Why so gloomy? A Bayesian explanation of human pessimism bias in the multi-armed bandit task

Dalin Guo and Angela J. Yu
Department of Cognitive Science, University of California, San Diego, CA

1. Introduction

• Multi-armed bandit (MAB) task: exploration vs. exploitation, online learning, can be modeled as POMDP
• Dynamic Belief Model (DBM): Bayesian generative model for sequential data assuming abrupt change points
• Fixed Belief Model (FBM): DBM with no change point
• DBM predicts human behavior better than FBM in stationary environment: 2AFC, inhibitory control, visual search, MAB
• Why do humans persist in making non-stationary assumption?
• Human data: 4-armed bandit task, 4 reward environments (high/low reward abundance, high/low reward variance)
• Compare 3 models: DBM, FBM, and Reinforcement Learning
• Recover and explain human “pessimism bias” about reward rates

2. Experiment

• 107 UCSD students each played 200 15-trial 4-armed bandit (“ice-fishing”) games with binary outcomes (reward/no reward)
• Reward rates for all four arms were generated i.i.d. from four Beta distributions (1 for each environment): Beta(4, 2), Beta(30, 15), Beta(2, 4) and Beta(15, 30)
• Subjects shown 20 samples from the true distribution to inform decisions

3. Models

Dynamic Belief Model

DBM [1, 2] assumes the subject believes the reward rate undergoes discrete, un-signalized changes with a per-trial probability of 1-γ (contrary to experimental design). Generative model:

\[ p(R_t = r | \theta^{(t-1)}) = \beta(r+1, \theta^{(t-1)} ) + (1 - \gamma p(0)) \]

Recognition model (Bayes’ Rule, updates only for chosen arm):

\[ p(\theta_t | R_t = r, \theta^{(t-1)}) = p(\theta | R_t = r, \theta^{(t-1)}) \]

The reward belief is thus a weighted sum of the posterior and the prior \( p(\theta_t | R_t = r, \theta^{(t-1)}) \)

Fixed Belief Model

FBM [2] assumes reward rates fixed during the game (consistent with experimental design); can be viewed as a special case of DBM: \( \gamma = 1 \).

The prior \( p(\theta_t | R_t = r, \theta^{(t-1)}) \) enters only once (on trial 1) and fades in influence

Reinforcement Learning (RL)

Delta-rule updating [3]:

\[ \theta_t = \theta_{t-1} + \alpha (R_t - \theta_{t-1}) \]

Two free parameters, \( \alpha \) and \( \theta_0 \), which we call “prior” as shorthand.

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 RL has no means of injecting a prior bias on each trial [4].

Softmax decision policy

The choice probabilities are modeled by softmax:

\[ p(D_t = k) = \frac{e^{\theta_t + \phi_{0,1} R_t}}{\sum_k e^{\theta_t + \phi_{0,1} R_t}} \]

Optimal policy

The optimal policy can be computed via dynamic programming, though previously we showed humans do not behave optimally [2].

4. Results

The three models predict different shift rates:

• DBM: high probability of having changed to a lower reward rate ⇒ readily shifts away from a previously winning arm
• FBM: estimates follow long term stats ⇒ reluctant to switch
• RL: constant learning rate ⇒ slower than DBM to adjust

5. Discussion

• Four Models simulated with varying assumed prior mean
• Diamond markers: x-estimated prior mean, y-human performance

References


Acknowledgement

We thank Shunan Zhang, Henry Qiu, Alvia Tran, Joseph Schilz, and numerous undergraduate research assistants who helped in the data collection. We thank Samer Sabri for helpful input with the writing. This work was in part funded by an NSF CRCSN grant (BCS-1130946) to AY.