強化学習の社会性：バンディット問題と目標設定理論
Sociality in Reinforcement Learning: Bandit problems and goal-setting theory

高橋 達二
Tatsuji Takahashi
東京電機大学, 理研 AIP
Tokyo Denki University, RIKEN AIP
tatsuji.takahashi@gmail.com

概要

不確実性の下の環境探索と適切な行動の学習に関し、従来は最も適切な (環境から得られる報酬を最大化) 行動の獲得が評価される。しかし実世界課題では多くの場合、単位を取る (60 点以上獲得)、黒字化する (損益を 0 以上に) などの目標の達成との関係で行動が評価される。そこで、バンディット問題において目標設定理論の予測 (具体的で高い目標がパフォーマンスを向上させる) が正しいかどうかを検証する。結果は、目標設定理論を弱く支持した。

キーワード：希求水準, 満足化, 限定合理性

1. はじめに

人間や動物の学習の大部分は個体群や社会の他個体から・他個体とともに行われ、また学習の結果も他個体に向けて・対して出力される。この意味で人間や動物の学習は本質的に社会的であるが、その点のモデリングはまだ萌芽段階である (例外として Najar et al. (2020)、レビューとして内藤 and 亀田 (2022) がある)。

学習の一形式として、エージェントが自ら試行錯誤をし、適切な行動の系列を学んでいく、という意味で、最も自律性の高い機械学習技術である強化学習は、デジタルゲームやボードゲームで人間のパフォーマンスを超え、また生成 AI のチューニングにも用いられる。ここでは、近年認知心理学でよく用いられるようになった、強化学習で最もシンプルな多本腕バンディット問題の実験を通じて、目標設定理論が不確実な強化学習形式でも妥当であるのかを検証する。

2. 目標設定理論

目標設定理論 (goal-setting theory) とは、タスクにおけるパフォーマンスと、タスク内での目標の設定との関係を扱う組織行動学・心理学の理論である (Locke and Latham (1990) の 1,2 章にこの理論の歴史と基本的な発見がまとまっている)。行動の理解には目標志向性が重要であり、人間の行動が目標や意図に導かれ、制御されている、という前提の下、目標に影響する要因や、目標が行動やパフォーマンスに対して持つ関係を特定しようとする。より概念的に言えば、エージェントが目標を設定し（不均衡を創造・導入し)、そしてその目標と現状との不均衡を解消しようとするプロセスに着目する。また目標（未来についての考え・欲される最終状態）が行動に因果的な役割を持つとする。

目標設定理論の基本的な経験的結果として、人間は「できる限り頑張れ」という非具体的な目標 (後述の実験条件では “最”) や、「楽な目標を達成せよ」と言われる (“易”) よりも、達成が困難だが不可能ではない、つまり野心的で、かつ具体的な「目標を達成せよ」と言われ (“適”)、それにコミットする方がパフォーマンスが上がる。目標は具体的・特定的である方が良く (“最” よりも “適” が良い)、また困難だが不可能ではないような高い目標がパフォーマンスを向上させる (“低” よりも “適” が良い)。

3. 多本腕バンディット問題

K 本腕バンディット問題 (K-armed bandit problems) (Sutton and Barto, 2018) は、人間や動物が、不確実性の下で、どのように環境を探索し、試行錯誤をして適応していくかを調べるためのタスクとして、神経科学や心理学でも最近盛んに用いられている (Daw et al., 2006; Ohta et al., 2021)。

本研究で扱う報酬が有りか無しかの二値的な場合 (ベルヌーイバンディット) は以下のようなものである。参加者の目の前には 4 つの異なる色のカードのデッキがある ($K = 4$)。各デッキには 10 枚の同色のカードが積まれている。カードは、裏面は単色であり、おもて面は「あたり」「はずれ」の 2 種類である。色 $i(= 1, 2, 3, 4)$ によって、当たりの枚数（つまり当たりの比率 $p_i \in [0, 1]$）が異なっているが、p_i は未知である。色を一つ選び、デッキの一番上のカードを裏返すという行動選択は、T 回行うことができる。おもて面がそのまま結果である。また、毎回全 4 デッキの左右の並びも、各デッキのカードの順序もシャッフルされる（復元抽出)。よって、p_i は常に一定である。

