Decision making in 2-arm bandit problem

Hannah Chen hechen@eng.ucsd.edu
Sonali Rahagude srahagud@ucsd.edu
Sneha Venkatesh Yelimeli svyelime@eng.ucsd.edu
Chun Fan c9fan@ucsd.edu
Yashodhan Karandikar ykarandi@eng.ucsd.edu

1 Background

The stochastic multi-armed bandit problem is an important model for studying the exploration-exploitation trade-off in reinforcement learning. The goal of the decision-maker here is to obtain the maximum number of rewards over a number of trials. This report presents an empirical study of the popular 2-armed bandit problem. Our work is motivated from [1]. The paper presents a variety of existing heuristic methods for a finite horizon 2-arm bandit problem. It then builds on the idea of latent full state modeling and presents a new, simplified model that captures the results of the latent full state model. It uses a Bayesian model where the way people balance exploration and exploitation depends on his or her assumptions about the distribution of reward rates.

2 Goals

The goals of our experimental setup are 3-fold: generate optimal data for the bandit problem given different environment settings, fit parameter values of different heuristic models to optimal model to get best heuristic decision sequences, and the comparison of these with the optimal and human data.

3 Implementation

3.1 Experiment Setup

Our dataset [6] consists of 50 experiments on 3 environments: 'plentiful'($\alpha = 4, \beta = 2$) , 'neutral'($\alpha = 1, \beta = 1$) and 'scarce'($\alpha = 2, \beta = 4$). α and β define the count of prior successes and failures, respectively. Each experiment consists of 8 trials and the reward rates μ_k for arm k is drawn from the beta distribution, i.e. $\mu_k \sim Beta(\alpha, \beta)$. Thus, we draw a total of 100 reward rates in this setting (50 for each arm). For the i^{th} trial in the t^{th} experiment, if arm k is chosen, the reward R_i^t is determined by the outcome of a Bernoulli using the reward rate μ_k^t, so $R_i^t \sim Bernoulli(\mu_k^t)$.

3.2 Optimal Model

For a small number of trials and a fixed environment, the optimal solution can be obtained by following a dynamic programming approach as given in [5],[4]. The state of the tth trial can be defined as $S_t =< s_1, s_2, f_1, f_2 >$, where s_1 & s_2 are the number of successes for arm 1 & 2, respectively, and f_1 & f_2 are the number of failures for arm 1 & 2, respectively, up to trial t. For a given environment α, β, we can recursively define the value function as the expected future reward for state S_t plus the

current reward at state S_t as follows :

$$V_t(s_1, s_2, f_1, f_2) = \max_{i \in \{1,2\}} \Big(\frac{s_i + \alpha}{s_i + f_i + \alpha + \beta} V_{t+1}(..., s_i + 1, ...) + \frac{f_i + \beta}{s_i + f_i + \alpha + \beta} V_{t+1}(..., f_i + 1, ...) + \frac{s_i + \alpha}{s_i + f_i + \alpha + \beta} \Big)$$

The state value function for the last trial $V_T = \max_{i \in \{1,2\}}(\frac{s_i+\alpha}{s_i+f_i+\alpha+\beta})$ completes the recursive definition. Using this definition, the value function is computed for all enumerable states from $t = 7 \ldots 1$. To obtain the optimal decision, the argument that maximizes the state value function is selected, thus giving the best arm to be chosen for each possible state.

To obtain a set of decisions for each trial, we sample the reward rate μ_i for each arm i from the known Beta distribution of that environment. The optimal decision i for each trial combined with the reward sampled from the Bernoulli distribution given by μ_i gives us the next state.

3.3 Heuristic Models

We implement the five heuristics used in [2] and summarize the key parameters and their meanings in the following table:

Heuristic	**Key Parameter**	**Meaning**
ϵ-greedy	ϵ	probability of exploration
ϵ-decreasing	ϵ_0	probability of exploration in 1st trial
WSLS	γ	probability of staying after winning and shifting after losing
Full latent state	θ	multifaceted, see [1]
τ-switch	τ	trial # for switching from exploration to exploitation

3.4 Heuristic Model Comparisons

We use grid search to find the key parameter or the environment parameters for each heuristic. To do this, we define an L-value similar to a likelihood function of a Bernoulli function. This comparison differs from the original paper which infers parameters by calculating a posterior predictive average agreement via sampling. For the optimal data, we perform grid search over the key parameter and the heuristic decision sequences that yield the highest L-value; while for the human data, we assume that the key parameter value is identified from the optimal data, and grid search over the α', β' parameters and the heuristic decision sequences that yield the highest L-value. Without loss of generality, the following approach is explained based on the ϵ-greedy heuristic.

We derive our method for model fitting to optimal and human data based on three key factors: (1) there are a finite number of possible decision sequences; (2) we constraint the possible combinations of environment variables, α' and β'; and (3) an assumption of conditional independence of the underlying arm's reward distribution and the action sequence. Each decision maker has a continuously updating belief of the environment variabls, α', β'. In the case of the optimal decision maker, α', β' match the true environment.

