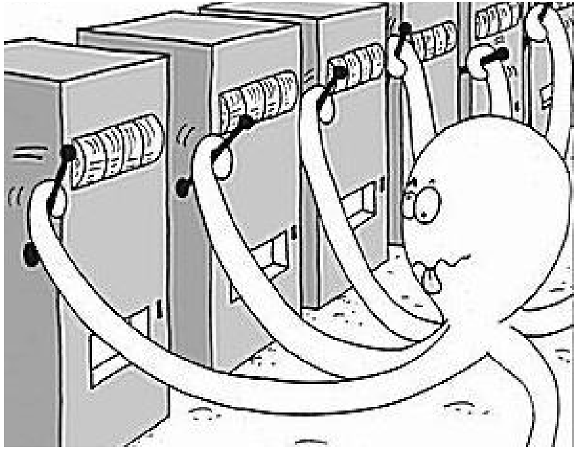
Decision Making in 2-Arm Bandit Problems

Hannah Sonali Sneha
Chun Yashodhan

The 2 -arm bandit problem

- The problem
  - 2 independent machines
  - one agent
  - Decision - which machine to choose
  - Exploration vs Exploitation trade-off
- For every machine k
  - State $\boldsymbol{S_t^k}$ after t transitions
  - Reward Rate $\boldsymbol{R^k}$



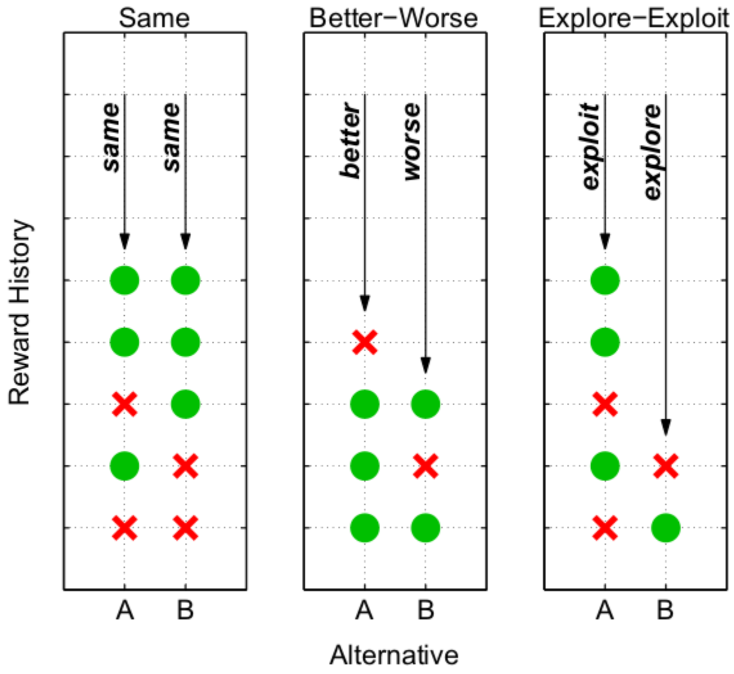
Heuristic Models

- ϵ - Greedy
  - Exploration with probability ϵ
  - Exploitation with probability $1 - \epsilon$
- ϵ - Decreasing
  - Decreasing value of ϵ
- Win-stay lose-shift
  - Stay after winning, shift after losing
  - Both with probability γ

Full Latent State Model

- Latent state for each trial
- Flexible switching between exploration and exploitation
- Decisions based on 3 different situations
  - Same
  - Better-worse
  - Explore-exploit

