Reinforcement Learning

Chao Lan
Revisit: Markov Decision Process Problem

Given reward and transition probability, identify optimal policy and utility.

\[ P(s' \mid s, a) \]
Reinforcement Learning Problem

*Unknowing* reward and transition probability, identify optimal policy and utility.

Assume the unknown environment is fixed, and there is an agent that can explore it (with given actions).
Three Questions

Q1. Given an arbitrary policy, how to identify utility?
   - passive RL: three methods

Q2. How to identify optimal policy?
   - active RL: Q-Learning

Q3. How to reduce search space?
   - utility approximation
Q1. Given a policy, how to identify utility?
[Q] Can we apply Monte Carlo?

\[ U(1,1) \approx \frac{0.72 + 0.72 - 1.16}{3} = 0.093 \]
Method 1: Monte Carlo

\[ U(1,1) \approx \frac{(0.72+0.72-1.16)}{3} = 0.093 \]
[Q] Can we apply policy iteration (step 2)?

1. Randomly initialize policy.

\[ U_i(s) = R(s) + \gamma \sum_{s'} P(s' | s, \pi_i(s)) U_i(s') \]

2. Fix policy, optimize utility by solving n linear equations.

3. Fix utility, optimize policy using the Bellman equation.

4. Repeat 2-4 until convergence.

Step 2 is also called “policy evaluation”. 
[Q] Can we apply policy iteration (step 2)?

1. Randomly initialize policy.

2. Fix policy, optimize utility using $n$ linear equations:

$$U_i(s) = R(s) + \gamma \sum P(s' | s, \pi_i(s)) U_i(s')$$

   two quantities are unknown

3. Fix utility, optimize policy using the Bellman equation.

4. Repeat 2-4 until convergence.
Discussion

Can we estimate \( R(s) \) and \( P(s'|s,a) \) from sample trials?

\[
\begin{align*}
P(s'=(1,2) \mid s=(1,1), a=\text{UP}) &= \ \\
P(s'=(2,1) \mid s=(1,1), a=\text{UP}) &= \ \\
P(s'=(1,1) \mid s=(1,1), a=\text{UP}) &=
\end{align*}
\]
Discussion

Can we estimate \( R(s) \) and \( P(s'|s,a) \) from sample trials?

\[
P(s'=(1,2) \mid s=(1,1), a=UP) = \frac{2}{3}
\]

\[
P(s'=(2,1) \mid s=(1,1), a=UP) = \frac{1}{3}
\]

\[
P(s'=(1,1) \mid s=(1,1), a=UP) = 0
\]

\[\text{R(s) will be collected after s is visited in a trial and no need to be updated afterwards.}\]
P((1,2) | (1,1), UP) = (2+1) / (3+1)

P((2,1) | (1,1), UP) = (1+0) / (3+1)

P((1,1) | (1,1), UP) = (0+0) / (3+1)

And whenever we collect a new trial, just update...
Method 2: Adaptive Dynamic Programming

1. Put agent at \( s = s_0 \). Set \( N(z|s,a) = N(s,a) = 0 \) for all states \( s, z \). Set utility and \( R \) empty.

2. Agent takes action \( policy(s) \) and moves to \( s' \).

3. If \( s' \) is newly visited, set \( R(s') = r \) and \( U(s') = r \), where \( r \) is observed reward at \( s' \).

4. Else, update transition probability, by incrementing \( N(s'|s,a) \) and \( N(s,a) \) by 1 and 
   \[
   Pr(z|s,a) = \frac{N(z|s,a)}{N(s,a)}, \text{ for all neighboring state } z \text{ including } s'
   \]

5. Update utility, by solving \( n \) linear equations based on update-to-date \( Pr \) and \( R \).

6. Update \( s = s' \) and repeat 2-6 until convergence. (If agent terminates, put it back to \( s_0 \).)

ADP alternates between policy-evaluation and transition probability estimate.
Example: estimate utility of the following policy
Initialization

1. place agent at initial state (1,1).
2. set all counters to zero.
3. newly visit (1,1), set $R(1,1)=r$ and $U(1,1)=r$. 
Iteration 1

