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

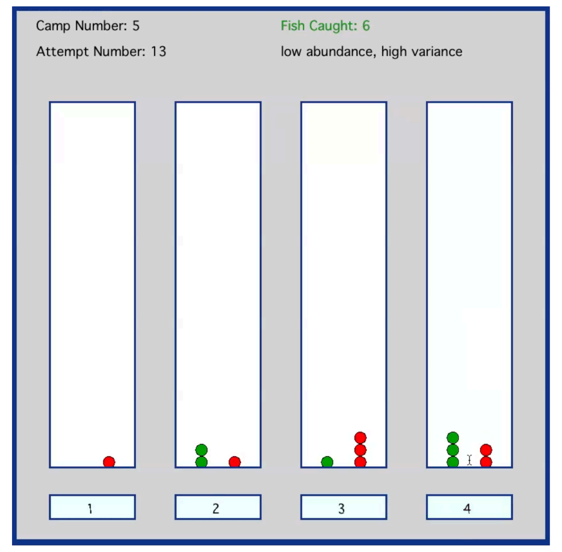
Dalin Guo and Angela J. Yu

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



Exploratory Fishing Report

5	7	5	7	7
7	6	7	7	8
6	8	7	6	7
7	5	7	7	6

It appears that this lake has high abundance, low variance of fish.

- 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 their prior beliefs
- 32 subjects reported their estimates of the reward rates of the unseen arms at the end of each game

3. Models

Dynamic Belief Model

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

$$p(\theta_k^t = \theta \mid \theta_k^{t-1}) = \gamma \delta(\theta_k^{t-1} - \theta) + (1-\gamma) p^0(\theta)$$

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

$$p(\theta_k^t \mid \mathbf{R}^t, \mathbf{D}^t) \propto p(R_t \mid \theta_k^t) p(\theta_k^t \mid \mathbf{R}^{t-1}, \mathbf{D}^{t-1}),\ \text{if } D_t = k$$

$$p(\theta_k^t \mid \mathbf{R}^t, \mathbf{D}^t) = p(\theta_k^t \mid \mathbf{R}^{t-1}, \mathbf{D}^{t-1}),\ \text{if } D_t \neq k$$

The reward belief is thus a weighted sum of the posterior and the prior $p^0(\theta)$ (repeatedly injected on each trial due to non-stationarity):

$$p(\theta_k^t = \theta \mid \mathbf{R}^{t-1}, \mathbf{D}^{t-1}) = \gamma p(\theta_k^{t-1} = \theta \mid \mathbf{R}^{t-1}, \mathbf{D}^{t-1}) + (1-\gamma) p^0(\theta)$$

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 $p^0(\theta)$ enters only once (on trial 1) and fades in influence

Reinforcement learning (RL)

Delta-rule updating [3]:

$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \epsilon (R_t - \hat{\theta}_k^{t-1})$$

Two free parameters, ϵ and $\hat{\theta}_k^0$, 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

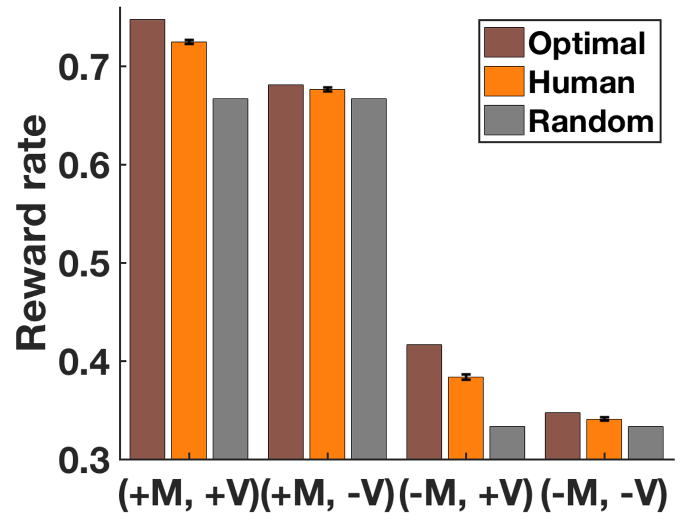
The choice probabilities are modeled by softmax:

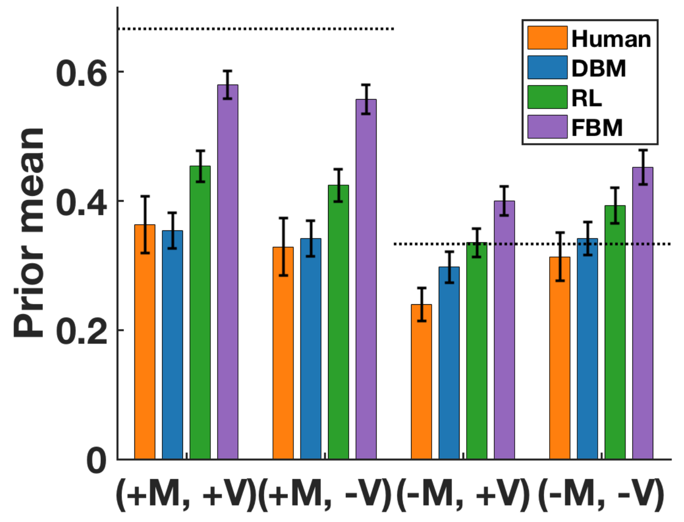
$$p(D_t = k) = \frac{(\hat{\theta}_k^t)^b}{\sum_i^K (\hat{\theta}_i^t)^b}$$