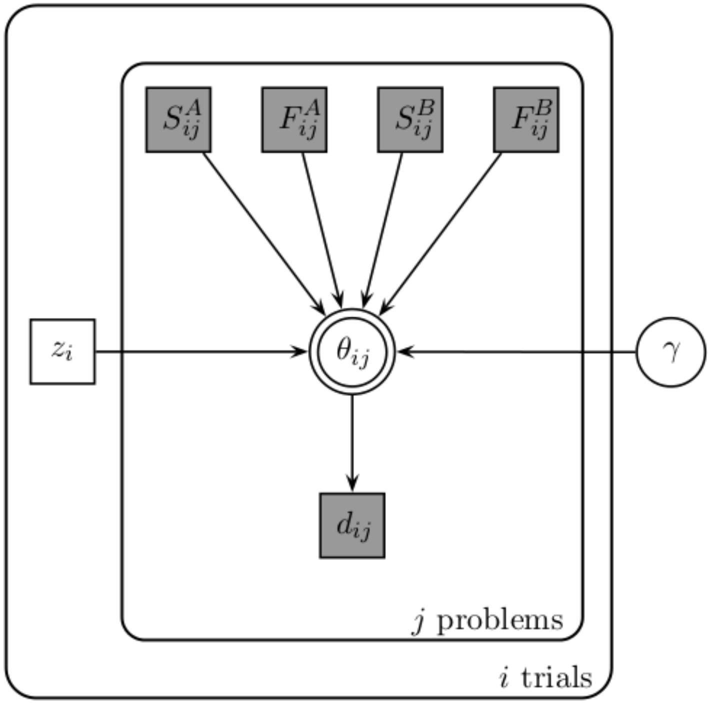
Full Latent State Model



$$\theta_{ij} = \begin{cases} 1/2, & \text{if A is same} \\ \gamma, & \text{if A is better} \\ 1-\gamma, & \text{if A is worse} \\ \gamma, & \text{if A is search and } z_i = 0 \\ 1-\gamma, & \text{if A is search and } z_i = 1 \\ \gamma, & \text{if A is stand and } z_i = 1 \\ 1-\gamma, & \text{if A is stand and } z_i = 0 \end{cases}$$

$$d_{ij} = Bernoulli(\theta_{ij})$$

Full Latent State Model



S : number of successes

F : number of failures

z_i : latent state variable

d_{ij} : decision

θ_{ij} : Bernoulli distribution for the decision

τ - switch model

- A simplification of the full latent state model
- Latent state z_i changes from exploration to exploitation only once during the game, after τ trials

Goals

- ➢ Generate optimal data
- ➢ Compare heuristic to optimal data
- ➢ Compare heuristic to human data
    - OMR dataset, 2011

Experimental Setup

- 50 games
- Each game has 8 trials
- 10 subjects
- Environment Settings
  - Define α (prior successes), β (prior failures)
    - $\alpha > \beta$: Plentiful
    - $\alpha = \beta$: Neutral
    - $\alpha < \beta$: Scarce

Generating Optimal Data

- Formulation of the problem
  - States : (s_1, s_2, f_1, f_2)
  - ***s****$_i$* = successes from arm ***i***, ***f****$_i$* = failures from arm ***i***
  - Environment settings : α , β
  - $V_t(s_1, s_2, f_1, f_2)$: Expected payoff after ***t*** trials
  - Recursive definition

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

Generating Optimal Decisions

- Enumerate all $V_t(s_1, s_2, f_1, f_2)$ and corresponding decision matrix $D_t(s_1, s_2, f_1, f_2)$
- Make a forward pass
  - Start with state S (0,0,0,0) and randomly pick the first action k
  - Sample reward $r_k \sim Bernoulli(\mu_k)$
  - Set next state as $nextState = (...., s_k + r_k, ..., f_k + 1 - r_k, ...)$
  - Thus, action at next trial t, action = $D_t(nextState)$
  - Repeat from Step 2 for all trials
- Output optimal decision sequence and corresponding reward sequence

Key Heuristic Parameters

Heuristic Method	Key Parameter	Meaning
ϵ - Greedy	ϵ	probability of exploration
ϵ - Decreasing	ϵ_0	probability of exploration in 0-th trial
WSLS	γ	probability of staying after winning and shifting after losing
Full Latent State	θ	Multifaceted
τ -switch	τ	trial # for switching from exploration to exploitation

Methods of Comparison

- Forward method: simulate model
  - Grid search for parameter based on decision sequence match percentage
  - Fit parameters from human/optimal data
  - Generate possible sequence
  - Compare sequences to optimal model sequence
- Fit models to optimal and human data
  - Define L-value

