Reinforcement Learning Lab

Odelia Schwartz,

2020

Rescorla-Wagner rule (1972)

- Minimize difference between received reward and predicted reward
- Binary variable u (1 if stimulus is present; 0 if absent)
- Predicted reward v
- Linear weight w

$$v = wu$$

- If stimulus u is present:

$$v = w$$

based on Dayan and Abbott book

Rescorla-Wagner rule (1972)

- Minimize squared error between received reward r and predicted reward v:

  $$(r-v)^2$$

(average over presentations of stimulus and reward)

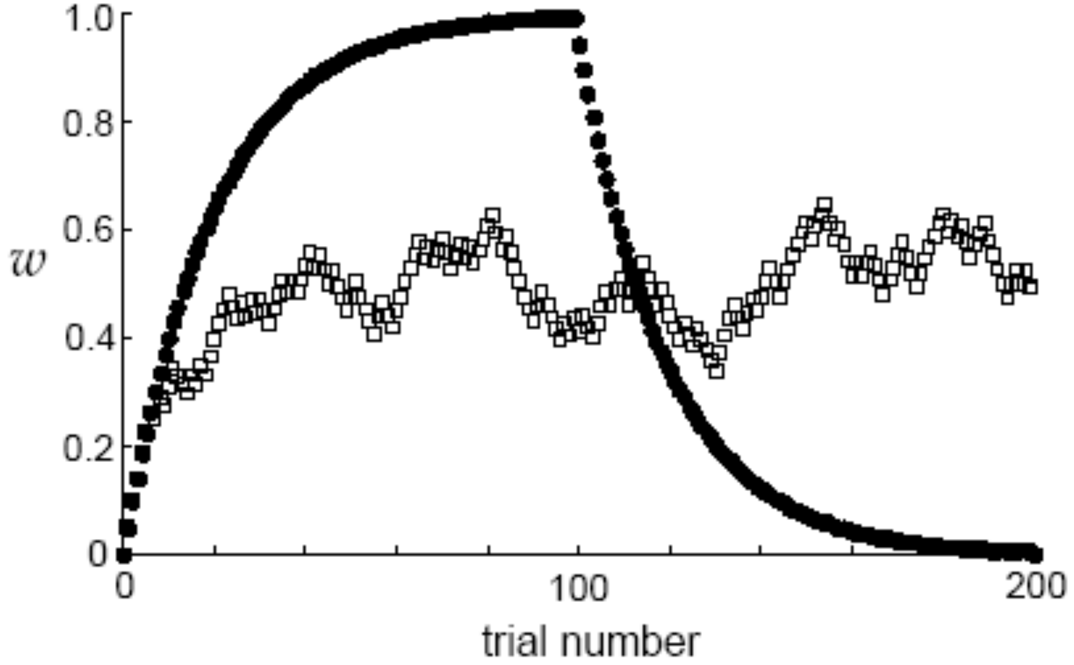
- Update weight:

