報酬と失敗コストを導入した数当てゲーム

吉岡 陸[1,a)]　櫻井 祐子[1,b)]　小山 聡[2,c)]　篠田 正人[3,d)]

概要：本論文では，数当てゲームおいて，回答者に対して失敗コストを導入した新たなゲームを提案し，そのゲームでの戦略を分析する．より具体的には，出題者は n 個の数字から 1 つ選び，回答者はその数字を当てるという数当てゲームにおいて，回答者は数字を当てれば報酬を得られる一方で，失敗する度にペナルティを支払わなければならないとする．本論文では，このゲームで，出題者の Min-Max 戦略と回答者の Max-Min 戦略を数学的に算出するだけでなく，Minimax Q 学習とクラウドソーシングを用いたアンケートを用いて分析を行う．

Number-Guessing Game introducing a Reward and a Failure Cost

Riku Yoshioka[1,a)]　Yuko Sakurai[1,b)]　Satoshi Oyama[2,c)]　Masato Shinoda[3,d)]

Abstract: We propose a new variant of number-guessing games by cost of failure then consider the strategies in this game. In more detail, a codemaker selects 1 number from 1 to n as her private information then a codebreaker guesses the number. While a codebreaker receives the the number as her reward When she hits the number selected by codemaker, she have to pay a cost for each failed guess. We first mathematically obtain a codemaker's Min-Max strategy and a codebreaker's Max-Min strategy and explore them using Minimax Q Learning. We also show the experimental result using crowdsourcing.

1. はじめに

数当てゲームは，出題者はいくつかの数字を用いた数字列を回答者に対して分からないように選択し，回答者は出題者によって選択された数字列を与えられたヒントをもとに当てるというゲームである，代表的な数当てゲームとして，Master Mind や Hit-and-Blow が古くから知られている．また，近年も数多くのユーザがオンラインゲームで楽しんでいる．これらのゲームでは，出題者が回答者の回答に対して使われている数字や場所の正しい個数を答え，このヒントをもとに回答者が修正して回答を正解となるまで繰り返す．数当てゲームは数学的にも興味深い性質を持つため，ゲームやその一般化に関していくつかの分析が行われている [1], [2], [3], [5], [6].

本論文では，出題者からのヒントを与えない数当てゲームを考える．本ゲームでは，異なる n 個の数字を対象とする．出題者はその中から数字を 1 つ選択し，回答者はその数字を当てる．回答者は当てた数字の額に対応した報酬を受け取ることができるとする．回答者は数を当てるまで回答することができるが，数当てに失敗したときはコストを支払うとする．このように，我々は，失敗コストを考慮した新たな数当てゲームを提案する．

直感的には，出題者ができるだけ小さい数を選択した方がよく，一方，回答者はなるべく大きい数を回答することが予想される．しかしながら，我々は失敗コストに関して条件を付与した下で，出題者の Min-Max 戦略と回答者の Max-Min 戦略を算出した結果，出題者は各数字を等確率で選び，一方，回答者は，出題者が等確率で各数字を選ぶにも関わらず，リスクを避けて，低い数字を選ぶ確率が多いことが分かった．

[1] 名古屋工業大学
Nagoya Institute of Technology
[2] 北海道大学
Hokkaido University
[3] 奈良女子大学
Nara Women's University
[a)] r.yoshioka.272@stn.nitech.ac.jp
[b)] sakurai@nitech.ac.jp
[c)] oyama@ist.hokudai.ac.jp
[d)] shinoda@cc.nara-wu.ac.jp

表 1 回答者の利得表 ($\alpha_1 > \alpha_2 > \alpha_3$).
括弧内の記号は各手番を取る確率を示す.

