Forgetful Bayes and myopic planning: Human decision making in a bandit setting

Shunan Zhang, Angela J. Yu, {s6zhang, ajyu}@ucsd.edu

UC San Diego

Introduction

How do people achieve long-term goals in an uncertain environment, via repeated trials and noisy observations?

Multi-armed bandit problem

- limited number of trials
- reward generated by a stationary, hidden Bernoulli
- the goal is to maximize total number of rewards

Balance between

- **exploration** selecting an arm to gain more information
- **exploitation** selecting an arm known to have relatively high expected reward

Our model

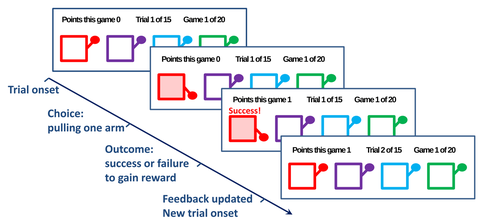
- learning component: **Dynamic Belief Model** [1]
  - Bayesian iterative inference assuming local patterns in a sequence of observations
- control component: **Knowledge Gradient** algorithm [2]
  - myopic approximation to the optimal policy
  - assuming next step being the last exploratory choice
  - chooses option that maximizes the future cumulative reward

Human Data [3]

Behavioral experiment

- 451 subjects, 4 arms, 15 trials, 20 games
- reward rates fixed, iid from Beta(2, 2)
- participants played *same* set of games, order randomized
- participants were instructed the reward rates were drawn from the same environment, and fixed

Schematic experimental design



Bayesian Inference Model

For arm k, denote the outcome by R_k and reward rate by θ_k

$$R_k^t \overset{iid}{\sim} \text{Bernoulli}\left(\theta_k\right)$$

$$\theta_k \overset{iid}{\sim} \text{Beta}\left(\alpha_0, \beta_0\right)$$

The **belief state**, i.e. the posterior distribution of θ_k given observations, is

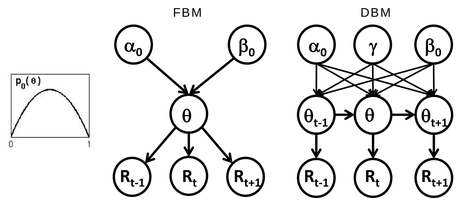
$$q_k^t\left(\theta_k^t\right) := \Pr\left(\theta_k^t | S_k^t, F_k^t\right)$$

where S_k^t and F_k^t are the cumulative numbers of successes and failures from arm k at t

The belief state can be calculated via Bayes' rule

$$q_k^t\left(\theta_k^t\right) \sim \Pr\left(R_k^t | \theta_k^t\right) \Pr\left(\theta_k^t | S_k^{t-1}, F_k^{t-1}\right)$$

Temporal Dynamics



Fixed Belief Model (FBM)

- assuming all reward rates fixed throughout the game
- $\Pr\left(\theta_k^t = \theta | S_k^{t-1}, F_k^{t-1}\right) = q_k^{t-1}$
  i.e. $\text{Beta}\left(\alpha_0 + S_k^{t-1}, \beta_0 + F_k^{t-1}\right)$

Dynamic Belief Model (DBM)

- assuming reward rates can undergo discrete changes according to a hidden Markov model
- updated belief is a weighted sum of last trial's belief and the generic prior

$$\Pr\left(\theta_k^t = \theta | S_k^{t-1}, F_k^{t-1}\right) = \gamma q_k^{t-1}(\theta) + (1-\gamma) p_0(\theta)$$

where $p_0 := \text{Beta}\left(\alpha_0, \beta_0\right)$

Decision Policies

The Optimal Algorithm

- computed via Bellman's dynamic programming principle

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

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

- optimal policy computed off-line and then applied to data

Knowledge Gradient (KG)

- expected increase in the best value among all arms if k were to be chosen, assuming next step is the last exploration
  - expectation taking over all possible outcomes incurred by pulling k

$$v_k^{\text{KG},\,t} = \mathbb{E}\left[\max_{k'} \hat{\theta}_{k'}^{t+1} \mid D^t = k,\, \mathbf{q}^t\right] - \max_{k'} \hat{\theta}_{k'}^t$$

- KG decision rule

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

ϵ-Greedy

- with probability ϵ, chooses randomly (exploration), otherwise chooses the greatest estimated value (exploitation)

$$\Pr\left(D^t = k \,|\, \epsilon,\, \hat{\theta}^t\right) = \begin{cases} (1-\epsilon)/M^t & \text{if } k \in \text{argmax}_{k'} \hat{\theta}_{k'}^t \\ \epsilon/(K - M^t) & \text{otherwise} \end{cases}$$

