Uncertainty and Exploration in a Restless Bandit Problem

Term Paper

Harsh Arora (20218266) Pratham Shukla (17816497)

Abstract

Decision-making problems are characterized by the exploration-exploitation tradeoff. An agent, at all times, needs to balance between exploring new options and exploiting known options, to reach the optimal solutions and maximize cumulative reward. Restless bandits are a version of the multi-armed bandits setting with variable expectancies of arm rewards. Daw et al. (2006) found that exploration in the restless bandits task was not driven by the uncertainty associated with the arms, which contradicts the logical explanation for exploration (Cohen et al., 2007). In this paper, Speekenbrink & Konstantinidis (2015) spot the possible issues with Daw's (2006) formulation and suggest an alternate way to incorporate uncertainty in the decision strategy. They use probability of maximum utility for a choice rule and find strong evidence in its favor.
In our replication of their paper, we consider three out of many models that the authors estimated, and corroborate with the findings. Though we failed to reproduce all the results exactly, we found evidence reinforcing probability of maximum utility as a possible choice strategy of the subjects.

1 Introduction

Decision-making in new and stochastic environments is always characterized by a dilemma between desire for rewards and desire for information. This ubiquitous phenomenon is termed as the exploration-exploitation tradeoff (Sutton & Barto, 1998, Wilson et al., 2020). It refers to the balance between exploiting options with known payoffs and exploring for new options with possibly higher payoffs. This is often modeled as the multi-armed bandits problem (Acuna & Schrater, 2008). Typically, an agent is expected to learn the probabilistic reward structure underlying each arm and maximize its cumulative reward over a fixed horizon in time.

For standard bandit problems, the mean reward of each arm is fixed. In such a case, an optimal decision strategy is obtainable by dynamic programming (Berry & Fristedt, 1985) or by calculating the Gittins index (Gittins, 1979). But real-life environments are often characterized by options with changing expected rewards. For example, when choosing between restaurants to dine out, one might want to revisit a restaurant that did not offer a great experience previously, allowing for the possibility that the quality may have improved over time. These problems, consisting of arms with changing means are called the 'restless' bandit problems. Optimal decision strategies for such cases are most times, difficult to find (Papadimitriou & Tsitsiklis, 1999).

Daw et al. (2006) investigated human decision-making in a restless bandit setting and found that exploration does not depend on the uncertainty associated with each arm. This is disappointing because, ideally, uncertainty should drive the exploration in decision-making problems (Cohen et al., 2007). However, Knox et al. (2012) could find an optimal decision strategy for a two-armed restless bandit problem in which the better arm kept flipping after every few trials. Their subjects acted similarly, and followed a policy that tracked the uncertainty in the choices. This raises doubts about

whether the role of uncertainty remained unobserved due to the way uncertainty was defined and modeled in Daw's (2006) task, or because of the difference in task designs. In this paper, the authors, Speekenbrink and Konstantinidis (2015) use a similar task and implement an alternate way to account for uncertainty in the choice process.

Let's assume a simple version of the restless bandits that we implement in this paper to illustrate the alternative decision-making model that accounts for uncertainty in the choice process. The reward structures are,

$$R_j(t) = \mu_j(t) + \epsilon_j(t) \qquad \epsilon_j(t) \sim N(0, \sigma_\epsilon)$$
$$\mu_j(t) = \mu_j(t-1) + \zeta_j(t) \qquad \zeta_j(t) \sim N(0, \sigma_\zeta)$$

An ideal Bayesian learner updates its belief about the reward structures based on the choices ($C_{1:t}$) made and the rewards ($R_{1:t}$) obtained in the past. In our case, we will use a Kalman Filter (Kalman, 1960; Kalman & Bucy, 1961) to provide these posterior distributions of rewards. These posteriors will then be used to derive a prior distribution for the expected rewards, which can further be used to estimate a prior predictive distribution for the rewards in the next trial. For each arm, this distribution is obtained as follows,

$$p(R_j(t)|R_{1:t}, C_{1:t}) = \int p(R_j(t+1)|\mu_j(t+1))p(\mu_j(t+1)|R_{1:t}, C_{1:t})d\mu_j(t+1)$$