1. take action UP at (1,1), end up moving right.
2. newly visit (2,1), set R(2,1)=r and U(2,1)=r
3. increment corresponding counters by 1
   \[ N[(1,1),\text{up}] = 1, \quad N[(2,1) \mid (1,1), \text{UP}] = 1 \]
4. update transition probability
   \[
   \begin{align*}
   & Pr[(2,1) \mid (1,1), \text{UP}] = 1/1 = 1 \\
   & Pr[(1,2) \mid (1,1), \text{UP}] = 0/1 = 0 \\
   & Pr[(1,1) \mid (1,1), \text{UP}] = 0/1 = 0
   \end{align*}
   \]
Iteration 2

1. take action LEFT at (2,1), end up moving left.
2. (1,1) is not newly visit, do nothing
3. increment corresponding counters by 1
   \[ N[(2,1), \text{left}] = 1, \quad N[(1,1) | (2,1), \text{left}] = 1 \]
4. update transition probability
   \[ \Pr[(1,1) | (2,1), \text{left}] = 1/1 = 1 \]
   \[ \Pr[(2,1) | (2,1), \text{left}] = 0/1 = 1 \]
Iteration 3

1. take action UP at (1,1), end up moving up.
2. newly visit (1,2), set R(1,2)=r and U(1,2)=r
3. increment corresponding counters by 1
   \[ N[(1,1),up] = 2, \quad N[(1,2) \mid (1,1), up] = 1 \]
4. update transition probability
   \[ Pr[(2,1) \mid (1,1), UP] = 1/2 = 0.5 \]
   \[ Pr[(1,2) \mid (1,1), UP] = 1/2 = 0.5 \]
   \[ Pr[(1,1) \mid (1,1), UP] = 0/2 = 0 \]
Explore until 1st trial terminates.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N[(1,1),up] = 2</td>
<td>N[(2,1)</td>
<td>(1,1),up] = 1</td>
</tr>
<tr>
<td>N[(2,1),left] = 1</td>
<td>N[(1,1)</td>
<td>(2,1),left] = 1</td>
</tr>
<tr>
<td>N[(1,2),up] = 1</td>
<td>N[(1,3)</td>
<td>(1,2),up] = 1</td>
</tr>
<tr>
<td>N[(1,3),right] = 1</td>
<td>N[(2,3)</td>
<td>(1,3),right] = 1</td>
</tr>
<tr>
<td>N[(2,3),right] = 1</td>
<td>N[(3,3)</td>
<td>(2,3),right] = 1</td>
</tr>
<tr>
<td>N[(3,3),right] = 2</td>
<td>N[(3,2)</td>
<td>(3,3),right] = 1</td>
</tr>
<tr>
<td>N[(3,2),up] = 1</td>
<td>N[(3,3)</td>
<td>(3,2),up] = 1</td>
</tr>
</tbody>
</table>

Diagram: 9x9 grid with start and end points marked.
Transition Probability \( P[z|s,a] = \frac{N[z|s,a]}{N[s,a]} \).

