1L5-1

強化学習を用いた繰り返しゲームにおける戦略の学習の高速化

Acceleration of learning strategies in repeated games using reinforcement learning

藤田 渉 *1　森山 甲一 *2　福井 健一 *2　沼尾 正行 *2
Wataru Fujita　Koichi Moriyama　Ken-ichi Fukui　Masayuki Numao

*1大阪大学 大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University

*2大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

In the real world, people interact with each other in various ways. To analyze such interaction, it is usually modeled as games. Reinforcement learning algorithms are widely studied to obtain suitable strategies in games. However, existing algorithms require a lot of interactions, i.e., they learn slowly. In this work, we construct an algorithm that learns quickly as well as maximizes payoffs in various games. It adopts two different algorithms in the early and late stages, respectively. We have conducted an experiment in which the proposed agents played ten kinds of games in self-play and with other agents. The result shows that the proposed algorithm learns quickly and gains large payoffs in nine games.

1. 序論

人間は社会において生きていく中で様々な意思決定を行う．そして，人々の意思決定は互いに影響しあい，複雑に干渉し合っている．そのような社会的状況では，人々の間には利害の対立や競争，協力など多様な相互関係が混在する．このような人々の関係を研究するために，それらをモデル化した「ゲーム」を定義し，ゲームにおいて合理的な意思決定を分析する「ゲーム理論」が広く研究されている．

また，人間は状況や局面に応じて試行錯誤を重ねて，学習する．自分の欲求を満たす状況を望み，不快な状況から逃げようと試みる．このような人間の思考をモデル化したものに「強化学習」がある．

本研究では，「相互に干渉しあう複数の個体」が「自己の利益となる行動を学習する」状況をモデル化するために，強化学習アルゴリズムを搭載したエージェントがゲームを行う状況を考える．様々なゲームにおいて高い報酬を獲得することができるアルゴリズム [1] [7] も存在するが，学習するために多くの試行錯誤を必要とする．現実世界は多様で複雑な状況で構成されているため，どのような状況においても素早く学習し，合理的な判断を下さなくてはならない．したがって，どのようなゲームにおいても利用できる範囲が広く，高速な学習が可能なアルゴリズムが必要である．

本研究では，様々なゲームにおいて自己の利益が最大となる行動を素早く学習できる強化学習アルゴリズムを構築することを目的とする．

2. 関連研究

人々の意思決定の相互干渉をモデル化した「ゲーム」と，試行錯誤を重ねることによって自己を取り巻く状況に適応する学習手法「強化学習」について紹介する．

連絡先: 藤田 渉，大阪大学産業科学研究所沼尾研究室，
〒 567-0047 大阪府茨木市美穂ヶ丘 8-1，
Tel: 06-6879-8426，Fax: 06-6879-8428，
E-mail: fujita@ai.sanken.osaka-u.ac.jp

2.1 ゲーム

人間は自己の欲求を満たすように行動するために意思決定を行う．しかし，ある一人の人が目的を達成できるかどうかはその人自身の意思だけではなく周囲の人々の意思決定にも依存している．社会における様々な意思決定の相互関係はゲーム理論 [6] として数理的に広く研究されている．

ゲーム理論では，複数の意思決定主体がそれぞれの目的を達成するために相互依存している状況を「ゲーム」と定義し，これを解析する．この「ゲーム」において意思決定を下すプレイヤーは自身を取り巻く状況に応じて行動を選択する．プレイヤーは自己の戦略に則って行動し，プレイヤー同士の行動の組み合わせによって各々のプレイヤーに与えられる利得値が決まる．自己の利得は自分の意思決定だけでなく他のプレイヤーにも依存しているので，自己の利得を最大化するためには，他のプレイヤーの行動を充分に考慮しなければならない．

2.2 強化学習

強化学習 [4] は，現在の自己の状態を観測し，受け取った報酬から自分が取るべき行動を決定する問題を扱う学習手法である．意思決定主体をエージェントと呼び，エージェントの外部全てから構成される制御対象を環境と呼ぶ．強化学習は一般的に多くの試行錯誤を必要とし，行動の収束が遅いために現実世界の事象に当てはめることが難しい．まして，他のエージェントと相互に干渉し合う場合には，より適用することが難しくなる．

3. 提案手法

本研究では，既存の強化学習アルゴリズムよりも素早く，高い報酬を獲得するために，2 つのアルゴリズムが得意とする部分をそれぞれが担当し，ゲームの途中でアルゴリズムを切り替える手法を提案する．提案手法のアルゴリズムを J-algorithm (J-alg) と呼称する．J-alg は内部に 2 つのアルゴリズムを持ち，ゲームのラウンド数に応じて 2 つの区分を持つ．1 つが探索パートを担当する S' algorithm, もう一つが定常パートを担当する BM algorithm [7] である．次にそれぞれのアルゴリズムについて述べる．