When choosing an arm, an agent can rely either on a greedy strategy, or another that induces exploration. Even though greedy strategies offer the best reward based on the running estimates of rewards, it is often wise to rely on explorative strategies and check whether some other arm has surpassed the current maximum. If an arm is left unchosen for some time, the associated uncertainty with the arm increases, and so does the probability that it yields the maximum reward. Not just uncertainty, the probability that an arm provides the maximum reward is a function of the expected reward as well. An arm that was previously closer to the best arm, would have a higher probability of being the best arm when left unchosen over time. The use of probability of maximum reward as a choice rule was introduced by Thompson (1933) and is commonly known as Thompson sampling. Daw et al.'s (2006) paper could not find evidence for uncertainty-driven exploration using an exploration bonus that increased linearly with the variance of the prior. The probability of maximum utility combines expected reward and uncertainty in a much sophisticated way, and has outperformed heuristic strategies (Viappiani, 2013). In this paper, the authors use a similar strategy and evaluate its fit against others.

2 Method

A four-armed restless bandits task, similar to Daw et at. (2006) was used. There were four versions of the task, dictated by a 2X2 setting of stable/variable X trend/no trend combinations. They were ST (stable-trend), VT (volatile-trend), SN (stable-no trend) and VN (volatile-no trend). The participants were expected to make more exploratory decisions in variable blocks than stable blocks. The exploration was expected to be lesser for participants who are able to tap the trend in the random walk.

2.1 Participants

Thirty participants (age = 21.1 (2.56), sex: 21 male, 9 female, handedness: 4 left, 26 right) volunteered for the experiment. Each participant was randomly assigned to one of the four conditions. The data for one right-handed male was removed after it appeared as an outlier in the regret calculations.

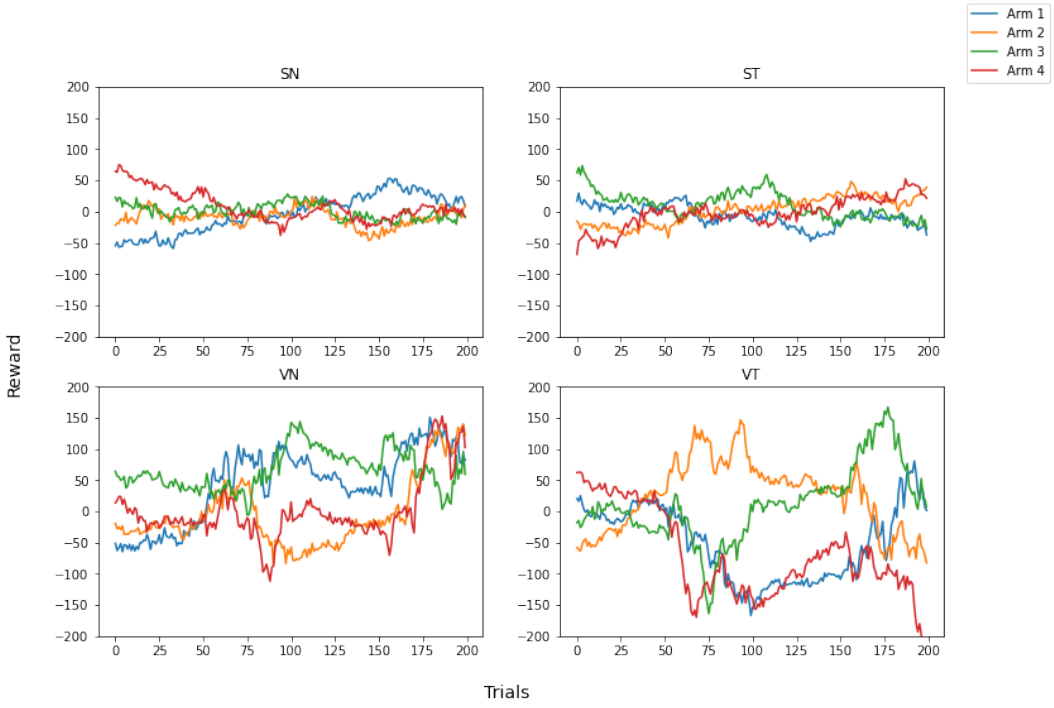
2.2 Task

The four-armed bandit task consisted of 200 trials. Each slot machine yielded some reward on pulling, and the aim of the participant was to maximize their cumulative reward by the end of the 200 trials. The reward structure underlying the arms is described by the equations,

