Cheap but Clever: Human Active Learning in a Bandit Setting

Shunan Zhang **Angela J. Yu**
(s6zhang, ajyu@ucsd.edu)
Department of Cognitive Science, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093-0515

Abstract

How people achieve long-term goals in an imperfectly known environment, via repeated tries and noisy outcomes, is an important problem in cognitive science. There are two interrelated questions: how humans *represent information*, both what has been learned and what can still be learned, and how they *choose actions*, in particular how they negotiate the tension between exploration and exploitation. In this work, we examine human behavioral data in a multi-armed bandit setting, in which the subject choose one of four "arms" to pull on each trial and receives a binary outcome (win/lose). We implement both the Bayes-optimal policy, which maximizes the expected cumulative reward in this finite-horizon bandit environment, as well as a variety of heuristic policies that vary in their complexity of information representation and decision policy. We find that the *knowledge gradient algorithm*, which combines exact Bayesian learning with a decision policy that maximizes a combination of immediate reward gain and long-term knowledge gain, captures subjects' trial-by-trial choice best among all the models considered; it also provides the best approximation to the computationally intense optimal policy among all the heuristic policies.

Keywords: Bandit problems; human decision making; human active learning; knowledge gradient

Introduction

How humans achieve long-term goals in an imperfectly known environment, via repeated tries and noisy outcomes, is an important problem in cognitive science. The computational challenges consist of the learning component, whereby the observer updates his/her representation of knowledge and uncertainty based on continual observations, and the control component, whereby the observer chooses an action that somehow balances between the need to obtain immediate reward and to obtain information that assists long-term reward accumulation.

A classical task setting used to study sequential decision-making under uncertainty is the multi-armed bandit problem (Robbins, 1952). The bandit problems are a family of reinforcement-learning problems where the decision maker must choose among a set of arms on each trial: the reward gained on each trial both has intrinsic value and informs the decision maker about the relative desirability of the arms, which can help with future decisions. In the basic bandit setting, each arm has an unknown probability of generating a reward on each trial. The problem is called *finite horizon* if the total number of trials is finite; it is called *infinite horizon* if the number of trials is infinite, in which case one either discounts future rewards or tries to maximize average reward per unit time. In this work, we focus on stationary, non-discounted, finite-horizon bandit problems, where the underlying reward rates are independent and identically (iid) distributed across the arms.

Bandit problems elegantly capture the tension between exploration (selecting an arm about which one is ignorant) and exploitation (selecting an arm that is known to have relatively high expected reward), which is manifest in many real-world decision-making situations involving noise or uncertainty. Bandit problems have been well studied in many fields, including statistics (Gittins, 1979), reinforcement learning (Kaebling, Littman, & Moore, 1996; Sutton & Barto, 1998), economics (Banks, Olson, & Porter, 2013, e.g.), psychology and neuroscience (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Cohen, McClure, & Yu, 2007; Steyvers, Lee, & Wagenmakers, 2009; Lee, Zhang, Munro, & Steyvers, 2011). There is no analytical solution to the general bandit problem, though properties about the optimal solution of special cases are known (Gittins, 1979). For relatively simple, finite-horizon problems, the optimal solution can be computed numerically via dynamic programming (Kaebling et al., 1996), but its computational complexity grows exponentially with the number of arms and with the time horizon. In the psychology literature, a number of heuristic policies, with varying levels of complexity in the learning and control processes, have been proposed as possible strategies used by human subjects (Daw et al., 2006; Cohen et al., 2007; Steyvers et al., 2009; Lee et al., 2011). Most models assume that humans either adopt simplistic policies that retain little information about the past and sidestep long-term optimization (e.g. win-stay-lose-shift and ε-greedy), or switch between an exploration and exploitation mode either randomly (Daw et al., 2006) or discretely over time as more is learned about the environment (Steyvers et al., 2009).

