Problem set 2 - Tutorial

Nicolás Bruno

March 6th 2020

Supervisor: Manuel Beiran

1 PROBLEM 1: *Rescorla-Wagner Model*

This project will focus on modeling classical conditioning learning using the Rescorla-Wagner Model. This model reproduced the behavior of an animal while learning the association between a stimulus and its reward. This led to the Rescorla-Wagner rule that is an equation for updating the expected reward (v) associated with the stimulus u.

$$w \rightarrow w + \epsilon \delta u \tag{1}$$

Where ϵ is the learning rate and $\delta = r - v$ that is the prediction error and w is the parameter that the animal is trying to learn.

1.a

First, 50 trials were generated in which during the first half of them a stimulus was presented and the animal received a reward of $r = 1$. During the last 25 trials, the stimulus was still present, however, no reward was given.

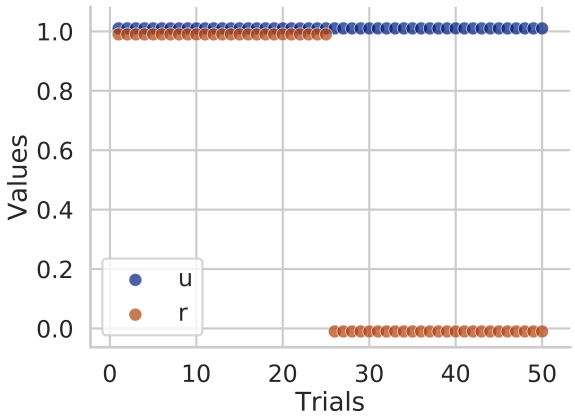

Figure 1: Each point in the plot represents each trial where u is the stimulus and r the reward. The plot shows the quality of the reward and if the stimulus was present $u = 1$ or absent $u = 0$

1.b

Now, the Rescorla-Wagner model was added to the stimulation and rewarding program use in **a** in order to predict the estimated reward of v.

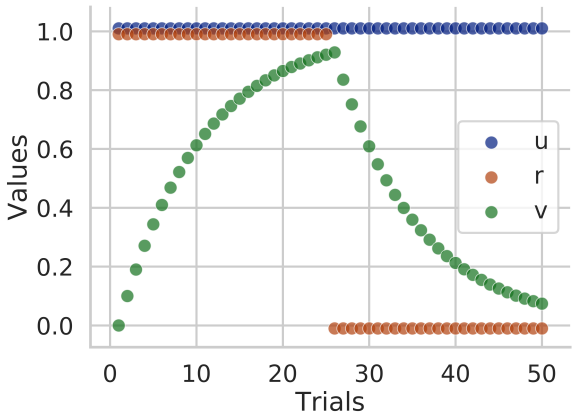

Figure 2: Each point in the plot represents each trial where u is the stimulus, r the reward and v is the predicted reward value.

It can be observed the learning curve during the first 25 trials were the reward was present, in which over the trials the prediction v starts to get closer and closer to the real value of $r = 1$. And in the last 25 trials, an extinction curve can be observed, where the value of the prediction starts to gets closer to the new value of the reward.

1.c

Furthermore, the effect of different ϵ learning parameters were tested over the model.

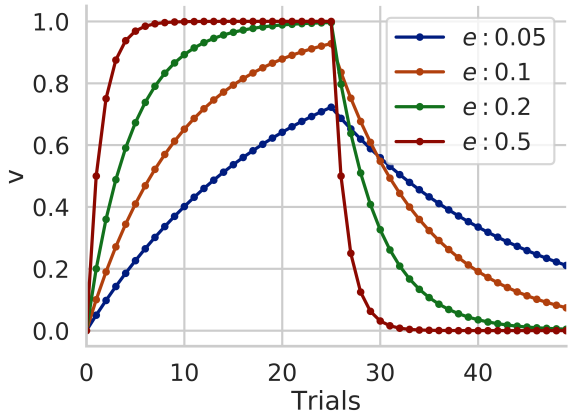

Figure 3: Each point in the plot represents each trial and v which is the predicted reward value. Each different line represents the model learning curve with different ϵ learning rate parameter.

