Personality and Behavioral Predictors of Human Exploration in a Bandit Task

Shunan Zhang, Alvita Tran, Angela J. Yu {s6zhang, aetran, ajyu}@ucsd.edu University of California, San Diego, Department of Cognitive Science

Summary

Evaluate a set of models of human learning and action selection in a *bandit* setting.

- How do people balance *exploration* and *exploitation*?
- How is it related to optimal behavior?
- Are there personality correlates?

Models considered

- **Dynamic Belief Model** [1] – Bayesian iterative inference assuming local patterns in a sequence of observations
- Different decision policies for balancing exploration and exploitation

Major results

- Subjects over-explore (compared to the optimal behavior)
  - smaller Softmax parameter β — more *matching* (as opposed to *maximizing*) of value in action selection
  - especially in scarce environment
- Forgetful Bayes learning partially explains over-exploration; optimal policy exploits more, simulated under subjects' beliefs
- Less random exploration is positively correlated with total reward achieved in bandit task, inhibitory control (separate behavioral metric), and locus of control

Behavioral Experiment

- 44 subjects, 15 trials, 4 arms, 50 games
- Reward rates fixed, iid sampled from the *environment*: Beta(4,2), Beta(2,4)
- In each environment, subjects played *same* games, with order of games randomized
- Payment in proportion to total reward

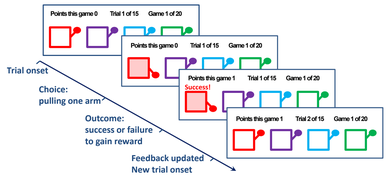

Figure 1: Schematic experimental design

Learning Model

For each arm k, the outcome R_k is Bernoulli with reward probability θ_k

Posterior of θ_k given observed sequence ($\mathbf{x}^{t-1}$) can be computed via Bayes' Rule:

$$q_k^t(\theta_k^t) \sim \Pr\left(R_k^t|\theta_k^t\right) \Pr\left(\theta_k^t|\mathbf{x}^{t-1}\right)$$

Dynamic Belief Model (DBM)

- assumes reward rates have discrete changes according to a hidden Markov model
- updates belief as a weighted sum of last trial's belief and the generic prior

$$\Pr\left(\theta_k^t = \theta|\mathbf{x}^{t-1}\right) = \gamma q_k^{t-1}(\theta) + (1-\gamma)q^0(\theta)$$

Decision Models

Softmax

- converts values into action probabilities

$$\Pr(D^t = k \,|\, \beta,\, \hat{\theta}^t) = \frac{(\hat{\theta}_k^t)^\beta}{\sum_{k'}(\hat{\theta}_{k'}^t)^\beta}$$

Knowledge Gradient [3]

- assumes next observation being the last exploratory choice for comparing values
- 1-step look-ahead approximation to optimal solution

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

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

τ-Switch

- before trial τ, chooses randomly
- after τ, chooses the best value

ϵ-Greedy

- with probability ϵ, chooses randomly
- otherwise, chooses the best value

Win-stay, Lose-shift

Softmax: Best Fitting Decision Model for All Subjects

Analyses

- Fit each decision model with DBM learning to human data (augmenting deterministic models into probabilistic models)
- Compare models using BIC, per-trial agreement
- Calculate subjects' beliefs using inferred parameters and observed sequence
- Simulate the optimal policy under the subjects beliefs and DBM γ values

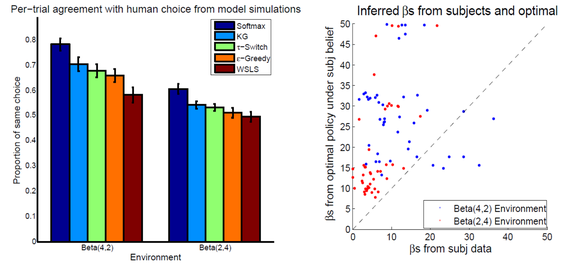

Figure 3: Left: proportions of choices consistent with human for each model simulated under MAP parameters, trial-to-trial. Error bars: s.e.m. Right: inferred subjects' βs are in general smaller than βs inferred from the optimal behavior simulated under subjects' beliefs

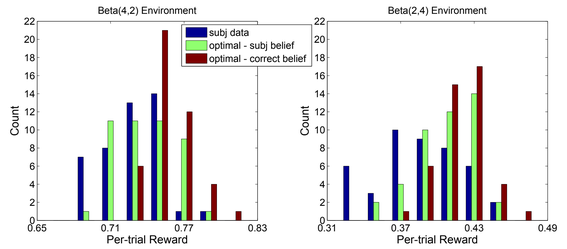

Figure 2: Subjects achieve less total reward than the optimal behavior simulated under subjects' beliefs of the environment and suboptimal settings of γ

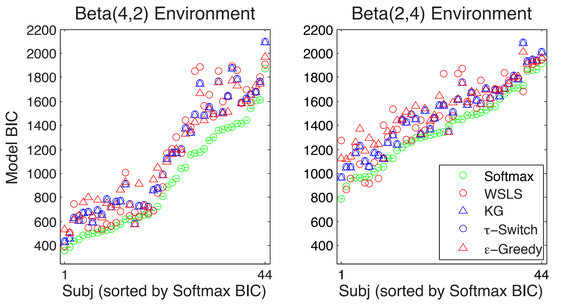

Figure 4: Softmax has the smallest Bayesian information criterion (BIC) across all models in both environments

Optimal Algorithm

Computed via Bellman's dynamic programming principle [2]

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

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

Use backward dynamic programming to compute the optimal policy off-line

Personality Correlates

Subjects did a separate experiment

- 12 cognitive tasks (memory, control, etc.)
- extent of belief-update after receiving positive vs negative feedback
- personality traits (anxiety, self-control, etc.)
- β is positively correlated with Stroop score, locus of control, and extent of belief change after positive feedback ($p<.01$)
- average reward in bandit task is positively correlated with β

References

[1] Yu,A & Cohen,J (2009). Sequential effects: Superstition or rational behavior? *NIPS* 21.
[2] Kaebling, Littman, & Moore (1996). Reinforcement learning: A survey. *J of Artificial Intelligence Res.*
[3] Ryzhov,I, Powell,W & Frazier,P (2012). The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*.