Learning All Optimal Policies with Multiple Criteria

Leon Barrett & Srini Narayanan

University of California, Berkeley

July 7, 2008
Motivation

- Standard Reinforcement Learning: single reward
- Multi-criterion learning, reduce to standard RL
- We lift to solve over all preferences at once
  - Can view all optimal policies
  - Can change preferences at runtime, without relearning
Reinforcement Learning: Important Components

- Maximize expected discounted reward
  - Summarize with $V$ and $Q$

- Bellman equations: recurrence
  - $Q^*(s, a) = \mathbb{E}[R(s, a) + \gamma V^*(s')]$
  - $V^*(s) = \max_a Q^*(s, a)$
Reward Decomposition

- Arbitrary choices
- Or twiddle to get desired behavior

We make weights explicit:

\[ R(s, a) = \vec{R}(s, a) \cdot \vec{w} \]
Q-Values in Space!

- $V(s_0)$
- $Q(s_0, a_0)$
- Each policy gives one value
Q-Values in Space!

- $V(s_0)$
- $Q(s_0, a_0)$
- Each policy gives one value
Q-Values in Space!

- $V(s_0)$
- $Q(s_0, a_0)$
- Each policy gives one value
Q-Values in Space!

Each policy gives one value.

- \( V(s_0) \)
- \( Q(s_0, a_0) \)
Revised Recurrences

\[ Q^*(s, a) = \mathbb{E}[\vec{R}(s, a) + \gamma V^*(s')] \]

\[ V^*(s) = \text{hull} \bigcup_a Q^*(s, a) \]
Given a $\vec{w}$

- Extract optimal value by taking max
- For all $\vec{w}$, solution identical to standard RL
  - Because max in any direction must be on hull
### Example Results

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>![Image 5]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>![Image 6]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Leon Barrett & Srini Narayanan
Example Results

Leon Barrett & Srini Narayanan
Example Results

Leon Barrett & Srini Narayanan
Example Results

Leon Barrett & Srini Narayanan
Complexity

- $O(n^d)$ for high dimension
- Efficient for 2D and 3D

Efficiency tricks
- Witnesses: check with previous hull
- Constrain $\vec{w}$ space
POMDPs

Rewrite as POMDP

\[ P(w) \leftrightarrow \overrightarrow{w} \]
Contributions

- New class of results: *all* optimal policies
  - Via convex hull version of Bellman recurrence
  - Complete view of useful policy space
  - On-line preference switching
Future Work

- Combine with POMDPs
- Inverse problem: determine range of $w$
  - Extract agent preferences
- Different discounting rates $\gamma$
  - Approximate hyperbolic discounting
Thanks To

- My guinea pigs

Louis  Milo  Chester