Computational Cognitive Neuroscience Coursework

Reinforcement Learning for Computational Psychiatry. March 2026

Lecturer: Peggy Seriès
Coursework Teaching Assistant: Jeffrey Scholes

Practicalities

Deadline is 23rd March 2026 at 12:00 Noon (standard late policies apply). Please submit your report as a **single PDF** on Gradescope through the Learn page. Marking will be done through Gradescope; **please match the pages of your report to the questions on Gradescope** (2 marks will be deducted if you fail to do so).

- Please keep your submission anonymous. The submission should be no longer than **6 pages**. References, appendices and code do not count toward this limit.
- You **do** have to **include code in your report**. Please attach it to the end of your PDF as an appendix.
- Plots should always include axis labels and units. Figures should always have a caption and be referenced in the main text. The presentation and format will count for 10% of the final mark. Note that using default settings for plots is probably not the best idea (e.g. the line-widths are often too thin).
- Show your work. If there are more than one (sensible) ways to calculate something be sure to describe how you did it. Be concise and precise in how you report your results. Don't include lots and lots of separate graphs: you can superimpose different graphs in the same plot.
- Copying results is not allowed. It is okay to ask for help from your friends. However, this help must not extend to copying code or written text that your friend has written, or that you and your friend have written together. You are assessed on the basis of what you are able to do by yourself. Similarly, we ask you NOT to use Generative AI (e.g., ChatGPT) to write the code and descriptions that you will submit. Please adhere to the University guidelines on the limited, acceptable uses of GenAI in academia. Those failing to do so will be penalized and reported to the University's Academic Misconduct Officers.
- Your programming style will not be assessed but your code should be readable/decipherable in a reasonable amount of time. If your code is incorrect, you may still be able to get points by realising that the results are not correct and describing how you think they should look like.
- It is recommended that you try to complete the tasks in the order they are given. Please clearly indicate which question you are answering in your report.
- If something is not clear please create a new post on **Piazza** so that the answer can be shared with all the other students. If you are unsure whether your question can be shared with all the students, you can create a private Piazza question which, in case, can be made public by the course staff.

Marks for Style of Presentation (10 marks)

You should present your findings in the style of a scientific report. By this we mean that the document should include text broken into appropriate sections (e.g., by question number and/or topic), equations (if relevant), figures (with clearly labelled axes, legible font size similar to the font in the main text, with figure captions explaining their content, and with numbering/lettering of figures and subpanels in order to refer to them in the main text), and references (if relevant).

You should consider how to best communicate your findings in a clear and concise way to the reader. This includes designing figures appropriately. For example, by making use of subplots and plotting multiple things on the same graph to save space or for aid of visual comparison, setting the axis limits appropriately, using figure legends and colour schemes to delineate different things plotted on the same graph, etc. There is no need to aim for a 6 page report if you can convey your findings more succinctly! Marks may be deducted if you provide excessively long-winded, incoherent, or unstructured answers. To take these factors into account, 10 marks have been allocated for the quality of presentation with regards to the overall layout and style, the quality of figures, and the scientific clarity and precision of written answers.

Assignment (90 marks)

For this coursework we draw inspiration from Crawley et al., 2020 (1), who examined flexible behaviour in children, adolescents, and adults. The authors were interested in examining learning processes in those with Autism Spectrum Disorder (ASD) compared to those with typical development (TD) across developmental stages of life using a probabilistic reversal learning paradigm. The Autism Diagnostic Observation Schedule and Autism Diagnostic Interview-Revised were used alongside clinical diagnoses to improve diagnostic stability.

In our hypothetical study, we will use scores from the Autism-Spectrum Quotient questionnaire (AQ-50) (2), a 50-item self report questionnaire used as a screening for autism-spectrum traits (see here for an online version). We are also interested in investigating if ASD is related to less flexible behaviour (i.e. adapting behaviour when feedback changes), and differences in feedback sensitivity. Impairments in flexible behaviour have been proposed to relate to core features of ASD, for example, restricted, repetitive behaviours (sameness) (1). In the file, item scoring (see (2) for detail if interested) has already been applied and each of the questions is scored 0 or 1. Participant scores can thus be in the range 0-50. The AQ-50 also includes subscales assessing 5 different areas, each made up of a subset of questions relating to social skills, attention switching, attention to detail, communication, and imagination.

