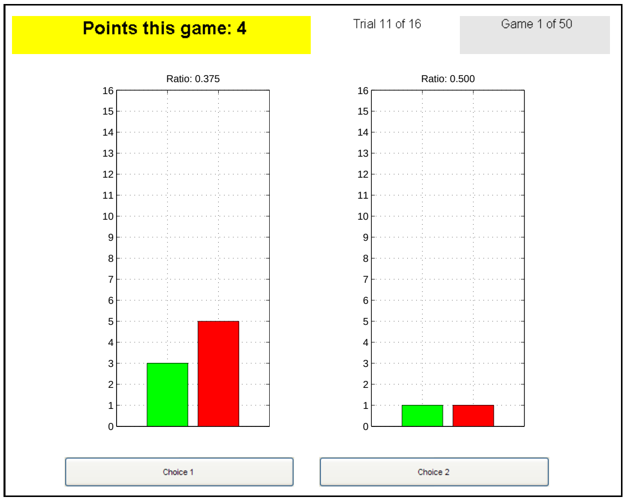
Human and Optimal Exploration and Exploitation in Bandit Problems

Shunan Zhang (szhang@uci.edu)
Michael D. Lee (mdlee@uci.edu)
Miles Munro (mmunro@uci.edu)
Department of Cognitive Sciences, 3151 Social Sciences Plaza A
University of California, Irvine, CA 92697-5100 USA

Abstract

We consider a class of bandit problems in which a decision-maker must choose between a set of alternatives—each of which has a fixed but unknown rate of reward—to maximize their total number of rewards over a short sequence of trials. Solving these problems requires balancing the need to search for highly-rewarding alternatives with the need to capitalize on those alternatives already known to be reasonably good. Consistent with this motivation, we develop a new model that relies on switching between latent *searching* and *standing* states. We test the model over a range of two-alternative bandit problems, varying the number of trials, and the distribution of reward rates. By making inferences about the latent states from optimal decision-making behavior, we characterize how people should switch between searching and standing. By making inferences from human data, we attempt to characterize how people actually do switch. We discuss the implications of our findings for understanding and measuring the competing demands of exploration and exploitation in decision-making.

Keywords: Bandit problems, exploration versus exploitation, reinforcement learning, Bayesian graphical models, human decision-making, optimal decision-making



Points this game: 4

Trial 11 of 16

Game 1 of 50

Figure 1: An example bandit problem, with two alternatives and 16 total trials. After 10 trials, the first alternative on the left has 2 successes (lighter, green bar) and 5 failures (darker, red bar), while the alternative on the right has 1 success and 1 failure.

Bandit Problems

Bandit problems, originally described by Robbins (1952), present a simple challenge to a decision-maker. They must choose between a known set of alternatives on each of a series of trials. They are told each of the alternatives has a fixed reward rate, but are not told what the rates are. Their goal is just to maximize the total reward they receive over the series of trials. In this paper, we focus on short finite-horizon versions of the bandit problem, involving just a small number of trials.

As an example of the challenge posed by these sorts of bandit problems, consider the situation shown in Figure 1. Here there are two alternatives, and 16 total trials available to attain rewards. After 10 trials, one alternative has been chosen 8 times, and returned 3 successes and 5 failures, while the other alternative has been tried just 2 times, for 1 success and 1 failure. Which alternative should be chosen on the 11th trial? Choosing the first alternative exploits the knowledge that it quite likely returns rewards at a moderate rate. Choosing the second alternative explores the possibility that this alternative may be the more rewarding one, even though much less is known about it.

As this example makes clear, finite-horizon bandit problems are psychologically interesting because they capture the tension between exploration and exploitation evident in many real-world decision-making situations. Decision-makers must try to learn about the alternatives, which requires exploration, while simultaneously satisfying their goal of attaining rewards, which requires exploitation. In this way, studying human performance on bandit problems addresses basic questions, including how people search for information, how they adapt to the information they find, and how they optimize their behavior to achieve their goals.

Human performance on bandit problems has been studied from a variety of psychological perspectives. Early studies used models and experimental manipulations motivated by theories of operant conditioning (e.g., Brand, Wood, & Sakoda, 1956); later studies were informed by economic theories with a focus on deviations from rationality in human decision-making (e.g., Banks, Olson, & Porter, 1997; Meyer & Shi, 1995); most recently human performance on the bandit problem has been a topic of interest in cognitive neuroscience (e.g., Cohen, McClure, & Yu, 2007; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006) and probabilistic models of human cognition (e.g., Steyvers, Lee, & Wagenmakers, in press).

