2R-08

Sarsaエージェントによる囚人のジレンマゲームでの相互協調の継続回数

百武佳輝 *1　森山甲一 *1　武藤敦子 *1　松井藤五郎 *2　犬塚信博 *1
*1 名古屋工業大学　*2 中部大学

1 はじめに

囚人のジレンマゲーム [1] で長期に渡って報酬を最大化するには相互協調が継続することが必要である。しかし強化学習エージェントのように自分の報酬の最大化を目的とする場合、協調を選択することは稀である。

これまで強化学習の手法の一つである Q 学習エージェントについて、相互協調発生後の価値関数の挙動から相互協調の継続回数について議論がされてきた [2]。ブートストラップ型強化学習で行動価値を学習する手法は方策オフ型の Q 学習と方策オン型の Sarsa に分けられる [3] ため、本研究では Sarsa における相互協調の継続回数を求めることで、行動価値を学習するブートストラップ型強化学習の囚人のジレンマゲームにおける挙動を議論する第一歩とする。

2 準備

2.1 Sarsa

行動選択を ε-greedy 法で行う状態数 1 の Sarsa のアルゴリズムは以下の通りである。

1. 行動 A_t を ε-greedy 法で決定する
2. 報酬 R_t を獲得する
3. 次の行動 A_{t+1} を ε-greedy 法で決定する
4. Q 値を下記の式で更新する
   $Q_{t+1}(A_t) = Q_t(A_t) + \alpha\left(R_t + \gamma Q_t(A_{t+1}) - Q_t(A_t)\right).$
5. 行動 A_t を A_{t+1} に更新して 2 に戻る

プロセス 3,4 に注目すると、Q 値の更新前に次の行動を決定しなくてはならないことに注意を要する。

2.2 囚人のジレンマゲーム

囚人のジレンマゲームとはゲーム理論で扱う 2 人 2 行動ゲームの 1 つである。2 人の取れる行動は協調 (C) と裏切り (D) の 2 つであり、2 人の行動の組み合わせによって利得 T, R, P, S のいずれかが与えられる (表 1)。ただし、$T > R > P > S$, $2R > T + S$ である。

表 1: 利得表

A \ B	C	D
C	R, R	S, T
D	T, S	P, P

3 相互協調の継続回数

3.1 導出条件・導出手順

相互協調の継続回数を求めるために、まずその崩壊を定義する。Sarsa は行動選択が 1 ステップ前の Q 値を用いて行われるため、1 回 $Q(D)$ の値が $Q(C)$ の値より大きくなるだけでは、次の行動が協調になる可能性がある。従って相互協調の崩壊を「相互協調中に 2 回連続して裏切りを選択する」と定義する。また下記の条件を設定する。

1. 自分が裏切りを選択する 1 ステップ前の $Q(C)$ は最大値とする
2. 相互協調開始時の $Q(D)$ の値を $P/(1-\gamma)$ とする

以下、相互協調が発生してから相互協調が崩壊するまでのゲームのプレイ回数の期待値を求める。導出の手順は以下の通りである。

1. 1 回の裏切りで増加する $Q(D)$ の変化量を求める
2. $Q(D)$ が増加して相互協調が崩壊するまでの裏切りの選択回数を求める
   (a) 自分が裏切りを選択して崩壊するとき
   (b) 相手が裏切りを選択して崩壊するとき
3. ゲームのプレイ回数の期待値を求める

3.2 相互協調が崩壊するまでのゲームのプレイ回数

l 回目の裏切りが起こる時刻を $t+m_l$ と置く。この時の $Q(D)$ の変化量 $\Delta Q_l(D)$ は次のように求められる。

$$\Delta Q_l(D) = \alpha(1-\alpha)^{l-1}\left(T + \gamma Q_t(C) - \left(1 - \frac{\alpha}{1-\alpha}\gamma^2\right)Q_t(D)\right). \tag{1}$$

また 1 回裏切りが起こってから再び裏切りが発生するまでのゲームのプレイ回数の期待値 n は幾何分布の期待値より以下で求められる。

$$n = \frac{4}{\varepsilon(2-\varepsilon)}. \tag{2}$$

Length of mutual cooperation in Prisoner's Dilemma Games played by Sarsa agents

Yoshiki Momotake*1, Koichi Moriyama*1, Atsuko Mutoh*1, Tohgoroh Matsui*2 and Nobuhiro Inuzuka*1
*1Nagoya Institute of Technology, Nagoya 466-8555, Japan
*2Chubu University, Kasugai 487-8501, Japan

3.2.1 自分の裏切りで相互協調が崩壊するとき

相手は常に協調を選択すると仮定する。この時、$Q_{t+m_l+1}(D) > Q_{t+m_l+1}(C)$ となる l を求める。

$$l > 1 + \frac{1}{\log(1-\alpha)} \log \frac{(1-\alpha\gamma)\left(Q_t(D)+X\right) - \frac{1-\alpha\gamma}{1-\gamma}R}{(1-\alpha-\alpha\gamma)X}. \quad (3)$$

ただし $X \equiv (T + \gamma Q_t(C) - (1 - \frac{\alpha}{1-\alpha}\gamma^2)Q_t(D))$ とする。l は自然数のため、最小値 l_{min} は式 (3) の右辺を L とおくと $l_{min} = \lceil L \rceil$ となる。

