Animal learning theory

Based on [Sutton and Barto, 1990, Dayan and Abbott, 2001]

May 18, 2020
Stimulus-response:

- Unconditioned stimulus (US): A stimulus that unconditionally, naturally, and automatically triggers a response.
- Unconditioned response (UR): an unlearned response that occurs naturally in reaction to the unconditioned stimulus.

For example, if the smell of food is the unconditioned stimulus, the feeling of hunger in response to the smell of food is the unconditioned response.
Classical conditioning: A so-called conditioned stimulus is a previously neutral stimulus that, after becoming associated with the US, eventually comes to trigger a conditioned response CR similar to UR.

Example: rabbit eye blink condition
- US air puff to rabbit eye, UR is eye blink
- CS is sound of a buzzer
- After CS-US pairing, the buzzer alone yields CR=UR eye blink
Trial level and Real-time theories

Trial level theories ignore temporal aspects of individual trial and updates apply at end of trial. Popular in experiments
- Rescorla-Wagner model [Sutton and Barto, 1990]
- learning trial level actions [Dayan and Abbott, 2001] (DA) Chapter 9.3

Real-time theories consider detailed timing aspects of the stimuli and rewards
- TD learning (DA 9.2)
- introduction to RL (Kaelbling)
- introduction to control theory, Bellman equation, (Kappen)
- exploration/exploitation, bandit problem (Kaelbling)
- RL algorithms: value iteration, policy iteration, Q learning, actor critic, Dyna (Kaelbling)
- Maze example and Morris water maze DA 9.4

1

1Last week: Trial level theories, TD learning, Intro to RL, Intro to control theory
This week: learning trial level actions, TD learning, control theory, bandit problem
Next week: RL algorithms, water maze

Bert Kappen
Rescorla-Wagner (RW) model

Learns relation between CS and US. Central idea is that learning occurs whenever events violate expectations.

In particular, when expected US $\bar{V}$ differs from the actual US level $\lambda$:

$$V_i := V_i + \Delta V_i$$

$$\Delta V_i = \beta \left( \bar{V} \right) \frac{\alpha_i X_i}{\text{reinforcement saliency}}$$

$$\bar{V} = \sum_i V_i X_i$$

$X_i = 0, 1$ to indicate absence/presence of $\text{CS}_i$. $\alpha_i$ is saliency (attention) of $\text{CS}_i$ (fixed).

$V_i$ is the association strength of stimulus $\text{CS}_i$ to US.

$\lambda$ denotes the 'level of US' (reward), "combining, in an unspecified way, the USs intensity, duration, and temporal relationship with the CS"\(^2\)

\(^2\)For instance, strength, duration of air puff and timing relative to CS.
RW is linear regression

Consider sequence \( \{(X_{i,t}, \lambda_t), t = 1, \ldots, T\} \) with \( X_{i,t} \) the value of stimulus \( i \) \((i = 1, \ldots, n)\) at time \( t \) and \( \lambda_t \) the reward at time \( t \). Define error

\[
C = \frac{1}{2} \sum_{t=1}^{T} \left( \lambda_t - \sum_{i=1}^{n} X_{i,t} V_i \right)^2
\]

Minimizing \( C \) wrt \( V_i \) gives best predictive model.

\[
\Delta V_i = -\beta \frac{\partial C}{\partial V_i} = \beta \left( \lambda_t - \sum_{i} X_{i,t} V_i \right) X_{i,t}
\]

Identical to RW rule with identical saliencies \( \alpha_i = 1 \).

After convergence to \( V_i^* \) the expected/predicted reward ('level of US') for stimulus \( X_i, i = 1, \ldots, n \) is

\[
\lambda = \sum_{i=1}^{n} V_i^* X_i
\]
Rescorla-Wagner model

Rescorla Wagner rule for multiple inputs can predict various phenomena:

- **Blocking**: learned $s_1 \rightarrow r$ prevents learning of association $s_2 \rightarrow r$
- **Inhibition**: $s_2$ reduces prediction when combined with any predicting stimulus
Foraging of bees

Foraging of bees as example of learning behavior with immediate reward:
• bees visit flowers of two colors (blue, yellow), preferring color with higher reward (nectar)
• when rewards are swapped, bees adjust their preference

Quantity of nectar $r_{b,y}$ is stochastic from (fixed) distribution $q_{r,b}(r)$. This is like a two-armed bandit problem.

The policy of the bees is given by a probability $p_{b,y}$ with $p_b + p_r = 1$. We can parametrize this as

$$p_b = \frac{e^{\beta m_b}}{e^{\beta m_b} + e^{\beta m_y}} \quad p_y = \frac{e^{\beta m_y}}{e^{\beta m_b} + e^{\beta m_y}}$$

The parameters $m_{b,y}$ are called action values, that do no require normalization.

