Challenges in Teaching Learners With Preferences

- Object-world gathering game
  - * yields reward 1.0, + yields reward 0.9
  - Learner’s preference: Avoid frequent proximity of green cells (≤ 1-cell distance)

• Providing demonstrations from optimal behavioral policy π∗ can lead to arbitrarily bad learner’s performance!

A Teacher and an IRL Learner Without Preferences

• MDP M = (S, A, P, R, γ) with rewards R(s) = (w∗, φ(s))
• Teacher T provides demonstrations using policy πT
• Policy π has reward R(π) = (w∗, μ(π)), where μ(π) = E[∑s∈S γφ(s) | π]
• Learner receives demonstrations and outputs π∗ s.t. ||μ(π∗) − μ(πT)|| ≤ ε
• This ensures that R(π∗) ≥ R(πT) − ε

Learner-aware Teaching for Known Constraints

Learner-aware teaching for hard preferences: AWARE-CMDP

• Define a set of feasible reward feature expectations Cπ = {μ(π) | g(μ(π)) ≤ 0}
• Optimal teaching policy — solution of constrained MDP:
  \[ \max_{π∗} \langle w∗, μ(π∗) \rangle \text{ s.t. } μ(π∗) ∈ Cπ \]
• Theorem. The value of learner-aware teaching can be arbitrarily high, given by
  \[ \max_{π∗} \langle w∗, μ(π∗) \rangle − \langle w∗, \text{Proj}_{Cπ}(μ(π∗)) \rangle \]
• For linear g(·), the above problem can be solved via linear programming

Learner-aware teaching for soft preferences: AWARE-BL

• Optimal teaching problem can be formulated as a bi-level optimization:
  \[ \max_{π∗} \langle w∗, μ(π∗) \rangle \text{ s.t. } \|π∗\| ≤ \epsilon \]
• Here IRL(π, μ(π∗)) stands for the IRL problem solved by the learner
• Optimal teaching policy is a softmax policy satisfying the learner’s constraints
• A challenging non-convex optimization problem
• Proposed a gradient-based optimization approach

Experimental Results

Experimental setup

- Object-world gathering environment:
  - Rewards: + yields 1.0, + yields 0.9, - yields 0.2
  - Two “green” distractors at 0-cell and 1-cell distance to the + objects
  - Two “yellow” distractors at 1-cell and 2-cell distance to the + objects
  - Discount factor γ = 0.99
- Learners with soft preferences (Cπ = 5, Cπ = 10) and δsoft = 0
- Environment and learners’ preferences for 5 different learners L1, . . . , L5

For instance, L2 has two preference features indicating whether there is a green cell at a distance of 0-cells or 1-cell, respectively

Learner-aware teaching for known constraints

• Learners’ rewards inferred from learner-agnostic teacher (AGNOSTIC)

<table>
<thead>
<tr>
<th>Teacher</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>L4</th>
<th>L5</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGNOSTIC</td>
<td>7.99 ± 0.02</td>
<td>0.01 ± 0.00</td>
<td>0.01 ± 0.00</td>
<td>0.01 ± 0.00</td>
<td>0.00 ± 0.00</td>
</tr>
<tr>
<td>AWARE-BL</td>
<td>8.00 ± 0.02</td>
<td>7.20 ± 0.01</td>
<td>4.86 ± 0.30</td>
<td>3.15 ± 0.27</td>
<td>1.30 ± 0.07</td>
</tr>
</tbody>
</table>

Further Results

• Algorithms for learner-aware teaching with unknown constraints
• Additional experimental results
• Formal statements, proofs, and derivations