Human learning and decision-making in the bandit task: Three wrongs make a right

Dalin Guo (dag082@ucsd.edu)
Cognitive Science, UC San Diego, 9500 Gilman Dr.
La Jolla, CA 92093, USA

Angela J. Yu (ajyu@ucsd.edu)
Cognitive Science, UC San Diego, 9500 Gilman Dr.
La Jolla, CA 92093, USA

Abstract

Humans and animals frequently need to make choices among options with imperfectly known reward outcomes. In neuroscience, this is often studied using the multi-armed bandit task, in which subjects repeatedly choose among bandit arms with fixed but unknown reward rates, thus negotiating a tension between exploitation and exploration. Here, using a modified version of the bandit task in which we query subjects reward expectations of unchosen arms, we investigate how general reward availability in the environment affects human prior beliefs. Based on self-report data and computational modeling of behavioral data, we obtain converging evidence that human subjects systematically under-estimate reward availability. Additional computational analyses reveal that this under-estimation compensates for two other apparent suboptimalities in human behavior, namely a default assumption of environmental non-stationarity, and the use of a simplistic decision policy. This result represents a concrete instance in which multiple sub-optimalities in brain computations synergistically interact to achieve much better-than-expected behavioral outcome. This work raises the intriguing possibility that many apparently isolated limitations in brain computation and representation may actually work together to achieve highly intelligent behavior in a broader context, and also sheds light on computationally efficient algorithms that could be adopted by artificial intelligence systems.

Keywords: Multi-armed Bandit; Reinforcement Learning; Bayesian Model

Introduction

Humans and animals frequently have to make choices among options with imperfectly known outcomes. This is often studied using the multi-armed bandit task (Cohen et al., 2007), in which the subject repeatedly chooses among bandit arms with fixed but unknown reward probabilities. The observer learns how rewarding an arm is by choosing it and observing whether it produces a reward, thus each choice pits exploitation against exploration since it affects not only the immediate reward outcome but also the longer-term information gain. Previously, it has been shown that human learning in the bandit task is well captured by a Bayesian ideal learning model (Zhang & Yu, 2013), the Dynamic Belief Model (DBM) (Yu & Cohen, 2009), which assumes the reward distribution to undergo occasional, unsignaled changes – this occurs despite the reward rates' actually being fixed during a game. While this finding was consistent with the default (and incorrect) non-stationarity assumption humans make in a variety of other psychological tasks (Yu & Cohen, 2009; Shenoy et al., 2010; Ide et al., 2013; Yu & Huang, 2014; Zhang & Yu, 2013), it has remained nevertheless rather mysterious why humans would persist making this assumptions despite inconsistent environmental statistics.

In this work, we present and model human behavioral data in a variant of the bandit task, in which we vary reward abundance and variability in different environments. We aim to examine how humans adapt their decision-making to the different reward environments. Specifically, we focus on whether human subjects have veridical prior beliefs about reward rates. To gain greater computational insight into human learning and decision making, we compare the ability of DBM and a number of alternative models in their ability to capture the human data. Specifically, we consider two Bayesian learning models, DBM and Fixed Belief Model (FBM) (Yu & Cohen, 2009), coupled with a softmax decision policy. Besides Bayesian learning, we also include a simple reinforcement learning rule (RL), the delta rule (Rescorla & Wagner, 1972), which has been widely used in the neuroscience literature (Behrens et al., 2007). DBM is related to RL in that the stability parameter in DBM also controls the exponential weights as the learning rate in RL does, but they are not mathematically equivalent. For the decision policy, we employ the softmax policy, which is popular in psychology and neuroscience, and has been frequently used to model human behavior in the bandit task (Daw et al., 2006), and Knowledge Gradient, which previously found to capture the human behavior in bandit task the best among a few decision making models (Zhang & Yu, 2013), not including softmax.

