

シグナリングゲームにおけるプレイヤーの強化学習モデル

Reinforcement Learning Model for Human Behavior of Signaling Games

千邑 峻明^{*1} 荒井 幸代^{*1}
Toshiaki Chimura Sachiyo Arai

^{*1}千葉大学大学院融合理工学府 都市環境システム

Department of Urban Environment Systems, Graduate School of Science and Engineering, Chiba University

This paper examines the applicability of the reinforcement learning schema for modelling player's decision-making process within a signaling game context where one player has information the other player does not have. This situation of asymmetric information is very common in the realworld. Though many applications of signaling games have been developed to solve economic problems, the previously proposed models could not reproduce the human way of signaling. We show some interesting empirical results concerning the refinement of equilibria by the proposed reinforcement learning model.

1. はじめに

市場における取引や契約において情報の非対称性がある場合、情報をもつ側がもたない側に情報を開示する行動をとることをシグナリングという。シグナリングは、ゲーム理論においてシグナリングゲームとして研究が進められてきた。しかし、被験者実験 [Brandts-Holt 92] の結果、既存の均衡概念とその精緻化ではシグナリングゲームにおけるプレイヤーの行動を十分に予測できない場合があることがわかっている。一方、ゲーム理論における他のゲームでは、強化学習モデルにより被験者実験におけるプレイヤーの行動を再現したという報告がある。そこで本論文では、強化学習モデルを用いて Brandts らの被験者実験の結果を再現する。

2. 準備

2.1 シグナリングゲーム

図 1 に Brandts らの被験者実験に用いられたシグナリングゲームのゲーム木を示す。利得は省略している。プレイヤーは送り手か受け手の役割をもつ。ゲームは次の 4 段階からなる。

Step1. 送り手 (S) のタイプ決定

確率 $p(t) = \{p(t_H), p(t_L)\} = \{2/3, 1/3\}$ にしたがって $T = \{t_H, t_L\}$ から送り手のタイプ t が決定する。

Step2. 送り手によるメッセージ選択

送り手は自身のタイプ t を確認し、 $M = \{m_H, m_L\}$ からメッセージ m を選び、受け手に送信する。

Step3. 受け手 (R) による反応選択

受け手はメッセージ m を確認し、 $A = \{a_H, a_L\}$ から反応 a を選ぶ。

Step4. 利得確定

送り手は利得 $U_S(t, m, a)$ を、受け手は利得 $U_R(t, m, a)$ を受け取る。

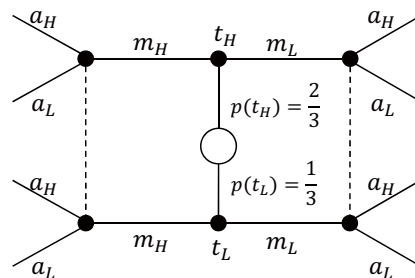


図 1: シグナリングゲームのゲーム木

なお、図 1 中のノードを結ぶ点線は受け手の情報集合を表す。このシグナリングゲームは、受け手の情報集合が二つのノードからなる不完全情報ゲームである。

2.1.1 完全ベイジアン均衡

完全ベイジアン均衡はシグナリングゲームにおける均衡概念の一つである。以下、本論文で均衡とは完全ベイジアン均衡を指す。ここで、受け手は信念に基づいて行動を選択する。信念とは、各情報集合において実際にどのノードにいるかを表す確率分布である。

定義 シグナリングゲームにおける完全ベイジアン均衡は、式 (1) を満たす送り手の戦略 σ^* 、式 (2) を満たす受け手の戦略 ρ^* 、式 (3) を満たす均衡経路上の信念 μ の組 (σ^*, ρ^*, μ) である。

$$\forall t \in T \quad \sigma^*(\cdot|t) \in \arg \max_{\sigma} U_S(t, \sigma, \rho^*) \quad (1)$$

$$\forall m \in M \quad \rho^*(\cdot|m) \in \arg \max_{\rho} \sum_t \mu(t|m) \cdot U_R(t, m, \rho) \quad (2)$$

$$\mu(t|m) = \frac{p(t)\sigma^*(m|t)}{\sum_{t' \in T} p(t')\sigma^*(m|t')} \quad (3)$$

