Cognitive Neuroscience II

Prof. Dr. Andreas Wendemuth

Lehrstuhl Kognitive Systeme

Institut für Elektronik, Signalverarbeitung und Kommunikationstechnik

Fakultät für Elektrotechnik und Informationstechnik
Otto-von-Guericke Universität Magdeburg

http://iesk.et.uni-magdeburg.de/ko/

Lecture 12

- Classical conditioning
  - Rescorla Wagner Rule
  - Temporal Difference Learning

Classical Conditioning

- Classical: Reinforcers delivered independently of actions taken by the animal

- Stimulus u
- Expected reward r, R
- Weight w
- Predicted reward v

Rescorla – Wagner rule (1972)

- Ansatz (multiple stimuli **u**): $v = \mathbf{w} * \mathbf{u}$
- Minimize square error (stochastic gradient)
  $\langle r - \mathbf{w} * \mathbf{u} \rangle^2$
- Rule: $\Delta \mathbf{w} = \varepsilon \delta \mathbf{u}$ with $\delta = r - v$.
  Has form of a delta-rule (same derivation)
- If steps are small, solution like differential eq.:

$$\tau_w \frac{d\mathbf{w}}{dt} = \delta \mathbf{u} = r\mathbf{u} - \mathbf{C} * \mathbf{w} \quad \text{with} \quad \mathbf{C} = \mathbf{u}\mathbf{u}^T$$

Ex 1 (Similar to Cha. 6 / Ex 2):

- Show that the differential Rescorla-Wagner rule

$$\tau_w \frac{d\mathbf{w}}{dt} = \delta\mathbf{u} = r\mathbf{u} - \mathbf{u}\mathbf{u}^T * \mathbf{w}$$

has the solution

$$\mathbf{w}(t) = \mathbf{u}_\perp + \mathbf{u}\left[\frac{r}{|\mathbf{u}|^2} - a\exp\left(-\frac{|\mathbf{u}|^2}{\tau_w}t\right)\right]$$

where the constant scalar *a* and the constant vector $\mathbf{u}_\perp$ (perpendicular to **u**) have to be chosen according to boundary conditions.

- Solve for *a* and $\mathbf{u}_\perp$:
  1) at t=0, with **w**(t=0) = **0**,
     then evolve for t=0...t_1, with r=1
  2) new *a* for t >t_1, with r = 0, then evolve
- Compare with fig.9.1 (following page)

Behaviour

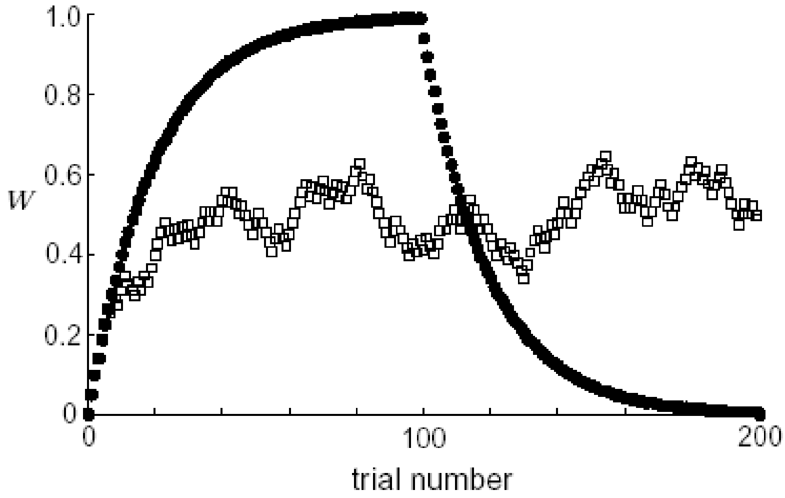

Figure 9.1: Acquisition and extinction curves for Pavlovian conditioning and partial reinforcement as predicted by the Rescorla-Wagner model. The filled circles show the time evolution of the weight w over 200 trials. In the first 100 trials, a reward of $r = 1$ was paired with the stimulus, while in trials 100-200 no reward was paired ($r = 0$). Open squares show the evolution of the weights when a reward of $r = 1$ was paired with the stimulus randomly on 50% of the trials. In both cases, $\epsilon = 0.05$.

