THE MULTI-ARMED BANDIT PROBLEM

- Decision-maker sequentially chooses one among a set of competing alternative options called *arms*.
- Goal: maximize total cumulative reward
- Fundamental tradeoff: explore (collect more information to reduce uncertainty) vs. exploit (use current information to maximize reward)

FEATURES OF HUMAN DECISION-MAKING

We want to capture these features in a model:

1. **Ambiguity bonus**: human behavior in a 2-bandit task is explained by the following heuristic:
   \[ Q_i = \Delta R_i + A \Delta I_i \]
   \[ \Rightarrow \text{Use a heuristic-based model} \]
2. **Stochasticity**: human decision-making is noisy.
   \[ \Rightarrow \text{Use stochastic action selection} \]
3. **Finite-horizon effects**: both the ambiguity bonus and stochasticity are sensitive to the time horizon \( T \).
   \[ \Rightarrow \text{Make heuristic and decision functions of } T \]
4. **Familiarity with the environment**: human decision-making is affected by knowledge about the environment.
   \[ \Rightarrow \text{Use Bayesian inference} \]
5. **Environmental structure effects**: humans learn about the structure of the world.
   \[ \Rightarrow \text{Use correlated priors} \]

REGRET ANALYSIS OF UCL

For the non-informative prior, i.e., \( \delta^2 \to 0^+ \), the UCL algorithm achieves a logarithmic regret. In particular, the regret \( R_T \) satisfies the following uniform upper bound: \( R_T(T) \leq \sum_{i=1}^{N} \Delta_i \left( \frac{8 \sigma^2}{\Delta_i^2} + 2 \right) \log T + 4 \frac{\sigma^2}{\Delta_i} \left( -\log \log T + 1 - \log 2 \right) + 2 \).

THE GAUSSIAN MULTI-ARMED BANDIT PROBLEM

- \( N \)-armed bandit problem. At each time \( t \), pick arm \( i_t \).
- Each arm \( i \) has associated Gaussian reward \( \sim N(\mu_i, \sigma_i^2) \).
- Objective: \( \max_{m_j} J, J = \mathbb{E} \left[ \sum_{t=1}^{T} r_t \right] = \sum_{t=1}^{T} m_{i_t} \).
- Equivalently, minimize expected regret
   \[ J_R = \mathbb{E} \left[ \sum_{t=1}^{T} \Delta_{i_t} \right] = \sum_{i=1}^{N} \Delta_i \mathbb{E} \left[ n_i^T \right], \]
   \[ n_i^T = \text{cumulative number of times option } i \text{ has been chosen up to time } T, m_{i_t} = \max_i m_i, \Delta_i = m_{i_t} - m_i \]

- Lai and Robbins (1985) showed that any algorithm obeys
  \[ \mathbb{E} \left[ n_i^T \right] \geq \left( \frac{1}{D(p_i||p_{i^*})} + o(1) \right) \log T, \]
  where \( D(p_i||p_{i^*}) \) is Kullback-Liebler divergence.
- So regret \( J_R(T) \) grows at least logarithmically in time:
  \[ J_R(T) \geq C \log T \]
  where \( C \) is the sum of the above constant factors.

THE UPPER CREDIBLE LIMIT ALGORITHM

- Prior on \( m_i \): \( N(\mu_0, \sigma_0^2) \)
- Posterior mean and variance:
  \[ \mu_i = \frac{\delta^2 \mu_0^i + n_i^i \mu_i^i}{\delta^2 + n_i^i}, \sigma_i^2 = \frac{\sigma_s^2}{\delta^2 + n_i^i} \]
  where \( \delta^2 = \sigma_s^2/\sigma_0^2 \).
- Heuristic function
  \[ Q_i^t = \mu_i + \frac{\sigma_s}{\delta^2 + n_i^i} \Phi^{-1} \left( 1 - \frac{1}{K_i} \right). \]
- Stochastic UCL extension: use softmax action selection with temperature \( \nu_i = \frac{\nu}{\log T} \), select option \( i \) with probability
  \[ P_i = \exp(Q_i^t/\nu_i) \bigg/ \sum_{j=1}^{N} \exp(Q_j^t/\nu_i). \]
- Extends naturally to correlated priors (smoothness of environment parametrized by length scale \( \lambda \))
- Letting \( K, \nu \) be increasing functions of \( T \), e.g., \( K = \sqrt{2\pi e(\log T)^c} \), \( c \geq 0 \), captures trends in human subject data
- Stochastic UCL models human performance in spatial bandit tasks with parameters \((\mu_0, \sigma_0, \lambda, \nu)\)

HUMAN SUBJECT DATA

We ran a spatial bandit task on Amazon Mechanical Turk and collected data from 417 subjects. Three classes of performance emerge. These can be captured by the Stochastic UCL algorithm with appropriate parameter values.