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

Daniel Acuña (ACUNA@cs.umn.edu)
Department of Computer Science and Engineering
University of Minnesota
Minneapolis, MN 55455 USA

Paul Schrater (SCHRATER@umn.edu)
Department of Psychology and Computer Science and Engineering
University of Minnesota
Minneapolis, MN 55455 USA

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 Bayesian models of human behavior for the Multi-Armed Bandit Problem (MAB) on experimental data perform better than previous accounts.

Keywords: Human sequential decision making; Exploration/exploitation; Bayesian modeling; Multi-armed Bandit Problem; Gittins index

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 tradeoff, because of the conflicting desires for maximizing information (i.e., exploration) and maximizing reward (i.e., exploitation). Bayesian optimal solutions to this problem are notoriously intractable (Lusena, Goldsmith, & Mundhenk, 2001). However, humans engage in sophisticated sequential decision-making behavior in everyday life. The ability to perform correct decisions has high-impact in the quality of life. Good decision-makers may correctly choose products, services, job, and retirement plans. On the other hand, bad decision-makers may not foresee risks and develop addictions, gambling problems, or obsessions. 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 Multi-Armed Bandit Problem (MAB) offers a good opportunity to test human sequential decision-making given its simplicity and widely-known optimal solution through the Gittins Index. In MAB problems, at each decision time the decision-maker must select one of several 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 investigated human decision making for MAB problems, and have generated mixed results. Earlier studies suggested human choices reflect inaccurate Bayesian updating with suboptimalities in exploration–different studies found under-exploration (Meyer & Shi, 1995; Horowitz, 1973), over-exploration (Anderson, 2001) and both (Banks, Olson, & Porter, 1997). Recently, Gans, Knox, and Croson (2007) looked at the predictive performance of a large class of optimal and heuristic choice mechanisms and found human choice behavior is 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 on non-Bayesian decision-making. One of the critical difficulties with this conclusion is that the Bayesian updating used by the optimal comparison model contains 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. (Gans et al., 2007). We find that a belief-update model that estimates the mean and variance of the reward process with limited memory and look ahead model provides the best predictive performance.

Modeling beliefs in MAB problems

In MAB problems, selecting arm i at time k generates a reward x_k^i. Our aim is to construct a action selection policy that maximizes the expected total discounted future reward

$x_0 + \gamma x_1 + \gamma^2 x_2 + \ldots$, where γ ($0 \leq \gamma < 1$) is a *discounting factor* that allows the infinite sum to converge. Given that the nature of the arm is unknown, we say the arm is at a belief state π. For most common reward processes, this state is represented by the sufficient statistics of the reward dynamics. We utilize this state to model the probabilistic transition $p(\pi_{k+1}|\pi_k, x_k)$ between states and the expected reward $p(x_k|\pi_k)$. In building the policy, if we select arm i at a time t steps in the future, we expect a reward $\gamma^t r(\pi_{k+t})$, where $r(\pi) = \mathbb{E}[x|\pi]$ is the expected reward of state π, and $\mathbb{E}$ is the expectation operator. Thus, MAB problems are special cases of Markov Decision Processes, and hence have an associated Bellman equation (Bellman, 1957) that can be solved in principle via dynamic programming. However, Gittins (Gittins, 1989) 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.

The Gittins Index

The virtue of Gittins' 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 $\tau - 1$ (a stopping time), divided by the total discounted time up to $\tau - 1$. More precisely,

$$\nu_i(\pi^i) = \sup_{\tau>0} \mathbb{E}_{\pi^i}\left[\sum_{t=0}^{\tau-1} \gamma^t r(\pi^i_t)\right] \Big/ \mathbb{E}_{\pi^i}\left[\sum_{t=0}^{\tau-1} \gamma^t\right] \quad (1)$$

$$= (1-\gamma) \sup_{\tau>0} \mathbb{E}_{\pi^i}\left[\sum_{t=0}^{\tau-1} \gamma^t r(\pi^i_t)\right] \Big/ (1 - \mathbb{E}_{\pi^i}[\gamma^\tau]). \quad (2)$$

