Bayesian and Approximate Bayesian Modeling of Human Sequential Decision-Making on the Multi-Armed Bandit Problem

Abstract

In this paper we investigate human exploration/exploitation behavior in sequential-decision making tasks. Previous studies have suggested that people are suboptimal at scheduling exploration, and heuristic decision strategies are better predictors of human choices than the optimal model. By incorporating more realistic assumptions about subject’s knowledge and limitations into models of belief updating, we show that optimal Bayesian and approximate Bayesian models of human behavior for the Multi-Armed Bandit Problem (MAB) outperform the best heuristic methods on experimental data for 2-arm, 3-arm, and 4-arm bandit problems. Moreover, we show that Bayesian modeling is more consistent to the exploratory and exploitative human behavior by disaggregating the fitting performance of decision sequences into several phases.

1 Introduction

Sequential decision-making in uncertain environments form an important class of problems in which the agent must simultaneously learn about the environment while choosing among uncertain alternatives to gather reward. Balancing these demands is called the exploration/exploitation trade-off, because the choices of when to gather more information (i.e., exploration) conflict with choices of when to use the information already gathered (i.e., exploitation). Bayesian optimal solutions to this problem are notoriously intractable, even in trivial cases [1, 2]. Nevertheless, humans engage in sophisticated sequential decision-making behavior in everyday life from a first year college student taking relevant courses to his major to a driver deciding the best path to home in a new town. However, it remains relatively unknown how people explore and exploit and whether human behavior is near-optimal, in part due to the lack of optimal solutions for comparison.

The simpler Multi-Armed Bandit Problem (MAB) offers a unique opportunity to test human sequential decision-making given its simplicity and widely-known optimal solution through the Gittins Index Theorem. The agent faces a set of $M$ independent state spaces, called arms. Each arm has only one action available, called pull. Only one arm can be pulled at each decision time, and therefore the agent has to plan the order in which the arms will be pulled. In MAB problems, at each decision time the decision-maker must select one of $M$ of potentially reward-generating processes (called arms) defined over independent state spaces. Selecting an option changes the state and stochastically generates a reward with unknown probability. The key requirements for a MAB problem are that the options are independent from each other, and that the states of unplayed arms are frozen until played again.

Only a small number of previous studies have been performed on human decision making with MAB problems with mixed results. Earlier studies suggested human choices reflect inaccurate Bayesian updating with suboptimals in exploration—different studies found under-exploration [3, 4], over-exploration [5] and both [6]. Recently, Gans et. al[7] looked at the predictive performance of a large class of optimal and heuristic choice mechanisms and found human choice behavior best predicted by a simple choice heuristic. The composite picture suggested by these studies is that
human exploration/exploitation behavior is non-optimal and perhaps based non-Bayesian decision-making. One of the critical difficulties with this conclusion is that the Bayesian updating used by the optimal comparison model contain unrealistic assumptions if used as a model for human belief-updating. In particular, people do not have infinite memory, infinite look ahead, nor precise encoding of reward and outcome events. In addition, it is unclear whether humans adopt the assumption that unplayed arms have frozen states in these experiments.

The purpose of this paper is to compare human performance on MAB problems with models of belief-updating that better reflect human abilities. In particular, we compare belief-update models based on Bernoulli reward processes, Normal reward processes with known variance, and Normal reward processes with unknown variance to human choice behavior. We compute Gittins Indices with limited memory and look ahead, and test their ability to predict human choices over a variety of 2-arm, 3-arm, and 4-arm bandit problems. We compare the performance of our models against the best predictive models found in Gans et. al[7]. We find that a belief-update model that estimates the mean and variance of the reward process with limited memory model provides the best predictive performance.

2 Modeling beliefs in MAB problems

In MAB problems, each arm $i$ is characterized by a state $x^i_k$ at time $k$. The goal is choose an action policy that maximizes total discounted future expected reward. For example, if the policy would select arm $i$ at a time $t$ steps in the future, an expected reward $\gamma^t r^i(x^i_{k+t})$ will be anticipated, where $\gamma$ is the discounting factor, with $0 \leq \gamma < 1$. In addition, the belief about the state of the $i^{th}$ arm would evolve according to a probabilistic transition model $P(x^i_{k+1} | x^i_k)$. Thus, MAB problems are special cases of Markov Decision Processes, and hence have an associated Bellman equation [8] that can be solved in principle via dynamic programming. However, Gittins [9] proved that the solution to MAB problems takes the form of an index for each arm, called the Gittins Index, and that the optimal action at each decision time is to pull the arm with highest index.

