Informatics 1 Cognitive Science

Lecture 27: Reinforcement Learning Part 1

Matthias Hennig

School of Informatics
University of Edinburgh
mhennig@inf.ed.ac.uk

Overview

Classical Conditioning

Classical Conditioning

What is Conditioning?

An everyday example:

What is Conditioning?

An everyday example:

- Notification sounds on your phone: starts as neutral stimulus.

What is Conditioning?

An everyday example:

- Notification sounds on your phone: starts as neutral stimulus.
- You realise some messages are rewarding.

What is Conditioning?

An everyday example:

- Notification sounds on your phone: starts as neutral stimulus.
- You realise some messages are rewarding.
- Repeated pairing with rewarding messages triggers expectation.

What is Conditioning?

An everyday example:

- Notification sounds on your phone: starts as neutral stimulus.
- You realise some messages are rewarding.
- Repeated pairing with rewarding messages triggers expectation.
- You start checking your phone because it is often rewarding (maybe even without receiving a notification).

What is Conditioning?

An everyday example:

- Notification sounds on your phone: starts as neutral stimulus.
- You realise some messages are rewarding.
- Repeated pairing with rewarding messages triggers expectation.
- You start checking your phone because it is often rewarding (maybe even without receiving a notification).

Conditioning is learning regularities between events:

What is Conditioning?

An everyday example:

- Notification sounds on your phone: starts as neutral stimulus.
- You realise some messages are rewarding.
- Repeated pairing with rewarding messages triggers expectation.
- You start checking your phone because it is often rewarding (maybe even without receiving a notification).

Conditioning is learning regularities between events:

- a cue predicts an outcome, or

What is Conditioning?

An everyday example:

- Notification sounds on your phone: starts as neutral stimulus.
- You realise some messages are rewarding.
- Repeated pairing with rewarding messages triggers expectation.
- You start checking your phone because it is often rewarding (maybe even without receiving a notification).

Conditioning is learning regularities between events:

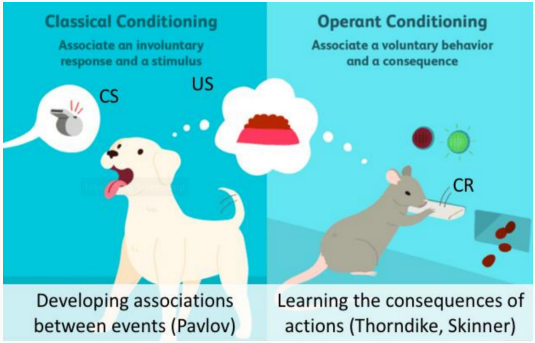
- a cue predicts an outcome, or
- an action changes what happens next.

What is Conditioning?

Conditioning links cue $\rightarrow$ expectation (**classical conditioning**) and action $\rightarrow$ outcome (**operant conditioning**).

Cues can be **rewarding** or **aversive**, and they may be **external** or **internal** (vicarious).

Classical and Operant (Instrumental) Conditioning



Classical conditioning: unconditioned stimulus (food) + conditioned stimulus (bell) $\rightarrow$ unconditioned response (salivation) becomes conditioned response

Operant conditioning: action (lever press) + unconditioned stimulus/reward (food) $\rightarrow$ action becomes more likely

Terminology Change for Reinforcement Learning (RL)

Stimulus + reward $\rightarrow$ expectation of reward

- Classical conditioning often describes learning in terms of **stimulus $\rightarrow$ response**.
- Reinforcement learning shifts the focus to **stimulus $\rightarrow$ reward prediction**.
- Key idea: behaviour reflects an internal **expectation of reward**.
- Therefore, RL models are usually phrased in terms of **stimuli**, **rewards**, and **expected reward**.

Classical Conditioning: The Rescorla-Wagner Rule

V_t: associative strength between CS and US at time t. Update rule:

$$\underbrace{V_t}_{\text{new}} = \underbrace{V_{t-1}}_{\text{old}} + \alpha \underbrace{(R - V_{t-1})}_{\delta}$$

- R: received reward
- $\alpha \in (0, 1)$: learning rate
- $\delta = R - V_{t-1}$: **prediction error**

Known as δ-**rule**: update $\propto$ surprise.

Classical Conditioning: The Rescorla-Wagner Rule

V_t: associative strength between CS and US at time t. Update rule:

$$\underbrace{V_t}_{\text{new}} = \underbrace{V_{t-1}}_{\text{old}} + \alpha \underbrace{(R - V_{t-1})}_{\delta}$$

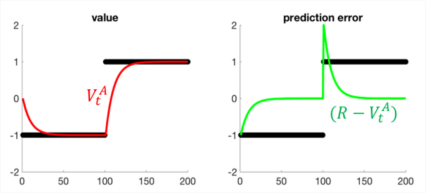
- R: received reward
- $\alpha \in (0, 1)$: learning rate
- $\delta = R - V_{t-1}$: **prediction error**

Known as δ-**rule**: update $\propto$ surprise.

