Modeling human learning in a combinatorial bandit task

Guangyu Deng (gydeng@stu.pku.edu.cn)
Peking-Tsinghua Center for Life Sciences, Peking University
Academy for Advanced Interdisciplinary Studies, Peking University
Beijing, 100871, China

Haoyang Lu (luhy@pku.edu.cn)
School of Psychological and Cognitive Sciences, Peking University
Beijing 100871, China

Yi-Long Lu (luyilong@pku.edu.cn)
School of Psychological and Cognitive Sciences, Peking University
Beijing 100871, China

Hao Yan (hao_y@bjmu.edu.cn)
Peking University Sixth Hospital and Peking University Institute of Mental Health
Beijing 100191, China

Hang Zhang (hang.zhang@pku.edu.cn)
School of Psychological and Cognitive Sciences, Peking University
Beijing 100871, China

Abstract

We propose a combinatorial variant of the multi-armed bandit task that allows an agent to select combinations of multiple arms, instead of only one of the arms. The resulting action space is thus multi-dimensional and larger than usual, highlighting the exploration–exploitation dilemma and the credit assignment problem. To model human learning in this task, we develop a learning model based on the policy-gradient (PG) algorithm of reinforcement learning, an algorithm that often excels value learning algorithms in complex action spaces, and examines the mathematical properties of its updating rule. In an experiment using this new task (*N* = 42), we find nearly half of the participants are better fit by the PG model and exhibit behavioral patterns that value learning may have difficulty accounting for.

Keywords: reinforcement learning; computational modeling; multi-armed bandit; policy gradient method

Introduction

The multi-armed bandit (MAB) task is a widely-used testbed for both reinforcement learning (RL) algorithms and human cognition (Lattimore & Szepesvári, 2020; Schulz, Franklin, & Gershman, 2020). In a typical MAB task, multiple arms are available, each corresponding to one latent reward distribution. The agent must select one arm a time, and may learn to adjust their action policy based on the outcomes. Although extensive research using the MAB task has revealed important features of human learning behavior, its limited action space may fail to detect learning mechanisms (including abnormal ones) that operate in daily life's large action spaces (Wise, Emery, & Radulescu, 2024). One previous approach to create a more realistic action space is to use options whose rewards correlate with their visual or spatial features (Stojić, Schulz, Analytis, & Speekenbrink, 2020; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018). In this study, we proposed a combinatorial extension to the multi-armed bandit (MAB) task. Furthermore, we developed a policy learning model that could potentially learn more efficiently in the expanded action space of our proposed task. We also examined the mathematical properties of this model and tested it in a new experiment, comparing it with a classic value learning model in explaining human behavioral patterns.

The combinatorial bandit task

In the task, the agent can select multiple arms simultaneously in each trial (Figure 1). With 4 arms available and the option to select no arms, there are $2^4 = 16$ possible actions. The reward for a combination is determined by summing the rewards drawn from the selected arms, with each arm behaving as in a standard MAB task. While combining multiple arms to maximize rewards might resemble tasks in the representation learning literature that involve learning relevant features for multiple dimensions (e.g. Song, Niv, & Cai, 2020), our task is more versatile in its reward structure, which emphasizes risky decisions and is as extensible as the original MAB.

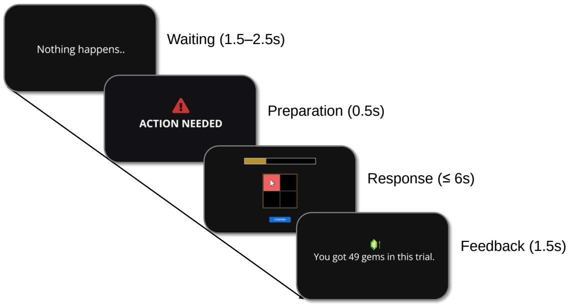

Figure 1: One trial of the combinatorial bandit task. During the response phase, participants clicked on the squares (arms) of the 2 by 2 panel to form a combinatorial action.

Experimental design Four arms were used for our experiment and simulation. Each arm's reward was drawn from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with $\mu \in \{35, 15, -5, -25\}$ and $\sigma = 15$, and the means were randomly allocated to the four arms before each session. During each trial of a session, the means of either $(35, -25)$ or $(15, -5)$ had a probability of $p = 0.05$ to be interchanged. Selecting no arms (all squares *OFF*) invoked the default choice, which followed a one-down-three-up reward schedule (starting at 50, max: 50, min: 10) that decreased by 10 when chosen and increased by 10 after three successive non-choices, resetting after each change. Across trials, the panel was randomly initialized to an *OFF* or *ON* state to avoid default choice abuse. Each participant completed four identical, independent 60-trial sessions. Each trial had a time limit of 6 s for participants to choose arms and confirm their choice, with a time-out punishment of $r = -50$. Participants were instructed to find the effective arms to maximize their token (gem) earnings.

