National Research Council Canada | Conseil national de recherches Canada

UNIVERSITY OF WATERLOO
FACULTY OF ENGINEERING
Department of Electrical & Computer Engineering

WATERLOO.AI
WATERLOO ARTIFICIAL INTELLIGENCE INSTITUTE

Balancing Information with Observation Costs

In Deep Reinforcement Learning

Colin Bellinger, Andriy Drozdyuk, Mark Crowley, Isaac Tamblyn
colin.bellinger@nrc-cnrc.gc.ca
https://web.cs.dal.ca/~bellinger/

Thursday 2 June, 2022

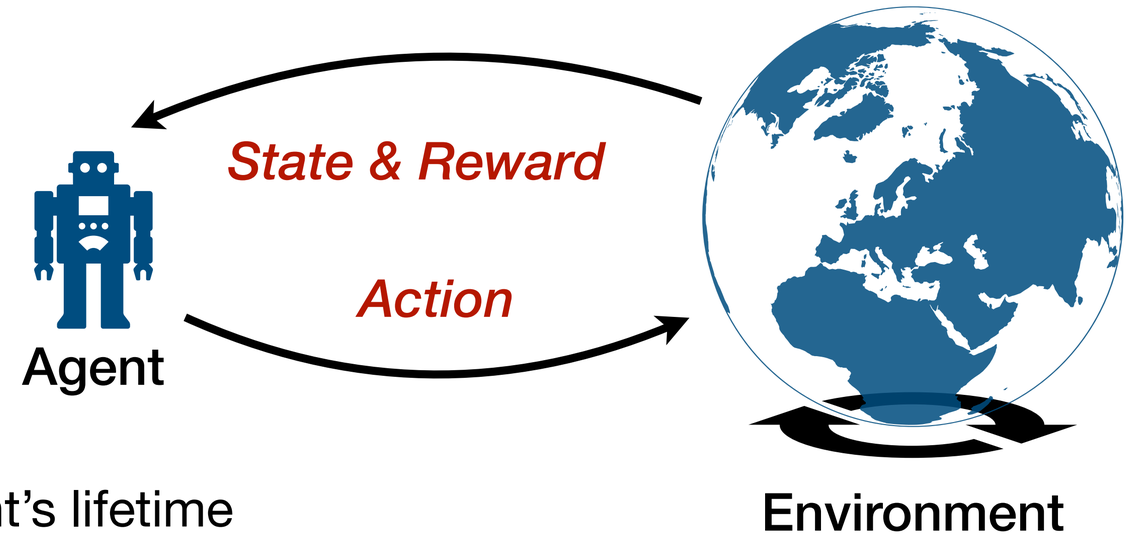
Outline

- Reinforcement learning background
- Measurement costs in reinforcement learning
- Expanded reinforcement learning framework
- Empirical evaluation with off-the-shelf deep reinforcement learning
- Conclusion

Overview of RL

Key Aspects

- RL applies to sequential decision making problem
- At each times step:
  - *Agent* selects an *action* to take
    - Actions changes the *environment*
  - *Environment* responds with a *reward* and new measurement of the *state* of the environment
- RL algorithms
  - Goal: optimize the sum of rewards over the agent's lifetime
  - Learn a *policy* to select actions given the current observation of the environment



Motivation

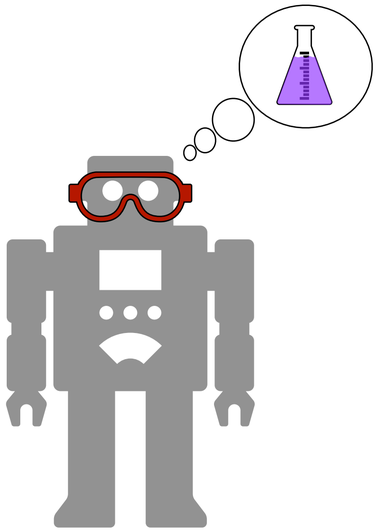
Measurement Costs in RL

- In some applications have high measurement costs
  - Materials design, automated chemistry, robotics, medical diagnostics
  - *Quality of an agent's policy* is impacted by:
    - Efficiency with which it achieves the goal
    - Number of costly measurements it required
- The agent must learn to *balance the need for information with the cost of information*
  - More low-confidence steps with fewer measurements can be optimal
  - RL framework does not provide a mechanism to learn this behaviour