One common finding is that people often switch flexibly between exploration and exploitation, often choosing alternatives in proportion to their reward rate, unless they are given strong incentives to maximize their reward by repeatedly choosing the most-rewarding alternative (e.g., Shanks, Tunney, & McCarthy, 2002). Typically, these experiments involve a large number of trials, and so one plausible explanation for sub-optimal probability matching is that people are allowing for the possibility that rewards rates might change over time. This seems less likely to be a confounding consideration in short-horizon bandit problems, and so we are especially interested to know if people switch between exploration and exploitation for these problems.

Accordingly, in this paper we develop and evaluate a probabilistic model that assumes different latent states guide decision-making for short-horizon bandit problems. These latent states give emphasis either to searching the environment, or to choosing the same alternative repeatedly, and so dictate how a decision-maker solves the dilemma in our introductory example, where a well-understood but only moderately-rewarding alternative must be compared to a less well-understood but possibly better-rewarding alternative. Using the optimal decision process, and human data, for a range of bandit problems we apply our model to understand the best way to switch between searching and standing, and how people actually do switch, for short horizon two-alternative bandit problems.

The outline of the paper is as follows. In the next section, we present our model, including its implementation as a probabilistic graphical model. We then report an experiment collecting human and optimal decisions for a range of bandit problems. Next, we use the behavioral data and our model to make inferences about the optimal way to switch between searching and standing, and how people actually do switch. Finally, we draw some conclusions relating to simpler latent state models suggested by our analysis.

A Latent State Model

Bandit problems have been widely studied in the fields of game theory and reinforcement learning (e.g., Berry, 1972; Berry & Fristedt, 1985; Gittins, 1979; Kaebling, Littman, & Moore, 1996; Macready & Wolpert, 1998; Sutton & Barto, 1988). One interesting idea coming from established reinforcement learning models is that of a latent state to control exploration versus exploitation behavior.

In particular, the 'ε-first' heuristic (Sutton & Barto, 1988) assumes two distinct stages in bandit problem decision-making. In trials in the first 'exploration' stage, alternatives are chosen at random. In the second 'exploitation' stage, the alternative with the best observed ratio of successes to failures from the first stage is chosen. The demarcation between these stages is determined by a free parameter, which corresponds to the trial at which exploration stops and exploitation starts.

Our Model

Our model preserves the basic idea of a latent exploration or exploitation state guiding decision-making, but makes two substantial changes. First, we allow each trial to have a latent state, introducing the possibility of switching flexibly between exploration and exploitation to solve bandit problems. In our model, for example, it is possible to begin by exploring, then exploit, and then return for an additional period of exploration before finishing by exploiting. Indeed, any pattern of exploration and exploitation, changing trial-by-trial if appropriate, is possible.

Second, we implement exploration and exploitation behavior using a more subtle mechanism than just random search followed by deterministic responding. In particular, for the two-alternative bandit problems we consider, our model distinguishes between three different situations,

- The *Same* situation, where both alternatives have the same number of observed successes and failures.

- The *Better-Worse* situation, where one alternative has more successes and fewer failures than the other alternative (or more successes and equal failures, or equal successes and fewer failures). In this situation, one alternative is clearly better than the other.

- The *Search-Stand* situation, where one alternative has been chosen much more often, and has more successes but also more failures than the other alternative. In this situation, neither alternative is clearly better, and the decision-maker faces a dilemma. Choosing the better-understood alternative corresponds to standing; choosing the less well-understood alternative corresponds to searching.[1]

Within our model, which alternative is chosen depends on the situation, as well as the latent search or stand state. For the *same* situation, both alternatives have an equal probability of being chosen. For the *better-worse* situation, the better alternative has a high probability, given by a parameter γ, of being chosen. The probability the worse alternative is chosen is $1-\gamma$.

[1] Intuitively, our notion of searching is a form of exploration, and our notion of standing is a form of exploitation. We use the new terms, however, to emphasize that our search and stand decisions have formal characterizations that are different definitions of exploration and exploitation in reinforcement learning algorithms. For example, ε-first uses simple random choices as a model of exploration, whereas our approach is based on choosing specifically the alternative that is less well known in a search-stand situation.