M^t: number of ties in greatest estimated value at t

- maximizes the immediate gain with a constant rate, otherwise searching for information by random selection

Win-stay, Lose-shift

- chooses the same arm if it continues to produce a reward
- shifts to other arms (equal probabilities) following a failure

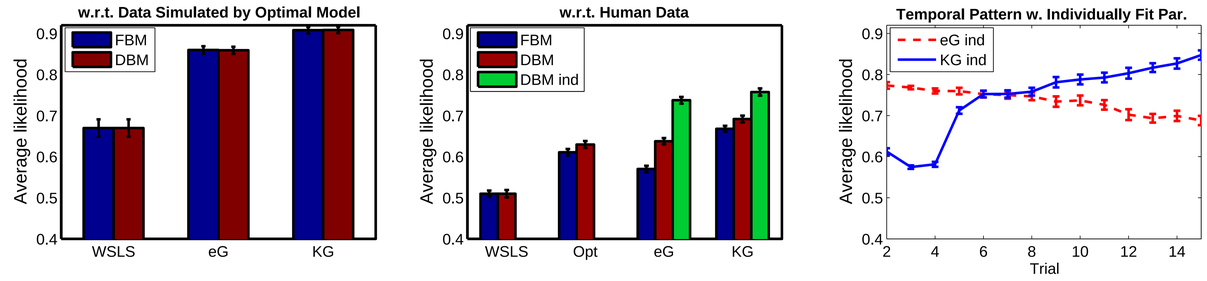
Model Inference and Evaluation

MCMC to infer a posterior distribution of all the model parameters: Gibbs for ϵ, Metropolis for the others

The **Average likelihood**: likelihood of consistent choices (between model and data, normalized by the set size of all consistent choices) of choices given each observed state of the game, based on MAP estimates of parameters

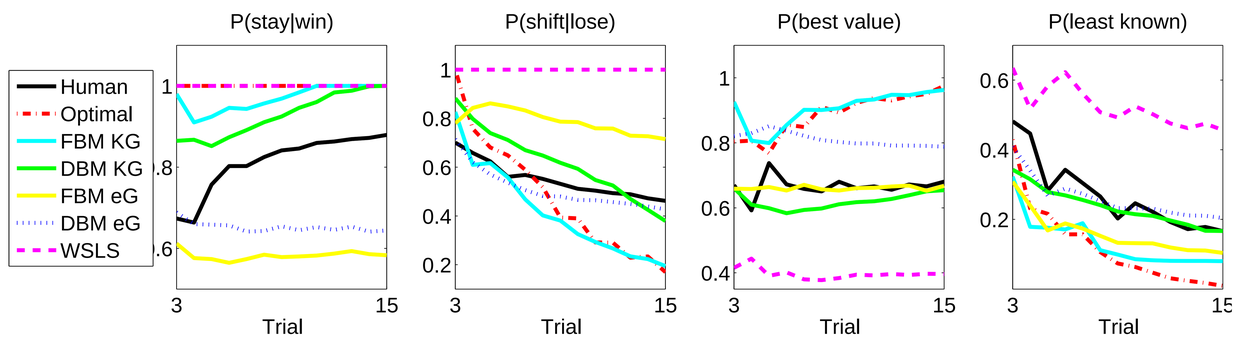
Leave-one(game)-out cross validation to calculate the average likelihood

Model Comparison of Trialwise Choices: vs Optimal Model and Human Data



- Fitted average γ is .81 – people believe the world has changed in every 5 steps
- KG has the best fit to data simulated by the optimal algorithm and human data
- DBM outperforms FBM for human data for all policies
- Models may differ in their ability of capturing early/late trials

Trialwise Behavioral Pattern from Model Simulations



- Compared to the optimal, human had more exploratory decisions, e.g. smaller probability of win-stay and choosing the best value
- Proportion of exploratory decisions decreases over trial
- DBM with KG captures the trial-wise patterns best among all models

Summary

Human subjects tend to explore more often than policies that optimize the specific utility of the bandit problems

- DBM attributes it to the belief of a stochastically changing environment causing the sequential effects due to recent trial history
- KG decision rule favors uncertain options during early trials (to increase information gain)

References

[1] Yu, A. J. & Cohen, J. D. (2009). Sequential effects: Superstition or rational behavior? *NIPS*

[2] Frazier,P.I. & Powell, W.B. (2008). The Knowledge-gradient stopping rule for ranking and selection, *Proc. of the 2008 Winter Simulation Conference*

[3] Ryzhov, I., Powell, W & Frazier, P. (2012). The knowledge gradient algorithm for a general class of online learning problems, *Operations Research*

[4] Steyvers, M., Lee, M. D. & Wagenmakers, E.-J. (2012). A Bayesian analysis of human decision-making on bandit problems, *Journal of Mathematical Psychology*