Here, we analyze a new model for human bandit choice behavior, based on the *knowledge gradient* (KG) algorithm, which has been developed by Frazier, Powell, and Dayanik (2008) to solve problems in operations research. At each time step, the KG policy chooses, conditioned on previous observations, the option that maximizes future cumulative reward gain. It is based on the myopic assumption that the next observation is the last exploratory choice, used to learn about the environment, and all remaining choices will be exploitative, choosing the option with the highest expected reward by the end of the next trial. Note that this myopic assumption is only used in reducing the complexity of computing the predicted value of each option, and not actually implemented in practice – the algorithm may end up executing arbitrarily many non-exploitative choices. Despite a certain greedy aspect to the KG control policy, it is not completely short-sighted. In particular, it tends to explore more when the number of trials left is large, because finding an

arm with even a slightly better reward rate than the currently best known one can lead to a large cumulative advantage in future gain; on the other hand, when the number of trials left is small, KG tends to exploit and stay with the currently best known option, because it knows that finding a slightly better option will not lead to large improvement, while the risk of wasting time on a bad option is high. KG is also known to be exactly optimal in certain special cases (Frazier et al., 2008), such as when there are only two arms.

KG has several advantages over previously proposed algorithms. Unlike the simple heuristic algorithms such as win-stay-lose-shift and ε-greedy, and in common with the other Bayesian learning algorithms (Daw et al., 2006; Steyvers et al., 2009; Lee et al., 2011), KG uses a sophisticated Bayesian posterior distribution as its belief state at each time step. Unlike the other Bayesian learning algorithms, KG gracefully and gradually transitions from primarily exploring to primarily exploiting over the course of a finite-horizon bandit experiment. Also unlike previously proposed algorithms, which typically assumes that the stochastic component of action selection is random or arbitrary, KG also provides a more sophisticated and discriminating way to explore, by normatively combining immediate reward expectation and long-term knowledge gain. On the other hand, in contrast to the optimal algorithm, which scales exponentially in computational complexity with respect to the number of remaining timesteps, KG is computationally much simpler, incurring a constant cost regardless of the number of timesteps left.

In the following, we first describe the experiment, then describe all the learning and control models that we consider. We then compare the performance of the models both in terms of agreement with human behavior on a trial-to-trial basis, and in terms of computational optimality.

Data

Participants

A total of 451 participants completed a series of bandit problems as part of 'testweek' at the University of Amsterdam.

Experimental procedure

Each participant completed 20 bandit problems in sequence, all problems had 4 arms and 15 trials. The reward rates for all games were generated independently from a Beta$(2,2)$ distribution, and were all done prior to data collection. All participants thus played games with the same sets of reward rates, but the order of the games was randomized. Participants were aware that the reward rates in all games were drawn from the same environment, but they were not told its form, i.e. Beta$(2,2)$. A representation of the basic experimental interface is shown in Fig 1.

Modeling Methods

There exist multiple levels of complexity and optimality in both the learning and the decision components of decision making models of bandit problems. For the learning component, we examine whether people learn any abstract representation of the environment at all, and if they do, whether

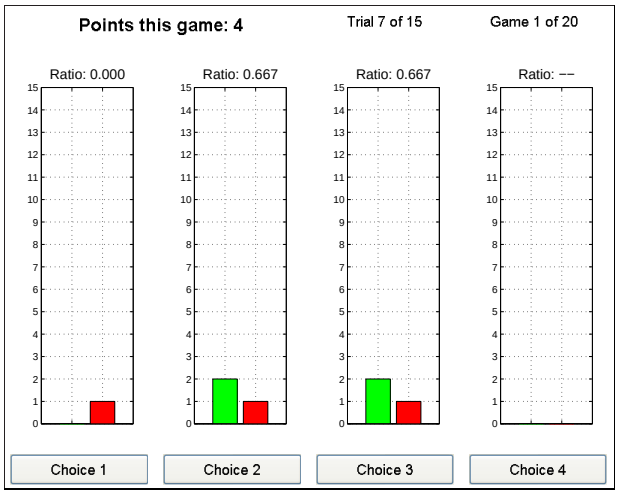

Figure 1: Experiment interface. The four panels correspond to the four arms, each of which can be chosen by pressing the corresponding button. In each panel, successes from previous trials are shown as green bars, and failures as red bars. At the top of each panel, the ratio of successes to failures, if defined, is shown. The top of the interface provides the count of the total number of successes to the current trial, index of the current trial and index of the current game.

they only keep a mean estimate (running average) of the reward rate of the different options, or also uncertainty about those estimates, or indeed more complex meta-information, such as the general abundance/scarcity of rewards. The decision component can also differ in complexity in at least two respects: the objective the decision policy tries to optimize (e.g. reward versus information), and the time-horizon over which the decision policy optimizes its objective (e.g. greedy versus long-term). In this section, we introduce models that encompass different combinations of learning and decision policies.

