安田女子大学紀要　50，233-240　2022.

Q学習手法の学習における4,1,2-ゲーム使用の提案

山　下　明　博

Proposal for Using 4,1,2-Games on Learning Q-learning

Akihiro YAMASHITA

安田女子大学家政学部造形デザイン学科

要　　旨

Q学習は、代表的な強化学習手法の1つである。この手法を学ぶときに、従来、3,3,3-ゲームと呼ばれる三目並べが題材として選ばれることが多かった。しかし、三目並べでQ学習を理解しようとすると、棋譜データ量が膨大なことがその理解を妨げるという問題点があった。本稿は、m,n,k-ゲームを複数回実行するプログラムを開発するとともに、勝ち、負け、引き分けがすべて現れるm,n,k-ゲームの中で、最も棋譜データ量が少ないゲームを選定し、選定した4,1,2-ゲームを、Q学習手法の学習のために使用することを提案するものである。

キーワード：Q学習、強化学習、m,n,k-ゲーム、Bellman最適方程式、三目並べ

はじめに

Q学習は、代表的な強化学習方法の1つである。この手法を学ぶときに、従来、3,3,3-ゲームと呼ばれる三目並べが題材として選ばれることが多かった。しかし、三目並べでQ学習の手法を理解しようとすると、棋譜データ量やQテーブルの大きさが巨大であることにより、その理解が妨げられるという問題点があった。

本稿は、勝ち、負け、引き分けがすべて現れるm,n,k-ゲームの中で、最も棋譜データ量が少ない4,1,2-ゲームを、三目並べの代わりに、Q学習の理解のために使用することを提案するものである。

Ⅰ．Q学習における三目並べ

1．強化学習

強化学習（Reinforcement Learning）は機械学習（Machine Learning）手法の一つで、システムが自ら試行錯誤しながら最適な制御を実現する手法である。強化学習の考え方自体は、すでに1950年代から存在していたが、強化学習が注目されるようになったのは、2016年、強化学習を用いたGoogle DeepMindの 囲 碁AI「AlphaGo」 が、囲碁のトップ棋士に完勝するという成果を挙げたときからである。そして、その後も囲碁AIの研究は続き、「AlphaZero」は、過去の棋譜やビッグデータを必要とせず、自己対局によって強化学習を行い、全くの初心者の状態から、人間がまったく勝利できない水準まで8時間で到達するという成果を挙げている。

強化学習では、環境とエージェントの相互作用を考える。図１に、強化学習におけるエージェントと環境の相互作用を示す。エージェントは、環境の状態 s を見て、行動 a を実行する。環境は、エージェントの行動 a に応じて状態 s を更新し、行動の結果として報酬 r をエージェントに返す。

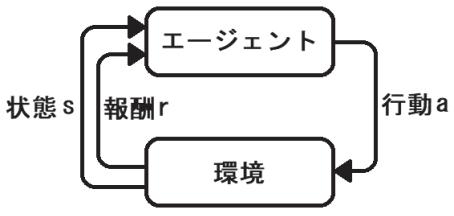

図１　強化学習におけるエージェントと環境の相互作用

強化学習の目的は、環境を攻略するエージェントを育て、総報酬をなるべく多くする行動の選び方を模索することである。そして、数学的には、Bellman最適方程式を解くことになる。

$$V^*(s) = \max_a Q^*(s,a),$$
$$Q^*(s,a) = \sum_{s'} P^a_{s,s'} \left(R^a_{s,s'} + \gamma V^*(s') \right)$$

2．Q学習

Q学習（Q-learning）は、代表的な強化学習方法の１つであり、TD学習（Temporal-Difference Learning）に属する。その起源は、1989年にクリス・ワトキンズ（Chris Watkins）が発表した論文[1]である。

TD学習では、モデルに関する情報が不要となる。そのため、行動1回単位で価値関数を更新できるという特徴がある。

Q学習では、次式により行動価値関数Qを更新していく。αは学習率、γは割引率である。

$$Q(s_t,a_t) \leftarrow (1-\alpha)Q(s_t,a_t) + \alpha \left(r_{t+1} + \gamma \max_a Q(s_{t+1},a) \right)$$

更新を行う度に、Qは、Bellman最適方程式のQ*に近づいていく。また、目的の方策は、Qを用いて、次式のように求めることができる。

$$\pi(s,a) = \begin{cases} 1 & (a = \underset{a}{\operatorname{argmax}} Q(s,a)), \\ 0 & (\text{otherwise}) \end{cases}$$