  $$w \longrightarrow w+\varepsilon(r-v)u$$

  ε learning rate

Also known as delta learning rule: $\delta = r - v$

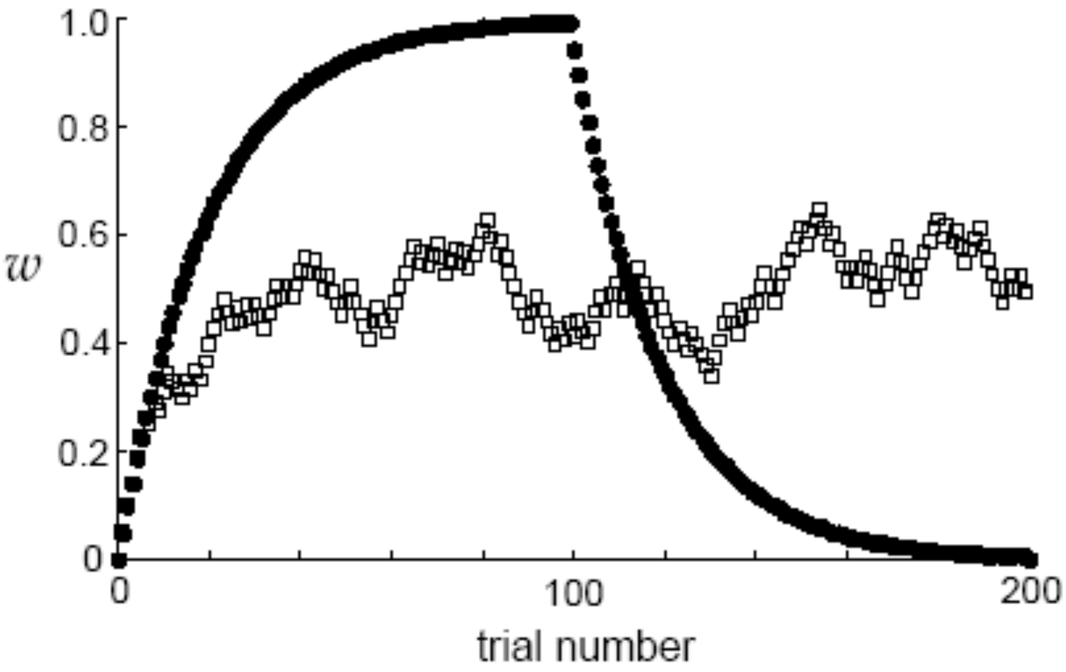
Acquisition and extinction



- Solid: First 100 trials: reward (r=1) paired with stimulus; next 100 trials no reward (r=0) paired with stimulus (learning rate .05)
- Dashed: Reward paired with stimulus randomly 50 percent of time

From Dayan and Abbott book

Acquisition and extinction



Code in lab
Produces solid and
we will also generate
dashed

- Solid: First 100 trials: reward (r=1) paired with stimulus; next 100 trials no reward (r=0) paired with stimulus (learning rate .05)
- Dashed: Reward paired with stimulus randomly 50 percent of time

From Dayan and Abbott book

Temporal Difference Learning

Want $v_t = r_t + r_{t+1} + r_{t+2} + r_{t+3} \ldots.$

(here t represents time within a trial; reward can come at any time within a trial. Sutton and Barto interpret v_t as the prediction of total future reward expected from time t onward until the end of the trial)

Based on Dayan slides; Daw slides

Temporal Difference Learning

Want $v_t = r_t + r_{t+1} + r_{t+2} + r_{t+3} \ldots.$

(here t represents time within a trial; reward can come at any time within a trial. Sutton and Barto interpret v_t as the prediction of total future reward expected from time t onward until the end of the trial)

Prediction error:

$$\delta_t = (r_t + r_{t+1} + r_{t+2} + r_{t+3} \ldots) - V_t$$

Temporal Difference Learning

Want $v_t = r_t + r_{t+1} + r_{t+2} + r_{t+3} \ldots.$

(here t represents time within a trial)

But we don’t want to wait forever for all future rewards…

$$r_{t+1}; r_{t+2}; r_{t+3} \ldots.$$

Temporal Difference Learning

Want $v_t = r_t + r_{t+1} + r_{t+2} + r_{t+3} \dots.$

(here t represents time within a trial)

Recursion "trick": $v_t = r_t + v_{t+1}$

Based on Dayan slides; Daw slides

Temporal Difference Learning

From recursion want:

$$v_t = r_t + v_{t+1}$$

Error:

$$\delta_t = r_t + v_{t+1} - v_t$$

Temporal Difference Learning

From recursion want: $v_t = r_t + v_{t+1}$

Error: $\delta_t = r_t + v_{t+1} - v_t$

Update: $v_t \rightarrow v_t + \varepsilon(r_t + v_{t+1} - v_t)$

$= (1- \varepsilon)v_t + \varepsilon(r_t + v_{t+1})$

RV versus TD

- Rescorla-Wagner error: (n represents trial)

$$\delta_n = r_n - v_n$$

- Temporal Difference Error: (t is time within a trial)

$$\delta_t = r_t + v_{t+1} - v_t$$

Updates are causal

RV versus TD

- Rescorla-Wagner error: (n represents trial)