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. For a given arm, rename the numerator and denominator in equation (2) by $R^\tau(\pi)$ and $W^\tau(\pi)$, respectively. Let $\tau^* - 1$ denote an optimal stopping time. The exploration bonus exceeds its expected reward payoff by

$$(1-\gamma)\frac{1 - W^{\tau^*}(\pi)}{W^{\tau^*}(\pi)} R^{\tau^*}(\pi).$$

The term $(1-\gamma)$ converts discounted reward into the undiscounted reward the agent expects to receive, because the Gittins Index (1) is defined in terms of the rate undiscounted expected *reward* is accrued. By dropping it, we recover the expected *value* $R^{\tau *}(\pi)/W^{\tau *}(\pi)$, which is more commonly used in classical dynamic programming literature (Bellman, 1957).

Using the fact that discounting is equivalent to a probability that a reward process will terminate, Sonin (Sonin, 2007) showed that $W^\tau(\pi)$ could be interpreted as the probability of reward termination within time $\tau - 1$. Thus, the Gittins index gives a bonus to arms it believes will survive the time $\tau^* - 1$. 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 reliability of each arm's payoff using strategies that are suboptimal for the experimentally imposed task.

Modeling human belief updates Because previous investigations of human behavior in MAB problems have almost exclusively focused on Bernoulli reward processes, we had our observers choose between Bernoulli arms that generated sequences $\{x^i_1, x^i_2, \ldots, x^i_n\}$ of independent and identically-distributed random values taking either R, with $R \in [0, 100]$ fixed, or 0 with probability θ, and $1 - \theta$, respectively. Without lost of generality, we develop our analysis with R equals 1. It is easy to see that for any $R \in \mathbb{R}$, the Bernoulli reward process index $\nu(\pi, R)$ of a state π is equal to $R \times \nu(\pi, 1)$, where $\nu(\pi, 1)$ is the index for $R = 1$.

To compute Gittins index using the calibration method. This method *calibrates* an arm by comparing it with a *standard* bandit process, which has one state and a constant reward λ. The method works by finding the supremum amount of reward λ such that we would be indifferent on whether to play the standard arm or the calibrated arm. For a given arm, the model assumes the x_is are drawn from a parametric distribution indexed by θ with a density function $f(\cdot|\theta)$. The prior density for θ is denoted by π. The base for the calibration process is the Bellman equation rewritten as:

$$\mathrm{U}(\lambda, \pi) = \max\left[\frac{\lambda}{1-\gamma}, r(\pi) + \gamma \int \mathrm{U}(\lambda, \pi_x) f(x|\pi)\, dx\right], \quad (3)$$

where π_x denotes the posterior $\pi(\theta|x)$, and $f(\cdot|\pi) = \int f(\cdot|\theta)\pi(\theta)d\theta$. We solve equation (3) for different values of λ as π varies throughout the family of posterior distributions. Note that when both arguments are equal inside the brackets in (3) for certain λ, the Gittins index is λ.

Although this an infinitely recursive equation, the influence of a h-nested equation is diminished by $O(\gamma^h)$. Thus, the procedures to compute the indices are extremely accurate approximations. In particular, we employ the methods described in (Gittins, 1989) until a horizon of $h = 2000$ when an "infinite" recursion is required.

To simulate human choice data in our experiments, we compute Gittins indices for a set of models that are Bayesian but differ from the generating model. We introduced three belief models: the optimal model based on the generating process, and two other models that simulate inaccuracy in the belief. These inaccurate models are Gaussian approximations to the Beta-Bernoulli model, which is the optimal. One model estimates θ assuming a standard deviation, and the other estimate both. It is important to notice that all models enjoy posterior consistency (Diaconis & Freedman, 1986). Namely, as more reward is observed, the posterior of each model will almost surely converge to an infinitesimal neighborhood around the true parameter θ. The Gaussian models would potentially take longer to converge though. However, we believe it is easier –and even more natural– to adopt a belief about θ that is symmetric like the Gaussian because it allows to separately estimate the standard deviation and the mean. Also, the optimal belief for a Bernoulli reward, a Beta distribution, has a complicated shape and both the mean and standard deviation are intricately related.

