Reinforcement learning in neuroscience - II

Rita Almeida

PhD course 3045, VT 2018

Karolinska Institutet

Prediction and control

- Prediction: Estimation of value functions given a policy

- Control: Finding an optimal policy

Classical conditioning

Pavlovian paradigm:

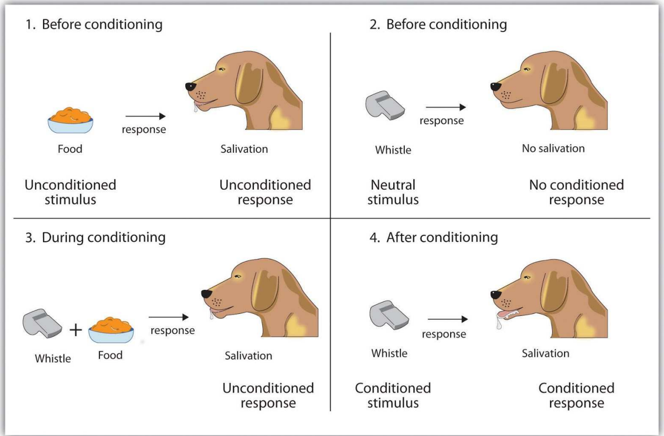


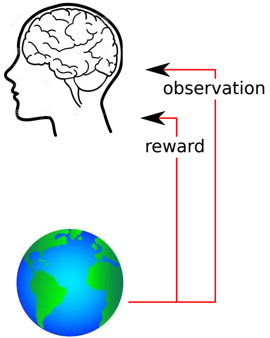
Figure from http://catalog.flatworldknowledge.com

- Does it relate with reinforcement learning?...

Classical conditioning

- How does it relate with reinforcement learning?

  - learning of values of stimuli through experience
  - learning based on reward
  - learning without instruction
  - no acting!



Rescorla-Wagner model of learning

- Learning happens when events are not predicted.

  Change in value is proportional to difference between actual and predicted outcome (prediction error).

For the Pavlovian paradigm:

$$V_{new}(S) = V_{old}(S) + \eta(r - V_{old}(S))$$

- S the conditioned stimuli - sound
- r the unconditioned stimulus - food
- η learning rate

- Predictions due to different stimuli are summed linearly.

$$V_t(S_i) = V_{t-1}(S_i) + \eta(r - \sum_j V_{t-1}(S_j))$$

Rescorla-Wagner model of learning

The Rescorla-Wagner model explains other types of conditioning:

- blocking ($S_1 \to r \quad S_1 + S_2 \to r \quad S_1 \to r' \quad S_2 \to .'$)
- overshadowing ($S_1 + S_2 \to r \quad S_1 \to \alpha_1 r' \quad S_2 \to \alpha_2 r'$)
- inhibitory
  ($S_1 \to r \quad S_1 + S_2 \to . \quad S_3 \to r \quad S_1 \to r' \quad S_3 \to r' \quad S_2 + S_3 \to .'$)

but it also has shortcomings:

- does not account for the sensitivity of conditioning to temporal contingencies,
- it does not explain second order conditioning
  ($S_1 \to r \quad S_2 \to S_1 \to r \quad S_2 \to r'$),
- it does not explain extinction of inhibition

Temporal difference learning

Idea of RL: Maximize total reward, not only immediate reward.

- Predict all future reward.
- Total reward depends on a sequence of choices.

Value of a sate s is the expected future reward. Given $s_t = s$:

$$V(s) = E[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + ...|s_t = s]$$

where $\gamma \leq 1$ discounts the effect of rewards distant in time

And in a recursive form:

$$V(s) = E[r_{t+1}|s_t = s] + \gamma E[V(s_{t+1})|s_t = s].$$

Temporal difference learning

Then, a prediction error can be defined as:

$$\delta = E[r_{t+1}|S_t] + \gamma E[V(S_{t+1})|S_t] - V(S_t).$$

If δ is estimated it can be used in the common update rule:

$$\mathit{NewEstimate} \leftarrow \mathit{OldEstimate} + \mathit{LearningRate}[\mathit{Target} - \mathit{OldEstimate}]$$

Applied to the value:

$V(s) \leftarrow V(s) + \eta \quad \delta$

Temporal difference learning

Problem: to calculate $E[\quad]$ one needs $P(r|s_t = s)$ and $P(s_{t+1}|s_t = s)$.