$$\delta_n = r_n - v_n$$

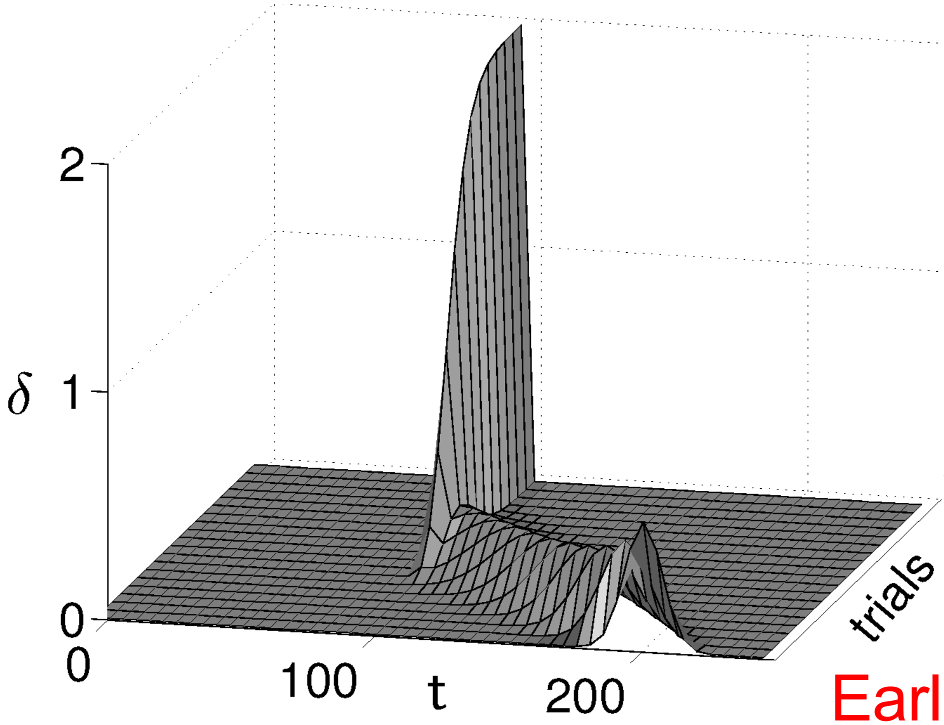
- Temporal Difference Error: (t is time within a trial)

$$\delta_t = r_t + v_{t+1} - v_t$$

We repeat this learning for many trials…

Updates are causal

Temporal Difference Learning

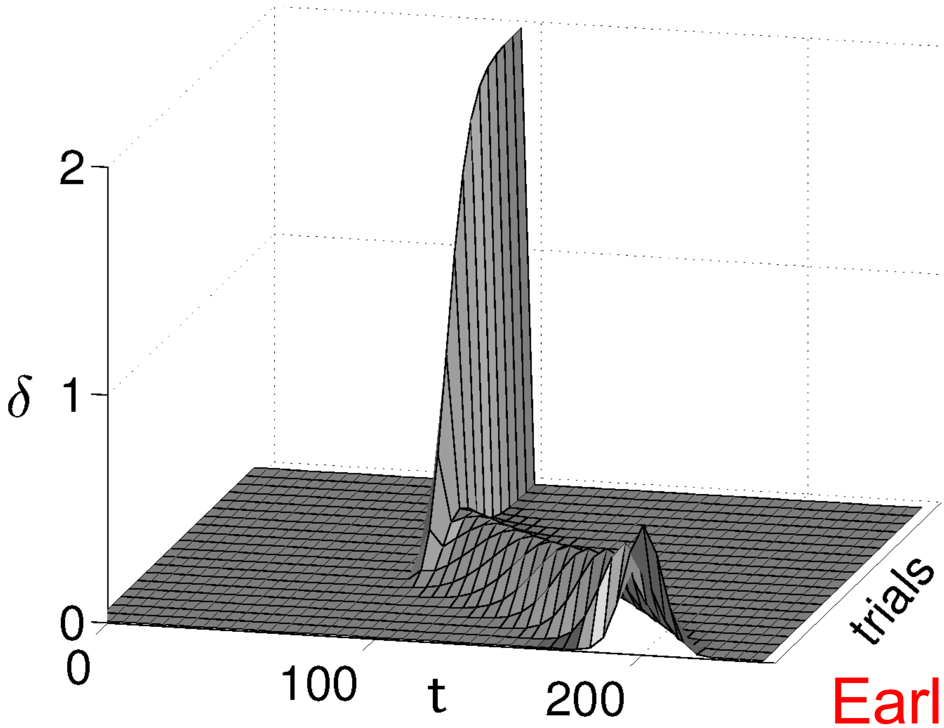


Late trials: error when stimulus

Early trials: error when reward

Dayan and Abbott Book: Surface plot of prediction error (stimulus at 100; reward at 200)