Methodology for Model Fitting

➢ Define L-value as

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

➢ Decision Rule

$$stay_t = I(D^t_{model} == D^{t-1}_{data})$$

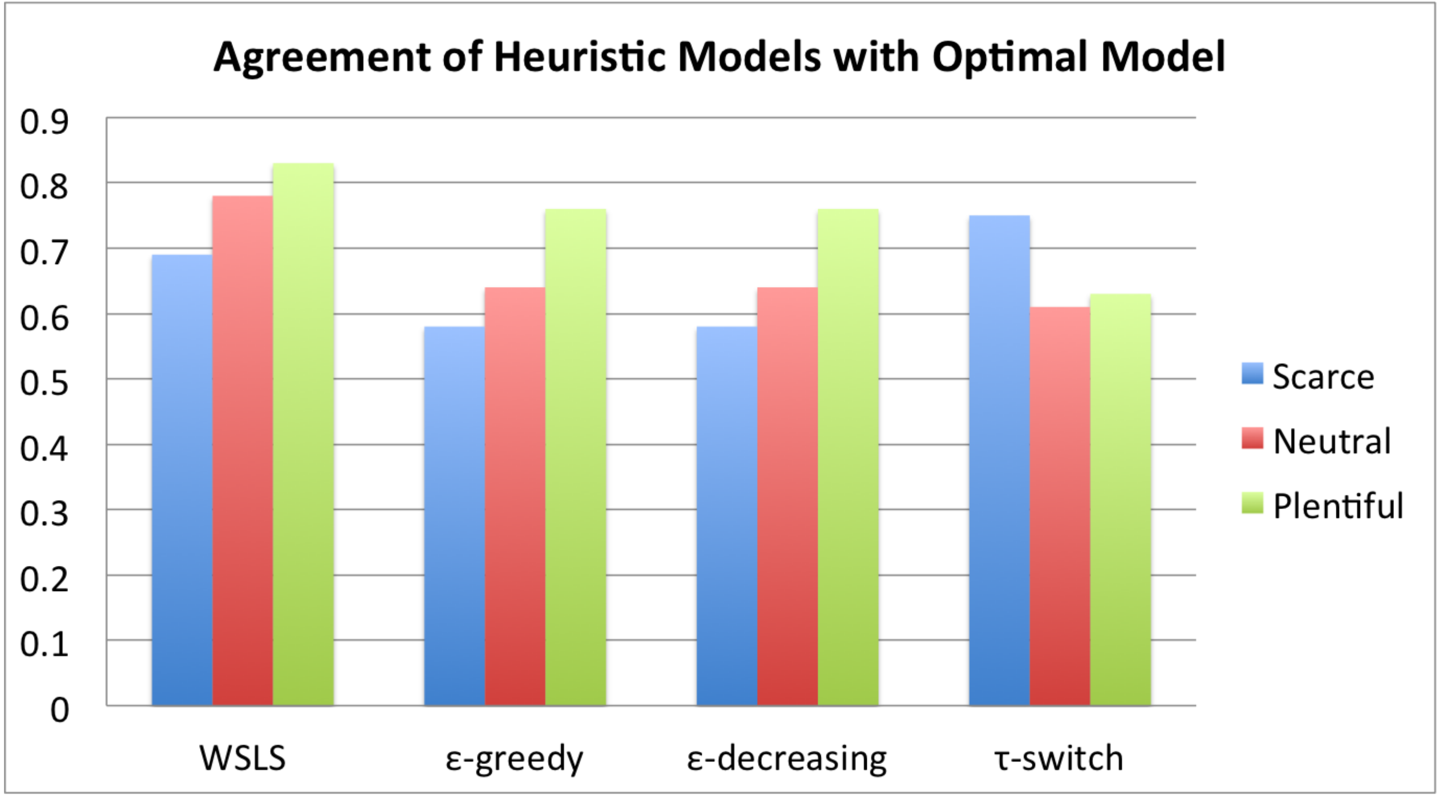
Model fitting to optimal data

- Given: Parameters α and β of the Beta distribution, optimal decisions d_{opt}, optimal rewards $r_{opt.}$
- For each value v of heuristic model parameter
  - For each decision sequence d_{model}
    - compute L-value
- Select (v, d_{model}) which maximizes L-value
- Compute percent match between d_{model} and d_{opt}