It can be seen that as much bigger the ϵ parameter the faster that the v value gets to the r value, this means that it takes fewer trials to predict correctly the value of the reward.

1.d Partial Conditioning

In partial conditioning was tested the effect on the prediction of a random reinforcement program, where the reward was sometimes present but others not in a random manner.

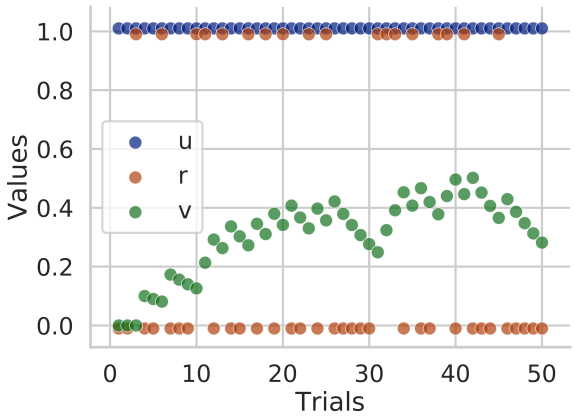

Figure 4: Each point in the plot represents each trial where u is the stimulus, r the reward and v is the predicted reward value.

It can be observed in **Figure 4** that when the reward is not always present the learning rate takes much more time, and v never gets close to r over these 50 trials. This indicates that the random reinforcement program produces more uncertainty to the model, therefore, it is less precise in its prediction of the reward.

1.e Blocking

Afterward, it was tested the Blocking effect, this effect consists in not learning the association of second stimuli and the reward when the second stimuli are added after it was already learned that the first stimulus predicts the reward. To test this, a second stimulus u_2 was added to the model to the last 25 trials.

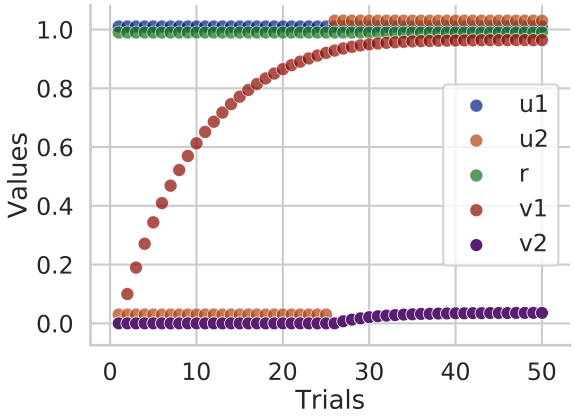

Figure 5: Each point in the plot represents each trial where u_1 is the first stimulus, u_2 indicates the presence or absence of the second stimulus, r the reward, v_1 is the predicted reward value for u_1 and v_2 indicants the predicted value for u_2.

It was observed that the predictions v_2 it does not change much when the stimulus u_2 was present in comparison to when it was absent. Furthermore, the v_1 learning curve was not affected by the presentation of the u_2. This Rescorla-Wagner model, therefore, is good for modeling the blocking effect as it can be observed in **Figure 5**.

1.f Overshadowing

Lastly, the overshadowing effect was modeled. Overshadowing stands for an effect found in Classical Conditioning in which the association to one of the stimuli to the reward it is learned faster than to the other stimulus, e.g. this can be due to one of the stimuli having a greater salience than the other so the animal gives more importance to the most salient. In order to model this effect, a second stimulus u_2 was added from the beginning with a learning rate of $\epsilon = 0.2$ while the learning rate of u_1 was $\epsilon = 0.1$.