Optimal policy

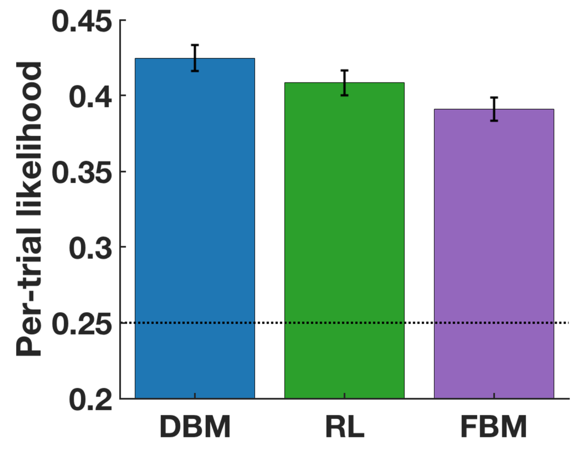
The optimal policy can be computed via dynamic programming, though previously we showed humans do not behave optimally [2].

4. Results

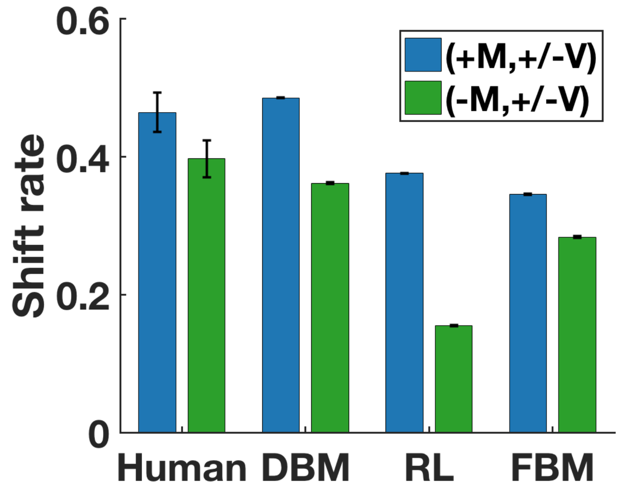




- (+M, -V): high mean and low variance environment
- Human performance close to optimal policy and higher than chance level (chance level equals to prior mean)
- Reported reward rate lower than the true mean
- DBM recovers prior mean most similar to human report



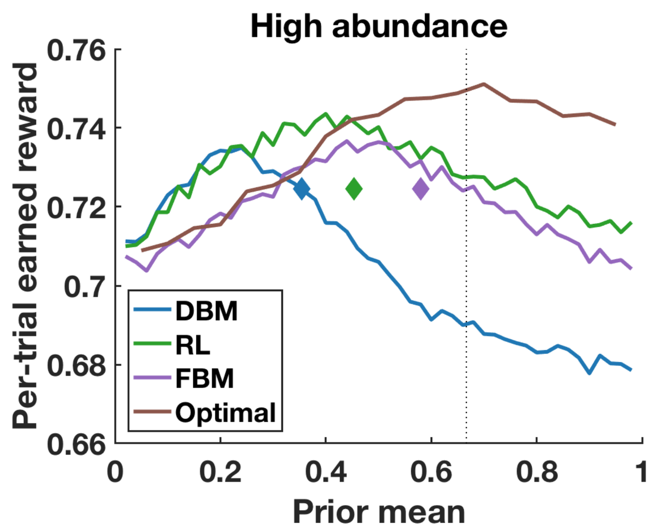
- 10-fold cross validation
- DBM achieves significantly higher per-trial likelihood
- DBM achieves lower BIC/AIC
- Further evidence that humans assume non-stationarity by default



- Shift rate: shifting to other arms after a failure preceded by three consecutive successes
- DBM predicts shift rate most similar to human data

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



- Four Models simulated with **varying assumed prior mean**
- Diamond markers: x-estimated prior mean, y-human performance
- Dotted line: true prior mean

- DBM predicts human performance the best
- FBM optimal with lower prior ⇒ compensates for simplified exploration policy (softmax)
- DBM optimal with even lower prior ⇒ compensates for (incorrect) non-stationary assumption
- Highest reward achievable by DBM and FBM quite close

5. Discussion

- Humans underestimate prior reward rates (pessimism bias)
- This underestimation recoverable by DBM, not FBM or RL
- Reward rate underestimation may help with performance, though human assumption a compromise between veridical representation of environmental statistics and optimizing reward
- Multiple human sub-optimalities combine to achieve better performance than might be expected

Acknowledgement

We thank Shunan Zhang, Henry Qiu, Alvita 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 CRCNS grant (BCS-1309346) to AJY.

References

[1] A J Yu and J D Cohen. Sequential effects: Superstition or rational behavior? Advances in Neural Information Processing Systems, 21:1873–80, 2009.

[2] S Zhang and A J Yu. Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting. Advances in Neural Information Processing Systems, 26, 2013.

[3] R A Rescorla and A R Wagner. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A H Black and W F Prokasy, editors, Classical Conditioning II: Current Research and Theory, pages 64–99. Appleton-Century-Crofts, Mew York, 1972.

[4] C Ryali, G Reddy, and A J Yu. Demystifying excessively volatile human learning: A bayesian persistent prior and a neural approximation. Advances in Neural Information Processing Systems, 2018.