E-019

四マスゲームを用いたミニマックスQ学習の動力学解析
Analysis of the dynamics of the Mini-Max Q learning using the 4-cell game

杉原 慎司† 伊藤 昭‡ 寺田 和憲‡
Shinji Sugihara Akira Ito Kazunori Terada

1. はじめに

ゼロ和２人ゲームでは双方自己の行動を変えても自己の得点を改善できないという意味での最適解 (Nash 均衡解) が必ず存在するが，それは多くの場合複数の行動を一定の確率で選択する混合戦略となる．しかしながら，状態の関数として最適行動を学習するという従来の Q 学習においては混合戦略を学習することができない．これに対してゲーム理論においてミニマックス原理により最適解を学習する MiniMaxQ 学習 [1] が Littman によって提案されている．しかしながら，論文で提案されている課題では最適戦略が学習されていないように見える．

そこで本研究では，四マスゲームという最適戦略が手計算により求めることができるゲームを用いて，MiniMaxQ 学習が最適解を学習するための条件を理論，実験を比較することで検証し，MiniMaxQ 学習を利用する時の注意点を明らかにする．

2. ゲーム概要

四マスゲームは Littman の論文 [1] において使われているミニサッカーゲームを最適戦略が計算できるように簡単化したものである．

四マスゲームは図１のフィールド上で行われる．A，B はそれぞれプレイヤーであり，○で囲まれたプレイヤーがボールを保持している．図はゲーム開始時の状態 (初期状態) である．両プレイヤーは停止，横移動，縦移動の３行動を選択することができる．プレイヤーはボールを保持して前に進む (A は右に，B は左に) とゴールしたと見なされ得点を得る．その後，両プレイヤーは初期位置に戻され，ボールはランダムに A,B に渡される．

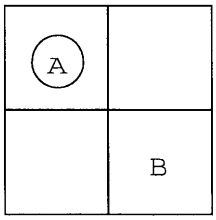

図 1: ゲームフィールド

プレイヤーが相手プレイヤーのいるマスへ移動しようとした場合は，プレイヤーは元の位置に留められ，またボールを保持している場合はボールが相手プレイヤーへ渡される．両プレイヤーが同時に同じマスに移動しようとした場合は，ランダムに選択された一方のプレイヤーが目的位置に移動し，他方は元の位置に留まる．もし元の位置に留められたプレーヤがボールを保持していた場合は，ボールは相手プレーヤに渡される．

ボールを保持したプレイヤーが相手がミスをするまで停止を続けるのを防ぐために，各移動タイミングで確率 $d(\ll 1)$ の割合でゲームは引き分け (draw) となる．引き分けとなると，プレイヤーは初期位置に戻されボールはランダムに A，B に渡される．

†岐阜大学大学院工学研究科
‡岐阜大学工学部

3. 四マスゲームでの最適戦略

このゲームの状態は配置，ボール保持の組合せにより 12 通り (ゴール状態を除く) 考えることができる．しかし，このゲームにおいては対象性を考慮すると図 2，図 3，図 4 のように状態 S_1，S_2，S_3 の３つの状態が実質的に相異なる．各状態での可能な行動，停止，横移動，縦移動を S，H，V とし $a_i \in A$ ={S,H,V} とする．

状態 $S_i(i = 1, 2, 3)$ での状態価値を次のように定義する．このゲームはゼロ和ゲームであるので両プレイヤーにとっての最適戦略 (Nash 均衡解) が存在する．また，ゲーム理論によるとゼロ和ゲームの Nash 均衡解は複数存在するとしても，その利得は全ての均衡解で共通となる．よって，これを最適戦略の状態価値とすることができる．ゼロ和ゲームのため B にとっての状態価値は A にとっての状態価値の符号を逆にしたものになる．

A にとっての S_i の状態価値を V_i とする．状態 S_1 においてゲームの性質よりプレイヤーの選択すべき行動は A が H，B が V であることがわかる (図 2)．このとき両プレイヤーは同位置に移動しようとするため 1/2 の確率でどちらかが元の位置に留まる．B が留まる場合，A は報酬を得る．A が留まる場合，ボールが B に渡り状態は S_2 になる．よって，価値関数の割引率を γ，報酬 r を 1 とすると

$$V_1 = \frac{1}{2}(1-d)(1-\gamma V_2) \qquad (1)$$

が成り立つ．なお，$(1-d)$ は確率 d で状態価値 0 の初期状態に戻るために付加された因子である．

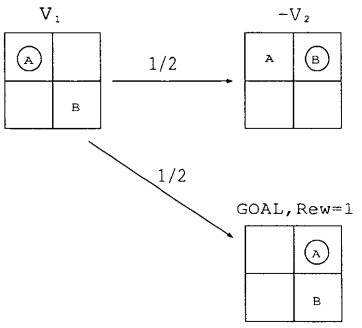

図 2: 状態 S_1 とその遷移先

状態 S_2 において A は S，V，H を，B は S，V を確率的に選択することが合理的である．それぞれの行動によって遷移する状態を図 3 に示す．このとき，A が S，V，H を選択する確率を $1-p-q$，q，p，B が S，V を