Experiment

We recruited 107 UCSD students to participate in a four-armed, binary bandit task, whereby the reward rates in four environments were identically and independently sampled from four Beta distributions: Beta(4, 2), Beta(2, 4), Beta(30, 15) and Beta(15, 30). The reward rates for the 50 games (15 trials each) were pre-sampled, and randomized for each subject. The cover story was that is an ice fishing contest, where

the four arms represent four fishing holes. Participants are informed that the different camps they fish from reside on four different lakes that vary in (a) overall abundance of fish, and (b) variability of fish abundance across locations. At the outset of each environment, we tell them the lake's fishing conditions (high/low abundance, high/low variance) and provide samples from the distribution (a fishing report showing the number of fish caught out of 10 attempts at 20 random locations in the lake). A subset of 32 subjects were required to report the reward rate of the never-chosen arms at the end of each game.

The reported reward rates are shown in Fig. 1A. Human subjects reported estimates of reward rate significantly lower than the true generative prior mean ($p < .01$), except in low abundance and low variance environment ($p = 0.2973$). The average reported estimates across the four reward environments are not significantly different ($F(3,91) = 1.78, p = 0.157$, see Fig. 2A), indicating that humans do not alter their prior belief about the reward rates even when provided with both explicit (verbal) and implicit (sampled) information about the reward statistics of the current environment. In spite of systematically underestimating expected rewards, our subjects appear to do well in the task. The actual total reward accrued by the subjects are only slightly lower than the optimal algorithm utilizing correct Bayesian inference and the dynamic-programming-derived decision policy; humans also perform significantly better than the chance level attained by a random policy ($p < .001$), which is equal to the generative prior mean of the reward rates. Thus, subjects are actually experiencing sample reward rates that are higher than the generative prior mean (since they perform much better than the random policy); nevertheless, they significantly underestimate the mean reward rate.

Models

Model description

We denote the reward rate of arm k at time t as θ_k^t. We denote the reward outcome at time t as $R_t \in \{0,1\}$, and $\mathbf{R}^t = [R_1, R_2, \ldots, R_t]$. We denote the decision at time t as D_t, $D_t \in \{1,2,3,4\}$, and $\mathbf{D}^t = [D_1, D_2, \ldots, D_t]$.

Dynamic belief model (DBM). The generative dynamics is

$$p(\theta_k^t = \theta | \theta_k^{t-1}) = \gamma\delta(\theta_k^{t-1} - \theta) + (1-\gamma)p^0(\theta), \quad (1)$$

where $p^0(\theta)$ is the assumed prior distribution.

The posterior reward rate distribution given the reward outcomes up to time t can be computed iteratively as

$$p(\theta_k^t | \mathbf{R}^t, \mathbf{D}^t) \propto p(R_t | \theta_k^t) p(\theta_k^t | \mathbf{R}^{t-1}, \mathbf{D}^{t-1}), \text{ if } D_t = k \quad (2)$$

$$p(\theta_k^t | \mathbf{R}^t, \mathbf{D}^t) = p(\theta_k^t | \mathbf{R}^{t-1}, \mathbf{D}^{t-1}), \text{ if } D_t \neq k \quad (3)$$

The predictive reward rate distribution at time t given the outcomes up to time $t-1$ is:

$$p(\theta_k^t = \theta | \mathbf{R}^{t-1}, \mathbf{D}^{t-1}) = \gamma p(\theta_k^{t-1} = \theta | \mathbf{R}^{t-1}, \mathbf{D}^{t-1}) + (1-\gamma)p^0(\theta). \quad (4)$$

The expected (mean predicted) reward rate of arm k at trial t is $\hat{\theta}_k^t = \mathbb{E}[\theta_k^t | \mathbf{R}^{t-1}, \mathbf{D}^{t-1}]$.

Fixed belief model (FBM). FBM assumes stationarity and can be viewed as a special case of DBM, with $\gamma = 1$.