<table>
<thead>
<tr>
<th>Transition</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N[(1,1),\text{up}] )</td>
<td>2</td>
</tr>
<tr>
<td>( N[(2,1),\text{left}] )</td>
<td>1</td>
</tr>
<tr>
<td>( N[(1,2),\text{up}] )</td>
<td>1</td>
</tr>
<tr>
<td>( N[(1,3),\text{right}] )</td>
<td>1</td>
</tr>
<tr>
<td>( N[(2,3),\text{right}] )</td>
<td>1</td>
</tr>
<tr>
<td>( N[(3,3),\text{right}] )</td>
<td>2</td>
</tr>
<tr>
<td>( N[(3,2),\text{up}] )</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Transition</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P[(2,1)</td>
<td>(1,1),\text{up}] )</td>
</tr>
<tr>
<td>( P[(1,1)</td>
<td>(2,1),\text{left}] )</td>
</tr>
<tr>
<td>( P[(1,3)</td>
<td>(1,2),\text{up}] )</td>
</tr>
<tr>
<td>( P[(2,3)</td>
<td>(1,3),\text{right}] )</td>
</tr>
<tr>
<td>( P[(3,3)</td>
<td>(2,3),\text{right}] )</td>
</tr>
<tr>
<td>( P[(3,2)</td>
<td>(3,3),\text{right}] )</td>
</tr>
<tr>
<td>( P[+1</td>
<td>(3,3),\text{right}] )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Transition</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N[(2,1)</td>
<td>(1,1),\text{up}] )</td>
</tr>
<tr>
<td>( N[(1,1)</td>
<td>(2,1),\text{left}] )</td>
</tr>
<tr>
<td>( N[(1,3)</td>
<td>(1,2),\text{up}] )</td>
</tr>
<tr>
<td>( N[(2,3)</td>
<td>(1,3),\text{right}] )</td>
</tr>
<tr>
<td>( N[(3,3)</td>
<td>(2,3),\text{right}] )</td>
</tr>
<tr>
<td>( N[(3,2)</td>
<td>(3,3),\text{right}] )</td>
</tr>
<tr>
<td>( N[+1</td>
<td>(3,3),\text{right}] )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Transition</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P[(2,1)</td>
<td>(1,1),\text{up}] )</td>
</tr>
<tr>
<td>( P[(1,1)</td>
<td>(2,1),\text{left}] )</td>
</tr>
<tr>
<td>( P[(1,3)</td>
<td>(1,2),\text{up}] )</td>
</tr>
<tr>
<td>( P[(2,3)</td>
<td>(1,3),\text{right}] )</td>
</tr>
<tr>
<td>( P[(3,3)</td>
<td>(2,3),\text{right}] )</td>
</tr>
<tr>
<td>( P[(3,2)</td>
<td>(3,3),\text{right}] )</td>
</tr>
<tr>
<td>( P[+1</td>
<td>(3,3),\text{right}] )</td>
</tr>
<tr>
<td>N[(1,1),up] = 2</td>
<td>N[(2,1)</td>
</tr>
<tr>
<td>----------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>N[(2,1),left] = 1</td>
<td>N[(1,1)</td>
</tr>
<tr>
<td>N[(1,2),up] = 1</td>
<td>N[(1,3)</td>
</tr>
<tr>
<td>N[(1,3),right] = 1</td>
<td>N[(2,3)</td>
</tr>
<tr>
<td>N[(2,3),right] = 1</td>
<td>N[(3,3)</td>
</tr>
<tr>
<td>N[(3,3),right] = 2</td>
<td>N[(3,2)</td>
</tr>
<tr>
<td>N[(3,2),up] = 1</td>
<td>N[(3,3)</td>
</tr>
</tbody>
</table>

![Diagram](image-url)
2nd Trial of Exploration: Iteration 1

N[(1,1),up] = 2+1
N[(2,1),left] = 1
N[(1,2),up] = 1
N[(1,3),right] = 1
N[(2,3),right] = 1
N[(3,3),right] = 2
N[(3,2),up] = 1

N[(2,1)|(1,1),up] = 1
N[(1,1)|(2,1),left] = 1
N[(1,3)|(1,2),up] = 1
N[(2,3)|(1,3),right] = 1
N[(3,3)|(2,3),right] = 1
N[(3,2)|(3,3),right] = 1
N[+1|(3,3),right] = 1
N[(3,3)|(3,2),up] = 1

N[(2,1)|(1,1),up] = 1+1
N[(1,2)|(1,1),up] = 1
N[(2,1),left] = 1
N[(1,2),up] = 1
N[(1,2)|(1,3),right] = 1
N[(2,3)|(2,3),right] = 1
N[+1|(3,3),right] = 1
N[(3,3)|(3,2),up] = 1
2nd Trial of Exploration: Iteration 2

\[
\begin{align*}
N[(1,1),\text{up}] &= 2+1 \\
N[(2,1),\text{left}] &= 1 \\
N[(1,2),\text{up}] &= 1+1 \\
N[(1,3),\text{right}] &= 1 \\
N[(2,3),\text{right}] &= 1 \\
N[(3,3),\text{right}] &= 2 \\
N[(3,2),\text{up}] &= 1 \\
N[(2,1)|(1,1),\text{up}] &= 1 \\
N[(1,1)|(2,1),\text{left}] &= 1 \\
N[(1,3)|(1,2),\text{up}] &= 1+1 \\
N[(2,3)|(1,3),\text{right}] &= 1 \\
N[(3,3)|(2,3),\text{right}] &= 1 \\
N[(3,2)|(3,3),\text{right}] &= 1 \\
N[(3,3)|(3,2),\text{up}] &= 1 \\
N[(2,1)|(1,1),\text{up}] &= 1 \\
N[(1,1)|(2,1),\text{left}] &= 1 \\
N[(1,3)|(1,2),\text{up}] &= 1+1 \\
N[(2,3)|(2,3),\text{right}] &= 1 \\
N[(3,3)|(3,2),\text{up}] &= 1 \\
N[(3,3)|(3,2),\text{up}] &= 1 \\
N[(1,2)|(1,1),\text{up}] &= 1+1 \\
N[(1,3)|(1,2),\text{up}] &= 1+1 \\
N[(2,3)|(2,3),\text{right}] &= 1 \\
N[(3,3)|(3,2),\text{up}] &= 1 \\
N[+1|(3,3),\text{right}] &= 1
\end{align*}
\]
### 2nd Trial of Exploration: Iteration 3

