Temporal Difference Learning In-Depth Math: Computational Neuroscience

Flo Martinez Addiego (fam53@georgetown.edu)

November 2021

Abstract

In Neuroscience, Reinforcement Learning (RL) is a learning process in which an animal (or person!) will use previous experiences to improve the outcome of future choices. We can think of this in terms of *how* does the brain make decisions?

This is extra information not covered in lectures, but different task modules (e.g. writing this out or responding to my phone notifications) are typically represented in a brain structure known as the prefrontal cortex (PFC). A secondary structure known as the anterior cingulate cortex (ACC) is going to help us switch between task modules and choose the one that is more favorable.

The reward pathway in the brain goes from the **Ventral Tegmental Area** (VTA) to the **Nucleus Accumbens** to the **Prefrontal Cortex**. Dopamine (DA) is the main neurotransmitter in this pathway. The VTA has a lot of dopaminergic neurons and projects to the frontal areas. More specifically it is what is going to initiate the dopaminergic reward signal. The nucleus accumbens is related in turn to the anticipation of pleasure. The Prefrontal Cortex will help evaluate the reward and plays a role in higher-order cognition (making decisions). (*Side Bar: one of the commonalities of all drugs of abuse is an increase in extracellular DA concentration*).

The **orbitofrontal cortex** (OFC) responds to stimulus-award associations and computes the value/valence of a reward. There is evidence that the OFC can associate conditioned stimuli with their reward and is modulated by desire for the stimulus (e.g. Consider the bell to food for a dog scenario. If the dog is super hungry, the OFC lights up a ton because it really values the food! Conversely, if the dog is very full, it has reduced OFC activation when exposed to the food = devaluing the food!). Likewise the **amygdala** and **ventral striatum** (we will come back to the ventral striatum) also reflect the value of a stimulus.

1 The Rescorla-Wagner Equation And Temporal Difference Learning: Introduction

The Rescorla-Wagner Model is really good at modeling conditioning (except when it's not). But, before getting into instances when it's not good at modeling conditioning, we should talk about what *exactly* the Rescorla-Wagner equation does: the Rescorla-Wagner model is an example of classical conditioning where we look at associations between conditioned and unconditioned stimuli. (The conditioned stimulus is a learned stimulus and the unconditioned stimulus is any stimulus that naturally triggers a response). Rescorla-Wagner specifies the amount of learning (the change in the predicitve value of a stimulus) and claims that it depends on the difference between expected and actual reward.
Let's consider an example: we have a dog, a bell, and food. The bell is our conditioned stimulus (has no value), the food is our unconditioned stimulus

(yummy). Before training, ringing the bell has no real effect on the dog. Additionally, showing the food causes the dog to salivate (unconditioned response). There is no association between the conditioned and unconditioned stimuli.Before learning, we do not expect a reward, but we get one! This results in a positive prediction error and dopamine neurons will fire after the stimulus is presented, reinforcing the association between the conditioned and unconditioned stimuli. During training, we can ring the bell and then present food (and do that over and over such that the bell perfectly predicts the presentation of food!). Therefore, later on when we ring the bell, the dog will salivate (conditioned response). That is, after learning, we expect a reward. Since the stimulus now perfectly predicts the reward, there will be an increase in DA firing after the stimulus is presented.
The Rescorla-Wagner equation is modeled by:

$$v = wu \tag{1}$$

where v is the expected reward; u = stimulus; w = weight/salience.
Learning is done by adjusting the weight to minimize error between predicted reward and actual reward. That is, we want actual reward - expected reward to be zero. This can be done by adjusting the weight. Why does this make sense? Well, going back to our equation (1), we know the stimulus is constant and won't change, so to change our prediction, we change our w.
The prediction error (introduced above conceptually), δ is expressed by:

$$\delta = r - v \tag{2}$$

where r is actual reward and v is again the expected/predicted reward
The Learning Rule is expressed by:

$$w + \epsilon\delta u \tag{3}$$

where ϵ is the learning rate and δ is our prediction error
Rescorla-Wagner **FAILS** in secondary conditioning. Why? Consider: stimulus1 = bell; reward = food; stimulus2 = light. After training, stimulus1 perfectly predicts the reward! The δ is therefore zero and no learningcan happen! There is no predictive value for any second stimulus because we do not include time...How do we fix this? In comes **temporal difference learning**!

2 Temporal Difference Learning

Let's introduce time into the scenario.
Let Actual Future Rewards = R(t) where t is the particular time in the trial which is represented by $0 \leq t \leq T$. We can now write a new equation:

$$R(t) = \sum_{\tau=0}^{T-t} r(t+\tau) \tag{4}$$

Why do we need a τ? Well, τ will allow us to keep track of temporal predictions after time t (ex a prediction at a particular time).

For example, Let current time t = 2 and the end time T = 4. Let's plug into our equation and see how it goes from there:

$$R(t) = \sum_{\tau=0}^{T-t} r(t+\tau); R(2) = \sum_{\tau=0}^{4-2} r(2+\tau) = r(2+0) + r(2+1) + r(2+2) \quad (5)$$

What is this saying? At time 2, the actual future rewards I will receive is a sum of the actual future rewards at distinct time points

The expected future reward, v(t) is based on how we have been learning over time up until this time point, t. So, we need to be able to look back into the past. Recall equation 1. Now, we are going to add time into it just like we did for the actual future rewards.

$$v(t) = \sum_{\tau=0}^{t} w(\tau)u(t-\tau) \quad (6)$$

This tells us two things! (1): w(τ) lets us know how the weights have been previously updated up until and including time t and (2) u(t-τ) well let us know what the preceding cues have been. *Recall: τ is just counting from 0 to t. It's just a counter.*

Just as we have adapted the actual and predicted rewards to include time, we need to update the learning rule.

$$w(\tau) = w(\tau) + \epsilon(t)\delta(\tau)u(t-\tau) \quad (7)$$

Importantly, δ(t) = actual total reward - predicted total reward. This can be expressed as:

$$\delta(t) = \sum_{\tau=0}^{T-t} r(t+\tau) - v(t) \quad (8)$$

the first half is the rewrad we actually get and the second part is the prediction for the future at a particular time.

But! There's a huge problem! r(t + τ) requires that we look into the future and know the exact reward we will get. This is obviously impossible. So, what do we do? We rewrite and approximate!

$$\sum_{\tau=0}^{T-t} r(t+\tau) = r(t) + \sum_{\tau=0}^{T-t-1} r(t+1+\tau) \quad (9)$$

where r(t) is the current reward and r(t + 1 + τ) is everything after time t and approximates v(t+1)

So, we can rewrite this as:

$$\sum_{\tau=0}^{T-t} r(t+\tau) \approx r(t) + v(t+1) \tag{10}$$

and plug back into the equation for delta.

$$\delta(t) = r(t) + v(t+1) - v(t) \tag{11}$$

This also tells us two things: (1) r(t) + v(t+1) is the actual future reward and v(t) is the predicted, so actual - predicted. Also, v(t+1) - v(t) tells us how much does expected reward change from one timepoint to another.