Reinforcement Learning in the Brain

• Reading: Y Niv, *Reinforcement learning in the brain*, 2009.

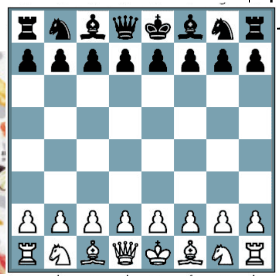
Reinforcement learning (RL):

- 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.

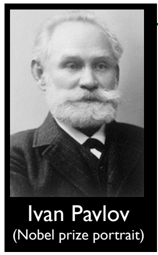
- thought to be a good model of how learning is occurring in the brain.

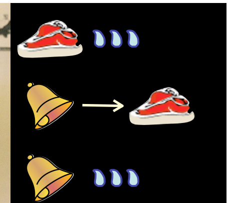
Maximizing reward as a guide to decision-making

• Decision making at all levels
• Reinforcement learning : maximize reward and minimize punishments; Sutton 1978; Sutton & Barto, 1990, 1998.
• 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=ZlZekx1P1g4

• autoshaping
http://www.youtube.com/watch?v=cacwAvgg8EA

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.

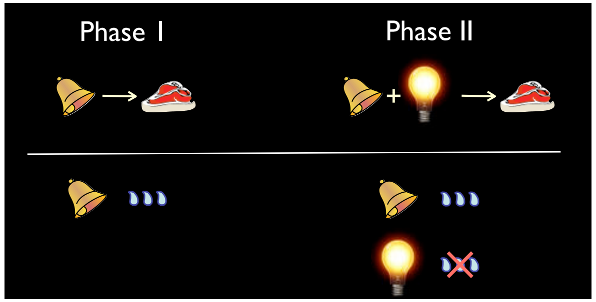
Change in value is proportional to the difference between actual and predicted outcome

$$V_{new}(CS_i) = V_{old}(CS_i) + \eta \left[\lambda_{US} - \sum_i V_{old}(CS_i) \right].$$

• 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.



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.

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 $P(r|S_t)$
• Useful quantity to predict is the expected sum of all future rewards, given current state S_t, = value of state S, $V(S_t)$

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

• 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

$$\begin{aligned}
V(S_t) &= E[r_t|S_t] + \gamma E[r_{t+1}|S_t] + \gamma^2 E[r_{t+2}|S_t] + \ldots = \\
&= E[r_t|S_t] + \gamma \sum_{S_{t+1}} P(S_{t+1}|S_t)\left(E[r_{t+1}|S_{t+1}] + \gamma E[r_{t+2}|S_{t+1}] + \ldots\right) = \\
&= P(r|S_t) + \gamma \sum_{S_{t+1}} P(S_{t+1}|S_t) V(S_{t+1})
\end{aligned}$$

Temporal Difference (TD) learning (2)

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

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

• 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 (sampling):

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

~ current reward+next prediction - current prediction

Temporal Difference (TD) learning (3)

• Resulting learning rule:

$$V_{new}(S_t) = V_{old}(S_t) + \eta(r_t + \gamma V(S_{t+1}) - V(S_t)).$$

• Incorporating Rescorla-Wagner idea that predictions due to different stimuli are additive:

$$V_{new}(S_{i,t}) = V_{old}(S_{i,t}) + \eta \left[r_t + \gamma \sum_{S_k @ t+1} V_{old}(S_{k,t+1}) - \sum_{S_j @ t} V_{old}(S_{j,t}) \right],$$

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

Instrumental conditioning: adding control

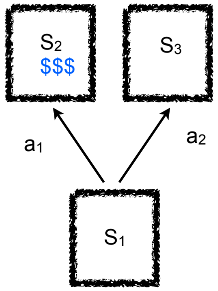
• Animals not only learn associations between stimuli and reward but also between actions and reward
• Learning to select actions that will increase the probability of rewarding events and decrease the probability of aversive events.
• rat lever pressing in boxes -- operant conditioning (Skinner)