参加者が通常指示されるのは当たり獲得回数の最大化である。つまり、4 色のうちからいずれかの色を選んでいくことを通じて、どの色が当たりやすいのかを見極めつつ、当たりの獲得回数をなるべく多くしなくてはならない。ある色の選択に固執してしまうと、より当たりやすい色を見落とす必要がある。なので、ある程度は、これまで相対的に試してこなかったものを中心に、色々な色を試してみる必要がある (探索 exploration が必要である)。しかし他方で、常になるべく当たりを得る必要があるので、これまでの当たり外れの経験から、各色の当たりやすさを判断し、最も当たりやすそうな色をなるべく多く選択することも重要である (知識利用 exploitation が必要である)。

4. 実験

バンディット問題における目標設定の効果を調べた 2 つの実験設定とその結果について述べる。比較するのは、「適切な目標」、「低すぎる目標」、「具体的ではないできるだけ頑張れという目標 (最適化目標)」の三つの違いが様々な環境 (報酬確率の設定) においてパフォーマンスに与える効果であり、以下の実験の記述ではそれぞれ “適”、“低”、“最” と略記される。2 実験の共通項目を述べた上で、それぞれの結果を述べる。

4.1 実験 1, 2 の共通設定

実験参加者はクラウドソーシングサイト CrowdWorks で募集された。参加者の性別や年齢の情報は直接は取得しなかった。実験は AWS 上の自作プログラムで実施された。参加者は PC を操作してタスクを完遂した。

参加者は実験 1 (2) では表 1 の 9 (6) のグループの一つにランダムに割り当てられた。グループは二つの変数によって定義されており、「略号」はそれら 2 変数を二つの漢字で表したものとしている。その二つの変数とは、3 通りの環境（報酬確率）と 3 つの目標（希求水準の有無と値）であり、3×3 (2×3) で 9 (6) グループとなる。

行動数は前述の例同様 4 で、報酬確率は表では百分率で記している。三つの確率の 4 つ組があり “易” しいのが $(0.2, 0.2, 0.4, 0.8)$ である。最適である報酬確率 0.8 の行動は、0.2 または 0.4 を報酬確率とする他の 3 つの行動との区別が容易である。“難” しいのが $(0.7, 0.7, 0.8, 0.9)$ であり、“激” しく難しいのが $(0.8, 0.8, 0.8, 0.9)$ である。

目標は、与えられない場合 (“最”) は「なるべく多くの当たりを獲得するように」と指示される。目標は与えられる場合、“低” すぎるか、目標の達成が最適行動の選択を意味する「“適” 切」な場合かの 2 種類である。目標は、「行動選択」に対する当たり回数の比率の閾値として与えられる。具体的な目標の値は、それぞれの報酬確率に対して設定される。易環境では “低” はどの 4 つの行動の報酬確率よりも低い 0.1 であり、“適” はその目標を果たすには最高の報酬確率を持つ 0.8 の腕を選択する他ないような、 0.6 という比率として与えられる。同様に、難・激環境の “低” は 0.5 であり、“適” は 0.85 である。つまり “難適” と “激適” の 2 グループの参加者にとっては、報酬確率 0.9 の最適行動をある程度多く選択することだけが (長期的・平均的には) 目標達成の手段となりうる。

表 1: 実験 1,2 のデザイン。参加者は実験 1 では 9 グループ、実験 2 では 6 グループにランダムに割り当てられた。報酬確率は % 値、回数は行動選択回数、人数は参加者数である。