Measurement Costs In RL

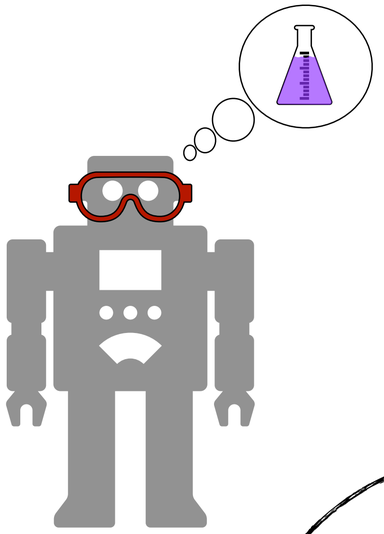
Problem Demonstration

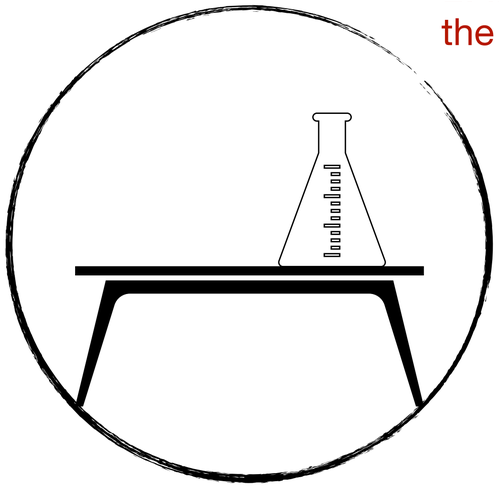
1. RL scientist with an goal



Measurement Costs In RL

Problem Demonstration

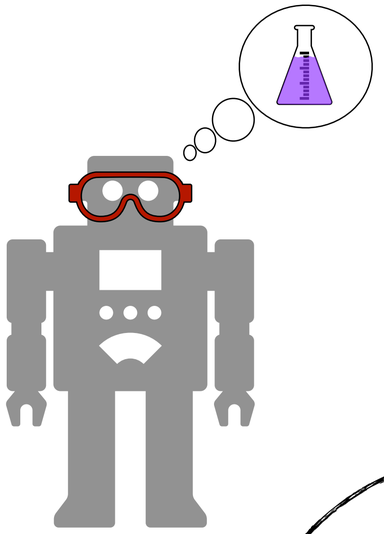


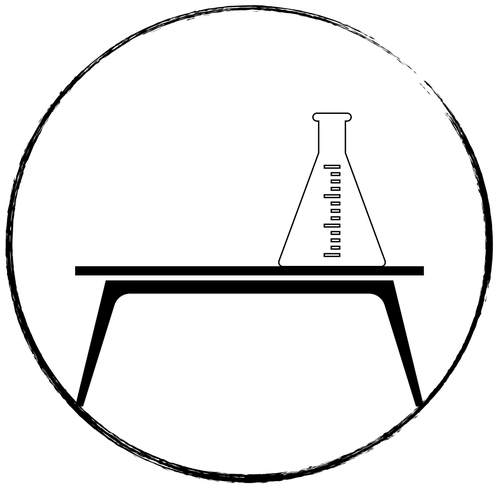


2. Initial state of the environment

Measurement Costs In RL

Problem Demonstration

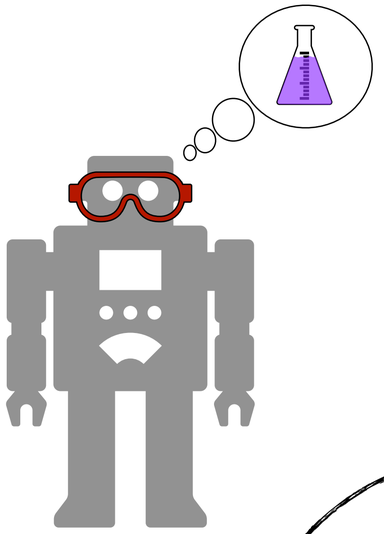




