Human and Agent Cooperative Learning

Matt Taylor

University of Alberta: Intelligent Robot Learning Lab (irll.ca)
Fellow-in-Residence: Alberta Machine Intelligence Institute (Amii.ca)
AI Redefined: Research Director (AI-R.com)

Canada CIFAR AI Chair, Amii
RL Applications

Developing vs. PoC vs. MVP

American Options
Exercise Policy

Data Center Cooling

Balloon Control

Tokamak Fusion

AlphaGO

Stock Trading

Water Treatment
How do we deploy more RL solutions?

- Better understand how to identify, de-risk, and tackle real world problems
  - Challenges of Real-World Reinforcement Learning
- Understand how and when to “cheat” by using external information
  - Existing agents/data
  - Human knowledge
Cooperative Learning
Cooperative Learning

Agent → Agent

- Offline / Batch RL
- Transfer Learning
- Curriculum Learning / Meta RL
- Advice
Cooperative Learning

Agent → Agent

Human → Agent

• Offline / Batch RL
• Demonstrations
• Curriculum Learning / Meta RL
• Advice
Cooperative Learning

Agent → Agent
Human → Agent
Agent → Human

• Intelligent Tutoring Systems
Agent → Agent Bootstrapping

Prior agent is optimal OK

How much data do I need?
How good is my policy? (OPE)
How sure am I about the policy’s performance? (HCOPE)

Agent → Agent Transfer

What to transfer? (From whom to transfer?)

How to transfer?

When to transfer?


- Offline / Batch RL
- Transfer Learning
- Curriculum Learning / Meta RL
- Advice
Agent → Agent Online Learning

Curriculum Learning
Meta RL


Muslimani, Lewandowski, Luo, Schuurmans

- Offline / Batch RL
- Transfer Learning
- Curriculum Learning / Meta RL
- Advice
Agent → Agent Advice

On-demand advice

Who initiates?
When do they provide?
Is there a cost?
Are there multiple teachers?

Agent → Agent Advice


- Offline / Batch RL
- Transfer Learning
- Curriculum Learning / Meta RL
- Advice
Multi-Agent Advisor Q-Learning.
Subramanian, S., Taylor, K. Larson, & M. Crowley. JMLR-22
ADvising Multiple Intelligent Agents (ADMIRAL)
Improving MARL sample efficiency

- We introduce: **Multi-agent action advising** for MARL
  - General methods (**no assumption** on advisor or type of environment)
  - **Two practical algorithms** to learn from advisor
  - **Principled method** to evaluate the advisor
  - **Theoretical guarantees** of convergence
Human → Agent

Reward signal?

No
• Demonstrations
• Feedback
• Preferences

Yes
• Demonstrations
• Feedback
• Preferences
• Action Advice
• Shaping Rewards

Human → Agent Feedback (!R)

Thomaz & Breazeal 2006: Anticipator
TAMER, Knox & Stone 2009

- Offline / Batch RL
- Demonstrations
- Curriculum Learning / Meta RL
- Advice
Human → Agent Feedback (!R)

Thomaz & Breazeal 2006: Anticipator
TAMER, Knox & Stone 2009: Numeric, Return
SABL, Loftin+ 2015

Feedback history $h$


... 

Really make sense to assign numeric rewards to these?
Human → Agent Feedback (!R)

Thomaz & Breazeal 2006: Anticipator
TAMER, Knox & Stone 2009: Numeric, Return
SABL, Loftin+ 2015: Categorical
COACH, McGlashlin+ 2017
Feedback can be Relative

Avoid  OK  Goal
Feedback can be Relative

Improving Condition:  bad → bad → alright
Steady Condition:    alright → alright → alright
Degradating Condition: good → good → alright

Advantage Function!
Human → Agent Feedback (!R)

- Offline / Batch RL
- Demonstrations
- Curriculum Learning / Meta RL
- Advice

Thomaz & Breazeal 2006: Anticipator
TAMER, Knox & Stone 2009: Numeric, Return
SABL, Loftin+ 2015: Categorical
COACH, McGlashlin+ 2017: Advantage Function
Human → Agent

Reward signal?

