Continuous state: $V(s)$
Continuous state: $\mathbb{Q}(s, a)$
Continuous state: $Q(s, a)$

$pos = 0.3, \; vel = -0.3$

$Q(s, forward) = 60$

$Q(s, reverse) = 75$

$Q(s, neutral) = 68$
Continuous state: $Q(s, a)$; Discrete actions

$pos = 0.3, \ vel = -0.3$

$Q(s, \text{forward}) = 60$

$Q(s, \text{reverse}) = 75$

$Q(s, \text{neutral}) = 68$
Continuous state, continuous action: $Q(s,a)$
Continuous state, continuous action: $Q(s,a)$
Episodic tasks

Discounting: $\gamma$

$V_{\pi_1}(s_0) = \ldots$

$V_{\pi_2}(s_0) = \ldots$
Episodic tasks

Discounting: \( \gamma \)

\[ \nu_{\pi_1}(s_0) = 0 + 0 - 3 \gamma^2 + 10 \gamma^3 \]

\[ \nu_{\pi_2}(s_0) = 0 + 0 + 2 \gamma^3 + 0 + 10 \gamma^5 \]

which policy is better?
Episodic tasks

Discounting: $\gamma$

$V_{\pi_1}(s_0) = 0 + \gamma - 3\gamma^2 + 10\gamma^3$

$V_{\pi_2}(s_0) = 0 + 0 + 2\gamma^3 + 0 + 10\gamma^5$

which policy is better?

$\gamma = 1$: $V_{\pi_1}(s_0) = 7$  $V_{\pi_2}(s_0) = 12$

$\gamma = 0.5$: $V_{\pi_1}(s_0) = 1.175$  $V_{\pi_2}(s_0) = 0.5625$

Two meaning of $\gamma$:

1)

2)
Episodic tasks

Discounting: \( \delta \)

\[ v_{\pi_1}(s_0) = 0 + D - 3\delta^2 + 10\delta^3 \]

\[ v_{\pi_2}(s_0) = 0 + 0 + 2\delta^3 + 0 + 10\delta^5 \]

which policy is better?

\( \delta = 1 \):

\[ v_{\pi_1}(s_0) = 7 \quad v_{\pi_2}(s_0) = 12 \]

\( \delta = .5 \):

\[ v_{\pi_1}(s_0) = 1.175 \quad v_{\pi_2}(s_0) = .5625 \]

Two meanings of \( \delta \):

1) interest/inflation
2) probability of episode ending \((1-\delta)\)
Continuing tasks

Discounting: \( \delta \)

\[ V_{\pi_1}(s_0) = \]

\[ V_{\pi_2}(s_0) = \]
Continuing tasks

Discounting: \( \gamma \)

\[
v_{\pi_1}(s_0) = 0 + 0 - 3\gamma^2 + 10\gamma^3 + \\
0 + 0 - 3\gamma^6 + 10\gamma^7 + \\
0 + 0 - 3\gamma^{10} + 10\gamma^{11} + \cdots
\]

\[
v_{\pi_2}(s_0) = 0 + 0 + 0 + 2\gamma^3 + 0 + 10\gamma^5 + \\
0 + 0 + 0 + 2\gamma^9 + 0 + 10\gamma^{11} + \cdots
\]

which policy is better?
Continuing tasks

Discounting: $\gamma$

$$v_{\pi_1}(s_0) = 0 + 0 - 3\delta^2 + 10\delta^3 + 0 + 0 - 3\delta^6 + 10\delta^7 + 0 + 0 - 3\delta^{10} + 10\delta^{11} + \cdots$$

$$v_{\pi_2}(s_0) = 0 + 0 + 0 + 2\delta^3 + 0 + 10\delta^5 + 0 + 0 + 2\delta^9 + 0 + 10\delta^9 + \cdots$$

which policy is better?

Discrete state (tabular): Depends on $\gamma$

Continuous state (function approx.): Might not depend on $\gamma$!
Continuing tasks

Average reward RL

\[ r(\pi_1) = \]

\[ r(\pi_2) = \]

which policy is better?
Continuing tasks

Stand
Clap
Wave

Average reward RL

\( r(\pi_1) = \frac{7}{4} \)

\( r(\pi_2) = \frac{12}{6} = 2 \)

which policy is better?
Differential value function

\[ r(\pi^*_2) = 2 \quad \text{estimated by algorithm in the book: \( \beta \)} \]

Differential semi-gradient SARSA (R-learning)

\[ V_\pi(s) = \sum_a \pi(a|s) \sum_{s'} \rho(s' | s, a) \left[ r(s) - r(\pi) + V_\pi(s') \right] \]
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_1$</td>
<td>+1</td>
<td>-3</td>
<td>-3</td>
<td>+10</td>
<td>+10</td>
<td>-3</td>
</tr>
<tr>
<td>$\pi_2$</td>
<td>+1</td>
<td>-3</td>
<td>-3</td>
<td>+10</td>
<td>+10</td>
<td>-3</td>
</tr>
</tbody>
</table>