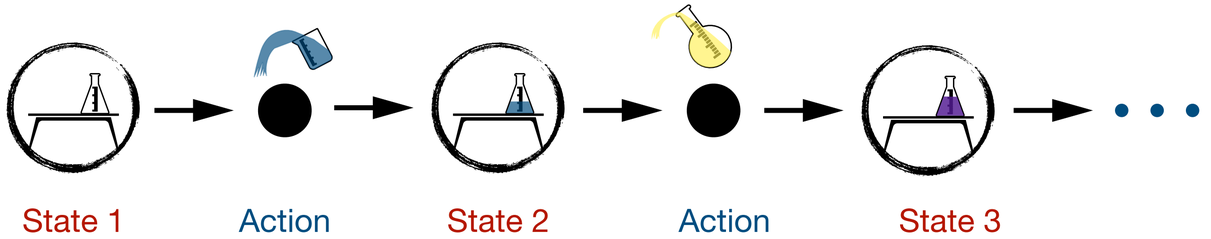
3. Actions available to the scientist agent

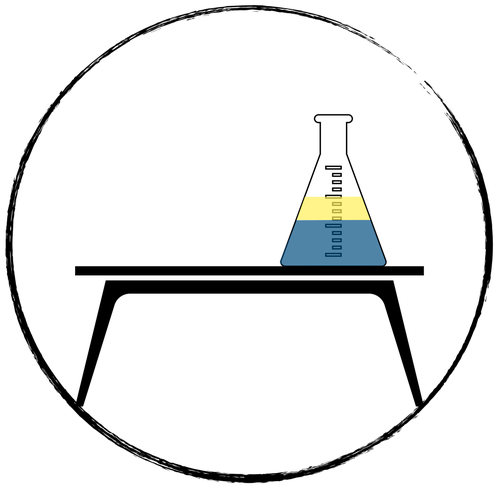
Measurement Costs In RL

Problem Demonstration



3. Actions available to the scientist agent

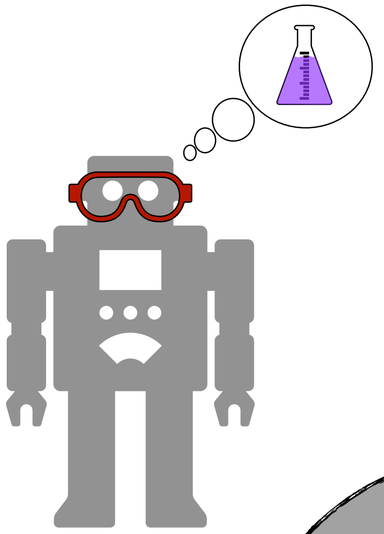


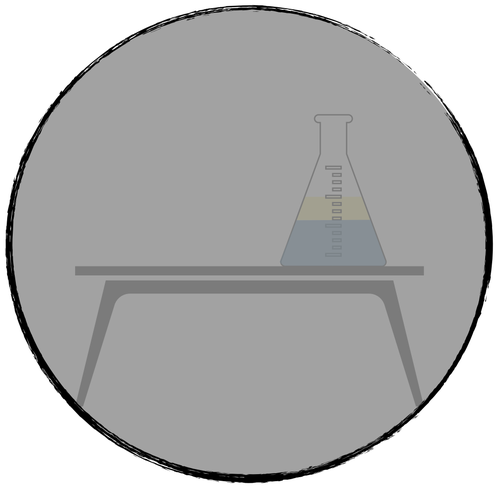


The *agent* sequential *applies actions and sees* the *updated state* of the environment

