ON NEURAL CORRELATES OF REINFORCEMENT LEARNING

the role of dopamine in planning and action

Genela Morris
Dept. of Neurobiology
Haifa University
gmorris@sci.haifa.ac.il
Suggested reading

- Barto AG & Sutton RS. *Reinforcement Learning: An introduction*. MIT Press, Cambridge MA (1988): Ch. 3, Ch. 6 + some of Ch. 2
- Figures from research papers are referenced throughout the presentation
Reinforcement learning
the basics

Supervised learning –
all knowing teacher, detailed feedback

Reinforcement learning –
scalar (correct/incorrect) feedback

Unsupervised learning –
self organization
Reinforcement learning: The law of effect

“The Law of Effect is that: Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur”

Edward Lee Thorndike (1911)
Early attempts at modeling

- By associative rules
- Classical conditioning
Properties of classical conditioning

(Pavlov 1927)

• Acquisition.

• Partial Reinforcement (probabilistic).

• Generalization.

• Interstimulus Interval (ISI) effects.

• Intertrial Interval (ITI) effects.
So far...

- A simple association (coincidence, Hebbian) model can explain the phenomenon.

- Acquisition.
- Partial Reinforcement (probabilistic).
- Generalization.
- Interstimulus Interval (ISI) effects.
- Intertrial Interval (ITI) effects.
Classical conditioning

The Elements:
- **US:** Unconditioned stimulus
- **UR:** Unconditioned response
- **NS:** Neutral stimulus
- **CS:** Conditioned stimulus
  - **CS1:** Conditioned stimulus 1
  - **CS2:** Conditioned stimulus 2
- **CR:** Conditioned response
Properties of classical conditioning

(Cnt’d)

- Conditioned Inhibition
- latent inhibition
- Relative validity (Wagner 1968).
- Blocking (Kamin 1968)
- ...

CS must RELIABLY predict US
Learning occurs not because two events co-occur, but because that co-occurrence is otherwise UNPREDICTED
Rescorla-Wagner rule (1972)

Learning to predict reward $R$ given stimulus $U=1$

Goal: Form a prediction $V$ of the reward of the form:

$$V = \omega U$$

And learn to change $\omega$:

$$\Delta \omega = \varepsilon (R-V)U$$

After learning of consistent pairing: $\omega = R$

Where:

$U =$ CS availability $(0,1)$;
$V =$ reward prediction;
$R =$ reward availability $(0,1)$:

$\omega =$ weight of the connection between $U$ and $V$
$\varepsilon =$ learning rate
$R-V =$ prediction error
Blocking with Rescorla Wagner

• Given U1, U2 and R, after U1 has been learnt:
  • $\omega_1 = R$
  • $V = \omega_1 U_1 + \omega_2 U_2$

R  0

• Prediction error: $R - V = 0$
  And no learning occurs for $\omega_2$
Critical problems, for control

1. Exploration/exploitation
Solutions, for control

1. Variability in response policy
   1. Greedy $\leftarrow \rightarrow$ Random (gambling)
   2. Based on expected return
Decision behaviour, theory and practice

maximizing probability-matching

monkeys?

\[
\frac{C_{right}}{C_{right} + C_{left}} = \frac{R_{right}}{R_{right} + R_{left} \cdot \frac{\theta_{left}}{\theta_{right}}}
\]

\( R = \text{reward} \)

\( C = \text{choice} \)
Monkeys’ decisions: probability matching
... whether optimal or not

• Actions are related to their consequences
Critical problems in reinforcement learning (and in Rescorla-Wagner)

2. Temporal credit assignment
TD learning - solution for temporal credit assignment

1. Estimate value of current state \( V_t = r_t + \gamma V_{t+1} + \ldots \) : (discounted) sum of expected rewards

2. Measure ‘truer’ value of current state: reward at present state + estimated value of next state \( r_t + \gamma V_{t+1} \)