Bayesian Learning in Beta Environments

The observations are generated independently and identically (iid) from an unknown Bernoulli distribution for each arm. We consider two Bayesian learning scenarios, either subjects have a fixed belief about the distribution from which the Bernoulli rates are drawn ("basic learning"), or they do meta-learning about the parameters of that distribution over time ("meta learning"). We explore the two scenarios below. In either case, we assume the distribution that generates the Bernoulli rates is a Beta distribution, Beta(α, β), which is a conjugate prior, and whose two hyper-parameters, α and β are determined by the total number of rewards and failures experienced so far, plus any pseudo-counts associated with the prior.

Basic Learning Suppose we have K arms with reward rates, $\theta_k^g, k = 1, \cdots, K$, which are independent and identically drawn from Beta(α, β) for the gth game. On the tth trial, if the kth arm is chosen, a reward is attained with a

Bernoulli distribution, $R_k^{t,g} \sim \text{Bernoulli}\left(\theta_k^g\right)$. Let $\mathbf{S}^{t,g}$ and $\mathbf{F}^{t,g}$ be vectors of the number of successes and failures attained from each arm at the tth trial of the gth game. The model learns the individual reward rates using Bayes' Rule:

$$\begin{aligned}\Pr\left(\theta^g \,|\, \alpha, \beta, \mathbf{S}^{t,g}, \mathbf{F}^{t,g}\right) &\propto \Pr\left(\mathbf{S}^{t,g}, \mathbf{F}^{t,g} \,|\, \theta^g\right) \Pr\left(\theta^g \,|\, \alpha, \beta\right) \\ \theta^g &\sim \text{Beta}\left(\alpha, \beta\right) \\ S_k^{t,g} &\sim \text{Binomial}\left(S_k^{t,g} + F_k^{t,g}, \theta_k^g\right)\end{aligned}$$

The learner's belief state at the trial t of the game g, $\mathrm{B}^{t,g}$, is the set of posterior Beta distributions for each arm, and the mean reward on each arm, based on the observed sequence, is $\hat{\theta}^{t,g} = (\alpha + S_k^{t,g})/(\alpha + \beta + S_k^{t,g} + F_K^{t,g})$.

Meta Learning We also consider the case that subjects may use observations to learn about the true environmental reward distribution (the true Beta distribution), corresponding to the general abundance/scarcity of resources in the environment. In this case, observing an outcome on any arm will affect the posterior distribution on all arms because of the correlation induced by shared hyper-parameters of the environment (Gelman, Carlin, Stern, & Rubin, 2004):

$$\Pr\left(\theta^g, \alpha, \beta | \mathbf{S}^{t,g}, \mathbf{F}^{t,g}\right) \propto \Pr\left(\mathbf{S}^{t,g}, \mathbf{F}^{t,g} \,|\, \theta^g\right) \Pr\left(\theta^g \,|\, \alpha, \beta\right) \Pr(\alpha, \beta)$$

The belief state on trial t of game g, $\mathrm{B}^{t,g}$, is a joint posterior distribution over the reward rates and environmental parameters, conditioned on the observed sequence.

Decision Policies

We consider five different decision policies. We first describe the optimal model, and then the four heuristic models with increasing levels of complexity.

The Optimal Model The learning and decision problem for bandit problems can be instantiated as a Markov Decision Process with a finite horizon (Kaebling et al., 1996). Due to the low dimensionality of the bandit problem here (i.e. small number of arms and number of trials per game), the optimal policy, up to a discretization of the belief state, can be computed numerically according to Bellman's dynamic programming principle. Let $V^t(\mathbf{S}^t, \mathbf{F}^t)$ be the expected total future reward on trial t. The optimal policy should satisfy the following iterative property:

$$V^{t,g}(\mathbf{S}^{t,g}, \mathbf{F}^{t,g}) = \max_k \mathbb{E}\left[V^{t+1,g}(\mathbf{S}^{t+1,g}, \mathbf{F}^{t+1,g})\right] + \hat{\theta}_k^{t,g}$$

and the optimal decision, $D^{t,g}$, is decided by

$$D^{t,g}(\mathbf{S}^{t,g}, \mathbf{F}^{t,g}) = \text{argmax}_k \mathbb{E}\left[V^{t+1,g}(\mathbf{S}^{t+1,g}, \mathbf{F}^{t+1,g})\right] + \hat{\theta}_k^{t,g}$$

