BIO 450/IBS 534

Spring 2025

Emory University

**Reinforcement learning homework: 2-armed bandit (100 points total)**

1. What is a reward prediction error? How is it calculated? (**10 points**)

Reward prediction error is the difference between actual reward received and expected reward, which represent how surprising / how different is the reward.

1. In a Q-learning model, how do prediction errors relate to Q values? (**10 points**)

In Q-learning model, the prediction error is used to update Q value, which represent the expected rewards of a action.

The Q-values are updated using learning rate \* RPE

Q(s,a)←Q(s,a)+α⋅RPE = Q(s,a)+α(Reward – Q(s,a))

1. Learning in a stable environment (Experiment 1). Each simulated data file represents a single subject with 200 trials of a 2-armed bandit task with binary outcomes (win $1, win nothing). Fit the model to the data from the 5 “subjects” from Experiment 1 (SampleDataExp1\_1b-5b) and examine the output, displayed in the MATLAB window.

Include the best-fitting parameters and fit for each subject. (**15 points**)

A graph of a graph of a graph

AI-generated content may be incorrect.A graph of a graph of a graph

AI-generated content may be incorrect.A graph of a graph of a graph

AI-generated content may be incorrect.A graph of a graph of a graph

AI-generated content may be incorrect.A graph of different numbers and graphs

AI-generated content may be incorrect.

**Note for a-e**. If you run the script with the option “makeplots = 1;” you will see 3 plots for each subject showing the RPE, Q value, and probability of choosing each option on each trial. The values for option/action 1 are blue, and the values for option/action 2 are red.

* 1. Generally, how do RPE and Q values change over time? (**5 points**)

Generally, RPE and Q value decrease in oscillation and converge to around 0.5

* 1. What types of trials have the largest positive prediction error? Negative prediction error? (**5 points**)

The largest positive prediction error occur when a unexpected award is received (when subject underestimate the probability of reward). Largest negative prediction error occur when reward was expected was not received (when subject overestimate the probability of reward).

* 1. How would you describe the behavior of the subjects with the lowest learning rates? (**5 points**)

With lowest learning rate, the subject would update Q very slowly, which can be interpret that the subject rely more on prvious experience and less sensitive to changes, so their expectiation is less adaptive.

* 1. How would you describe the behavior of the subjects with the highest learning rates? (**5 points**)

With highest learning rate, the subject would update Q very quickly, which can be interpret that the subject rely less on prvious experience and very sensitive to changes, so their expectiation is very adaptive and immediately reflected in Q value.

* 1. From the data, what would you estimate as the probability of reward for Option 1? Option 2? (**5 points**)

The estimated probability of reward for each option depend on Q value, and overtime, the Q value would converge toward the probability of reward. Therefore, I estimate probability of option 1 and option 2 to around 50%.

* 1. What would you expect behavior to look like with a learning rate of 1? A learning rate of 0? (**5 points**)

A subject with learning rate of 1 would completely flip the decision/expectation based on the most recent trial. Therefore, their behavior would be highly unstable.

A subject with learning rate of 0 would never update its decision, completely undependent on past experience, and not learning from the experience.

4. Learning in a volatile environment (Experiment 2): Each data file contains 200 simulated trials from a 2-armed bandit task with binary outcomes (win $1, win nothing) with reversals (i.e. the option that more frequently gives rewards reverses throughout the task). Fit the model to the data from 5 subjects from Experiment 2 (SampleDataExp2\_1b-5b) and examine the output.

Include the best-fitting parameters and fit for each subject. (**15 points**)

A graph of a graph of a graph

AI-generated content may be incorrect.A graph of a graph of a graph

AI-generated content may be incorrect.A graph of a graph of a graph

AI-generated content may be incorrect.A graph of a graph of a graph

AI-generated content may be incorrect.A graph of a graph of a graph

AI-generated content may be incorrect.

* 1. Looking at the data, how would you describe the timing of reversals? (i.e. when do you think they occur?) (**10 points**)

When there’s a sudden shift in RPE (largest negative prediction error when previous optimal choice stop yielding reward; largest positive prediction error when the other option become more rewarding). Also when there’s a adjustment of Q value decline of the previously preferred option.

* 1. How does the environment affect optimality of learning rates? Is there a benefit to a lower learning rate? A higher learning rate? (**10 points**)

The lower learning rate is preferred in stable environment which prevent overaction of the trial, which could help maintain a consistant and accurate estimate of reward that help subject achieve more reward.

The higher learning rate is preferred in violatile environment which encourage sudden shift of option / rapid adapation (which lower learning rate wouldn’t react so quickly), makes it beneficial to obtain more reward.