Other features

- *Blocking*: Second stimulus is blocked by a trained first one (no δ)
- *Inhibitory conditioning*: Second stimulus is presented together with trained first only in absence of reward, inhibits a trained first one ($0 < w_1 = -w_2$)
- *Overshadowing*: Sharing reward between two stimuli ($0 < w_1 = w_2$)

Full list of other features

Paradigm	Pre-Train	Train	Result
Pavlovian		$s \to r$	$s \to$ ‘r’
Extinction	$s \to r$	$s \to \cdot$	$s \to$ ‘$\cdot$’
Partial		$s \to r$ $\quad s \to \cdot$	$s \to \alpha$‘r’
Blocking	$s_1 \to r$	$s_1 + s_2 \to r$	$s_1 \to$ ‘r’ $\quad s_2 \to$ ‘$\cdot$’
Inhibitory		$s_1 + s_2 \to \cdot$ $\quad s_1 \to r$	$s_1 \to$ ‘r’ $\quad s_2 \to -$‘r’
Overshadow		$s_1 + s_2 \to r$	$s_1 \to \alpha_1$‘r’ $\quad s_2 \to \alpha_2$‘r’
Secondary	$s_1 \to r$	$s_2 \to s_1$	$s_2 \to$ ‘r’

Table 9.1: Classical conditioning paradigms. The columns indicate the training procedures and results, with some paradigms requiring a pre-training as well as a training period. Both training and pre-training periods consist of a moderate number of training trials. The arrows represent an association between one or two stimuli (s, or s_1 and s_2) and either a reward (r) or the absence of a reward ($\cdot$). In Partial and Inhibitory conditioning, the two types of training trials that are indicated are alternated. In the Result column, the arrows represent an association between a stimulus and the expectation of a reward (‘r’) or no reward (‘$\cdot$’). The factors of α denote a partial or weakened expectation, and the minus sign indicates the suppression of an expectation of reward.

Ex 2

- For the list of features on the previous page, show which (all?) of these can be explained by the Rescorla-Wagner rule with 2 weights.

  [Hint: follow the examples which were given]

Temporal Difference Learning

- Total trial time T
- Predicting Future Reward $R(t) = \left\langle \sum_{\tau=0}^{T-t} r(t+\tau) \right\rangle$
- Stimuli u over a range of time are weighted: (Sutton and Barto 1990)

$$v(t) = \sum_{\tau=0}^{t} w(\tau) u(t-\tau)$$

Rule derivation (Dayan 1992)

- Error function:

$$< R(t) - v(t) >^2 = < \sum_{\tau=0}^{T-t} r(t+\tau) - \sum_{\tau=0}^{t} w(\tau)u(t-\tau) >^2$$

- Stochastic gradient

$$\frac{\partial < R(t) - v(t) >^2}{\partial w(\alpha)} = < \sum_{\tau=0}^{T-t} r(t+\tau) - \sum_{\tau=0}^{t} w(\tau)u(t-\tau) > * u(t-\alpha)$$

- Rule $\Delta \mathbf{w}(\tau) = \varepsilon \delta(t) \mathbf{u}(t-\tau)$; $\delta(t) = < \sum_{\tau=0}^{T-t} r(t+\tau) > - v(t)$

Introducing temporal difference

- Have $\delta(t) = <\sum_{\tau=0}^{T-t} r(t+\tau)> - v(t)$ where

$$<\sum_{\tau=0}^{T-t} r(t+\tau)> = r(t) + <\sum_{\tau=0}^{T-(t+1)} r((t+1)+\tau)> = r(t) + v(t+1)$$

  i.e. prediction is used in formula again.
- Hence prediction error $\delta(t) = r(t) + v(t+1) - v(t)$ where $\Delta v(t) = v(t+1) - v(t)$ is called the *temporal difference term*.
- Allows to predict future rewards.

Example(1): u(t=100) = 1, r(t=200)=1

- Need to learn R(t):
  0 (t<100), 1 (t=100..200), 0 (t>200)
- First trial: w(t) = 0, v(t) = 0. Hence δ(t=200)=1 and $\Delta\mathbf{w}(\tau) = \varepsilon\delta(t)\mathbf{u}(t-\tau) = \varepsilon\mathbf{u}(200-\tau) = \varepsilon[\tau = 100]$
- So we have the first predictor w(100)=ε and
  $$v_1(t) = \sum_{\tau=0}^{t} w(\tau)u(t-\tau) = \varepsilon u(t-100)$$
- This predicts only at t=200.