回答者＼出題者	α_1 (p_1)	α_2 (p_2)	α_3 (p_3)
$\alpha_1 \to \alpha_2 \to \alpha_3$ (q_1)	α_1	$\alpha_2 - k$	$\alpha_3 - 2k$
$\alpha_1 \to \alpha_3 \to \alpha_2$ (q_2)	α_1	$\alpha_2 - 2k$	$\alpha_3 - k$
$\alpha_2 \to \alpha_1 \to \alpha_3$ (q_3)	$\alpha_1 - k$	α_2	$\alpha_3 - 2k$
$\alpha_2 \to \alpha_3 \to \alpha_1$ (q_4)	$\alpha_1 - 2k$	α_2	$\alpha_3 - k$
$\alpha_3 \to \alpha_1 \to \alpha_2$ (q_5)	$\alpha_1 - k$	$\alpha_2 - 2k$	α_3
$\alpha_3 \to \alpha_2 \to \alpha_1$ (q_6)	$\alpha_1 - 2k$	$\alpha_2 - k$	α_3

2. 同時手番数当てゲーム

本章で，報酬と失敗コスト付きの新たな数当てゲームの提案を行う．ここでは，回答者と出題者共にゲーム開始前に全ての意思決定を行う，同時手番ゲームを考える．

ゲームを具体的に示すために，3 つの数字における数当てゲームを対象にゲームの定義を与える．なお，数字の種類を増やした，一般の場合においても議論の拡張が可能である．

2.1 ゲームの定義

正解の候補となる数字は α_1，α_2，α_3 とする ($\alpha_1 > \alpha_2 > \alpha_3$)．回答者は回答が外れるごとに出題者へペナルティ k を支払う．本ゲームはゼロサムゲームであり，表 1 に回答者の利得表を示す．出題者は $\alpha_1, \alpha_2, \alpha_3$ からいずれかの数字を選ぶ．回答者は回答の順序として 6 通りから 1 つを選択する．なお，ペナルティ k は，

$$\begin{aligned} k &> \alpha_1 - \alpha_2 \\ k &> \alpha_2 - \alpha_3 \end{aligned} \tag{1}$$

を満たすとする．この条件を付与しない場合，出題者が後述する Min-Max 戦略において選択しない数字が生じる．

表 1 について簡単に説明する．例えば，出題者が α_3 を選んでいるとき，回答者は α_3 を当てるまで回答を続けることとなる．もしこのとき，回答者が $\alpha_1 \to \alpha_3 \to \alpha_2$ という手番を選んだ場合，第 1 回目の回答として α_1 を言うが，間違っているため，ペナルティ k が生じる．第 2 回目の回答は α_3 であるため，ここでゲームが終了する．回答者が得る報酬は $\alpha_3 - k$ であり，出題者の損失は $\alpha_3 - k$ である．また，出題者が α_2 を選んでいるとき，回答者が同じように $\alpha_1 \to \alpha_3 \to \alpha_2$ という手番を選んだ場合，第 1 回目の回答として α_1 を言うが，間違っているため，ペナルティ k が生じる．第 2 回目の回答は α_3 であるが，間違っているため，更にペナルティ k が生じる．第 3 回目の回答まで進み，最後に α_2 を当てることになる．回答者が得る報酬は $\alpha_2 - 2k$ であり，出題者の損失は $\alpha_2 - 2k$ である．

2.2 均衡戦略の理論的分析

本節では，提案ゲームの均衡戦略の理論的分析を行う．

出題者は損失の最小化を目指し，最大の損失の保証水準を可能な限り小さくする戦略 (Min-Max 戦略) を取る．一方，回答者は利益の最大化を目指し，相手がどのような戦略を選んでも確実に得られる利得の保証水準を可能な限り大きくする戦略 (Max-Min 戦略) を取る．

表 1 に示すように，出題者が α_1，α_2，α_3 をそれぞれ選ぶ確率を p_1，p_2，p_3 とし，$\vec{p} = (p_1, p_2, p_3)$，$\sum_{i=1}^{3} p_i = 1$ とする．回答者は，利得表に記載されている手番をそれぞれ確率 q_j で選ぶとし，$\vec{q} = (q_1, \ldots, q_6)$，$\sum_{j=1}^{6} q_j = 1$ を満たすように決定する．