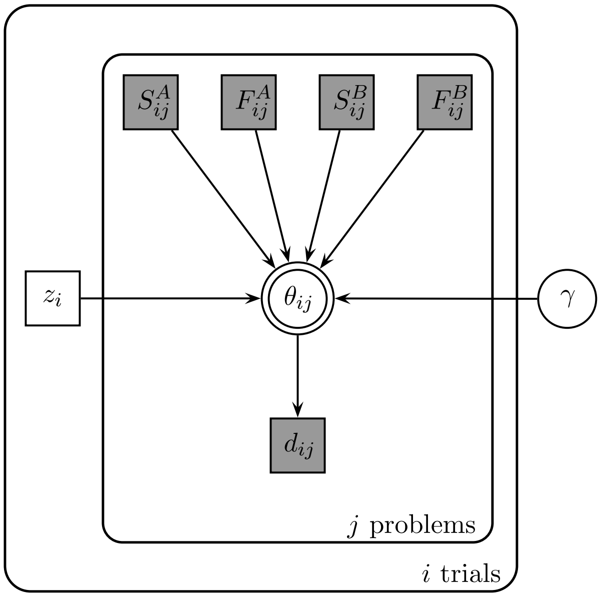


Figure 2: Graphical representation of the latent state model.

For the *search-stand* situation, the exploration alternative will be chosen with the high probability γ if the decision-maker is in a latent search state, but the exploitation alternative will be chosen with probability γ if the decision-maker is in the latent stand state. In this way, the latent state for a trial controls how decisions are made each time the decision-maker encounters a search-stand situation.

Graphical Model Implementation

We implemented our model as a probabilistic graphical model in WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), which makes it easy to do fully Bayesian inference using computational methods based on posterior sampling. The graphical model is shown in Figure 2, using the same notation as Lee (2008).

The encompassing plates show the repetitions for the trials within each problem, and the multiple problems completed by a decision-maker. The square shaded nodes S^A_{ij}, S^B_{ij} , F^A_{ij} and F^B_{ij} are the observed counts of successes and failures for alternatives A and B on the ith trial of the jth problem. The unshaded node γ is the 'accuracy of execution' parameter, controlling the (high) probability that the deterministic heuristic described by our model is followed. The unshaded z_i nodes are the discrete latent indicator variables, with $z_i = 0$ meaning the ith trial is in the explore state, and $z_i = 1$ meaning it is in the exploit state. We assumed uninformative priors $\gamma \sim \text{Uniform}(0,1)$ and $z_i \sim \text{Bernoulli}(1/2)$.

The double-bordered θ_{ij} node is a deterministic function of the S^A_{ij}, S^B_{ij} , F^A_{ij}, F^B_{ij}, γ and z_i variables. It gives the probability that alternative A will be chosen on the ith trial of the jth problem. According to our model, this is

$$\theta_{ij} = \begin{cases} 1/2 & \text{if A is same} \\ \gamma & \text{if A is better} \\ 1-\gamma & \text{if A is worse} \\ \gamma & \text{if A is search and } z_i = 0 \\ 1-\gamma & \text{if A is search and } z_i = 1 \\ \gamma & \text{if A is stand and } z_i = 1 \\ 1-\gamma & \text{if A is stand and } z_i = 0. \end{cases}$$

The shaded d_{ij} node is the observed decision made, $d_{ij} = 1$ if alternative A is chosen and $d_{ij} = 0$ if alternative B is chosen, so that $d_{ij} \sim \text{Bernoulli}(\theta_{ij})$.

In this way, the graphical model in Figure 2 provides a probabilistic generative account of observed decision behavior. It is, therefore, easy to use the model to make inferences about latent search and stand states from decision data. In particular, the posterior distribution of the z_i variable represents the probability that a decision-maker has a latent search versus stand state on the ith trial. In the next section, we describe an experiment that provides both human and optimal data suitable for this type of analysis.

Experiment

Participants

We collected data from 10 naive participants (6 males, 4 females).

Stimuli

We considered six different types of bandit problems, all involving just two alternatives. The six bandit problem types varied in terms of two trial sizes (8 trials and 16 trials) and three different environmental distributions ('plentiful', 'neutral' and 'scarce') from which reward rates for the two alternatives were drawn.

Following Steyvers et al. (in press), we defined these environments in terms of Beta (α,β) distributions, where α corresponds to a count of 'prior successes' and β to a count of 'prior failures'. The three environmental distributions are shown in Figure 3, and use values $\alpha = 4$, $\beta = 2$, $\alpha = \beta = 1$, and $\alpha = 2$, $\beta = 4$, respectively.