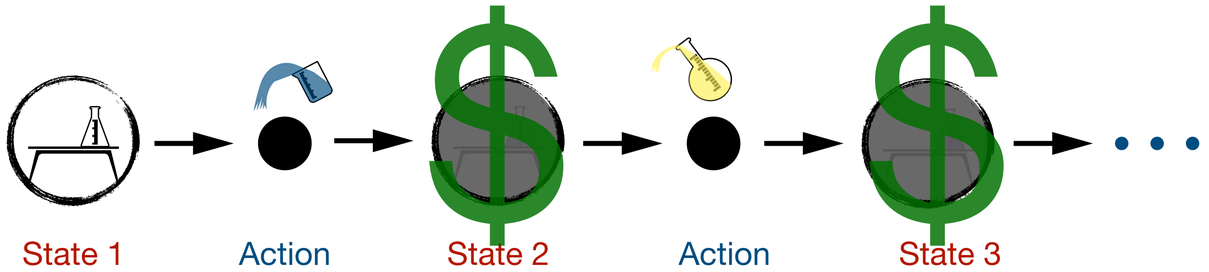
Measurement Costs In RL

Problem Demonstration





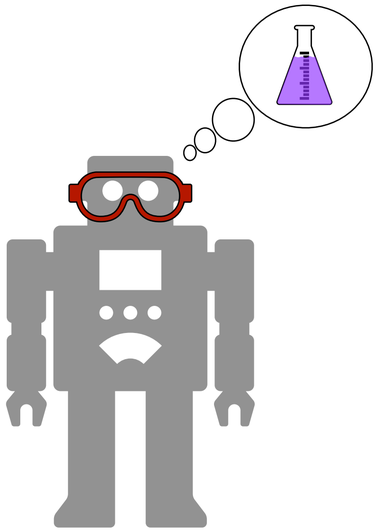
3. Actions available to the scientist agent



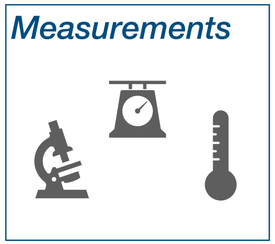
- ***Constantly measuring*** the next state can be ***costly***
- ***May not be necessary*** for an agent with good intuition

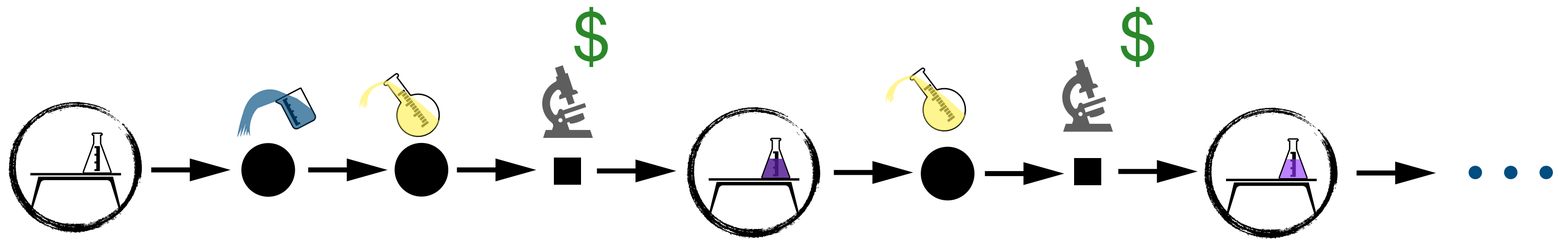
Measurement Costs In RL

Problem Demonstration



4. State measurements actions

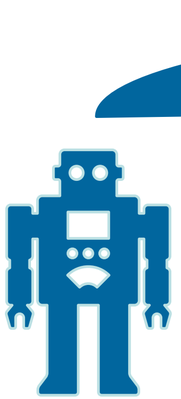




- State ***measurements are promoted to*** special ***actions***
- Agents can ***learn to measure when*** the ***benefit from additional information is greater than the cost***

Proposed RL Framework

With Explicit Measurement Actions and Costs



$$\pi_o :\rightarrow (p, m)$$

$$o_{t+1}, r_{t+1}$$

$$o_{t+1} = \begin{cases} P(s_{t+1} \mid s_t, p_t) \cap 1, & \text{if } m = 1 \\ \text{last\_measured} \cap 0, & \text{if } m = 0 \end{cases}$$

$$r_{t+1} = \begin{cases} R(s_t, p_t) - cost, & \text{if } m = 1 \\ R(s_t, p_t), & \text{if } m = 0 \end{cases}$$

$$\text{last\_measured} = \begin{cases} s_{t+1}, & \text{if } m = 1 \end{cases}$$