We define the L-value as :

$$\begin{aligned} L(Action, Reward|Data) &= P(Action, Reward|Data) \\ &= \prod_t P(f(\epsilon), f(\mu_k)|Data) \\ &= \prod_t P(f(\epsilon)|Data) P(f(\mu_k)|Data) \\ &= \prod_t \epsilon^x (1-\epsilon)^{1-x} \mu_k^y (1-\mu_k)^{1-y} \end{aligned} \tag{1}$$

where data corresponds to the observed decision and reward sequence of a decison maker encoded in the binary value for x and y, respectively. The random variables for action and reward are functions depending on the heuristic parameter, ϵ, and arm k's reward probability, μ_k. Additionally the actions are based on rules defined by each heuristic. Here we make the assumption that the probabiliy of the random variables Action and Reward given data are independent. This assumption stems from our intuition that given data for a decision sequence, the probabilities for ϵ and μ are independent.

3.4.1 Fitting Heuristic Decision to Optimal Data

We assume the optimal decision-maker has perfect knowledge of the environment, i.e. the α and β values are known, and thus the arm's reward probability. This makes sense as we want to characterize a heuristic's decision making process and see how well it compares to the best possible decision. To find the best fit for each of the heuristic methods, we perform grid search for its key parameter value from all enumerated sequences. For example, the best ϵ parameter is found by averaging the results of the grid search which resulted in the maximum L-value. Finally that particular heuristic decision sequence is compared to the optimal decision sequence using match percentage for assessment. A comparison of these decisions with the optimal decisions gives us an idea of each heuristic's performance.

3.4.2 Fitting Heuristic Decision to Human Data

The assumption of conditional independence of ϵ and μ may not hold true given human data, but if a heuristic is indeed capturing the best decision sequence, a human's expectation of μ would match the arm's true reward probability.

Using this, we grid search over three factors that yield the maximum L-value: a sequence of decisions, and the human's belief of the two environment variables. There are a finite number of decision sequences, specifically 2^8 for an 8-trial experiment, and we constrain our environment with $\alpha' + \beta' = 10$ to simplify the search process. Finally, we fit the human data to the optimal heuristic parameter and sequence to determine a measure of agreement averaged over all games.

3.5 Potential Design Issues

Originally we did grid search based on decision match percentage compared to the optimal model decision. However, averaging our parameter value for multiple iterations yielded results with wide variation. One reason match percentage gave inconclusive results is because many different parameter values could result in 100% match. If the best match results in just one decision being off (e.g. a match percentage of 87.5%), there could eight different decision sequences that differ from the optimal sequence by one. This mean there could be *at least* eight different ϵ parameters that result in that match percentage. We could improve our method by tracking all the different decision sequences that could result from a particular match percentage and the corresponding ϵ. Then each identified parameter choice would have to be simulated under the heuristic and compared. Regardless of the resulting increase in complexity, we would still need to define a method to compare all these parameter choices, and it is not clear how to define a way to differentiate the parameters for identifying an optimal value.

Thus, we formulated the idea for L-value. While we ultimately report tentative results using the method of L-value for comparing grid search parameters, we recognize it is not fully valid for model fitting to optimal data because we cannot make the assumption of conditional independence of ϵ and μ. For human data, μ' is an expectation of the arm's reward probability and that belief is continuously updated over the course of the trial; nowhere is it used to determine the reward because the data is already observed.

However, the optimal data consists of all enumerable states and the optimal decision sequence to take for each state. The reward data is drawn from the arm's reward probability so for an optimal model that data is probabilistic depending on the true μ. In the optimal model, μ' is also a continuously updated expectation of the true reward probability. Even though every state can be enumerated, it is not clear how to choose which state to transition to without drawing the outcome of an action from the true μ, so there are many L-values that could be maximal depending on the state.

4 Results

4.1 Comparison with optimal data

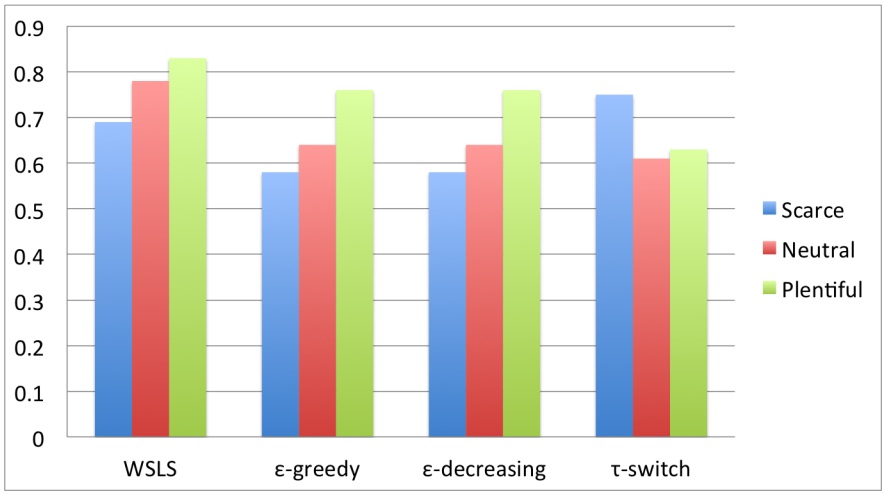

Figure 1: *Agreement of Heuristic Models with Optimal Model*

The grid search over parameter values of the heuristic models yields the best results at extreme values of the parameters which is hard to interpret. Therefore, we choose the parameter value given in [2].Figure 1 shows the agreement of each of the heuristic models with the decisions obtained using the optimal model, for the three environment settings. The agreement with optimal is maximum for the 'plentiful' environment for most of the heuristics, which agrees with the results in [2].