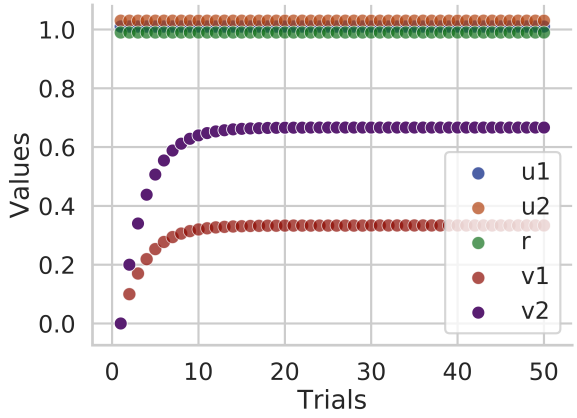

Figure 6: Each point in the plot represents each trial where u_1 is the first stimulus, u_2 indicates the presence or absence of the second stimulus, r the reward, v_1 is the predicted reward value for u_1 and v_2 indicates the predicted value for u_2

It can be seen in **Figure 6** the overshadowing effect, as it was expected the v_2 value is higher than v_1. This was expected because the learning rate was higher, thus, the animal would learn that u_2 predicts more information about the reward than u_1. However, u_2 does not explain alone the total of the reward, therefore, this is explained by the sum of both v_1 and v_2.

1.g Conclusions of problem 1.

The Rescorla-Wagner model is good for modeling Classical Conditioning learning with one or two stimuli.

2 PROBLEM 2: *Simple decision strategy for flower sampling by bees*

In the following exercise it was study the Softmax-Gibbs strategy to calculate the probability that a bee chooses a flower over the other based on the mental estimate (m) of the flowers reward. The function consist on:

$$p_b = \frac{1}{1 + exp(\beta(m_y - m_b))} \tag{2}$$

Where p_b is the probability of choosing the blue flower and β is the exploration-exploitation parameter, this indicates how much the bee will exploit the already known flower or how much will explore new possibilities.

2.a

First of all, the effect of the β parameter over the probability of choosing the blue flower was studied by plotting the p_b as a function of β with a fix $m_y - m_b$ difference.

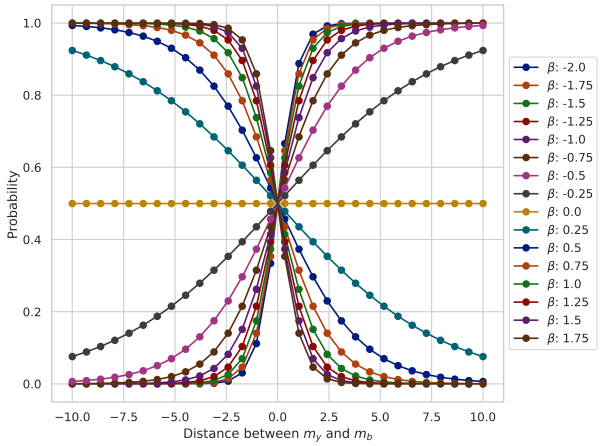

Figure 7: Plot of p_b as a function of the β exploration-exploitation parameter, with different fix distances between m_y and m_b. Each line represents a different value of β

It can be observed from **Figure 7** that for positive β values if the distance between m_y and m_b is positive, which indicates that m_y is higher, the probability of choosing the blue flower is really low. This is because the reward associated with the yellow flower is greater and as bigger the β is the more the exploitation of the best rewarding flower will be. However, when the m_b was bigger than m_y (i.e. the distance was negative) the probability of choosing the blue flower is high. For negative β interestingly the bee would choose the least rewarding flower, which implies that the bee will be making the choice that is the least beneficial on purpose. Afterward, the inverse plot was made in order to test the effect of the distance between m_y and m_b with a fixed β parameter.