<table>
<thead>
<tr>
<th>Position</th>
<th>Value</th>
<th>Position</th>
<th>Value</th>
<th>Position</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(1,1), \text{up}$</td>
<td>$2+1$</td>
<td>$(2,1)</td>
<td>(1,1), \text{up}$</td>
<td>$1$</td>
<td>$(1,2)</td>
</tr>
<tr>
<td>$(2,1), \text{left}$</td>
<td>$1$</td>
<td>$(1,1)</td>
<td>(2,1), \text{left}$</td>
<td>$1$</td>
<td>$(1,3)</td>
</tr>
<tr>
<td>$(1,2), \text{up}$</td>
<td>$1+1$</td>
<td>$(1,3)</td>
<td>(1,3), \text{right}$</td>
<td>$1$</td>
<td>$(2,3)</td>
</tr>
<tr>
<td>$(1,3), \text{right}$</td>
<td>$1 + 1$</td>
<td>$(2,3)</td>
<td>(2,3), \text{right}$</td>
<td>$1$</td>
<td>$(3,3)</td>
</tr>
<tr>
<td>$(2,3), \text{right}$</td>
<td>$1$</td>
<td>$(3,3)</td>
<td>(3,3), \text{right}$</td>
<td>$1$</td>
<td>$(3,2)</td>
</tr>
<tr>
<td>$(3,3), \text{right}$</td>
<td>$2$</td>
<td>$(3,3)</td>
<td>(3,2), \text{up}$</td>
<td>$1$</td>
<td>$(+1)</td>
</tr>
</tbody>
</table>
### 2nd Trial of Exploration: Iteration 3

<table>
<thead>
<tr>
<th>State</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N[(1,1),\text{up}] = 2+1 )</td>
<td>N[(2,1)</td>
</tr>
<tr>
<td>( N[(2,1),\text{left}] = 1 )</td>
<td>N[(1,1)</td>
</tr>
<tr>
<td>( N[(1,2),\text{up}] = 1+1+1 )</td>
<td>N[(1,3)</td>
</tr>
<tr>
<td>( N[(1,3),\text{right}] = 1 + 1 )</td>
<td>N[(2,3)</td>
</tr>
<tr>
<td>( N[(2,3),\text{right}] = 1 )</td>
<td>N[(3,3)</td>
</tr>
<tr>
<td>( N[(3,3),\text{right}] = 2 )</td>
<td>N[(3,2)</td>
</tr>
<tr>
<td>( N[(3,2),\text{up}] = 1 )</td>
<td>N[(3,3)</td>
</tr>
<tr>
<td>( N[(3,3),\text{up}] = 1 )</td>
<td></td>
</tr>
</tbody>
</table>
2nd Trial of Exploration: Iteration 3

\[
\begin{align*}
N[(1,1), up] &= 2 + 1 \\
N[(2,1), left] &= 1 \\
N[(1,2), up] &= 1 + 1 + 1 \\
N[(1,3), right] &= 1 + 1 + 1 \\
N[(2,3), right] &= 1 \\
N[(3,3), right] &= 2 \\
N[(3,2), up] &= 1
\end{align*}
\]

\[
\begin{align*}
N[(2,1)|(1,1), up] &= 1 \\
N[(1,1)|(2,1), left] &= 1 \\
N[(1,3)|(1,2), up] &= 1 + 1 + 1 \\
N[(2,3)|(1,3), right] &= 1 \\
N[(3,3)|(2,3), right] &= 1 \\
N[(3,2)|(3,3), right] &= 1 \\
N[(3,3)|(3,2), up] &= 1
\end{align*}
\]
2nd Trial of Exploration till Termination.