2.1 The Gittins Index

The virtue Gittins’s solution is that only information from a particular arm’s dynamics is required to compute that arm’s index. Moreover, the solution has a number of interpretations that help clarify how an optimal decision-maker schedules exploratory and exploitative moves. In particular, the Gittins Index for an arm $i$ can be viewed as an optimal value function for playing only that arm, that encodes the ratio between the expected reward if the arm is pulled until a best time $\tau$ (a stopping time), divided by the total discounted time up to $\tau$. More precisely,

$$\nu_i = \sup_{\tau > 0} \frac{E_x \left[ \sum_{t=0}^{\tau-1} \gamma^t r^i(x^i_{t+1}) | x^i_0 \right]}{E_x \left[ \sum_{t=0}^{\tau-1} \gamma^t | x^i_0 \right]}$$  \hspace{1cm} (1)$$

Thus at each choice point the optimal decision maker assesses the maximal reward rate (per unit discounted future time) expected for each arm. The solution has an interpretation in terms of exploration bonuses. Rename the numerator and denominator in equation 1 by $R_i(\tau)$ and $W_i(\tau)$ respectively. Let $\tau^*$ denote an optimal stopping time. The bonus value $B$ of arm $i$ exceeds its expected payoff by

$$B = \frac{R_i(\tau^*)}{W_i(\tau^*)} - R_i(\tau^*) = \frac{1}{W_i(\tau^*)} - R_i(\tau^*)$$

Using the fact that discounting is equivalent to a probability that a reward process will terminate, Sonin[10] showed that $W_i(\tau)$ could be interpreted as the probability of reward termination within time $\tau$. Thus the Gittin’s index gives a bonus to arms it believes will survive the time $\tau$. We believe this provides an important consideration for computing near-optimal solutions. A near-optimal agent should compute both the expected reward for an arm, and maintain an estimate of the reliability of the arm’s payoff. Our modifications of the optimal solution are motivated by the idea that humans may be estimating the the reliability of each arm’s payoff using strategies that are suboptimal for the experimentally imposed task.
2.1.1 Modeling human belief updates

Because previous investigations of human behavior in MAB problems has almost exclusively focused on Bernoulli reward processes, we had our observers choose between Bernoulli arms that generated sequences \( \{x_1, x_2, \ldots, x_n\} \) of independent and identically-distributed random variables taking values 1 or 0 with probability \( \theta \), and \( 1 - \theta \), respectively. To simulate human choice data in our experiments we compute value functions for a set of models that are Bayesian but differ from the generating model. For a given arm, the model assumes the \( x_i \)'s are drawn from a parametric distribution indexed by \( \theta \) with a density function \( f(\cdot|\theta) \). The prior density for \( \theta \) is denoted \( \pi \).

To compute Gittens indices for these models, we use the calibration method. This method calibrates an arm by comparing it with a standard bandit process, which has one state and a constant reward \( \lambda \) to compute an equivalent constant reward for each arm. The method works by finding the supremum amount of reward \( \lambda \) such that we would still prefer not to play the standard process. This amount \( \lambda \) is exactly the Gittins Index. The base for the calibration process is the Bellman equation rewritten as:

\[
U(\lambda, \pi) = \max \left[ \frac{\lambda}{1-\gamma}, r(\pi) + \gamma \int U(\lambda, \pi_x) f(x|\pi) \, dx \right], \tag{2}
\]

where \( \pi_x \) denotes the posterior \( \pi(\theta|x) \). We solve equation (2) for different values of \( \lambda \) as \( \pi \) varies throughout the family of posterior distributions. Note that when both arguments inside the brackets are equal for \( U(\lambda, \pi) \), it is indifferent to pull the standard process or the actual arm.

We introduced three belief update models, the optimal belief update based on the generating process (Bernoulli reward process). Two other models were introduced to simulate the effects tracking the reliability of each arm. In addition, assuming that the observer is completely represented by the mean \( \mu \) of the generating process. For a given arm, the model assumes the \( x_i \) are drawn from a parametric distribution indexed by \( \theta \) with a density function \( f(\cdot|\theta) \). The prior density for \( \theta \) is denoted \( \pi \).

\[
U(\lambda, \pi) = \max \left[ \frac{\lambda}{1-\gamma}, r(\pi) + \gamma \int U(\lambda, \pi_x) f(x|\pi) \, dx \right], \tag{2}
\]

Therefore, the calibration equation (2) becomes

\[
U(\lambda, \alpha, \beta) = \max \left[ \frac{\lambda}{1-\gamma}, \frac{\alpha}{\alpha + \beta} (1 + \gamma U(\lambda, \alpha + 1, \beta)) + \frac{\beta}{\alpha + \beta} \gamma U(\lambda, \alpha, \beta + 1) \right].
\]

Normal reward process with unknown mean and known variance: Here, the reward function follows a normal distribution with unknown mean, but known variance \( \sigma^2 \). The state may be completely represented by the mean \( \bar{x} \) and size \( n \) of the sample.