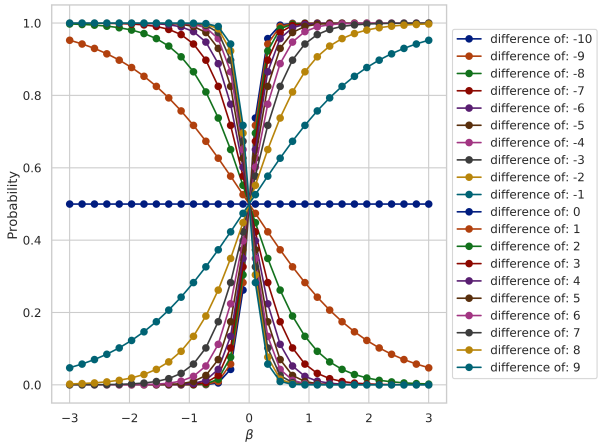

Figure 8: Plot of p_b as a function of the distances between m_y and m_b., with different fixed β exploration-exploitation parameter. Each line represents a different value of $m_y - m_b$

Similarly, it can be observed that for positive distances the probability of choosing the blue flower is really low, nevertheless, for negative distances the p_b is high. This indicates that the highest rewarding flower is the one that is the most exploited when β is high. However, for low β even if one flower gives more rewards the probability of choosing that flower is never 1, i.e. the bee will explore more even knowing than one flower gives greater rewards.

2.b Dumb bee

Here it was model the probability of choosing each flower for a bee that does not learn from its experience, then, 'dumb' bee. The $m_y = 5$ and $m_b = 0$ was constant over the 2 days of training.

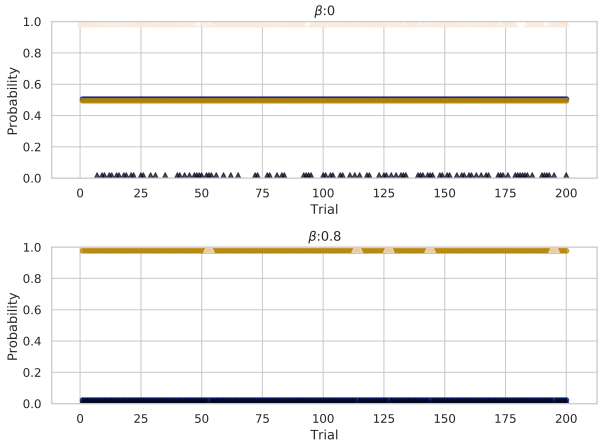

Figure 9: Plot of p_b for a $\beta = 0$ (upper) and a $\beta = 0.8$ (lower). A choice of 1 indicates that the yellow flower was chosen and a choice of 2 that the blue flower was selected.

It can be drawn from **Figure 9** that as the bee does not learn from its experience the probability of choosing one flower is constant over all the trials. It was also observed that when $\beta = 0$ the exploration was too high, so, the probability of choosing each flower was 0.5 even though the yellow flower gave a greater reward. Nonetheless, when $\beta = 0.8$ as m_y was higher than m_b, the exploitation of the most rewarding flower was very high and then the probability of choosing the yellow flower was 1 over all the trials.

2.c Smart bee

Finally, an online update rule was used for modeling the bee's learning of the rewards of each flower. The online update rule, is the learning rule from Rescorla-Wagner model in **Equation 1**. The learning rate used was of $\epsilon = 0.2$. The initial assumptions were the same that for 'Dumb bee', but afterwards the rewards for the first day were $r_y = 2$ and $r_b = 8$ and for the second day were $r_y = 8$ and $r_b = 2$