均衡経路とは各プレイヤーが均衡戦略を用いる場合の、ゲーム木における初期点と頂点を結ぶ枝の系列である。なお、均衡経路上にない情報集合における信念はどのような確率分布であってもよい。

連絡先: 千邑峻明, 千葉大学大学院融合理工学府 都市環境システム, 千葉市稲毛区弥生町 1-33, aeta2065@chiba-u.jp

表 1: game 1 と game 3R の均衡

名称	送り手の戦略 (t_H 時のメッセージ, t_L 時のメッセージ)	受け手の戦略 (m_H に対する反応, m_L に対する反応)	受け手の信念
均衡 1	(m_H, m_H)	(a_H, a_H)	$\mu(t_L m_L) \geq 1/2$, $\mu(t_L m_H) = 1/3$
均衡 2	(m_L, m_L)	(a_H, a_H)	$\mu(t_L m_L) = 1/3$, $\mu(t_L m_H) \geq 1/2$

2.1.2 直観的基準 [Cho-Kreps 87]

シグナリングゲームには複数の均衡が存在する場合がある。その中で問題のある均衡を排除する作業を均衡の精緻化といい、代表的な精緻化手法の一つが直観的基準である。

定義 メッセージ m に対して信念 μ をもつ受け手の最適反応の集合 $BR(\mu, m)$ を式 (4) で定義する。

$$BR(\mu, m) \equiv \arg \max_a \sum_{t \in T} \mu(t|m) U_R(t, m, a) \quad (4)$$

ある均衡において、タイプ t の送り手が獲得する利得を $U_S^*(t)$ とする。均衡経路上にない各メッセージ m に対し、 $J(m)$ を式 (5) を満たすすべてのタイプの集合として定義する。

$$U_S^*(t) > \max_{a \in BR(T \setminus J(m), m)} U_S(t, m, a) \quad (5)$$

いずれかの m に対し、式 (6) を満たすタイプ t' が存在するならば、この均衡は直観的基準を満たさない。

$$\min_{a \in BR(T \setminus J(m), m)} U_S(t', m, a) > U_S^*(t') \quad (6)$$

3. Brandts らの被験者実験

Brandts らは、表 1 に示す二つの純粋戦略の均衡^{*1}が存在し、均衡 1 だけが直観的基準を満たすシグナリングゲームを複数構築し被験者実験を行った。

Brandts らは、はじめに game 1 とよばれるゲームを構築し実験した。game 1 の利得表を表 2 に示す。Brandts らは game 1 に対して送り手側、受け手側ともに 4 人の被験者を用意し、1 ラウンドごとにランダムで 4 ペアをつくりゲームをプレイさせた。game 1 で実施された 4 回の実験において、送り手が選択したメッセージと受け手が選択した反応の組（以下、プレイパターン）の生起率と集計値を図 2 に示す。図 2 より、送り手は自身のタイプが t_H の場合 m_H を、 t_L の場合 m_L を送信する傾向にある。受け手はメッセージ m_H に対して a_H を、 m_L に対して a_L をとる傾向にある。これは均衡と直観的基準による精緻化では予測されない結果である。

その後、Brandts らは game 3R とよばれるゲームを構築し実験した。game 3R の利得表を表 3 に示す。Brandts らは game 3R に対して、送り手側、受け手側ともに 6 人の被験者を用意し、1 ラウンドごとにランダムで 6 ペアをつくりゲームをプレイさせた。game 3R で実施された 2 回の実験における、各プレイパターンの生起率と集計値を図 3 に示す。図 3 より、送り手は自身のタイプが t_H の場合は m_L を、 t_L の場合はほぼランダムにメッセージを送信する傾向にある。受け手はメッセージ m_H に対してはほぼランダムに応答し、 m_L に対しては a_H をとる傾向にある。これも均衡と直観的基準による精緻化では予測されない結果である。