			実験 1		実験 2	
グループ	報酬確率	希求水準	回数	人数	回数	人数
易最	[20,20,40,80]	なし	150	99	—	0
易低	“	0.1	“	98	—	0
易適	“	0.6	“	102	—	0
難最	[70,70,80,90]	なし	“	109	600	113
難低	“	0.5	“	110	“	111
難適	“	0.85	“	133	“	102
激最	[80,80,80,90]	なし	“	103	“	103
激低	“	0.5	“	113	“	116
激適	“	0.85	“	108	“	112
		合計:		998		653

実験の流れは 3 段階で

1. 練習 1: デッキ、色、当たり外れの確率比較に基づく意思決定の基本の理解確認を行う (図 1a)。
2. 練習 2: 行動数 3（3 色）の易しいバンディット問題での練習 (図 1b) 。報酬確率は $(0.1, 0.1, 0.9)$ となっており、最適な行動である 0.9 のデッキは見つけやすい (その色はランダムに割り当てられている)。最適行動を 5 回連続で選択すると、この練習フェイズは終了する。
3. タスク: 割り当てグループ (表 1) に応じた本番のバンディット問題 (図 1c)。

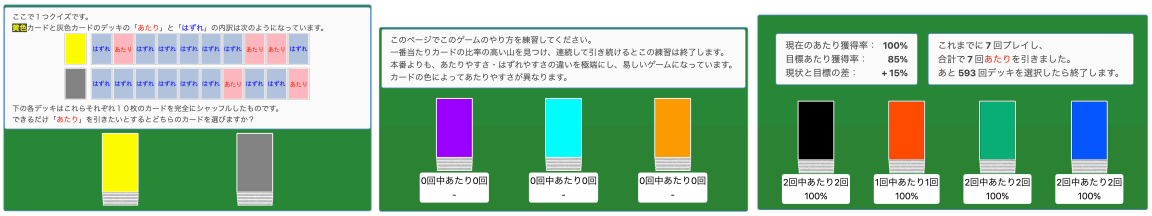
となる。

4.2 実験 1

1,000 人の参加者への謝金は 165 円 (税込) であった。このうち 2 人の実験実施に関してサーバー上の実験プログラム実行上の不具合があったため、998 人のデータを分析する。行動選択は 150 回であった。

4.2.1 結果

基本的な結果として、各グループの平均相対精度 (relative accuracy) を図 2 に示す。これは通常の精度 (accuracy) をより詳細にした指標であるので、精度から説明する。ある行動選択の精度は、その行動 i が最適行



(a) 練習 1 の画面　(b) 練習 2 の画面　(c) タスクの画面。

図 1: 練習 1 (a)、練習 2 (b)、本番タスク (c) の画面。(c) は希求水準は 0.85、選択回数は 600 回の例、つまり実験 2 のグループ “難適” や “激適” の場合である。

動 (もっとも報酬確率の高い行動、つまり $p_i = p_{\max}$) であれば 1、そうでなければ 0 である。相対精度は、その行動が最適行動でも報酬確率最低の行動でもない場合には、0 以上 1 以下の値として $(p_i - p_{\min})/(p_{\max} - p_{\min})$ とする。例えば $(0.2, 0.2, 0.4, 0.8)$ の環境で報酬確率 $0.2, 0.4, 0.8$ のそれぞれの行動を選んだ時の相対精度は $0, 1/3, 1$ となり、最適行動以外を選んでいるときの選択の相対的な良さも評価できる点で精度よりも優れている。他方で、後述のリグレットと異なり、値が $[0, 1]$ に収まり、学習が成功裡に進めば値が 1 に近づいていくという性質を持つため、異なる環境や参加者の間での比較がしやすい。

図 4 にタスク終了時の「最終リグレット」の値のグループごとの分布を示す。リグレットとは、タスクを見渡した立場からの「後悔」、期待損失の総計のことで、“易” 環境で報酬確率 $0.2, 0.4, 0.8$ のそれぞれの行動を一回選んだ場合の期待損失は $0.6, 0.2, 0$ となる。行動選択終了時点での累積 (実験 1 では 150 回) が最終リグレットである。

