Temporal difference learning

Marc Vöhringer Carrera

December 2, 2023

1 Introduction

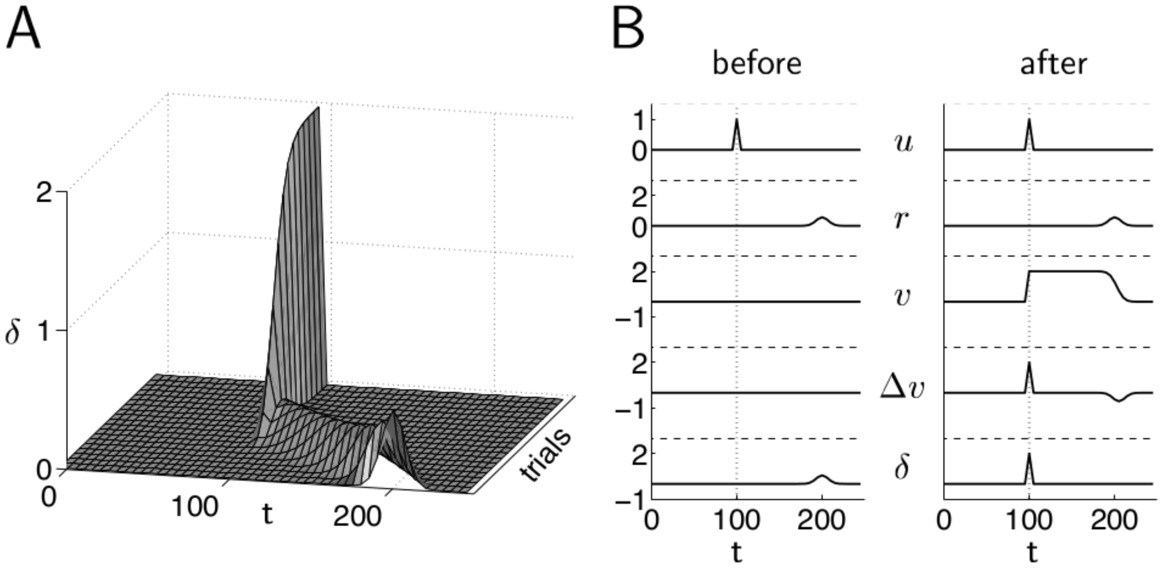

Figure 1: Learning to predict a reward.(A) The surface plot shows the prediction error $\delta(t)$ as a function of time within a trial, across trials. In the early trials, the peak error occurs at the time of the reward ($t = 200$), while in later trials it occurs at the time of the stimulus ($t = 100$). (B) The rows show the stimulus $u(t)$, the reward $r(t)$, the prediction $v(t)$, the temporal difference between predictions $\Delta v(t-1) = v(t) - v(t-1)$, and the full temporal difference error $\delta(t-1) = r(t-1) + \Delta v(t-1)$. The reward is presented over a short interval, and the prediction v sums the total reward. The left column shows the behavior before training, and the right column, after training. $\Delta v(t-1)$ and $\delta(t-1)$ are plotted instead of $\Delta v(t)$ and $\delta(t)$ because the latter quantities cannot be computed until time $t+1$, when $v(t+1)$ is available. [1]

Fig. 1A shows the prediction error δ in temporal difference learning as a function of time and trials ina a Pavliovian experimental paradigm. As one can observe, the prediction error δ before the first trial starts off as a Gaussian function at time $t = 200$, corresponding to the time at which the reward is delivered. The responses of the DA cells correlate with the prediction error and they can be measured to have a Gaussian shape in time when the reward is delivered, before the conditioned response is established. That is why the reward function $r(t)$ takes a Gaussian shape in this model. As the conditioned response is learned and trials progress, the Gaussian peak in the prediction error δ moves to earlier times t, decreases in amplitude and increases in width. At some point, when the center of the Gaussian reaches the time at which the conditioned stimulus is presented ($t = 100$), the Gaussian begins to fade away and an impulse function

$$i(t) = \begin{cases} h & t = 100 \\ 0 & t \neq 100 \end{cases} \tag{1}$$

with a maximum amplitude of $h = 2$ at $t = 100$ rapidly but continuously emerges.

