ISSN 2186-5647

1 次政策評価と 0 次政策改善による Soft Actor-Critic の提案

日大生産工　○小渕 敏生　　日大生産工　山内　ゆかり

1. まえがき

強化学習はビデオゲーム、ロボット制御、自動運転など、幅広い領域で成功を収めている。代表的なアルゴリズムとしてActor Critic法がある。

そのほかに強化学習問題をブラックボックス最適化とみなし、0次的に、すなわち1次の勾配情報を用いずに最適な方針を直接探索する方法がある。0次最適化手法の例としてEvolution Strategies:ES[1]やGenetic Algorithm:GAをディープニューラルネットワークに適用するものがあり、それらは一般的な強化学習とも競争力があることが示されている。Yuheng Lie らは1次強化学習手法と0次強化学習手法をActor Criticに統合したZEROTH-ORDER ACTOR-CRITIC(ZOAC)[2]を提案し、効率、性能、ロバスト性において0次および1次のベースラインアルゴリズムを上回ることが報告がされている。しかし従来研究で使用されているon-policyなActor-Criticでは勾配更新のたびに新たなサンプルを必要とするためサンプル効率が悪いという問題がある。

本研究では，サンプル効率が悪いという問題において1次政策評価と0次政策改善によるSoft Actor-Critic[3]:Soft ZOACを提案し、迷路問題おける計算機実験により提案手法とSoft ZOACを比較し、実際にサンプル効率や学習の安定性ついて報告する。

2. 従来研究

2-1 Actor Critic

Actor Criticは実際の行動を決定し実行する行動器(アクター)とアクターを評価する評価器(クリティック)によって学習を行う手法である。Actor Criticの大まかな流れを図1で示す。

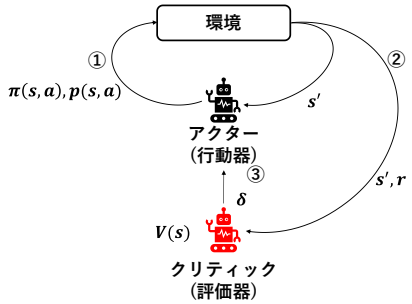

図 1. Actor Critic

クリティックでは得られた報酬や遷移先の状態を用いてTD誤差δを以下の式で求める。

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

評価値の更新は以下の式となる。

$$V(s_t) = V(s_t) + \alpha(r_{t+1} + \gamma V(s_{t+1}) - V(s_t)) \quad (2)$$

アクターでは行動優先度p(s,a)をクリティックから得られるTD誤差を用いて更新する。更新する式は以下の式となる。

$$p(s,a) = p(s,a) + \alpha\delta \quad (3)$$

2-2 進化戦略(ES)

Q学習や方策勾配法などの強化学習の方法に新たに変わるものとしてESがある。これはブラックボックス関数の最適化手法の一つである。ESの概要を図2として以下に示す。

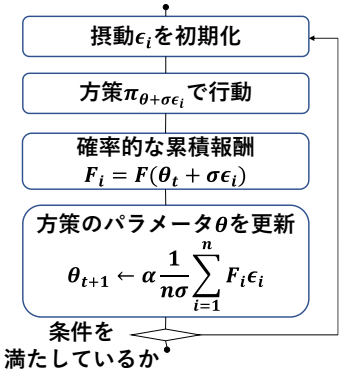

図 2. ES

ESの特徴として方策のパラメータθに対してn個の摂動ϵ_iを与え行動をすることで直接θを改善できる。上記の点からESではゴールするまでの行動回数の長さの影響を受けないことやBack Propagationが不要という利点がある。

一方で摂動による近傍探索を行うので学習の収束まで時間がかかるという点がある。

2-3 ZEROTH-ORDER ACTOR-CRITIC

全体的な枠組みとして初めにロールアウト収集を行う。ロールアウト収集では並列化されたエージェントに対して新しいランダムな方向$\epsilon_{i,j}$がサンプリングされ、行動方策が摂動する。

次に1次政策評価を行う。長さTの軌道において、各目標値$\hat{G}_t$は状態s_tで以下のように求める。

$$\hat{G}_t = V_W(s_t) + \sum_{k=0}^{T-t-1} (\gamma\lambda)^k [r_{t+k} + \gamma V_W(s_{t+k+1}) - V_W(s_{t+k})] \quad (4)$$

政策評価の目的関数は次のように求める。

Soft Actor-Critic with first-order policy evaluation and zero-order policy improvement

Toshiki OBUCHI and Yukari YAMAUCHI

$$J_{critic}(\mathrm{w}) = \mathbb{E}_{(s,\hat{G})}\{\frac{1}{2}[V_W(\mathrm{s}) - \hat{G}]^2\} \quad (5)$$