本論文では前述の game 1 と game 3R を対象とし、強化学習モデルにより被験者の行動が再現可能か検証する。

表 2: game 1 の利得表

S \ R	t_H		S \ R	t_L	
	a_L	a_H		a_L	a_H
m_L	20, 75	120, 125	m_L	60, 125	140, 75
m_H	60, 75	140, 125	m_H	20, 125	100, 75

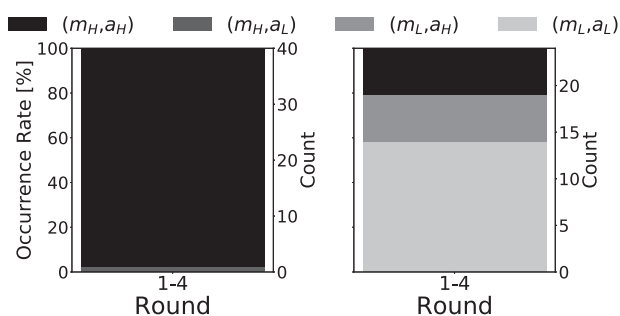
(a) 送り手が t_H の場合 (b) 送り手が t_L の場合

図 2: game 1 における被験者のプレイパターン

表 3: game 3R の利得表

S \ R	t_H		S \ R	t_L	
	a_L	a_H		a_L	a_H
m_L	160, 75	160, 175	m_L	10, 175	190, 75
m_H	10, 75	190, 175	m_H	100, 175	160, 75

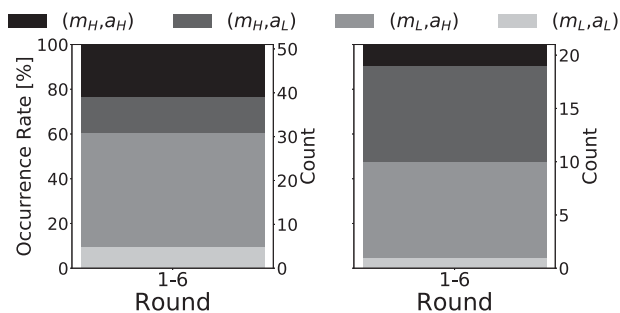
(a) 送り手が t_H の場合 (b) 送り手が t_L の場合

図 3: game 3R における被験者のプレイパターン

*1 Brandts らは逐次均衡 (Sequential Equilibrium) とよばれる別の均衡概念を用いて均衡を導出しているが、本論文で扱うゲームでは完全ベイジアン均衡でも同じ均衡が導出される。

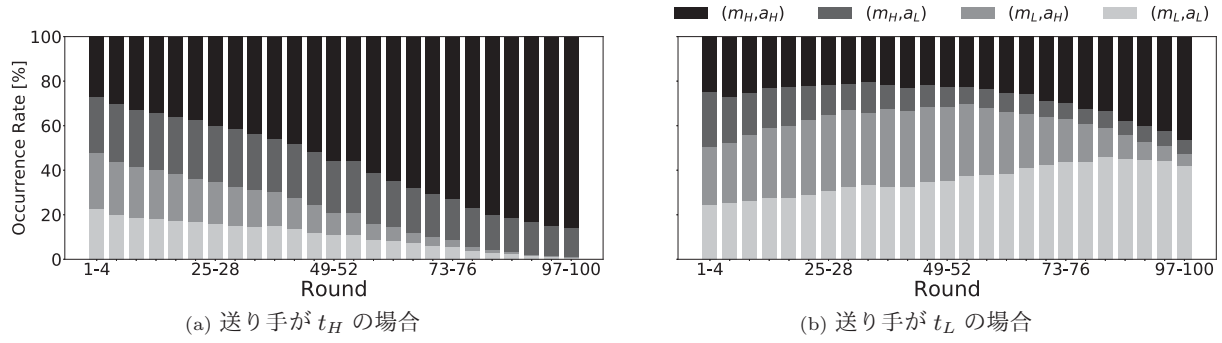


図 4: game 1 におけるエージェントのプレイパターンの変化 (500 試行の平均)