Fig. 1B shows the conditioned stimulus u, reward r, prediction v, temporal difference between predictions Δv and prediction error δ as a function of time before and after the conditioned response has been fully established.

In this report, temporal difference learning is explored under variation of several parameters, such as the learning rate, the shape (width) of the time dependent reward function (corresponding to the response of the DA cells before learning is established), the timing between the conditioned stimulus and the reward, as well as probability and number of rewards. The focus of this investigation lies on the prediction error $\delta(t)$ and the conditioned response, corresponding to the integral of the prediction $v(t)$ over time.

2 Replication of the plot

First, the plot in Fig. 1A is replicated. To do this, the temporal difference learning algorithm is implemented in a python jupyternotebook [2]. The plot is replicated with the following parameters: learning rate = 0.2, 800 trials, a standard deviation of the Gaussian reward function of $\sigma_r = 2$. The resulting plot is displayed in Fig. 2.

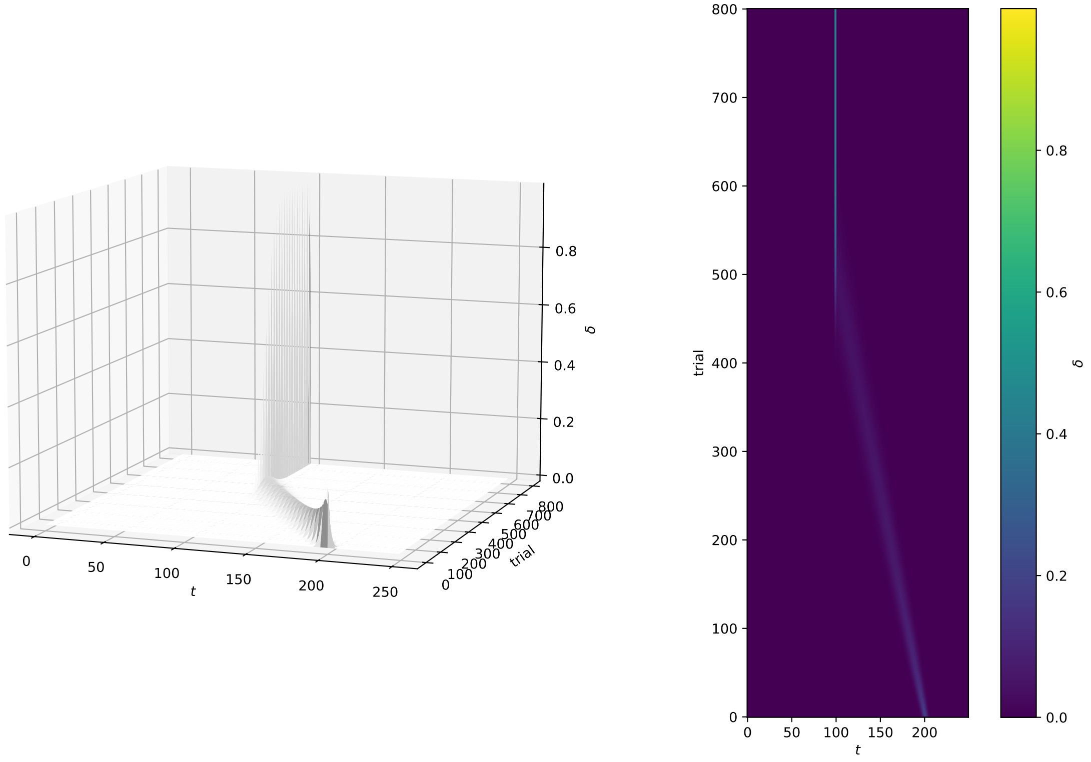

Figure 2: Replication of Fig. 1A.

