Bridging the Imitation Gap by Adaptive Insubordination

Luca Weihs¹,*  Unnat Jain²,*  Iou-Jen Liu²  Jordi Salvador¹  Svetlana Lazebnik²  Ani Kembhavi¹,³  Alexander Schwing²

¹ Allen Institute for AI (PRIOR Team)
² University of Illinois at Urbana-Champaign
³ University of Washington, Seattle
* Equal contribution
Imitation gap: Some experts are too smart to teach

Define the imitation gap for RL agents

Introduce ADVISOR

Describe our empirical results
Imitation gap: Some experts are too smart to teach

“Obviously…” Move forward Turn left Turn left/right Turn right
Imitation gap: Some experts are too smart to teach

“Obviously…”

- Move forward
- Turn left
- Turn left/right
- Turn right

Student’s Perspective
Imitation gap: Some experts are too smart to teach

“Obviously…”

Move forward
Turn left
Turn left/right
Turn right

Student’s Perspective
Imitation gap: Some experts are too smart to teach

“Obviously…”

Move forward    Turn left    Turn left/right    Turn right

Student’s Perspective
Imitation gap: Some experts are too smart to teach

Environment state

\[ f(s_1) = f(s_2) = f(s_3) = f(s_4) = 0 \]
If

Prop. 1. A student’s policy is the average of the teacher’s policy.

Then

$$\pi^{IL}(o) = \mathbb{E}[\pi^{exp}(S) \mid f(S) = o]$$
Prop. 1: A student’s policy is the average of the teacher’s policy.

If

\[ \pi^{IL} = \arg \min_{\text{student policies } \pi} \mathbb{E}[ \text{CrossEntropy}(\pi, \pi^{\text{exp}}) ] \]

Then

\[ \pi^{IL}(o) = \mathbb{E}[ \pi^{\text{exp}}(S) \mid f(S) = o ] \]
Prop. 1: A student’s policy is the average of the teacher’s policy.

\[ \pi^{IL}(o) = \mathbb{E}[ \pi^{exp}(S) \mid f(S) = o ] \]
Prop. 1: A student’s policy is the average of the teacher’s policy.

\[
\pi_{IL} \big( o \big) = \mathbb{E} \left[ \pi^{\text{exp}} \big( S \big) \mid f(S) = o \right]
\]
Experts using partial obs.

**Pros**
- No gap

**Cons**
- Expensive
- Challenging to code

**Imitation gap:**
In many realistic settings, even when optimizing perfectly,
\[ \pi_{IL} \neq \pi_{exp} \]

\[ g \left( \text{“Find the TV”} \right) = ? \]
Adaptive Insubordination (ADVISOR)

Intuition: When you can’t imitate the teacher, ignore them and self-study.

In practice: Adaptively reweigh IL and RL losses during training.

\[ \mathcal{L}(s) = \]

\[ w(s) \approx \begin{cases} 
1 & \text{Student can, in principle, imitate teacher at state } S \\
0 & \text{Otherwise} 
\end{cases} \]
Adaptive Insubordination (ADVISOR)

When you can’t imitate the teacher, ignore them and self-study.

In practice: Adaptively reweigh IL and RL losses during training.

Otherwise:

How do we estimate $\tau_\lambda$?

Key idea: Try to imitate using an auxiliary policy. When you can’t, use RL. When you can, use IL.

$$L(s) = w \cdot L_{RL}(s)$$

$$w(s) \approx \begin{cases} f(s) = 0 & \text{partial observation} \\ \text{enc} & \text{encoder} \end{cases}$$
Empirical Results

Exhaustive evaluation in discrete gridworlds:

- 15 RL & IL baselines,
- 9 tasks,
- careful hyperparameter control

<table>
<thead>
<tr>
<th>Tasks →</th>
<th>Training routines ↓</th>
<th><strong>PD</strong></th>
<th><strong>LAVA CROSSING</strong></th>
<th><strong>WALL CROSSING</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Base Ver.</td>
<td>Corrupt Exp.</td>
<td>Faulty Switch</td>
</tr>
<tr>
<td>RL</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>IL</td>
<td>-0.59</td>
<td>0.88</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>IL &amp; RL</td>
<td>-0.17</td>
<td>0.94</td>
<td>0.74</td>
<td>0.04</td>
</tr>
<tr>
<td>Demo. Based</td>
<td>-0.09</td>
<td>0.96</td>
<td>0.2</td>
<td>0.02</td>
</tr>
<tr>
<td>ADV. Based (ours)</td>
<td>1</td>
<td>0.96</td>
<td>0.94</td>
<td>0.77</td>
</tr>
</tbody>
</table>

ADVISOR succeeds in continuous and visually complex environments

<table>
<thead>
<tr>
<th>Tasks →</th>
<th><strong>PointGoal Navigation</strong></th>
<th><strong>ObjectGoal Navigation</strong></th>
<th><strong>Cooperative Navigation</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SPL @10%</td>
<td>SPL @100%</td>
<td>SPL @10%</td>
</tr>
<tr>
<td>Training routines ↓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RL only</td>
<td>30.9</td>
<td>54.7</td>
<td>6.7</td>
</tr>
<tr>
<td>IL only</td>
<td>30.1</td>
<td>68.7</td>
<td>3.8</td>
</tr>
<tr>
<td>IL + RL static</td>
<td>48.9</td>
<td>71.5</td>
<td>6.5</td>
</tr>
<tr>
<td>ADVISOR (ours)</td>
<td><strong>57.7</strong></td>
<td><strong>77.1</strong></td>
<td><strong>11.9</strong></td>
</tr>
</tbody>
</table>
If your expert is too smart to teach, treat them as an ADVISOR