Markov Decision Processes and Reinforcement Learning

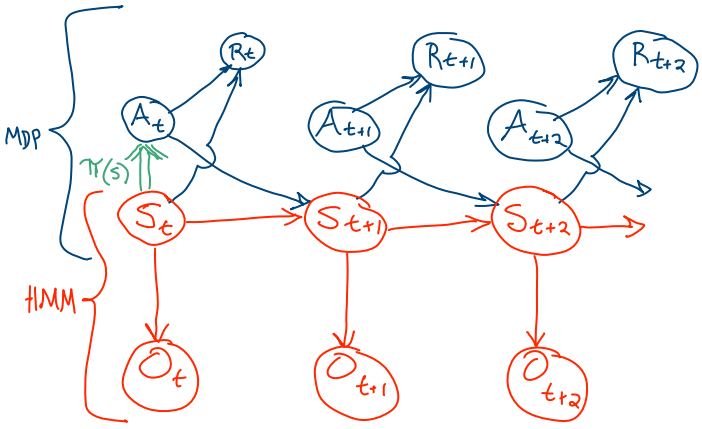
Part II

Machine Learning 10-701

April 28, 2010

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

HMM, Markov Process, Markov Decision Process



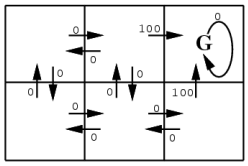
Immediate rewards r(s,a)

State values V*(s)

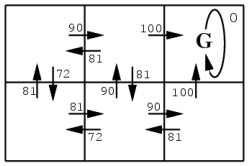
State-action values Q*(s,a)

$$V^*(s) = E[r(s, \pi^*(s))] + \gamma E_{s'|s,\pi^*(s)}[V^*(s')]$$

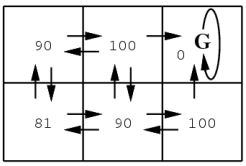
Bellman equation.



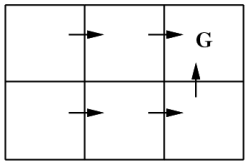
$r(s,a)$ (immediate reward) values



$Q(s,a)$ values



$V^*(s)$ values



One optimal policy

Q Learning for Deterministic Worlds

For each s, a initialize table entry $\hat{Q}(s,a) \leftarrow 0$

Observe current state s

Do forever:

- Select an action a and execute it
- Receive immediate reward r
- Observe the new state s'
- Update the table entry for $\hat{Q}(s,a)$ as follows:

$$\hat{Q}(s,a) \leftarrow r + \gamma \max_{a'} \hat{Q}(s',a')$$

- $s \leftarrow s'$

$\hat{Q}$ converges to Q. Consider case of deterministic world where see each $\langle s, a \rangle$ visited infinitely often.

Proof: Define a full interval to be an interval during which each $\langle s, a \rangle$ is visited. During each full interval the largest error in $\hat{Q}$ table is reduced by factor of γ

Let $\hat{Q}_n$ be table after n updates, and Δ_n be the maximum error in $\hat{Q}_n$; that is

$$\Delta_n = \max_{s,a} |\hat{Q}_n(s,a) - Q(s,a)|$$

For any table entry $\hat{Q}_n(s,a)$ updated on iteration $n+1$, the error in the revised estimate $\hat{Q}_{n+1}(s,a)$ is

$$\begin{aligned}
|\hat{Q}_{n+1}(s,a) - Q(s,a)| &= |(r + \gamma \max_{a'} \hat{Q}_n(s',a')) \\
&\quad -(r + \gamma \max_{a'} Q(s',a'))| \\
&= \gamma |\max_{a'} \hat{Q}_n(s',a') - \max_{a'} Q(s',a')| \\
&\leq \gamma \max_{a'} |\hat{Q}_n(s',a') - Q(s',a')| \\
&\leq \gamma \max_{s'',a'} |\hat{Q}_n(s'',a') - Q(s'',a')| \\
|\hat{Q}_{n+1}(s,a) - Q(s,a)| &\leq \gamma \Delta_n
\end{aligned}$$

Use general fact:
$|\max_a f_1(a) - \max_a f_2(a)| \leq \max_a |f_1(a) - f_2(a)|$

Nondeterministic Case

Q learning generalizes to nondeterministic worlds

Alter training rule to

$$\hat{Q}_n(s,a) \leftarrow (1-\alpha_n)\hat{Q}_{n-1}(s,a) + \alpha_n [r + \max_{a'} \hat{Q}_{n-1}(s',a')]$$

where

$$\alpha_n = \frac{1}{1 + visits_n(s,a)}$$

Can still prove convergence of $\hat{Q}$ to Q [Watkins and Dayan, 1992]

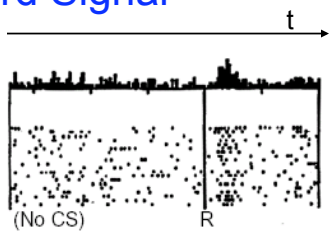
MDPs and Reinforcement Learning: Further Issues

- What strategy for choosing actions will optimize
  - learning rate? (*explore* uninvestigated states)
  - obtained reward? (*exploit* what you know so far)
- Can we bound sample complexity?
  - R-Max learns with δ, ε bounds in polynomial number of actions
- *Partially observable* Markov Decision Processes
  - state is not fully observable
  - maintain probability distribution over possible states you're in
- Convergence guarantee with function approximators?
  - our proof assumed a tabular representation for Q, V
- Correspondence to human learning?