We will utilise a similar probabilistic reversal learning paradigm as used by Crawley et al., 2020. During our task, participants chose one of two coloured shapes (vertical yellow bars or horizontal blue bars), both of which had a chance to lead to the reward outcome with an 80/20 reward/punishment contingency. Positive feedback (reward) consisted of a green smiley face, while negative feedback (punishment) consisted of a red, frowning face. The aim for participants was to learn which stimulus most frequently resulted in a smiley face, which may change throughout the task.

Introduction

Three csv files containing simulated data for 50 adult participants can be found on the CCN open-course page. The experiment consists of 100 trials per participant, where the first 50 trials are the *acquisition* phase, and the last 50 are the *reversal* phase, where outcome contingencies swap for stimuli. Rewards are encoded as 1, and punishments are encoded as -1. At each trial, participants make a choice between the two shapes (vertical yellow bars or horizontal blue bars) and learn the option with the "best" outcome over time. For each participant, their first choice was selected by the experimenter to be the "best" option throughout the acquisition phase. During the reversal phase, this option was set to be the worst option (and the initially incorrect stimulus became the best choice). For example, if a participant chose the vertical yellow bars on the first trial, this stimulus would have an 80% chance of giving the reward outcome (and a 20% chance of the bad outcome) in the acquisition phase, while the horizontal blue bars would have a 20% chance of giving the reward outcome (and an 80% chance of the bad outcome). These probabilities then swapped in the reversal phase. Participants were aware that there was a good and a bad action to choose, but did not know which was which and were not aware of the assigned probabilities or their changes – i.e., they had to infer them from the observed outcomes of their actions, with the aim of maximizing reward.

Crawley et al. 2020 (1) utilised a variety of reinforcement learning models to model participant learning behaviour. We will apply several similar models here, starting with the simplest, shown in Box 1:

Box 1: Temporal Difference Reinforcement Learning (Q-learning)

Following standard Q-learning (Sutton & Barto, 1987 (3)), the expected value (Q) of the *chosen* action on the next trial ($t+1$) is updated as follows:

$$Q_C(t+1) = Q_C(t) + \alpha \cdot \delta(t) \tag{1}$$

where $C \in \{Y, B\}$ denotes the *chosen* option (yellow or blue stimulus). The value of the *unchosen* action remains unchanged. The trial-by-trial prediction error $\delta(t)$ is computed as the difference between the actual outcome (O) and the expected value (Q_C):

$$\delta(t) = O(t) - Q_C(t) \tag{2}$$

The probability of choosing the yellow stimulus, Y, on the *next* trial is modelled with a softmax function incorporating inverse temperature ($\beta > 0$):

$$P_Y(t+1) = \frac{e^{\beta Q_Y(t+1)}}{e^{\beta Q_Y(t+1)} + e^{\beta Q_B(t+1)}} \tag{3}$$

The expected values of both actions are initialized as follows:

$$Q_Y = Q_B = 0.55$$

The hyperparameters α and β are assumed to remain *constant* across trials.

Question 1. Exploring the data (8 marks)

Load and explore the provided data. For each participant, compute their total AQ-50 score, as well as each *subscale score* for the following sets of questionnaire items:

- *Social skills*: 1, 11, 13, 15, 22, 36, 44, 45, 47, 48
- *Attention switching*: 2, 4, 10, 16, 25, 32, 34, 37, 43, 46
- *Attention to detail*: 5, 6, 9, 12, 19, 23, 28, 29, 30, 49
- *Communication*: 7, 17, 18, 26, 27, 31, 33, 35, 38, 39
- *Imagination*: 3, 8, 14, 20, 21, 24, 40, 41, 42, 50

High AQ scores may support an ASD diagnosis, and have been shown to predict diagnosis of ASD (4). In the structured interviews, the first 25 participants were found to have a clinical diagnosis of ASD, while the last 25 were found to be typically developing adults (TD).

Task 1a): Compute the mean, standard deviation and median of the total score, as well as each of the subscale scores. Your answers should be rounded to 3 significant figures. Do this for: i) all participants, ii) the 25 ASD subjects and for the 25 TD subjects separately. Would you say that the AQ-50 scores align with the given diagnoses? Within the ASD group, which subscale shows the most variability? What might this variability imply?