3.1 S' algorithm

S' algorithm (S'-alg) は Satisficing algorithm (S-alg) [3] の「相手プレイヤーとの協調行動を素早く学習する」という長所に着目し，そこに今回の提案手法のアルゴリズムの探索パートを担うように手を加えたアルゴリズムである．S'-alg は自分の満足度 (aspiration level) を報酬によって更新し，満足度を超える報酬が得られる状態に留まろうとするアルゴリズムである．ここで，相手プレイヤーが自分にとって最も報酬が少なくなるような行動をとった時に自分が最大の報酬を獲得するような戦略をマキシミニ戦略 π^{MM} と定義し，この戦略をとった時に獲得できる報酬をマキシミニ値 v^{MM} とする．S'-alg は以下の通りである．

1. 満足度の初期値 (α^0) を報酬の最大値 R_{max} から $2R_{max}$ の間の値にランダムに設定する
2. 次を繰り返す
   (a) 行動 a^t を選択

$$a^t \longleftarrow \begin{cases} a^{t-1} & \text{if} \quad (r^{t-1} \geq a^{t-1}), \\ b & \text{if} \quad (r^{t-1} = v^{MM}), \\ \text{ランダム行動} & \text{otherwise.} \end{cases} \tag{1}$$

   ただし，b は確率 p で a^{t-1} 以外の行動をランダムに選択，$(1-p)$ で全ての行動からランダムに選択する.$(0 \leq p \leq 1)$

   (b) 報酬 r^t を受け取り，満足度 α^t を更新

$$\alpha^{t+1} \longleftarrow \lambda\alpha^t + (1-\lambda)r^t. \tag{2}$$

ここで t は時刻を表し，$\lambda \in (0,1)$ は学習率である．

3.2 BM algorithm

BM algorithm (BM-alg) [7] は M-Qubed [1] と S-alg [3] を Boltzmann multiplication [5] を用いて組み合わせたアルゴリズムである．得意不得意が相補的な 2 つの異なるアルゴリズムを組み合わせることによって，最終的な利得を最大化する戦略を学習することを目的とする．エージェントが選択した行動に基づいて，組み合わせた 2 つのアルゴリズムが同時に行動の価値や満足度の更新を行う．Boltzmann multiplication は構成するアルゴリズム j の方策 π_j^t に基づいて，各行動の選択確率を乗算して，ボルツマン分布によりエージェントの方策を決定する．各行動の選好度は

$$p^t(s^t, a[i]) = \prod_j \pi_j^t(s^t, a[i]) \tag{3}$$

で計算される．エージェントのアンサンブル戦略は

$$\pi^t(s^t, a[i]) = \frac{p^t(s^t, a[i])^{\frac{1}{\tau}}}{\sum_k p^t(s^t, a[k])^{\frac{1}{\tau}}} \tag{4}$$

で計算される．s^t は時刻 t における状態，$a[i]$ は行動の選択肢，$\pi^t(s,a)$ は時間 t でプレイヤーが状態 s でとる行動 a の確率，τ は温度パラメータである．

ここで，M-Qubed について説明する．M-Qubed [1] は，状態における行動の価値を算出し，自己の最低限の報酬を確保しつつ，高い報酬を獲得するために相手プレイヤーと協調することを学習するアルゴリズムであり，多くのゲームにおいて優れた戦略を獲得することができる．M-Qubed では自己と他者の行動の組み合わせを状態 s と定義し，状態 s における行動 a の価値 $Q(s,a)$（Q 値）を学習するため Sarsa [2] が使用されている．Sarsa の更新則は以下の式で表される．

$$Q^{t+1}(s^t, a^t) = Q^t(s^t, a^t) + \alpha[r^t + \gamma V^t(s^{t+1}) - Q^t(s^t, a^t)] \tag{5}$$

$$V^t(s) = \sum_{a \in A} \pi^t(s,a) Q^t(s,a) \tag{6}$$

a^t は時刻 t に実際にとった行動，r^{t+1} は a^t によりもたらされる正規化された報酬，α は学習率，γ は割引率である．M-Qubed はこの更新則によって Q 値を求めた後,「利益追求」,「損失回避」,「積極的探索」を目的とした 3 つの戦略によってエージェントの方策を決定する．