Bernoulli reward process: For a Bernoulli reward process, we will have a prior over θ as a Beta distribution with a density in the interval [0,1]

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

where $\alpha > 0$ and $\beta > 0$, and the prediction

$$f(x|\alpha,\beta) \quad = \quad \begin{cases} \frac{\beta}{\alpha+\beta} & \text{if } x = 0 \\ \frac{\alpha}{\alpha+\beta} & \text{if } x = 1 \end{cases}$$

The state may be completely represented by α and β. The reward is

$$r(\alpha,\beta) = \int_0^1 \theta \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\,d\theta = \frac{\alpha}{\alpha+\beta}.$$

Therefore, the calibration equation (3) becomes

$$\mathrm{U}(\lambda,\alpha,\beta) = \max\left[\frac{\lambda}{1-\gamma}, \frac{\alpha[1+\gamma\mathrm{U}(\lambda,\alpha+1,\beta)]+\beta\gamma\mathrm{U}(\lambda,\alpha,\beta+1)}{\alpha+\beta}\right] \quad (4)$$

Normal reward process with unknown mean and known variance: Here, the reward function follows a normal distribution with unknown mean, but known variance σ^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

$$\mathrm{U}(b\lambda+c, b\bar{x}+c, b\sigma) = c(1-\gamma)^{-1} + b\mathrm{U}(\lambda,\bar{x},n,\sigma), \quad (5)$$

where $b > 0, c \in \mathbb{R}$, and

$$\nu(\bar{x},n,\sigma) = \bar{x} + \sigma\nu(0,n,1). \quad (6)$$

Without lost of generality, we assume a known variance $\sigma^2 = 1$, and therefore leave the true mean μ as the only unknown parameter. We take π_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

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

and

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

Thanks to (5) and (6), we only need to solve the following calibration equation for $\bar{x} = 0$ to find any other index

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

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

$$f(x|\bar{x},s,n) \propto \left(1 + \frac{n}{n+1}\frac{(x-\bar{x})^2}{(n-1)s^2}\right)^{-n/2},$$

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

Similarly to (5) and (6) (Gittins, 1989), the following property $\nu(\bar{x},s,n) = \bar{x} + s\nu(0,1,n)$ holds. Therefore, the calibration method is simplified to

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

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

Restricting the memory and look ahead for computing the Gittins Index

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.

We model limited memory by assuming that subjects keep in memory the last m reward observations to infer the belief of a state. Note, we assume subjects do not build their beliefs from the entire reward history, but rather use a limited history of rewards. We incorporate the idea of limited memory in an additional step performed before the Gittins index is computed. This step consists of building up the state of the arm from the last m observations as

$$\pi_k(\theta|x_1,x_2,\ldots,x_m) = \pi_0(\theta)\prod_{i=1}^{m} f(x_i|\theta), \quad (7)$$

with $\pi_0(\theta)$ being fixed. Thereafter, the Gittins Index is computed as before, using the limited memory state π_k.

Additionally, we can easily incorporate a limited look ahead in the computation of the Gittins Index by allowing a maximum of h steps in the recursive computation. We use the following modified calibration equation

$$\mathrm{U}(\lambda,\pi,h) = \max\left[\frac{\lambda}{1-\gamma}, r(\pi)+\gamma\int \mathrm{U}(\lambda,\pi_x,h-1)\,f(x|\pi)\,dx\right] \quad (8)$$

when $h>0$, and $\mathrm{U}(\lambda,\pi_x,0)=\mathrm{S}(\lambda,\pi_x)$, and

$$\begin{aligned} S(\lambda,\pi) &= \max\left[\frac{\lambda}{1-\gamma}, \sum_{t=0}^{\infty}\gamma^t r(\pi)\right] && (9)\\ &= \max\left[\lambda, r(\pi)\right]/(1-\gamma) && (10)\end{aligned}$$

is a myopic expected future reward based on current state π.

*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 (7) and considering only the subsequence $x_{n-m},x_{n-m+1},\ldots,x_n$. If there are n_0 occurrences of 0 and n_1 occurrences of 1 in this subsequence, the state of the process new state of the process is $\alpha=n_1+1$ and $\beta=n_0+1$. To restrict the look ahead, we use the modified calibration (8) with $\mathrm{S}(\lambda,\alpha,\beta)=\max\left[\lambda,\alpha(\alpha+\beta)^{-1}\right]/(1-\gamma)$.