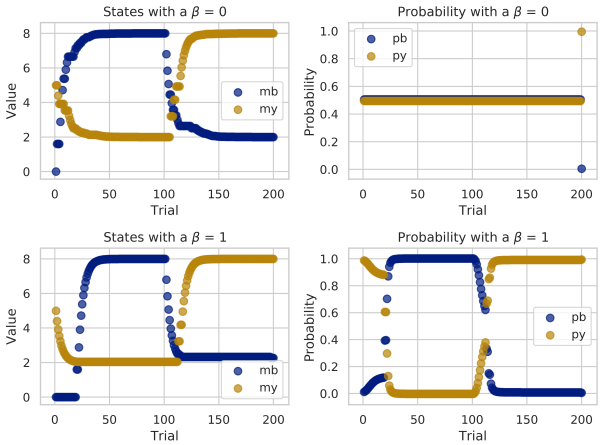

Figure 10: (Right) Plots of the estimated value for m_b and m_y for for a $\beta = 0$ (upper) and a $\beta = 0.8$ (lower). (Left) Plot of p_b and p_y for a $\beta = 0$ (upper) and a $\beta = 0.8$ (lower).

In the right plots of **Figure 10** it can be observed the learning rate of the prediction of the rewards for each flower. The learning and extinction curve are similar to the ones observed for Problem 1. It does not seem to be a significant difference for the learning curves when using different β values. However, in the left plots, it is observed a clear difference in the probabilities of choosing each flower depending on the β. In the upper-left plot, it is observed that the probability of choosing each flower is independent of the estimate of the reward. This is due to the fact that as observed in the Dumbe bee when using a $\beta = 0$ the reward does not matter to the bee and will choose the flower at random. On the other side, when a $\beta = 1$ is used (lower-left) it is observed that the probability of choosing each flower it behaves similarly to the estimate of the reward (m) on the plot in the right side.

2.d Conclusions of problem 2

It can be concluded that the β parameter is of a great importance when making modeling behavior because the choice probability it depends directly on this parameter. So, tuning this parameter it seems to be the greatest challenge when using the Softmax-Gibbs policy.

3 PROBLEM 3 *The drift diffusion model of decision-making.*

The drift diffusion model is used for modelling two alternative force choice tasks. This model allows you to model the reaction times, the percentage of correct and wrong responses, the noise of the subject and the sensory and motor processing time. The final differential equation of the model could be:

$$x(t + \Delta t) = x(t) + (m_A - m_B)\Delta t + \sigma\eta(t)\sqrt{\Delta t} \tag{3}$$

3.a

First, 10 runs of the DDM were performed with an $x(0) = 0$, $\Delta t = 0.1$ms, a noise level $\sigma = 0.5$ms.

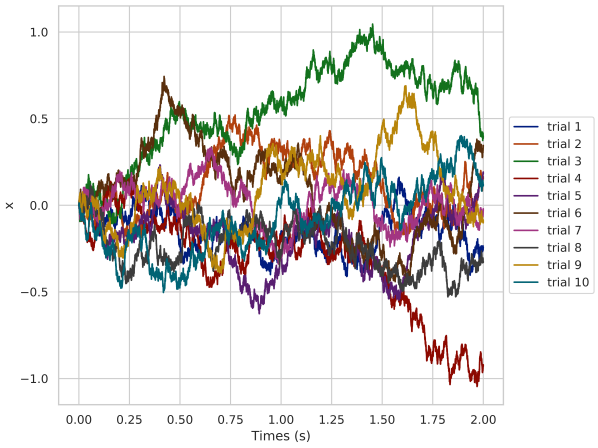

Figure 11: Each line represents a different trial/run of the DDM.

It can be observed that a good value of μ for this model could be 0.5.

3.b

Now, using the value of $\mu = 0.5$, obtained observing the trials in (a), the reaction times were recorded for each response. A response was considered A when the integration was equal to μ and a response was considered B when the integration was $-\mu$.

