Reinforcement Learning

Andrew W. Moore
Associate Professor
School of Computer Science
Carnegie Mellon University
www.cs.cmu.edu/~awm
awm@cs.cmu.edu
412-268-7599

Note to other teachers and users of these slides. Andrew would be delighted if you found this source material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or the following link to the source repository of Andrew's tutorials: http://www.cs.cmu.edu/~awm/tutorials.

Comments and corrections gratefully received.

Predicting Delayed Rewards in a Discounted Markov System

Prob(next state = S5 | this state = S4) = 0.8 etc…

What is expected sum of future rewards (discounted) ?

\[ \sum_{t=0}^{\infty} \gamma^t R(t+1) = \text{SSR} \]

Just Solve It!

We use standard Markov System Theory

Learning Delayed Rewards...

All you can see is a series of states and rewards:

\[ S_1(R=0) \rightarrow S_2(R=0) \rightarrow S_3(R=4) \rightarrow S_2(R=0) \rightarrow S_4(R=0) \rightarrow S_5(R=0) \]

Task: Based on this sequence, estimate \( J^*(S_1), J^*(S_2), \ldots, J^*(S_6) \)

Idea 1: Supervised Learning

Assume \( \gamma = 1/2 \)

At \( t=1 \) we were in state \( S_1 \) and eventually got a long term discounted reward of \( 0 + 0.5 + 0.25 + 0.125 \cdots = 1 \)

At \( t=2 \) in state \( S_2 \) ltdr = 2

At \( t=3 \) in state \( S_3 \) ltdr = 4

At \( t=4 \) in state \( S_2 \) ltdr = 0

Supervised Learning ALG

- For \( t=0,1, \ldots, T \), compute \( J[t] = \sum_{i=0}^{\infty} \gamma^t r[t+i] \)
- Compute \( J^*(S_t) = \) mean value of \( J[t] \) among all transitions beginning in state \( S_t \) on the trajectory

\[ \text{MATCHES}(S_t) = \{ r[S_{t+1}] = S_{t+2} \} \]

\[ J^*(S_t) = \frac{\sum_{r[S_{t+1}]=S_{t+2}} J[t]}{\text{MATCHES}(S_t)} \]

You’re done!

Supervised Learning ALG for the timid

If you have an anxious personality you may be worried about edge effects for some of the final transitions. With large trajectories these are negligible.
**Online Supervised Learning**

Initialize:  
\[ \text{Count}[S_i] = 0 \quad \forall S_i \]
\[ \text{Sum}[S_i] = 0 \quad \forall S_i \]

Eligibility\([S_i] = 0 \quad \forall S_i \]

Observe:  
When we experience \( S_i \) with reward \( r \) do this:
\[ \forall j \quad \text{Elig}[S_j] \leftarrow \gamma \text{Elig}[S_j] \]
\[ \text{Elig}[S_i] \leftarrow \text{Elig}[S_i] + 1 \]
\[ \forall j \quad \text{Sum}[S_j] \leftarrow \text{Sum}[S_j] + r \text{Elig}[S_j] \]
\[ \text{Count}[S_i] \leftarrow \text{Count}[S_i] + 1 \]

Then at any time,  
\[ J^*(S_i) = \frac{\text{Sum}[S_i]}{\text{Count}[S_i]} \]

---

**Online Supervised Learning Economics**

Given \( N \) states \( S_1 \cdots S_N \), OSL needs \( O(N) \) memory.

Each update needs \( O(N) \) work since we must update all \( \text{Elig}[\ ] \) array elements.

| Idea: Be sparse and only update/process \( \text{Elig}[\] elements with values > \( \xi \) for tiny \( \xi \) |
| There are only \( \log(\frac{1}{\xi})/\log(\frac{1}{\gamma}) \) such elements |

Easy to prove:  
\[ A \rightarrow T \rightarrow \infty, J^* (S_i) \rightarrow J^*(S_i) \quad \forall S_i \]

---

**Certainty-Equivalent (CE) Learning**

Idea: Use your data to estimate the underlying Markov system, instead of trying to estimate \( J \) directly.

Let’s grab OSL off the street, bundle it into a black van, take it to a bunker and interrogate it under 600 Watt lights.