Participants Fifty participants were recruited through the Prolific online platform. After completing the study, they were paid a base reward of £4.90 and a performance-based bonus ranging from £0.00 to £0.80. Eight participants were excluded due to low effort (average clicks per trial less than 0.25). Data from the remaining 42 participants entered further analysis.

Learning effects Within-session learning was evident in the first 10 trials of each session, with the probability of selecting the best arm increasing with trials (logistic regression, $\mathrm{OR} = 1.094$, $z = 5.179$, $p < 0.001$). A learning effect was also observed across sessions, with the mean expected reward (standard deviation in parentheses) increasing from 20.85 (8.47) in session 1 to 22.97 (9.32), 22.07 (9.46), and 23.04 (8.75) in sessions 2, 3, and 4, respectively.

Computational modeling

The combinatorial action space exaggerates two classic problems in RL: the exploration–exploitation dilemma and the credit assignment problem. In addition to the classic need to balance exploiting good actions and exploring potentially better ones (Mehlhorn et al., 2015), the agent must also tradeoff between obtaining more accurate information from fewer arms and less accurate information from more arms. Mean-

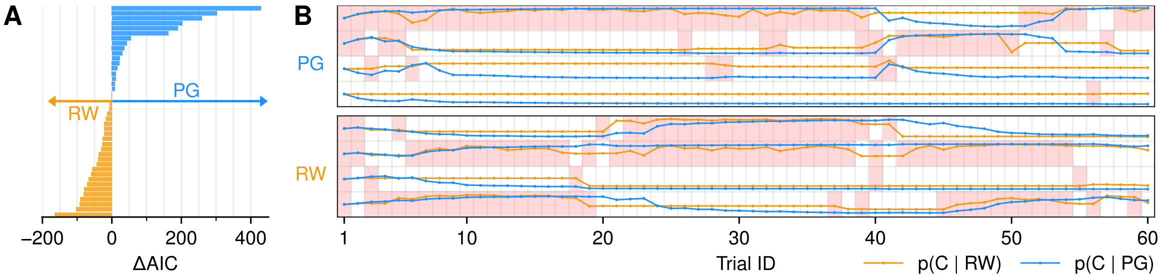

Figure 2: Modeling results comparing the policy-gradient (PG) and Rescorla–Wagner (RW) models. (A) Model comparison using the difference in $\mathrm{AIC}_{\mathrm{RW}} - \mathrm{AIC}_{\mathrm{PG}}$. Each bar denotes one participant. Participants better fit by PG and RW are respectively in blue and yellow. (B) Data vs. model prediction for one session each from two example participants whose choices were better fit by PG (upper) and RW (lower). Each row denotes an arm, with colored cells indicating selected arms. The superimposed lines show the model-predicted probabilities of selecting each arm, indicated by the relative height within each row.

while, with a common feedback for multiple arms, the credit assignment to each arm is essential for learning.

Value-based learning algorithms, which reduces to the Rescorla–Wagner (RW) learning rule (Rescorla & Wagner, 1972) in one-step decision tasks as ours, have been frequently used to model human learning in MAB. However, value learning algorithms such as RW typically do not provide efficient ways to explore the large number of actions resulting from combinatorial choices. Their even attribution of the prediction error to each arm can also be inefficient.

Policy learning algorithms, an alternative family of RL algorithms, offer relatively smooth learning and flexibility in tasks with higher-dimensional or even continuous action spaces (Sutton & Barto, 2018). We hypothesize that such models can capture some aspects of human learning in our task.

The policy gradient model The policy-gradient (PG) model developed for our task is based on the REINFORCE formulation of the policy gradient method (Williams, 1992):

$$H \leftarrow H + \eta\,(R_t - B_t)\,\nabla\pi_t(A_t|H)/\pi_t(A_t|H), \qquad (1)$$

where the vector $H \in \mathbb{R}^4$ denotes the agent's preference for the four arms, η is the learning rate, R_t is received reward in trial t, A_t is action in trial t, the ∇ term is gradient with respect to H, and $B_t = (1-\alpha)B_{t-1} + \alpha R_t$ with parameter α serves as a baseline to reduce variance. The policy is parameterized using a softmax function, which normalizes the action preferences $H(A)$ across all possible actions A. In simulations, the model with parameters $\eta = 0.053$ and $\alpha = 0.2$ achieves a mean reward level of 22.56 (SD = 11.83), comparable to the behavioral data.