Fortunately, the normal reward process has a location and scale parameter behavior, represented by the mean and standard deviation, respectively. This means that the calibration equation and therefore the Gittins Index can be computed for any mean and standard deviation by translating and scaling the results from a previous computation. Because of this, the convenient following properties hold

\[
U(b\lambda + c, b\bar{x} + c, b\sigma) = c(1-\gamma)^{-1} + bU(\lambda, \bar{x}, n, \sigma) \quad (b > 0, c \in \mathbb{R}) \tag{3}
\]

and

\[
\nu(\bar{x}, n, \sigma) = \bar{x} + \sigma \nu(0, n, 1). \tag{4}
\]

Without lost of generality, we assume a known variance \( \sigma^2 = 1 \), and therefore leave the true mean \( \mu \) as the only unknown parameter. We take \( \pi_0 \) as the improper uniform density over all the real line.
and $f(x|\mu)$ as the normal distribution $N(\mu, 1)$. The density for a new value $x$ is

$$f(x|\bar{x}, n) = \int_{-\infty}^{\infty} f(x|\mu)\pi_n(\mu|\bar{x})\,d\mu$$

$$= \left(\frac{n}{2\pi(n+1)}\right)^{1/2} \exp\left\{-\frac{n}{2(n+1)}(x-\bar{x})^2\right\},$$

and

$$r(\bar{x}, n) = \int_{-\infty}^{\infty} xf(x|\bar{x}, n)\,dx = \bar{x}.$$ 

Thanks to (3) and (4), we only need to solve the following calibration equation for $\bar{x} = 0$ to find any other index

$$U(\lambda, 0, n) = \max\left[\frac{\lambda}{1-\gamma}, \gamma \int_{-\infty}^{\infty} U(\lambda - \frac{x}{n+1}, 0, n+1) f(x|0, n)\,dx\right].$$

**Normal reward process with unknown mean and variance:** In this case, the unknown parameters are the mean $\mu$ and the standard deviation $\sigma$. We choose a convenient improper prior $\pi_0 \propto 1/\sigma (\sigma > 0)$. The density function $f(x|\mu, \sigma)$ will be the usual normal distribution $N(\mu, \sigma)$.

The state of the process can be fully characterized by the sample mean $\bar{x}$, standard deviation $s$, and size $n$. The results are similar to the known variance case,

$$f(x|\bar{x}, s, n) = \int_{0}^{\infty} \int_{-\infty}^{\infty} f(x|\mu, \sigma)\pi_n(\mu, \sigma|\bar{x}, s, n)\,d\mu\,d\sigma \propto \left(1 + \frac{n}{n+1} \frac{(x-\bar{x})^2}{(n-1)s^2}\right)^{-n/2},$$

where $s^2$ is the unbiased sample variance, and the expected reward $r(\bar{x}, s, n) = \bar{x}$.

The property $\nu(\bar{x}, s, n) = \bar{x} + \nu(0, 1, n)$ holds (from (3), and (4)) and therefore the calibration method is simplified to

$$U(\lambda, 0, 1, n) = \max\left[\frac{\lambda}{1-\gamma}, \gamma \int_{-\infty}^{\infty} s_x U\left(\lambda - \frac{x(n+1)^{-1}}{s_x}, 0, n+1\right) f(x|0, 1)\,dx\right],$$

where $s_x = (n^{-1}(n-1)s^2 + (n+1)^{-1}(x-\bar{x})^2)^{1/2}$.

### 3 Restricting the order and stopping time of a reward process

The problem with previous accounts of Gittins Index as a model for human behavior is that they assume infinite memory and look ahead. Here, we show how more appropriate models with limited memory and future inference can be easily derived from the original Gittins Index.