選択する確率を r，$1-r$ とすると

$$V_2 = (1-d)\{rp + \gamma r(qV_2 + (1-p-q)V_1) + \gamma(1-r)(-pV_2 + qV_1 + (1-p-q)V_2)\} \quad (2)$$

が成り立つ．

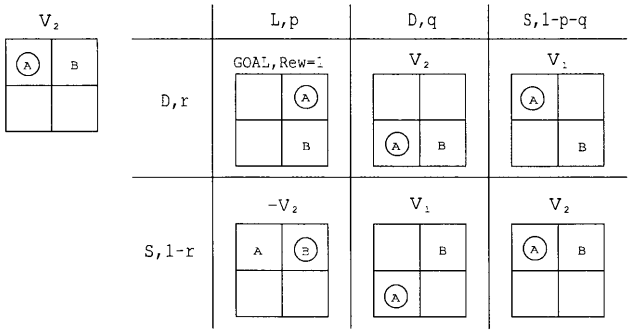

図 3: 状態 S_2 とその遷移先

状態 S_3 においては A は H を選択することによって B の選択に関係なくゴールすることができる．よって

$$V_3 = 1 - d \quad (3)$$

が成り立つ．

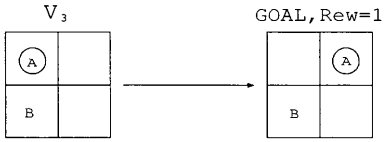

図 4: 状態 S_3 とその遷移先

以上で定義した状態価値を用いると行動価値 $Q(S_i, a^A, a^B)$，すなわち状態 S_i において A が行動 a^A，B が行動 a^B を選択し，以後最適行動をとるとしたときの行動価値を求めることができる．

状態 S_1

$$Q_1 = Q(S_1) = (1-d)\begin{pmatrix} \gamma V_1 & \gamma V_3 & \gamma V_2 \\ 1 & 1 & \frac{1-\gamma V_2}{2} \\ \gamma V_2 & -\frac{1-\gamma V_2}{2} & \gamma V_1 \end{pmatrix}$$

状態 S_2

$$Q_2 = Q(S_2) = (1-d)\begin{pmatrix} \gamma V_2 & \gamma V_2 & \gamma V_1 \\ -\gamma V_2 & 1 & 1 \\ \gamma V_1 & \gamma V_3 & \gamma V_2 \end{pmatrix}$$

状態 S_3

$$Q_3 = Q(S_3) = (1-d)\begin{pmatrix} \gamma V_3 & \gamma V_1 & \gamma V_3 \\ 1 & 1 & 1 \\ -1 & \gamma V_2 & \gamma V_3 \end{pmatrix}$$

なお，縦が A の戦略，横が B の戦略，また添字の順序は {S, H, V} とする．

ここで，p^A，p^B を A，B の行動確率ベクトルとすると，

$$V_i = \max_{p^A} \min_{p^B} p^A Q_i p^B \quad (4)$$

となる．これを用いて具体的に p^A，p^B を求めると，状態 S_1 では

$$p^A = (0,1,0), p^B = (0,0,1)$$

となり，式 (4) は式 (1) となる．

同じく状態 S_3 では，

$$p^A = (0,1,0), p^B = (p,q,r)$$

となり，式 (3) となる．

状態 S_2 は少し複雑である．$p^A = (1-p-q, p, q)$ として

$$\min_{p^B} p^A Q_i p^B = (1-d)\min(M_s, M_h, M_v) \quad (5)$$
$$M_s = \gamma((1-p-q)V_2 - pV_2 + qV_1)$$
$$M_h = \gamma((1-p-q)V_2 + qV_3) + p$$
$$M_v = \gamma((1-p-q)V_1 + qV_2) + p$$