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

Actor/Critic Methods

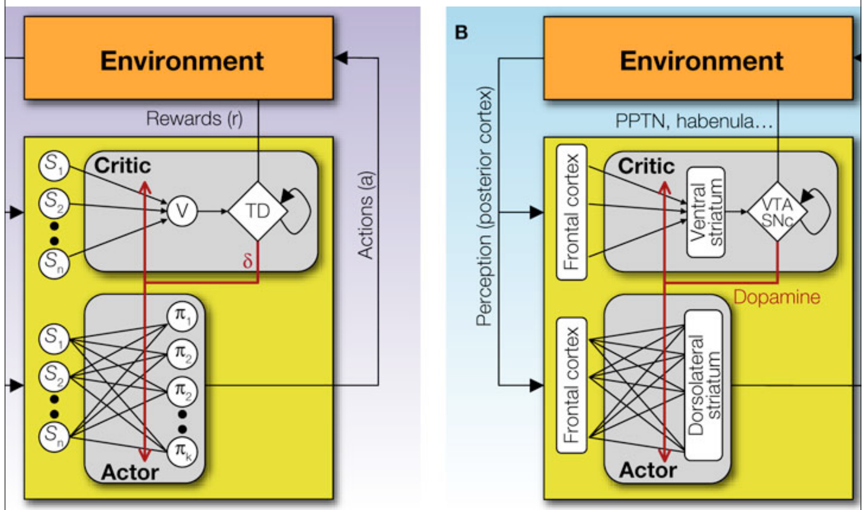
• How can such action selection be learned?

• 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:

$$\pi(S,a) = p(a|S). \qquad \pi(S,a)_{new} = \pi(S,a)_{old} + \eta_\pi \delta_t$$



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:

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

~ current reward+ prediction of next best action- current prediction

Machine learning applications of Q learning

Google DeepMind

LETTER

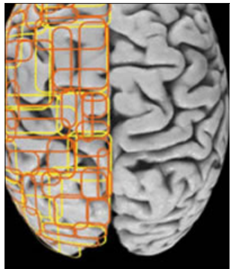
doi:10.1038/nature14236

Human-level control through deep reinforcement learning

Volodymyr Mnih[1]*, Koray Kavukcuoglu[1]*, David Silver[1]*, Andrei A. Rusu[1], Joel Veness[1], Marc G. Bellemare[1], Alex Graves[1], Martin Riedmiller[1], Andreas K. Fidjeland[1], Georg Ostrovski[1], Stig Petersen[1], Charles Beattie[1], Amir Sadik[1], Ioannis Antonoglou[1], Helen King[1], Dharshan Kumaran[1], Daan Wierstra[1], Shane Legg[1] & Demis Hassabis[1]

https://www.youtube.com/watch?v=V1eYniJ0Rnk

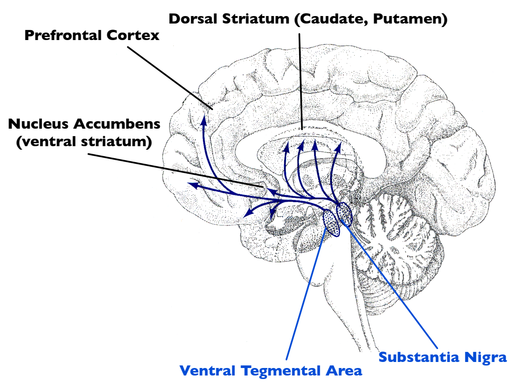
A recent application of Q-learning to deep learning, by Google DeepMind has been successful at playing some Atari 2600 games at expert human levels. Preliminary results were presented in 2014, with a paper published in February 2015 in Nature.



How does the brain do reinforcement learning ?

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

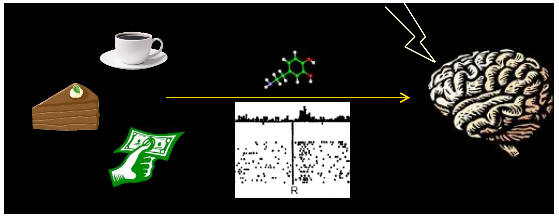
What is Dopamine ?