3 Influence of the learning rate on the prediction error

To investigate the influence of various parameters on the prediction error δ, it is analyzed by fitting the superposition of a Gaussian function and in impulse function over time for each trial. The results of these fits for varying learning rate are shown in Fig. 3.

From the plot it is apparent that the amplitude h of the impulse function (the peak at $t = 100$) increases over trials like a logistic function. The larger the learning rate, the sooner the amplitude reaches its maximum value. Additionally, the width of the logistic function decreases with the learning rate, meaning that the growth of the amplitude for larger learning rates is more abrupt.

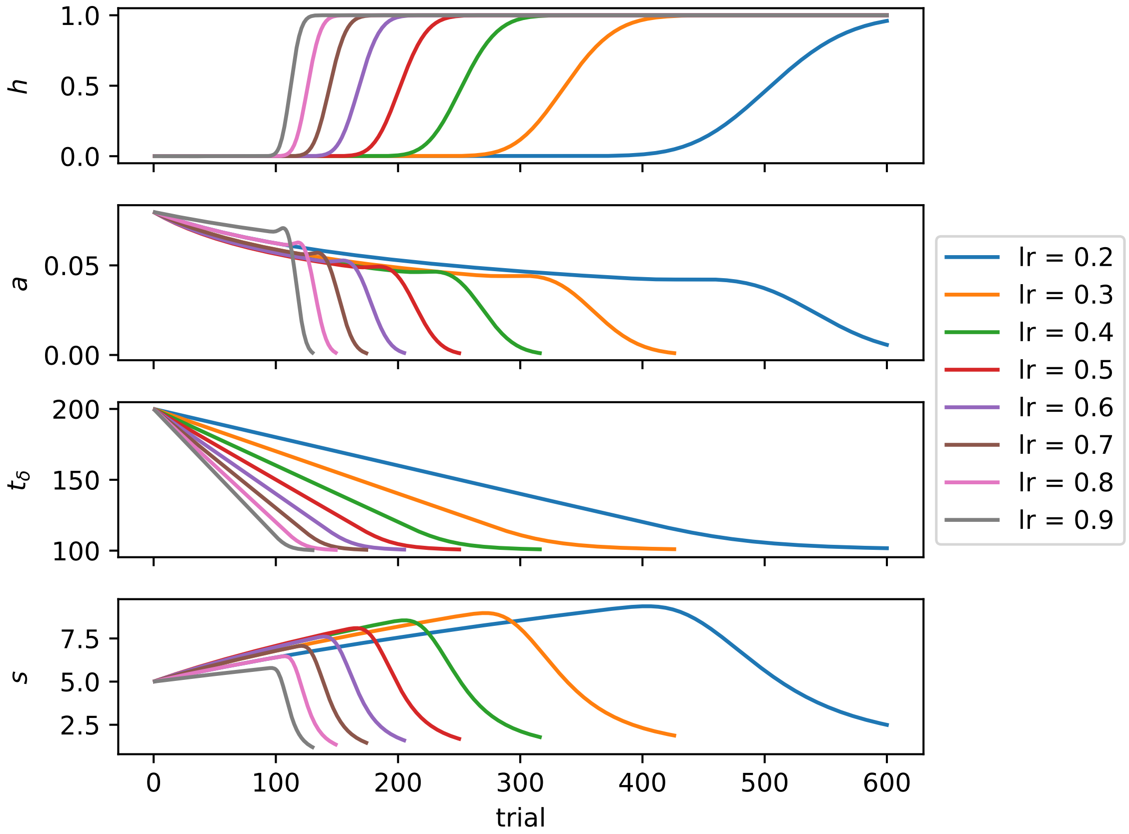

Figure 3: Results of the fits to the prediction error δ for varying learning rates.