We solve the equation using dynamically programming, backward in time from the last time step, whose value function and optimal policy are known for any belief state, i.e. any setting of posterior Beta distribution for each of the arms: it always choose the arm with the highest expected reward, $\hat{\theta}^{T,g}$, and the value function is just that expected reward. In the simulations, we compute the optimal policy offline, for any conceivable setting of belief state on each trial (up to a fine discretization of the belief state space), and then apply the computed policy for each sequence of choice and observations that each subject experiences. We use the term "the optimal solution" to refer to the specific solution under $\alpha = 2$ and $\beta = 2$, which is the true experimental design.

Win-Stay-Lose-Shift WSLS does not learn any abstract representation of the environment, and has a very simple decision policy. It assumes that the decision-maker continues to choose an arm following a reward, but shifts to other arms (with equal probabilities) following a failure to gain reward.

ε-Greedy The ε-greedy model assumes that decision-making is driven by a parameter ε that controls the balance between random exploration and exploitation inherent in bandit problems. On each trial, with probability ε, the decision-maker chooses randomly (exploration), otherwise chooses the arm with the greatest estimated reward rate (exploitation). ε-Greedy keeps simple estimates of the reward rates, but does not track the uncertainty of the estimates. It is not sensitive to the horizon, maximizing the immediate gain with a constant rate, otherwise searching for information by random selection[1].

We call the situation $k \in \text{argmax}_{k'} \hat{\theta}_k^{t,g}$ 'case 1', and the ε-greedy model is implemented as

$$\Pr\left(D^{t,g} = k \,|\, \varepsilon, \hat{\theta}^{t,g}\right) = \begin{cases} (1-\varepsilon)/M^{t,g} & \text{if case 1} \\ \varepsilon/(K - M^{t,g}) & \text{otherwise} \end{cases}$$

where $M^{t,g}$ is the number of arms with the greatest estimated value at the tth trial of the gth game.

ε-Infomax The ε-infomax model is similar to the ε-greedy model in that it chooses the arm with the greatest estimated reward rate with probability $1-\varepsilon$, and explores with probability ε. The difference is that, instead of random selection for exploration, it selects the arm that results in the largest reduction in the expected total entropy. In our study, the arms are independent given the same environmental distribution, and the policy reduces to choose the arm with the largest uncertainty. We use $S_k^{t,g} + F_k^{t,g}$ as an approximate, simple measure of the uncertainty associated with arm k given the state of the game. In this model, an arm may be chosen when one of the two cases applies: in case 1, it has the greatest estimated reward rate; in case 2, it does not have the greatest estimated reward rate, but has the least number of times being chosen. We implement ε-infomax as

$$\Pr\left(D^{t,g} = k \,|\, \varepsilon, \hat{\theta}^{t,g}\right) = \begin{cases} (1-\varepsilon)/M^{t,g} & \text{if case 1} \\ \varepsilon/N^{t,g} & \text{if case 2} \\ 0 & \text{otherwise} \end{cases}$$

where $M^{t,g}$ and $N^{t,g}$ are the number of arms that satisfy case 1 and 2, respectively, at the tth trial of the gth game.

The ε-infomax model uses both the mean estimates and measure of uncertainty as criteria for action selection. It is a

[1]The ε-Greedy model has a variant, ε-decreasing, where the probability of random selection decreases over trials. However, previous studies found that ε-decreasing had a poor fit to the same data when compared with the ε-greedy model (Zhang & Lee, 2010), so we only consider the ε-greedy model in this study.

greedy heuristic, maximizing the immediate reward gain at a constant rate.