Task 1b): It is suggested that an AQ-50 score of at least 32 is indicative of ASD (2) (5). If this rule had been used as the only diagnostic criterion, how many TD subjects would have been diagnosed as having ASD? How many subjects with ASD would have received no diagnosis?

Task 1c): Turning to the experimental part of the study, compute:

i): How many participants selected the yellow stimulus first.

ii): The "total smiley faces" and "total frown faces" received by each participant. Report the mean value of each of these quantities across individuals in the ASD/TD groups.

Task 1d): Using all participants, produce a trial-by-trial plot showing the probability of choosing the stimulus that was initially correct in the acquisition phase, i.e. for each trial: (number of participants choosing the initially correct stimulus at trial, t)/(total number of participants). Place p(initially correct) on the y axis and trial on the x axis. Clearly indicate the acquisition and reversal phases on this plot.

Briefly describe what this plot indicates about learning and behavioural flexibility of our participants. Did they seem to perform the task well?

Question 2. Simulations (7)

Since we are working with a generative model, we are able to simulate data, which can be extremely useful. Use equations (1) to (3) to write a function that lets you generate data for a single simulated participant from known parameters (the parameter values should be an input to the function). Generate the outcomes with probabilities which are changing across trials as described in the introduction.

Hint: You can take inspiration from the code from the lectures to help you get started.

For a hypothetical agent that always chooses the yellow stimulus first, simulate 100 choices with parameter settings $\alpha = 0.3, \beta = 2.5$ a number of times. (Choose a reasonable number so you can average the simulations). Illustrate the average evolution of values Q_Y and Q_B. Illustrate the average evolution of the difference in Q values of the two stimuli (i.e. show how $Q_Y - Q_B$ changes, on average, over the course of the simulated experiments). Very briefly explain what is observed and why the shape of this evolution makes sense.

Question 3. Exploring parameter settings (6)

Simulate 100 trials several times for a number of different parameter settings. Systematically vary settings of α and β to explore how different values affect the average net reward (number of smiley faces received) at the end of the simulation. Plot the average net reward as a function of the parameter settings for α and β, which means there will be three dimensions. Choose sensible ranges for the parameters (e.g. $0 < \alpha < 1$ and $0 < \beta < 5$).

Briefly describe your precise approach and comment on how the expected performance during the experiment is related to different settings of the parameters.

Question 4. Likelihood function (6)

To find the parameter values that best capture each participant's behaviour, we need to define the likelihood of the observed data under the model. Write a function that takes as input the data (choices and outcomes) for an individual and a vector of parameters (learning rate and inverse temperature). The function should return the negative log likelihood (NLL) of the observed data under the model, where θ is the parameter vector containing α and β, T is the total number of trials, and c_t is the individuals choice at time t.

$$NLL = -\sum_{t=1}^{T} \log p(c_t | Q_t, \theta).$$

Hint: You can use the code from the lectures to help you get started.

Report the NLL for the 10th and 20th participants using parameter setting $\alpha = 0.3, \beta = 2.5$.

Question 5. Model fitting (8)

Task 5a): Find the parameters that minimize the NLL for each individual: Pass your NLL function and a set of starting parameters (use $\alpha = 0.3, \beta = 2$) to a optimization function performing minimization.

Hint: For Python, you could have a look at `scipy.optimize` and particularly the Nelder-Mead algorithm.

Task 5b): Calculate and report mean and variance of the fitted parameter values for learning rate and inverse temperature. Illustrate the results: Plot the participant index on the x-axis, parameter values on the y-axis. Can any difference between the ASD and TD groups be seen from the plot? Comment briefly on what you observe. Do all the values make sense? If not, how can you adjust your minimisation so that they do? If your values do make sense, what have you done to ensure this?

Task 5c): Calculate and report the Pearson's correlation coefficient between estimated parameters (i.e. between α and β across all participants). Also, calculate and report the Pearson's correlation coefficient between estimated parameters separately for participants within each ASD and TD group.

Question 6. Parameter recovery (8)

We now want to check the reliability and identifiability of our parameter estimates.