$$R_j(t) = \mu_j(t) + \epsilon_j(t) \qquad \epsilon_j(t) \sim N(0, \sigma_\epsilon)$$
$$\mu_j(t) = \lambda\mu_j(t-1) + \kappa_j + \zeta_j(t) \qquad \zeta_j(t) \sim N(0, \sigma_\zeta(t))$$

The initial averages of the arms were $\mu_j(1) = [-60, -20, 20, 60]$. A decay parameter, $\lambda = 0.9836$ was used to avoid runaways by pulling the means closer to zero. In the ST and VT conditions, the trend parameter, κ_j, was added. The values corresponding to the initial averages were $\kappa_j = [0.5, 0.5, -0.5, -0.5]$. For SN and VN conditions, $\kappa_j = 0$. There were two variances in the reward equation, the error variance, σ^2_ϵ, and the innovation variance, $\sigma^2_\zeta(t)$. The error variance was 16 for all trials of all blocks. The innovation variance was 16 everywhere except trial numbers 51 - 100 and 151 - 200 in the VT and VN conditions, where it was 256. An example of the rewards in each of these conditions is shown in Figure 1.

Figure 1: Example Rewards in the Four-Armed Restless Bandits Task



2.3 Procedure

The experiment was designed and hosted for online participation on Cognition (2020). Participants were asked to imagine themselves in a casino in front of a wall with four slot machines. They were told that each machine yields some reward on pulling. At any point, some machines could be better than others, but the yield of each of the machines could change. Before each trial there was fixation with duration randomly chosen from $[750, 1000, 1250, 1500, 1750]$ms There were images of 4 slot machines on the screen arranged as if on the corners of a rectangle. They were mapped to the keys 'Q', 'P', 'M' and 'Z' in the same order in which the keys appear on the keyboard. After a button was clicked, the reward appeared on the slot machine corresponding to the key for 1.5sec. Throughout the task, a counter indicated the total points collected thus far.