The Gaussian travelling from right to left in time has amplitude a, mean value t_δ and standard deviation s. As can be seen in the plot the Gaussian travels faster from right to left in time over trials if the learning rate is larger. At the time where the amplitude a reaches zero and the mean value t_δ reaches 100, the amplitude h of the impulse peak reaches its maximum value. The width s of the Gaussian initially increases and starts decreasing again as soon as the amplitude h of the impulse peak starts growing.

4 Stochastic rewards

For stochastic rewards, the reward is only delivered with probability p in each trial and is zero otherwise. To investigate the influence of this paradigm, we will look at the conditioned response, which is

$$\text{conditioned response} \propto \int_0^T \mathrm{d}t v(t). \tag{2}$$

At a learning rate of 0.5 and with $\sigma_r = 5$, the conditioned response takes the form displayed in Fig. 4. It is apparent that the normalized conditioned response first increases linearly until it reaches its steady-state value p and then fluctuates around that value (except for $p = 1$ where it stays constant). The fluctuations are described by a random walk and when increasing the learning rate one can observe that the fluctuations also increase.

5 Unprecise timing

In real experiments, the timing between the stimulus and the reward is always prone to an error. To simulate unprecise timing, the mean of the reward function in time is randomly distributed across

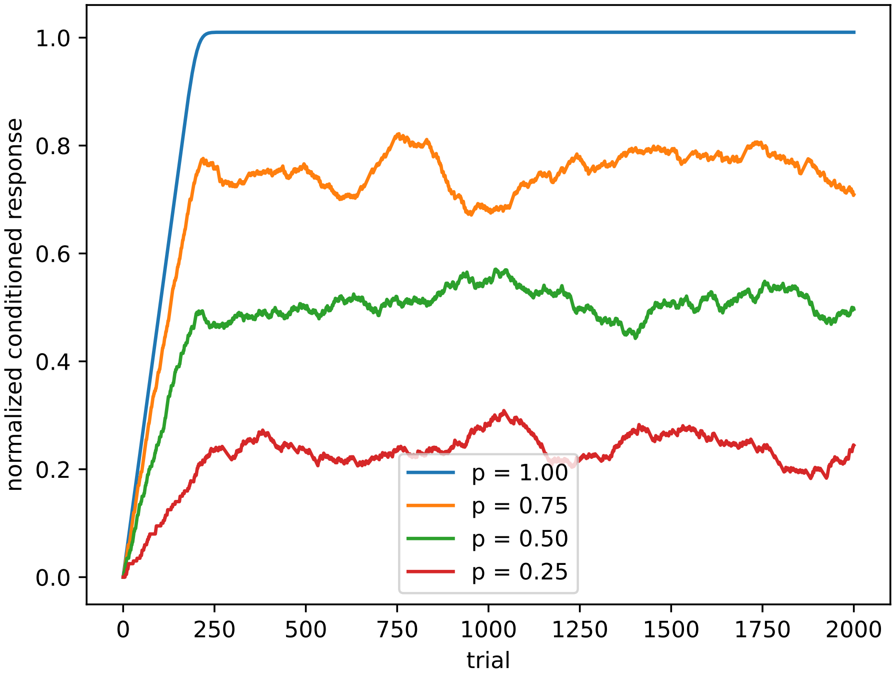

Figure 4: Stochastic rewards.