利益追求

M-Qubed はマキシミニ戦略 π^{MM} を取るかどうかを考慮しながら，現在の状態における最大の Q 値を持つ行動を選択する純粋戦略をとる．

損失回避

被損失が許容可能な損失を超過するまで，その状態における最大の Q 値を持つ行動を選択し，超過してからはマキシミニ戦略を選択する．

積極的探索

利益追求戦略はその瞬間において高い報酬を獲得できるが，より高い未来の報酬を考慮しない近視眼的な戦略に陥りがちである．また，損失回避戦略は高い報酬をもたらす相手プレイヤーとの協調行動を導くことができない．この問題を解決するために Q 値の初期値を最大の可能割引報酬 $1/(1-\gamma)$ に設定することで，M-Qubed は大局的な戦略を学習する．

M-Qubed はこれら 3 つの戦略を組み合わせて最終的な戦略を生成する．

3.3 J-algorithm

J-alg を構成する 2 つのアルゴリズムは優秀だが，それぞれ問題点がある．BM-alg の問題点として，複数の戦略を持つので戦略の配分の決定や学習に時間が掛かり，相手プレイヤーとの協調行動により唯一つの状態が最適解となるゲームにおいて，平均獲得報酬が少なくなることが挙げられる．一方，S'-alg の問題点として，相手プレイヤーがグリーディな戦略を行うアルゴリズムだった場合，アルゴリズム内の満足度が減少し，低い報酬に満足してしまい一方的に搾取されてしまうことが挙げられる．

これらの問題を解決するため，損失回避戦略により相手プレイヤーに搾取されることを回避し，積極的探索戦略により最適な状態を探索することができる BM-alg によって S'-alg の欠点を補い，協調行動を繰り返しの早い段階で学習することができる S'-alg によって BM-alg の欠点を補うことにより，様々なゲームで素早く利得和を最大にする解をもたらす戦略，すなわち最適戦略を学習する強化学習アルゴリズムを構築する．

J-alg は探索パートに S'-alg を用い，定常パートに BM-alg を用いる．探索パートはゲーム開始から最大で 500 ラウンドまでとおいた．探索パート内で，自分と相手プレイヤーの行動が収束した場合，J-alg は定常パートへと移行する．移行する時に繰り返した行動の価値を高く設定し，BM-alg の戦略に基づいて行動を選択する．500 ラウンドを超過しても自分と相手プレイヤーの行動が収束しなかった場合，状態における行動の価値を再設定せずに，BM-alg が内部の戦略に基づいて行動を開始する．

ゲームの早い段階では，相手プレイヤーとの協調行動を学習することができる S'-alg が探索パートを担当し，仮に相手プレイヤーに搾取される行動を取られていたとしても，定常パートで自己の損失を回避する戦略をとることができる BM-alg が挽回をすることが期待される．

4. 実験

提案手法により構築したアルゴリズムの性能を確認するために，Crandall と Goodrich の論文 [1] より引用した 10 種類のゲームを用いて行った実験について述べる．

BM-alg の割引率を $\gamma = 0.95$，状態として用いる過去の行動の数を $\omega = 1$，温度パラメータを $\tau = 0.2$ といた．S'-alg の学習率を $\lambda = 0.99$，確率を $p = 0.3$ に設定した．ゲームを行い，行プレイヤーと列プレイヤーが 30 回同じ行動を繰り返した時を，両者の行動が収束したと定義する．表 1 は Crandall と Goodrich の論文 [1] で用いられている 10 種類の 2 人 2 行動非零和行列ゲームである．太字はプレイヤーの利得和を最大にする解を表している．

	c	d
a	**1.0,1.0**	0.0,0.0
b	0.0,0.0	0.5,0.5

(a) Common interest game (CIG)

	c	d
a	**1.0,0.5**	0.0,0.0
b	0.0,0.0	**0.5,1.0**

(b) Coordination game (CG)

	c	d
a	**1.0,1.0**	0.0,0.75
b	0.75,0.0	0.5,0.5

(c) Stag hunt (SH)

	c	d
a	0.0,1.0	**1.0,0.67**
b	0.33,0.0	0.67,0.33

(d) Tricky game (TG)

	c	d
a	**0.6,0.6**	0.0,1.0
b	1.0,0.0	0.2,0.2

(e) Prisoner's dilemma (PD)

	c	d
a	0.0,0.0	**0.67,1.0**
b	**1.0,0.67**	0.33,0.33

(f) Battle of the sexes (BS)

	c	d
a	**0.84,0.84**	0.33,1.0
b	1.0,0.33	0.0,0.0

(g) Chicken (Ch)

	c	d
a	0.84,0.33	0.84,0.0
b	0.0,1.0	**1.0,0.67**

(h) Security game (SG)

	c	d
a	0.0,0.0	**0.0,1.0**
b	**1.0,0.0**	0.0,0.0

(i) Offset game (OG)

	c	d
a	**1.0,0.0**	**0.0,1.0**
b	**0.0,1.0**	**1.0,0.0**

(j) Matching pennies (MP)