4. 提案法

4.1 強化学習

強化学習とは、意思決定主体であるエージェントが、ある環境内においてその時点での状態を観測し、数値化された報酬をもとにとるべき行動を決定する枠組みである。

本論文での強化学習モデルにおける行動価値の更新式を式 (7) に示す。学習率 α は学習の速さを表すパラメータである。

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r - Q(s, a)] \quad (7)$$

行動選択には Boltzman 選択を用いる。Boltzman 選択において、状態 s で行動 a を選択する確率 $P(a | s)$ を式 (8) に示す。

$$P(a | s) = \frac{\exp(Q(s, a)/\tau)}{\sum_{a' \in A} \exp(Q(s, a')/\tau)} \quad (8)$$

温度パラメータ τ は選択のランダム性を表す。 $\tau \rightarrow \infty$ の場合は行動選択はランダムになり、 $\tau \rightarrow 0$ の場合は行動価値が最大の行動を選択する。

4.2 エージェントモデル

シグナリングゲームにおける送り手と受け手を、前節の強化学習モデルにより学習するエージェントとする。以下、それぞれを送り手エージェントと受け手エージェントとよぶ。

送り手エージェントの状態は状態集合 $S_s = \{s_{t_H}, s_{t_L}\}$ から確率 $p(t) = \{p(s_{t_H}), p(s_{t_L})\} = \{2/3, 1/3\}$ で決定する。 s_{t_H} は自身のタイプが t_H であること、 s_{t_L} は t_L であることに対応する。行動集合は $A_s = \{a_{m_H}, a_{m_L}\}$ である。 a_{m_H} はメッセージ m_H を、 a_{m_L} は m_L を送信する行動である。報酬は表 2, 3 に示す送り手が獲得するゲームの利得とする。

受け手エージェントの状態は送り手エージェントの送信したメッセージに応じて状態集合 $S_r = \{s_{m_H}, s_{m_L}\}$ のいずれかが入力される。行動集合は $A_r = \{a_H, a_L\}$ である。報酬は表 2, 3 に示す受け手が獲得するゲームの利得とする。

5. 実験

5.1 実験設定

game 1 では送り手エージェントと受け手エージェントをそれぞれ 4 エージェント用意し、1 ラウンドごとにランダムで 4 ペアをつくりゲームを行う。game 3R では 6 エージェント用意し同様に 1 ラウンドごとにランダムで 6 ペアをつくりゲームを行う。パラメータは全エージェント学習率 $\alpha = 0.5$ 、温度パラメータ $\tau = 100 \cdot 0.995^{\text{round}-1}$ とした。