Task 6a): Sample 50 pairs of parameter values (α, β) from a multivariate normal distribution. Choose sensible numbers for the mean of this distribution and describe how you chose them; choose small numbers for the variance; set the covariance to zero. Illustrate the sampled values and highlight and exclude (resample) nonsensical values.

Task 6b): Use the sampled parameter values to simulate 50 sets of data (as in Question 2). Fit new parameter values to these simulated data sets (as in Question 5). Calculate, report and illustrate the Pearson correlation between the parameter values you used to simulate the data and the parameter values that you obtained from fitting the model to the simulated data. Repeat this process 5 times and report the Pearson correlation each time (there is no need for you to plot the sampled values all 5 times, the correlation coefficients are enough).

Task 6c): Comment briefly. Does this parameter recovery simulation meet your expectations? Explore and describe how the number of trials and the number of simulated data sets affect the performance of the parameter recovery (be sure to maintain the same task structure, i.e. first half of trials = acquisition, second half = reversal).

Question 7. Alternative model - reward-punishment (9)

Some studies have suggested that impairments in flexible behaviour may be related to reduced punishment-specific learning. To assess whether this may be present in our participants, we introduce a new model that extends upon the Q-learning model shown in Box 1. Here, we split the learning rate in two, using one for value updates following reward outcomes, and another for value updates following punishment outcomes (see Box 2).

Task 7a): Fit this reward-punishment model to the data of each participant, as in Question 5. In this model, we use different learning rate parameters based on outcome value (see Box 2).

Task 7b): For each participant you should now have two negative log likelihood values: One for the basic model (Question 5), and one for the reward-punishment model (Question 7). Compare these negative log likelihood values between models. Explain what you observe. Does it make sense?

Task 7c): For each participant, compute Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) scores for each model. For what percentage of participants does the BIC score improve when we have specific learning rates for reward/punishment? And the AIC score? Briefly discuss what this indicates.

Box 2: Reward-punishment model

Similar to the Q-learning model shown in box 1, however, the value update Equation (1) is now:

$$Q_C(t+1) = \begin{cases} Q_C(t) + \alpha^{\text{rew}}\delta(t), & \text{if } O(t) > 0 \\ Q_C(t) + \alpha^{\text{pun}}\delta(t), & \text{if } O(t) < 0 \end{cases} \tag{4}$$

The initial expected values of both stimuli (Q_Y, Q_B) are once again 0.55, and the value of the *unchosen* action remains unchanged. Prediction errors and choice probabilities are computed in the same fashion as the first model (using Equations 2 and 3). Each participant thus has 3 parameter estimates (α^{rew}, α^{pun}, and β).

Box 3: Model selection criteria

For your reference, these are the formulas used to compute AIC and BIC:

$$\text{AIC} = 2k - 2\ln(L)$$

$$\text{BIC} = k\ln(n) - 2\ln(L)$$

where k is the number of parameters in the model, L is the maximized value of the model's likelihood function (i.e., $\ln(L)$ is the negative of the minimized NLL), and n is the number of samples used to train the model. Generally, lower AIC/BIC scores indicate a better fit.

Going forward, we will consider the model with separate learning rate parameters for reward/punishment outcomes, as this model has been shown to outperform the basic Q-learning model shown in Box 1 e.g. (1).

Question 8. Group comparison (4)

Use a two sample t-test (describe which exact test you are using) to test whether the estimated parameter values (Question 7) are significantly different across groups (recall the first 25 participants are your ASD group). Report the t statistic, degrees of freedom and p value if applicable. Explain briefly how we can interpret the results and how this relates to our hypotheses.

Question 9. Alternative model - Experience-weighted attraction-dynamic learning rate (EWA-DL) model (9)

It has been suggested that reduced flexible task behaviour may also result from becoming insensitive to novel information over time, leading to a perseveration of previously rewarded choices, or a complete failure to switch during the reversal phase (1) (6). To assess, we introduce the experience-weighted attraction model (as used by Crawley et al., 2020 (1)) which allows us to capture a dynamic learning rate that changes over time (see box 3). Specifically, this allows us to assess if participants (ASD or TD) become more reliant on past experience and less sensitive to new feedback as time goes on (i.e. if learning slows down dramatically across trials).

Implement simulation and negative log likelihood functions for this new model. Fit the model to the data, using $\rho = 0.3$, $\varphi = 0.5$, and $\beta = 2$ as initial values.