Knowledge Gradient The knowledge gradient (KG) algorithm (Ryzhov, Powell, & Frazier, 2012) is an approximation to the optimal policy, by pretending only one more exploratory measurement is allowed, and assuming all remaining choices will exploit what is known after the next measurement. It evaluates the expected change in each estimated reward rate, if a certain arm were to be chosen, based on the current belief state. Its mathematical expression is

$$v_k^{\mathrm{KG},t} = \mathbb{E}\left[\max_{k'} \hat{\theta}_{k'}^{t+1} \,|\, D^t = k, B^t\right] - \max_{k'} \hat{\theta}_{k'}^t$$

The first term is the expected largest reward rate on the next step if the kth arm were to be chosen, with the expectation taken over all possible outcomes of choosing k. The KG decision rule is

$$D^{\mathrm{KG},t,g} = \arg\max_k \hat{\theta}_k^{t,g} + (T-t-1)\, v_k^{\mathrm{KG},t,g} \qquad (1)$$

The first term of Equation 1 denotes the expected immediate reward by choosing the kth arm at t of the gth game, whereas the second term reflects the expected gain of total remaining reward from $t+1$ to the last trial of the current game. The formula for calculating $v_k^{\mathrm{KG},t,g}$ for the binary bandit problems can be found in Chapter 5 of Powell and Ryzhov (2012).

Model Implementation and Agreement Calculation

We used model *agreement* as a measure of how well it captures experimental data, which was calculated as the average per-trial likelihood, conditioned on the observed game states. We fit the models and calculated model agreement across all participants.

WSLS is a fully deterministic paradigm, so the per-trial likelihood is 1 for a win-stay decision, $1/3$ for a lose-shift decision, and 0 otherwise. All other models have at least two free parameters, α and β, and the ε-greedy and the ε-infomax models have one additional parameter, ε. We implemented the KG, ε-greedy and ε-infomax models as Bayesian graphical models under both learning frameworks. We used a vague prior for the environmental parameters, $\Pr(\alpha,\beta) = (\alpha+\beta)^{5/2}$, as suggested by Gelman et al. (2004), because it is uniform on the psychologically interpretable reparameterization, $\alpha/(\alpha+\beta)$ and $(\alpha+\beta)^{-1/2}$. We used uniform prior for ε. Model inference used combined sampling algorithm, with Gibbs sampling of ε, and Metropolis sampling of α and β. All chains contained 3000 steps, with a burn-in size of 1000. All chains converged according the R-hat measure (Gelman et al., 2004). We calculated the model agreement as the proportion of same choices between the model and the data, based on the full posterior predictive distribution of choices given each observed state of the game. For this study, we implemented the optimal model only with basic learning because of the heavy computational load.

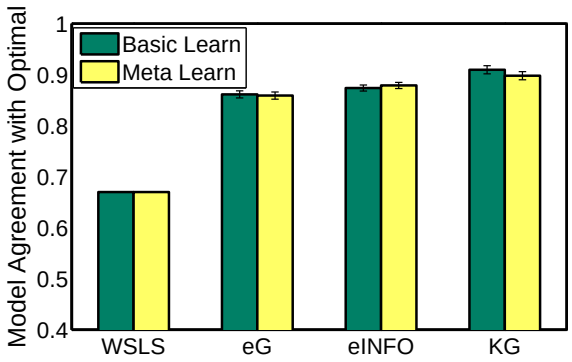

Figure 2: Model agreement with data simulated by the optimal solution under the correct prior of the environment. Each bar shows the agreement of a model combining the corresponding decision policy and the learning framework. For the ε-greedy (eG), ε-infomax (eINFO) and the KG models, the error bars show the standard errors of the average agreement based on a 4-fold cross-validation. WSLS has no parameters to fit and does not rely on any learning framework.

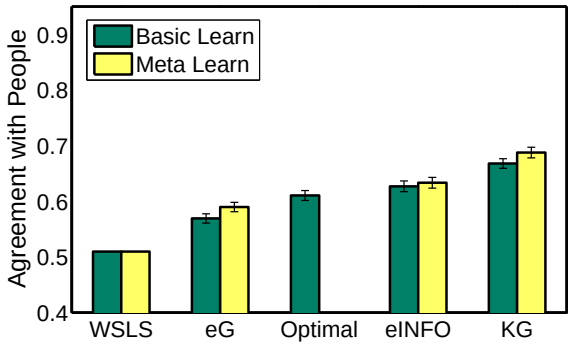

Figure 3: Model agreement with human data. The figure is generated in the same way as for Figure 2, except for that the optimal model was only implemented with basic learning for this study.