4.2 Comparison with human data

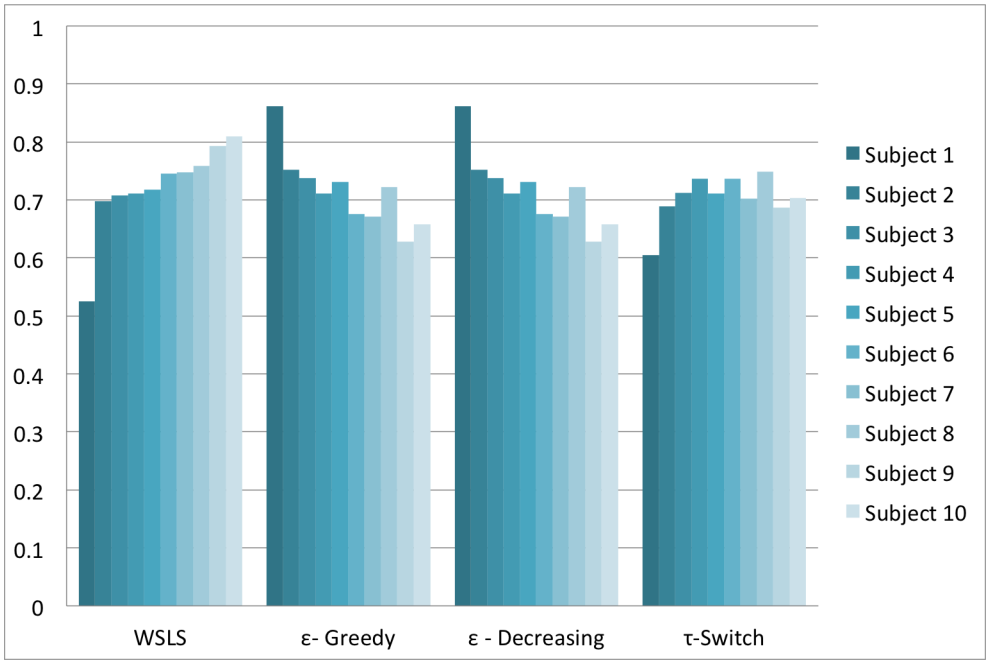

Figure 2: *Agreement of Heuristic Models with Human Data*

Figure 2 shows the agreement of the heuristics with the human data for each of the 10 subjects. The 10 bars for the WSLS have been sorted in increasing order of agreement, while the bars for the remaining heuristics follow this sorted order of the participants.

5 Conclusions

The project consisted of 3 goals: generating optimal data for different environments, and evaluating agreement of different heuristic models and the $\tau - switch$ model with the optimal data as well as the given human data. The first step was to understand how different environments can be obtained by different parameter settings of the beta distribution $Beta(\alpha, \beta)$ and then implementing sampling rewards for the optimal model. The decisions and rewards obtained from the optimal model make intuitive sense, depending on the prior confidence in the environment ($\alpha + \beta$), and expected rate of the rewards ($\alpha/(\alpha + \beta)$). We found that our implementation does agree with the results of the paper [2] to some extent. The fact that we use a different comparison metric (match percentage) instead of posterior predictive agreement used in [2] might be the reason for the differences.

References

[1] *Human and Optimal Exploration and Exploitation in Bandit Problems*, Shunan Zhang, Michael D Lee, Miles Munro, 2009.

[2] *Psychological models of human and optimal performance in bandit problems*, Michael D Lee, Shunan Zhang, Miles Munro, Mark Steyvers, 2011.

[3] *Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting* Shunan Zhang, Angela Yu 2013.

[4] *A Bayesian Analysis of Human Decision-Making on Bandit Problems*, Mark Steyvers, Michael D Lee, Eric-Jan Wagenmakers 2008.

[5] *Reinforcement learning: a survey*, Leslie Pack Kaelbling, Michael L. Littman, Andrew W. Moore, Journal of Artificial Intelligence Research, 1996.

[6] *Dataset: Finite-horizon two-alternative bandit problems*, Michael D Lee, Shunan Zhang, Miles Munro, Mark Steyvers, 2011, `http://www.cmr.osu.edu/browse/datasets?pid=63&sid=90:psychological-models-of-human-and-optimal-performance-on-bandit-problems`.