<table>
<thead>
<tr>
<th>State</th>
<th>Observations of LTDR</th>
<th>( J(S_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_1 )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( S_2 )</td>
<td>2 0</td>
<td>4</td>
</tr>
<tr>
<td>( S_3 )</td>
<td>0 0</td>
<td>0</td>
</tr>
<tr>
<td>( S_4 )</td>
<td>0 0</td>
<td>0</td>
</tr>
</tbody>
</table>

There’s something a little suspicious about this (efficiency-wise)

---

**C.E. Method for Markov Systems**

Initialize:  
\[ \text{Count}[S_i] = 0 \quad \forall S_i \]
\[ \text{Sum}[R(S_i)] = 0 \quad \forall S_i \]
\[ \text{Trans}[S_i,S_j] = 0 \quad \forall S_i \text{Trans}[S_i,S_j] \]

When we are in state \( S_i \), and we receive reward \( r \), and we move to \( S_j \):
\[ \text{Count}[S_i] \leftarrow \text{Count}[S_i] + 1 \]
\[ \text{Sum}[R(S_i)] \leftarrow \text{Sum}[R(S_i)] + r \]
\[ \text{Trans}[S_i,S_j] \leftarrow \text{Trans}[S_i,S_j] + 1 \]

Then at any time,  
\[ r^m(S_i) = \frac{\text{Sum}[R(S_i)]}{\text{Count}[S_i]} \]
\[ P^m = \text{Estimated Prob(next = } S_j \mid \text{this = } S_i) \]
\[ = \frac{\text{Trans}[S_i,S_j]}{\text{Count}[S_i]} \]

---

**C.E. for Markov Systems (continued) ...**

So at any time we have  
\[ r^m(S_i) \text{ and } P^m \text{ (next=} S_j \mid \text{this=} S_i) \]
\[ \forall S_i S_j \]

So at any time we can solve the set of linear equations  
\[ J^m(S_i) = r^m(S_i) + \gamma \sum_{S_j} P^m(S_i,S_j) J^m(S_j) \]

In vector notation,  
\[ J^m = \gamma P^m J^m + \gamma P^m J^m \]
\[ \implies J^m = (\gamma P^m + \gamma P^m) J^m \]
where \( J^m, P^m \) are vectors of length \( N \)  
\( P^m \) is an \( N \times N \) matrix  
\( N = \# \text{ states} \)
C.E. Online Economics
Memory: \(O(N^2)\)
Time to update counters: \(O(1)\)
Time to re-evaluate \(J^{est}\)
- \(O(N^3)\) if use matrix inversion
- \(O(N^2k_{\text{CRIT}})\) if use value iteration and we need \(k_{\text{CRIT}}\) iterations to converge
- \(O(Nk_{\text{CRIT}})\) if use value iteration, and \(k_{\text{CRIT}}\) to converge, and M.S. is Sparse (i.e. mean # successors is constant)

Certainty Equivalent Learning
Memory use could be \(O(N^2)\)!
And time per update could be \(O(Nk_{\text{CRIT}})\) up to \(O(N^3)\)!
Too expensive for some people.
Prioritized sweeping will help, (see later), but first let's review a very inexpensive approach

Why this obsession with onlineiness?
I really care about supplying up-to-date \(J^{est}\) estimates all the time.
Can you guess why?
If not, all will be revealed in good time…

“One Backup C.E.” Economics
Space : \(O(N^2)\)
Time to update statistics : \(O(1)\)
Time to update \(J^{est}\) : \(O(1)\)
- Good News: Much cheaper per transition
- Good News: Contraction Mapping proof (modified) promises convergence to optimal
- Bad News: Wastes data

Less Time: More Data
Limited Backups
- Do previous C.E. algorithm.
- At each time step observe \(S_i(r)\rightarrow S_j\) and update \(\text{Count}[S_i], \text{SumR}[S_i], \text{Trans}[S_i,S_j]\)
- And thus also update estimates \(r_i^{\infty}, P_i^{\infty} \ \forall f \in \text{outcomes}(S_i)\)
  But instead of re-solving for \(J^{est}\), do much less work.
  Just do one “backup” of \(J^{est}[S_i]\)
  \[J^{est}[S_i] \leftarrow r_i^{\infty} + \gamma \sum_f P_i^{\infty} J^{est}[S_f] \]

Prioritized Sweeping
[Moore + Atkeson, ’93]
Tries to be almost as data-efficient as full CE but not much more expensive than “One Backup” CE.
On every transition, some number (\(\beta\)) of states may have a backup applied. Which ones?
- The most “deserving”
- We keep a priority queue of which states have the biggest potential for changing their \(J^{est}(S_i)\) value
Where Are We?
Trying to do online Jest prediction from streams of transitions

<table>
<thead>
<tr>
<th>Space</th>
<th>Jest Update Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised Learning</td>
<td>$O(N_s)$</td>
</tr>
<tr>
<td>Full C.E. Learning</td>
<td>$O(N_{so})$</td>
</tr>
<tr>
<td>One Backup C.E. Learning</td>
<td>$O(N_{so})$</td>
</tr>
<tr>
<td>Prioritized Sweeping</td>
<td>$O(N_{so})$</td>
</tr>
</tbody>
</table>

What Next?
Sample Backups!!!