- This knowledge is usually not available.
- Information can be accumulated by sampling.

The current values r_{t+1} and $V_t(s_{t+1})$ can be used:

$$V(s_t) \leftarrow V(s_t) + \eta(r_{t+1} + \gamma V(s_{t+1}) - V(s_t))$$

And the temporal difference prediction error is:

$$\delta = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

- Over time total expected values of events can be learned even in stochastic environments with unknown dynamics.

Temporal difference learning algorithm

- Initialize V(s) at some value
- Repeat for each episode
  - Initialize s
  - Repeat for each step t
    - do action a given s, according to policy
    - observe reward r and next state s'
    - $V(s) \leftarrow V(s) + \eta(r + \gamma V(s') - V(s))$
    - $s \leftarrow s'$

Dopamine pathways

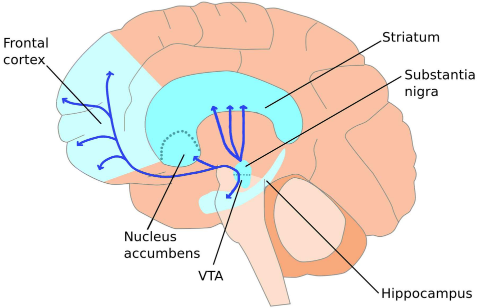


Figure from OIST (www.oist.jp)

Dopamine is thought to have a role in:

- movement control
- reward and motivation
  - substance abuse
- other:
  - working-memory
  - schizophrenia ...

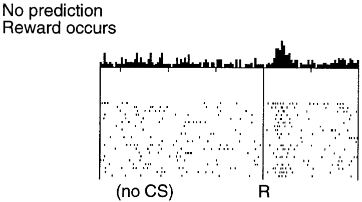
- Dopamine was first hypothesized to be the reward system,
- now is associated with prediction error of reward learning.

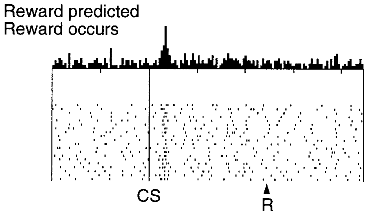
Dopamine and prediction error

Phasic activity of dopamine neurons proposed as reflecting prediction error:

- series of experiments by Schultz and coworkers,
- theoretical work by Montague, Dayan and coworkers.

- Before learning - reward unexpected
- After learning - reward expected
- After learning - no reward is unexpected





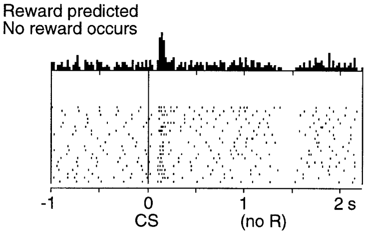


Figure adapted from Schultz 1998.

Dopamine and prediction error - second order conditioning

- Dopamine neurons response transfers to earliest predictive stimulus.

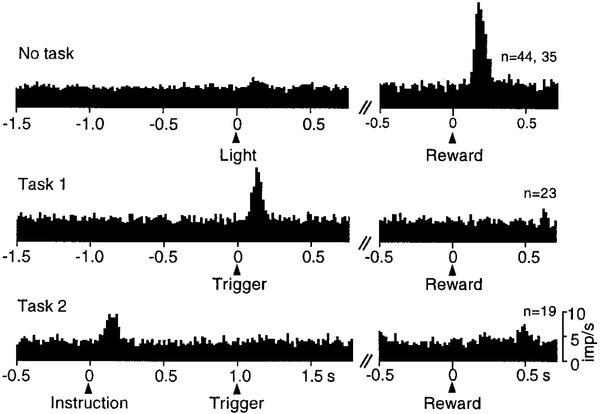


Figure adapted from Schultz 1998.

- Dopamine neurons activity reflects reward probability.