• 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)

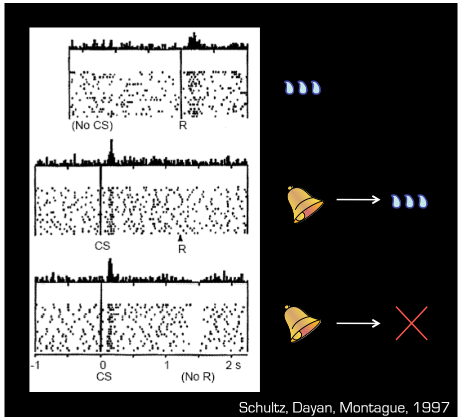
• Initial idea: dopamine might represent reward signals
• antipsychotic drugs (dopamine antagonists) cause anhaedonia
• brain self stimulation by rats http://www.youtube.com/watch?v=7HbAFYiejvo
• dopamine important for reward mediated conditioning



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.

DopamineResponse = RewardOccurred – RewardPredicted.

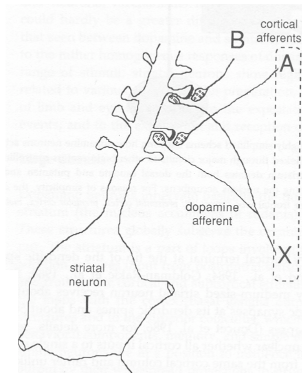


Dopamine and Prediction

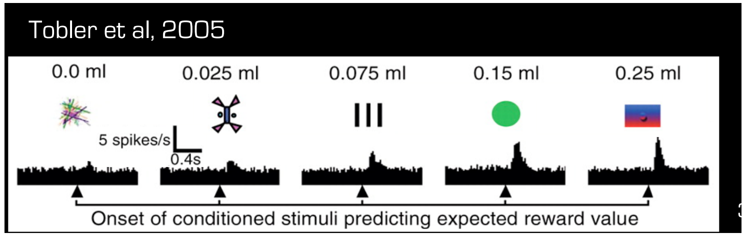
• The idea: dopamine encodes prediction error (Montague, Dayan, Barto, 1996)

• provided normative basis for understanding not only why dopamine neurons fire when they do, but also what the function of these firing might be.

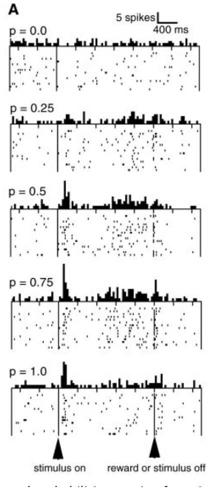
• evidence for dopamine dependent, or dopamine gated plasticity in synapses between cortex and striatum.



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



• checking that size of response at onset of CS is proportional to reward probability (Fiorillo et al, Science 2003)



fMRI data

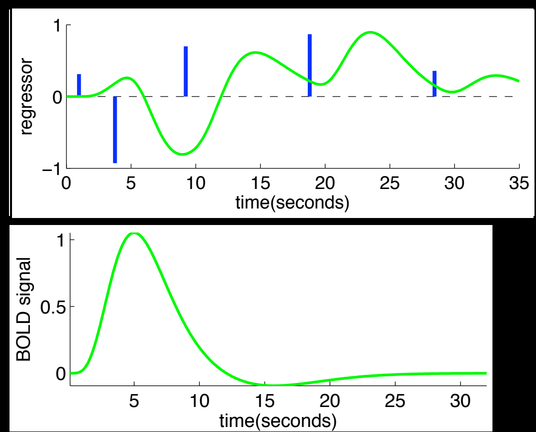
• fMRI to study the underpinnings of RL in the human brain

• 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.

short aside: functional magnetic resonance imaging (fMRI)



Application to Psychiatry

doi:10.1093/brain/awm173 Brain (2007), 130, 2387–2400

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

P. R. Corlett,[1] G. K. Murray,[1,2] G. D. Honey,[1] M. R. F. Aitken,[3] D. R. Shanks,[4] T. W. Robbins,[3] E. T. Bullmore,[1,2] A. Dickinson[3] and P. C. Fletcher[1]

- 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

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

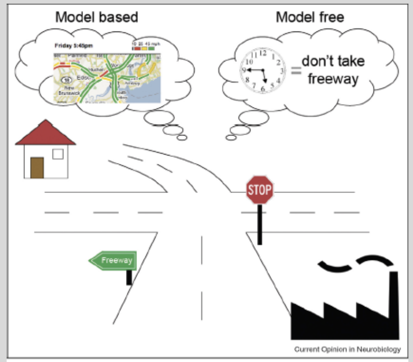


Figure 1: Two ways to choose which route to take when traveling home from work on friday evening.

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.