Temporal Difference Learning



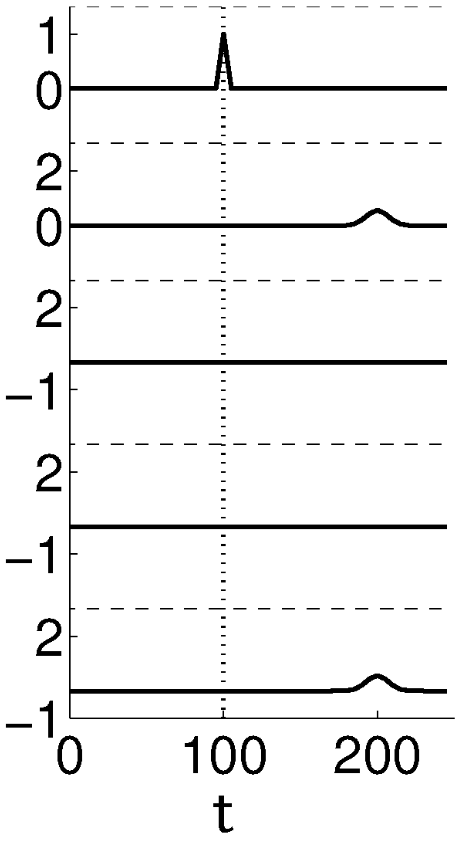
Late trials: error when stimulus

Early trials: error when reward

Dayan and Abbott Book: Surface plot of prediction error (stimulus at 100; reward at 200)