4.3 実験 2

実験 2 は、実験 1 の “難”・“激” 環境では選択回数 150 回という短さのため明確なパフォーマンス差が出なかったという可能性を検証するために行われた。660 人の参加者への謝金は 495 円 (税込) であった。実験プログラムの不具合のため、653 人分のデータを用いた。実験 1 との違いは、行動選択回数が 4 倍の 600 回となっており、また環境として “難” と “激” のみ扱っているという 2 点のみである。実験 1 と同様、平均相対精度と最終リグレット (600 回行動選択完了時) の結果をそれぞれ図 3, 5 に示す。

5. 議論

実験 1,2 の結果から、多本腕バンディット問題における目標設定の効果について議論する。実験 1 ではリグレットの比較により、パフォーマンスの高さ (リグレットの小ささを指標としている) では易適 > 易最 > 易低という結果が得られた。このうち t 検定 5% 水準で有意差があるものは「易適 > 易低」のみであった ($p = 0.0178$)。“難” 環境においても平均値では同様に難適 > 難最 > 難低という関係があるが、有意ではない。“激” 環境では激最 > 激適 > 激低と順位が入れ替わっているが、有意性はない。このようにはっきりした目標設定の効果が “易” においてのみ見られたのは、他の二つの環境では、行動ごとの報酬確率の差が微細なため、150 回の選択回数では不足していたことが理由として考えられる。

そこで、選択回数を 4 倍の 600 回に増やした実験 2 で同様の分析をしてみると、平均値では難適 > 難最 > 難低、激適 > 激最 > 激低という目標設定理論の予測通りの結果になっている。ただし同様の基準で有意差があるのは難適 > 難低のみ ($p = 0.001$) であった。

結果は目標設定理論の仮説に沿った傾向を見せているが、支持は弱い。その原因としては、そもそもベルヌーイバンディットの動作の分散の大きさがある。シミュレーションでアルゴリズムの性能を比較する際にも、シミュレーション数を 100 回以上にした上で、数千回以上の選択回数としないとはっきりした差が出ないのが通例である。この点については分散を抑えるために、実際の当たり外れが報酬確率と近くなるような確率の設定の仕方 (理論上の報酬の分布を作りシャッフルにより確率化) することができる。また、“易” 以外の “難”・“激” 環境ではそもそも、実験 2 のように選択回数を増やしても、最適行動を見つけることが難しすぎたということが考えられる。最適行動を見つける、つまりせいぜい 0.2 や 0.1 の報酬確率の差を検知し、それを確立する (最適行動が最適であることを十分に検証した上でその選択を続ける) ことに十分な意味が見出せなければ、そのような困難な行動方策を取ること自体が難しいと思われる。これらの点については今後の実験で操作・検証していく。

