Human behavior in contextual multi-armed bandit problems

Universitat Pompeu Fabra Barcelona

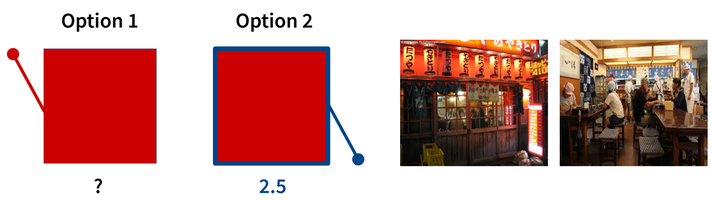
Hrvoje Stojic, Pantelis P. Analytis, Maarten Speekenbrink
hrvoje.stojic@upf.edu, analytis@mpib-berlin.mpg.de, m.speekenbrink@ucl.ac.uk

MAX-PLANCK-GESELLSCHAFT

UCL

Question

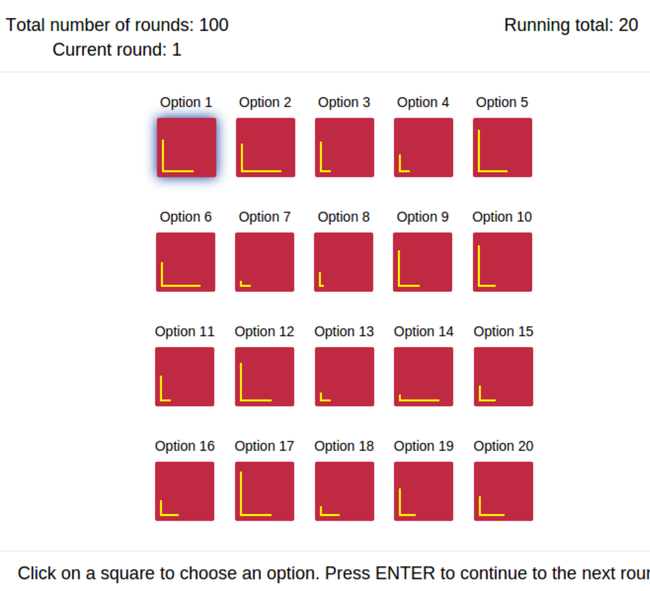
- Multi-armed bandit (MAB) tasks and reinforcement learning (RL) models play a crucial role in the study of decision making.
- However, realistic decision environments have multi-dimensional alternatives where features might be related to their values.



- The MAB paradigm poorly represents such situations – no room for learning the structure and generalization to new decision situations.
- We developed a new contextual multi-armed bandit (CMAB) task characterized by multi-dimensional stimuli.
- We examined human behavior in the task and tested RL models that rely on function learning to learn the structure of the situation.

The CMAB task

- Repeated choices between alternatives with two features – horizontal and vertical line.
- Feedback – value of the chosen alternative j in trial t, $R_j(t)$.
- Goal – maximize the cumulative reward.



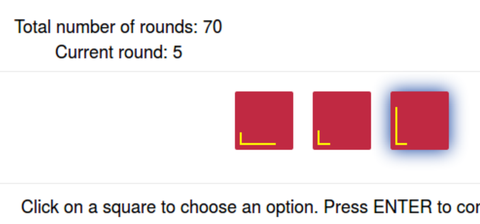
- Rewards $R_j(t)$ are a function of feature values:

$$R_j(t) = 2 \times x_{1,j} + 1 \times x_{2,j} + \epsilon_j(t).$$

- $\epsilon_j(t) \sim N$, independently for arm and trial.

Experimental design

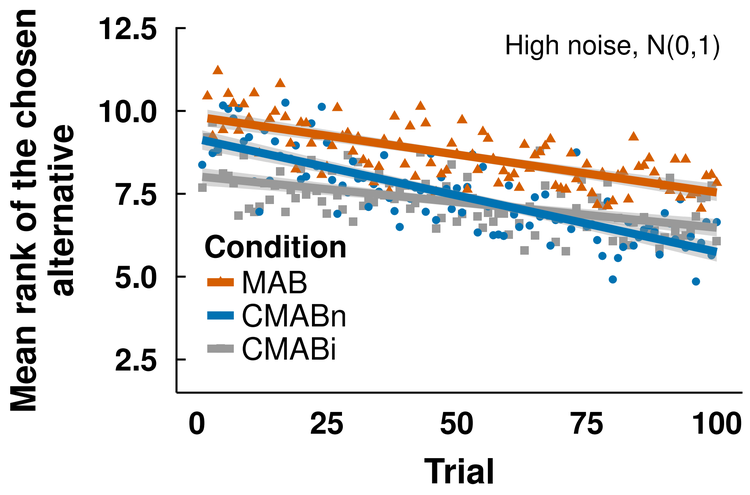
- **Training phase** – between subject design, CMAB task and MAB task, 20 alternatives and 100 trials.
- **Test phase** – one shot choices, 3 alternatives, 70 trials, no feedback – function learning test.