ただし、Qの更新が局所に偏るとうまく学習できないので、学習時に取る行動を、
ε-greedy法[2]でランダム性を加えて選ぶようにする。

また、Q学習をコンピュータ上に実装する場合、全ての状態sと行動aに対応するQの値を、Qテーブルという名称で配列として保持する。そのため、可能な状態や行動の数が膨大な問題では、必要なメモリーの容量および収束までの計算量が爆発してしまい計算不可能になる。

3．三目並べ

三目並べというゲームがある。これは、2人のプレイヤーが対戦するゲームであり、３×３のマス目に、先手プレイヤー（以下、先攻と称する）は○、後手プレイヤー（以下、後攻と称する）は×のシンボルを交互に書き込み、先に縦・横・斜めいずれかのマス目でシンボルを３つ揃えたら勝ちというルールである。

図２に、三目並べのゲームの進行例を示す。先攻が、横のマス目に○を３つ揃えたので、先攻の勝ちである。しかし、先攻、後攻がそれぞれ最善を尽くした場合、三目並べのゲームの結果はすべて引き分けとなる。

ゲーム論においては、２人のプレイヤーがm×nのマス目からできた盤面上に自分のシンボルを交互に書き込み、先に自分のシンボルを縦・横・斜めのいずれかにk個並べたプレイヤーが勝利となるゲームを、m,n,k-ゲームと総称している。三目並べは3,3,3-ゲームとなる。m,n,k-ゲームは、二人零和有限確定完全情報ゲーム[4]である。

4．三目並べによるQ学習

Q学習という強化学習のアルゴリズムを学ぶ際に、3,3,3-ゲームである三目並べが題材として選ばれることが多い。それは、三目並べのゲームの中で扱う状態 s の数が高々３の９乗（19683）通り、行動 a の数が９通りであり、一般的な性能のパソコン上で動作するプログラムでも扱える程度のQテーブルに収まるゲームであるからである。また、２人が対戦するゲームとして、勝ち、負け、引き分けの３つのパターンがすべて現れることも、三目並べが題材に選ばれる理由の一つである。片方のユーザが決して勝つことのできないゲ

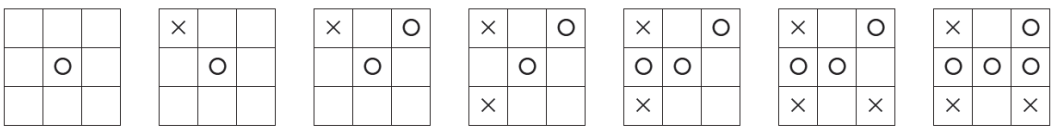

図２　三目並べのゲームの進行例

ームであれば、それは、学習の題材には不向きである。

しかし、強化学習を学ぼうとする者が、三目並べによるQ学習を理解しようとするとき、棋譜データ量が膨大なことがその理解を妨げるという問題点がある。三目並べは、わずか縦３マス、横３マスの盤面上のゲームではあるが、状態ｓと行動ａをすべて紙に書き出そうとすると、データ量が膨大なものになり、書き出すことは困難である。そのため、本当にすべての場合を網羅しているかを確認することは難しい。

５．Q学習の理解のための提案

1.4で述べた、Q学習の理解に三目並べを題材として選ぶ場合の棋譜データが膨大になるという問題点を解決するために、本稿では、4,1,2-ゲームをQ学習の理解のために使用することを提案するものである。

4,1,2-ゲームは、縦４マス、横１マスの盤面に、２人のプレイヤーが○と×を交互に書き込み、先に縦・横・斜めいずれかのマス目でシンボルを２つ揃えたら勝ちというルールをもつゲームである。勝ち、負け、引き分けのパターンがすべて現れるm,n,k-ゲームの中では、4,1,2-ゲームは最も棋譜データ量が少なく、ゲームの中で扱う状態ｓの数が高々３の４乗（81）通り、行動ａの数が４通りにしか過ぎない。そのため、すべての棋譜データを紙に書き出し、すべての場合を網羅しているかを容易に確認することができる。

