4P-06

マルチエージェント強化学習における進化的探査率調整法とその拡張

岡野 拓哉† 野田 五十樹‡
東京工業大学 †‡
(独) 産業技術総合研究所 †‡

1 はじめに

本稿では, 我々が以前提案した, 動的環境のマルチエージェント強化学習 (以下,MARL) における探査率の調整法に関して、資源共有問題の一種である待ち行列問題を用いて実験的に評価を行った.

シングルエージェントによる強化学習と同様に, 探査率は MARL の性能を大きく左右する重要なパラメータである. このパラメータの調整法に関しては, 数多く研究されてきた. 多くの研究はシングルエージェントのみを対象にした探査率の調整法や, 静的な環境での調整法などであった.

しかし, エージェント学習を実社会の問題に応用するとした場合には, 環境が動的であり, 複数のエージェントによる同時学習の場合も考慮しなければならない. そのため, 動的な環境下の MARL における探査率の調整法が, 必要不可欠である. ところが, そのような探査率の調整法は未だに確立していない. そこで, 本稿では, 動的環境の MARL の探査率調整法を提案し, 評価を行う. 本稿では全エージェントは $\epsilon - greedy$ 選択を用いて行動選択をすることを仮定する. つまり, 各エージェントの探査率である ϵ を調整することとする.

2 Win or Update Exploration Rate

我々は以前, 探査率の調整法として"Win or Update Exploration rate"(以下,WoUE)[1] を提案した. WoUE は全体平均の報酬を全エージェントに通知することにより, 各エージェントが自律的に探査率を調整する.

Investigation of Evolutionarily Adaptation of Exploration Rate in Multi-agent Reinforcement Learning and extends

†Takuya Okano,Tokyo,Tech,AIST
<okano.t.ac@m.titech.ac.jp>
‡Itsuki Noda,Tokyo,Tech,AIST <i.noda@aist.go.jp>

各エージェントは全エージェントの報酬の分布の分散を最小化するように探査率更新する. 一定期間の全エージェントの報酬の分布の分散 σ_G^2 は以下のように定式化できる.

$$\sigma_G^2 = \frac{1}{N}\sum^{N}(\mu_G - r_i(\epsilon_i))^2 \tag{1}$$

エージェント数を N, G は一定期間の全エージェントの平均報酬の分布, μ_G を G の平均, ϵ_i は一定期間でのエージェント i の探査率, $r_i(\epsilon_i)$ はエージェント i が探査率 ϵ_i 一定期間学習行動を行い得た報酬 $r_i(\epsilon_i)$. ここで $r_i(\epsilon_i)$ を 知識利用による報酬 r_{ia}, 探査による報酬 r_{ib} で分解すると,

$$r_i(\epsilon_i) = (1-\epsilon_i)r_{ia} + \epsilon_i r_{ib} \tag{2}$$

と表現できる. 式 1 に式 4 を代入しエージェント i の探査率 ϵ_i により微分し, 報酬の分布の探査率に対する勾配を求めると以下のようになる

$$-\frac{\partial\sigma_G^2}{\partial\epsilon_i} = \frac{2}{N}(\mu_G - r_i(\epsilon_i))(r_{ib} - r_{ia}) \tag{3}$$

探査率を更新するエージェントはこの勾配を用いて, 分散の最小化を目指し, 探査率を更新する. 従って, エージェント i の探査率の更新量 $\Delta\epsilon_i$ は以下のように設定する

$$\Delta\epsilon_i \leftarrow (\mu_G - r_i(\epsilon_i))(r_{ib} - r_{ia}) \tag{4}$$

この式 4 は各エージェントにとっての外部の情報は全エージェントの平均報酬であるため, 更新フェーズに全エージェントに平均報酬を通知するだけで各エージェントは探査率を更新することが可能である. そして, 本手法では平均以下の報酬を得ているエージェントのみが探査率を更新するため $(\mu_G - r_i(\epsilon_i))$ は必ず正の値である. よってより不利なエージェントが積極的に探査率を更新することを示す. また,$(r_{ib} - r_{ia})$ により, 知

識利用より探査をした時のほうが報酬が大きくなれば, 探査率を増加させ, 逆に探査より知識利用による報酬が大きい場合は, 探査率を減少させる作用がある. これにより, 環境の変化に応じて探査率を更新することができる.

Algorithm 1 は本手法のアルゴリズムである.

Algorithm 1 WoUE

```
Initialze N agents
for cycle = 1 → end_cycle do
    for all Agents do
        select action, using ε-greedy
    end for
    evaluate all agent reward
    if interval MOD cycle = 0 then
        broadcast μ_G to each agent
        for all Agents do
            if r(ε_t) < μ_G then
                calculate Δε by (4)
                ε_{t+1} ← ε_t + TΔε
            end if
        end for
    end if
end for
```

3 実験結果

資源共有問題の報酬関数を待ち行列問題にて用いる関数に設定し実験的に評価を行った. 結果として, 本手法によって環境の変化率に応じて, 探査率が推移することを確認した. 本実験では様々な環境の変化率にて実験を行った。環境の変化率別の探査率の推移を図 1 に示す. 本手法による探査率の推移は環境の変化率が高く, 高い探査率が必要な時に比較的高い探査率を得ることができていることがわかる. また, 環境の変化率が低く, 小さな探査率が必要な時には徐々に小さな探査率に収束することがわかる.

4 終わりに

本稿では, 我々が以前提案した動的環境のマルチエージェント強化学習における探査率調整法 WoUE を待ち行列問題用いて実験的に評価を行った. WoUE は平均報酬以下の報酬を得ているエージェントのみが探査率を更新する探査率調整法である.

実験結果として, 環境の変化率に応じて探査率を調整することが可能であることがわかった. つまり, 高い探査率が必要な環境では比較的高い探査率, 低い探査率が必要な環境では比較的低い探査率へ探査率が推移することを確認した.

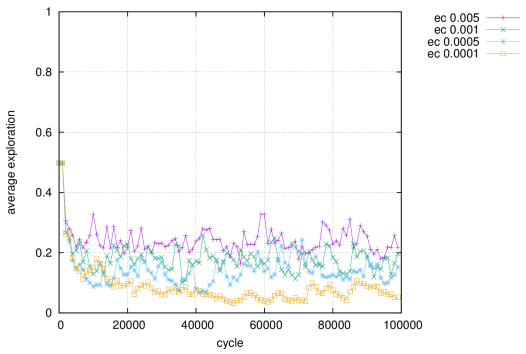

図 1: 環境の変化率別の探査率の推移.

参考文献

[1] T.Okano and I.Noda:Investigation of Evolutinarily Adaptation of Exploration Rate in Multi-agent Reinforcement Learning .*International Symposium on Artificial Life and Robotics*,(2016) 発表予定