Reinforcement Learning (RL). The update rule is

$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \varepsilon(R_t - \hat{\theta}_k^{t-1}), \quad (5)$$

Softmax decision policy. Softmax assumes the choice probabilities among the options to be normalized polynomial functions of the estimated expected reward rates:

$$p(D_t = k) = \frac{(\hat{\theta}_k^t)^b}{\sum_i^K (\hat{\theta}_k^t)^b}, \quad (6)$$

Optimal policy. The multi-armed bandit problem can be viewed as a Markov decision process, where the state variable is the posterior belief after making each observation. The optimal solution to the problem considered here can be computed numerically via dynamic programming (Zhang & Yu, 2013; Averbeck, 2015), where the optimal learning model is FBM with the correct prior distribution. Previously, it has been shown that human behavior does not follow the optimal policy (Zhang & Yu, 2013); nevertheless, it is a useful model to consider in order to assess the performance of human subjects and the various other models in terms of maximal expected total reward.

Knowledge Gradient (KG). The knowledge gradient decision-making policy (Frazier & Yu, 2008) is an approximation to the optimal policy, which is much cheaper computationally. The algorithm is myopic as to compute the knowledge gain, it commits to one more exploratory decision at any given trial, and assumes to exploit in all remaining trials given the knowledge after that single additional exploration. The knowledge gain is computed as

$$v_k^t = \mathbb{E}[\max_{k'} \theta_{k'}^{t+1} | D_t = k, \mathbf{R}^{t-1}, \mathbf{D}^{t-1}] - \max_{k'} \theta_{k'}^t \quad (7)$$

The decision rule takes into account of current expected reward rates, knowledge gain, and horizon. The value function is computed as:

$$V_k^t = \hat{\theta}_k^t + (T - t - 1)v_k^t \quad (8)$$

Under DBM assumption, the horzion is computed as the min of the expected horizon (before a change point occur) and current horizon:

$$t' = \min\{\frac{1}{1-\gamma}, T - t\} - 1 \quad (9)$$

The original KG algorithm is deterministic, to allow decision noise, we add another layer of softmax decision rule with extra parameter b.

KG is previously found to be the best model among several decision policies (Zhang & Yu, 2013) (not including softmax). However, in a later study, softmax was found to explain human data better than KG (Harlé et al., 2015).

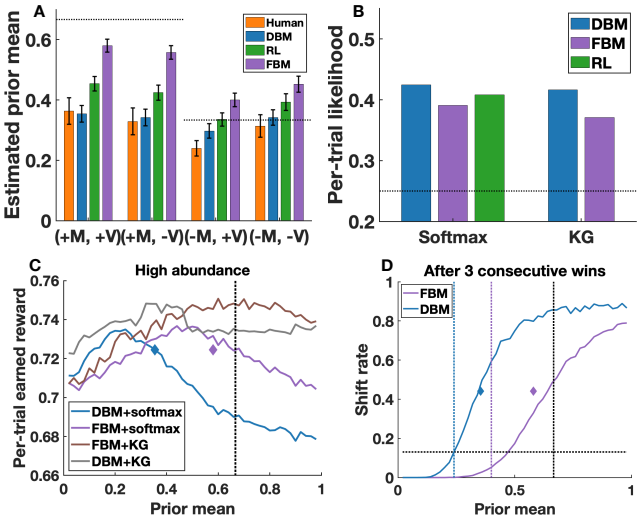

Figure 1: (A) Reported reward rate estimates by human subjects (orange), and fitted prior mean of DBM (blue), FBM (purple), and RL (green). Dotted lines: the true generative prior mean (0.67/0.33 for high/low abundance environments). Error bars: s.e.m. across participants or validation runs. (+M, -V) denotes high mean (abundance), low variance, and so on. (B) Averaged per-trial likelihood of 10-fold cross validation of three learning models coupled with two decision models. Dotted line: the chance level (0.25) (C) Reward rates achieved in high abundance and high variance by different models: DBM+softmax (blue), FBM+softmax (purple), DBM+KG (grey), FBM+KG (brown) and optimal policy (brown). The diamond symbols represent the actual reward per trial earned by human subjects (y-axis) vs. the fitted prior mean (x-axis) of the two models. Vertical dotted lines: true generative prior mean (D) shift rate (choosing different arms as the last trial) after three wins proceed by a loss of FBM+softmax and DBM+softmax models. The diamond symbols represent the actual shift rate by human subjects (y-axis) vs. the fitted prior mean (x-axis) of the two models.