図３に、4,1,2-ゲームのすべての棋譜を示す。

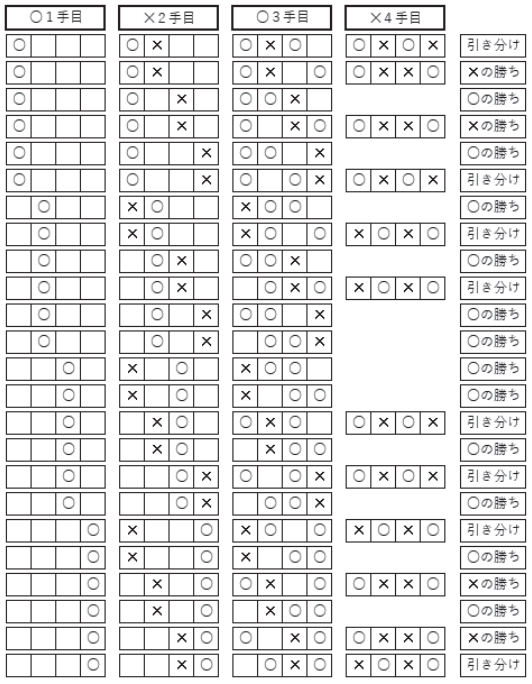

図３　4,1,2-ゲームの全棋譜

Ⅱ．Q学習によるm,n,k-ゲームプログラムの開発

１．開発の目的

今回、m≦４、n≦４、k≦４という条件の下で、m,n,k-ゲームのルールに従って、ランダム、先読みランダム、Q学習等のプレイヤーによる対戦を複数回実施し、m,n,k-ゲームの性質を調べるプログラムを開発した。

このプログラムにより、ランダムで手を決める２プレイヤー間の対戦を実施し、勝ち、負け、引き分けがどの程度の比率で出現するかについて調べた。また、Q学習により強化学習を行った２プレイヤー間の対戦を実施し、両プレイヤーが最善手を打ち続けた場合、先手必勝か、後手必勝か、引き分けになるかというゲームの性質を調べた。

２．開発環境

m,n,k-ゲームプログラムは、Windows10上のPython3.9.2で開発を行い、Q学習の成果は、pickleライブラリで直列化した。

３．プレイヤー

m,n,k-ゲームのルールに従ってプレイするプレイヤーの種類は、ランダム、先読みランダム、Q学習の3種類を用意した。

(1) ランダムプレイヤー（Random）

ランダムプレイヤーは、盤面のうち打つことができる位置を乱数で選び、先攻なら○、後攻なら×を打つものである。

(2) 先読みランダムプレイヤー
（LookAheadRandom）

先読みランダムプレイヤーは、次の打ち手で勝負に勝つかどうかを先読みし、勝つことがわかったらその位置に打ち、それ以外なら、ランダムプ

レイヤーと同じ手を打つものである。

(3) Q学習プレイヤー（Qlearning）

Q学習プレイヤーは、ランダムプレイヤーと200万回対戦し、状態sと行動aに対応するQ値を更新する学習を繰り返すことにより、強いゲームプレイヤーとして成長させたものである。Q学習の初期には、ε-greedy法でランダム性を加えて選ぶようにした。

200万回対戦し学習した後、他のプレイヤーと対戦するときは、$\varepsilon = 0$で、最善手を選択するようにしている。

4．プログラム実行例

今回開発したm,n,k-ゲームプログラムにより、三目並べの場合、各プレイヤーをそれぞれ対戦させた場合、どのような戦績になるかについて述べる。

（1）Random、（2）LookAheadRandom、（3）Qlearningの各プレイヤーを相互に10万回ずつ対戦させた結果は以下の通りである。1が先攻、2が後攻を示す。

その結果、先攻のほうが後攻より勝利する対戦が多いこと、プレイヤーの強さではRandom < LookAheadRandom < Qlearningの順であること、Q学習同士で両プレイヤーが最善手を打ち続けた場合、すべて引き分けになることを読み取ることができた。

5．プログラムの制限

今回、m,n,k-ゲームのルールに従って対戦するプログラムを開発するにあたり、m≦４、n≦４、k≦4という条件を設定した。これは、Q学習を行う上で、m,n,k-ゲームの中で扱う状態 s の数が高々３のm*n乗通り、行動 a の数がk通りであり、m、n、kが大きくなるにつれて、必要なメモリーの容量および収束までの計算量が爆発してしまい計算不可能になるからである。

本プログラムでは、afterstateを作成することにより状態 s と行動 a の組み合わせを集約したり、状態と行動をPythonの辞書機能で保持したりする工夫を行い、m≦４、n≦４、k≦４の範囲で動作するようにした。