Mathematical properties A key feature of the model is that rarer actions experience larger net preference changes. This follows the component-wise learning rule derived from Eq. 1 and additive action preferences, with the preference for component H^i updated by

$$H^i \leftarrow H^i + \eta\,(R_t - B_t)\sum_{A\in\mathcal{A}^i}(\delta(A_t, A) - \pi_t(A)), \qquad (2)$$

where $\mathcal{A}^i$ denotes the set of all actions that include the i-th arm, and $\delta(A_t, A) = 1$ if $A_t = A$ and otherwise 0. For a chosen arm, Eq. 2 shows that its net change is proportional to $1 - \sum_{A\in\mathcal{A}^i}\pi_t(A)$, where $\sum_{A\in\mathcal{A}^i}\pi_t(A)$ can be viewed as the marginal probability of selecting the i-th arm. This equation has several implications for the learning speed and credit assignment. First, when the outcome is desirable ($R_t \geq B_t$), the preference for a chosen arm increases quickly if it was previously low but slowly if it was already high. This effectively creates individual learning rates for different arms, in contrast to the common learning rate used in RW. Second, preference updating occurs similarly for each unchosen arm. When the absence of a preferred arm yields a desirable outcome, its preference would dramatically decrease. For large action spaces, such preference updating can be more efficient than RW learning. It may also parsimoniously explain the recent finding of error-driven value updating of unchosen actions (Ben-Artzi, Kessler, Nicenboim, & Shahar, 2023).

Third, when two arms are chosen with $H^i < H^j$, Eq. 2 implies that for an undesirable outcome, the less preferred i-th arm is blamed, while for a desirable outcome, the more preferred j-th arm's contribution is ignored. This property can be generalized to unchosen arms or more than two arms. A PG agent may be trapped if the best arm becomes the worst, but may escape the trap and detect changes more quickly through parameters that encourage exploration.

Model fitting and comparison To see whether value learning or policy learning better explains human behaviors in our experiment, we also implemented an RW model that evaluates the value of each arm. For each participant, we fit the RW and PG models to the choices in the last three sessions using maximum likelihood estimation and compared the two models' goodness-of-fit. Figure 2 highlights substantial individual differences in model fitness, with the PG model outperforming the RW model for 19 out of 42 participants, as indicated by large differences in $\Delta\mathrm{AIC}$. For the example participant better fit by the PG model (Figure 2B, upper), the probability of selecting a long-unchosen arm decreased with trials, consistent with PG but not RW predictions. Conversely, for the example participant better fit by the RW model (Figure 2B, lower), the RW model more accurately captured rapid valuation or devaluation. Further analysis of the PG model's distinctive behavioral patterns is needed to fully understand its unique contributions.

Acknowledgements

This study was partly supported by the Clinical Medicine Plus X – Young Scholars Project, Peking University, the Fundamental Research Funds for the Central Universities (PKU2023LCXQ023) for HY and HZ. HY was partly supported by the National Natural Science Foundation of China (82071505). HZ was partly supported by the National Natural Science Foundation of China (32171095), National Science and Technology Innovation 2030 Major Program (2022ZD0204803), and funding from Peking-Tsinghua Center for Life Sciences.

References

Ben-Artzi, I., Kessler, Y., Nicenboim, B., & Shahar, N. (2023). Computational mechanisms underlying latent value updating of unchosen actions. *Science Advances*, *9*(42), eadi2704. doi: https://doi.org/10.1126/sciadv.adi2704

Lattimore, T., & Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge: Cambridge University Press. doi: https://doi.org/10.1017/9781108571401

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., ... Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191–215. doi: https://doi.org/10.1037/dec0000033

Rescorla, RA., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II: Current Research and Theory* (Vol. 2).

Schulz, E., Franklin, N. T., & Gershman, S. J. (2020). Finding structure in multi-armed bandits. *Cognitive Psychology*, *119*, 101261. doi: https://doi.org/10.1016/j.cogpsych.2019.101261

Song, M., Niv, Y., & Cai, M. B. (2020). Learning what is relevant for rewards via value-based serial hypothesis testing. In *42nd Annual Meeting of the Cognitive Science Society: Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020.*

Stojić, H., Schulz, E., Analytis, P. P., & Speekenbrink, M. (2020). It's new, but is it good? How generalization and uncertainty guide the exploration of novel options. *Journal of Experimental Psychology: General*, *149*(10), 1878–1907. doi: https://doi.org/10.1037/xge0000749

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction, 2nd ed.* Cambridge, MA, US: The MIT Press.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*(3), 229–256. doi: https://doi.org/10.1007/BF00992696

Wise, T., Emery, K., & Radulescu, A. (2024). Naturalistic reinforcement learning. *Trends in Cognitive Sciences*, *28*(2), 144–158. doi: https://doi.org/10.1016/j.tics.2023.08.016

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915–924. doi: https://doi.org/10.1038/s41562-018-0467-4