The Impact of Task Underspecification in Evaluating Deep Reinforcement Learning

Vindula Jayawardana , Catherine Tang, Sirui Li, Dajiang Suo, Cathy Wu

MIT

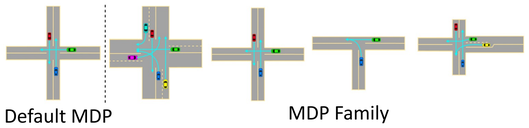
Correspondence: vindula@mit.edu

Introduction

- Standard practice in evaluating Deep Reinforcement Learning (DRL) for task-specific RL involves using a few instances of Markov Decision Processes (MDPs) to represent the task.
- However, many tasks induce a large ***family of MDPs*** owing to variations in the underlying environment.
- We refer to this phenomenon as the **task underspecification** problem in DRL evaluations.
- The select MDP instances may thus inadvertently cause overfitting, lacking the statistical power to draw conclusions about the method's true performance across the family.
- We show that in comparison to evaluating DRL methods on select MDP instances, evaluating the MDP family often yields a substantially different relative ranking of methods, casting doubt on what methods should be considered state-of-the-art.

Curse of Variety in Task-Specific RL

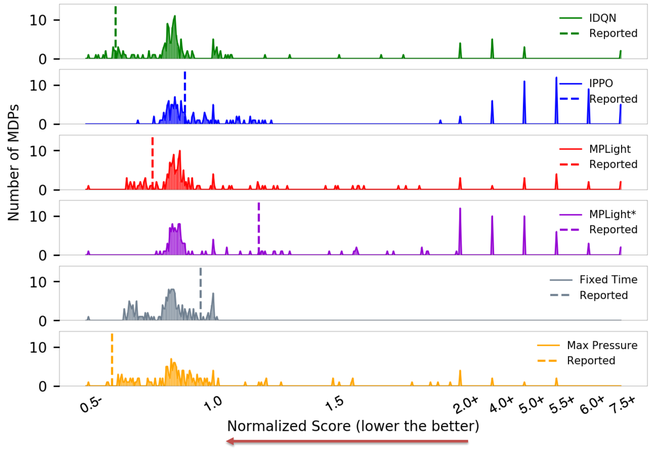
- One of the main challenges for task-specific reinforcement learning is the sheer diversity of the problem instances or, formally, possible MDP instances.
- For example, in traffic signal control, variations may stem from intersection geometries and traffic flow levels.



- However, most of the recent works on task-specific reinforcement learning have overlooked this requirement and use a few select MDP instances to evaluate the DRL algorithms.

Case Study: Traffic Signal Control

- Evaluate DRL for traffic signal control using RESCO benchmark.
  - **Four DRL algorithms**: *IDQN, IPPO, MPLight, MPLight**
  - **Two traditional algorithms**: *Fine-tuned Fixed Time, Max Pressure*
- Performance evaluated based on normalized average delay per vehicle.
  - Score normalization based on untuned-fixed time controller.
- 345 signalized intersections in Salt Lake City in Utah, are binned to 164 unique signalized intersections (164 unique MDPs)



- **Observation 1:** The performance of the methods can vary based on the chosen MDP.
- **Observation 2:** For most methods, the reported performance of the method in past evaluations overestimates the true performance of the method.

Overall Performance Across an MDP Family

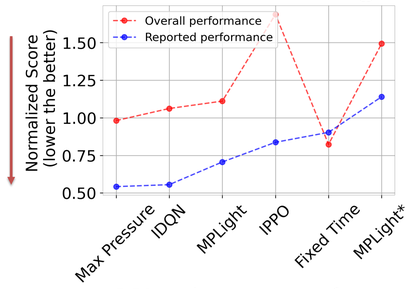
$$\mathbb{E}[X_R] = \sum_{i=1}^{|U|} s_{R,i} \times p_i$$

$s_{R,i}$ normalized performance score of method *R* on MDP *i*

p_i normalized importance score of MDP *i*

U MDP family

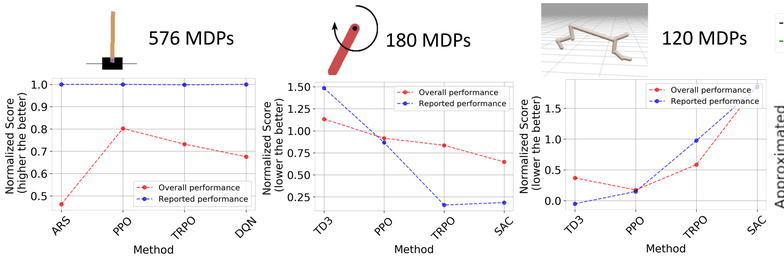
Re-evaluation: Traffic Signal Control



Rank	Reported	Overall (Ours)
1	Max Pressure	Fixed Time
2	IDQN	Max Pressure
3	MPLight	IDQN
4	IPPO	MPLight
5	Fixed Time	MPLight*
6	MPLight*	IPPO

DRL no longer outperforms any traditional methods!

Similar Phenomenon in Popular Benchmarks Tasks

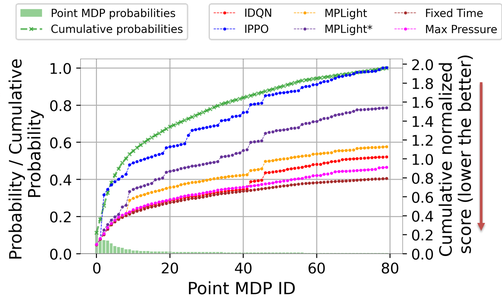


Challenges in MDP Family-based Evaluations

Challenge	Current Approach	Recommendation
Lack of evaluations benchmarks	Arbitrarily created few MDP instances as benchmarks.	Create benchmarks of MDP families with parameterized MDPs
Lack of emphasis on all MDP instances' performances	Tables with performances of selected MDP instances	Use performance profiles
Large families of MDPs with limited computational budgets	Arbitrarily selected few MDP instances for evaluations.	Use performance approximations

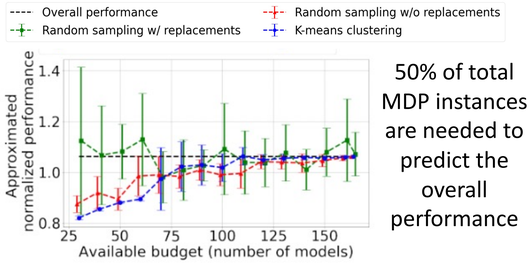
Performance Profiling

- A performance profile illustrates what MDPs instances are most probable in the distribution and how each of the MDP instances contributes to the final estimate of the performance.



Performance Approximations

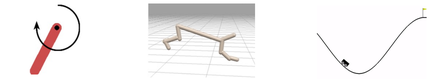
- Standardized performance approximations can be used to tackle evaluating under a computational budget.



50% of total MDP instances are needed to predict the overall performance

Evaluations Across Tasks

- Do MDP instances-based evaluations affect standard DRL evaluations ***across*** tasks?



Evaluation combinations: 180 x 120 x 120

Method	Rank 1	Rank 2	Rank 3	Rank 4
SAC	42.56%	29.09%	18.50%	9.85%
TRPO	26.23%	24.99%	30.45%	18.33%
PPO	18.75%	29.54%	27.74%	23.97%
TD3	12.46%	16.38%	23.31%	47.85%
Reported	SAC	TRPO	PPO	TD3