Neural Representation of Value, Reward and Expectation

Peter Dayan
Gatsby Computational Neuroscience Unit
Nathaniel Daw Daphna Joel Read Montague Yael Niv

Conditioning

- Pavlovian conditioning
  - prediction learning
  - temporal difference (TD) prediction error
- instrumental conditioning
  - active choices to maximize rewards; minimize punishment
  - actions 'stamped-in' by reinforcement
  - actions & habits

Conditioning

prediction: of important events
control: in the light of those predictions

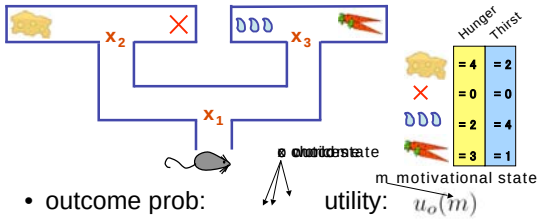
- Ethology
  - optimality
  - appropriateness
- Psychology
  - classical/operant conditioning
- Computation
  - dynamic progr.
  - Kalman filtering
- Algorithm
  - TD/delta rules
  - simple weights
- Neurobiology

neuromodulators; amygdala; OFC
nucleus accumbens; dorsal striatum

Plan

- phasic dopamine
  - TD prediction error for long term reward
  - TD prediction error for actor/critic
  - SARSA learning signal for Q values
- tonic dopamine
  - long run average rate of reward (Niv et al)
  - vigour controller
- striatum, amygdala, OFC

Reinforcement Learning

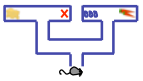


- outcome prob: utility: $u_o(m)$
- from x_3 average value for choice c best choice: c^*

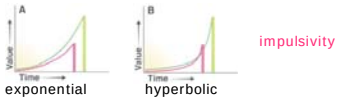
$$Q_c(x,m) = \sum_o Q_c(x,o)\,u_o(m)$$

$$c^*(x,m) = \arg\max_c \{Q_c(x,m)\}$$

RL Methods

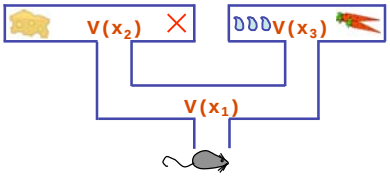


- Given delayed outcomes:
  - utility function



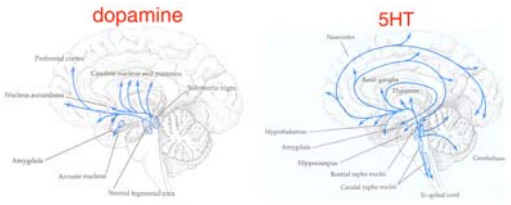
  - or average case:
  - plus transition probabilities: $\mathcal{T}_{xy}(c)$

TD: Consistency over Time



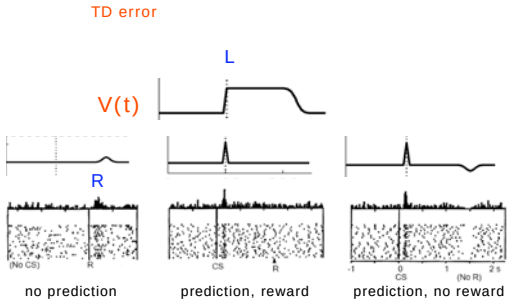
- want:
- error:
- more generally:

Where to Look?



- general: excitability, signal/noise ratios
- specific: prediction errors, uncertainty signals
  - drug addiction; self-stimulation;
  - psychiatry; neurology

TD Error and Dopamine



dopamine cells in VTA/SNc

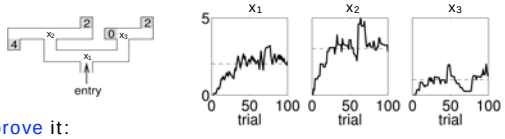
Schultz *et al*

Prediction Pathway

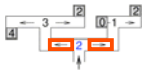
Actor/Critic

start with policy:

evaluate it:



improve it:



1

-1

thus choose L more frequently than R

Actor-Critic Pathway

- stimulus-response links
- inflexible
  - bad for motivational changes
  - good for modularity
  - subtle: no natural convergence

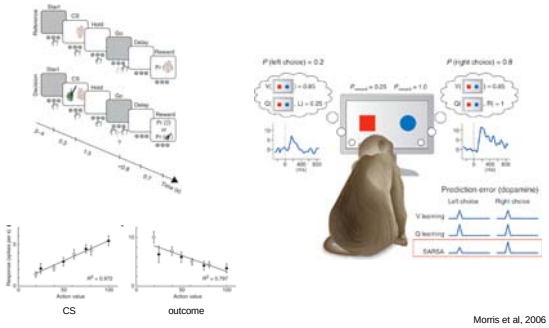
Anatomically

Q-learning

- action-values $Q_c(x, m)$
  - well-defined asymptotic limit
- Q-learning (Watkins)

- SARSA (Rummery et al)

Dopamine Implications



Morris et al, 2006

Average Reward RL (Niv)

Compute differential values of actions

Differential value of taking action L with latency τ when in state x

$\bar{\rho}$ = average rewards minus costs, per unit time

$Q_{L,\tau}(x_1)$ = Rewards – Costs + Future Returns

- steady state behavior (not learning dynamics)
- deriv: min:

(Extension of Schwartz 1993)

Average Reward Cost/benefit Tradeoffs

1. Which action to take?

⇒ Choose action with largest expected reward minus cost

2. How fast to perform it?

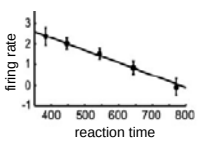
- slow → less costly (vigour cost)
- slow → delays (all) rewards
- net rate of rewards = cost of delay (opportunity cost of time)

⇒ Choose rate that balances vigour and opportunity costs

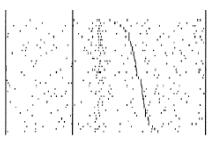
explains faster (irrelevant) actions under hunger, etc

Tonic dopamine hypothesis

…explains effects of phasic dopamine on response times



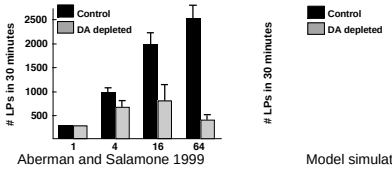
Satoh and Kimura 2003



Ljungberg, Apicella and Schultz 1992

Tonic dopamine = Average reward rate

1. explains pharmacological manipulations
2. dopamine control of vigour through BG pathways



Aberman and Salamone 1999

Model simulation

- eating time confound
- context/state dependence (motivation & drugs?)

NB. phasic signal RPE for choice/value learning

Summary

- phasic dopamine as TD prediction error
  - value at time of CS
  - SARSA?
  - amygdala; mPFC; (lateral habenula)
  - opponency (and serotonin)?
  - uncertainty?
- tonic dopamine as average reward rate
  - vigour
- integration with goal-directed value?
- Pavlovian effects on instrumental actions