Reinforcement Learning in the brain


Animals learn predictions -- Pavlovian conditioning

- animals learn predictions
- conditioned suppression
  - http://www.youtube.com/watch?v=Zlf6dx1P1q4
- autoshaping
  - http://www.youtube.com/watch?v=8scwAvg8CEA

Rescorla & Wagner (1972)

- Most influential model of animal learning, explains puzzling behavioural phenomena such as blocking, overshadowing and conditioned inhibition.
- The idea: error-driven learning:
  Learning occurs only when events violate expectations.

Change in value is proportional to the difference between actual and predicted outcome

\[ V_{\text{new}}(CS_i) = V_{\text{old}}(CS_i) + \eta \left[ \lambda V_{\text{old}}(US) - \sum V_{\text{old}}(CS_i) \right]. \]

- learning only occurs when events not predicted
- predictions due to different stimuli are summed to form the total prediction in a trial.
How do we know that animals use an error-correcting rule?

- blocking
- interpretation: the bell fully predicts the food and the presence of the light adds no new predictive information – therefore no association develops to the light.

Temporal Difference (TD) learning (1)

- Consider a succession of states $S$, following each other with $P(S_{t+1} | S_t)$
- Rewards observed in each state with probability $P(r | S_t)$
- A useful quantity to predict is the expected sum of all future rewards, given current state $S_t$, $V(S_t)$

$$V(S_t) = E \left[ r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots | S_t \right] = E \left[ \sum_{i=t}^{\infty} \gamma^i r_i | S_t \right]$$

- Discount factor introduced to make sure that the sum is finite, but also humans and animals prefer earlier rewards to later ones
- Incorporating probabilities $P(S_{t+1} | S_t)$ and $P(r | S_t)$, we get recursive form

$$V(S_t) = E \left[ r_t + \gamma E [r_{t+1} | S_t] + \gamma^2 E [r_{t+2} | S_t] + \ldots \right]$$

= $E [r_t | S_t] + \gamma \sum_{S_t} P(S_{t+1} | S_t) E [r_{t+1} | S_{t+1}] + \gamma^2 \sum_{S_t} P(S_{t+1} | S_t) E [r_{t+2} | S_{t+2}] + \ldots$

= $P(r | S_t) + \gamma \sum_{S_t} P(S_{t+1} | S_t) V(S_{t+1})$

Limitations of Rescorla & Wagner (1972)

- does not extend to 2d order conditioning.
  - A→B→reward; A gains reward predictive value

- Basic unit of learning = conditioning trial as discrete temporal object fails to account for the temporal relations between condition and unconditional stimuli within a trial

- TD learning as a means to overcome these limitations = extension of Rescorla Wagner to take into account timing of events.

Temporal Difference (TD) learning (2)

- When estimated values are incorrect, there is a discrepancy between 2 sides of equation: prediction error:

$$\delta_t = P(r | S_t) + \gamma \sum_{S_t} P(S_{t+1} | S_t) V(S_{t+1}) - V(S_t)$$

- Prediction error is a natural signal for improving estimates $V(S_t)$, giving

$$V(S_t)_{new} = V(S_t)_{old} + \eta \cdot \delta_t$$

= Optimal learning rule, basis of “dynamic programming”.

- One problem: assumes knowledge of $P(S_{t+1} | S_t)$ and $P(r | S_t)$ which is unreasonable in basic learning situations.

- Model-free Approximation which can be formally justified:

$$\delta_t = r_t + \gamma V(S_{t+1}) - V(S_t)$$

~ current reward + next prediction - current prediction
Temporal Difference (TD) learning (3)

- Resulting learning rule:
  \[ V_{\text{new}}(S_t) = V_{\text{old}}(S_t) + \eta (r_t + \gamma V(S_{t+1}) - V(S_t)). \]

- Incorporating Rescorla-Wagner idea that predictions due to different stimuli are additive:
  \[ V_{\text{new}}(S_{a,j}) = V_{\text{old}}(S_{a,j}) + \eta \left[ r_t + \gamma \sum_{S_{k+1}} V_{\text{old}}(S_{k+1}) - \sum_{S_{j+1}} V_{\text{old}}(S_{j+1}) \right]. \]