Dopamine As Reward Signal

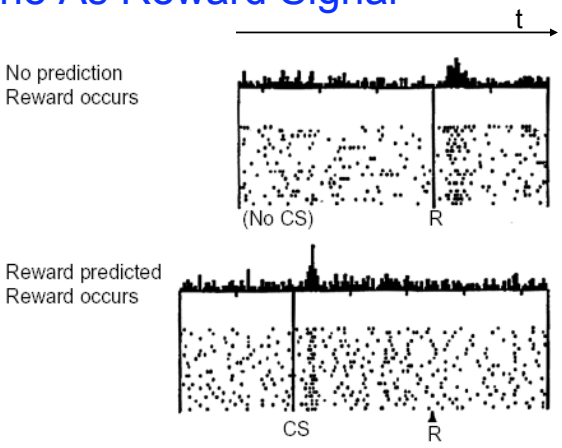
[Schultz et al., *Science*, 1997]

No prediction
Reward occurs



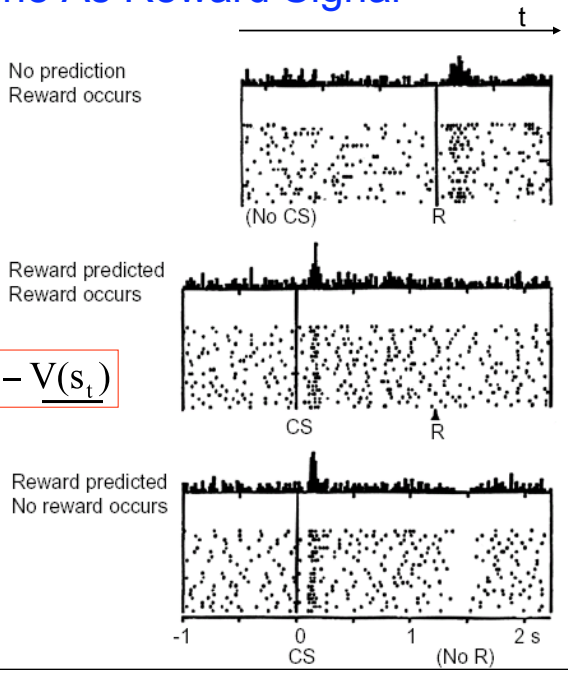
Dopamine As Reward Signal

[Schultz et al., *Science*, 1997]



Dopamine As Reward Signal

[Schultz et al., *Science*, 1997]



$$\text{error} = r_t + \gamma V(s_{t+1}) - V(s_t)$$

RL Models for Human Learning

[Seymore et al., Nature 2004]

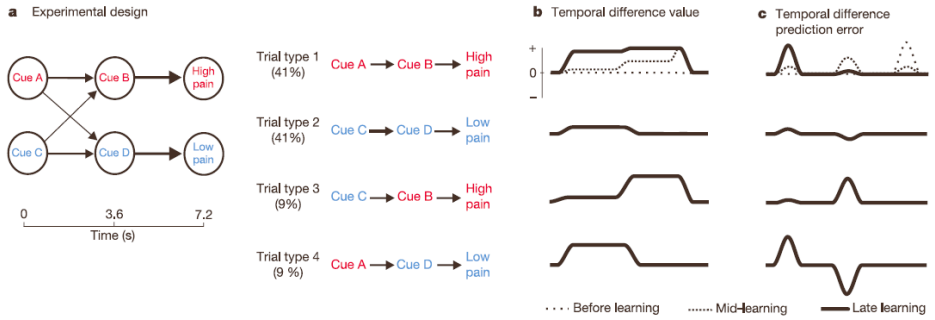
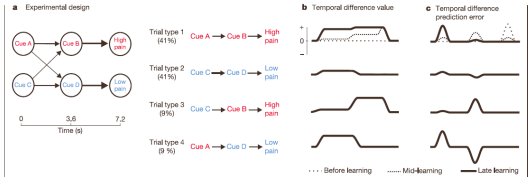


Figure 1 Experimental design and temporal difference model. **a**, The experimental design expressed as a Markov chain, giving four separate trial types. **b**, Temporal difference value. As learning proceeds, earlier cues learn to make accurate value predictions (that is, weighted averages of the final expected pain). **c**, Temporal difference prediction error; during learning the prediction error is transferred to earlier cues as they acquire the ability to make predictions. In trial types 3 and 4, the substantial change in prediction elicits a large positive or negative prediction error. (For clarity, before and mid-learning are shown only for trial type 1.)