As in Question 5, illustrate the fitted parameter values.

Box 3: EWA-DL model

The EWA-DL model introduces an accumulating experience weight, $n_C(t)$, for each stimulus that grows across trials, causing the model to rely increasingly on past experience and less on new outcomes. This results in changes to updating over trials and may potentially explain impairments in flexible behaviour.

A separate experience weight term for each stimulus is updated at every trial with an experience decay parameter ρ. Previous value estimates also decay via the φ parameter.

$$n_C(t+1) = n_C(t) \times \rho + 1 \tag{5}$$

$$Q_C(t+1) = (Q_C(t) \times \varphi \times n_C(t) + O(t))/n_C(t+1) \tag{6}$$

Note that both φ and ρ can take values in the range (0, 1). The initial expected values of both stimuli (Q_Y, Q_B) are once again 0.55, and the initial value of the experience weights for both stimuli ($n_Y(0)$, $n_B(0)$) are 1. The value and experience weight of the *unchosen* stimulus remains unchanged. Choice probabilities are computed in the same fashion as the first model (using Equation 3). Each participant thus has 3 parameter estimates (ρ, φ, and β).

Question 10. Model comparison (7)

We would now like to compare the two alternative models (Questions 7 and 9).

Task 10a): For each participant, and for both models, compute AIC and BIC scores. Sum up the participant scores for each model (i.e. for each model you will have a single BIC and a single AIC score). Report the results. Comment briefly. Which model would you choose as the 'best' model?

Task 10b): For the parameters of the new EWA-DL model, compute group comparisons, as in Question 8. Report t statistic, degrees of freedom and p value if applicable. Do you find statistically significant differences for a parameter between the ASD and TD groups? How could these be interpreted? What might we conclude if the data was real?

Question 11. Model recovery and confusion matrix (10)

We now want to check the reliability of our model comparison procedure using model recovery simulations. This will give us some indication about how much we can trust our results from Question 10. For the models from questions 7 and 9, simulate data (using the functions you already implemented) multiple times. For each of these simulated data sets, fit each model and use model comparison (i.e., AIC and BIC scores) to choose the best model. Display your results in two confusion matrices – one based on AIC, one on BIC. Comment briefly. (Make sure to explain your approach and results in appropriate detail.)

Question 12. Discussion (8)

Summarize and discuss the findings of your report. Did your model fitting procedures demonstrate notable differences between the ASD and TD groups? How may these differences, if any have been found, relate to symptoms of ASD? Point out any limitations of your experiments that may affect the interpretability of your results. Write a conclusion for your report, as you would expect to find in a scientific paper.

References

[1] Crawley D, Zhang L, Jones EJH, Ahmad J, Oakley B, San José Cáceres A, et al. Modeling flexible behavior in childhood to adulthood shows age-dependent learning mechanisms and less optimal learning in autism in each age group. PLoS Biology. 2020 Oct;18(10):e3000908. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7591042/.

[2] Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E. The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. Journal of Autism and Developmental Disorders. 2001 Feb;31(1):5-17.

[3] Sutton RS, Barto AG. A Temporal-Difference Model of Classical Conditioning. Proceedings of the Annual Meeting of the Cognitive Science Society. 1987;9(0). Available from: https://escholarship.org/uc/item/9ps125p9.

[4] Woodbury-Smith MR, Robinson J, Wheelwright S, Baron-Cohen S. Screening adults for Asperger Syndrome using the AQ: a preliminary study of its diagnostic validity in clinical practice. Journal of Autism and Developmental Disorders. 2005 Jun;35(3):331-5.

[5] Ashwood KL, Gillan N, Horder J, Hayward H, Woodhouse E, McEwen FS, et al. Predicting the diagnosis of autism in adults using the Autism-Spectrum Quotient (AQ) questionnaire. Psychological Medicine. 2016 Sep;46(12):2595-604. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC4988267/.

[6] den Ouden HEM, Daw ND, Fernandez G, Elshout JA, Rijpkema M, Hoogman M, et al. Dissociable Effects of Dopamine and Serotonin on Reversal Learning. Neuron. 2013 Nov;80(4):1090-100. Available from: https://www.cell.com/neuron/abstract/S0896-6273(13)00789-7.