Differential value function

$\gamma(\pi_2) = 2$

$V(A) = ?$
$V(B) = -2$
$V(C) = 0$
$V(D) = ?$
$V(E) = ?$
$V(F) = ?$

$V_\pi(s) = \frac{1}{a} \max_{a} \min_{a'} \rho(s', r(s', a)) [r - r(\pi) + V_\pi(s')]$
Differential value function

![Diagram showing a grid with actions and rewards]

Can this be $V$?

<table>
<thead>
<tr>
<th>State</th>
<th>$V(A)$</th>
<th>$V(B)$</th>
<th>$V(C)$</th>
<th>$V(D)$</th>
<th>$V(E)$</th>
<th>$V(F)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_1$</td>
<td>-4</td>
<td>-2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

$$V_\pi(s) = \max_{a} \min_{s'} \rho(s', r(s, a)) \left[ r + V_\pi(s') - r(\pi) \right]$$
Differential value function

\[ r(\pi_2) = 2 \]

Can this be \( V \)?

\[ V(A) + V(B) + \ldots + V(F) = 2 \]

But avg. value of a cycle must be 0....

\[ V^\pi(s) = \frac{1}{n} \sum_{a \in A} \sum_{s' \in S} \rho(s', r(s, a) \left[ r - r(\pi) + V^\pi(s') \right] \]
Differential value function

What's the steady state distribution of $\pi_2$?

Can this be $V$?

$$V(A) + V(B) + \ldots + V(F) = 2$$

But avg. value of a cycle must be 0,....

$$v_\pi(s) = \frac{1}{\alpha} \tau_\pi(al_s) \leq \rho(s',r|s,a)[r - r(\pi) + v_\pi(s')]$$
Differential value function

![Diagram of a grid with states and transitions]

\[ r(\pi_e) = 2 \]

What's the steady state distribution of \( \pi_e \)?

Is this MDP ergodic? (what does ergodic mean?)

\[ V(A) = -4 - \frac{1}{3} = -\frac{13}{3} \]

\[ V(B) = -2 - \frac{1}{3} = -\frac{7}{3} \]

\[ V(C) = 0 - \frac{1}{3} = -\frac{1}{3} \]

\[ V(D) = 2 - \frac{1}{3} = \frac{5}{3} \]

\[ V(E) = 2 - \frac{1}{3} = \frac{5}{3} \]

\[ V(F) = 4 - \frac{1}{3} = \frac{11}{3} \]

Can this be \( V \)?

\[ V(A) + V(B) + \ldots + V(F) = 2 \]

But avg. value of a cycle must be 0, ...

\[ V_\pi(s) = \sum_a \pi(a|s) \sum_{s'} \rho(s',r|s,a) \left[ r - r(\pi) + V_\pi(s') \right] \]
Differential value function

What's the steady state distribution of $\pi_z$?

Is this MDP ergodic? (i.e. does every policy have a steady state distribution independent of $S_0$?)

Can this be $V$?

$V(A) + V(B) + ... + V(F) = 2$

But avg. value of a cycle must be 0. . . . . .

$V_\pi(s) = \sum_a \pi(a|s) \sum_{s'} \rho(s', r|s, a) [r - r(\pi) + \nu_\pi(s')]$
What's the steady state distribution of \( \pi' \)?

Is this MDP ergodic? YES

*IF transitions are at least slightly stochastic\(^\dagger\)*

\[
v_\pi(s) = \frac{1}{a} \sum_{a} \pi'(a|s) \sum_{s', \alpha} \rho(s', \alpha|s, a) \left[ r(s') + v_\pi(s') \right]
\]
What's the steady state distribution of $\pi_2$?
Is this MDP ergodic? YES

What about this MDP? YES*

What about this MDP?

*IF transitions are at least slightly stochastic(?)

$$v_\pi(s) = \sum_a \pi(als) \sum_{s',r} p(s',r|s,a)[r - r(\pi) + v_\pi(s')]$$
What's the steady state distribution of $T_e$?  
Is this MDP ergodic? YES

What about this MDP? YES*

What about this MDP? NO

There are different stationary distributions for the same policy.  
A transition can happen that affects where the agent can get (eventually).

* IF transitions are at least slightly stochastic $(\triangleright)$

\[
v_\pi(s) = \sum_a \pi(a|s) \sum_{s', r} p(s', r|s, a) [r - r(\pi) + v_\pi(s')] \]

Differential value function