<table>
<thead>
<tr>
<th>Location</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>N[(1,1),up]</td>
<td>2+1</td>
</tr>
<tr>
<td>N[(2,1),left]</td>
<td>1</td>
</tr>
<tr>
<td>N[(1,2),up]</td>
<td>1+1+1</td>
</tr>
<tr>
<td>N[(1,3),right]</td>
<td>1+1+1</td>
</tr>
<tr>
<td>N[(2,3),right]</td>
<td>1+1+1</td>
</tr>
<tr>
<td>N[(3,3),right]</td>
<td>2+1</td>
</tr>
<tr>
<td>N[(3,2),up]</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Path Condition</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>N[(2,1)</td>
<td>(1,1),up]</td>
</tr>
<tr>
<td>N[(1,1)</td>
<td>(2,1),left]</td>
</tr>
<tr>
<td>N[(1,2)</td>
<td>(1,2),up]</td>
</tr>
<tr>
<td>N[(2,3)</td>
<td>(1,3),right]</td>
</tr>
<tr>
<td>N[(3,3)</td>
<td>(2,3),right]</td>
</tr>
<tr>
<td>N[(3,2)</td>
<td>(3,3),right]</td>
</tr>
<tr>
<td>N[(3,3)</td>
<td>(3,2),up]</td>
</tr>
</tbody>
</table>

Diagram of the exploration path with start and end points marked.
<table>
<thead>
<tr>
<th>Transition</th>
<th>Transition</th>
<th>Transition</th>
<th>True Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>N[(1,1),up] = 2+1</td>
<td>N[(2,1)</td>
<td>(1,1),left] = 1</td>
<td>N[(1,1]</td>
</tr>
<tr>
<td>N[(2,1),left] = 1</td>
<td>N[(1,1)</td>
<td>(2,1),up] = 1</td>
<td>N[(1,1)</td>
</tr>
<tr>
<td>N[(1,2),up] = 1+1+1</td>
<td>N[(1,3)</td>
<td>(1,2),up] = 1+1+1</td>
<td>N[(1,2)</td>
</tr>
<tr>
<td>N[(1,3),right]=1+1+1</td>
<td>N[(2,3)</td>
<td>(1,3),right]=1+1</td>
<td>N[(1,2)</td>
</tr>
<tr>
<td>N[(2,3),right]=1+1+1</td>
<td>N[(3,3)</td>
<td>(2,3),right] = 1+1</td>
<td>N[(2,3)</td>
</tr>
<tr>
<td>N[(3,3),right] = 2+1</td>
<td>N[(3,2)</td>
<td>(3,3),right] = 1</td>
<td>N[+1</td>
</tr>
<tr>
<td>N[(3,2),up] = 1</td>
<td>N[(3,3)</td>
<td>(3,2),up] = 1</td>
<td>N[(3,3)</td>
</tr>
</tbody>
</table>
Performance of ADP
Discussion: Can we speed up ADP?

1. Initialization.

2. Take action $a$, and move from $s$ to $s'$.

3. If $s'$ is new, $R(s')=r$ and $U(s')=r$.

4. else, increment $N(s'|s,a)$ and $N(s,a)$ by 1, and $Pr(z|s,a) = N(z|s,a) / N(s,a)$.

5. update utility, by solving $n$ linear equations (from Bellman eq. with fixed policy).

6. update $s=s'$ and repeat 2-6 until convergence.
Revisit Bellman Equation with Fixed Policy

\[ U_i(s) = R(s) + \gamma \sum_{s'} P(s' \mid s, \pi_i(s)) U_i(s') \quad \Rightarrow \quad U(1,1) = R(1,1) + \gamma \Delta \]

\[ \Delta = P\{(1,1)\mid(1,1),\text{up}\} \times U(1,1) + P\{(1,2)\mid(1,1),\text{up}\} \times U(1,2) + P\{(2,1)\mid(1,1),\text{up}\} \times U(2,1) \]
How to estimate $\Delta$ if we observe $(1,1)\rightarrow(1,2)$ for the first time in the above trial?

$$\Delta = P\{(1,1)|(1,1), up\} \cdot U(1,1) + P\{(1,2)|(1,1), up\} \cdot U(1,2) + P\{(2,1)|(1,1), up\} \cdot U(2,1)$$
How to estimate $\Delta$ if we observe $(1,1) \rightarrow (1,2)$ for the first time in the above trial?