Procedure

We collected within-participant data on 50 problems for all six bandit problem conditions, using a slight variant of the experimental interface shown in Figure 1. The order of the conditions, and of the problems within the conditions, was randomized for each participant. All $6 \times 50 = 300$ problems (plus 5 practice problems per condition) were completed in a single experimental session, with breaks taken between conditions.

Optimal Performance

Given the α and β parameters of the environmental distribution, and the trial size, it is possible to find the optimal decision-making process for a bandit problem. This is achieved via dynamic programming, using a recursive approach well understood in the reinforcement learning

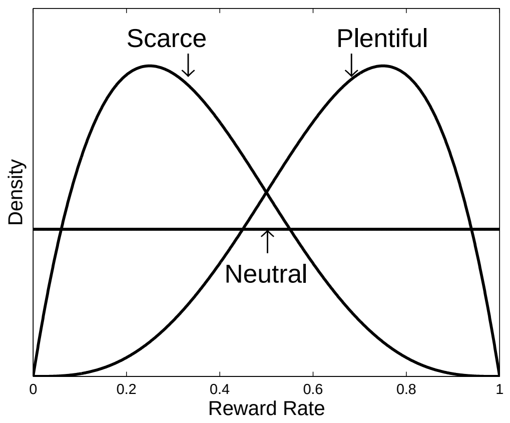


Figure 3: The plentiful, neutral and scarce environmental distributions of reward rates.

Table 1: Posterior predictive agreement between the latent state model, and the optimal and human decision-makers (DMs), for the three environments and two problem sizes.

	Plentiful		Neutral		Scarce	
DM	8	16	8	16	8	16
Optimal	.95	.93	.95	.94	.92	.90
PH	.96	.94	.92	.92	.84	.90
ST	.99	.87	.94	.84	.93	.80
AH	.89	.89	.76	.75	.71	.73
MM	.92	.88	.92	.93	.90	.94
SZ	.92	.94	.95	.92	.88	.91
MY	.94	.95	.92	.93	.89	.88
EG	.94	.91	.90	.90	.85	.89
MZ	.97	.91	.92	.88	.93	.86
RW	.89	.90	.86	.80	.84	.80
BM	.93	.88	.92	.87	.89	.90

literature (e.g., Kaebling et al., 1996). Using this approach, we calculated optimal decision-making behavior for all of the problems completed by our participants.

Modeling Analysis

We applied the graphical model in Figure 2 to the optimal and human decision data, for all six bandit problem conditions. For each data set, we recorded 1,000 posterior samples from the joint distribution of the unobserved variables. We used a burn-in also of 1,000 samples, and multiple independent chains, to assess convergence.

Basic Results

Descriptive Adequacy A basic requirement of any cognitive model is that it can fit the observed data reasonably well. To test the descriptive adequacy of the latent state model, we used a standard Bayesian approach and evaluated its posterior predictive fit to the to all of the human and optimal decision-making data (i.e., the agreement between the model and data averaged over the posterior distribution of the parameters). The levels of agreement are shown in Table 1. It is clear that the latent state model is generally able to fit both human and optimal behavior very well. There are some small suggestive differences—scarce environments seem, for example, to be a little less well described, as does one participant (AH)—that are worthy of future investigation, but do not affect our broad analyses in this paper.

Latent States Having checked the descriptive adequacy of the latent state model, our main interest is in the change between latent search and stand states, as shown by the inferred model parameters.[2] The basic results needed to address this question are summarized by the posterior mean of the z_i indicator variables, which approximates the posterior probability that the ith trial uses the stand state.

Figure 4 shows the posterior means of the z_i variables for the optimal decision process, and all 10 participants, in all six experimental conditions. The experimental conditions are organized into the panels, with rows corresponding the plentiful, neutral and scarce environments, and the columns corresponding to the 8- and 16-trial problems. Each bar graph shows the probability of an stand state for each trial, beginning at the third trial (since it is not possible to encounter the search-stand situation until at least two choices have been made). The larger bar graph, with black bars, in each panel is for the optimal decision-making data. The 10 smaller bar graphs, with gray bars, corresponds to the 10 participants within that condition.

Analysis

The most striking feature of the pattern of results in Figure 4 is that, to a good approximation, once the optimal or human decision-maker first switches from searching to standing, they do not switch back. This is remarkable, given the completely unconstrained nature of the model in terms of search and stand states. All possible sequences of these states over trials are given equal prior probability, and all could be inferred if the decision data warranted.