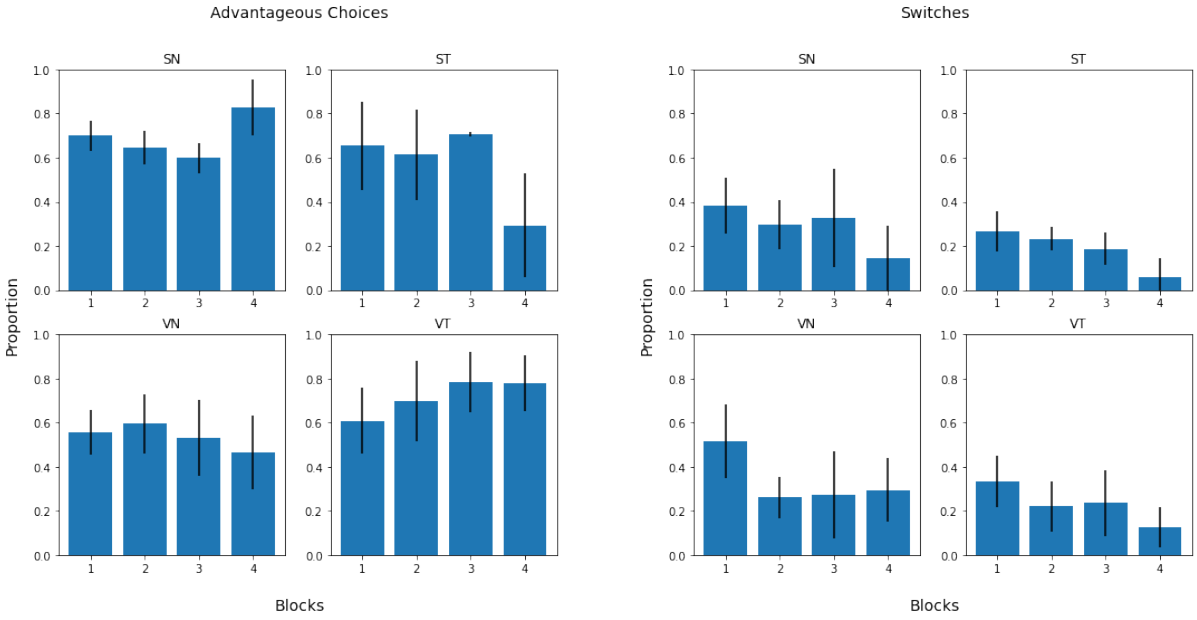
3 Behavioral Results

We focused on two metrics to analyze the behavior of the participants, they are, the proportion of advantageous choices and the proportion of switches. Proportion of advantageous choices is the ratio of trials in each block when the best arm was selected to the total trials in the block. And, the proportion of switching is the total number of times a participant chooses a different option than they did in the previous trial, over the number of trials in that block.

A serious analysis could have been done by estimating a generalized linear mixed effects model with Block, Trend and Volatility as fixed effect regressors. However, while estimating using *statsmodels* (2010), the program crashed because of exceeded RAM on Google Colab (2019). Such analysis would have also taken care of the non-Gaussian nature of the variables under interest. Hence, we proceed with only a short qualitative assessment of the behavior. Volatility and Trend, both affect the discriminability of the the arms. The visible differences moving from no trend to trend conditions should be increased advantageous choices and decreased switching, given that the trend increases discriminability. This same pattern is obtained, as shown in Figure 2. The only exception to the trend arises between SN and ST condition where the proportion drops substantially in Block 4. Here, the effect should have actually been the highest because of maximum displacement of the random walk due to the trend. A possible reason could be the lack of data. Participants' random assignment to conditions was not balanced and ST had only 3 of 29 data points. That also explains the high variance. Hence, conclusions made from ST can not be taken as strong indicators.

Since increased volatility would inflate uncertainty estimates, it shall result in increased exploration, which directly implies increased switching. The expected patterns are visible in Figure 2, but with changes less significant in Block 3. This could possibly be because, after a block of high volatility, the arms may actually have been left farther apart and hence more discriminable. The proportion of advantageous choices, as expected, decreases going from low to high volatile blocks, but visibly insignificantly for Block 3. This can be argued in the same way as before, that high volatility block leads to enhanced discriminability.

Though not calculated here, statistical measures to these differences could have been calculated by hypothesis testing. One may have assumed binomial distribution for the proportion of advantageous choices. If the switches were taken as count rather than proportions, then assuming Poisson distribution could have been assumed for them.



(a) Advantageous Choices

(b) Switches

Figure 2: Proportion of advantageous choices and switches by blocks for each condition

4 Analysis

4.1 Models

In decision-making tasks, an agent typically assigns a subjective value to its options and updates them at every timestep. But the perception of reward and punishment is not symmetric, according to Prospect theory (Tversky & Kahneman, 1992), people are loss-averse, i.e., for the same magnitude of reward and punishment, the loss is perceived asymmetrically greater than the gain. Building on these lines, the authors define utility as,

$$u(t) = \begin{cases} R(t)^{\alpha} & R(t) \geq 0 \\ -\lambda |R(t)|^{\alpha} & R(t) < 0 \end{cases}$$

where $\alpha > 0$. When $\alpha < 1$, the utility curve is concave for gains and convex for losses. This implies risk-aversion and risk-seeking respectively. The parameter $\lambda \geq 0$ accounts for loss-aversion. In the original paper (Speekenbrink and Konstantinidis, 2015), the authors used a variety of learning and choice rules and fit data to them. In this replication, we chose three models, each of which performs the best for their learning rule. These models are,

Kalman Filter - Probability of Maximum Utility (Bayes - PMU). This model employed Bayesian updating using Kalman Filter (Kalman, 1960) for learning, and probability of maximum utility for choice.