First, we can limit the amount of memory by imposing a $K$-order decision process. After $x_1, x_2, \ldots, x_n$ observations, the state of the process would have a $K$-order assumption

$$\pi_n(\theta|x_1, x_2, \ldots, x_n) \equiv \pi_n(\theta|x_{n-K}, x_{n-K+1}, \ldots, x_n)$$

$$\propto \pi_{n-K}(\theta) \prod_{i=n-K}^{n} f(x_i|\theta), \quad (K > 0)$$

where $\pi_{n-K}(\theta) = \pi_0(\theta)$. For the calibration method, we can incorporate limited look ahead by modifying the equation $U(\lambda, \pi)$ to

$$U(\lambda, \pi, N) = \begin{cases} \max_{S(\pi)} \left[\frac{\lambda}{1-\gamma}, r(\pi) + \gamma \int U(\lambda, \pi, N-1) f(x|\pi)\,dx\right] & \text{if } N \geq 1 \\ S(\pi) & \text{otherwise} \end{cases},$$

where $S(\pi)$ is myopic future discounted total reward based on the state $\pi$ when the calibration process stops at $N = 0$.  

4
Bernoulli Reward Process: After observing the sequence \( x_1, x_2, \ldots, x_n \), we can limit the memory by plugging the density function \( f(x_i|\theta) = \theta^{x_i}(1 - \theta)^{1-x_i} \) into (6) and considering only the subsequence \( x_{n-K}, x_{n-K+1}, \ldots, x_n \). If there are \( n_0 \) occurrences of 0 and \( n_1 \) occurrences of 1 in the subsequence, we change the state of the process from \((\alpha, \beta)\) to \((n_1 + 1, n_0 + 1)\).

To restrict the look ahead, we use the modified calibration (7) and set a myopic influence of future discounted total reward in the horizon as \( S(\alpha, \beta) = \sum_{t=0}^{\infty} \gamma^t r(\alpha, \beta) = \alpha(\alpha + \beta)^{-1}(1 - \gamma)^{-1}. \)

We will denote this restricted Gittins Index for the Bernoulli Reward Process \( B(K, T) \).

Normal reward process with unknown mean and known variance: We restrict the memory by considering the last \( K \) elements of the reward sequence \( r(x_1), r(x_2), \ldots, r(x_n) \), and therefore we change the state of the process \((\bar{x}, n)\) to \((\bar{y}, K + 1)\), where \( \bar{y} = \sum_{i=n-K}^{n} r(x_i)(K + 1)^{-1} \). We will denote this restricted Gittins Index \( N_\sigma(K) \).

Normal reward process with unknown mean and variance: Similarly to the normal reward process with known variance, we change the state \((\bar{x}, s, n)\) to \((\bar{y}, s_2, K + 1)\), where \( s_2 = \sum_{i=n-K}^{n} (r(x_i) - \bar{y})^2(K + 1)^{-1} \). We will denote this restricted Gittins Index \( N_s(K) \).

4 Experiments

We test four paid subjects (graduate students, averaging 30.4 years old) on 10 2-arm, 20 3-arm, and 30 4-arm bandit problems with different configuration of probabilities and rewards drawn from \( \{0.1, 0.2, \ldots, 0.6\} \). At the beginning of the experiment, we emphasize that each arm has a non-zero probability of payoff, and that they should assume a fifty percent chance of winning at first.

Each arm is shown in the screen as a square over which a click means a pull. The amount of reward is shown blinking over the place of the click, green colored with a non-zero reward, and gray-colored with a zero reward. The total amount of reward given by an arm is shown inside the corresponding square. Additionally, the color of the square varies according to the winning and losses of it. There is a gradient from red (only losses), through yellow (equal number of winnings and losses), and green (only winnings). The number of pulls left (shared by all the arm) is show at the top-center of the screen. We establish purely-practical time limit of 150 seconds to each problem, a limit that was never reached.

Note that the notion of discount future was treated as an interest rate of \( 1/\gamma - 1 \) applied to the total reward gathered. We believe that this is a much more powerful, yet theoretically equivalent way to get the subject to perform greedy decision at some arbitrary degree. The standard way to give the notion of discounted future is to say that, at every decision time, there is a \( 100(1-\gamma)\% \) chance of abruptly ending the process. The problem with this approach is that the experiment needs to be finished at some random point, and therefore the analysis cannot be made in a uniform manner throughout subjects. On the contrary, using the notion of interest rate, we can fix the number of pullings and make it explicit to the subject. For every problem, we establish a number of 100 pullings, with a interest rate of 5% (or, equivalently, a discounting rate \( \gamma \approx 0.9525 \)).