Temporal Difference Learning
Only maintain a Jest array… nothing else

So you’ve got $Jest(S_1), Jest(S_2), \ldots, Jest(S_N)$
and you observe $S_i \rightarrow S_j$ what should you do?

Can You Guess?

TD Learning
$S_i \rightarrow S_j$
We update $Jest(S_i)$
We nudge it to be closer to expected future rewards

$$Jest(S_i) \leftarrow (1-\alpha)Jest(S_i) + \alpha[\text{Expected future rewards}]$$

$$= (1-\alpha)Jest(S_i) + \alpha [r + \sum_j \text{Jest}(S_j)]$$

$\alpha$ is called a “learning rate” parameter. (See “$\eta$” in the neural lecture)

Simplified TD Analysis
Let’s estimate it with TD

Let’s run TD-Learning, where

$$J_t = \text{Estimate } Jest(S_0) \text{ before the } t\text{'th trial.}$$

From definition of TD-Learning:

$$J_{t+1} = (1-\alpha)J_t + \alpha r_t$$

Useful quantity: Define

$$\sigma^2 = \text{Variance of reward} = E[(r_t - J^*)^2]$$

$$= \sum_{k=1}^{M} p(k) (r^{(k)} - J^*)^2$$
Remember $J^* = \mathbb{E}[r_t]$, $\sigma^2 = \mathbb{E}[(r_t - J^*)^2]$

$$J_{t+1} = \alpha r_t + (1-\alpha)J_t$$

Thus...

$$\lim_{t \to \infty} \mathbb{E}[J_t] = J^*$$

Is this impressive??

And it is thus easy to show that ....

$$\lim_{t \to \infty} S_t = \lim_{t \to \infty} \mathbb{E}[J_t - J^*]^2 = \frac{\alpha \sigma^2}{2 - \alpha}$$

• What do you think of TD learning?
• How would you improve it?

Decaying Learning Rate

| This Works: | $\alpha_t = 1/t$ |
| This Doesn't: | $\alpha_t = \alpha_0$ |
| This Works: | $\alpha_t = \beta/(\beta+t)$ [e.g. $\beta=1000$] |
| This Doesn't: | $\alpha_t = \beta \alpha_{t-1}$ ($\beta<1$) |

IN OUR EXAMPLE, .... USE $\alpha_t = 1/t$

Remember $J' = \mathbb{E}[r_t]$, $\sigma^2 = \mathbb{E}[(r_t - J')^2]$

$$J_{t+1} = \alpha r_t + (1-\alpha)J_t$$

Write $C_t = (1-\alpha)J_t$ and you’ll see that

$$C_{t+1} = r_t + C_t \quad \text{so} \quad J_{t+1} = \left(\frac{1}{1-\alpha}\right)r_t + J_t$$

And...

Decaying Learning Rate con't...

$$\mathbb{E}[J_t - J^*]^2 = \sigma^2 + \left(J_0 - J^*\right)^2$$

so, ultimately $\lim_{t \to \infty} \mathbb{E}[J_t - J^*]^2 = 0$
A Fancier TD...

Write \( S[t] = \) state at time \( t \)

Suppose \( \alpha = 1/4 \) \( \gamma = 1/2 \)

Assume \( J^\text{est}(S_{23}) = 0 \) \( J^\text{est}(S_{17}) = 0 \) \( J^\text{est}(S_{44}) = 16 \)

Assume \( t = 405 \) and \( S[t] = S_{23} \)

Observe \( S_{23} \rightarrow S_{17} \) with reward 0

Now \( t = 406 \), \( S[t] = S_{17} \), \( S[t-1] = S_{23} \)

Observe \( S_{17} \rightarrow S_{44} \)

Now \( t = 407 \), \( S[t] = S_{44} \)

\( J^\text{est}(S_{23}) = \), \( J^\text{est}(S_{17}) = \), \( J^\text{est}(S_{44}) = \)

INSIGHT: \( J^\text{est}(S_{23}) \) might think

\( \text{I gotta get me some of that !!!} \)

(r=0)

TD(\( \lambda \)) Comments

TD(\( \lambda = 0 \)) is the original TD

TD(\( \lambda = 1 \)) is almost the same as supervised learning (except it uses a learning rate instead of explicit counts)

TD(\( \lambda = 0.7 \)) is often empirically the best performer

- Dayan’s proof holds for all \( 0 \leq \lambda \leq 1 \)
- Updates can be made more computationally efficient with “eligibility” traces (similar to O.S.L.)
- Question:
  - Can you invent a problem that would make TD(0) look bad and TD(1) look good?
  - How about TD(0) look good & TD(1) bad??