[Seymore et al., Nature 2004]

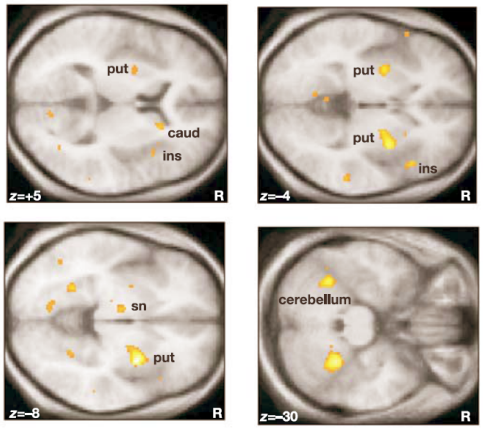


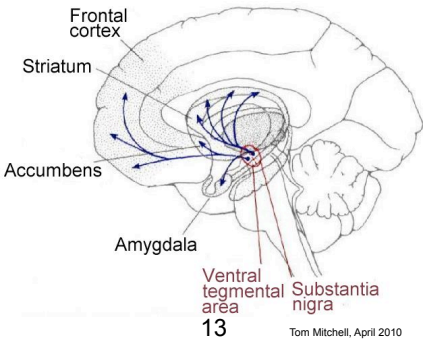
Figure 2 Temporal difference prediction error (statistical parametric maps). Areas coloured yellow/orange show significant correlation with the temporal difference

One Theory of RL in the Brain

from [Nieuwenhuis et al.]

- Basal ganglia monitor events, predict future rewards
- When prediction revised upward (downward), causes increase (decrease) in activity of midbrain dopaminergic neurons, influencing ACC

- This dopamine-based activation somehow results in revising the reward prediction function. Possibly through direct influence on Basal ganglia, and via prefrontal cortex



Summary: Temporal Difference ML Model Predicts Dopaminergic Neuron Acitivity during Learning

- Evidence now of neural reward signals from
  - Direct neural recordings in monkeys
  - fMRI in humans (1 mm spatial resolution)
  - EEG, MEG in humans (1-10 msec temporal resolution)
- Dopaminergic responses encode Bellman error
- Some differences, and efforts to refine the model
  - How/where is the value function encoded in the brain?
  - Study timing (e.g., basal ganglia learns faster than PFC ?)
  - Role of prior knowledge, rehearsal of experience, multi-task learning?

You Have Learned a Lot!

- Learning as function approximation
- Learning as optimization
- Classification, Regression
- Bayes optimal classifiers
- Discriminative vs. Generative
- Bias-variance decomposition
- Cross validation, overfitting
- VC dimension, PAC bounds
- Conditional independence
- Bayes nets
- Unsupervised, Semi-supervised
- EM
- Dimensionality reduction
- …

- Decision trees
- K nearest neighbor
- Naïve Bayes
- Logistic regression
- Linear regression
- Neural networks
- Mixture of Gaussians
- Hidden Markov Models
- SVM's
- Boosting
- PCA
- Spectral clustering
- Structure learning
- MDPs, Reinforcement learning
- …

Tom Mitchell, April 2010

BIG PICTURE

- Improving the performance at some task though experience!!! ☺
  - before you start any learning task, remember the fundamental questions:

What is the learning problem?

From what experience?

What model?

What loss function are you optimizing?

With what optimization algorithm?

Which learning algorithm?

With what guarantees?

How will you evaluate it?

55

What next?

- Machine Learning Department Seminar: http://calendar.cs.cmu.edu/ml/google_seminar
- Machine Learning Lunch talks: http://www.cs.cmu.edu/~learning/
- Intelligence Seminars: http://www.cs.cmu.edu/~iseminar/

- Journal:
  - JMLR – Journal of Machine Learning Research (free, on the web)

- Conferences:
  - ICML: International Conference on Machine Learning
  - NIPS: Neural Information Processing Systems
  - COLT: Computational Learning Theory
  - UAI: Uncertainty in AI
  - AIStats: intersection of Statistics and AI
  - Also AAAI, IJCAI and others

- Some MLD courses:
  - 10-708 Probabilistic Graphical Models (Fall)
  - 10-705 Intermediate Statistics (Fall)
  - 11-762 Language and Statistics II (Fall)
  - 10-702 Statistical Foundations of Machine Learning (Spring)
  - 10-725 Optimization (Spring)
  - ...

What Else Next?

- Poster session, Tuesday May 4, 3-6pm, NSH Atrium
- Final project report due: Wednesday May 5, midnight, email to 10701-instructors@cs.cmu.edu
- Final exam: Friday May 7, 5:30-8:30pm, DH 2302

thank you for your hard work!