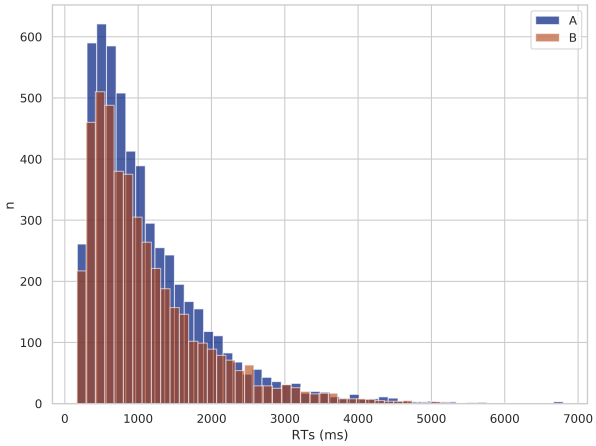

Figure 12: Histogram of RTs for A and B in seconds for a 1000 simulations using DDM.

As it was expected the RTs for each different outcome did not differ significantly. However, for the slight difference of 0.05 between m_A and m_B it can be observed that more A responses were given.

3.c

Lastly, the probability of choosing A obtained with the DDM was compared to the probability obtained with the Softmax-Gibbs algorithm (Equation 2). Both models were performed for 1000 trials and their probabilities were plotted against different values of m_e ($m_A - m_B$). The β parameter for the Softmax model used was $\beta = 2\mu/\sigma^2$.

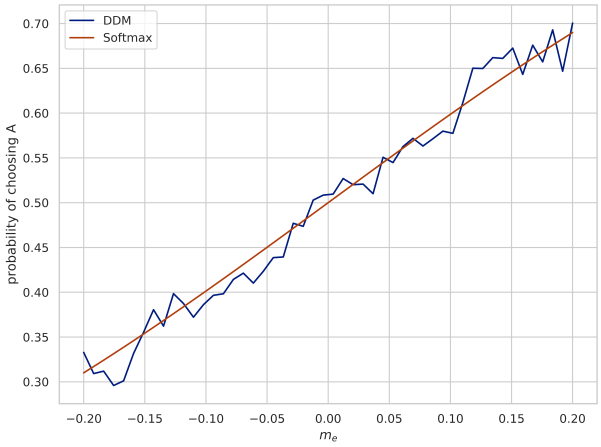

Figure 13: Probability plot for choosing A depending on the m_e (difference between m_A and m_B). The blue line represents the drift-diffusion model and the red line the Softmax-Gibbs model.

In **Figure 13** can be observed that both models predict similar probabilities for each difference in reward. However, the drift-diffusion model includes more noise to the model.

3.d Conclusions of problem 3

In conclusion, the DDM is a more realistic model, it is good for calculating RTs and probabilities of choosing a response. However, it is computationally much more expensive than the Softmax model for the latter purpose.

4 PROBLEM 4: Reinforcement learning in a maze

This problem consists of applying Reinforcement Learning algorithms to model a rat behavior in a maze. This maze is composed of 8 different states: A, B, C, D, E, F, G, H. The rat enters the maze through A and leaves it though H. Then, depending on the intermediate states to which the rat is going, different rewards will receive.

4.a

First, it was model a rat that would go to each state at random with a 50% chance of choosing each direction in the maze. A total of 100 trials were performed.

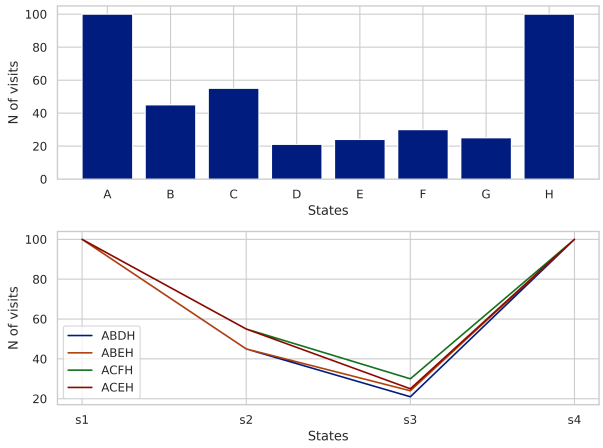

Figure 14: (upper) Bar plot representing the times each state was visited. (lower) each line represents the number of times that the rat was at each state, order by the path that it followed.