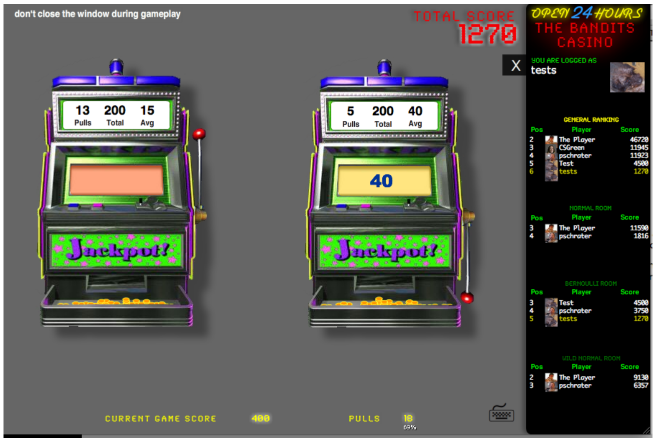
*Normal reward process with unknown mean and known variance***:** We restrict the memory by considering the last m elements of the reward sequence $x_1,x_2,\ldots,x_n$, and therefore we change the state of the process $(\bar{x},n)$ to $(\bar{y},m+1)$, where $\bar{y}=\sum_{i=n-m}^{n}x_i(m+1)^{-1}$. We restrict the look ahead by using $S(\lambda,\bar{x},n)=\max\left[\lambda,\bar{x}\right](1-\gamma)^{-1}$. The parameters $\bar{x}_0$ and σ_0 denote the prior belief about mean and standard deviation, respectively.

*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 $(\bar{y},s_2,m+1)$, where $s_2=\sum_{i=n-m}^{n}(x_i-\bar{y})^2(m+1)^{-1}$. We restrict the look ahead by using $S(\lambda,\bar{x},s,n)=\max\left[\lambda,\bar{x}\right](1-\gamma)^{-1}$.

Experiments

We test five paid subjects (graduate students, averaging 29 years old) on 10 2-arm, 20 3-arm, and 30 4-arm bandit problems. The payoff configurations were random samples from $\{0.2,\ldots,0.8\}$, the rewards drawn from the interval $[0,100]$, and discount is factor 0.98. To better compare the models ability to predict subjects data, the payoff configurations and the stopping times of each problem was the same for all subjects, but were randomly presented.

We emphasized to subjects that each arm has a non-zero probability of payoff and that it remains static throughout each game. We described that a game may suddenly stop after any pull with 2% probability. We told them that the probability of stopping does not increase within a game, but that to survive the n-th pull they need to have survived all the previous pullings. Additionally, we show to subjects the quantile of people that survive until the 20, 40, 60, 80, 100, 200, and 300-th pull, which is 66.7%, 44.5%, 29.7%, 19.8%, 13.2%, 1.75%, and 0.23%, respectively.


don't close the window during gameplay

YOU ARE LOGGED AS
tests

GENERAL RANKING

Pos	Player	Score
2	The Player	46720
3	CSGreen	11945
4	pschrater	11923
5	Test	4500
6	tests	1270

NORMAL ROOM

Pos	Player	Score
3	The Player	11590
4	pschrater	1816

BERNOULLI ROOM

Pos	Player	Score
3	Test	4500
4	pschrater	3750
5	tests	1270

WILD NORMAL ROOM

Pos	Player	Score
2	The Player	9130
3	pschrater	6357

Figure 1: Game screenshot

Each arm is shown in the screen as a slot machine. Subjects pull a machine by pressing a key in the keyboard. When pulled, an animation of the lever is shown, 200 msec later the reward appears in the machine's screen, and a sound mimicking dropping coins lasts proportionally to the amount gathered. No sooner than 1 secs, the subject can pull again.

A machine has several cues, some redundant, to help subjects keep track of previous rewards. At the top, the machine shows the number of pulls, total reward, and average reward per pull so far. The machine's screen changes the color according to the average reward, from red (zero points), through yellow (fifty points), and green (one hundred points). The machine's total reward is shown as a pile of coins underneath it

The total score and total pulls within a current game is shown at the bottom of the screen. The total score throughout the games is shown at the top-right. Additionally, a ranking with the score of the other players is shown to the left. A typical game with two arms is shown in Figure 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.