となるが，$M_s \leq M_h$ であるため，

$$V_i = (1-d)\max_{p^A}\min(M_s, M_v) \quad (6)$$

となる．

M_s，M_v の関数形を考えると，$M_v = M_s$, $p+q=1$ のところで式 (6) は最大値となり，q，V_2 が次のように求まる．

$$q = \frac{\gamma V_2 + 1}{\gamma V_1 + 1} \quad (7)$$

$$V_2 = (1-d)\frac{\gamma(\gamma V_2^2 + V_1)}{\gamma V_1 + 1} \quad (8)$$

当然ながら，式 (7) と $p = 1-q$ を併せると式 (2) になる．また，式 (1) とあわせることで，最終的に次のように V_2 を求めることができる．

$$V_2 = (1-d)\frac{2 + \bar{\gamma} + \bar{\gamma}^2 - \sqrt{4 + 4\bar{\gamma} + 5\bar{\gamma}^2 - 10\bar{\gamma}^3 + \bar{\gamma}^4}}{6\bar{\gamma}^2} \quad (9)$$

ここで，$\bar{\gamma} = (1-d)\gamma$ である．

一方，B から見ると次式が成り立つ．

$$V_i = \min_{p^B}\max_{p^A} p^A Q_i p^B \quad (10)$$

これを直接解いても良いが図 3 を用いると $p^B = (1-r, 0, r)$ となり

$$V_2 = (1-d)\min_r \max(N_s, N_h, N_v) \quad (11)$$
$$N_s = \gamma(V_2(1-r) + V_1 r)$$
$$N_n = -\gamma(1-r)V_2 + r$$
$$N_v = \gamma(V_1(1-r) + V_2 r)$$

となる．これを解くと，

$$r = \frac{\gamma(V_1 + V_2)}{1 + \gamma V_1} \quad (12)$$

と，式 (8) になる．

以上で割引率 γ，確率 d の関数として行動価値 Q，状態価値 V_i を求めることができた．このゲームがゼロ和ゲームであることを考えると，ここで求めた Nash 均衡解が最適解となる．よって，ボール保持状態での最適戦略は状態 S_1 では行動 H，状態 S_2 では確率 r，p，q により定まる行動，状態 S_3 では行動 H となる．

4. 学習理論

4.1 MiniMaxQ 学習

MiniMaxQ 学習では，自己の行動 $a^s \in A^s$，相手の行動 $a^o \in A^o$ を引数として持つ行動価値関数 $Q(S, a^s, a^o)$ を導入する．Littman の主張は，ゼロ和ゲームでは両エージェントが次式に従って学習すれば，$Q(S, a^s, a^o)$ は双方にとって最適な Nash 均衡解の行動価値に収束するというものである．

$$Q(S_t, a^s, a^o) \leftarrow (1-\alpha)Q(S_t, a^s, a^o) + \alpha(r_{t+1} + \gamma \max_{\{p(a^{s*})\}} \Sigma_{a^{s*}} p(a^{s*}) \Sigma_{a^{o*}} p(a^{o*}|S) Q(S_{t+1}, a^{s*}, a^{o*})) \quad (13)$$

右辺第２項の意味は，こちらの $Q(S, a^s, a^o)$ を最小化する行動を相手が選択すると仮定して，その結果が最大となるように自己の行動を選択することを意味する．具体的には，こちらが行動 a^s を選択する確率を $p(a^s)$ とすると，相手は行動価値 $\sum_{a^s} p(a^s) Q(S, a^s, a^o)$ を最小化する行動 $a^o = a^o(\{p(a^s)\})$ を選択するものと考える．したがって，こちらは確率分布 $\{p(a^s)\}$ を変化させて自己の行動価値を最大化する確率政策をとることになる．すなわち

$$\max_{\{p(a^s)\}} \sum_{a^s} p(a^s) Q(S_{t+1}, a^s, a^o(\{p(a^s)\})) \quad (14)$$

で定まる $\{p(a^s)\}$ を選択することになる．ここで max は $\{p(a^s)\}$ を変化させて，その後に続く式を最大化することを意味する．

4.2 SimpleQ 学習

SimpleQ 学習では，自己の行動 $a^s \in A^s$ を引数として持つ行動価値関数 $Q(S, a^s)$ を使用する．また，Q の更新式として式 (15) を使用する．

$$Q(S_t, a^s) \leftarrow (1-\alpha)Q(S_t, a^s) + \alpha(r_{t+1} + \gamma \max_{a^s} Q(S_{t+1}, a^s)) \quad (15)$$