3. TD error \( \delta_t = r_t + \gamma V_{t+1} - V_t \)

4. Use TD error to improve 1 \( V_t^{k+1} = V_t^k + \eta \delta_t \)

where: \( V_t = \text{value of the state reached at time } t \text{ in iteration } k \)
\( r_t = \text{reward given at time } t \); \( \eta = \text{learning rate} \), \( \delta = \text{prediction error} \)
TD error:  \[ \delta_t = r_t + \gamma V_{t+1} - V_t \]
TD error: \( \delta_t = \gamma V_{t+1} - V_t + r_t \)
Basal ganglia - anatomy
Intracranial self stimulation

Activates reward circuits
The midbrain dopamine system
Dopamine and acetylcholine meet in the striatum

Monkey

Mouse

Berlin 2004
Facts to remember (1)

- Basal ganglia receive cortical input
- Basal ganglia project to frontal cortex
- Dopamine and acetylcholine localization
The midbrain dopamine system

Schultz et al, 
Probabilistic instrumental conditioning task

\[ \delta_t = \gamma V_{t+1} - V_t + r_t \]

*Morris et al., Neuron 43(1): 133-143, 2004*
DA response
Dopamine population response to cue

The figure illustrates the dopamine population response to a cue in a task with different stages: start, cue, hold, go, delay, and reward. The x-axis represents time in seconds (-2.4 to +0.7), and the y-axis represents spikes per second (0 to 5). The graph shows different colored lines representing different probability levels (p=0.25, p=0.5, p=0.75, p=1) with corresponding data points. The graph also includes a scatter plot with conditional reward probability on the x-axis and spikes per second on the y-axis. The data point n=114 is indicated.
Dopamine population response - reward

-2.4  + 0.3  +1.5  +<0.8  +0.7 time(s)

Spikes/s

Conditional reward probability

-200  0  400

Spikes/S

11

7

11

7

3

0.25  0.5  0.75  1

Pr

reward

Haifa University
Dopamine population response – reward omission
Instrumental conditioning - results

• Responses to visual cue are correlated with future reward probability
• Responses to reward are inversely correlated with reward probability
• Responses to reward omission are indifferent to reward probability

Dopamine neurons provide an accurate TD signal (but only in the positive domain)
... and it can cause long term plasticity of cortico-striatal synapses

... and it can cause long term plasticity of cortico-striatal synapses

Shen et al., Science 321:848-851 2008
Facts to remember 2

• DA neurons provide a TD error signal
• To the cortico (state) striatal (action) synapses
• And DA modulates synaptic plasticity
Control - Adding action

The agent has to:

- Learn to predict reinforcement
- Know the state-action-state transitions

\[ \text{state value} \]

\[ \text{behavioural policy} \]
Solution 1: actor/critic networks

% Diagram

Actor \rightarrow \text{State} \rightarrow \text{Reward} \rightarrow \text{Critic} \rightarrow \text{TD} \rightarrow \text{Action} \rightarrow \text{Environment}
How can the dopamine signal contribute to decision behaviour?

• Long term policy-shaping effect through synaptic plasticity

• Immediate effect on action

\[ P_{\text{action}} = \frac{1}{1 + e^{-m\delta(t)+b}} \]
Monkeys’ decisions: probability matching

\[ R_{\text{right}} / (R_{\text{right}} + R_{\text{left}}) \]

\[ R^2 = 0.884 \]
The two armed bandit task

Monkeys’ decisions: probability matching
Lost in translation?

reward \[\xrightarrow{\times} \] behaviour

dopamine response \[\xrightarrow{\quad} \] plasticity in action circuits
Monkeys’ decisions: shaping by dopamine
Dopamine neurons during decision

stimulus  decision  action

DA  DA

[Diagram showing the flow from stimulus to decision to action with dopamine (DA) release at both decision and action stages.]

Reference: decision
Are DA neurons aware of future choice?

Hi (explore)
Lo (exploit)
The learning is of state-action values