Reinforcement Learning Analysis Plan for a Two-Armed Bandit Task

Luis Luarte

2025 - 08 - 01

1. Relevance of the Task

The two-armed bandit task is a classic paradigm for studying the neural and computational mechanisms of decision-making under uncertainty. It provides a formal framework for investigating how individuals balance the exploration of new, uncertain options with the exploitation of known, rewarding options. This exploration-exploitation trade-off is fundamental to adaptive behavior, from simple foraging to complex economic choices. This analysis aims to model participants' decision-making processes using a Q-learning framework to uncover the individual parameters that govern their learning and choice strategies.

2. Data Acquisition and Preprocessing from `lab.js`

The experimental task, built with `lab.js`, will generate a `.csv` file for each participant. This file contains the trial-by-trial data necessary for the reinforcement learning analysis.

2.1. Key Variables from Experimental Data

The primary data for the model will be extracted from the following columns in the output `.csv` file:

- `participantId`: A unique identifier for each subject. This will be used to group data for individual parameter estimation.
- `trial_index`: The trial number within the task.
- `response`: The participant's choice on a given trial, recorded as a string (e.g., `'left'` or `'right'`).
- `currentReward`: The numerical reward value received by the participant as a consequence of their choice on that trial.

2.2. Data Transformation

Before fitting the model, the raw data must be preprocessed. For each participant, we will create two vectors:

1. **Action Vector (a_t)**: The categorical `response` data will be converted into a numerical format. For instance, `'left'` can be coded as 1 and `'right'` as 2. This creates the sequence of actions $a_1, a_2, ..., a_T$ for each participant.
2. **Reward Vector (r_t)**: The `currentReward` column will be used directly as the sequence of rewards $r_1, r_2, ..., r_T$.

These two vectors, a_t and r_t, form the complete behavioral data sequence for each participant and will serve as the direct input for the model fitting procedure.

3. Computational Modeling of Choice Behavior

To analyze the choice behavior, we will use a Q-learning model. This model assumes that participants learn the expected value (Q-value) of each option through experience and use these values to guide their decisions.

3.1. Q-Value Estimation

After each trial t, where an action a_t is chosen and a reward r_t is received, the model computes a prediction error, δ_t:

$$\delta_t = r_t - Q_t(a_t)$$

This prediction error is then used to update the Q-value for the chosen action, scaled by the learning rate, α:

$$Q_{t+1}(a_t) = Q_t(a_t) + \alpha \cdot \delta_t$$

The Q-values for unchosen actions remain unchanged: $Q_{t+1}(a) = Q_t(a) \quad \forall a \neq a_t$.

3.2. Action Selection: The Softmax Function

The model uses a softmax function to translate Q-values into action probabilities, governed by the temperature parameter, τ:

$$P_t(a = i) = \frac{e^{Q_t(i)/\tau}}{\sum_{j=1}^{N} e^{Q_t(j)/\tau}}$$

Where N is the number of choices (here, $N = 2$). The τ parameter quantifies the stochasticity of choices.

4. Parameter Estimation Procedure

The free parameters of the model, α and τ, will be estimated for each participant by fitting the model to their sequence of actions (a_t) and rewards (r_t) obtained from the `lab.js` task. We will use Maximum Likelihood Estimation (MLE) to find the parameter values that maximize the probability of observing the participant's actual choice sequence.

4.1. Trial-by-Trial Estimation of Temperature (τ_t)

To test the primary hypothesis, which involves a trial-by-trial predictor, we need to estimate a dynamic temperature parameter, τ_t.

1. An optimal, stable learning rate (α) is estimated for each participant across all trials.
2. Using this fixed α and the participant's empirical data (a_t, r_t), we reconstruct the trial-by-trial Q-values for each option.
3. For each trial t, we use MLE to find the τ_t that best explains the specific choice a_t, given the reconstructed Q-values:

$$\tau_t = \arg\max_{\tau} P_t(a_t | Q_t, \tau)$$

This procedure yields a time-series of decision stochasticity (τ_t) for each participant.

5. Statistical Analysis Plan for Hypothesis Testing

The final step is to test the core hypothesis: that the variability of spontaneous behavior predicts the stochasticity of subsequent economic decisions.

5.1. Variables for Statistical Model

- **Dependent Variable (VDEC)**: The estimated trial-by-trial temperature, τ_t. This variable, derived from the Q-learning model, quantifies the randomness of the decision-making process on each trial.
- **Independent Variable (VESP)**: A measure of spontaneous behavior variability. Following the research plan, this would be derived from eye-tracking data (e.g., standard deviation of saccadic length) collected during the fixation cross period that precedes each choice in the `lab.js` task. Note that the task was not designed for eye-tracker data collection, fixation cross was just set as a placeholder for now, however, this is the ultimate goal of our reasearch project.

5.2. Statistical Model

Given the nested structure of the data (trials within participants), a Linear Mixed-Effects Model (LMM) is the appropriate statistical method. This approach allows us to account for individual differences while estimating the relationship between our variables of interest. The model will be specified as:

$$\tau_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})\text{VESP}_{ij} + \beta_2\text{Control}_{ij} + \epsilon_{ij}$$

Where:

- τ_{ij} is the temperature for trial i of subject j.
- β_0 is the fixed intercept.
- β_1 is the fixed effect of VESP, representing our main hypothesis. A significant positive β_1 would support the hypothesis that greater behavioral variability predicts more stochastic/exploratory decisions.
- Control_{ij} represents any control variables (e.g., trial number, age, gender).
- u_{0j} and u_{1j} are the random intercept and random slope for subject j, accounting for individual differences in baseline τ and in the strength of the VESP-τ relationship.
- ϵ_{ij} is the residual error.