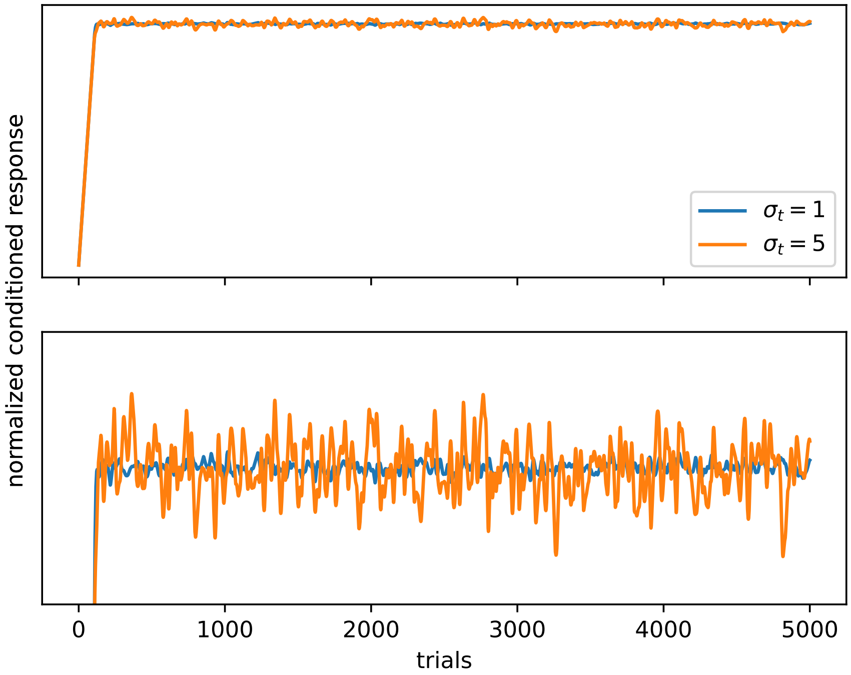

Figure 5: Unprecise timing of the reward. The plots show the conditioned response as a function of trials at different y-axis scales.

trials. This is done by sampling from a normal distribution with a width of σ_t. The results of this simulation is shown in Fig. 5. As one can see the conditioned response fluctuates around its steady-state value much like in the case of stochastic rewards. If the precision of the time interval between stimulus and reward increases (decreasing σ_t), the fluctuations decrease.

6 Multiple rewards

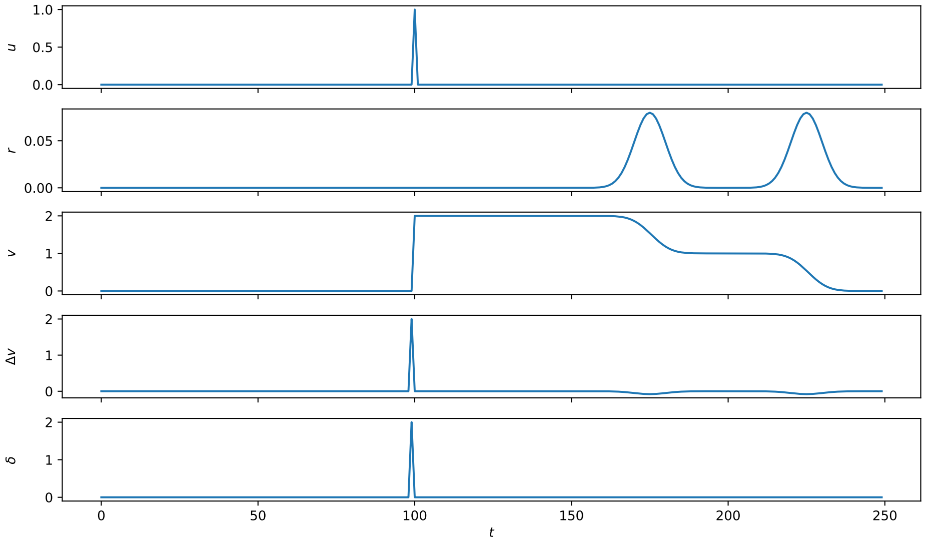

Figure 6: Functions describing the state of the system after learning for multiple rewards.

If multiple rewards are administered, the shape of the functions describing the state of the system change. This is well visualized in Fig. 6, which can be compared to Fig. 1B.