In **Figure 14**, it can be observed that the number of the visit of states A and H is 100, as always the rat needs to go through these states. Next, the percentages for B and C are close to 50%, and the percentages for the D, E, F, and G are approximately 25%.

4.b

Afterwatds, the TD-Learning algorithm was applied,which is based on the the Rescorla-Wagner model (Equation 1), in order to model the predictions of the reward that the rat is learning while going through the maze each time. The equation used was:

$$V(s_t) \rightarrow V(s_t) + \epsilon[r(s_t) + V(s_t + 1) - V(s_t)] \tag{4}$$

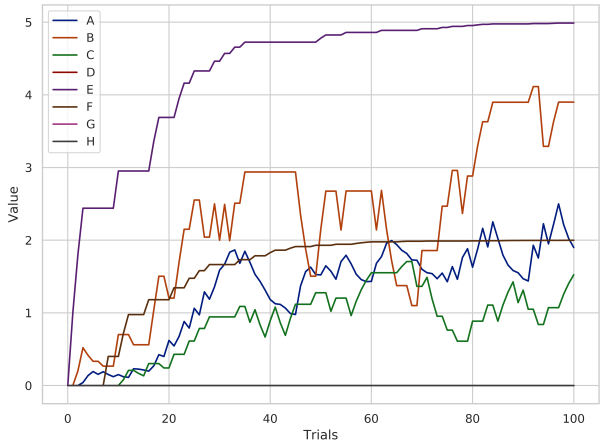

Figure 15: Each line represents the change in the prediction of the reward across all trials for each different state

Based on **Figure 15** it is observed that the prediction value for the states it learned faster when closer to the reward, and when it is farther away the prediction curve is noisier. This indicates that the temporal distance between the stimuli and the reward plays an important role in the learning of the future reward.

4.c

Finally, it was added to the model that the choice instead of being random depends on the probability obtained through the Softmax-Gibbs policy. Here, the rat makes choices based on its knowledge of the value of each state. Two different β values were used 0.1 and 1.

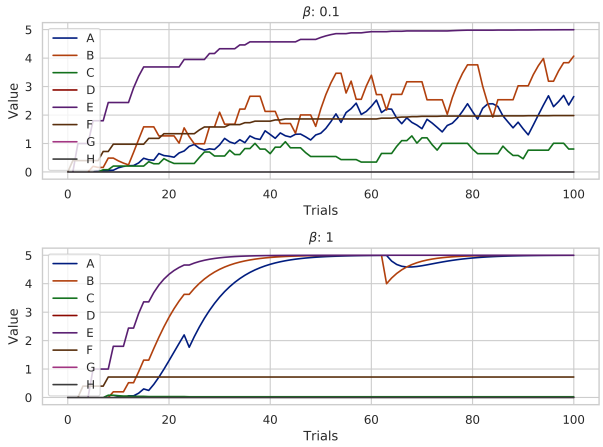

Figure 16: Each line represents the change in the prediction of the reward across all trials for each different state. (upper) The softmax model was applied with $\beta = 0.1$, (lower) the softmax model was applied with $\beta = 1$

It can be observed in **Figure 16** the rat behavior when $\beta = 0.1$ (upper plot) is pretty much similar to the results obtained in **Figure 15**. This supports the idea that a lower β will make the rat more explorative, thus, their choices will be closer to randomness. On the other hand, when $\beta = 1$ the learning is faster and rapidly starts exploiting the better rewarding states and does not even finish learning the prediction of the value from the other states.

4.d Conclusions problem 4

In conclusion, when combined the Softmax policy with TD-Learning a good model of the behavior of a rat in a maze can be performed. One of the problems that can be observed from the previous results when the β is high, is that as soon as the rat finds a good rewarding state will exploit it and will not explore even if this will make the rat find an even better state.