$$\Delta = P\{(1,1)|(1,1),up\} \cdot U(1,1) + P\{(1,2)|(1,1),up\} \cdot U(1,2) + P\{(2,1)|(1,1),up\} \cdot U(2,1)$$

$$\Delta = 0 \cdot U(1,1) + 1 \cdot U(1,2) + 0 \cdot U(2,1)$$

$$\Delta = U(1,2)$$
Now we have an estimate of Bellman equation based on $(1,1) \rightarrow (1,2)$ in a trial.

\[
U(1,1) = R(1,1) + \gamma \Delta = R(1,1) + \gamma \cdot U(1,2)
\]
Suppose we maintain update-to-date estimates of all utilities, including $U(1,1)$ and $U(1,2)$. Let $V(1,1)$ be our current estimate of utility at $(1,1)$. How can we update $V(1,1)$ as informed by the estimated Bellman equation?

$$U(1,1) = R(1,1) + \gamma \times U(1,2)$$
Update $V(1,1)$ cautiously towards $U(1,1)$

\[
V(1,1) = V(1,1) + \alpha \left[ U(1,1) - V(1,1) \right]
\]

\[
= V(1,1) + \alpha \left[ R(1,1) + \gamma \cdot U(1,2) - V(1,1) \right]
\]
Method 3: Temporal Difference Learning

Idea: update utility of $s$ after the agent moves from $s$ to $s'$.

$$U^\pi(s) \leftarrow U^\pi(s) + \alpha(R(s) + \gamma U^\pi(s') - U^\pi(s))$$

\[ V(1,1) = V(1,1) + \alpha \left[ R(1,1) + \gamma U(1,2) - V(1,1) \right] \]

$\alpha$ is often referred to as "learning rate".
Method 3: Temporal Difference Learning

1. Put agent at \( s = s_0 \). Set learning rate \( \alpha \).

2. Agent takes action \( \text{policy}(s) \) and moves to \( s' \).

3. If \( s' \) is newly visited, set \( R(s') = r \) and \( U(s') = r \).

4. Update utility of \( s \) by (not \( s' \)!
   not else!)

\[
U^\pi(s) \leftarrow U^\pi(s) + \alpha (R(s) + \gamma U^\pi(s') - U^\pi(s))
\]

5. Update \( s = s' \) and repeat 2-5 until convergence. (If agent terminates, put it back to \( s_0 \).)

Initialize utilities arbitrarily.
Example

\[ U^\pi(s) \leftarrow U^\pi(s) + \alpha(R(s) + \gamma U^\pi(s') - U^\pi(s)) \]

<table>
<thead>
<tr>
<th>U(1,1)</th>
<th>U(2,1)</th>
<th>U(1,2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Initialization

1. place agent at initial state (1,1).
2. choose a learning rate, e.g., $\alpha = 0.6$.
3. newly visit (1,1), set $R(1,1)=r$ and $U(1,1)=r$.

\[
U^\pi(s) \leftarrow U^\pi(s) + \alpha (R(s) + \gamma U^\pi(s') - U^\pi(s))
\]

<table>
<thead>
<tr>
<th>U(1,1)</th>
<th>U(2,1)</th>
<th>U(1,2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Iteration 1

1. take action UP at (1,1), end up moving right.

2. newly visit (2,1), set \( R(2,1) = r \) and \( U(2,1) = r \)

3. update utility at (1,1)

\[
U(1,1) \leftarrow U(1,1) + \alpha \{ R(1,1) + \gamma U(2,1) - U(1,1) \}
\]

<table>
<thead>
<tr>
<th>( U(1,1) )</th>
<th>( U(2,1) )</th>
<th>( U(1,2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r + \alpha^*\gamma^*r )</td>
<td>( r )</td>
<td>0</td>
</tr>
</tbody>
</table>
Iteration 2

1. take action LEFT at (2,1), end up moving left.
2. (1,1) is not newly visit, do nothing
3. update utility at (2,1)

\[ U(2,1) \leftarrow U(2,1) + \alpha \left( R(2,1) + \gamma U(1,1) - U(2,1) \right) \]

<table>
<thead>
<tr>
<th>U(1,1)</th>
<th>U(2,1)</th>
<th>U(1,2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>r + \alpha \gamma r</td>
<td>r + \alpha \gamma (r + \alpha \gamma r)</td>
<td>0</td>
</tr>
</tbody>
</table>
Iteration 3

1. take action UP at (1,1), end up moving up.