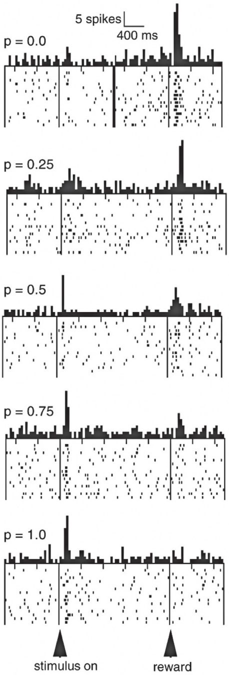
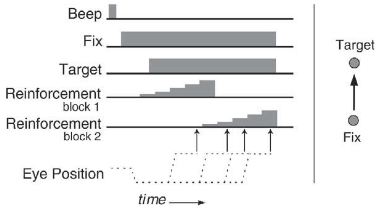


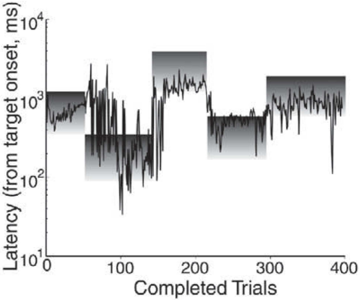
Figure adapted from Fiorillo et al 2003.

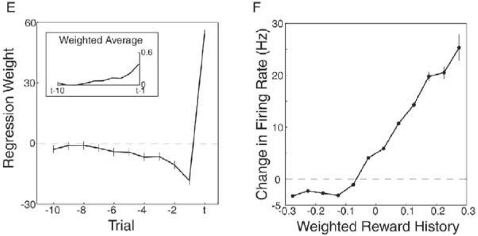
Dopamine and prediction error - history of reward

- Dopamine neurons activity reflects history of previous rewards (for reward higher than expected).



$$y = \beta_0 r_t + \beta_1 r_{t-1} + \cdots + \beta_{10} r_{t-10} + k$$





Figures adapted from Bayer and Glimcher 2005.

Dopamine response

Dopamine release after stimulation of the axon:

- rise in extracelullar dopamine concentration of several nM in few ms,
- concentration becomes homogeneous on a sphere, max diffusion after 75 ms, over a diameter of 7-12 μm, 80 nM,
- reuptake takes concentration to baseline values after a few 100 ms.

Dopamine activity after a reward predicting event:

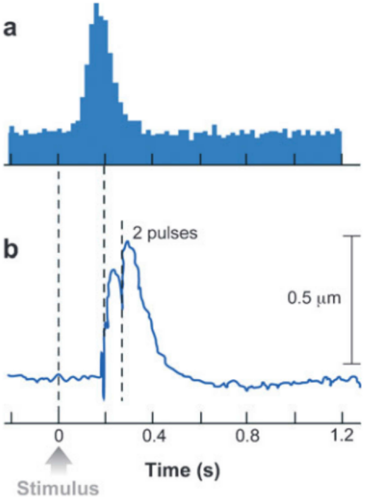


Figure from Schultz, 2007.

Dopamine and synaptic plasticity

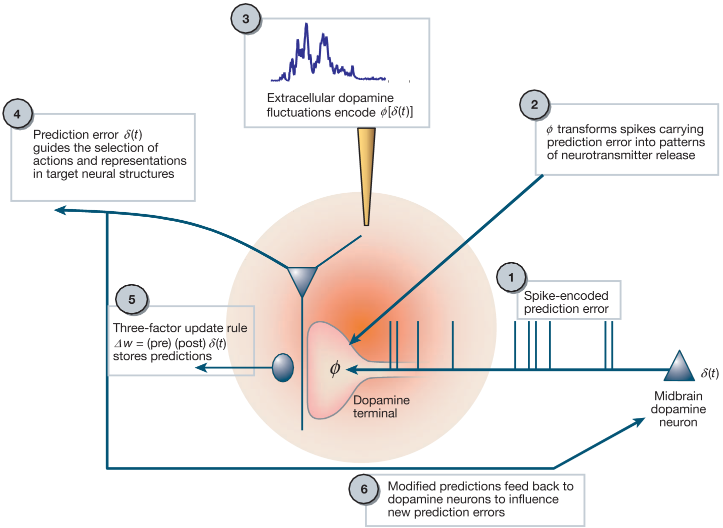


Figure from Montague et al. 2004.

Operant or instrumental conditioning

Introducing actions

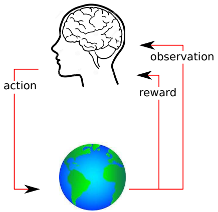


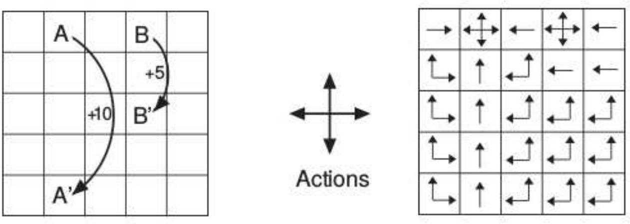
Figure from Scientific American.