Other models for comparison Additionally, we compare the performance of the proposed models with the two best models from Gans, Knox, and Croson (Gans et al., 2007).

Exponential Smoothing: Exponential Smoothing is a simple model of for belief updating that discounts earlier sample information in favor of the most recent sample using a weighting factor ξ, where $0<\xi<1$. Beliefs about the value of the reward are initialized to a prior e_0. After each reward x_n at time n, the exponential smoother updates an estimate of the value of the arm using $e_n=\xi x_n+(1-\xi)e_{n-1}$. It is an appropriate model for agents that believe the environment is changing, because it correctly considers that early reward information has less influence in the current estimate of expected reward.

Hot Hand: The Hot Hand, a model that captures subjects beliefs in performance streaks, is a simple heuristic model that has shown to be surprisingly good at modeling human behavior in Bernoulli trials. In the 2-arm problem, the hot hand model switches arms after an unsuccessful pull. We can relax the 1-pull rule by allowing a tolerance of k losses. It is important to notice that the Hot Hand model is not an index: it does not assign a number to each arm. We assign an index 1 to the arm that is 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.

Performance Evaluation Assuming subjects have perfect internal discrimination between their indices, our observation of their decisions contains noise from unobservables in measurements, and proxy or instrumental variables. Hence, our decision analysis has a random component that may be modeled by a probability function. We assume a common analysis of event histories in which the random noise is i.i.d. distributed according to a double exponential distribution (Gumbel distribution) (Allison, 1982).

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), the Gumbel distribution over a subject decision boils down to a logistic distribution for the model response (Gans et al., 2007). This logistic distribution can be interpreted as *how likely* the model's decision is with respect to the subject's.

At each decision time, a model ν will assign an index to each arm. Let $\pi^1, \pi^2, \ldots, \pi^l$ be the states of the arms after interaction with a subject. Let $\nu(\pi^1), \nu(\pi^2), \ldots, \nu(\pi^l)$ be the indices assigned to arms by the model ν. Finally, likelihood of a model decision given subject' decision z is

$$\frac{\exp\{\eta\nu(\pi^z)\}}{\sum_{j=1}^{l}\exp\{\eta\nu(\pi^j)\}}. \tag{11}$$

This model response measure adds an additional level of hypotheses across a subset of decisions to construct a framework for nested hypothesis testing. This new nested spaced is characterized by the parameter η (with $0 < \eta < \infty$). This allowed us to assess the performance of a model while accounting for the variability either across subjects or within subjects. From this likelihood (11), we can easily build an aggregate log likelihood for the full set of decisions. Equivalently, we use the negative log-likelihood (nLL) to evaluate model performance.

To account for the complexity of a model, we utilize BIC and cross validation. The BIC measure (Bayesian hypothesis testing (Kass & Raftery, 1995)) penalizes the nLL by the number of free parameters (d) using the formula $\text{BIC} = 2 \times \text{nLL} + d \times \ln(n)$, where n is the size of the dataset. Cross validation (CV) is a common statistical procedure that partitions the data into subsets such that the data used to fit the model are in one subset and the data to test it are in another. We use 10-fold cross validation with nLL as the performance measure.

For nLL, BIC, and CV, we search for the best model's parameters for given set of decisions in two steps. First, we discretize the parameter coarsely to find candidate parameter values. Second, we search for an η using non-linear optimization such that the nLL is minimized. We did not find any significant difference in our conclusions when η was fit across subjects or within subjects. The results reported consider one η for all subjects.

Results

The dataset consists of 18098 decisions spanned over 3228 decisions for 2-arm problems, 5873 decisions for 3-arm problems, and 8997 decisions for 4-arm problems. Table 1 details the models performance.

Exponential smoothing best fits subjects decisions on 2-arm problems, which is consistent with the literature. Notably, our limited Gittins index for Bernoulli with unlimited memory (subject should have a clear idea of the average payoff) and 2-step look ahead outperforms the popular Hot Hand heuristic. However, the optimal Gittins index (not shown in the table) is worse than Hot Hand at modeling subject choices, confirming previous results that simple strategies are more predictive of human choices than the optimal 2-armed bandit strategy.