Results

Model Agreement with the Optimal Solution

As shown in Figure 2, the KG algorithm, under either learning framework, is able to approximate the optimal solution well in terms of the average number of correct predictions. In this sense, the KG policy is 'process optimal'. ε-infomax outperforms the ε-greedy model, which implies that smarter exploration for information gain increases the optimality of the heuristic. The simple WSLS model achieves model agreement well above 60%. In fact, both WSLS and the optimal model do win-stay with probability 1. The only situation that WSLS does not resemble the optimal behavior is when it shifts away from an arm that the optimal solution would otherwise stay with.

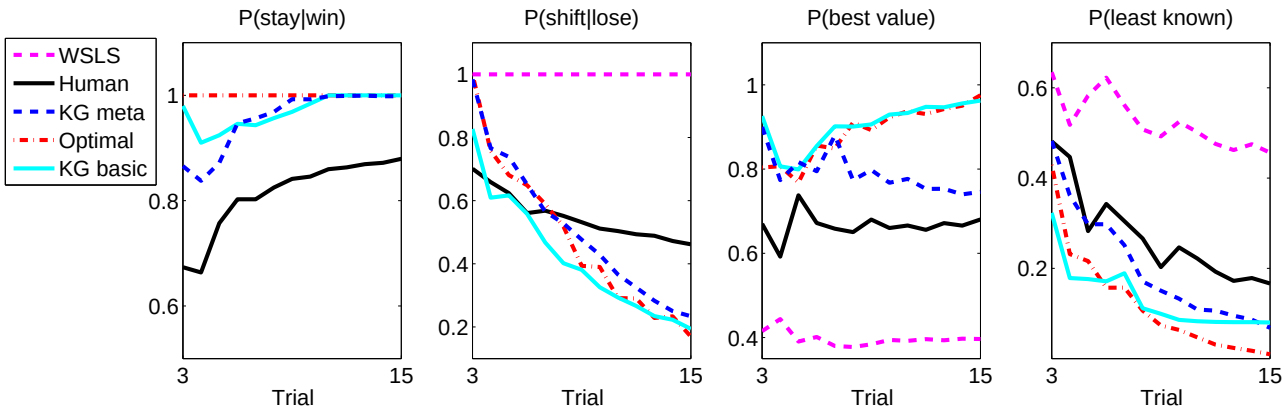

Figure 4: Behavioral patterns in the human data and the simulated data from a selection of the best- and worst-performing models. The four panels show the trial-wise probability of win-stay, lose-shift, choosing the greatest estimated value, and choosing the least known when it is an exploration trial, respectively. Probabilities are calculated based on simulated data from each model at their MAP estimate, and are averaged across all games and all participants. The optimal solution shown here uses the correct prior Beta $(2,2)$.

Model Agreement with the Human Data

Figure 3 shows the average model agreement with human data. Overall, the type of decision policy, other than the learning framework, makes significant differences in the model agreement. However, a decision policy tends to do better under the meta learning framework — the ε-greedy model and the KG model have significantly greater model agreement with meta learning.

We next break down the overall behavioral performance into four finer measures: how often people adopt win-stay and lose-shift, how often they exploit, and whether they use random selection or search for the greatest amount of information during exploration. We compare three of our models that have the highest agreement with human data on these additional behavioral criteria. Figure 4 shows the model analysis results. We show the patterns of the human subjects, the optimal solution, the best performing decision policy (KG) under both learning frameworks, and the simplest WSLS.

The first panel probably contains the most interesting results. It shows the trialwise probability of staying with the same arm following a previous success. People show clear sub-optimality by not staying with the same arm after an immediate reward. In fact, obtaining a reward from any arm should always increase the estimated value of the chosen arm. Under the basic learning framework where unchosen arms do not change in value, this means the optimal decision process should always do win-stay. This is consistent with the curve of the optimal solution. As implied by Equation 1, KG considers the likelihood of an arm surpassing the known best value upon chosen, and weights this knowledge gain more heavily in the early stage of the game. In general, during the early trials, it chooses the second-best arm with a certain probability, not necessarily depending on the previous outcome. This explains the drop of the win-stay probability of KG during the early trials. When the learner is also updating its knowledge of the environment, a previous success will cause the environment to appear more rewarding, making other arms more likely to surpass the current best arm.

The second panel shows the trialwise probability of shifting away following a previous failure. People, the optimal solution, and KG show a decline in this probability over trial. When the horizon is approaching, it becomes increasingly important to stay with the arm that is known to be reasonably good, even if it may occasionally yield in a failure, because it is increasingly important to maximize the reward on the current trial.

