Reinforcement learning (RL):
- an area of machine learning inspired by behaviorist psychology, concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.
- thought to be a good model of how learning is occurring in the brain.

Maximizing reward as a guide to decision-making
- Decision making at all levels
- Why is this hard: (1) rewards/ punishment may be delayed; (2) outcome may depend on series of actions (credit assignment problem)
- Need learning of predictions of events and actions

Animals learn predictions — Pavlovian conditioning
- Animals learn predictions
- Classical conditioning: pairing of a CS with a US
- example: conditioned suppression
  [http://www.youtube.com/watch?v=OMwStQn8b5k](http://www.youtube.com/watch?v=OMwStQn8b5k)
- autosshaping
  [http://www.youtube.com/watch?v=mjnxv4ygElA](http://www.youtube.com/watch?v=mjnxv4ygElA)
Rescorla & Wagner model of classical conditioning (1972)

- Most influential model of animal learning, explains puzzling behavioural phenomena such as blocking, overshadowing and conditioned inhibition.
- describe changes in associative strength ($V$) between a signal (conditioned stimulus CS) and subsequent stimulus (unconditioned stimulus US)
- 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

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

Learning only occurs when events not predicted
- predictions due to different stimuli are summed to form the total prediction in a trial.

Limitations of Rescorla & Wagner (1972)

- does not extend to 2d order conditioning.
- A->B->reward; where A gains reward predictive value
- Basic unit of learning = conditioning trial as discrete temporal object fails to account for the temporal relations between CS and US stimuli within a trial
- Temporal Difference (TD) learning as a means to overcome these limitations = extension of Rescorla-Wagner to take into account timing of events.

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 $P(S_{t+1} \mid S_t)$
- Rewards observed in each state with probability $P(r \mid S)$
- Useful quantity to predict is the expected sum of all future rewards, given current state $S_t$, value of state $S$, $V(S)$

$$V(S_t) = E \left[ r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ... \mid S_t \right] = E \left[ \sum_{i=0}^{\infty} \gamma^i r_{t+i} \mid S_t \right]$$

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

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

$$E \left[ r_t \mid S_t \right] + \gamma \sum P(S_{t+1} \mid S_t) \left( E \left[ r_{t+1} \mid S_{t+1} \right] + \gamma E \left[ r_{t+2} \mid S_{t+1} \right] + ... \right) =$$

$$P(r \mid S_t) + \gamma \sum P(S_{t+1} \mid S_t) V(S_{t+1})$$
Temporal Difference (TD) learning (2)

- When estimated values are incorrect, there is a discrepancy between 2 sides of equation: prediction error:
  \[ \delta_t = P(r_t|s_t) + \gamma \sum_{t=1}^{\infty} P(s_{t+1}|s_t) V(s_{t+1}) - V(s_t). \]
  prediction error is a natural signal for improving estimates \( V(s_t) \), giving
  \[ V(s_t)_{\text{new}} = V(s_t)_{\text{old}} + \eta \cdot \delta_t, \]
- Optimal learning rule, basis of "dynamic programming".
- One problem: assumes knowledge of \( P(s_{t+1}|s_t) \) and \( P(r_t|s_t) \) which is unreasonable in basic learning situations.
- Model-free Approximation which can be formally justified (sampling):
  \[ \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \]
  - current reward + next prediction - current prediction

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)

http://www.youtube.com/watch?v=1_clJqHtHIA  (Interview of Skinner)

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:
  \[ V_{\text{new}}(s_t) = V_{\text{old}}(s_t) + \eta \left[ r_t + \gamma \sum_{s_{t+1}} V_{\text{old}}(s_{t+1}) - \sum_{s_{t+1}} V_{\text{old}}(s_{t+1}) \right]. \]
- This is TD learning rule as proposed by Sutton & Barto (1990)

Actor/Critic Methods

- How can such action selection be learned?
- 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, learn to select 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:
\[ \pi(S,a) = p(a|S), \] \[ \pi(S,a)_{\text{new}} = \pi(S,a)_{\text{old}} + \eta \pi \delta_t \]
A recent application of Q-learning to deep learning, by Google DeepMind has been successful at playing some Atari 2600 games at expert human levels. Preliminary results were presented in 2014, with a paper published in February 2015 in Nature.

**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:
  \[
  Q(S_t, a_t)_{\text{new}} = Q(S_t, a_t)_{\text{old}} + \eta \delta_t
  \]
- TD prediction error:
  \[
  \delta_t = r_t + \max_a Q(S_{t+1}, a) - Q(S_t, a_t)
  \]
  - current reward + prediction of next best action - current prediction

**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

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.

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

- Initial idea: dopamine might represent reward signals
- Antipsychotic drugs (dopamine antagonists) cause anhedonia
- Brain self stimulation by rats
- Dopamine important for reward mediated conditioning

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.

Dopamine Response = RewardOccurred – RewardPredicted.

Schultz, Dayan, Montague, 1997

http://www.youtube.com/watch?v=7HbAFYiejvo
• checking that size of response at onset of CS is proportional to reward size

**Tobler et al, 2005**

<table>
<thead>
<tr>
<th>0.0 ml</th>
<th>0.025 ml</th>
<th>0.075 ml</th>
<th>0.15 ml</th>
<th>0.25 ml</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Graph" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Onset of conditioned stimuli predicting expected reward value**

• Actor/Critic architecture. Whether instrumental conditioning vs pavlovian condition, supporting an prediction errors.

**O Doherty et al (2004) show that FMRI correlates of prediction error**

**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.
Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions


- Frontal cortex responses in the patient group were suggestive of disrupted prediction-error processing.
- Across subjects, the extent of disruption was significantly related to an individual’s propensity to delusion formation.

Application to Psychiatry

Model based vs Model Free

- Debated how much human learning is “model-free” vs “model-based”
- Model free corresponds to habit, inflexible
- Possibly relevant to pathology

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.