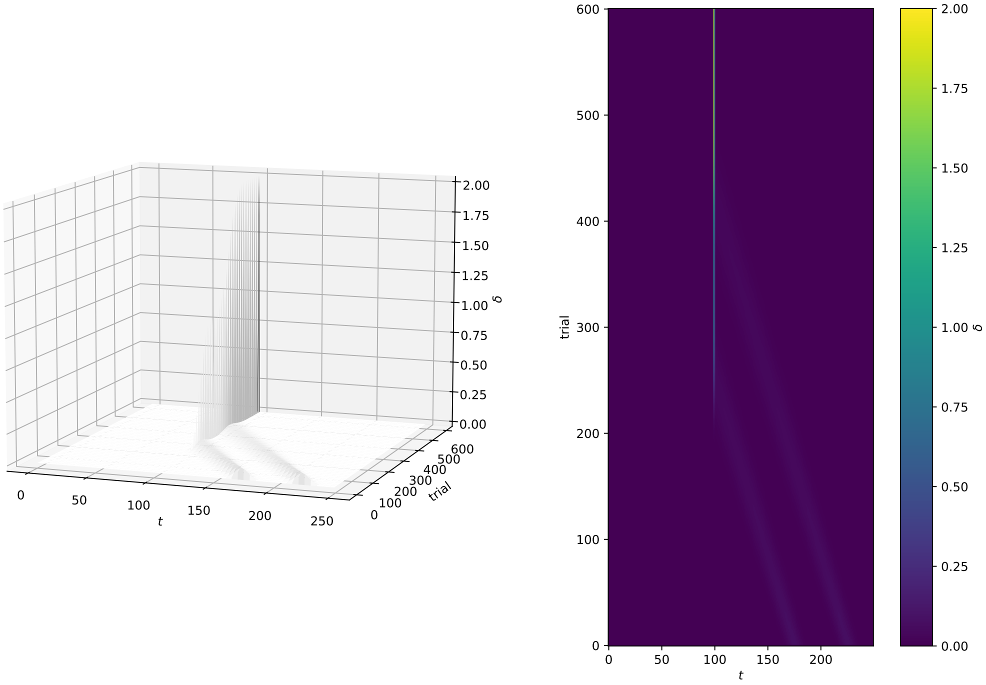

Figure 7: Behavior of δ for multiple rewards.

Fig. 7 depicts how the prediction error δ behaves when multiple rewards are present. One can observe that the impulse peak initially increases to one when the first Gaussian reaches $t = 100$, and afterwards further increases to an amplitude of two when the second Gaussian reaches it.

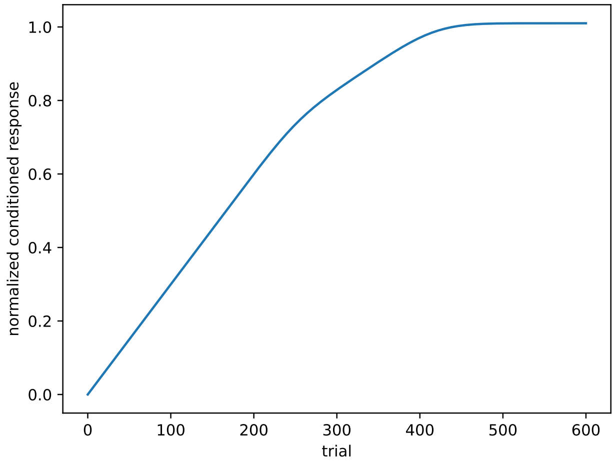

Figure 8: Conditioned response as a function of trial number for a multiple reward paradigm.

The conditioned response also behaves differently when multiple rewards are present. Its course over progressing trials is shown in Fig. 8, where one observes an initial linear increase, which reduces its steepness as soon as the first Gaussian from the prediction error δ fully transformed into the impulse peak, since only contributions from the second Gaussian now add to the increase of the conditioned response.