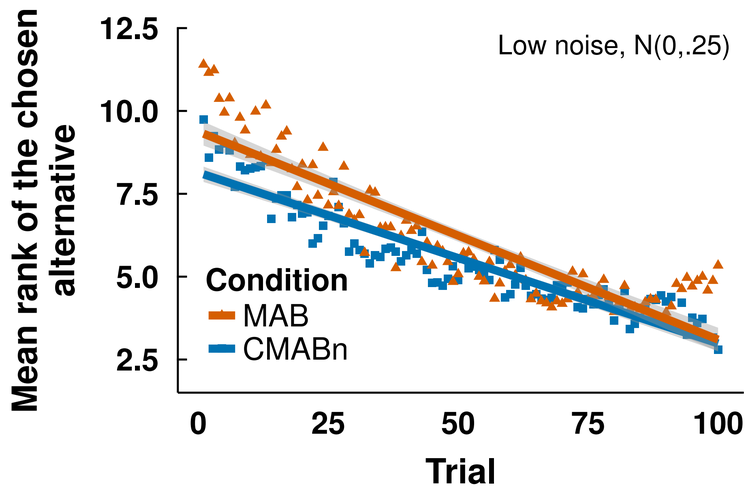


- Experiment 1 - High noise: $\epsilon_j(t)$ drawn from $N(0, 1)$, $N = 145$, Mturk, incentivised.
- Experiment 2 - Low noise: $\epsilon_j(t)$ drawn from $N(0, 0.25)$, $N = 143$, Mturk, incentivised.

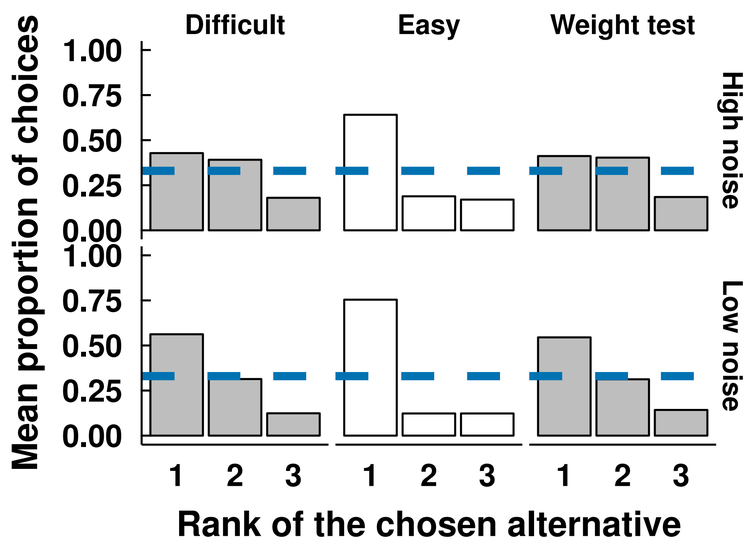
Data is available at: `http://dx.doi.org/10.6084/m9.figshare.1314099`, and the task can be tried at: `experimentnext.com/CMABvsMABexp1`