Prediction error updates:

- $\delta > 0$: outcome better than expected $\Rightarrow$ increase V
- $\delta = 0$: outcome as expected $\Rightarrow$ no change
- $\delta < 0$: outcome worse than expected $\Rightarrow$ decrease V

The Rescorla-Wagner Rule for Classical Conditioning



$$V_t = V_{t-1} + \alpha(R - V_{t-1})$$

First R was set to -1, and then changed to $+1$. Learning rate $\alpha = 0.1$.
Figure shows the learned associative strength $V(t)$ (left) and the prediction error $\delta(t)$ (right). When R, $\delta(t)$ becomes large because the outcome is surprising; later $V(t)$ approaches the new R and $\delta(t)$ decays to zero.

Prediction Errors in the Brain

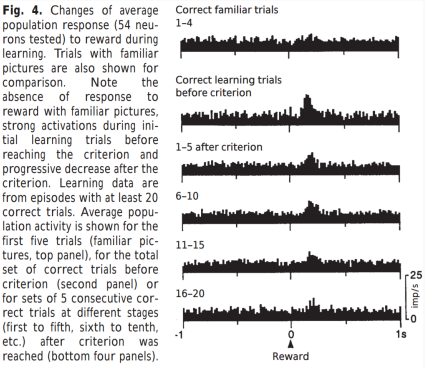
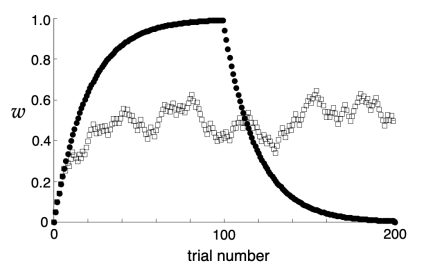


Fig. 4. Changes of average population response (54 neurons tested) to reward during learning. Trials with familiar pictures are also shown for comparison. Note the absence of response to reward with familiar pictures, strong activations during initial learning trials before reaching the criterion and progressive decrease after the criterion. Learning data are from episodes with at least 20 correct trials. Average population activity is shown for the first five trials (familiar pictures, top panel), for the total set of correct trials before criterion (second panel) or for sets of 5 consecutive correct trials at different stages (first to fifth, sixth to tenth, etc.) after criterion was reached (bottom four panels).

Visual discrimination task with reward: Dopamine neurons of the substantia nigra appear to signal the prediction error as predicted by the Rescorla-Wagner rule (Hollerman & Schultz, 1998, Nature Neuroscience, 1(4), 304-309).

Extinction and Partial Reinforcement



Extinction: reward is first present ($R = 1$) and then absent ($R = 0$) (filled circles).
Partial Reinforcement: reward is paired only 50% of the time (empty squares).

Blocking

- **Phase 1 (pre-training):** $v_1 \rightarrow R$ until $V_1 \approx R$.
- **Phase 2 (compound training):** $(v_1 + v_2) \rightarrow R$.
- **Test:** v_2 alone elicits little reward expectation (learning about v_2 is **blocked**).

When v_1 already predicts R, there is little **surprise** left for v_2 to learn from.

Rescorla-Wagner with two stimuli
Predicted reward:

$$V = V_1 + V_2$$

Prediction error:

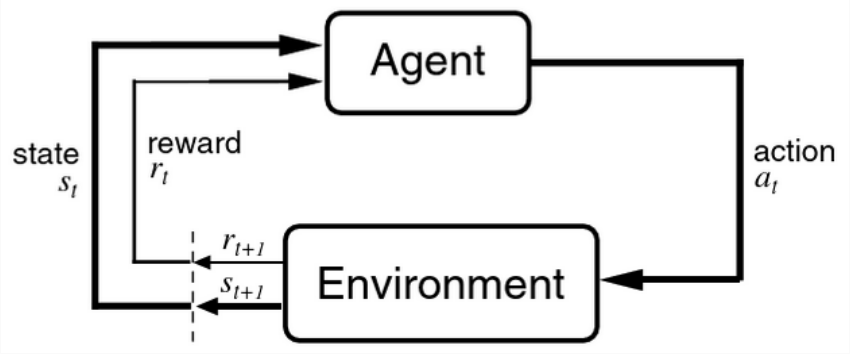
$$\delta = R - (V_1 + V_2)$$

Update (for any presented stimulus v_i):

$$V_i \leftarrow V_i + \alpha\,\delta$$

During Phase 2,
$V_1 \approx R \Rightarrow \delta \approx 0 \Rightarrow V_2$ hardly changes.

Reinforcement Learning: Approach



- RL agents have explicit goals, manifested through rewards (or punishments).
- RL agents act on the environment and collect information to inform the next action.

Summary

- Conditioning is learning stimulus/action $\rightarrow$ reward expectation.
- Learning is driven by **prediction errors**: outcomes better/worse than expected update expectations.
- With multiple cues, learning depends on surprise (e.g., **blocking** when a cue is already predictive).
- Prediction-error-like signals are observed in the brain (dopamine).