Lecture Notes
Reinforcement Learning

Hanneke den Ouden
October 12, 2016

Learning goals

In this lecture we will look at how neuroscience and computational model approaches together have brought about an understanding of the computational and neurobiological processes that drive reinforcement learning. After attending the lecture and studying the slides and lecture notes, you should:

1. Understand why we believe that Pavlovian conditionings is driven by error-based learning
2. Recall the equations and use these to explain the two basic features of the Rescorla-Wagner model
3. Understand the evidence is that midbrain dopamine firing reflects the reward prediction error
4. Explain the limitations of the Rescorla-Wagner model (ie. which phenomena it cannot explain)
5. Explain how temporal difference learning addresses these
6. Explain how goal-directed behaviour cannot be understood in terms of classic learning theory (ie. RW and TD learning)
7. Understand how goal-directed and habitual behaviour can be distinguished experimentally

1 Computations and neurobiology of simple conditioning

1.1 Dopamine & processing reward

Adaptive behaviour is characterised by the ability to seek out rewards and avoid punishments. Similarity of neural responses across a range of types of rewards suggest that there is a modality independent, generalised reward function that drives this ability. Early studies suggested an important role of dopamine in reward processing, using dopamine agonists (ie substances that mimick dopamine) and antagonists (substances that block the working of dopamine) in experimental animals.

The first cracks in this theory appeared with the observation that, neural dopamine firing to a rewarding stimulus reduces over time, even though the rewarding quality of a stimulus itself would not reduce,as evidenced by the fact that agents would still seek out these rewards (see e.g. Schultz et al. 2002 Science).

1.2 Conditioning

Pavlovian conditioning is the phenonenon by which a stimulus comes to elicit the response (conditioned response, CR) that is normally only elicted by the outcome this stimulus predicts (reward/unconditioned stimulus, US). For example, in Pavlov's classic experiment, a bell comes to elicit a salivation response, when this bell has been paired with food presentation.

definitions

unconditioned stimulus (US):	stimulus that evokes unconditioned response, e.g. food
unconditioned response (UR):	response automatically (ie not learnt) evoked by unconditioned stimulus, e.g. salivation with food
conditioned stimulus (CS):	stimulus that comes to predict US and evokes CR.
conditioned response(CR):	response evoked by conditioned stimulus, e.g. salivation after bell

Pavlovian conditioning

Normally, US (e.g. food) elicits UR (e.g. salivation):	*US→UR*
Repeated presentation of	*CS→US →UR*
Leads to CS eliciting CR	*CS→ CR*

Pavlovian conditioning can be explained by either simple associative learning (ie unconditioned stimuli that appear together with valuable outcomes, acquire value), but this cannot explain the phenomenon of blocking.

blocking

In blocking (see slides for illustration), a stimulus CS1 is paired with a reward US, and comes to elicit a conditioned response CR (eg salivation). Then in a next phase , a 2nd conditioned stimulus is added, say CS2. Now repeated pairing of CS1+CS2 with the US fails to lead to a conditioned response to CS2. CS2 is said to be 'blocked'. This effect cannot be understood in a framework of associative/Hebbian learning, which would predict that repeated co-presentation of CS2 and US should lead to conditioning of CS2.

1.3 Rescorla-Wagner model and error-based learning

Rescorla & Wagner (1972) proposed that to understand blocking (as well as a number of other conditioning phenomena), learning during conditioning is driven by errors in prediction. From a parsimony viewpoint this makes theoretical sense: only when something unexpected happens, should you update your beliefs about the world.

In the RW model, the expected value EV (e.g. the expected amount / probability of receiving food after hearing a bell) is updated with a weighted combination of the EV before observing the last outcome, and the discrepancy between that expectation and the observed outcome *r.* Here, α determines the respective weights, serving as a learning rate:

$$EV_{t+1} = (1-\alpha)\cdot EV_t + \alpha \cdot r_t$$

This can be reformulated, to show that this equates to updating the previous belief with a weighted prediction error r_t - EV_t :

$$EV_{t+1} = EV_t + \alpha \cdot \left(r_t - EV_t\right)$$

These equations illustrates how 1) the EV is a weighted average of past rewards, where rewards further in the past. You can verify this by writing out the equation above for EV at t = 1, t = 2, t = 3, t = 4, and 2) how learning is error driven by using the prediction error to update beliefs.