No
• Demonstrations
• Feedback
• Preferences

Yes
• Demonstrations
• Feedback
• Preferences
• Action Advice
• Shaping Rewards

Human → Agent Bootstrapping

Lay person
Subject Matter Expert
Programmer

Agent → Human: ITS

Convey information
Model user’s understanding
Model user’s learning

→ Sequential decision tasks
Agent → Human: ITS

How to practice
How to support
When to support

Fig. 1: The Mouselab-MDP paradigm. (Left) Participants click to reveal the value at future states. (Right) ITS provides feedback on each planning operation. The question mark represents optional elaborated feedback.

Agent → Human: Pilot Training

Shortage of pilots
Lots of knowledge
Hands on training

Figure 1: System Architecture

Program Synthesis

Write better code faster

Program Optimization with Locally Improving search (POLIS)
- A system for improving programs w.r.t. reward
- Local search algorithm exploits program structure
- Generate effective & short programs

Can You Improve My Code? Optimizing Programs with Local Search.
Fatemeh Abdollahi, Saqib Ameen, Levi Lelis, Taylor. IJCAI-23
POLIS

polis, plural poleis, ancient Greek city-state.... There were several hundred poleis, the history and constitutions of most of which are known only sketchily .... most ancient Greek history is recounted in terms of the histories of Athens, Sparta, and a few others.
def initial(o):
        action = 2
    elif o[1] > 1 and o[3] < -0.7 and o[0] < -0.05:
        action = 3
    elif o[1] > 1 and o[3] < -0.8 and o[0] > 0.1:
        action = 1
    elif o[0] < -0.15 and o[4] > 0.1:
        action = 3
    elif o[0] < 0.13 and o[4] < -0.1:
        action = 1
    elif o[1] < 0.8 and o[1] > 0.2:
        action = 2
    elif o[1] <= 0.2 and o[4] > 0.1:
        action = 3
    elif o[0] <= 0.2 and o[4] < -0.1:
        action = 1
    else:
        action = 0
    return action

def improved(o):
    if o[3] > -0.038:
        action = 0
    elif o[7] > 0.036:
        action = 2
    elif o[5] < -0.1:
        action = 1
    elif o[0] and o[6]:
        action = 2
    elif o[5] > 0.959:
        action = 0
    elif o[3] < -0.388:
        action = 2
    elif o[4] > 0.1:
        action = 3
    elif o[2] > 0.28:
        action = 1
    else:
        action = 3
    return action
... if o[1] == 4:
    if o[9] < 4:
        action = 2
    else:
        action = 0
else:
    action = 0
...

Objective function $F$

Improved program $p' \approx \text{argmax } F(p)$

Program Synthesis

Improve $p$
## Bottom-Up Search (BUS)

<table>
<thead>
<tr>
<th>Cost</th>
<th># Programs</th>
<th>Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>{x_1, x_2, x_3}</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>{\sqrt{x_1}, \sqrt{x_2}, \sqrt{x_3}}</td>
</tr>
<tr>
<td>3</td>
<td>75</td>
<td>{\sqrt{\sqrt{x_1}}, \sqrt{\sqrt{x_2}}, \sqrt{\sqrt{x_3}}, x_1 + x_1, \ldots, x_1 - x_1, x_1 - x_2, \ldots}</td>
</tr>
<tr>
<td>4</td>
<td>147</td>
<td>{\sqrt{\sqrt{\sqrt{x_1}}}, \ldots, \sqrt{x_1 + x_1}, \ldots, \sqrt{x_1 - x_1}, \ldots, \sqrt{x_1 + x_1}, \ldots, \sqrt{x_1 - x_1}, \ldots}</td>
</tr>
<tr>
<td>5</td>
<td>12K</td>
<td>{\ldots}</td>
</tr>
<tr>
<td>6</td>
<td>70K</td>
<td>{\ldots}</td>
</tr>
<tr>
<td>7</td>
<td>...</td>
<td>{\ldots, \sqrt{x_1 + x_2 + x_3}, \ldots}</td>
</tr>
</tbody>
</table>
Guided BUS: Probe  (Barke et. al. 2020)