Temporal Difference Learning

Before learning



u_t stimulus

r_t reward

v_t expected future reward

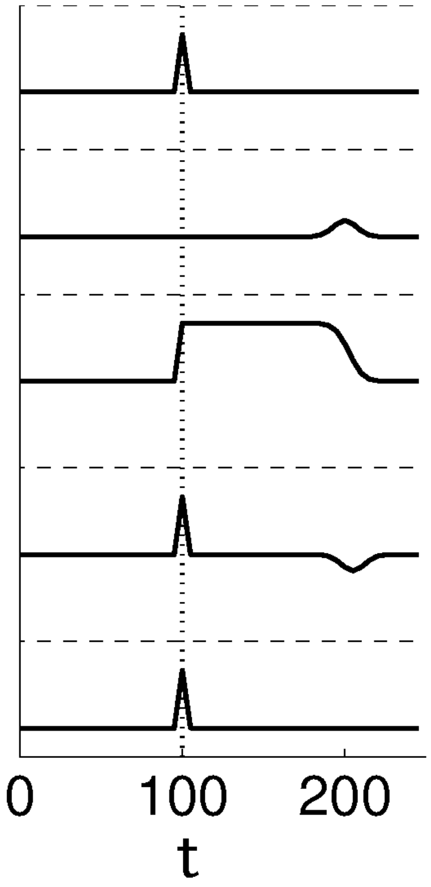
$v_{t+1} - v_t$

$$\delta_t = r_t + v_{t+1} - v_t$$

Temporal difference error

Temporal Difference Learning

After learning



u_t stimulus

r_t reward

v_t expected future reward

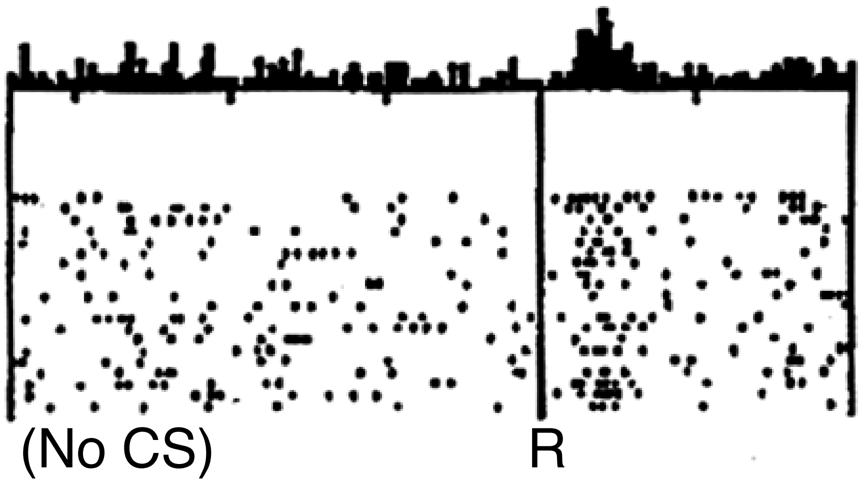
$v_{t+1} - v_t$

$\delta_t = r_t + v_{t+1} - v_t$

Temporal difference error

VTA Activity of dopaminergic neurons

No prediction
Reward occurs



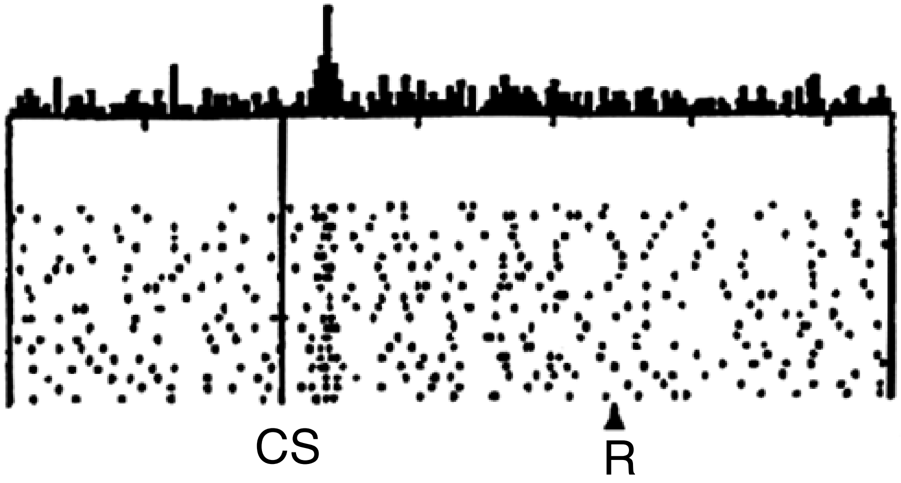
Before learning (= early trials), reward is given in experiment, but animal does not predict (expect) reward (why is there Increased activity after reward?)

Prediction error (and error when reward)

Schultz, Dayan, Montague, 1997

VTA Activity of dopaminergic neurons

Reward predicted
Reward occurs



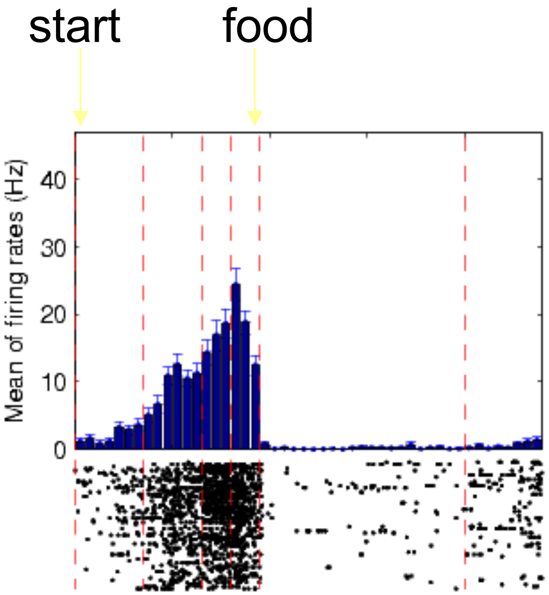
After learning, conditioned stimulus predicts reward, and reward is given in experiment

Prediction error flat when reward but note the error when stimulus presented

Schultz, Dayan, Montague, 1997

Temporal Difference Learning

Striatal neurons (activity that precedes rewards and changes with learning)



(Daw)

What about anticipation of future rewards?
(like the v variable)

From Dayan slides