表 1: 10 種類の 2 人 2 行動非零和行列ゲーム

4.1 実験 1

同一のアルゴリズムを持つエージェント同士がゲームを行った場合に，提案手法のアルゴリズムを搭載したエージェントが最適戦略を学習するかどうかを確認するため，表 1 の 10 種類の 2 人 2 行動非零和行列ゲームを 5 万回繰り返す実験を 50 回行った．そして，開始時から受け取る全ての報酬の平均，平均獲得報酬を比較した．平均獲得報酬はゲームにおける最適戦略への収束が早ければ早いほど大きい値を示す．表 2 に 10 種類のゲームにおける平均獲得報酬を各ゲームの最大利得和で割り正規化した値を示す．値が 1 に近いほどそのゲームにおいて最適戦略をとれていることを示している．これらを平均した値は数値が大きいほど多くのゲームで高い報酬を得たことを表す．

表 2: 各アルゴリズムを搭載したエージェント同士でゲームを行った時の平均獲得報酬を最大利得和で割り正規化した値

	J-alg	M-Qubed	S-alg	BM
CIG	0.999080	0.993804	0.999039	**0.999181**
CG	0.951450	0.847301	**0.997966**	0.867965
SH	**0.999182**	0.995857	0.999149	0.998505
TG	0.998033	0.814068	**0.998251**	0.828770
PD	0.998569	0.711794	**0.998626**	0.723831
BS	0.998686	0.910656	**0.998700**	0.917655
Ch	**0.998909**	0.934089	0.998900	0.931541
SG	**0.998612**	0.902031	0.718236	0.907887
OG	0.466739	0.473539	**0.499964**	0.478196
MP	1.000000	1.000000	1.000000	1.000000
平均	**0.940926**	0.858314	0.920883	0.865353

同じアルゴリズム同士で 10 種類のゲームを行った場合，J-alg は Offset Game (OG) を除く 9 種類のゲームにおいて最終的に最適戦略を学習している．J-alg は探索パートの S'-alg によって M-Qubed, BM よりも素早く相手プレイヤーとの協調行動を学習しており，多くのゲームにおいてこれらよりも高い報酬を獲得している．M-Qubed, BM は Prisoner's dilemma (PD) において特に結果が悪くなっているが，これはこの 2 つのアルゴリズム内の損失回避戦略が働き，互いにグリーディな行動をしたため，相手プレイヤーとの協調行動を学習することが遅くなっているためである．S-alg は Security Game (SG) と Offset Game (OG) を除く 8 種類のゲームにおいては高い報酬を得ているが，Security Game (SG) では次善の報酬に満足してしまい，そこで探索を止め，最適戦略を学習することができないため，他のアルゴリズムよりも得た報酬が少なくなっている．Offset Game (OG) は最適戦略を学習することが困難で，全てのアルゴリズムが未達である．

4.2 実験 2

表 1 の 10 種類のゲームにおいて各アルゴリズムを持つエージェントが総当り戦を行った場合の平均獲得報酬を比較し，J-alg の性能を確認する．表 1 の 10 種類の 2 人 2 行動非零和行列ゲームを 5 万回繰り返す実験を 50 回行った．プレイヤーの立場が非対称なゲーム（Tricky Game, Security Game）が存在するので，これらのゲームでは行プレイヤーと列プレイヤーを入れ替えてゲームをもう一度行った．表 3 に 10 種類のゲームでのそれぞれのアルゴリズムを搭載したエージェントが総当り戦を行った時の平均獲得報酬を各ゲームの最大利得和で割り正規化した値を示す．1 を超えた場合は，他のプレイヤーを搾取して各ゲームの最大利得和よりも高い報酬を獲得したことを示す．これらを平均した値は数値が大きいほど多くのゲームで高い報酬を得たことを表す．