自分の裏切りで1回 Q 値が反転すると $Q(D)$ は $Q(C)$ の最大値より大きくなるため、次に協調を選択しても再反転することはなく、以降は裏切りを選択し続けることになり、相互協調が崩壊する。従って反転してから3ステップで相互協調が崩壊するため、相互協調の継続回数 n_l は以下の通りである。

$$n_l = n \times l_{min} + 3. \quad (4)$$

3.2.2 相手の裏切りで相互協調が崩壊するとき

相手は相互協調が崩壊する直前に一度だけ裏切りを選択すると仮定する。自分が l' 回裏切った後、k ステップ目に相手に裏切られた時に $Q_{t+m_{l'}+k+1}(D) > Q_{t+m_{l'}+k+1}(C)$ となる l' を求める。

$$l' > 1 + \frac{1}{\log(1-\alpha)} \log \left(\frac{Q_t(D) + X - \frac{1-\alpha+\alpha\gamma}{1-\gamma}R - \alpha S}{(1-\alpha-\alpha\gamma(1-\alpha+\alpha\gamma)^{k-1})X} - \frac{\alpha\gamma(1-\alpha+\alpha\gamma)^{k-1}\left(Q_t(D) + X - \frac{R}{1-\gamma}\right)}{(1-\alpha-\alpha\gamma(1-\alpha+\alpha\gamma)^{k-1})X} \right). \quad (5)$$

l' は自然数のため、最小値 l'_{min} は式 (5) の右辺を L' とおくと $l'_{min} = \lceil L' \rceil$ となる。

相手の裏切りで一度 Q 値が反転してからは協調と裏切りが交互に選択される可能性がある。反転した時刻を τ として、$\tau+1$ で再び相互協調になる条件は以下の通りである。

$$Q_\tau(D) < \frac{1-\alpha}{1-\alpha\gamma}\left(\left(\frac{1}{1-\gamma} + \frac{\alpha^2}{1-\alpha}\right)R + \alpha S\right). \quad (6)$$

相手が協調を選択すると仮定したため、$Q(D)$ が条件 (6) を満たす場合、次のステップで協調を選択してから2回連続で裏切りを選択し、相互協調は崩壊する。満たさない場合は以降の行動で裏切りを選択する。従って反転してから最大3ステップで相互協調は崩壊するため、相互協調の継続回数 n'_l は以下の通りである。

$$n_{l'} = n \times l'_{min} + k + 3. \quad (7)$$

3.3 期待値の導出

相手は行動を ε-greedy 法で決定するとする。$l_{min} > l'_{min}$ のため、自分が l'_{min} 回裏切り後に l_{min} 回裏切るまでの間に相手に裏切られなければ自分の裏切りで相互協調が崩壊する。ゲームのプレイ回数を Δn とすると

$$\Delta n = n_l - n_{l'} \quad (8)$$

表 2: 実験結果

	計算結果	実験結果 1	実験結果 2	実験結果 3
l_{min}	6	1.72	2.49	4.05
l'_{min}	3	1.90	2.42	3.44
n	21.1	16.2	17.6	17.2
p	0.039	0.112	0.103	0.085
$1-p$	0.961	0.888	0.897	0.915
N	69.6	42.5	58.2	75.8

より、自分の裏切りで相互協調が崩壊する確率 p は

$$p = \left(1 - \frac{\varepsilon}{2}\right)^{\Delta n}. \quad (9)$$

式 (4), 式 (7), 式 (9) より相互協調の継続回数の期待値 N は、

$$N = p \times n_l + (1-p) \times n_{l'}. \quad (10)$$

4 実験結果

$T = 5,\ R = 4.3,\ P = 1,\ S = 0,\ \alpha = 0.25,\ \gamma = 0.1,\ \varepsilon = 0.1$ とし、2つの Sarsa エージェントを用いて 1000 回 ×1000 試行実験を行った結果を表 2 に示す。

各実験結果は同一の実験から取得したデータをまとめたものである。実験結果 1 はすべての実験結果を、実験結果 2 は最初の裏切り選択時に導出条件 1 を満たした時のみの実験結果を、実験結果 3 は導出条件 1,2 を満たした時の実験結果をまとめている。

実験結果 1 で $l_{min} < l'_{min}$ と計算結果と異なる結果が得られたのは、相互協調が崩壊した直後に相互協調が発生し、$Q(C)$, $Q(D)$ が減少しないために l_{min} のみが減少したためである。また N が計算結果より大きく下回ったのは、相互協調が発生して $Q(C)$ が上がりきらないで崩壊するために n_l, n'_l が小さいことや崩壊後に $Q(D)$ の値が下がりきらないで相互協調が発生したためであると考えられる。

5 おわりに

本研究では、Sarsa エージェントによる囚人のジレンマゲームにおいて、相互協調が発生してから崩壊するまでのゲームのプレイ回数の期待値を解析的に求めた。

謝辞

本研究の一部は、JSPS 科研費 JP19K12118 の助成を受けて行われた。

参考文献

[1] W. Poundstone, 松浦俊輔他訳, “囚人のジレンマ”, 青土社, 1995.

[2] K. Moriyama et al., “The Resilience of Cooperation in a Dilemma Game Played by Reinforcement Learning Agents”, *Proc. IEEE ICA*, 2017.

[3] R. S. Sutton and A. G. Barto, *Reinforcement Learning*, 2nd ed., MIT Press, 2018.