The fact that both optimal and human data lead to a highly constrained pattern of searching and standing states across trials reveals an important regularity in bandit problem decision-making. We consider this finding first in terms of optimal decision-making, and then in terms of human decision-making.

Optimal Decision-Making The optimal decision process results in Figure 4 show that it is optimal to be-

[2] We observed that the inferred γ parameter values were all close to 1, as expected, and do not report them in detail.

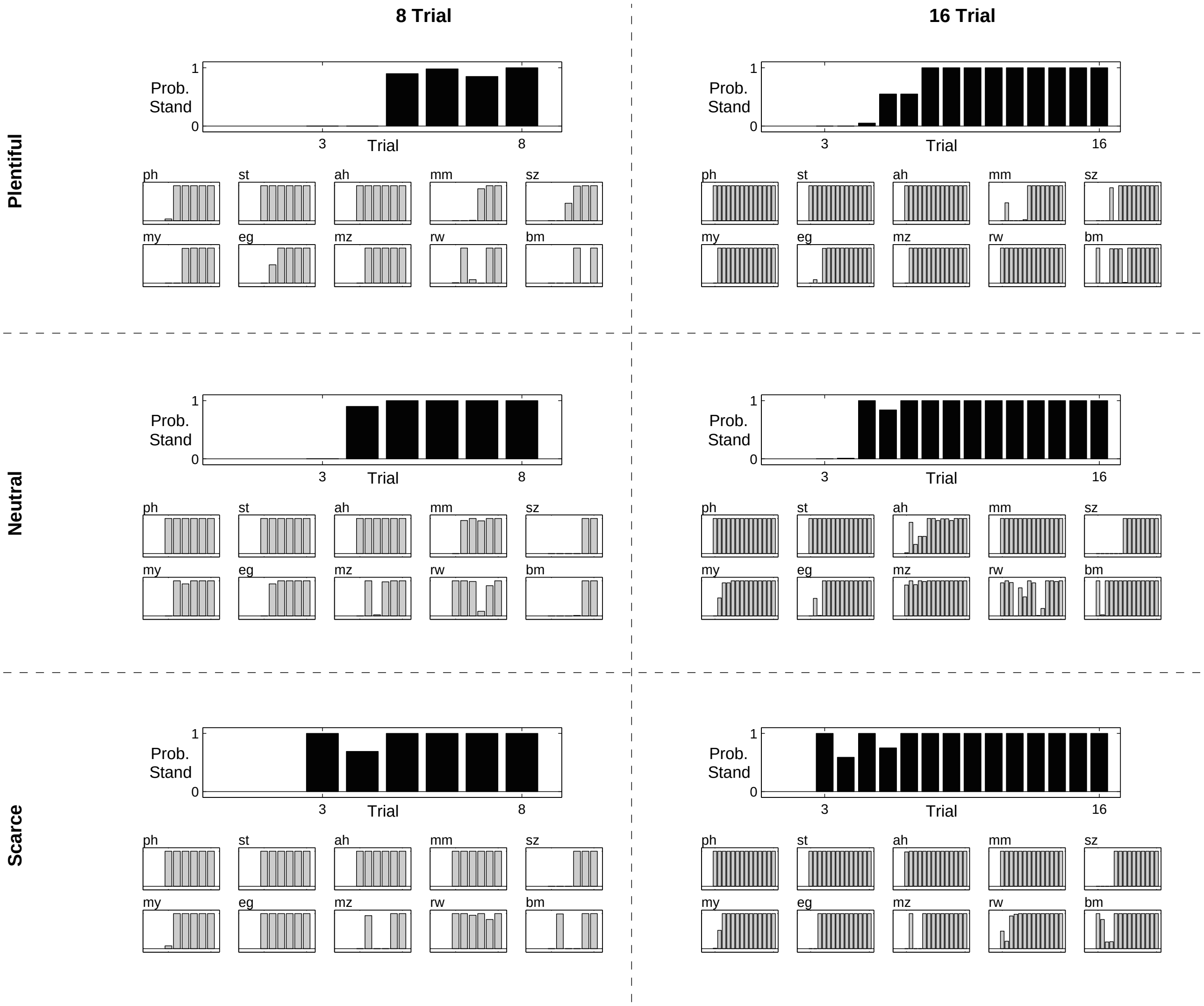


Figure 4: Each bar graph shows the inferred probabilities of the stand state over the trials in a bandit problem. Each of the six panels corresponds to an experimental condition, varying in terms of the plentiful, neutral or scarce environment, or the use of 8 or 16 trials. Within each panel, the large black bar graph shows the stand probability for the optimal decision-process, while the 10 smaller gray bar graphs correspond to the 10 participants.