2. newly visit (1,2), set $R(1,2)=r$ and $U(1,2)=r$

3. update utility at (1,1)

$$U(1,1) \leftarrow U(1,1) + \alpha \{ R(1,1) + \gamma U(1,2) - U(1,1) \}$$

<table>
<thead>
<tr>
<th>U(1,1)</th>
<th>U(2,1)</th>
<th>U(1,2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r + 2\alpha \gamma r - \alpha^2 \gamma r$</td>
<td>$r + \alpha \gamma^* (r + \alpha \gamma^* r)$</td>
<td>$r$</td>
</tr>
</tbody>
</table>
Explore until 1st trial terminates.

(1,1) \[ U(1,1) = r \]

(1,1) → (2,1) \[ U(1,1) \rightarrow U(1,1) + \alpha \{ R(1,1) + \gamma U(2,1) - U(1,1) \} \]

(2,1) → (1,1) \[ U(2,1) \rightarrow U(2,1) + \alpha \{ R(2,1) + \gamma U(1,1) - U(2,1) \} \]

(1,1) → (1,2) \[ U(1,1) \rightarrow U(1,1) + \alpha \{ R(1,1) + \gamma U(1,2) - U(1,1) \} \]

(1,2) → (1,3) \[ U(1,2) \rightarrow U(1,2) + \alpha \{ R(1,2) + \gamma U(1,3) - U(1,2) \} \]

(1,3) → (2,3) \[ U(1,3) \rightarrow U(1,3) + \alpha \{ R(1,3) + \gamma U(2,3) - U(1,3) \} \]

......

......
2nd Trial of Exploration

(1,1)   U(1,1) is result from 1st trial.
(1,1) → (1,2)  U(1,1) = U(1,1) + α \{R(1,1)+γU(1,2)-U(1,1)\}
(1,2) → (1,3)  U(1,2) = U(1,2) + α \{R(1,2)+γU(1,3)-U(1,2)\}
……  ……
[Q] True or False?

\[
\begin{align*}
U(1,1) & \rightarrow U(1,1) + \alpha \{R(1,1) + \gamma \cdot U(1,2) - U(1,1)\} \\
U(1,2) & \rightarrow U(1,2) + \alpha \{R(1,2) + \gamma \cdot U(1,3) - U(1,2)\} \\
& \cdots \\
U(2,3) & \rightarrow U(2,3) + \alpha \{R(2,3) + \gamma \cdot U(2,3) - U(2,3)\}
\end{align*}
\]
Simulation Results

Graphs showing simulation results with Utility estimates and RMS error in utility against the number of trials.
Adaptive Dynamic Programming

1. \( s=s_0 \). set \( N[s,a] \) and \( N[z|s,a] \).

2. take action \( a \) and moves to \( s' \).

3. If \( s' \) is new, set \( R(s') = r \) and \( U(s') = r \).

4. else, update \( N[s,a] \), \( N[z|s,a] \) and 
   \[
   \text{Pr}(z|s,a) = \frac{N(z|s,a)}{N(s,a)}
   \]

5. update utility (solve \( n \) linear equations)

6. \( s=s' \), repeat 2-5 until convergence.

Temporal Difference Learning

1. \( s=s_0 \). set \( \alpha \).

2. take action \( a \) and moves to \( s' \).

3. If \( s' \) is new, set \( R(s') = r \) and \( U(s') = r \).

4. Update \( U(s) \)
   \[
   U^\pi(s) \leftarrow U^\pi(s) + \alpha(R(s) + \gamma U^\pi(s') - U^\pi(s))
   \]

5. \( s=s' \), repeat 2-4 until convergence.
ADP vs TDL

Which is more robust, converge faster, or more computationally efficient?
ADP vs TDL

ADP is more robust, converges faster; TDL is more computationally efficient.
Passive RL: given a policy, how to identify utility?

Monte Carlo
- estimate utility from samples (in a black-box fashion)

Adaptive Dynamic Programming
- estimate transition probabilities in trials and identify utilities (by solving equations)

Temporal Difference Learning
- estimate transition probabilities in trials and update utilities
Discussion

Can we just estimate all transition probabilities (and rewards) and transform reinforcement learning to a MDP problem?

\[ P(s'=(1,2) \mid s=(1,1), a=UP) = \frac{2}{3} \]
\[ P(s'=(2,1) \mid s=(1,1), a=UP) = \frac{1}{3} \]
\[ P(s'=(1,1) \mid s=(1,1), a=UP) = 0 \]