On the left-hand side of the screen, the reward with interest is shown as a bar that raises from the bottom to the top of the screen with the numerical amount at the top. Each time the subjects pull, the interest rate is applied to the reward gathered.

4.1 Model fitting and performance evaluation

We fit the data using the Gittins Index assuming that the subject considers the process to be Bernoulli, or Normal. For the Bernoulli case, we fit all possible combinations of the model \( B(\bar{K} \in \{1, \ldots, 1, \infty\}, T \in \{1, \ldots, 5, \infty\}) \). For normal reward process with known variance, we fit \( N_\sigma(\bar{K} \in \{1, \ldots, 1, \infty\}) \). For the unknown variance case, we fit \( N_s(\bar{K} \in \{1, \ldots, 1, \infty\}) \).

4.1.1 Other models for comparison

Additionally, we compare the performance of the proposed models with the two best models from Gans, Knox, and Croson [7]. The Exponential Smoothing index that fits both \( e_0 \) and \( \xi \) showed the
best performance. On the other hand, the Hot Hand showed the second best performance by not precisely being an index, but a simple rule. This simplicity is very appealing to be tested as it is claimed to model human behavior.

**Exponential Smoothing:** Exponential Smoothing is a simple model of sequential decision making that discounts earlier sample information, and weights the more recent one by a complement of a discounting factor $\xi$.

At the beginning, the subject has a prior $e_0$ about the value of the reward. When the subject gets a reward $r(x_n)$ at time $n$, the subject update the index-like value of the arm using $e_n = \xi r(x_n) + (1 - \xi) e_{n-1}$, where $0 < \xi < 0$.

We fit the following Exponential Smoothing models $E(e_0 \in \{0, 0.1, \ldots, 1\}, \xi \in \{0.1, 0.2, \ldots, 0.9\})$.

**Hot Hand:** The Hot Hand model is even simpler than Exponential Smoothing. The term was coined in basketball where it is commonly belief that players have hot hands: a previous success in a free-throw will condition the success in the next free-throw. In the 2-AB problem, this means to switch the arm after an unsuccessful pull.

We can relax the rule by allowing a tolerance of $K$ losses, in which the original rule would have $K = 1$. It is important to notice that the Hot Hand model is not an index: it does not assign a number to each arm. We will simply assign an index 1 to the arm that is currently being played by Hot Hand and 0 to all others. If the last $K$ pulls are losses, then the model will switch to a different, random arm.

We denote the Hot Hand model with $K$-tolerance $H(K)$, and fit it to data with parameters $K \in \{1, 2, \ldots, 10\}$.

**4.1.2 Performance Evaluation**

At each decision time, a model $\nu$ will assign an index to each arm. Instead of counting how many times the model chooses the arm that the subject effectively pulls (i.e., the times it correctly assigns the highest index to it), we use an exponential distribution that accounts for how likely the model’s decision is with respect to the subject’s. If the indices assigned to arms are $\nu(\pi_1), \nu(\pi_2), \ldots, \nu(\pi_M)$ and the arm chosen by the subject is $z$, then

$$L(\pi_1, \pi_2, \ldots, \pi_M \mid z) = \frac{\exp\{\eta\nu(\pi_z)\}}{\sum_{j=1}^M \exp\{\eta\nu(\pi_j)\}}$$

is the model’s likelihood when arms are at states $\pi_1, \pi_2, \ldots, \pi_M$, where the constant $\eta$ is fit to the data.

To compare the hypotheses proposed by the models, we use the Bayesian hypothesis testing using the Schwarz Criterion or Bayesian Information Criterion (BIC) [11], $\text{BIC} = -2\log(L) + p \log(n)$, where $L$ is the likelihood of the hypothesis, $p$ is the number of parameters of the model, and $n$ is the sample size.

Moreover, it is very important to analyze the effectiveness of a model in the early steps of the decision process because these are mainly information-gatherers to decrease uncertainty. To better analyze this phenomenon, we disaggregate the data process in 10 phases of 10 decisions. For every phase and every model, we computed the likelihood of the subject’s choices for each model. We plug this likelihood in the BIC equation, and get the model’s effectiveness in a phase. Finally, we consider the overall model’s effectiveness in a problem as the average BIC across phases.