最後に0次政策改善を行う。アドバンテージ関数は以下のように書くことができる。

$$\hat{A}_N^{\pi_{\theta+\sigma\epsilon_{i,j}}} = \sum_{k=0}^{N-1}(\gamma\lambda)^k [r_{i,jN+k} + \gamma V_W(s_{i,jN+k+1}) - V_W(s_{i,jN+k})] \quad (6)$$

0次勾配は次のように求める。

$$\nabla_\theta J_{actor}(\theta) \approx \frac{1}{nH\sigma}\sum_{i=0}^{n}\sum_{j=0}^{H-1}\hat{A}_N^{\pi_{\theta+\sigma\epsilon_{i,j}}}\epsilon_{i,j} \quad (7)$$

ZOACの特徴としてCritic側では摂動を使用しない更新を行い、Actor側では摂動を使用した更新を行う。

2-4 Soft Actor-Critic(SAC)

通常の強化学習では以下の式で期待報酬を最大化することを目的としているがSACでは、期待報酬にエントロピー最大化項を加える。SACでの目的関数を式(8)として以下に示す。

$$J(\pi) = \sum_{t=0}^{T}\mathbb{E}_{(s_t,a_t)\sim p_\pi}[r(s_t,a_t) + \alpha H(\cdot|s_t)] \quad (8)$$

ここでαは温度パラメータで、どのくらいエントロピー項Hを考慮するかを決めるハイパーパラメータになる。式(8)を最大化することで、サンプル効率・安定性を兼ね備えたアルゴリズムを実現している。

3. 提案手法

on-policyなアルゴリズムでは勾配更新のたびに新たなサンプルを必要とするため、サンプル効率が悪いことが挙げられる。一方でoff-policyなアルゴリズムではサンプル効率は改善されるものの既存のアルゴリズムでは学習が不安定であることが問題に挙げられる。

そこで高いサンプル効率、より多彩な探索による学習の安定化を実現したSACをActor Criticの代わりに使用する手法を提案する。

4. 実験および検討

本研究で使用する学習環境のマップを図3に示す。

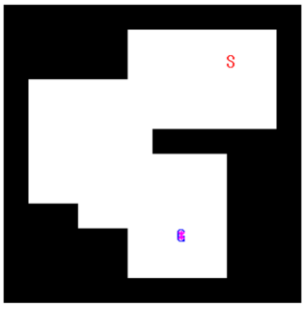

図 3. 実験環境

9×9の迷路を使用しスタートからゴールまで探索を繰り返す。ゴール時の報酬は1.0とし、ゴールをしていない行動に-0.01の報酬を与えている。ESとSACの総行動回数の結果を表1として以下に示す。

表 1. 総行動回数

	ES	SAC
総行動回数	585113	145212

総行動回数はESよりSACのほうが少なく短い距離で報酬を得ていることがわかる。10エピソードにおける平均行動回数を図4として以下に示す。

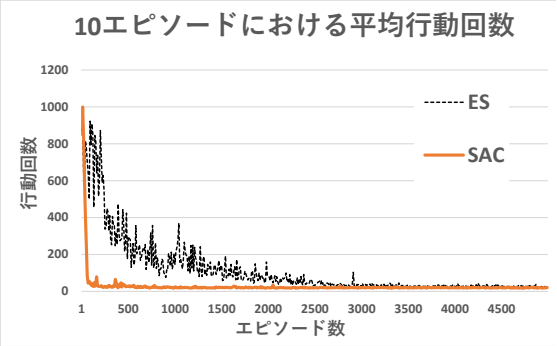

図 4.10 エピソードにおける平均行動回数

図4より学習初期の行動回数がESは多くSACは少なく学習初期における学習精度の違いがあることが分かる。学習最後のほうでは同じような精度があることが分かるので2つには学習初期における学習性能の差がある。

5. まとめ

本研究では、ZOACに対してSACを合わせる1次政策評価と0次政策改善によるSoft Actor-Criticの提案を行った。さらにESをZOACにして、SACを組み合わせることでさらなる迷路探索問題における学習精度が上がると考えている。

参考文献

[1] Tim Salimans, Jonathan Ho, Xi Chen, Symon Sidor, Ilya Sutskever, “Evolution Strategies as a Scalable Alternative to Reinforcement Learning” , arXiv preprint arXiv:1703.03864, 2017.

[2] Yuheng Lie, Jianyu Chen,Shengbo Eben Li, Sifa Zheng, “ZEROTH-ORDER ACTOR-CRITIC”, (2022)

[3] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, Sergey Levine, ” Soft Actor-Critic:Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor”,(2018)