このとき，回答者の報酬の期待値 $\pi(\vec{p}, \vec{q})$ は下記のように計算できる．

$$\begin{aligned} \pi(\vec{p}, \vec{q}) &= \{\alpha_1 - k(q_3 + q_5 + 2q_4 + 2q_6)\}p_1 \\ &= \{\alpha_2 - k(q_1 + q_6 + 2q_2 + 2q_5)\}p_2 \\ &= \{\alpha_3 - k(q_2 + q_4 + 2q_1 + 2q_3)\}p_3 \end{aligned} \tag{2}$$

ここで，出題者の立場に立って考えると，各手番の確率 p_1，p_2，p_3 の係数となっている値がそれぞれの確率を取るときの損失を決定する値である．もしこの 3 つの係数の値の中で大小関係が存在すれば，出題者はその値に応じて各確率を決定する，より具体的には，最も高い値を取る係数を持つ確率を低くし，最も小さい値を取る係数を持つ確率を高くする．従って，回答者は全ての係数が一致する確率を選ぶことが Max-Min 戦略となり，その時の期待利得は

$$\frac{\alpha_1 + \alpha_2 + \alpha_3 - 3k}{3} \tag{3}$$

となる．

また，各手番を取る確率 $\vec{q}$ について，いくつかの可能性が考えられる．例えば，$\alpha_1 - \alpha_3 < k/2$ のように，ペナルティ k の値が，$\alpha_1, \alpha_2, \alpha_1 - \alpha_3$ の差に比べて小さいときは

$$\begin{aligned} q_1 &= \frac{1}{6} - \frac{\alpha_2 - \alpha_3}{3k}, \quad q_2 = \frac{1}{6} + \frac{\alpha_2 - \alpha_3}{3k}, \\ q_3 &= \frac{1}{6} - \frac{\alpha_1 - \alpha_3}{3k}, \quad q_4 = \frac{1}{6} + \frac{\alpha_1 - \alpha_2}{3k}, \\ q_5 &= \frac{1}{6} - \frac{\alpha_1 - \alpha_2}{3k}, \quad q_6 = \frac{1}{6} + \frac{\alpha_1 - \alpha_3}{3k} \end{aligned}$$

がある．

また，k が $\alpha_1 - \alpha_2$，$\alpha_2 - \alpha_3$ と同程度の値のとき，

$$\begin{aligned} q_1 &= 0, \quad q_2 = \frac{1}{2} - \frac{2\alpha_1 - \alpha_2 - \alpha_3}{6k}, \\ q_3 &= 0, \quad q_4 = \frac{1}{2} - \frac{\alpha_2 - \alpha_3}{6k}, \\ q_5 &= 0, \quad q_6 = \frac{\alpha_1 - \alpha_3}{3k} \end{aligned}$$

がある．

一方，出題者の損失の期待値も式 (3) となり，このときの確率は下記となる．

$$p_1 = \frac{1}{3}, \quad p_2 = \frac{1}{3}, \quad p_3 = \frac{1}{3}$$

以上より，3 つの数字の場合，出題者は条件式 (1) を満たす利得表では，各値に依存せずに等確率で各数字を選ぶことが Min-Max 均衡戦略となる．一方，回答者の Max-Min 均衡戦略では，出題者が等確率で各数字を選ぶにも関わらず，リスクを避けて低い数字を選ぶ確率が高くなる．

なお，別の考察として，確率的に戦略を決定するのではなく，出題者と回答者各々が決定的に，つまり，1 つの Min-Max 戦略と Max-Min 戦略を決める場合，出題者は α_3 を取り，回答者は $\alpha_3 \to \alpha_2 \to \alpha_1$ を取ることとなる．

最後に，n 枚と一般的に考えた場合の結果を示す．数字を $\alpha_1, \alpha_2, \ldots, \alpha_n$ $(\alpha_1 > \alpha_2 > \ldots > \alpha_n)$ とし，1 回あたりのペナルティを k とする．このとき，回答者の報酬及び，出題者の損失の期待値は次のように計算できる．

$$\pi(\vec{p}, \vec{q}) = \frac{\sum_{l=1}^{n} \alpha_l - \sum_{l'=1}^{n-1} l' \times k}{n} \tag{4}$$