Behavioral results



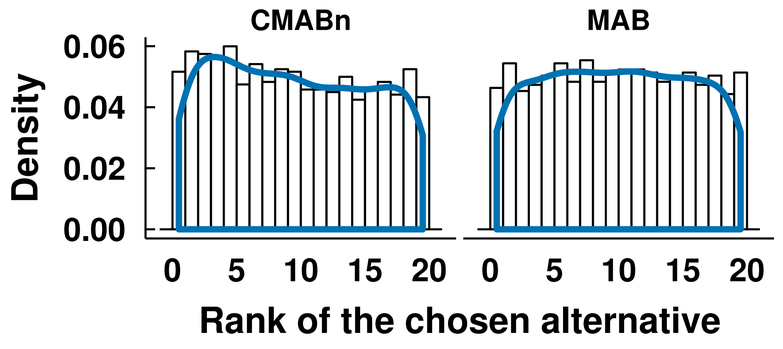


- **Top left and right:** People successfully learned to solve the task in all conditions. There were two CMAB conditions. In CMABi people were told features might be useful in deciding, while in CMABn they were not told anything about features. In the high noise experiment both main effects are significant, while in the low noise only the block effect.
- **Bottom right:** Test phase shows that people learned the function and used it to make choices – distributions were skewed toward alternatives with high function value, especially in the low noise case.



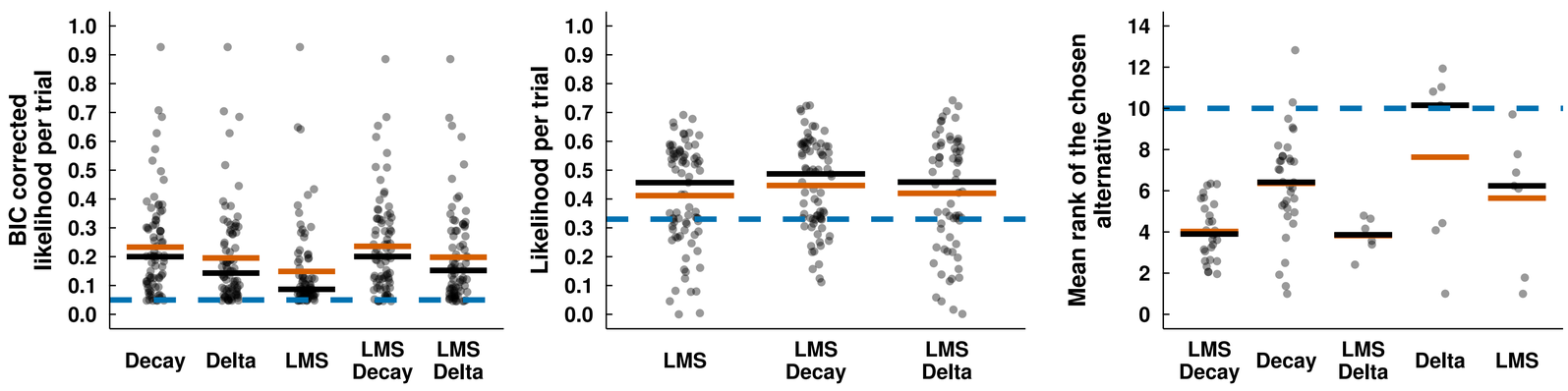
Exploration in the CMAB task - Low noise experiment

- Exploration-exploitation trade-off changes fundamentally in the CMAB task. Exploration should be guided by functional knowledge, skewed toward highly ranked alternatives, but also include the extremes.
- Figure shows an indication of this behavior. Ranks of chosen alternatives after switching to previously untried alternative have slightly bimodal distribution in the CMAB task.



Modeling results - Low noise experiment

- Three types of models were fitted: 1) Naïve RL (mean tracing) model using either decay or delta learning; 2) RL model with function learning based on least-mean-squares network (LMS) model; 3) Hybrid RL model combining both LMS and mean tracing through either decay or delta learning.
- **Left:** Hybrid LMS with decay and naïve decay RL model fit the behavior in the CMAB task the best. Red and black lines are means and medians, and blue dashed line is random choice.
- **Middle:** Only the LMS based models are able to predict choices in the test phase.
- **Right:** People best fitted with hybrid LMS models tend to perform better in the task.



Summary

- New CMAB task is closer to realistic situations rich in features, and we can study generalization, transfer of learning and novelty.
- We characterized human behavior in the CMAB task.
- We developed formal models and a hybrid LMS function learning and decay mean tracing model fit the choices well.

References

[1] Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In Proceedings of the 19th international conference on world wide web (661–670).

[2] Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. Journal of Experimental Psychology: General, 117(3), 227–247.

[3] Speekenbrink, M., & Shanks, D. R. (2010). Learning in a changing environment. Journal of Experimental Psychology: General, 139(2), 266–298.

[4] Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. Journal of Mathematical Psychology, 53(3), 168–179.

[5] Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction. Cambridge, MA, US: MIT Press.