Ⅲ．m,n,k-ゲームの性質の定義

m,n,k-ゲームは、2プレイヤー間で対戦するゲームであるが、両プレイヤーが最善手を打ち続けた場合、先手必勝か、後手必勝か、引き分けに分類することができる。これを、m,n,k-ゲームの性質と呼ぶことにする。

最善手を打ち続けるプレイヤーは、Q学習プレイヤー（Qlearning）でε=0を指定した場合である。そこで、m,n,k-ゲームプログラムを使用し、m≦４、n≦４、k≦４という条件の下で、Q学習プレイヤー同士で対戦を10万回ずつ行うことにより、両プレイヤーが最善手を打ち続けた場合、先手必勝か、後手必勝か、引き分けになるかというm,n,k-ゲームの性質を検討する。

ただし、mとnは交換可能であるので、検討の簡略化のために、m≧nと仮定する。

m,n,k-ゲームにおける先手必勝、後手必勝、引き分けという性質を表１にまとめた。「先」が先手必勝、「後」が後手必勝、「引」が引き分けを示している。

k= １の場合、すべて先手必勝である。また、

100000回試行結果	Random1	:58533	Random2	:28666	引分:	12801
100000回試行結果	Random1	:39688	LookAheadRandom2	:52015	引分:	8297
100000回試行結果	Random1	: 0	Qlearning2	:91698	引分:	8302
100000回試行結果	LookAheadRandom1	:81379	Random2	:11919	引分:	6702
100000回試行結果	LookAheadRandom1	:68548	LookAheadRandom2	:27070	引分:	4382
100000回試行結果	LookAheadRandom1	: 0	Qlearning2	:91587	引分:	8413
100000回試行結果	Qlearning1	:99453	Random2	: 0	引分:	547
100000回試行結果	Qlearning1	:99460	LookAheadRandom2	: 0	引分:	540
100000回試行結果	Qlearning1	: 0	Qlearning2	: 0	引分:	100000

表１　m,n,k-ゲームの性質

k=1

n＼m	1	2	3	4
1	先	先	先	先
2	先	先	先	先
3	先	先	先	先
4	先	先	先	先

k=2

n＼m	1	2	3	4
1	−	引	先	先
2	引	先	先	先
3	先	先	先	先
4	先	先	先	先

k=3

n＼m	1	2	3	4
1	−	−	引	引
2	−	−	引	引
3	引	引	引	先
4	引	引	先	先

k=4

n＼m	1	2	3	4
1	−	−	−	引
2	−	−	−	引
3	−	−	−	引
4	引	引	引	引

k=2のときも、盤面が小さ過ぎるときを除き、先手必勝である。k=3のときは、引き分けが多くなるが、mやnがともに大きいときは先手必勝になる。k=4のときは、すべて引き分けになる。

Ⅳ．Q学習の理解のための4,1,2-ゲームの提案

１．提案の目的

Q学習の理解に三目並べを題材として選ぶ場合、棋譜データ量が膨大になるという問題点はすでに述べた。そこで、棋譜データ量が三目並べより小さく、かつ、勝ち、負け、引き分けのパターンがすべて現れるm,n,k-ゲームが何であるかを見出すことにより、m,n,k-ゲームの理解、およびQ学習の理解を深めることが、本稿の目的である。

２．探索手法

棋譜データ量が小さく、かつ、勝ち、負け、引き分けのパターンがすべて現れるm,n,k-ゲームを探索する手法として、ランダムプレイヤー同士で対戦させることにする。これは、最善手を打ち続けるQ学習プレイヤーが決して打たないような、負けるかもしれない手を含めて、すべてのパターンを調べるためである。

表２に、m≦４、n≦４、k≦４という条件の下で、ランダムプレイヤー同士による対戦の結果を、先手勝利確率、後手勝利確率、引き分け確率

表２　ランダムプレイヤー同士による対戦結果

先攻勝利確率（%）

k=1

n＼m	1	2	3	4
1	100	100	100	100
2	100	100	100	100
3	100	100	100	100
4	100	100	100	100