Model comparison

Here, we compare the various models to human behavior, in order to identify the best (of those considered) formal description of the underlying psychological processes.

We first evaluate how well the three learning models fit human data. Since they have different numbers of parameters, we perform 10-fold cross-validation to avoid overfitting for comparison. We use per-trial likelihood as the objective function, calculated as $\exp(\log \mathcal{L}/N)$, where $\mathcal{L}$ is the maximum likelihood of the data, and N is the total data points. We fit prior weight ($\alpha+\beta$, related to precision) at the group level. We fitother parameters at the individual level, and separately for four reward environments.

Fig. 1B shows the held-out per-trial likelihood for softmax and KG with DBM, FBM, and RL, averaged across ten runs of cross-validation. Coupled with DBM or FBM, softmax achieves significantly higher per-trial likelihood than KG ($p < .001$) based on paired t-test, i.e. softmax decision policy behaves more like human subjects than KG. Coupled with softmax or KG, DBM achieves significantly higher per-trial likelihood than FBM ($p < .001$) and RL ($p < .001$) based on paired t-test, i.e., DBM predicts human behavior better than the other two learning models. This result corroborates previous findings (Zhang & Yu, 2013) that humans assume non-stationarity by default in the multi-armed bandit task, even though the reward structure is truly stationary, and they do not follow optimal or approximate optimal decision policy.

Next, we examine how well the learning models coupled with softmax can recover the underestimation effect observed in human participants. The reported estimation is on the arm(s) that they never chose at the end of each game, which is their belief of the mean reward rate before any observation, i.e., mathematically equivalent to the prior mean (DBM & FBM) or the initial value (RL). For simplicity, we will refer to them all as the prior mean. Fig. 1A shows the average fitted prior mean of the models. FBM recovers prior mean values that are well correlated with the true generative prior means ($r = +.96, p < 0.05$), and significantly different in the four environments ($F(3,424) = 13.47, p < .001$). The recovered prior means for RL are also significantly different in the four environments ($F(3,424) = 4.21, p < 0.01$). In contrast, the recovered prior means for DBM are not significantly different in the four environments ($F(3,424) = 0.91, p = 0.4350$), just like human estimates (Fig. 2A). DBM also recovers prior mean values in low abundance and high variance environment slightly lower than in other environments, similar to human reports. In summary, DBM allows for better recovery of human internal prior beliefs of reward expectation than FBM or RL.

Simulation results

Finally, we try to understand *why* humans might exhibit a "pessimistic bias" in their reward rate expectation. Fig. 1C shows the simulated average earned reward per trial in the high abundance environment, of the various models as a function of the assumed prior mean. The per-trial earned reward rates are calculated from the simulation of models/optimal policy under the same reward rates of the human experiment. We focus on the high variance and high abundance environments, since model performance is relatively insensitive to the assumed prior mean in other environments (not shown).

Firstly, consider the diamond symbols in Fig. 1C: the combination of human subjects' actual average per-trial earned reward (y-axis) and the fitted prior mean for each of the two models (x-axis, color-coded) is very close to DBM's joint predictions of the two quantities (blue lines), but very far away from FBM (purple line)'s joint predictions of the two quantities. This result provides additional evidence that DBM can predict and capture human performance better than the other two models.

More interestingly, while the FBM and KG (brown) achieves the highest earned reward when it assumes the correct prior

(as expected), when substitute the decision policy to softmax, or substitute the learning model to DBM, it achieves its maximum reward at a prior mean much lower than the true generative mean. This implies that one way to compensate for using the sub-optimal softmax policy, or having an incorrect nonstationary assumption, is to somewhat underestimate the prior mean. In addition, DBM and softmax achieves maximal earned reward with an assumed prior mean even lower than FBM and softmax, or DBM and KG, implying that even more prior reward rate underestimation is needed to compensate for the combination of softmax and DBM. We note that human participants do not assume a prior mean that optimizes the earning of reward (blue diamonds are far from the peak of the blue lines) – this may reflect a compromise between optimizing reward earned and truthfully representing environmental statistics.