- This is TD learning rule as proposed by Sutton & Barton (1990)

Instrumental conditioning: adding control

- Animals not only learn associations between stimuli and reward but also between actions and reward
- Learning to select actions that will increase the probability of rewarding events and decrease the probability of aversive events.
- Rat lever pressing in boxes -- operant conditioning (Skinner)

Actor/Critic Methods

- How can such action selection be learned?
- Problem of credit assignment
- RL: base action selection not only on immediate outcomes but also future value predictions.
- Barto (1983) shows that credit assignment problem can be solved by a learning system comprised of 2 neurons-like elements:
  - The critic, uses TD learning to construct values of states
  - The actor, selects actions at each state using prediction error.

Idea: if positive prediction error is encountered, current action has improved prospects for the future and should be repeated.

Learning of policies:
\[ \pi(S,a) = p(a|S), \quad \pi(S,a)_{\text{new}} = \pi(S,a)_{\text{old}} + \eta \delta_t \]
Q learning

- Watkins (1989)
- Alternative: explicitly learn the predictive value (future expected rewards) of taking an action at each state, = learn the value of state-action pairs $Q(S,a)$
- learning rule:
  $$Q(S_t,a_t)_{new} = Q(S_t,a_t)_{old} + \eta \delta_t$$

- TD prediction error:
  $$\delta_t = r_t + \max_a \gamma Q(S_{t+1}, a) - Q(S_t,a_t)$$

How does the brain do reinforcement learning?

- "the largest success of computational neuroscience", dopamine and prediction error

What is Dopamine?

- Parkinson's Disease: motor control/ initiation
- addiction, gambling, natural rewards
- also involved in: working memory, novel situations, ADHD, schizophrenia

Former idea: Dopamine signals reward (Wise, ‘80s)

- Initial idea: dopamine might represent reward signals
- neuroleptics (dopamine antagonists) cause anhedonia
- brain self stimulation by rats  
  http://www.youtube.com/watch?v=7HbAFYeivo
- dopamine important for reward mediated conditioning
New idea: phasic dopamine signals prediction error

- Schultz et al 90s
- monkeys underwent simple instrumental or pavlovian conditioning
- disappearance of dopaminergic response at reward delivery after learning
- if reward is not presented, response depression below basal firing at expected time of reward.

Schultz, Dayan, Montague, 1997

Monday, 8 March 2010

dopamine and prediction

- The idea: dopamine encodes prediction error (Montague, Dayan, Barto, 1996)
- provided normative basis for understanding not only why dopamine neurons fire when they do, but also what the function of these firing might be.
- evidence for dopamine dependent, or dopamine gated plasticity in synapses between cortex and striatum.

Schultz et al, 1993

Monday, 8 March 2010

Prediction error: stringent tests

- checking that size of response at onset of CS is proportional to reward size

Tobler et al, 2005

- Bayer & Glimcher, Neuron, 2005
- firing rates of dopamine neurons following delivery of reward encode a computation reflecting the difference between the current reward and a recency-weighted average of previous rewards

Monday, 8 March 2010
fMRI data

• fMRI to study the underpinnings of RL in the human brain
• model driven analysis -- search the brain for predicted hidden variables that should control learning and decision making, eg state values and prediction errors.

• prediction errors signals found in nucleus accumbens and orbito frontal cortex, both major dopaminergic targets.

• O Doherty et al (2004) show that FMRI correlates of prediction error signals can be dissociated in dorsal and ventral striatum according to whether instrumental conditioning vs pavlovian condition, -- supporting an Actor/Critic architecture.

Summary

• Optimal learning depends on prediction and control

• the problem: prediction of future reward

• the algorithm: TD learning

• neural implementation: dopamine dependent learning in cortico-striatal synapses in basal ganglia

• RL has revolutionised how we think of learning in the brain implications for the understanding of disorders, such as Parkinson’s and schizophrenia, as well as addiction.