Classical Conditioning IV: Temporal Difference (TD) learning

Understanding a new concept

Plan for this class

- (flipped) Puzzle 1: Second order conditioning
- (flipped) Puzzle 2: What does dopamine do in the brain?
- (In class) brainstorm to solve the puzzles
- How computational thinking can solve both puzzles:
  - Marr’s three levels
  - Deriving temporal difference learning
  - Evidence from the brain
Where were we?

\[ \Delta V(CS_i) = \eta[R_{US} - \sum_{j \in \text{trial}} V(CS_j)] \]

- Rescorla-Wagner learning suggests that we learn from prediction errors.
- We can think of the “learning rate” as an important factor determining how we balance old and new information.
  \[ V_{\text{new}}(CS) = (1 - \eta) \cdot V_{\text{old}}(CS) + \eta \cdot R \]
- Learning rate = Forgetting rate

but: second-order conditioning

phase 1: 🟢👉🏾

phase 2: 🟢👉🏾

test: 🟢❓

What does the Rescorla-Wagner model predict?
A. animals will salivate to the light
B. animals will not salivate to the light
C. if there are few phase 2 trials they will salivate, otherwise not due to extinction of the tone-food association
but: second-order conditioning

What do YOU think will happen?
A. animals will salivate to the light
B. animals will not salivate to the light
C. if there are few phase 2 trials they will salivate, otherwise not due to extinction of the tone-food association

animals learn that a predictor of a predictor is also a predictor! ⇒ not interested solely in predicting immediate reinforcement.
Challenge:

$$ RW: \ V_{\text{new}}(CS) = V_{\text{old}}(CS) + \eta[R_{\text{us}} - V_{\text{old}}(CS)] $$

- Can you modify the R-W learning rule to account for second order conditioning?
- Points to think about before class:
  - what is the fundamental problem here?
  - ideas how to solve it?

a hint and second puzzle: dopamine

Parkinson’s Disease
→ Motor control / initiation?

Drug addiction, gambling, Natural rewards
→ Reward pathway?
→ Learning?

Also involved in:
- Working memory
- Novel situations
- ADHD
- Schizophrenia
- …
the anhedonia hypothesis (Wise, ’80s)

- Anhedonia = inability to experience positive emotional states derived from obtaining a desired or biologically significant stimulus
- Neuroleptics (dopamine antagonists) cause anhedonia
- Dopamine is important for reward-mediated conditioning
but...

predictable reward

omitted reward

---

Schultz et al., 1997

---

dopamine

Schultz et al., 1997
so... two puzzles

- behavioral puzzle: second order conditioning
- neural puzzle: dopamine responds to reward-predicting stimuli instead of to rewards
- How to solve these puzzles?

understanding the brain: where do we start?!

David Marr (1945-1980) proposed three levels of analysis:
1. the problem (Computational Level)
2. the strategy (Algorithmic Level)
3. how its actually done by networks of neurons (Implementational Level)
The problem: optimal prediction of future reinforcement

\[
V(S_t) = E \left[ \sum_{i=1}^{\infty} r(S_{t+i}) \right]
\]
want to predict expected sum of future reinforcement

\[
V(S_t) = E \left[ \sum_{i=1}^{\infty} \gamma^{i-1} r(S_{t+i}) \right]
\]
want to predict expected sum of discounted future reinf. (0<\(\gamma\)<1)
The problem: optimal prediction of future reinforcement

\[ V(S_t) = E \left[ \sum_{i=1}^{\infty} r(S_{t+i}) \right] \]

want to predict expected sum of future reinforcement

\[ V(S_t) = E \left[ \sum_{i=1}^{\infty} \gamma^{i-1} r(S_{t+i}) \right] \]

want to predict expected sum of discounted future reinf. (\(0<\gamma<1\))

\[ V(S_t) = E \left[ \sum_{i=t+1}^{t_{end}} r(S_i) \right] \]

want to predict expected sum of future reinforcement in a trial/episode
Temporal Difference (TD) learning

Marr’s 3 levels:

The problem: optimal prediction of future reinforcement

The algorithm:

\[
V(S_t) = E[r(S_{t+1})] + V(S_{t+1})
\]

\[
V_{new}(S_t) = V_{old}(S_t) + \eta \left[ r(S_{t+1}) + V_{old}(S_{t+1}) - V_{old}(S_t) \right]
\]

temporal difference prediction error \( \delta_{t+1} \)

compare to:

\[
V_{new}(CS) = V_{old}(CS) + \eta \left[ R_{US} - V_{old}(CS) \right]
\]

how does this solve the two puzzles?

beginning of trial: \( r(S_t) = 0 \)

\( V(S_{t-1}) = 0 \)

\( \delta_t = V(S_t) \)

middle of trial: \( r(S_t) = 0 \)

\( V(S_{t-1}) = 0 \)

\( \delta_t = V(S_t) - V(S_{t-1}) \)

end of trial: \( V(S_t) = 0 \)

\( \delta_t = r(S_t) - V(S_{t-1}) \)
δ_t = r(S_t)

δ_t = V(S_t) - V(S_{t-1})

δ_t = 0 - V(S_{t-1})

what would happen with partial reinforcement?
what would happen in second order conditioning?
Summary so far...

- Temporal difference learning is a “better” version of Rescorla-Wagner learning
- derived from first principles (from definition of problem)
- explains everything that R-W does, and more (eg. 2nd order conditioning)
- basically a generalization of R-W to real time

The problem: optimal prediction of future reinforcement
The algorithm: temporal difference learning
Neural implementation: does the brain use TD learning?
we already saw this:

\[ \delta_t = r(S_t) \]

\[ \delta_t = V(S_t) - V(S_{t-1}) \]

\[ \delta_t = V(S_t) - 0 - V(S_{t-1}) \]

The idea: Dopamine encodes a temporal difference reward prediction error

(Montague, Dayan, Barto mid 90's)
**prediction error hypothesis of dopamine**

Bayer & Glimcher (2005)

**where does dopamine project to?**

**main target:** striatum in basal ganglia (also prefrontal cortex)

---

Fiorillo et al, 2003

Model prediction error hypothesis of dopamine.
the basal ganglia: afferents
inputs to striatum are from all over the cortex (and they are topographic)

Voorn et al, 2004

a precise microstructure
dopamine and synaptic plasticity

• prediction errors are for learning…
• cortico-striatal synapses show dopamine-dependent plasticity
• three-factor learning rule: need presynaptic + postsynaptic + dopamine to strengthen synapse

summary

Thinking computationally about prediction learning

• The problem: prediction of future reward
• An algorithm: temporal difference learning
• Neural implementation: dopamine dependent learning in BG
  ⇒ Solves our puzzles: explains dopaminergic firing patterns, 2nd order conditioning
  ⇒ Compelling account for the role of dopamine in classical conditioning: prediction errors drive prediction learning
How can I use this in real life?

• Asking a question in science? Ask yourself: at which of Marr’s levels of analysis should it be asked?
  What is the system doing?
  How is it doing it?
  What is the specific implementation?

• Knowing what we know about dopamine: if I can measure your dopamine when you see a stimulus, I can find out how much reinforcer you are expecting! (more on this next time)

if you are confused or intrigued:
additional reading

• Sutton & Barto (1990) - Time derivative models of Pavlovian reinforcement - shows step by step why TD learning is a suitable rule for modeling classical conditioning
• Niv & Schoenbaum (2008) - Dialogues on prediction errors - a guide for the perplexed
• Barto (1995) - adaptive critic and the basal ganglia - very clear exposition of TD learning in the basal ganglia

(all on Canvas)