる。計算式は以下のように書ける。

$$L^{CPI}(\theta) = \hat{E}_t\left[\frac{\pi_\Theta(a_t|s_t)}{\pi_{\Theta_{old}}(a_t|s_t)}\hat{A}_t\right] = \hat{E}_t[r_t(\theta)\hat{A}_t] \quad (11)$$

代理目的関数をクリッピングするClipped Surrogate Objectiveでは式(11)を(1-ε)以下、(1+ε)以上にならないようにクリッピングを行う。その計算を下に示す。

$$L^{CLIP}(\theta) = \hat{E}_t[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t] \quad (12)$$

クリップ処理を行うことで上限、下限両方で極端な更新を回避することができる。

3. 提案手法

方策勾配法では方策パラメータの更新方向はわかっても、適切な更新幅を決めることが困難なため、大きく更新してしまうと方策ネットワークが局所解に陥る不安定さ、逆に更新が全く進まず、学習が効率的に進まない可能性がある。そこでPPOのクリッピング処理を取り入れ、制約をかけて更新を行うことを提案する。

4. 実験および検討

実験環境は迷路問題を使用する。比較にはActor Critic、ZOAC、提案手法をそれぞれ実行する。検討には、行動回数、実行時間、収束の安定性を用いる。

5. まとめ

提案手法を取り入れることで、更新幅を決めるこができ、局所的な解に陥ることを防ぎつつ、適切な更新サイズを保証することでパラメータの劣化を防ぐことが期待される。また、さらなる学習効率の向上と安定化を図ることができるのではないかと考えている。

参考文献

[1] Yuheng Lei and Jianyu Chen and Shengbo Eben Li1 and Sifa Zheng ”ZEROTH-ORDER ACTOR-CRITIC”,(2022)

[2] Tim Salimans and Jonathan Ho and Xi Chen and Szymon Sidor and Ilya Sutskever” Evolution Strategies as a Scalable Alternative to Reinforcement Learning”(2017)

[3] John Schulman and Filip Wolski and Prafulla Dhariwal and Alec Radford and Oleg Klimov” Proximal Policy Optimization Algorithms”(2017)