**4.2 Results**

Table 1 details the models’ parameters that showed best average BIC across phases. In the two-arm bandit problems, Hot Hand does a pretty good job considering its simplicity, outperforming exponential smoothing. Actually, exponential smoothing is outperformed by Gittins Index with Normal reward process, both with known and unknown mean, and only wins over the Bernoulli reward process. In the three- and four-armed bandit problems, the Index for a Normal reward process with known variance outperformed all other models.
Table 1: Performance of each model with best parameters on the problem set. The performance is measured as the Mean BIC across phases.

<table>
<thead>
<tr>
<th>Problem set</th>
<th>Model</th>
<th>Mean BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-Arm Bandit Problems</td>
<td>H(K = 8)</td>
<td>16.65</td>
</tr>
<tr>
<td></td>
<td>Nσ(K = 7)</td>
<td>17.28</td>
</tr>
<tr>
<td></td>
<td>Nσ(K = 9)</td>
<td>17.72</td>
</tr>
<tr>
<td></td>
<td>E(ε0 = 0.1, ξ = 0.6)</td>
<td>21.64</td>
</tr>
<tr>
<td></td>
<td>B(K = 9, T = 2)</td>
<td>22.00</td>
</tr>
<tr>
<td>3-Arm Bandit Problems</td>
<td>Nσ(K = 9)</td>
<td>17.72</td>
</tr>
<tr>
<td></td>
<td>E(ε0 = 0.1, ξ = 0.85)</td>
<td>28.73</td>
</tr>
<tr>
<td></td>
<td>Nσ(K = 1)</td>
<td>29.16</td>
</tr>
<tr>
<td></td>
<td>B(K = ∞, T = 2)</td>
<td>29.52</td>
</tr>
<tr>
<td>4-Arm Bandit Problems</td>
<td>Nσ(K = 3)</td>
<td>24.78</td>
</tr>
<tr>
<td></td>
<td>H(K = 8)</td>
<td>27.62</td>
</tr>
<tr>
<td></td>
<td>Nσ(K = 3)</td>
<td>30.45</td>
</tr>
<tr>
<td></td>
<td>H(K = 9)</td>
<td>33.42</td>
</tr>
<tr>
<td></td>
<td>Nσ(K = 3)</td>
<td>34.12</td>
</tr>
<tr>
<td></td>
<td>E(ε0 = 0.1, ξ = 0.85)</td>
<td>34.37</td>
</tr>
<tr>
<td></td>
<td>B(K = ∞, T = 2)</td>
<td>35.21</td>
</tr>
</tbody>
</table>

Figure 1: Performance of each model with best parameters on the set of problems, disaggregated in 10 phases. The first phases are for exploration, while the later ones are for exploitation. The Gittins Index for the Normal reward process with known variance show consistent performance throughout phases, being the best in 3-AB and 4-AB problems.

Figure 1 details the BIC performance disaggregated in 10 phases. Taking a more careful look at Hot Hand in the 2-AB problems, we can see that it performs quite poorly at the three first phases suggesting a defective capacity to model the crucial exploratory phase. It is interesting to notice that the Normal reward process with unknown variance models consistently all the phases, suggesting an overwhelming capacity to model not only the exploratory, but also the exploitative phases.

5 Conclusions and future work

The good performance of the Normal reward process with unknown variance suggests that people may be performing approximate Bayesian inference with limited memory. As the Bernoulli reward process is actually the true process underlying the experiment, it assumes that subjects know both the generating process and reward possibilities a priori. Rather than attribute failures of the Bernoulli model to poor exploration/exploitation, we show that it is important to correctly configure the Bayesian models so that they reflect subjects’ prior understanding of the task and capabilities. Obviously, a larger-scale experiment needs to be performed, involving more arms and payoff con-
figurations that better target distinctions between update models. Clearly the early stages of decision making behavior are most informative. Disaggregating the BIC values to separately analyze different stages of the decision process proved useful, and suggest that designing experiments that focus on first phases of the reward process are critical to quantifying exploration behavior.

There is one important deviation from MAB optimality in human decision making we did not model: a relaxation of the “frozen state” assumption for unplayed arms. Faced with making decisions in a bandit problem, people may entertain the hypothesis that unplayed arms may change states. Bandit problems that allow state changes on non-selected arms are termed “Restless bandit” problems, for which only approximate solutions are known to exist[12, 13]. Recently, Daw et. al[14] performed a brain imaging study while subjects made choices between restless bandits. Although optimality was not tested, their subject’s choices were well fit by a model that used a Kalman filter to track estimates of each arm’s reward value, suggesting subjects may have changed their beliefs about unplayed arms at each decision time. Future work will target whether subject’s belief updating also involve estimating the stability of the reward processes while unplayed.

References