5. 実験１-行動価値の収束

MiniMaxQ 学習をさまざまな相手と対戦させることで学習させる．次に学習中の行動価値の変化を記録し，手計算で求めた最適解の行動価値と比較する．

なお，学習時の共通条件を次のようにする．学習時間 10^6step，$\gamma = 0.9$，$d = 0.01$，$\alpha = 0.001$ とする．

5.1 対戦戦略

MiniMaxQ 学習を行うための対戦相手としては次の３つの戦略を用いる．(1)Random 戦略：この戦略は名前の通り行動をランダムに選択する．(2)MiniMaxQ 学習戦略：この戦略は MiniMaxQ 学習であり，全く同じ戦略同士で対戦し学習を行う．なお，学習条件は $\gamma = 0.9$，$\alpha = 0.001$，$\varepsilon = 0.01$ である．(3)HandMade 戦略：この戦略は 3. において手計算によって求めた最適戦略の行動選択確率を基に行動する．なお，確率は $\gamma = 0.9$ の条件の基で計算しており $r = 0.4476$，$p = 0.0446$，$q = 0.9554$ である．また，学習時の MiniMaxQ 学習の行動選択に ε グリーディ手法を使用し (1)，(3) では $\varepsilon = 1.0$，(2) では $\varepsilon = 0.01$ である．

5.2 結果

表 1 に学習の結果を示す．Q 値二乗誤差の項目は最適な Q 値と学習後の Q 値の二乗誤差を示す．

	MiniMax 戦略		
学習相手	Random	MiniMax	HandMade
Q 値二乗誤差	2.51	116.68	235.67

表 1: 収束状況

表 1 を見ると Random 戦略との対戦によって学習した戦略では Q 値の二乗誤差が 2.51 となっており，ほぼ最適な Q 値に収束していることがわかる．他の学習においては最適な Q 値に収束していないことがわかる．これは Random 戦略以外の学習相手には行動に偏りがあり，全ての行動に対して学習することができないからである．例として図 5 に MiniMaxQ 学習同士の学習対戦時の状態 S_2｛ボール保持時｝での Q 値の推移を示す．なお，図に示す Q 値の推移グラフは横軸に step，縦軸に Q 値を 1000step 間隔で表示している．A X : B Y が A が行動 X を B が行動 Y を選択したときの行動対での Q 値を示す．また，グラフ横の矢印は最適戦略のときの Q 値を示している．なお，$X, Y \in \{S, H, V\}$ とする．

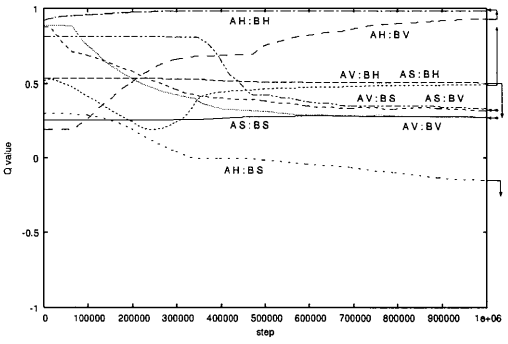

図 5: 状態 S_2 での Q 値の推移

図 5 を見ると，B が H の行動を選択するときの Q 値が大きく最適値よりずれていることがわかる．これは学習相手である B がこの状態の時に H を選択することがほとんどないからである．

また，表 2 に MiniMaxQ 学習同士の学習対戦後の戦略 (行動選択確率) を示す．なお，それぞれの状態には A が上方にいるとき，下方にいるときの２種類あるため左側に上方，右側に下方にいるときの確率を示す．

	学習確率		最適確率
p	0.0000	0.0507	0.0446
q	0.5562	0.9493	0.9554
r	0.5267	0.4882	0.4476

表 2: 行動選択確率

表 2 を見ることにより，Q 値が収束していないことで戦略も最適な戦略確率になっていないことがわかる．特に今回の場合は A が上方にいるときの戦略が最適な値でないことがわかる．

また，HandMade 戦略との学習においての Q 値は HandMade 戦略が選択することのない行動については

全く収束しておらず，その結果行動選択確率は最適な戦略確率になっていない．図6にHandMade戦略との学習対戦時の状態S_2{ボール保持時}でのQ値の推移を示す．