7 Conclusion

When playing around with the parameters of the simulations, the following observations were made:

- For precise/constant timing between stimulus and reward and $p = 1$
  - the prediction error δ consists of an impulse peak at $t = 100$ plus a Gaussian travelling backwards in time over the course of trials.
  - the travelling speed of the Gaussian increases with the learning rate.
  - the impulse peak appears when the Gaussian reaches $t = 100$.
  - the impulse peaks amplitude h as a function of trial number increases in a logistic manner, while the width of the logistic function decreases with the learning rate.
  - the impulse peaks amplitude h takes the value of the integral over the reward function.
- For precise/constant timing between stimulus and reward and stochastic rewards with probability p the normalized conditioned response
  - increases linearly across trials with some fluctuations (if $p < 1$) until it reaches its steady-state expectation value, corresponding to p.
  - fluctuates around p, while the fluctuations are described by a random walk and increase with the learning rate.
- For unprecise/fluctuating timing between stimulus and reward and $p = 1$ the conditioned response fluctuates once the steady-state is reached. The fluctuations decrease with the precision of the timing.

- For multiple rewards two Gaussian functions appear in the prediction error δ, which both eventually transform into the impulse peak at the time of the stimulus. The conditioned response increases linearly until the first Gaussian fully transformed into the impulse peak and afterwards decreases its rate of change by a factor of two, since only the second Gaussian contributes.

References

[1] Peter Dayan and Laurence Abbott. *Theoretical neuroscience*. 2001.

[2] Marc Vöhringer Carrera. Github repository. https://github.com/mavoeh/RL_hw1, 2023. [Online; accessed 02-December-2023].