Proposed RL Framework

With Explicit Measurement Actions and Costs

- The agent's objective is to maximize the ***costed return***
  - (Long-term discounted sum of rewards) - (The explicit measurement costs)

Standard RL objective

Balance with measurement costs

- Framework is implementation as a wrapper to the Open AI Gym
  - https://github.com/cbellinger27/active-measureGymWrapper
  - Enables off-the-shelf DRL algorithms to learning balance measurement costs with reward

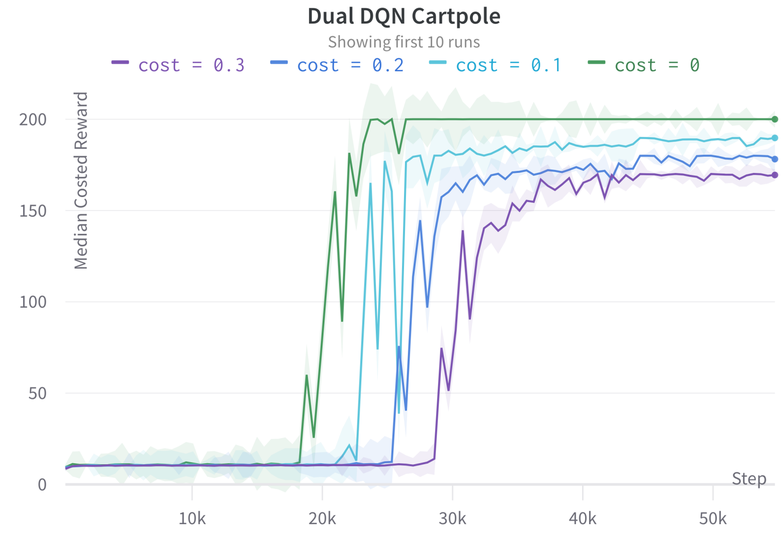
Experimental Method

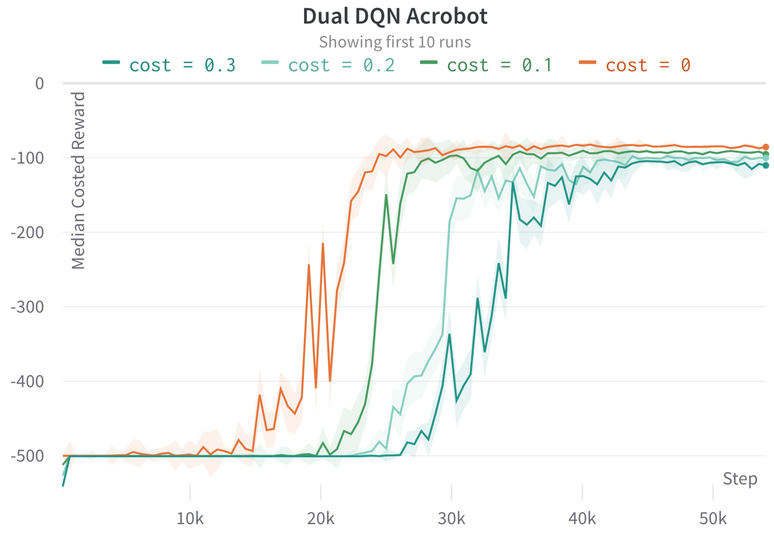
RL Algorithms and Analysis

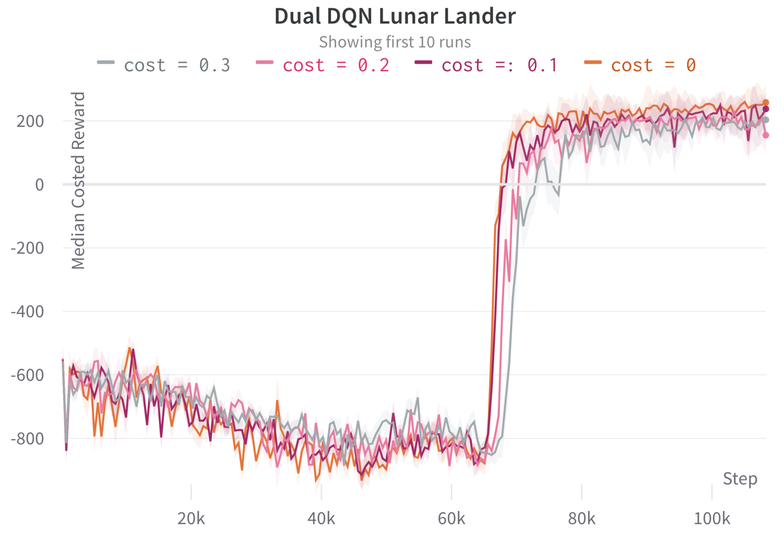
- *Off-the-shelf model-free DRL algorithms*
  - PPO (Schulman et al. 2017)
  - Dual DQN (Wang et al. 2016)
  - DRQN (Hausknecht, and Stone 2015)
- *Environments with explicit measurement actions and costs*
  - With classic controls problems from Open AI gym (Brockman et al. 2016)
    - Cartpole, Acrobot, and Lundar Lander
- *Evaluation metric*
  - Median costed reward on 20 independent trials
  - Measure pattern of the converged policies

