Neural Modelling exercise 4: Model fitting and Pavlovian biases

Hand-in by **Saturday, 7.12.24 (midnight)** to neuralmodelling24@gmail.com

If you are handing in as a two person team, make sure to put both of your names on your solution (please hand in only one report per team). Besides your responses to the questions in text form, your submission should contain your code. You could e.g. link to a Github repo or submit your report as a Jupyter notebook.

Pavlovian-instrumental interactions (Lecture 6)

Download the dataset .csv from Slack. It contains the data of 10 subjects (see column "ID" for the subject identifier), performing a go/no-go task, each for 600 trials. The column "cue" informs you about the presented trial type (see the "cue_mapping" variable in our template). The column "pressed" contains the response of the participant (0 is no-go, 1 is go) and "outcome" contains whether a reward was delivered (1), nothing was delivered (0), a punishment was given (-1).

- Recreate figure 2E of the paper "Go and no-go learning in reward and punishment: Interactions between affect and effect" with the data you have. Only the bar plots are important here, no need for error bars or significance tests.
- Program the log likelihood functions of the models 1 to 7 (including) presented in "Disentangling the Roles of Approach, Activation and Valence in Instrumental and Pavlovian Responding" (see Table 2 of that paper for the model numbering and relevant parameters). The paper uses these parameters
  - learning rate ϵ
  - feedback sensitivity β
  - the general feedback sensitivity β can be replaced by separate reward and punishment sensitivities ρ (we don't include a sensitivity for omission)
  - there can be different learning rates ϵ for reward, feedback omission, and punishment (the paper doesn't make use of omissions, so they use only two learning rates, you will need three)
  - there can be a general bias to approach $bias_{app}$, and a general bias to withhold responding $bias_{wth}$
- Create an additional model which takes into account Pavlovian biases. Use model 7 as a starting point for this. Add a parameter p to the model. To determine the action values to put into the softmax function for a given cue, take the Q-values, add the general bias to approach or withhold (as in

equation 1 of the paper), and add p to the Q-value for approaching if the maximum Q-value for the current cue is positive, or add p to the Q-value for withholding if the maximal Q-value for the current cue is negative. That is, equation (1) of the paper becomes:

$$W(s_t, a_t) = Q_t(s_t, a_t) + b(a_t) + p'_t(a_t)$$

$$\text{with } p'_t(a_t) = \begin{cases} p & \text{if } max_a Q_t(s_t, a) > 0 \text{ and } a_t = 1 \\ p & \text{if } max_a Q_t(s_t, a) < 0 \text{ and } a_t = 0 \\ 0 & \text{else} \end{cases}$$

though note that for this specific model we also overwrite $b(a_t)$ with a different bias.

- Optimize the models, by fitting all the parameters of each model to each individual subject, using the scipy minimize function. Pay attention to initialize the parameters to reasonable values and set sensible bounds for each parameter (since Q-values get turned into probabilities through a softmax, which uses an exponential function, you may have to limit some of the parameters to certain magnitudes, to prevent overflow errors). Given the number of models this can take some minutes, to save time you can e.g. only apply the logarithm at the end, rather than during every iteration of your for-loop.

- Sum up the optimized log-likelihoods across all subjects for each model. Use this and all other relevant values to compute the BIC score for each model (using e.g. the BIC equation of Wikipedia). What does this tell you about which model describes the data best?

- Compare the fitted ϵ_{app} and ϵ_{wth} for the last model. How do you interpret the difference in their means?

- Bonus: Fit the first subject 10 times with the last model, using different initial parameters. Create a scatter plot between the fitted $bias_{app}$ and $bias_{wth}$ across the fits. How do you explain this plot?