The parameter $\beta$ controls exploration-exploitation:
• $\beta = 0$: exploration, $\beta = \infty$: exploitation
Foraging of bees

Two learning strategies:
- learn expected reward and set policy $p_{y,b}$ accordingly (indirect actor)
- learn policy directly by maximizing expected reward (direct actor)
Indirect actor

Bee visits flowers (yellow and blue) with reward distributions $q_{y,b}(r)$.

Bee visits the flowers according to $p_{y,b}$ using current values of $m_{y,b}$.

$m_{y,b}$ are learned from past experience. For $m_b$, consider the cost function $E(m_b) = \frac{1}{2} \sum_r q_b(r)(m_b - r)^2$. Learning is done by gradient descend:

$$m_b := m_b + \Delta m_b \quad \Delta m_b = -\epsilon \frac{dE}{dm_b} = \epsilon (\langle r \rangle_b - m_b)$$

In stationary environment, $E(m_b)$ is minimized $m_b = \langle r \rangle_b$.

In reality, the bee uses stochastic gradient descent to learn $m_b$, replacing $\langle r \rangle_b$ by a single sample $r$ from $q_b(r)$, and similar for $m_y$:

$$\Delta m_b = \epsilon (r - m_b) \quad \text{when bee gets reward } r \text{ from visit to blue flower}$$
$$\Delta m_y = \epsilon (r - m_y) \quad \text{when bee gets reward } r \text{ from visit to yellow flower}$$

This is instance of the Rescorla-Wagner rule.
Indirect actor

\[ \langle r_b \rangle = 1, \langle r_y \rangle = 2 \] for first 100 visits with

\[ q_b(r) = \frac{1}{2} \delta_{r,0} + \frac{1}{2} \delta_{r,2}, \quad q_y(r) = \frac{1}{2} \delta_{r,0} + \frac{1}{2} \delta_{r,4}. \]

Rewards are reversed for second 100 visits.

A) Values of \( m_b \) (solid) and \( m_y \) (dashed) as a function of visits for \( \beta = 1 \). Because a fixed value of \( \epsilon = 0.1 \) was used, the weights do not converge perfectly to the corresponding average reward, but they fluctuates around these values.
Indirect actor

B-D. Cumulative visits to blue (solid) and yellow (dashed) flowers. B) When $\beta = 1$, learning is slow (initially both flowers are visited equally often), but ultimately the change to the optimal flower color is made reliably. C;D) When $\beta = 50$, sometimes the bee performs well (C), and other times it performs poorly (D) (first visits yellow, the blue for a while, then yellow again).

Note, that learning $m_b$ or $m_y$ only takes place when the bee visits the blue or yellow flower, respectively. For instance, in C, $m_b$ is hardly updated because the blue flower is not visited due to the large $\beta$.

This behavior is seen in real bees.
Indirect actor

The behaviour does not only depend on \( \langle r \rangle_{b,y} \). Real bees are risk averse.

A) Blue flowers provide 2 \( \mu l \) nectar, yellow flowers provide 6\( \mu l \) on 1/3 of the trials and zero otherwise. After 15 trials reversed. Mean preference of 5 bees for blue flowers over 30 trials, each consisting of 40 visits.

\[
\langle f(r) \rangle_b = \sum_r q_b(r)f(r) > \langle f(r) \rangle_y = \sum_r q_y(r)f(r)
\]

B) This can be modelled by replacing \( r \) by \( f(r) \) with \( f \) a concave function. Then

C) Model prediction of visits to blue using \( f (\epsilon = 0.3, \beta = 23/8) \)
Direct actor

Direct actor method estimates $m_{b,y}$ such as to maximize the expected future reward

$$\langle r \rangle = p_b \langle r \rangle_b + p_y \langle r \rangle_y$$

$$p_{y,b} = \frac{e^{\beta m_{y,b}}}{e^{\beta m_{y}} + e^{\beta m_{b}}}$$

We optimize $m_{b,y}$ by gradient ascent on $\langle r \rangle$:

$$\Delta m_{b} = \epsilon \frac{\partial \langle r \rangle}{\partial m_{b}} = \epsilon \beta p_b p_y \left( \langle r \rangle_b - \langle r \rangle_y \right)$$

$$\Delta m_{y} = \epsilon \frac{\partial \langle r \rangle}{\partial m_{y}} = \epsilon \beta p_b p_y \left( \langle r \rangle_y - \langle r \rangle_b \right)$$

where we used

$$\frac{dp_b}{dm_b} = \frac{dp_y}{dm_y} = -\frac{dp_b}{dm_y} = -\frac{dp_y}{dm_b} = \beta p_b p_y$$
Direct actor

