Uncertainty and Exploration in a Restless Bandit Problem

Extended Abstract

Harsh Arora (20218266) Pratham Shukla (17816497)

In new and stochastic decision-making environments, balancing exploration-exploitation is a ubiquitous optimization problem (Wilson et al., 2020). A prominent experimental paradigm used to study the exploration-exploitation tradeoff is the multi-armed bandit task. In this setting, an agent is typically expected to learn the probabilistic reward structure underlying each bandit arm and maximize its cumulative reward within a limited time horizon. For bandits with fixed mean rewards, an optimal decision-strategy is obtainable (Berry & Fristedt, 1985). But real-life environments are often characterized by changing expected rewards, requiring continuous optimization of the tradeoff. For such 'restless' bandits, Daw et al. (2006) found that exploration is independent of the uncertainty, which rather, should be a driving factor (Cohen et al., 2007). Knox et al. (2012), in a restricted setting with two arms, found evidence for uncertainty-driven exploration. In the presented study, Speekenbrink and Konstantinidis (2015) used a four-armed restless bandit setting to probe the tradeoff.

The bandits were similar to those in Daw et al. (2006), with Gaussian random walks for means. The four blocks were a 2X2 setting of stable-variable and trend-no trend characteristics. The trend provided a constant drift to the means as against the completely unpredictable random walk. The variable blocks were similar to the stable blocks except that they had higher volatility for fixed intervals of trials in between. The authors expected the exploration to increase in high volatility blocks, decreasing relatively for subjects who are able to notice the trends. To model human performance in the task, a variety of learning and choice rule combinations were used. The learning rules included Bayesian updating using Kalman filters, and the model-free decay rule and delta rule. The choice rules included epsilon-greedy, softmax and its variants, and choosing the arm with the probability of maximum utility.

Modeling results indicate that the Bayesian update rule along with the choice of arm with probability of maximum utility performed the best. The subjects estimated higher uncertainty for blocks with higher volatility, resulting in increased exploration. The trends could explain the switching behavior, but not the performance in the blocks. The paper provides strong evidence that exploration is driven by uncertainty, impacting the predicted utility distributions that determine choice. It also calls for future research on nuanced analysis of how uncertainty affects other model parameters, for example, the learning rate.

1 References

1. Wilson RC, Bonawitz E, Costa VD, Ebitz RB. Balancing exploration and exploitation with information and randomization. Curr Opin Behav Sci. 2021 Apr;38:49-56. doi: 10.1016/j.cobeha.2020.10.001. Epub 2020 Nov 6. PMID: 33184605; PMCID: PMC7654823.

2. Berry, D. A., & Fristedt, B. (1985). Bandit problems Sequential allocation of experiments. London Chapman Hall.

3. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. Nature. 2006 Jun 15;441(7095):876-9. doi: 10.1038/nature04766. PMID: 16778890; PMCID: PMC2635947.

4. Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 362(1481), 933–942. https://doi.org/10.1098/rstb.2007.2098

5. Knox, W. B., Otto, A. R., Stone, P., & Love, B. C. (2012). The nature of belief-directed exploratory choice in human decision-making. Frontiers in psychology, 2, 398. https://doi.org/10.3389/fpsyg.2011.00398

6. Speekenbrink M, Konstantinidis E. Uncertainty and exploration in a restless bandit problem. Top Cogn Sci. 2015 Apr;7(2):351-67. doi: 10.1111/tops.12145. Epub 2015 Apr 20. PMID: 25899069.