Animals will behave in order to get reward.

Reinforcement learning

- Learning by trial-and-error which action to choose
- Learning values of actions for a given state, $Q(s, a)$
- Policy π for behavior

Example for the gridworld:



- How to find the best policy?

Policy improvement

- On state s select $a \neq \pi(s)$ and then follow π.
- If $Q(s, a) > V^{\pi}(s)$ then the change leads to a policy π' better than π.
- Intuition: extending to all actions, a policy can be improved by taking better actions.

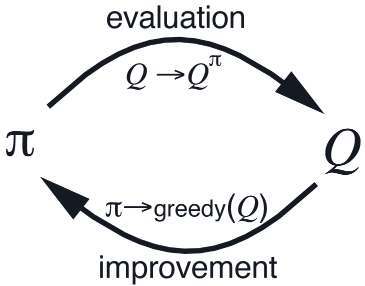


Figure adapted from Sutton and Barto 1998.

- Greedy policy: a policy that only selects actions with maximum value.

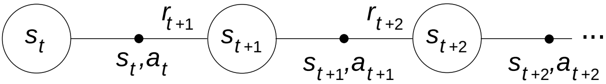
Trade off between exploration and exploitation

$Q(s, a)$ can be estimated iteratively but one has to assure that all actions continue to be selected.

- On-policy methods
  - Use a soft policy: $\pi(s, a) > 0$ for all a and s
  - For example ϵ-greedy policy chooses the greedy action but with probability ϵ chooses another random action.
- Off-policy methods
  - Different behavior and estimation policies
  - The estimation policy can be greedy as long as the behavioral policy continues to sample all actions.

Sarsa: on-policy TD control

- Policy iteration
- TD methods for evaluation / prediction



$$Q(s_t, a_t) \longleftarrow Q(s_t, a_t) + \eta[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

The rule uses the events s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1}.

- Sarsa converges to optimal policy and action-value if:
  - all s, a are visited an infinite number of times
  - the policy converges to greedy (e.g. ϵ-greedy with $\epsilon = 1/t$)

Q-learning: off-policy TD control

One step Q-learning:

$$Q(s_t, a_t) \longleftarrow Q(s_t, a_t) + \eta[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$$

- The action-value function approximates the optimal, independent of the policy followed.
- In some conditions convergence can be shown.

Actor/critic method

The critic evaluates using TD error:

$$\delta_t = r_{t+1} + \eta V(s_{t+1}) - V(s_t)$$

Given a_t, s_t:

- $\delta_t > 0 \rightarrow$ increase probability of selecting a
- $\delta_t < 0 \rightarrow$ decrease probability of selecting a

For example:

$$p(s_t, a_t) \leftarrow p(s_t, a_t) + \beta\delta_t$$

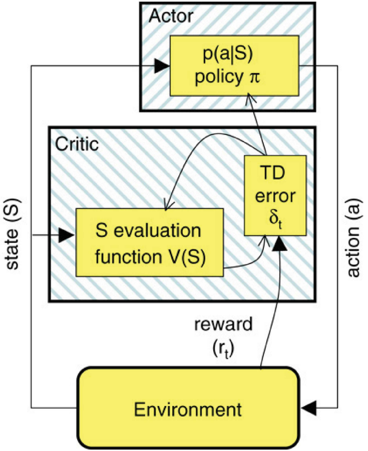


Figure from Niv, 2009.

References

- H. Bayer and P. Glimcher, Midbrain dopamine neurons encode a quantitative reward prediction error signal, *Neuron*, 2005.
- P. Dayan and L. Abbott, Theoretical neuroscience, 2001.
- C. Fiorillo et al., Discrete coding of reward probability and uncertainty by dopamine neurons, *Science*, 2003.
- P. Montague et al., Computational roles for dopamine in behavioural control, *Nature*, 2004.
- Y. Niv, Reinforcement learning in the brain, *Journal of Mathematical Psychology*, 2009.
- W. Schultz, Predictive reward signal of dopamine neurons, *Journal of Neurophysiology*, 1998.
- W. Schultz, Multiple dopamine functions at different time courses, *Annual Review of Neuroscience*, 2007.
- R. Sutton and A. Barto, Reinforcement learning, 1998.