ところで、本研究では当たり率の表示が行動に与える影響についての分析を行わなかった。各グループの約半数には当たり率 (と頻度情報) が表示され、残り

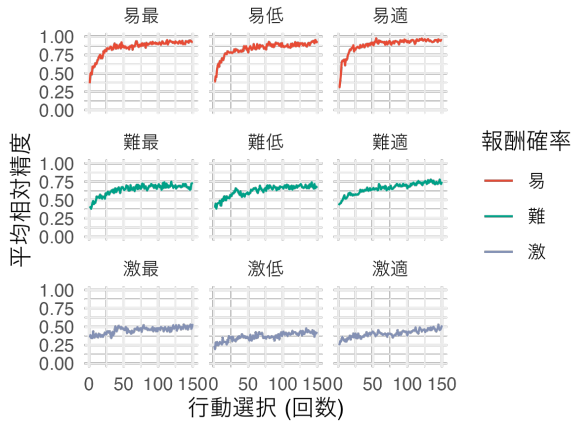


図 2: 実験 1 の各グループ平均相対精度の時間発展

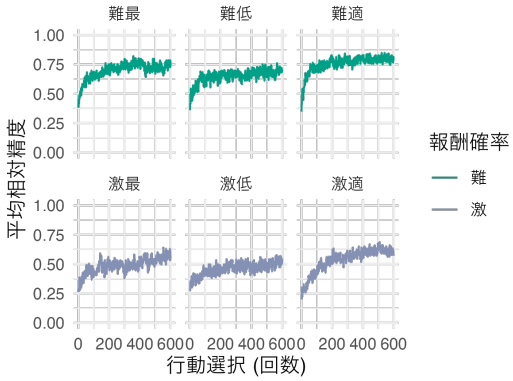


図 3: 実験 2 の平均相対精度の時間発展

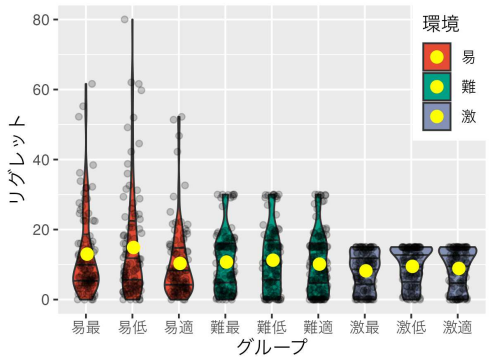


図 4: 実験 1 のタスク完了時 (行動選択 600 回完了) 時のリグレット、黄色の円は平均値。

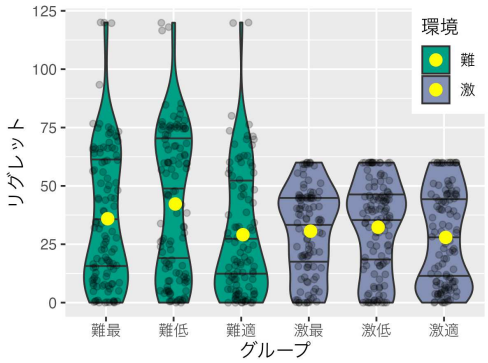


図 5: 実験 2 のタスク完了時 (行動選択 600 回完了) 時のリグレット、黄色の円は平均値。

には表示されない。表示の有無によるパフォーマンスの違いには一貫した傾向も有意差もなかったため、本論文では両者をマージして分析したが、両者で行動価値の更新方法や忘却の効果についての違いが予想される。また、本研究では、目標が参加者に与えられた場合は、その目標は「600 回の行動選択のうち (その 85% にあたる) 510 回の当たりを目指せ」のような、「報酬獲得率」についての閾値であった。これに対し、「当たり確率が 85% 以上の腕を見つけてください」という形式の目標 (希求水準) を与えることで、Simon のいう満足化 satisficing をターゲットとして扱うこともでき、強化学習状況における限定合理性についての研究が可能となる。この目的に対応するのは、従来の報酬最大化と異なる最適腕識別 (optimal arm identification) という問題枠組みで、A/B テストや臨床実験を含み、多くの実世界課題がこちらの目的を持つため重要であるが、先行研究がない。これら当たり率表示の有無や目的 (報酬最大化 vs. 最適腕識別) に関する違いの実験操作・分析を今後の課題とする。

文　献

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., and Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–9.

Locke, E. and Latham, G. (1990). *A Theory of Goal Setting & Task Performance*, volume 16. Prentice-Hall.

Najar, A., Bonnet, E., Bahrami, B., and Palminteri, S. (2020). The actions of others act as a pseudo-reward to drive imitation in the context of social reinforcement learning. *PLOS Biology*, 18(12):e3001028.

Ohta, H., Satori, K., Takarada, Y., Arake, M., Ishizuka, T., Morimoto, Y., and Takahashi, T. (2021). The asymmetric learning rates of murine exploratory behavior in sparse reward environments. *Neural Networks*, 143:218–229.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, 2 edition.

内藤, 碧. and 亀田, 達. (2022). 集合知を支える社会学習過程の合理的基礎. **認知科学**, 29(3):354–363.