Markov Decision Processes & Q values

PSY/NEU338: from animal learning to changing people’s minds

how to model instrumental conditioning?

• The problem: find the best behavioral policy (what to do in what situation)
• A bit more formally: Markov decision process (S,A,R,T)
more formally: MDPs

transitions: \( P(b|a, \text{left}) = 90\%; P(c|a, \text{left}) = 10\% \text{ etc. (wonky shopping cart)} \)

The Markov property

- The idea: given the current situation, history does not matter
- \( P(S_{t+1}|S_1, S_2, \ldots, S_t, a_1, a_2, \ldots, a_t) = P(S_{t+1}|S_t, a_t) \)
- \( P(r_t|S_1, S_2, \ldots, S_t, a_1, a_2, \ldots, a_t) = P(r_t|S_t, a_t) \)
- Examples of MDPs? Counter examples?
Stylized task: described fully by S,A,R,T

World: “You are in state 34. Your immediate reward is 3. You have 2 actions”
Robot: “I’ll take action 1”

World: “You are in state 77. Your immediate reward is -7. You have 3 actions”
Robot: “I’ll take action 3”

The task description requires no memory
(doesn’t mean that the decision maker does not use memory to solve the task!)

learning a policy for MDPs

(policy dependent) State values: $V^\pi(S) = E[\text{sum of future rewards} \mid S, \pi]$
computing the value of actions

(policy dependent) State-Action values: \( Q^\pi(S, a) = E[\text{sum of future rewards} | S, a, \pi] \)

- \( Q(S_0, L) = ? \)
- \( Q(S_0, R) = ? \)
- which action is better?

learning optimal policies

Optimal policy: in terms of future rewards; a policy that obtains the largest possible amount of reward overall

How to learn an optimal policy?

OPTION 1: “batch” algorithm:

- behave according to current policy
- estimate \( Q \) values based on experience
- improve policy based on these \( Q \) values
- repeat
learning optimal policies

Optimal policy: in terms of future rewards; a policy that obtains the largest possible amount of reward overall

How to learn an optimal policy?

OPTION 2: “online” algorithm:

- behave according to current Q values
- calculate prediction error after every action
- update Q value based on prediction error
- repeat

SARSA versus Q-learning

\[ \delta_{t+1} = R_{t+1} + Q(S_{t+1}, a_{t+1}) - Q(S_t, a_t) \] (SARSA)

or

\[ \delta_{t+1} = R_{t+1} + \max_a Q(S_{t+1}, a) - Q(S_t, a) \] (Q-learning)

- choose actions according to softmax: \( p(S, a) \propto e^{\beta Q(S,a)} \)
compare to:

summary so far...

- Modeling instrumental conditioning (action selection): several models proposed (Q learning, SARSA, Actor/Critic)
- in all cases reinforcement learning uses **predictive values** to inform **choice**
- remember: **all this works only in MDPs**
  (but many problems can be represented as MDPs or approximated by MDPs)
5 min break

does the brain really use actor/critic learning?
how can we tell the models apart?

- $\delta_{t+1} = R_{t+1} + Q(S_{t+1}, a_{t+1}) - Q(S_t, a_t)$ (SARSA)
- $\delta_{t+1} = R_{t+1} + \max_a Q(S_{t+1}, a) - Q(S_t, a_t)$ (Q-learning)
- $\delta_{t+1} = R_{t+1} + V(S_{t+1}) - V(S_t)$ (Actor/Critic)

what do dopamine prediction errors represent at trial onset?
what do dopamine prediction errors represent at trial onset?

\[ P_{\text{left choice}} = 0.2 \]
\[ P_{\text{right choice}} = 0.8 \]

\[ V(I) = 0.85 \]
\[ Q(L) = 0.25 \]

\[ V(R) = 0.85 \]
\[ Q(R) = 1 \]

\[ P_{\text{reward}} = 0.25 \quad P_{\text{reward}} = 1.0 \]

what do dopamine prediction errors represent at trial onset?

\[ \text{Response (spikes per s)} \]
\[ R^2 = 0.972 \]

\[ \text{Reward receipt} \]
\[ R^2 = 0.797 \]
summary so far...

- **In the brain**: evidence for division between prediction learning and policy learning (Actor/Critic)
- But: nature of prediction errors themselves suggests otherwise (they look more like SARSA)
- Not only do the models inform us about the brain, but the brain can inform us about the models!

How is this relevant to real life?

- Reinforcement learning depends on experience. Fundamentally. You don’t learn new values for what you don’t try out. Let that sink in.
- Implications for: anxiety
  (you avoid stuff you think has a low value.. thereby not giving yourself a chance to learn that your estimated value was wrong!)
- Implications for: avoiding risks
  (you don’t choose an option you think is worse.. thereby not giving yourself a chance to learn that it is actually not that bad)