The bee estimates $\langle r \rangle_b$ from the $N_b = Np_b$ visits to the blue flower:

$$\langle r \rangle_b \approx \frac{1}{Np_b} \sum_{\text{visits } b} r_b$$

$$\langle r \rangle_y \approx \frac{1}{Np_y} \sum_{\text{visits } y} r_y$$

The stochastic version is ($\epsilon \leftarrow \frac{\epsilon B}{N}$)

$$\Delta m_b = \epsilon p_y \sum_{\text{visits } b} r_b - \epsilon p_b \sum_{\text{visits } y} r_y$$

$$\Delta m_y = \epsilon p_b \sum_{\text{visits } y} r_y - \epsilon p_y \sum_{\text{visits } b} r_b$$

or

$$\Delta m_b = -\Delta m_y = \epsilon p_y r_b \quad \text{if } b \text{ is selected}$$

$$\Delta m_b = -\Delta m_y = -\epsilon p_b r_y \quad \text{if } y \text{ is selected}$$
For multiple actions we have

\[ \langle r \rangle = \sum_a p_a \langle r \rangle_a \quad p_a = \frac{e^{\beta m_a}}{\sum a' e^{\beta m_{a'}}} \]

The direct actor learning rule generalizes to, when action \( a \) is taken:

\[ \delta m_{a'} = \epsilon (\delta_{a,a'} - p_{a'}) r_a \quad \forall a' \]

(For instance, take \( a = b \) and \( a' = b \), \( y \) gives first line above).

We will use this in the actor-critic algorithm (DA 9.4).
Direct actor

Setting as before; two different runs. A,B) Successful learning. C,D) Failure to learn to switch due to larger difference $m_y - m_b$

Both direct and indirect actor may be too greedy and have poor exploration.
Temporal aspect play important role

Figure 1
Illustration of variations in a US’s reinforcing effect, $\lambda$, within a single trial. With a long duration US, a CS preceding its onset can become positively associated with the US, whereas a CS preceding its offset can become negatively associated.
Real-Time Theories of Eligibility

Conditioning depends on time interval between CS and US.

Mechanisms for coincidence of CS and US:
- *Stimulus trace* (Hull 1939) is internal representation build by the brain. Disadvantage of using stimulus trace for learning is that long inter-stimulus intervals (ISIs) require broad stimulus trace, which disagrees with fast, precisely timed responses.

- *Eligibility trace* $\bar{X}$ (Knopf 1972) is similar to stimulus trace but only used for learning
From RW to a real time theory

Left: Trial based reinforcement $\lambda$ is area under the curve. Right: Real-time reinforcement is future area under the curve.

Predicted reinforcement as future area under the curve is constant for all times prior to US (B). This disagrees with empirical observation.

*imminence weighted* predicted future reinforcement reduces long time association agrees with empirical observation (C).
Subject predicts at time $t$ the imminence weighted area (c) rather than the un-weighted area (a) which may be infinite.
Temporal difference learning

Consider time interval \(0 \leq t < \infty\) with \(u(t)\) stimulus, \(\lambda(t)\) the reward. The expected (discounted=imminence weighted) future reward is

\[
R(t) = \mathbb{E} \sum_{\tau=t}^{\infty} \gamma^{\tau-t} \lambda(\tau)
\]

The aim of reinforcement learning is to learn a signal \(v(t)\) that predicts the expected future reward \(R(t)\) from the past stimulus.

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

with \(w(\tau)\) adaptive parameters.

Minimize quadratic error \(C = \frac{1}{2} \sum_{t=0}^{\infty} (v(t) - R(t))^2\) with respect \(w(\tau)\) yields

\[
\Delta w(\tau) = -\epsilon \frac{\partial C}{\partial w(\tau)} = \epsilon \sum_{t=0}^{\infty} (R(t) - v(t)) u(t - \tau)
\]
Temporal difference learning

Problem is that we do not know $R(t)$. But we can use the recursive relation

$$R(t) = \mathbb{E} \sum_{\tau=t}^{\infty} \gamma^{t-\tau} \lambda(\tau) = \lambda(t) + \gamma \mathbb{E} \sum_{\tau=t+1}^{\infty} \gamma^{t-\tau-1} \lambda(\tau) = \lambda(t) + \gamma R(t+1) \approx \lambda(t) + \gamma v(t+1)$$