A Kalman Filter assumes an underlying Gaussian process, and estimates the mean and variance of the distribution according to the following equations,

$$E_j(t) = E_j(t-1) + \delta_j(t) K_j(t) [u(t) - E_j(t-1)]$$

where $\delta_j(t) = 1$ if arm j was chosen on trial t, and zero otherwise. The Kalman gain, K, is calculated as,

$$K_j(t) = \frac{S_j(t-1) + \sigma_\zeta^2}{S_j(t-1) + \sigma_\zeta^2 + \sigma_\epsilon^2}$$

The posterior variance, $S_j(t)$, is computed as,

$$S_j(t) = [1 - \delta_j(t)K_j(t)][S_j(t-1) + \sigma_\zeta^2]$$

The prior means and variances were initialized to $E_j(0) = 0$ and $S_j(0) = 1000$.

The probability of maximum utility choice rule has been discussed in the introduction. For an arm to be chosen, the prior predictive distribution of the probability of pairwise difference of rewards between this arm and others should be positive.

$$P(C(t) = j) = P(\forall k : u_j(t) \geq u_k(t)) = \int_0^\infty \Phi(M_j(t), H_j(t))$$

where Φ is the multivariate normal density function, and

$$M_j(t) = A_j E(t)$$

$$H_j(t) = A_j diag(S(t) + \hat{\sigma}_\epsilon^2) A_j^T$$

Here, $E(t)$ is the vector of prior expected utility on trial t, and $diag(S(t) + \hat{\sigma}_\epsilon^2)$ is a matrix with variance of the prior predictive distribution on the principal diagonal. $\hat{\sigma}_\epsilon^2$ is the observed variance assumed by the agent. The matrix A, is the pairwise difference, for example, for the first arm,

$$A_1 = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix}$$

Decay - Fixed Softmax (Decay - SM_f. This model employs the model-free decay rule and the fixed softmax rule for choice.
Decay rule (Ahn et al., 2008) assumes that the value estimate of each bandit decays towards zero,

$$E_j(t) = E_j(t-1) + \delta_j(t)\eta[u(t) - E_j(t-1)$$

where the decay parameter, $0 \leq \eta \leq 1$

The softmax choice (Sutton & Barto, 1998) rule maps the reward estimates to probabilities, as follows,

$$P(C(t) = j) = \frac{exp(\theta_0 E_j(t))}{\Sigma_{k=1}^4 exp(\theta_0 E_j(t))}$$

where θ_0 is the temperature parameter and can take any positive value.

Delta - Dynamic Softmax (Delta - SM_d). This model uses the model-free delta learning rule (Yechiam & Busemeyer, 2005), along with softmax for choice.
The delta update happens as follows,

$$E_j(t) = E_j(t-1) + \delta_j(t)\eta[u(t) - E_j(t-1)]$$

Model	△AIC	w(AIC)	n(AIC)	△BIC	w(BIC)	n(BIC)
Bayes - PMU	209.93 (87.34)	0.45 (0.46)	8	200.17 (87.23)	0.51 (0.47)	7
Decay - SM_f	223.34 (81.61)	0.29 (0.43)	12	216.81 (81.54)	0.27 (0.42)	12
Delta - SM_d	216.09 (94.87)	0.25 (0.40)	10	209.57 (94.80)	0.22 (0.39)	11

Table 1: AIC and BIC scores and weights for the models. Figures mentioned as mean (std. dev.)

Model	μ_0	σ_0	σ_ζ	σ_ϵ	η	θ_0	λ	α
Bayes - PMU	2.12	129.75	0.75	0.75			0.07	0.47
Decay - SM_f	0.08				0.54	0.62	1.41	0.29
Delta - SM_d	6.03				0.80	0.02	0.002	0.31

Table 2: Median values of the parameters obtained after fitting

where the parameter, $0 \leq \eta \leq 1$, is the learning rate.