<table>
<thead>
<tr>
<th>Cost</th>
<th># Programs</th>
<th>Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>{x_1, x_2, x_3}</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>{\sqrt{x_1}, \sqrt{x_2}, \sqrt{x_3}}</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>{\sqrt{x_1}, \sqrt{x_2}, \sqrt{x_3}}</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>{\sqrt[3]{x_1}, \ldots, x_1 + x_1, x_1 + x_2, x_1 + x_3, \ldots}</td>
</tr>
<tr>
<td>6</td>
<td>48</td>
<td>{\sqrt[3]{x_1}, \ldots, \sqrt{x_1 + x_2}, \ldots, x_1 + \sqrt{x_1}, \ldots, \sqrt{x_1 + x_1}}</td>
</tr>
<tr>
<td>7</td>
<td>93</td>
<td>{\ldots}</td>
</tr>
<tr>
<td>8</td>
<td>354</td>
<td>{\ldots}</td>
</tr>
<tr>
<td>9</td>
<td>3200</td>
<td>{\ldots}</td>
</tr>
<tr>
<td>10</td>
<td>...</td>
<td>{\ldots, \sqrt{x_1 + x_2 + x_3}, \ldots}</td>
</tr>
</tbody>
</table>
How does POLIS use BUS and Probe?

Program $p$
Line i
Objective function $F$
input-output examples

BUS or Probe synthesizes a new program for line i of $p$

evaluate $p$ with input-output example

is it among the top 20?

Yes

BUS* or Probe*

evaluate $p$ with $F$

No

Improved program for line i of $p$
Experimental details

Diagram:
- DQN agent
- Highlights
- Improve p
- Improve line i-th of p
- Program Synthesis BUS* or Probe*
- Bayesian Optimizer

Arrows:
- Program p
- Objective function F
- Improve p
- Improve line i-th of p
- POLIS
- Improved program p' ≜ argmax F(p)
- Restart
def initial(o):
        action = 4
        if o[1] == 4:
            if o[9] < 4:
                action = 2
            else:
                action = 0
        else:
            action = 0
    else:
        action = 3
    return action

def improved(o):
    if o[1] and o[3]:
        if o[1] and o[3]:
            action = 4
        else:
            if o[1] < 7.93:
                action = 2
            else:
                action = 0
    else:
        action = 2
    return action

Score ~6.8

Score ~39
Computational results: POLIS results

Lunar Lander

Highway
Research Question 1:
Can we teach people how to be better teachers?

Muslimani et al., 2021
Research Question 2: Can we adapt our algorithms to better learn from human teachers?

- Figure out what human feedback means?
Research Question 2: Can we adapt our algorithms to better learn from human teachers?

- Figure out what human feedback means?

Figure 2: The library of 16 environments is organized by the number of rooms and objects. There is a command list for each environment.

Peng et al., 2018
Research Question 3: Will Explainability Help?

- Explanations can help people select better agent and/or better anticipate agent’s actions

Davis-Pearson et al., under submission
Research Question 3: Will Explainability Help?

- Explanations can help people select better agent and/or better anticipate agent’s actions
- Knowing what the agents knows should let teacher better target how they help → seems obvious...

Davis-Pearson et al., under submission
Research Question 4: When is one type of help preferred?

- Teacher competence?
- Student capabilities?
- Speed of simulation?
- ...

Multi-agent, Multi-human Teaming

The **first platform** to allow the design, training, and deployment of complex **intelligence ecosystems**, mixing **humans and artificial agents** of various kinds.

It orchestrates heterogeneous ML & non-ML agents with real-time human interaction.
Human Input Parsing Platform for OpenAI Gym

In a web browser, human subjects can interact with Atari games, MuJuCo robots, etc.
- Give demonstrations
- Provide feedback
- Identify errors

Enable scaling up & out of HitL RL
- Built-in AWS support
- Integrate with MTurk

HIPPO GYM
hippogym.irll.ca
Conclusion: Many more questions!

We should cheat whenever possible

Lots of room for improvement

- Learning from agents/data
- Learning from humans
- Teaching humans

http://irll.ca
http://cogment.ai/