Lastly, To understand how DBM behaves differently than FBM, and thus how a lower prior mean helps DBM, and what it implies about human psychological processes, we consider the empirical/simulated probability of the participants/model switching away from a "winning" arm after it suddenly produces a loss (Fig. 1D). Since DBM assumes reward rates can change any time, a string of wins followed by a loss indicates a high probability of the arm switching to a lower reward rate. On the other hand, since FBM assumes reward rates to be stable, it depends more on long-term statistics to estimate an arm's reward rate. Give observations of many wins, which leads to a relatively high reward rate estimate as well as a relatively low uncertainty, a single loss should still induce in FBM a high probability of sticking with the same arm. Fig. 1C shows that the simulated shift rate of the two models (probability of a model to shift away from the previously chosen arm) exactly follow the pattern of behavior described above, that DBM (blue) always has a higher shift rates than FBM (purple). The diamond markers shows fitted prior mean on the x-axis, and human shift rate on the y-axis. Human subjects' shift rates are closest to what DBM predicts, which is what we would already expect from the fact that overall DBM has already been found to fit human data the best.

To further examine the effect of a lower prior mean, we compare the shift rates of DBM and FBM under different prior mean with the optimal policy. The dotted horizontal black line is the shift rate of optimal policy, and the vertical black line is the true prior mean. The vertical blue and pruple are the prior mean that maximizes reward rate for DBM and FBM (the best prior mean). With the true prior, both DBM and FBM shift away from the "winning" arm much more than the optimal policy, and this is mitigated by a lower prior. The best prior mean for DBM almost exactly reproduces the shift rate of the optimal policy, while the optimal prior mean for FBM yields lower shift rate than the optimal policy.

Acknowledgments

This work was in part funded by a NSF CRCNS grant (BCS-1309346) to AJY.

References

Averbeck, B. B. (2015). Theory of choice in bandit, information sampling and foraging tasks. *PLoS computational biology*, *11*(3), e1004164.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neurosci*, *10*(9), 1214–21.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? how the human brain manages the tradeoff between exploitation and exploration. *Philos Trans R Soc Lond B Biol Sci*, *362*(1481), 933-42.

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), 876-9.

Frazier, P., & Yu, A. J. (2008). Sequential hypothesis testing under stochastic deadlines. *Advances in Neural Information Processing Systems*, *20*.

Harlé, K. M., Zhang, S., Schiff, M., Mackey, S., Paulus*, M. P., & Yu*, A. J. (2015). Altered statistical learning and decision-making in methamphetamine dependence: evidence from a two-armed bandit task. *Frontiers in Psychology*, *6*(1910). (*Co-senior authors) doi: 10.3389/fpsyg.2015.01910

Ide, J. S., Shenoy, P., Yu*, A. J., & Li*, C.-S. R. (2013). Bayesian prediction and evaluation in the anterior cingulate cortex. *Journal of Neuroscience*, *33*, 2039-2047. (*Co-senior authors)

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning ii: Current research and theory* (p. 64-99). Mew York: Appleton-Century-Crofts.

Shenoy, P., Rao, R., & Yu, A. J. (2010). A rational decision making framework for inhibitory control. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems 23* (Vol. 23, pp. 2146–54). Cambridge, MA: MIT Press.

Yu, A. J., & Cohen, J. D. (2009). Sequential effects: Superstition or rational behavior? *Advances in Neural Information Processing Systems*, *21*, 1873-80.

Yu, A. J., & Huang, H. (2014). Maximizing masquerading as matching: Statistical learning and decision-making in choice behavior. *Decision*, *1*(4), 275-287.

Zhang, S., & Yu, A. J. (2013). Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting. *Advances in Neural Information Processing Systems*, *26*.