In a dynamic softmax (Busemeyer & Stout, 2002), the temperature parameter changes over time according to $\theta(t) = [t/10]^{\theta_0}$, hence, the final relation being,

$$P(C(t) = j) = \frac{exp(\theta(t)E_j(t))}{\Sigma_{k=1}^{4} exp(\theta(t)E_k(t))}$$

4.2 Model Estimation and Inference

For each individual, the parameters were estimated by maximizing likelihood using Nelder-Mead Simplex algorithm. The optimization was performed by the *constrNMPy* (2018) library in Python. To evaluate the fit, difference between the Akaike (AIC) and Schwarz (BIC) scores of the model of interest and a null, non learning model, were calculated (cf. Yechiam & Busemeyer, 2005). For the null model, the arms were assumed to be drawn from a multinomial distribution, parameters for which were also estimated for each participant. A higher value of these scores, △AIC and △BIC implies a better fit. From these scores, the Akaike and Scwarz weights, w(AIC) and w(BIC), were calculated (cf. Wagenmakers & Farrell, 2004). They compare the models in a relative probabilistic sense.

5 Results

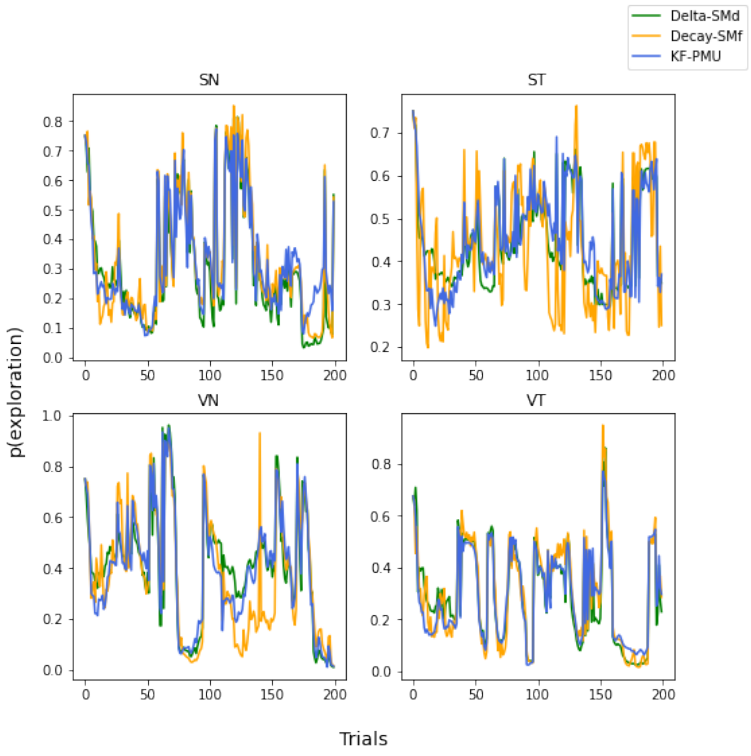
Table 1 consists of all fit measures. From the difference scores, the Decay - SM_f model gives the best fit, with the highest value for both, △AIC and △BIC. But, the weights convey a different picture. The average weights, w(AIC) and w(BIC) are maximum for the Bayes - PMU model. This discrepancy arises due to the fact that whenever Bayes - PMU fits the data the best, it does sweepingly better than the rest, resulting in very high weights. The other models fit best only with small margins. The number of participants that are best fit by each model, n(AIC) and n(BIC), are higher for the Decay - SM_f model. Of the above findings, all except the number of subjects that were fit best, agree with the original paper. For the authors, Bayes - PMU fitted more number of subjects.

The median parameter estimates are listed in Table 2. In our estimation, the parameters μ_0 and σ_0 have been added to account for the prior mean and variance. These parameters were not estimated in the original paper. The values of most parameters fall close to the ones reported in the paper with

considerable deviations for λ and α. All the three models depict risk aversions, given the values of α. The Bayes - PMU and Delta - SM_d show some evidence for risk aversion since $\lambda < 1$. The difference is possibly because for decay rule, the value estimates decay towards zero, making them more predictable and favourable. The qualitative results agree with the paper.