8 Appendix

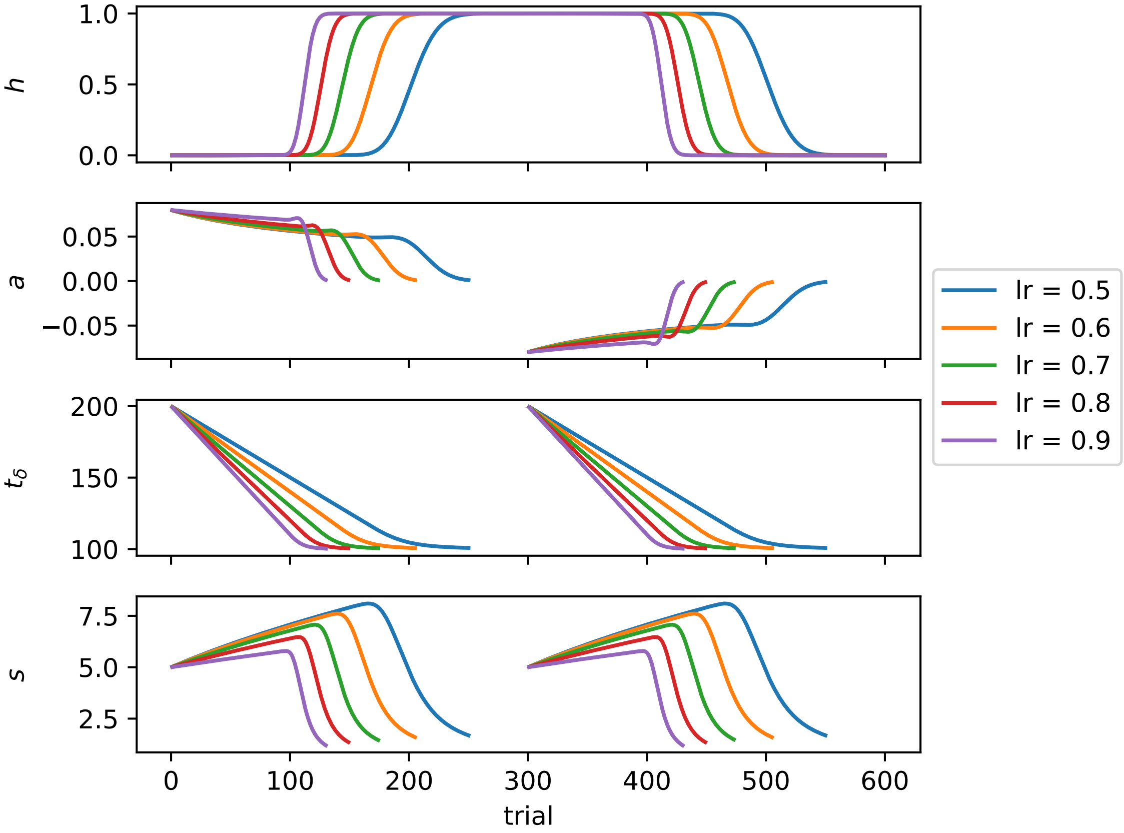

Figure 9: Fit parameters to the prediction error δ for experimental paradigm with extinction. A reward is delivered only throughout trials 0-300, afterwards no reward is delivered anymore.

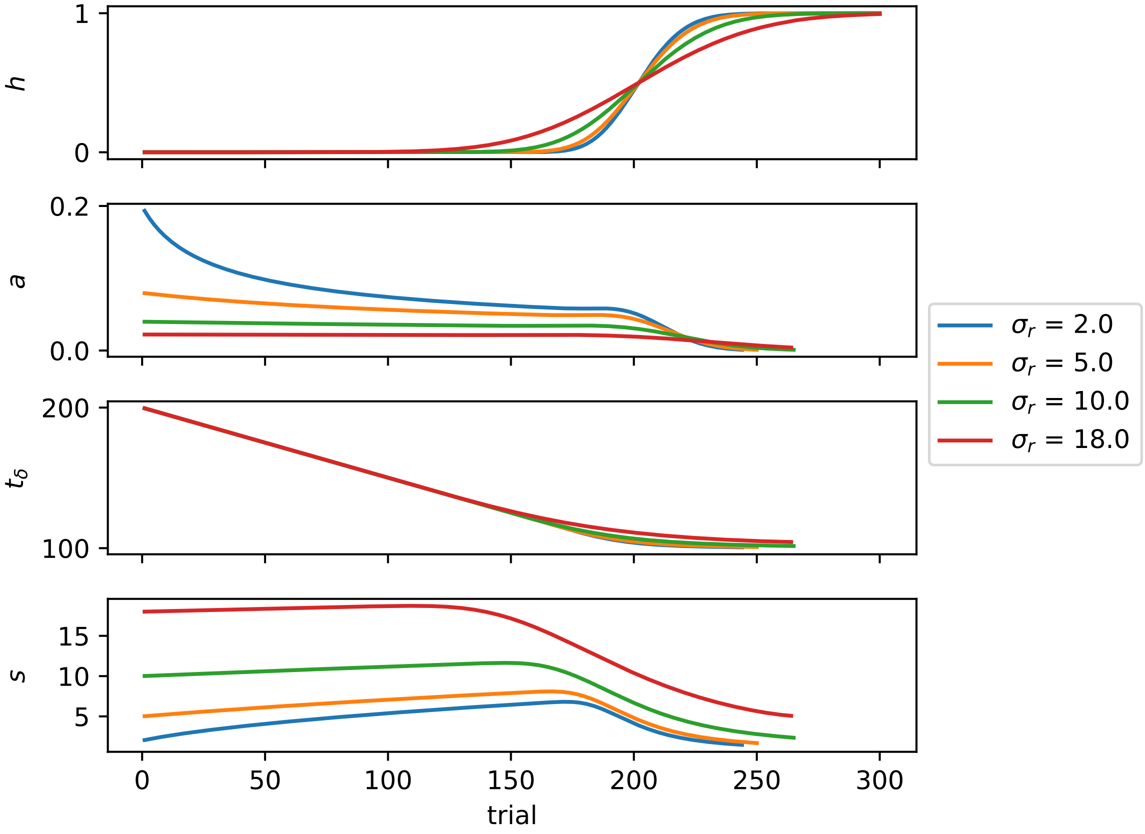

Figure 10: Fit parameters to the prediction error δ for varying width of the reward function σ_r.