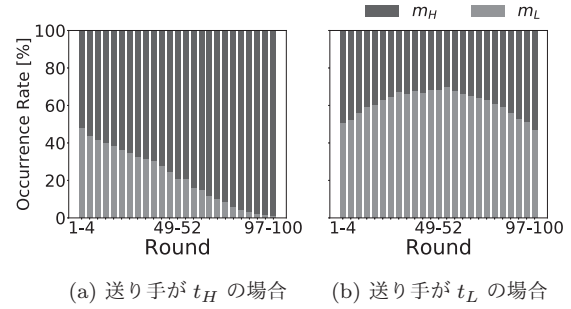


図 5: game 1 における送り手エージェントの行動選択の変化

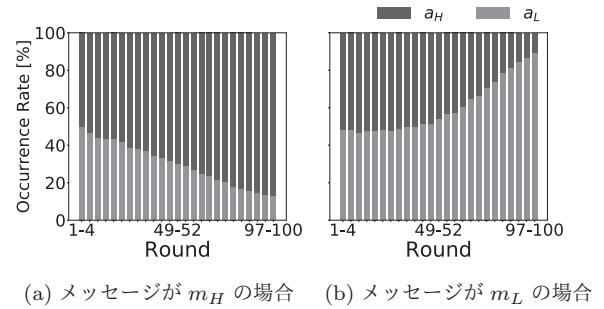


図 6: game 1 における受け手エージェントの行動選択の変化

5.2 実験結果 1 : game 1

図 4 に game 1 に対する 1 試行 100 ラウンドの実験 500 試行の、4 ラウンドごとのプレイパターンの平均生起率を示す。図 5 に送り手エージェントの平均行動選択率、図 6 に受け手エージェントの平均行動選択率を示す。図 4, 5, 6 はいずれも横軸がラウンド数、縦軸がプレイパターンおよび行動の生起率を表す。なお 4 ラウンド連続で全送り手エージェントが t_H および t_L の一方にしかならなかった場合と、メッセージ m_H および m_L の一方しか送信しなかった場合が生じた際には、その 4 ラウンドは除外して平均値を算出した。

game 1 に対する実験では、ラウンド数に乖離があるものの被験者実験と同様に t_H の場合は (m_H, a_H) が、 t_L の場合は (m_L, a_L) が一時的に最も高い生起率を示した。

図 5 より、送り手エージェントは自身のタイプ t_H の場合に m_H を送信する傾向にある。これは受け手の行動選択がランダムな場合、 m_H を送信した場合の期待利得 100 が m_L を送信した場合の期待利得 70 を上回るためだと考えられる。一方 t_L の場合には 53 - 56 ラウンドまで m_L を送信する割合が増加傾向

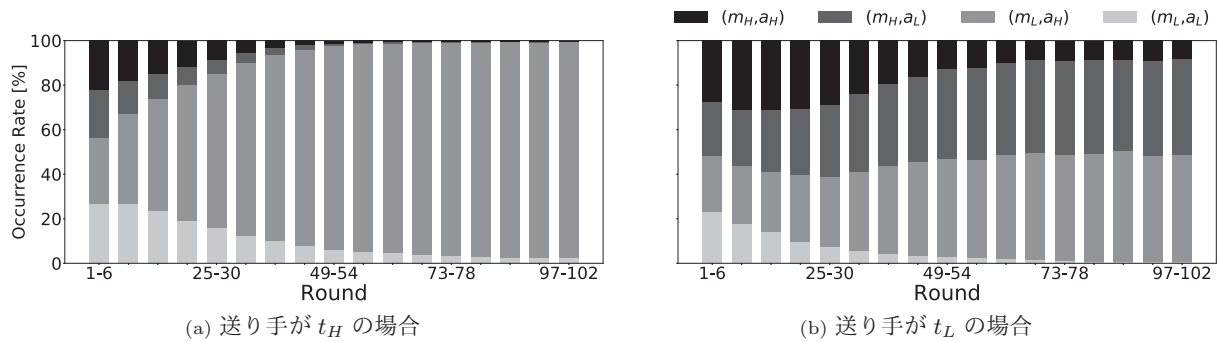


図 7: game 3R におけるエージェントのプレイパターンの変化 (500 試行の平均)