Learning M.S. Summary

<table>
<thead>
<tr>
<th>Model-Based</th>
<th>Space</th>
<th>J Update</th>
<th>Cost</th>
<th>Data Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised Learning</td>
<td>0(N_s)</td>
<td>0</td>
<td><img src="https://via.placeholder.com/150" alt="image" /></td>
<td></td>
</tr>
<tr>
<td>Full C.E. Learning</td>
<td>0(N_s)</td>
<td>0(N_s,N_a)</td>
<td><img src="https://via.placeholder.com/150" alt="image" /></td>
<td></td>
</tr>
<tr>
<td>One Backup C.E. Learning</td>
<td>0(N_s)</td>
<td>0(1)</td>
<td><img src="https://via.placeholder.com/150" alt="image" /></td>
<td></td>
</tr>
<tr>
<td>Prioritized Sweeping</td>
<td>0(N_s)</td>
<td>0(1)</td>
<td><img src="https://via.placeholder.com/150" alt="image" /></td>
<td></td>
</tr>
<tr>
<td>TD(0)</td>
<td>0(N_s)</td>
<td>0(1)</td>
<td><img src="https://via.placeholder.com/150" alt="image" /></td>
<td></td>
</tr>
<tr>
<td>TD(( \lambda ), ( 0&lt;\lambda \leq 1 ))</td>
<td>0(N_s)</td>
<td>0</td>
<td><img src="https://via.placeholder.com/150" alt="image" /></td>
<td></td>
</tr>
</tbody>
</table>

Learning Policies for MDPs

See previous lecture slides for definition of and computation with MDPs.

The task:

World: You are in state 34.
Your immediate reward is 3. You have 3 actions.

Robot: I’ll take action 2.

World: You are in state 77.
Your immediate reward is -7. You have 2 actions.

Robot: I’ll take action 1.

World: You’re in state 34 (again).
Your immediate reward is 3. You have 3 actions.
The Markov property means once you’ve selected an action the P.D.F. of your next state is the same as the last time you tried the action in this state.

The “Credit Assignment” Problem

I’m in state 43, reward = 0, action = 2
- * * * 39, * = 0, * = 4
- * * * 22, * = 0, * = 1
- * * * 21, * = 0, * = 1
- * * * 13, * = 0, * = 2
- * * * 54, * = 0, * = 2
- * * * 26, * = 100

Yippee! I got to a state with a big reward! But which of my actions along the way actually helped me get there?? This is the Credit Assignment problem.

It makes Supervised Learning approaches [e.g. Boxes (Michie & Chambers)] very, very slow.

Using the MDP assumption helps avoid this problem.
MDP Policy Learning

<table>
<thead>
<tr>
<th>Method</th>
<th>Space</th>
<th>Update Cost</th>
<th>Data Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full C.E. Learning</td>
<td>0(NsAo)</td>
<td>0(NsAo)</td>
<td>✔</td>
</tr>
<tr>
<td>One Backup C.E.</td>
<td>0(NsAo)</td>
<td>0(NsAo)</td>
<td>❌</td>
</tr>
<tr>
<td>Prioritized Sweeping</td>
<td>0(NsAo)</td>
<td>0(NsAo)</td>
<td>❌</td>
</tr>
</tbody>
</table>

- We’ll think about Model-Free in a moment...
- The C.E. methods are very similar to the MS case, except now do value-iteration-for-MDP backups

\[ J^\text{opt}(S_i) = \max_a \left[ r_i^\text{est} + \gamma \sum_{S_j, a' | S_j \in \text{SUCC}(S_i)} p_{i,S_j} a' J^\text{opt}(S_j) \right] \]

Choosing Actions

We’re in state \( S_i \)

We can estimate:

- \( r_i^\text{est} \)
- \( P_{i,S_j}(\text{next} = S_j \mid \text{this} = S_i, \text{action} a) \)
- \( J^\text{opt}(\text{next} = S_j) \)

So what action should we choose?