In general, the KG model with meta learning matches the second-order trend of human data. However, there still exists a big difference on the absolute scale, especially regarding the probability of staying with 'good' arms — in fact, the KG policy does win-stay and exploitation more often, and resembles the optimal solution more than the human data.

Model Performance in Cumulative Reward Collection

Fig 5 shows a comparison of the distribution of average reward per trial achieved by the participants, the optimal solution, and the knowledge gradient model. When playing at their best fit parameterization based on the human data, KG with meta learning and WSLS achieve nearly identical reward distributions as the participants. Moreover, if we let KG with meta learning forward play under the correct prior knowledge of the environment, i.e. $Beta\,(2,2)$, it is able to achieve a nearly identical distribution as the optimal solution.

Discussion

Our analysis supports the KG decision policy under the meta learning framework as a good fit to human data in bandit problems. Our result implies that people might learn the individual reward rates as well as the general environment,

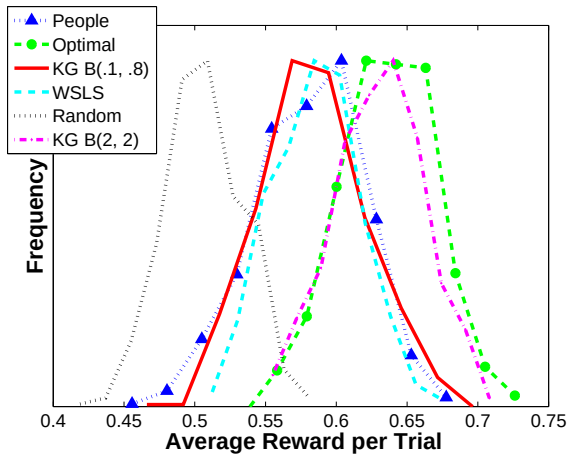

Figure 5: Average reward achieved by the KG model forward playing the bandit problems with the same reward rates. KG achieves similar reward distribution as the human performance, with KG playing at its maximum a posteriori probability (MAP) estimate, $\alpha = .1$ and $\beta = .8$. KG achieves the same reward distribution as the optimal solution when playing with the correct prior knowledge of the environment.

and the shared, latent environment induces a special type of correlation among the bandit arms. The meta learning framework is a psychologically sensible improvement to basic learning, because correct knowledge of the environment can be critical for achieving the best performance, especially when the environment can change over time or contexts. For the decision component, our results support the KG policy, which optimizes the semi-myopic goal of maximizing future cumulative reward while assuming only one more time step of exploration and strict exploitation thereafter (but does not actually ever carry out that policy). The KG model under the more general learning framework has the largest proportion of correct predictions of human data, and can capture the trial-wise dynamics of human behavioral reasonably well. KG achieves similar behavioral patterns as the optimal model, and is computationally tractable, making it a plausible algorithm for human learning and decision-making

One remaining puzzle why human subjects tend to explore more often than policies that optimize the specific utility of the bandit problems. One possibility is that people believe the task environment can undergo stochastic changesand exhibit sequential effects due to recent trial history, as in many other psychological task contexts (Yu & Cohen, 2009) . This would be an interesting line of future inquiry.

Acknowledgments

We thank Mark Steyvers and Eric-Jan Wagenmakers for sharing the data on bandit problems for our analysis, and three anonymous reviewers for their valuable comments that helped improve the paper.

References

Banks, J., Olson, M., & Porter, D. (2013). An experimental analysis of the bandit problem. *Economic Theory*, *10*, 55–77.

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.

Frazier, P., Powell, W., & Dayanik, S. (2008). A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, *47*, 2410-2439.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2 ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society*, *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., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, *12*, 164-174.

Powell, W., & Ryzhov, I. (2012). *Optimal learning* (1 ed.). Wiley.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*, 527-535.

Ryzhov, I., Powell, W., & Frazier, P. (2012). The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, *60*, 180-195.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*, 168-179.

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

Yu, A. J., & Cohen, J. D. (2009). Sequential effects: Superstition or rational behavior? In *Advances in neural information processing systems* (Vol. 21, p. 1873-1880). Cambridge, MA.: MIT Press.

Zhang, S., & Lee, M. D. (2010). Cognitive models and the wisdom of crowds. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 32th annual conference of the cognitive science society*. Austin, TX.