k=2

n＼m	1	2	3	4
1	−	0	67	50
2	0	67	78	75
3	67	78	71	68
4	50	75	68	66

k=3

n＼m	1	2	3	4
1	−	−	0	0
2	−	−	10	23
3	0	10	59	59
4	0	23	59	59

k=4

n＼m	1	2	3	4
1	−	−	−	0
2	−	−	−	3
3	−	−	−	8
4	0	3	8	31

後攻勝利確率（%）

k=1

n＼m	1	2	3	4
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0

k=2

n＼m	1	2	3	4
1	−	0	0	17
2	0	0	22	24
3	0	22	29	31
4	17	24	31	33

k=3

n＼m	1	2	3	4
1	−	−	0	0
2	−	−	0	14
3	0	0	29	38
4	0	14	38	41

k=4

n＼m	1	2	3	4
1	−	−	−	0
2	−	−	−	0
3	−	−	−	6
4	0	0	6	27

引き分け確率（%）

k=1

n＼m	1	2	3	4
1	100	100	100	100
2	100	100	100	100
3	100	100	100	100
4	100	100	100	100

k=2

n＼m	1	2	3	4
1	−	100	33	33
2	100	33	0	0
3	33	0	0	0
4	33	0	0	0

k=3

n＼m	1	2	3	4
1	−	−	100	100
2	−	−	90	63
3	100	90	13	3
4	100	63	3	0

k=4

n＼m	1	2	3	4
1	−	−	−	100
2	−	−	−	97
3	−	−	−	86
4	100	97	86	41

に分けて示す。

先手勝利確率、後手勝利確率、引き分け確率のすべてに対し、確率が0％か100％以外である(m,n,k）の組み合わせは、
(m,n,k)=(4,1,2)，(1,4,2)，(3,3,3)，(4,3,3)，(3,4,3)，(4,3,4)，(3,4,4)，(4,4,4）の８通りである。

そして、これらの組み合わせの中で、m×nが最も小さいのは、(m,n,k）=（4,1,2)，(1,4,2）の２通りである。これらは、縦と横を交換すれば、同じm,n,k-ゲームであるので、棋譜データ量が三目並べより小さく、かつ、勝ち、負け、引き分けのパターンがすべて現れるm,n,k-ゲームは、4,1,2-ゲームとなる。

4,1,2-ゲームは、４×１の盤面をもち、各マス目ごとに、「○」、「×」、「　」という３つの状態を取ることができ、3つの行動を取ることがる。したがって、取りうる状態と行動の数は、３の４乗に３をかけて243である。しかし、4,1,2-ゲームが、異なるプレイヤーが交互に手を打つゲームであるため、すべて「○」といった、存在しえないパターンもある。実際には、前掲の図３に示した通り、4,1,2-ゲームのすべての棋譜は、24通りである。

4,1,2-ゲームの場合、棋譜の数が少ないので書き出すことが可能であるが、3,3,3-ゲーム（三目並べ）の場合、取りうる棋譜の数が、３の９乗に３をかけた59049になる。これは、とても書き出すことができないほど膨大なデータである。

また、4,1,2-ゲームの場合、Q値の総数も少ないので書き出すことが可能である。図４に、4,1,2-ゲームの200万回Q学習実施後のQ値を示す。盤面に置かれたパターンは、[a,b,c,d]（０≦a≦２、０≦b≦２、０≦c≦２、０≦d≦２）で表し、その下の数値は、盤面の状態でのQ値を示している。なお、次に打つのは必ず○側（数値では１）になるように、先攻、後攻が入れ替わるとき、盤面の１と２を交換している。

結　　論

本稿では、Q学習によるm,n,k-ゲームの性質を調べたり、棋譜データ量が小さく、かつ、勝ち、負け、引き分けのパターンがすべて現れるm,n,k-ゲームを探索するために、m,n,k-ゲームプログラ