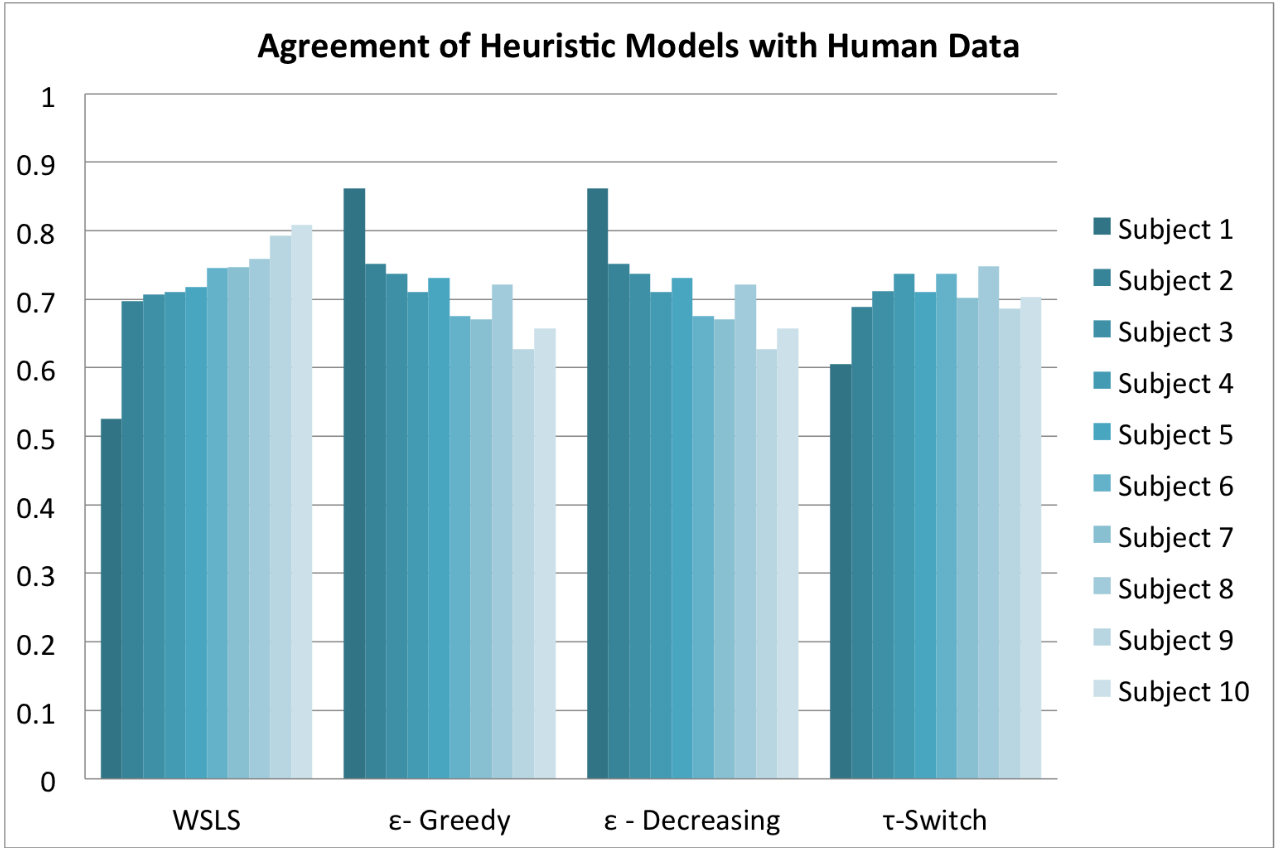
Model fitting to human data

- Given: Heuristic model parameters, optimal decisions d_{opt}, optimal rewards r_{opt}
- For each pair (α‘,β‘)
  - For each decision sequence d_{model}
    - Compute L-value
- Select (v, d_{model}) which maximizes L-value
- Compute percent match between d_{model} and d_{opt}

Results - Agreement with Optimal



Results - Agreement with Human Data



Discussion

- ➢ Heuristic models and optimal models
- ➢ Comparisons
- ➢ Limitations

Questions