また，出題者の Min-Max 均衡戦略は

$$\begin{aligned} k &> \alpha_l - \alpha_2 \\ &\cdots \\ k &> \alpha_{n-1} - \alpha_n \end{aligned} \tag{5}$$

の条件下において，

$$\frac{1}{n} \tag{6}$$

の等確率となる．

3. Min-Max Q 学習

強化学習における Q 学習と Min-Max 戦略を組み合わせた学習方法として，Min-Max Q 学習が提案されている [4]．本論文では，Min-Max Q 学習を用いて，出題者と回答者の戦略を決定した場合の戦略の推移を分析する．

本論文では，3 つの数字の場合を対象にして実験を行う．特に，ここでは，攻撃者の利得が全て正の場合での結果をしめす．具体的な数値としては，$\alpha_1 = 100$，$\alpha_2 = 80$，$\alpha_3 = 70$，$k = 30$ とした．図 1 に Q 値の推移，図 3 に出題者の各確率の推移，図 2 に回答者の各確率の推移，図 4 に期待利得の推移について，50,000 エピソードまでの結果を示す．

Q 値は 20,000 エピソードまでに収束しており，出題者の各確率も 1/3 に収束している．図 4 に示すように期待利得も収束している．一方，回答者の確率は収束していない．これは，6 つの変数に対して，3 つの制約条件しかないため，ある意味，変数が取りえる値の自由度が高いことからこのような結果となっていることが推察される．回答者の報酬の期待値は $(100 + 80 + 70 - 3 \cdot 30)/3 = 53.33\ldots$ であるが，強化学習の結果も報酬の期待値はほとんどこの数値に等しい結果が得られた．

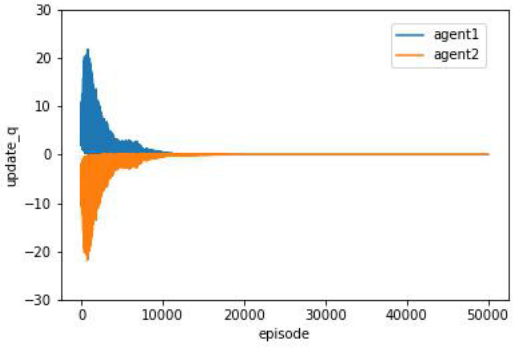

図 1 Q 値の推移

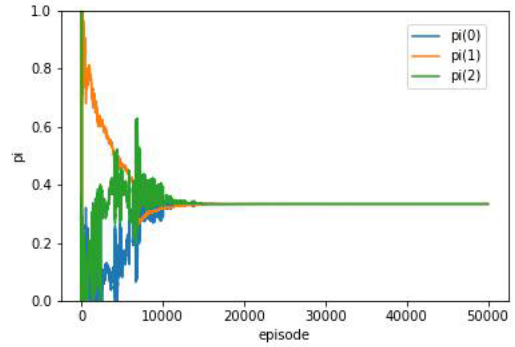

図 2 出題者の各確率の推移

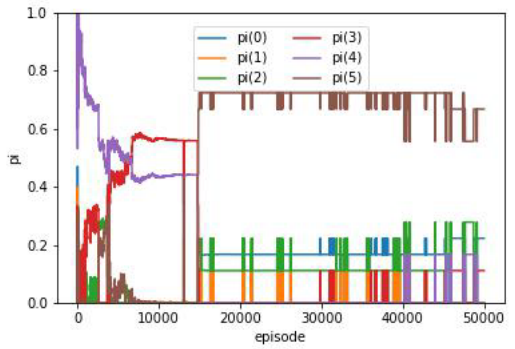

図 3 回答者の各確率の推移

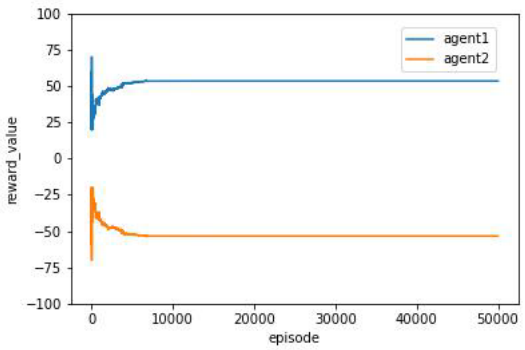

図 4 期待利得の推移

図 5 ランサーズのタスク画面

表 2 出題者の場合に取る手番 (各手番を選んだワーカ数)