Example (2): u(t=100) = 1, r(t=200)=1

$$v_1(t)=\sum_{\tau=0}^{t} w(\tau)u(t-\tau) = \varepsilon u(t-100)$$

- Second trial: $\delta(199) = r(199) + v(200) - v(199) =$

  t=199: $0 + \varepsilon u(100) - \varepsilon u(99) = \varepsilon$

$$\Delta\mathbf{w}(\tau) = \varepsilon\delta(199)\mathbf{u}(199-\tau) = \varepsilon^2\mathbf{u}(199-\tau) = \varepsilon^2[\tau = 99]$$

  t=200: $\delta(200) = r(200) + v(201) - v(200) =$

$$1 + \varepsilon u(101) - \varepsilon u(100) = 1 - \varepsilon$$

$$\Delta\mathbf{w}(\tau) = \varepsilon\delta(200)\mathbf{u}(200-\tau) = \varepsilon(1-\varepsilon)\mathbf{u}(200-\tau) = \varepsilon(1-\varepsilon)[\tau = 100]$$

- Predictor: $v_2(t)=\sum_{\tau=0}^{t} w(\tau)u(t-\tau) = \varepsilon(2-\varepsilon)u(t-100) + \varepsilon^2 u(t-99)$

Example (3): u(t=100) = 1, r(t=200)=1

$$v_2(t) = \sum_{\tau=0}^{t} w(\tau)u(t-\tau) = \varepsilon(2-\varepsilon)u(t-100) + \varepsilon^2 u(t-99)$$

starts predicting 1 step earlier, at t=199

- Third trial: t=198: $\delta(198) = r(198) + v(199) - v(198) = \varepsilon^2$

$$\Delta\mathbf{w}(\tau) = \varepsilon\delta(198)\mathbf{u}(198-\tau) = \varepsilon^3[\tau = 98]$$

t=199: $\delta(199) = r(199) + v(200) - v(199) = \varepsilon(2-\varepsilon) - \varepsilon^2$

$$\Delta\mathbf{w}(\tau) = \varepsilon\delta(199)\mathbf{u}(199-\tau) = 2\varepsilon^2(1-\varepsilon)[\tau = 99]$$

t=200: $\delta(200) = r(200) + v(201) - v(200) = 1 - \varepsilon(2-\varepsilon)$

$$\Delta\mathbf{w}(\tau) = \varepsilon\delta(200)\mathbf{u}(200-\tau) = \varepsilon - \varepsilon^2(2-\varepsilon)[\tau = 100]$$

- P.: $v_3(t) = \varepsilon(3 - 3\varepsilon + \varepsilon^2)u(t-100) + \varepsilon^2(3-\varepsilon)u(t-99) + \varepsilon^3 u(t-98)$

Temporal difference conclusion

- With every trial, the prediction starts 1 step *earlier*.
- The prediction *mass* also moves to earlier times. This happens the faster, the larger ε.
- Example: ε =1:

$$v_1(t) = u(t-100)$$

$$v_2(t) = u(t-100) + u(t-99)$$

$$v_3(t) = u(t-100) + 2u(t-99) + u(t-98)$$

Ex 3

- Study temporal difference learning with a program.
- Explain the behaviour of figure 9.2 (next page)

Effects of temporal difference learning

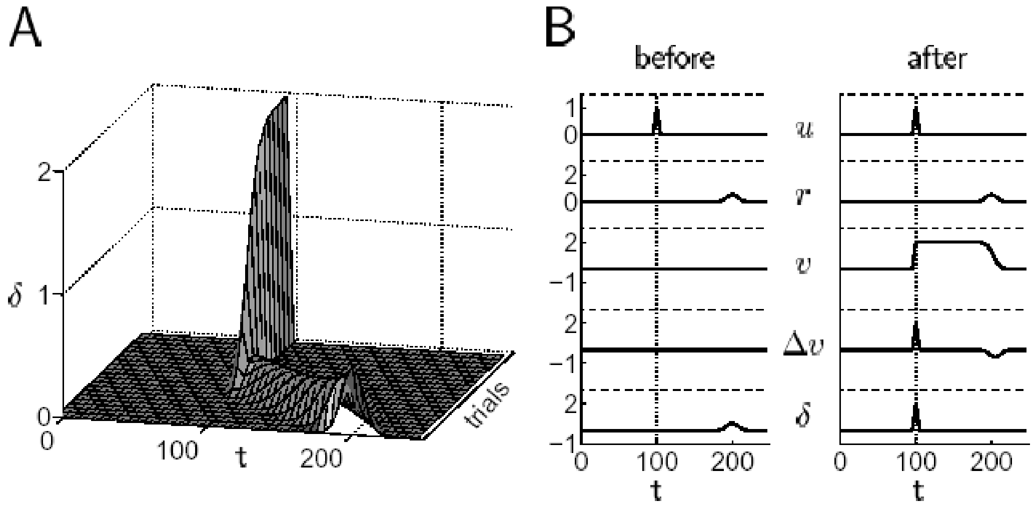

Figure 9.2: Learning to predict a reward. A) The surface plot shows the prediction error $\delta(t)$ as a function of time within a trial, across trials. In the early trials, the peak error occurs at the time of the reward ($t = 200$), while in later trials it occurs at the time of the stimulus ($t = 100$). (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.

Next lecture: Instrumental Conditioning

- Actions of the animal determines which reinforcement is provided