We approximate $R(t+1)$ by $v(t+1)$. Update at time $t$ becomes:

$$\Delta w(\tau) = \epsilon \sum_{t=0}^{\infty} (R(t) - v(t)) u(t - \tau) \approx \epsilon \sum_{t=0}^{\infty} \delta(t) u(t - \tau)$$

$$\delta(t) = \lambda(t) + \gamma v(t+1) - v(t)$$
Temporal difference learning

When $v(t)$ is not parametrized as above, but is estimated for each $t$ independently:

$$C = \frac{1}{2} \sum_{t=0}^{\infty} (R(t) - v(t))^2$$

$$\Delta v(t) = -\epsilon \frac{\partial C}{\partial v(t)} = \epsilon \sum_{t=0}^{\infty} (R(t) - v(t)) \approx \epsilon \sum_{t=0}^{\infty} (r(t) + \gamma v(t + 1) - v(t))$$

---

Figure 17
Intra-trial behavior of $\bar{V}$ for the TD model in a CSC experiment. The highest curve represents the theoretically correct predictions as given by Equation 4. The lower curves are the predictions generated by the TD model in a CSC experiment after various numbers of trials, as indicated. A different component CS is presented for every time step during the trial. The height at any intra-trial time represents the associative strength of the component CS presented at that time. The peak of the predictions is one time step before US onset—when the US is temporally closest but still lies entirely in the future. The US is 8 time steps long, and the intra-trial time shown is 40 time steps.
Inter-Stimulus-Interval Dependency

Rabbit CR (eye blink) to CS (sound) as a function of CS to US (puff) time interval shows good agreement with TD model.

Figure 5
The empirical ISI dependency for the rabbit startleting meridian response. Data is shown for both fixed-CS and delay conditioning (figure 6). The general shape of the ISI dependency is constant across species and response systems, but its time course varies substantially.

Figure 6
Temporal relationships in fixed-CS and delay conditioning. The indicated stimulus durations are commonly used in rabbit NM conditioning. These durations were also used to obtain the simulation data shown in figures 8, 13, and 18, under the interpretation that one simulation time step is equivalent to 50 ms.

Figure 18
ISI dependency of the TD model. Unlike the other models, the TD model’s ISI dependency is a good match to the empirical data in figure 5. These associative strengths were obtained after 80 trials. See figure 6 for the temporal relationships between stimuli.
Example of TD learning. DA Fig. 9.2

Consider interval $0 \leq t \leq T = 200$. Stimulus is $u(t) = \delta_{t,100}$

The value $v(t)$ is modelled for all time $t$ and is predicted from recent stimulus values in the time interval $[t - \tau_{\text{max}}, t]$. Thus Eq. 9.6 becomes

$$v(t) = \sum_{\tau=0}^{\tau_{\text{max}}} w(\tau)u(t - \tau) = w(t - 100) \quad 100 \leq t \leq 100 + \tau_{\text{max}}$$

$$= 0 \quad 0 \leq t < 100$$

When $\tau_{\text{max}} = 100$, one-to-one correspondence between $v(t)$ and $w(t - 100)$ on $100 \leq t \leq 200$. 

Delta rule Eq. 9.10 (NB sum is missing!) \(^3\)

\[
\Delta w(\tau) = \epsilon \sum_{s=0}^{\infty} \delta(s)u(s-\tau) = \epsilon \delta(\tau + 100) \quad 0 \leq \tau \leq 100
\]

\[
\delta(t) = r(t) + v(t + 1) - v(t)
\]

Reward is \( r = \delta_{t,200} \).

Define \( t = \tau + 100 \). We get equivalently:

\[
\Delta v(t) = \Delta w(\tau) = \epsilon \delta(t) \quad 100 \leq t \leq 200
\]

\(^3\)NB difference between \( \delta(t) \) in the temporal difference rule and \( \delta_{t,s} \), the Kronecker delta.
First iteration:  
\[ v(t) = 0 \quad \delta(t) = \delta_{t,200} \]

Second iteration:  
\[ v(t) = \epsilon \delta_{t,200} \quad \delta(t) = (1 - \epsilon) \delta_{t,200} + \epsilon \delta_{t,199} \]

Third iteration:  
\[ v(t) = (2\epsilon - \epsilon^2) \delta_{t,200} + \epsilon \delta_{t,199} \quad \delta(t) = \ldots \]
Dopamine

Activity of dopamine neurons in ventral tegmental area (VTA) of monkey encode $\delta$ in fig. 9.2.

- A: Left panels, trials locked to stimulus, right panels trials locked to reward. Top row is early trials, bottom row is late trials. VTA cells respond to reward in early trials and to stimulus in late trials.

- B Top: stimulus locked trials after learning as in bottom fig. A
- B bottom: withholding the (expected) reward yields inhibition

Monkey button press after stimulus (sound) to receive reward (fruit juice).
References