J-alg は内部の S'-alg により協調戦略を BM-alg により損失回避戦略をバランスよくとるため多くのゲームにおいて高い平均獲得報酬を示している．S-alg は相手プレイヤーと望む状態に背反が起きており，Coordination game (CG)，Battle of the sexes (BS)，Security game (SG)，Offset game (OG)，Matching pennies (MP) においてグリーディな戦略を取られ，一方的に搾取されたため，平均獲得報酬が少なくなっている．M-Qubed や BM は学習に多くのインタラクションを必要とするため，相手プレイヤーと上手く協調することができず，平均獲得報酬が低くなっている．M-Qubed は Battle of the sexes (BS) や Matching pennies (MP) において相手プレイヤーを

表 3: 各アルゴリズムを搭載したエージェントが総当り戦を行った時の平均獲得報酬を最大利得和で割り正規化した値

	J-alg	M-Qubed	S-alg	BM
CIG	0.998983	0.995743	**0.998991**	0.998603
CG	**1.048619**	0.915732	0.832476	0.905246
SH	0.999098	0.996896	**0.999148**	0.998226
TG	**0.952595**	0.863191	0.943683	0.868901
PD	**0.900618**	0.743598	0.871854	0.754596
BS	0.988460	**1.002762**	0.921269	0.951607
Ch	0.974938	0.933062	**0.975377**	0.926621
SG	**0.935857**	0.881564	0.810752	0.885159
OG	**0.555748**	0.497452	0.341784	0.538482
MP	1.041080	**1.051359**	0.858698	1.048864
平均	**0.939600**	0.888136	0.855403	0.887630

搾取する戦略を取り続け，高い平均獲得報酬を得ている．

表 2 と表 3 の全てのゲームの平均獲得報酬の平均から，J-alg は同じエージェント同士と総当り戦のどちらの場合でも最も高い報酬を得ていることがわかる．S-alg は同じエージェント同士では協調戦略を積極的に取ることにより成功を収めているが，グリーディな戦略を取る相手とゲームを行った場合，相手プレイヤーに搾取され報酬が低くなっている．M-Qubed と BM は学習に多くのインタラクションが必要であるため，ほとんどのゲームにおいて J-alg よりも平均獲得報酬が低くなっており，特に相手プレイヤーと協調戦略をとることが最適戦略であるゲームの時に，自己の利益を確保する戦略をとってしまうため低くなっている．

5. 結論

人々の意思決定が互いに影響し合う状況をモデル化した「ゲーム」において，意思決定主体が学習を行う場合に，自己の利益を最大化する戦略を獲得するアルゴリズムが広く研究されている．しかしながら，既存の強化学習アルゴリズムは様々なゲームにおいて高い報酬を獲得することができるが，最適な戦略を学習するために多くのインタラクションを必要とするという問題点があった．

本研究では，それぞれが得意な分野を持つ 2 つのアルゴリズムを組み合わせて，より素早く高い利得を獲得するアルゴリズムを構築した．提案したアルゴリズムを搭載したエージェント同士で 10 種類の非ゼロ和行列ゲームを行った時，9 種類のゲームにおいて高い報酬を獲得することができた．また，複数の異なるエージェントと総当り戦を行った場合でも最も高い平均獲得報酬を得ることができた．

今後の課題としては，Offset Game (OG) において自分と相手プレイヤーが協調行動をとり，両プレイヤーの利得和を最大にする戦略をとるアルゴリズムを構築すること，また，2 人 2 行動ゲームだけでなく n 人 n 行動ゲームにおいても最適戦略を高速で学習するアルゴリズムを構築することが挙げられる．

参考文献

[1] J.W. Crandall and M.A. Goodrich, “Learning to compete, coordinate, and cooperate in repeated games using reinforcement learning.”, Mach Learn, 82: 281–314, 2011.

[2] G.A. Rummery and M. Niranjan,“On-line Q-learning using connectionist systems”, Technical Report TR166, Cambridge University Engineering Department, 1994.

[3] J.L. Stimpson and M.A. Goodrich, “Learning To Cooperate in a Social Dilemma: A Satisficing Approach to Bargaining”, Proc. ICML, 728–735, 2003.

[4] R.S. Sutton and A.G. Barto, 三上貞芳・皆川雅章訳『強化学習』森北出版, 1998.

[5] M.A. Wiering and H. van Hasselt, “Ensemble Algorithms in Reinforcement Learning”, IEEE Trans Syst Man Cybern B, 38: 930–936, 2008.

[6] 岡田章『ゲーム理論 新版』有斐閣, 2011.

[7] 藤田渉, 森山甲一, 福井健一, 沼尾正行 “繰り返しゲームでの強化学習アルゴリズムの組み合わせによる協調行動の学習”, 人工知能学会全国大会（第 28 回）論文集, 4H1-4, 2014.