100	20
80	24
70	56

表 3 回答者の場合に取る手番 (各手番を選んだワーカ数)

100 → 80 → 70	25
100 → 70 → 80	7
80 → 100 → 70	12
80 → 70 → 100	15
70 → 100 → 80	10
70 → 80 → 100	31

4. クラウドソーシングを用いた実験

本章では，クラウドソーシングを用いて，人々がこのゲームを実施するとき，実際にどのような戦略を取るのかを調べる実験を行った．

図 5 に示すように，まず具体的なゲームの流れを説明し，その後，出題者の場合に取る手番，回答者の場合に取る手番を選択させた．また，その手番を取った理由も答えさせた．本実験ではワーカを 100 名集めた．ゲームの具体的な数値は，3 章で行った強化学習のインスタンスと同様に，$\alpha_1 = 100$，$\alpha_2 = 80$，$\alpha_3 = 70$，$k = 30$ とした．

表 2 に，出題者の場合において各手番を何人のワーカが選んだかを示す．表に示されるように，一番小さい値の 70 を選択したワーカが最も多かった．選択した理由として，損失が最も小さいという理由が多かった．一方，100 を選んだ理由として，回答者が一番最初に最も大きい数字である 100 を選ばないと思ったからという回答が複数あった．

表 3 に，回答者の場合において各手番を何人のワーカが選んだかの人数を示す．出題者の場合と異なり，回答者の立場となった場合，選択する手番に大きな偏りが生じなかった．特に，100 → 80 → 70 を選択したワーカが 25 名もいたことは興味深い．なるべく高い報酬を得たいと考えるという理由が最も多かった．一方，31 名のワーカが 70 → 80 → 100 を選んだ．そのの理由として，出題者はなるべく損失を少なくしたいと思うからという理由を記載したワーカが複数存在した．

理由として，感情論を記載するワーカも存在したが，出題者は損失を最小化したいとというような，論理的な考察を踏まえての回答が予想以上に多く存在した．

5. おわりに

本稿では，報酬と失敗コスト付きの新たな数当てゲームの提案を行った．さらに，同時手番ゲームとした場合の戦略の理論的解析，強化学習による均衡戦略の導出を行うと共に，クラウドソーシングを用いて，人々がこのゲームを行う際の戦略の分析を行った．

今後は，報酬と失敗コスト付きの数当てゲームを逐次手番ゲームとして行う際の戦略の分析を行う．さらに，人間がこのゲームを繰り返し実施する場合，どのような戦略を取るかの検証を行うと共に，理論的な解析との比較を行う．また，現在のゲームは未だ単純なゲームであるため，より複雑なゲームとして発展できないかについても検討を行う．

参考文献

[1] Huang, L.-T. and Lin, S.-S.: Optimal Analyses for 3×n AB Games in the Worst Case, *Advances in Computer Games (ACG-09)*, pp. 170–181 (2009).

[2] Jäger, G. and Peczarski, M.: The worst case number of questions in Generalized AB game with and without white-peg answers, *Discrete Applied Mathematics*, Vol. 184, pp. 20–31 (2015).

[3] Knuth, D. E.: The Computer as Master Mind, *Recreational Mathematics*, Vol. 9, No. 1, pp. 1–6 (1976).

[4] Littman, M. L.: Markov games as a framework for multi-agent reinforcement learning, *Proceedings of the 11th International Conference on International Conference on Machine Learning (ICML-94)*, pp. 170–181 (1994).

[5] 篠田正人：$3 \times n$ AB game の最適戦略，情報処理学会論文誌， Vol. 53, No. 6, pp. 1602–1607 (2012).

[6] 田中哲朗：数当てゲーム MOO の最小質問戦略と最強戦略，pp. 202–209 (1996).