**IDEA 1:**

\[ a = \arg \max_{a'} \left[ r_i + \gamma \sum_j p_{i,j,a'} J^\text{opt}(S_j) \right] \]

**IDEA 2:**  
- Any problems with these ideas?  
- Any other suggestions?  
- Could we be optimal?

Model-Free R.L.

Why not use T.D.?

Observe

\[ S_i \xrightarrow{a} S_j \]

update

\[ J^\text{opt}(S_i) \leftarrow \alpha \left( r_i + \gamma J^\text{opt}(S_j) \right) + (1 - \alpha) J^\text{opt}(S_i) \]

What’s wrong with this?

Q-Learning: Model-Free R.L.

[Watkins, 1988]

Define

\( Q^*(S_i,a) = \text{Expected sum of discounted future rewards if I start in state } S_i, \text{ if I then take action } a, \text{ and if I’m subsequently optimal} \)

Questions:

Define \( Q^*(S_i,a) \) in terms of \( J^* \)

Define \( J^*(S_i) \) in terms of \( Q^* \)

Q-Learning Update

Note that

\[ Q^*(S_i,a) = r_i + \gamma \sum_{S_j \in \text{SUCC}(S_i)} p_{S_i} a S_j \max_{a'} Q^*(S_j,a') \]

In Q-learning we maintain a table of \( Q^\text{est} \) values instead of \( J^\text{est} \) values…

When you see \( S_i \xrightarrow{a} S_j \) do...

\[ Q^\text{est}(S_i,a) \leftarrow \alpha \left[ r_i + \gamma \max_{a'} Q^\text{est}(S_j,a') \right] + (1 - \alpha) Q^\text{est}(S_i,a) \]

This is even cleverer than it looks: the \( Q^\text{est} \) values are not biased by any particular exploration policy. It avoids the Credit Assignment problem.

\[ V^*(s) = \max_{a'} Q(s,a') \]

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

\[ \alpha_n = \frac{1}{1 + \text{visits}_n(s,a)} \]
Q-Learning: Choosing Actions

Same issues as for CE choosing actions
- Exploration vs. Exploitation
- Don’t always be greedy, so don’t always choose: \( \arg \max_a Q(s, a) \)
- Don’t always be random (otherwise it will take a long time to reach somewhere exciting)
- Boltzmann exploration [Watkins]
  \[
  \text{Prob(choose action } a \text{)} = \exp \left( \frac{Q^\text{opt} (s, a)}{K} \right)
  \]
- Optimism in the face of uncertainty [Sutton ’90, Kaelbling ’90]
  - Initialize Q-values optimistically high to encourage exploration
  - Or take into account how often each s,a pair has been tried

Q-Learning Comments
- [Watkins] proved that Q-learning will eventually converge to an optimal policy.
- Empirically it is cute
- Empirically it is very slow
- Why not do Q(\( \lambda \))?
  - Would not make much sense [reintroduce the credit assignment problem]
  - Some people (e.g. Peng & Williams) have tried to work their way around this.

If we had time...
- Value function approximation
  - Use a Neural Net to represent \( \text{J}^\text{opt} \) [e.g. Tesauro]
  - Use a Neural Net to represent \( \text{Q}^\text{opt} \) [e.g. Crites]
  - Use a decision tree
    - with Q-learning [Chapman + Kaelbling ’91]
    - with C.E. learning [Moore ’91]
  - How to split up space?
    - Significance test on Q values [Chapman + Kaelbling]
    - Execution accuracy monitoring [Moore ’91]
    - Game Theory [Moore + Atkeson ’95]
    - New influence/variance criteria [Munos ’99]

If we had time...
- R.L. Theory
  - Counterexamples [Boyan + Moore], [Baird]
  - Value Function Approximators with Averaging will converge to something [Gordon]
  - Neural Nets can fail [Baird]
  - Neural Nets with Residual Gradient updates will converge to something
  - Linear approximators for TD learning will converge to something useful [Tsitsiklis + Van Roy]

TD-Gammon

Current Issues
- Making function approximation work
- Abstraction (options, macros, …), learning structure
- Partially Observable Markov Processes, POMDPs
What You Should Know

- Supervised learning for predicting delayed rewards
- Certainty equivalent learning for predicting delayed rewards
- Model free learning (TD) for predicting delayed rewards
- Reinforcement Learning with MDPs: What’s the task?
- Why is it hard to choose actions?
- Q-learning (including being able to work through small simulated examples of RL)