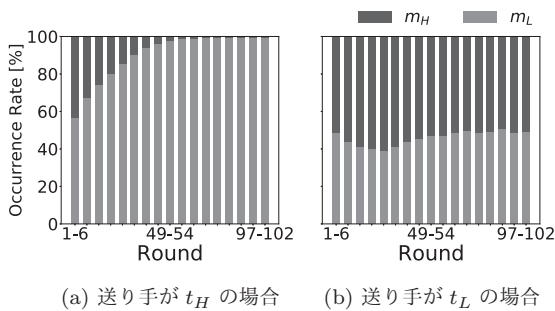


図 8: game 3R における送り手エージェントの行動選択の変化

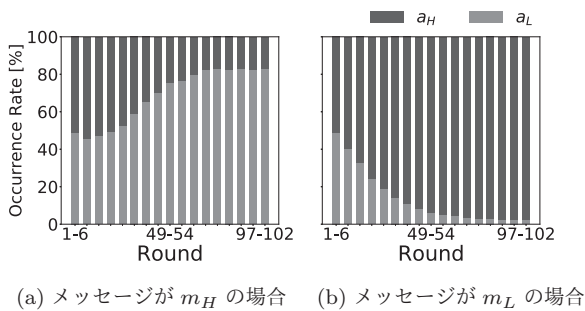


図 9: game 3R における受け手エージェントの行動選択の変化

向にある。これは m_L を送信した場合の期待利得 100 が m_H を送信した場合の期待利得 60 を上回るためだと考えられる。

図 6 より、受け手エージェントは、はじめの 1-4 ラウンドはメッセージに関わらずほぼランダムに反応を選択する傾向にある。しかし、送り手エージェントがタイプに応じて別々のメッセージを送信するようになると、次第に受け手エージェントはメッセージ m_L に対しては反応 a_L を、メッセージ m_H に対しては反応 a_H を選択するようになる。

以上の通り各役割のエージェントが学習した結果、送り手が t_H の場合は (m_H, a_H) 、送り手が t_L の場合は平均 85 - 90 ラウンドで (m_L, a_L) が最も生起率が高くなると考えられる。

5.3 実験結果 2 : game 3R

図 7 に game 3R に対する 1 試行 102 ラウンドの実験 500 試行の、6 ラウンドごとのプレイパターンの平均生起率を示す。図 8 に送り手エージェントの平均行動選択率、図 9 に受け手エージェントの平均行動選択率を示す。図 7, 8, 9 はいずれも横軸がラウンド数、縦軸がプレイパターンおよび行動の生起率を表す。なお 6 ラウンド連続で全送り手エージェントが t_H お

よび t_L の一方にしかならなかった場合と、メッセージ m_H および m_L の一方しか送信しなかった場合が生じた際には、その 6 ラウンドは除外して平均値を算出した。

game 3R に対する実験でもラウンド数に乖離があるものの、被験者実験と同様に t_H の場合は (m_L, a_H) が、 t_L の場合は (m_H, a_L) と (m_L, a_H) が高い生起率を示した。

図 8 より、送り手エージェントは自身のタイプ t_H の場合に m_L を送信する傾向にある。これは受け手の行動選択がランダムな場合、 m_L を送信した場合の期待利得 160 が m_H を送信した場合の期待利得 100 を上回るためだと考えられる。一方 t_L の場合には 31 - 36 ラウンドまで m_H を送信する割合がやや増加傾向にある。これは m_H を送信した場合の期待利得 130 が m_L を送信した場合の期待利得 100 を上回るためだと考えられる。ただしこの傾向は t_H の場合ほど顕著ではなく、試行により m_H か m_L の一方だけが半数以上を占める場合があり、各ラウンドの両メッセージの平均行動選択率はほぼ等しい。

図 9 より、受け手エージェントはメッセージ m_H に対しては 25 - 30 ラウンドまでほぼランダムに行動選択する傾向にある。一方 m_L に対しては、送り手エージェントの行動選択に対応して早期から反応 a_H を選択する。その後、メッセージ m_H に対しては反応 a_L を選択するようになる。

以上の通り各役割のエージェントが学習した結果、送り手が t_H の場合は (m_L, a_H) が、送り手が t_L の場合は (m_H, a_L) と (m_L, a_H) が高い生起率を示したと考えられる。

6. 結論および今後の課題

本論文では強化学習モデルによって、シグナリングゲームにおける代表的な被験者実験である Brandts らの被験者実験の結果を再現した。本論文で対象とした二つのゲームに関しては被験者実験においてプレイヤーがとる行動の傾向を再現することができた。今後の課題として、被験者と同じプレイ回数で学習可能なモデルの構築を挙げる。

参考文献

- [澤木 14] 澤木久之：シグナリングのゲーム理論，勁草書房 (2014)。
- [Brandts-Holt 92] Brandts and Charles A Holt : An experimental test of equilibrium dominance in signaling games, *The American Economic Review*, Vol.82, No.5, pp.1350-1365 (1992)。
- [Cho-Kreps 87] In-Koo Cho and David M Kreps : Signaling games and stable equilibria, *The Quarterly Journal of Economics*, Vol.102, No.2, pp.179-221 (1987)。