The amount of exploration in each of the models cannot be figured out only by looking at the parameters. Exploration at any trial is a function of the expected rewards, and hence the utility function and all the parameters it depends upon. To look at the exploratory behavior of each of the models, one can look at the probability of choosing a non-maximizing arm. Hence, the probability of choosing an arm that does not give the maximum reward was calculated for the fitted data for every participant and then averaged. From the figure, Bayes - PMU and Decay - SM$_f$ predict the highest exploration. It is observable that probability of exploration seems noisier for stable conditions but goes through a lot more cycles in the volatile condition. It is also visible that there is high chances for exploration in all conditions during the start of the task, which then decreases rapidly. The probability of exploration is also less in trend conditions than in the ones without. This validates the hypothesis that trends make it easier for subjects to tap the best arm and hence explore less. The above results agree with the paper, except that the Bayes - PMU model very clearly predicted more exploration for the authors in the stable condition.

Figure 3: Probability of Exploration for Each Condition



6 Conclusion

We found that our participants modulate exploitation based on conditions in the environment they are set. They switched more in volatile conditions and in no-trend completely random settings. Their behavior in the task was well emulated by two of our models, one with a decay learning rule that fits majority of the data, and the other with a Bayesian update that fits slightly lesser data, but very decidedly. This points to at least some evidence that exploration is driven by uncertainty. This result agrees with Knox et al. (2012), who used a much restricted task. And as predicted, disagrees with Daw et al. (2006), who could not find evidence for uncertainty-driven exploration. Given the simplicity of our Bayesian learner, it is prudent to assume that the actual learning could be a much nuanced version. This calls for future research into finer analyses of how uncertainty affects learning, decision-making and choice.

Link to the GitHub repository of this project - https://github.com/Pratham-04/CS786A-Spring-2020-21/tree/main/Project

7 References

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

2. 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.

3. Acuna, D., & Schrater, P. (2008). Bayesian modeling of human sequential decision-making on the multiarmed bandit problem. In B. C. Love, K. McRae, V. M. Sloutsky (Eds.), Proceedings of the 30th Annual Conference of the Cognitive Science Society (pp. 2065–2070). Austin, TX: Cognitive Science Society.

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

5. Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society. Series B, 41, 148–177.

6. Papadimitriou, C. H., & Tsitsiklis, J. N. (1999). The complexity of optimal queuing network control. Mathematics of Operations Research, 24, 293–305.

7. 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.

8. 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

9. 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

10. 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.

11. Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. Transactions of the American Society of Mechanical Engineers, Series D, Journal of Basic Engineering, 82, 35–45.

12. Kalman, R. E., Bucy, R. S. (1961). New results in linear filtering and prediction theory. Transactions of the American Society of Mechanical Engineers, Series D, Journal of Basic Engineering, 83, 95–108.

13. Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25, 285–294.

14. Viappiani, P. (2013). Thompson sampling for bayesian bandits with resets. In P. Perny, M. Pirlot, A. Tsouki Aas (Eds.), Algorithmic decision theory (Vol. 8176, pp. 399–410). Berlin: Springer.

15. Cognition.run (2020). https://www.cognition.run/'

16. Bisong E. (2019) Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-4470-8_7

17. Seabold, Skipper, and Josef Perktold. "statsmodels: Econometric and statistical modeling with python." Proceedings of the 9th Python in Science Conference. 2010.

18. Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5, 297–323.

19. Ahn WY, Busemeyer JR, Wagenmakers EJ, Stout JC. Comparison of decision learning models using the generalization criterion method. Cogn Sci. 2008 Dec;32(8):1376-402. doi: 10.1080/03640210802352992. PMID: 21585458.

20. Alexander Blaessle. "constrNMPy: A Python package for constrained Nelder-Mead optimization." 2018.

21. Yechiam, E., Busemeyer, J.R. Comparison of basic assumptions embedded in learning models for experience-based decision making. Psychonomic Bulletin Review 12, 387–402 (2005). https://doi.org/10.3758/BF03193783

22. Busemeyer JR, Stout JC. A contribution of cognitive decision models to clinical assessment: decomposing performance on the Bechara gambling task. Psychol Assess. 2002 Sep;14(3):253-62. doi: 10.1037//1040-3590.14.3.253. PMID: 12214432.

23. Wagenmakers, EJ., Farrell, S. AIC model selection using Akaike weights. Psychonomic Bulletin Review 11, 192–196 (2004). https://doi.org/10.3758/BF03206482