gin with searching, then transition (generally) abruptly to standing at some trial that depends on the nature of the environment, and remain in the stand state for all of the remaining trials. The plentiful and scarce environments for 16-trial problems show a few trials where there is uncertainty as to whether searching or standing is optimal but, otherwise, it seems clear that optimal decision-making can be characterized by a single transition from searching to standing.

It is also clear from Figure 4 that the optimal decision-making must be sensitive to the environment in switching from searching to standing. In particular, as environments have lower expected reward rates, the switch away from searching begins earlier in the trial sequence. For example, the optimal decision process for 8-trial problems essentially switches from searching to standing at the 5th trial in the plentiful environment, but at the 4th trial in the neutral environment, and the 3rd trial in the scarce environment.

Human Decision-Making While the regularity in switching might not be surprising for optimal decision-making, it is more remarkable that human participants show the same pattern. There are some exceptions—both participants RW and BM, for example, sometimes switch from standing back to searching briefly, before returning to standing—but, overall, there is remarkable consistency. Most participants, in most conditions, begin by searching, and transition at a single trial to standing, which they maintain for all of the subsequent trials.

However, while there is consistency over the participants in switching just once from searching to standing, there are clear differences between individuals in when that switch happens. For example, the participant SZ, in all of the conditions, switches at a much later trial than most of the other participants.

There also seem to be individual differences in terms of sensitivity to the environment. Some participants switch at different trials for different environments, while others—such as participant ST—switch at essentially the same trial in all experimental conditions.

Discussion

Our basic findings involve both a regularity and a flexibility in the way people (and optimal) decision-makers switch between exploration and exploitation in bandit problems. The regularity is that a beginning period of searching gives way to a sustained period of standing. The flexibility is that when this switch occurs depends on the individual decision-maker, the statistical properties of the reward environment, and perhaps the interaction between these two factors.

The obvious cognitive model suggested by our findings combines the regularity with the flexibility. We propose that decision-making on finite-horizon bandit problem can be modeled in terms of a single parameter, controlling when searching switches to standing. That is, rather than needing a latent state parameter for each trial, only a single switch-point parameter is needed, with all earlier trials following the searching state, and all later trials following the standing state. Such a model would be similar in spirit—but formally different in an important way—to the standard ε-first heuristic from reinforcement learning. It would combine the single switch-point with an analysis of bandit game situations (‘same’, ‘better-worse’, ‘search-stand’) that produces more focused and principled operational definitions of what it means for decision-maker to explore and exploit.

A priority for future research is to apply this new single-switch model to human and optimal behavior on bandit problems. Being able to make inferences about when people and optimal decision-makers switch from exploration to exploitation promises a direct way to assess individual differences in how people search their environment for information, and react to different distributions of reward in those environments.

Acknowledgments

This work is was supported by the Air Force Office of Scientific Research (FA9550-07-1-0082) to Michael Lee and Mark Steyvers. We thank Jun Zhang for suggesting we evaluate the type of model considered in this paper.

References

Banks, J., Olson, M., & Porter, D. (1997). An experimental analysis of the bandit problem. *Economic Theory*, *10*, 55–77.

Berry, D. A. (1972). A Bernoulli two-armed bandit. *The Annals of Mathematical Statistics*, *43*(3), 871–897.

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

Brand, H., Wood, P. J., & Sakoda, J. M. (1956). Anticipation of reward as a function of partial reinforcement. *Journal of Experimental Psychology*, *52*(1), 18–22.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? exploration versus exploitation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*, 933–942.

Daw, N. D., O’Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876–879.

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

Kaebling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1–15.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS: A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Macready, W. G., & Wolpert, D. H. (1998). Bandit problems and the exploration/exploitation tradeoff. *IEEE Transactions on evolutionary computation*, *2*(1), 2-22.

Meyer, R. J., & Shi, Y. (1995). Sequential choice under ambuigity: Intuitive solutions to the armed-bandit problem. *Management Science*, *41*(5), 817–834.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *55*, 527–535.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*, 233–250.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (in press). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*.

Sutton, R. S., & Barto, A. G. (1988). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.