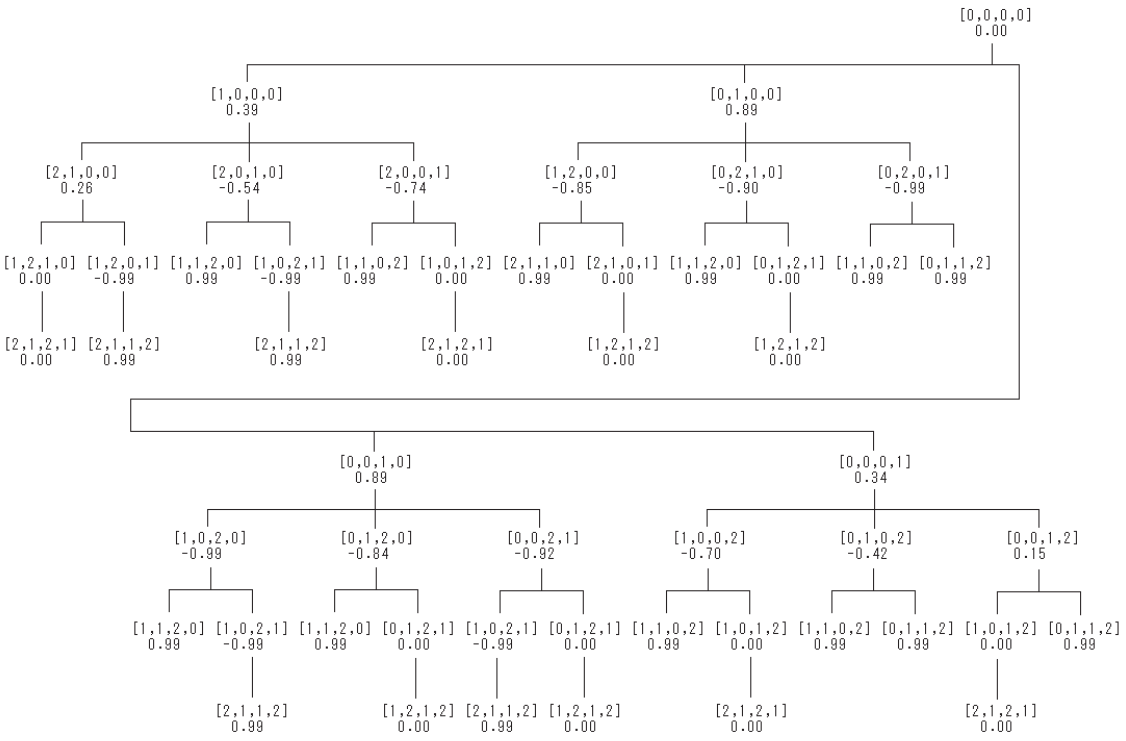

図４　4,1,2-ゲームの200万回Q学習実施後のQ値

ムを開発した。人間プレイヤー、ランダムプレイヤー、先読みランダムプレイヤー、Q学習プレイヤーを用意し、プレイヤーを組み合わせて複数回対戦させることができるため、m,n,k-ゲームを比較するのに大変役立った。

課題としては、Q学習の抱える問題でもあるが、盤面を構成するmとnを大きくすると、必要なメモリーの容量および収束までの計算量が爆発してしまい計算不可能になることが挙げられる。これについては、Q関数をニューラルネットワークで実装することで解決することができる。将来、DQNプレイヤーを実装することによって回避したいと考える。

また、Q学習の理解に使われることが多い三目並べの代わりに、棋譜データ量が三目並べより小さく、かつ、勝ち、負け、引き分けのパターンがすべて現れる4,1,2-ゲームを使うことを提案し、図３ですべての棋譜を、図７で200万回Q学習実施後のQ値を示した。これは、4,1,2-ゲームが単純であるから可能なことであり、三目並べでは大変困難な作業になる。Q学習を理解しようとする人の指針となれば幸いである。

注

1．Watkins, C.J.C.H.(1989), Learning from Delayed Rewards, PhD thesis, Cambridge University, Cambridge, England.

2．ε-greedy法は，強化学習において最適なアクションを効率よく学習するための手法であり，学習初期は学習結果（Q値）に基づいた行動を行わず，できるだけランダムに行動を行い行動に対する結果を広く確認し，学習が進むと，次第に学習結果に基づいた行動を行う手法のことである。

3．1899年，黒岩涙香が，禁手のない初期ルールの五目並べの必勝法を発表している。

4．二人零和有限確定完全情報ゲームは，ゲーム理論によるゲームの分類の1つであり，
・二人：プレイヤーの数が二人
・零和：プレイヤー間の利害が完全に対立し，一方のプレイヤーが利益を得ると，他方のプレイヤーがそれと同量の損失を受ける
・有限：ゲームが必ず有限の手番で終了する
・確定：ランダムな要素が存在しない
・完全情報：全ての情報が両方のプレイヤーに公開されている
という特徴を満たすゲームのことである。岡田章(1996),「ゲーム理論」，東京：有斐閣参照。

〔2021. 9. 16　受理〕

コントリビューター：染岡　慎一　教授
（造形デザイン学科）