Interestingly, on 3-arm and 4-arm problems, our limited Gittins index with Normal reward and known variance outperforms all other models, although the exponential smoother remains viable. The predictive performance of the Normal reward and known variance may result from its ability to represent key limitations in human sequential decision-making. In particular, we think subjects encode reward feedback with noise, which is not captured by a binary model reward model. However, subjects may be aware of the noise in their encoding process, consistent with the poor performance of the Gittins index with normal reward and unknown variance. The unknown variance model keeps very high index to arms with few samples, even though subjects should get a clear idea of payoff using only one successful sample. Note that without limited memory and limited lookahead, the modified Gittins index models would not account for subject's decisions.

Conclusions and future work

The good performance of the Normal reward process with known variance suggests that people may perform approximate Bayesian inference with limited memory and look ahead. As the Bernoulli reward process is actually the true process underlying the experiment, optimal performance requires that subjects know both the generating process and reward process a priori. Rather than attribute failures of the Bernoulli model to poor exploration/exploitation, we show

Table 1: Performance of models under 10-fold cross validation using negative log-likelihood (CV), negative log-likelihood (nLL), and BIC. The best parameters for each model according to BIC are shown in parenthesis.

	Set of Data			
Performance measure	**2 arms**	**3 arms**	**4 arms**	**All**
Exponential Smoothing				
CV	1983	4019	6025	12021
nLL	2002	4038	6044	12067
BIC ($e_0 = 0, \xi = 0.1$)	**4020**	8093	12106	24153
Hot Hand				
CV	3093	6196	9291	18558
nLL	3098	6201	9296	18612
BIC ($k = 1$)	6204	12410	18601	37233
Gittins Index Bernoulli Process				
CV	2382	4794	7194	14293
nLL	2392	4804	7204	14350
BIC ($m = \infty, h = 2$)	4800	9625	14426	28719
Gittins Index Normal Process with known variance				
CV	1993	3992	5986	12026
nLL	1996	3995	5989	12029
BIC ($\bar{x}_0 = 0, s = 10, m = 4, h = 3$)	4024	**8024**	**12014**	**24097**
Gittins Index Normal Process with unknown variance				
CV	3462	6959	10394	20860
nLL	3478	6975	10410	20876
BIC ($\bar{x}_0 = 0, m = 3, h = 3$)	6980	13976	20847	41781

that it is important to correctly configure the Bayesian models so that they reflect subject's prior understanding of the task and their capabilities. Obviously, a larger-scale experiment needs to be performed, involving more arms and payoff configurations that better target distinctions between update models.

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 (Whittle, 1988; Mora, 2001). Recently, Daw et al. (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006) 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

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, *13*, 61-98.

Anderson, C. (2001). *Behavioral models of strategies in multi-armed bandit problems*. Ph.d., California Institute of Technology, Pasadena, CA.

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

Bellman, R. E. (1957). *Dynamic programming*. Princeton: Princeton University Press.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876-879.

Diaconis, P., & Freedman, D. (1986). On the consistency of bayes estimates. *The Annals of Statistics*, *14*(1), 1–26.

Gans, N., Knox, G., & Croson, R. (2007). Simple models of discrete choice and their performance in bandit experiments. *Manufacturing and Service Operations Management*, *9*(4), 383-408.

Gittins, J. C. (1989). *Multi-armed bandit allocation indices*. Chichester [West Sussex] ; New York: Wiley.

Horowitz, A. D. (1973). *Experimental study of the two-armed bandit problem. ph.d. dissertation*. Chapel Hill, NC: University of North Carolina, Chapel Hill.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795.

Lusena, C., Goldsmith, J., & Mundhenk, M. (2001). Nonapproximability results for partially observable markov decision processes. *JAIR*, *14*, 83-103.

Meyer, R. J., & Shi, Y. (1995). Sequential choice under ambiguity: Intuitive solutions to the armed-bandit problem. *Management Science*, *41*, 817-83.

Mora, J. Niño. (2001). Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability*, *33*, 76-98.

Sonin, I. M. (2007). A generalized gittins index for markov chain and its recursive calculation. *submitted to Statistics and Probability Letters*.

Whittle, P. (1988). Restless bandits: activity allocation in a changing world. *Journal of Applied Probability*, *25A*, 287-298.