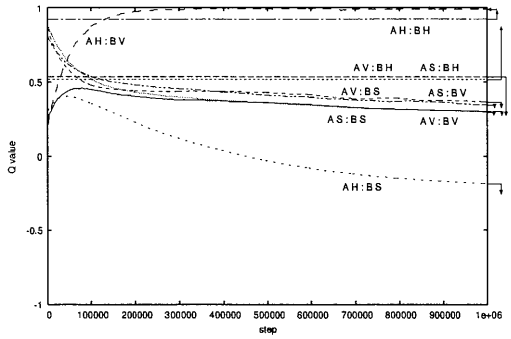

図6: 状態S_2でのQ値の推移

表3にHandMade戦略との学習対戦後の戦略を示す．この学習においてはrの行動選択確率が最適な確率になっていない．また，この表には現れていないがボール保持時の状態S_1での選択行動が$p(H) = 1.0$になっていない．

	学習確率		最適確率
p	0.0522	0.0351	0.0446
q	0.9478	0.9649	0.9554
r	0.9282	0.6242	0.4476

表3: 行動選択確率

6. 実験2-学習戦略の性能

学習後の戦略をいくつかの戦略と対戦させることで，収束時の戦略の性能を評価する．

なお，対戦時の共通条件を次のようにする．対戦時間10^6step，$d = 0.01$，$\alpha = 0.0$，$\varepsilon = 0.0$とする．

6.1 対戦戦略

学習後の性能を評価するための対戦相手としては次の3つの戦略を用いる．(1)Random戦略 (2)SimploQ学習戦略 (3)HandMade戦略

6.2 結果

表4に対戦結果を示す．数値はMiniMax戦略の各々の戦略に対しての勝率[%]を示す．

	MiniMax戦略		
学習相手	Random	MiniMax	HandMade
vs.HandMade	50.02	49.78	48.70
vs.Random	70.73	69.17	54.09
vs.SimpleQ	57.36	50.37	26.24

表4: 対戦結果

4を見ると全ての学習においてHandMade戦略と互格の戦いができていることがわかる．しかし，それ以外の対戦においては勝率に明確な差ができている．図7に差が一番大きく現れたSimpleQ学習戦略に対する各戦略の単位stepあたりの勝率を示す．

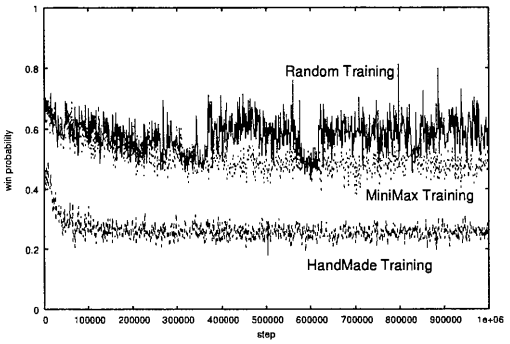

図7: SQとの対戦

図7よりHandMade戦略との対戦で学習したMiniMax戦略はSimpleQ学習戦略に対して全く対応仕切れていないことがわかる．また，MiniMaxQ学習同士の対戦で学習したMiniMax戦略はSimpleQ学習戦略が学習をほぼ完了したと思われるstep数においても互格の対戦ができている．そして，Random戦略との対戦で学習したMiniMax戦略はSimpleQ学習戦略が学習をほぼ完了したと思われるstep数においても勝率が5割を越えることができている．これは言いかえれば，SimpleQ学習では，最適戦略を学習できないことを意味する．

7. 考察

実験結果よりMiniMaxQ学習は全ての行動対を学習させることによって最適な戦略を学習することがわかった．しかし，逆に学習する相手が今回のHandMade戦略のように最適行動のみをする場合においては最適な戦略を学習しない，もしくは学習相手のみに特化した戦略になってしまうことがわかった．これはSimpleQ学習にも言えることだがそれを回避するためのMiniMaxQ学習であったはずなのに同じ結果となってしまっている．

8. まとめ

本実験によってMiniMaxQ学習が最適戦略を学習するためには互いが全ての行動を選択するような条件で多くのstep学習しなければならないことがわかった．しかし，この結果は本来敵である相手に最適行動以外の多様な行動をとることを求めており，MiniMaxQ学習が実際の対戦の中で学習をすることの困難さを表している．今後の課題としては学習段階での行動選択に条件を付加することなく最適な行動を学習できるような学習法を提案する必要があると思われる．

参考文献

[1] Michale L. Littman, "Markov games as a framework for multi-agent reinforcement learning" Proceedings of the 11th International Conference on Machine Learning (ML-94)

[2] Richard S. Sutton and Andrew G. Barto, "Reinforcement Learning" The MIT Press, 1998