Results

Median Costed Rewards



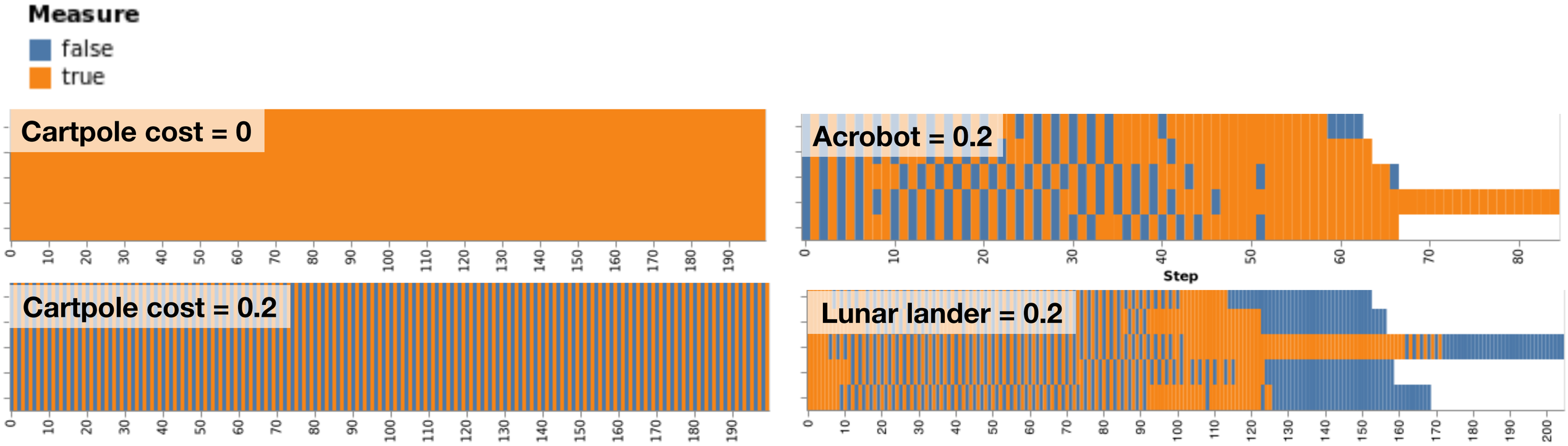




The agents solved each problem and reduced the number of measurements

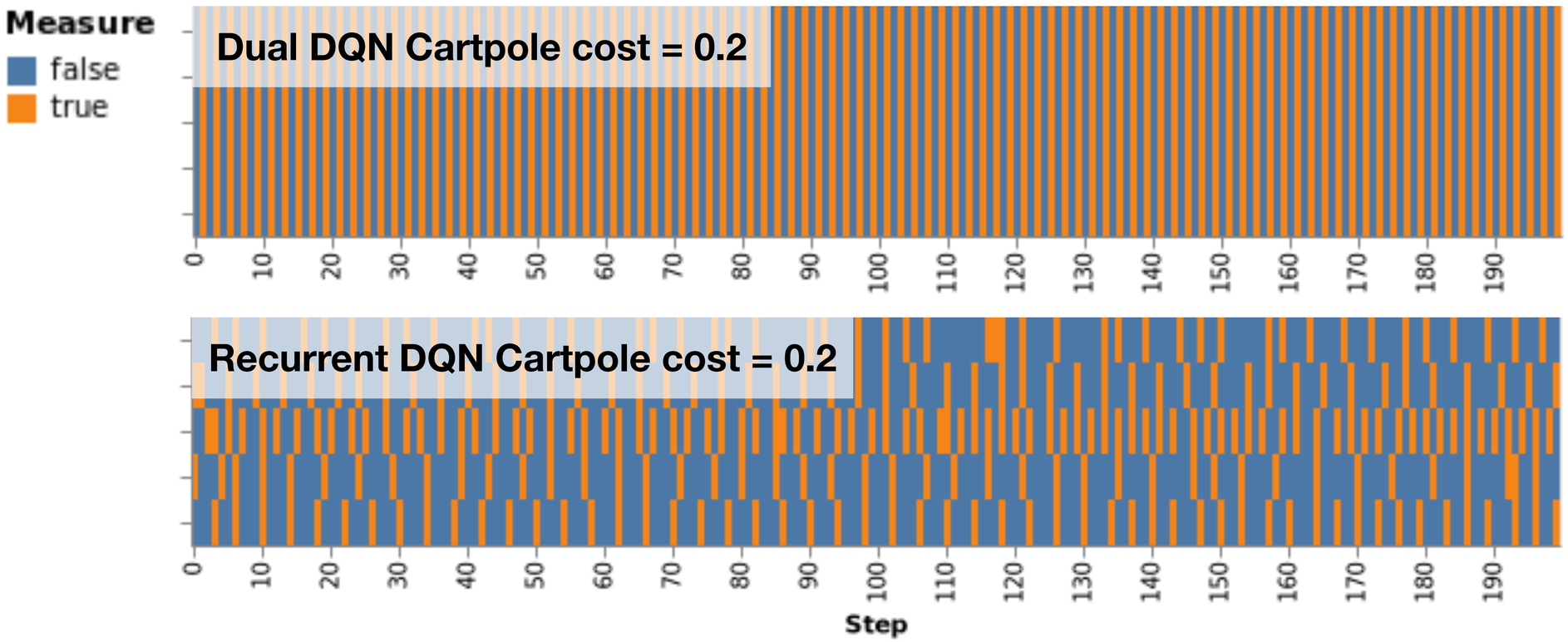
Results

Measurement Patterns



Results

Measurement Patterns



Conclusion

Future Work

- Improve efficiency of action space (action-pair) design
- Advance the use of recurrent agents
- Explore techniques from model-based RL and POMDPs
- Evaluating the framework in real-world applications
  - Simulated chemistry lab environment: https://www.chemgymrl.com/

Conclusion

Summary

- Goal: enable DRL algorithms to learn to balance the need for information with the cost of information
- Proposed an expanded RL framework with explicit measurement actions and costs
- Evaluated DQN, PPO and DRQN on the framework
  - DRL agents learned to reduce their measurements in reaction to cost
    - There is still room for improvement!

National Research Council Canada

Conseil national de recherches Canada

UNIVERSITY OF WATERLOO
FACULTY OF ENGINEERING
Department of Electrical & Computer Engineering

WATERLOO.AI
WATERLOO ARTIFICIAL INTELLIGENCE INSTITUTE

Thank you!

colin.bellinger@nrc-cnrc.gc.ca

https://web.cs.dal.ca/~bellinger/

https://caiac.pubpub.org/pub/0jmy7gpd

National Research Council Canada

Conseil national de recherches Canada

UNIVERSITY OF WATERLOO
FACULTY OF ENGINEERING
Department of Electrical & Computer Engineering

WATERLOO.AI
WATERLOO ARTIFICIAL INTELLIGENCE INSTITUTE