The phenomenon of blocking can be understood in terms of error-driven learning, when extending the equation to multiple stimuli, where the EV is now the sum of the predictions *w* of each of *i* stimuli that are presented:

$$EV_t = \sum_i w_{i,t}$$
$$w_{i,t+1} = w_{i,t} + \alpha\left(r_t - EV_t\right)$$

Assume i = [1 2], e.g. stim 1 = bell, and stim 2 = light. In blocking, first you condition on the bell, so that after learning, $w_{bell} = 1$, ie it fully predicts the reward. In the next phase, we add the light, which initially has a value $w_{light} = 0$, because there is no reason to think it is associated with food. Now we present the light and bell, followed by food.
So the reward $r_t = 1$ (ie food is presented)

The expected value for this trial is:

$$EV_t = w_{bell,t} + w_{light,t}$$
$$EV_t = 1 + 0 = 1$$

And so we can compute the update for each of the weights:
$$w_{bell,t+1} = w_{bell,t} + \alpha(1-1) = w_{bell,t} = 1$$
$$w_{light,t+1} = w_{light,t} + \alpha(1-1) = w_{light,t} = 0$$

In other words, w_{light} will stay 0

Thus, the conditioning phenomenon of blocking provides support for the theory that learning in conditioning is error-driven rather than simply the co-presence of two stimuli.

1.4 Dopamine as a reward prediction error

Schultz et al (1997) observed that midbrain dopamine firing in macaque monkeys appears to reflect a theoretically predicted reward prediction error, firing in proportion to the difference between the reward received and the rewarded expected: increased firing for an unexpected reward, no change for an expected reward, and decreased firing for an unexpected omission of reward.

Dopamine neurons in the midbrain (ventral tegmental area) fire at a baseline rate of 3-5 Hz. Their firing patterns during Pavlovian conditioning matches exactly the theoretical prediction error (PE = r_t-EV_t)from the Rescorla Wagner model (Schultz ea. 1997):

- Burst of firing when PE>0, i.e. outcome is better than expected
- Decrease / pause in firing when PE<0,

- No change in firing when PE = 0. This situation occurs both when a presented reward is completely predicted, and of course also when simply nothing happens (ie no rewards are expected nor presented)

The size of the burst / length of pause are proportional to how much the outcome is better/worse than expected (Fiorillo ea. 2003).

For example, assume a situation where a cue is followed by a reward 25% of the time, i.e. EV = 0.25.
- If a reward is presented, PE = 1-0.25 = 0.75, leading to a relatively large burst of firing
- If no reward is presented, PE = 0 – 0.25 = -0.25, leading to a relatively short pause in firing

In a situation where the cue is followed by a reward 75% of the time, i.e. EV = 0.75, thus
- If a reward is presented, PE = 1-0.75 = 0.25, leading to a relatively small burst of firing
- If no reward is presented, PE = 0 – 0.75 = -0.75, leading to a relatively long pause in firing

Final evidence that the dopamine firing pattern does not just look like a prediction error, but is indeed sufficient for learning to occur, was provided recently using optogenetics (Steinberg ea. Nature Neuroscience 2013). Here, stimulation of the dopamine neurons at the time of outcome delivery during a blocking paradigm was shown to be sufficient to unblock the blocked stimulus (CS2, see above):

Phase 1: CS1+reward → conditioned response to CS1
Phase 2: CS1+CS2 + reward + optogenetic stimulation of DA neurons→ conditioned response to both CS1 and CS2

Whereas in the original blocking experiment without the optogenetic stimuluation:
Phase 2: CS1+CS2 + reward → conditioned response to only CS1 but not CS2

Finally, the dopamine prediction errors are 'computed' in the following way: sensory information provides input about r_t by driving activity in the DA neurons, whereas GABAergic neurons inhibit the dopamine neurons, reflecting EV_t. So if the excitation>inhibition, ie $r_t > EV_t$, then the DA neurons will fire, and vice versa.

1.5 Evidence for dopamine reward prediction errors in humans

Also in humans there is evidence, from neuroimaging and psychopathology, that prediction errors are encoded by dopamine. This evidence is based on indirectly measuring neural activity with fMRI in the striatum, a subcortical target structure of the dopaminergic midbrain, so that when dopamine neurons in the midbrain fire, dopamine is released in the striatum.

Pharmacology: dopamine agonists and antagonists
Neural activity in the striatum scales with the size of the prediction error, such that it increases with larger prediction errors. This effect is enhanced when people receive dopamine agonists, and reduces when people receive dopamine antagonists.

Psychopathology : Parkinson's disease
Parkinson's disease is characterised by a progressive depletion of dopamine in the striatum (because the dopamine neurons die). Prediction error activity measured using fMRI is abolished in the part of the striatum (dorsal) where depletion is more advanced, compared to the part of the striatum (ventral), where dopamine levels are still intact.

2 multi-step choice

2.1 problems with Rescorla-Wagner

Second Order Conditioning (SOC)

The Rescorla-Wagner model cannot explain 2nd order conditioning: here once CS1 is conditioned to the reward, presentation of a 2nd CS preceding the first CS, leads to the 2nd CS to also acquire value and evoke a conditioned response. Thus, repeated presentation of

CS2→CS1→reward

Leads to

CS2→ conditioned response

Even though there has *never been a reward prediction error*! RW learning would predict here that because there is no prediction error, there is also no learning.

Dopamine firing to cue

Similarly, it cannot explain why if dopamine neurons reflect a reward prediction error, they would to fire to presentation of the CS1 cue (after the association of CS1 and reward has been learnt). At the time of the cue, no reward is presented, so how can there be a reward prediction error?

2.2 temporal difference learning

To understand the above 2 effects (2nd order conditioning and dopamine firing to the cue), we need to extend the RL framework to continous time. Sutton first introduced this idea of prediction based on continuous time, rather than trial bsed, proposing so-called temporal difference learning. Rather than reflecting expectation of reward at each precise point in time (as RW does), in TD learning the EV reflects the expected reward at any point in time and into the future. So what we try to predict is:

$$EV_t = r_t + r_{t+1} + r_{t+2} + r_{t+3}...$$

Or moving one step forward in time

$$EV_{t+1} = r_{t+1} + r_{t+2} + r_{t+3}...$$

And taking these together

$$EV_t = r_t + EV_{t+1}$$

In other words, our EV aims to predict the current reward plus any rewards into the future. Then, the prediction error becomes

$$PE = (r_t + EV_{t+1}) - EV_t$$

This means there are now 2 ways in which there can be a positive prediction error:
As in RW, when right now I got something better than I expected

$$r_t > EV_t,$$

But also when my expectation about the future got better:

$$EV_{t+1} > EV_t$$

So after conditioning, seeing a cue that tells me that in the future, I will be getting a reward, will evoke the same dopamine response as just getting that reward before conditioning!
This can also explain 2nd order conditioning, as now the CS2 comes to predict that I will get CS1, which predicts that I will get a reward. In other words, seeing CS2 makes my future unexpectedly better.

2.3 Pavlovian versus instrumental conditioning

Up to this point, we have focussed on learning stimulus-reward associations, rather than action-reward associations. Instrumental conditioning, i.e. learning to make particular actions to get rewards, rather than just learning an association, is thought to proceed in the same way as Pavlovian conditioning: When an action is followed by a reward, this action acquires value. Also here 2nd order learning is possible, where one action is followed by another action to ultimately lead to a reward. This is for example how we can teach dogs complicated tricks. In this final section, we will discuss how we can learn about learning from observing the actions an animal makes, rather than the automatic, conditioned responses like salivation.

2.4 goal-directed versus habitual learning

Latent learning shows that animals can learn about the structure of their environment and apply this knowledge, even when they do not get rewards at the time (e.g. mouse in the maze). Such 'knowledge about the environment' is something that is not a part of simple conditioning theories like RW / TD learning: in these models, all that counts is the value of a stimulus, not their identity.

RW/TD learning states that once an animal has learnt to make a particular response to get a reward, the response *itself* has acquired a positive value, where this value does not contain information about the identity of the reward, only its size (or probability). This makes the counter-intuitive prediction that the value of outcome that follows the response suddenly changes, behaviour would not be changed appropriately. Rather the new value needs to be learnt by repeated exposure.

For example, the prediction is that when a hungry rat learns to press a lever for food, the lever acquires value. Then if the rat is fed to satiety (so it no longer values food), according to TD learning, it would continue to press the lever.This prediction can be tested when an instrumental conditioning paradigm is augmented with a devaluation phase.

Instrumental phase: agent learns to make response for reward, e.g. press a lever for food
Devaluation phase: devalue reward, e.g. by associating food with being sick (via poison injection) – this is a stronger test than just satiety, because no the food actually becomes aversive
Test phase: test in extinction (i.e. with no feedback), whether the agent continues to make the learnt response (i.e. presses the lever), despite the fact that the outcome the agent has learnt normally follows this behaviour, is now aversive.

It turns out that the counter-intuitive prediction of TD learning, i.e. continued lever pressing after devaluation, only holds when the agent has been trained for a *long* period on the task. Then, behaviour becomes insensitive to devaluation. This is also known as 'habitual behaviour': actions that are not guided by the knowledge that it will lead to a particular outcome, but rather is performed habitually/automatically.

In contrast, when the animal is trained only moderately on the task, it will strongly reduce or even completely abolish the lever-pressing after the outcome has been devalued. This is known as goal-directed behaviour, because the agent takes into account the outcome of its actions when making a response, and will not make responses when they lead to an outcome that is no longer desired.

3 References & Further Reading

Schultz W, Dayan P, Montague PR. (1997) A neural substrate of prediction and reward. Science (course reading)

Dolan RJ, Dayan P. (2013) Goals and habits in the brain. Neuron (coourse reading)

Cohen JY, Haesler S, Vong L, Lowell BB, Uchida N. (2012) Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature

Sutton & Barto. Reinforcement learning – an introduction. Particularly relevant is chapter 6 on TD learning.
Online version of the book can be found here: http://incompleteideas.net/sutton/book/the-book.html

Rescorla, R.A. & Wagner, A.R. (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement, Classical Conditioning II, A.H. Black & W.F. Prokasy, Eds., pp. 64–99. Appleton-Century-Crofts.

CD Fiorillo, PN Tobler, W Schultz (2003) Discrete coding of reward probability and uncertainty by dopamine neurons. Science