1 BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

2

3 Penn UNIVERSITY of PENNSYLVANIA

Cautious Adaptation For RL in Safety-Critical Settings

International Conference on Machine Learning 2020

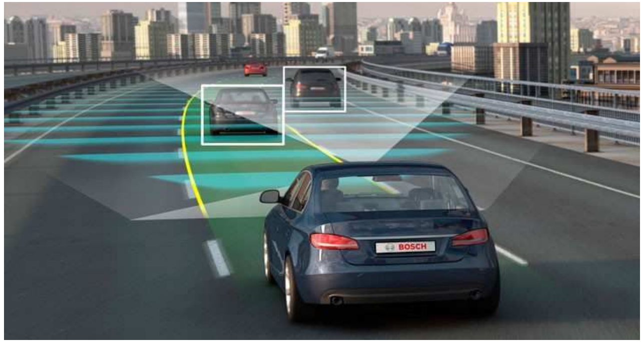
Jesse Zhang[1], Brian Cheung[1], Chelsea Finn[2], Sergey Levine[1], Dinesh Jayaraman[3]

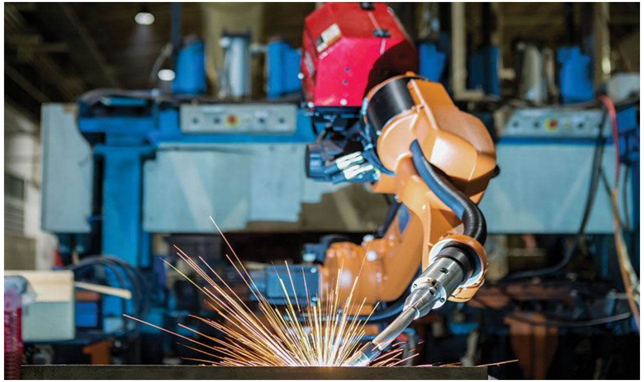
Outline

- Short overview (4 Minutes)
- In-depth talk(11 Minutes)

Introduction

- Real-world RL **hazardous** in safety-critical settings
- Hard to reset from real-life failures
- How to adapt to unseen environments safely?





Motivation

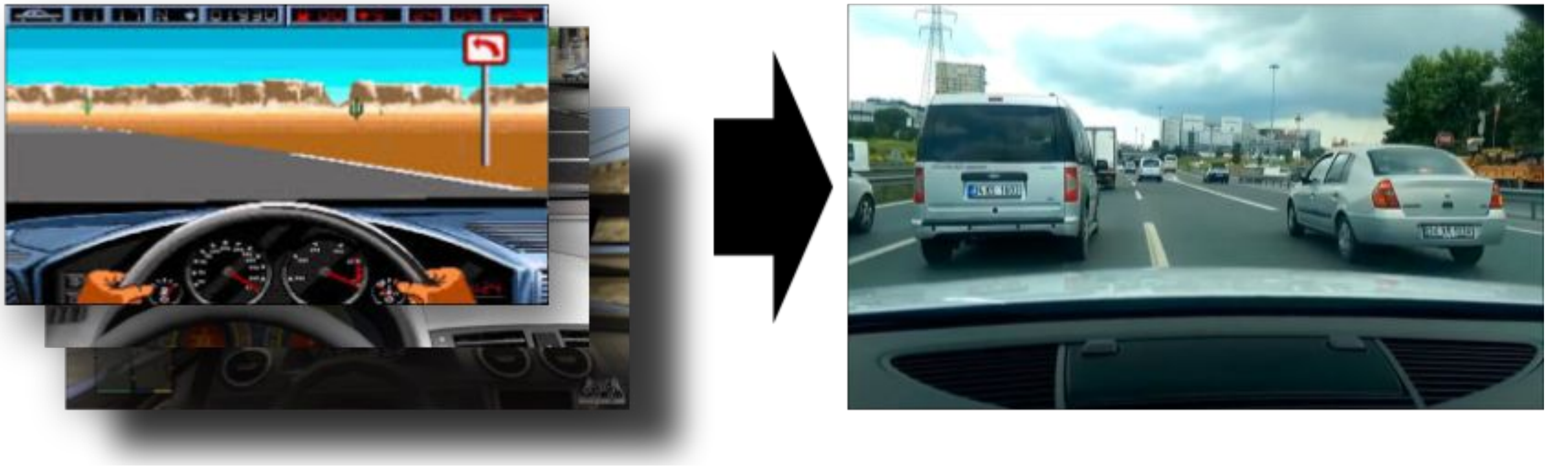
- How do humans adapt?

Motivation

- Safety-Critical Adaptation (SCA):
  - Pretraining: Sandbox environments
  - Adaptation: Safety-critical target environment

PRETRAINING

ADAPTATION



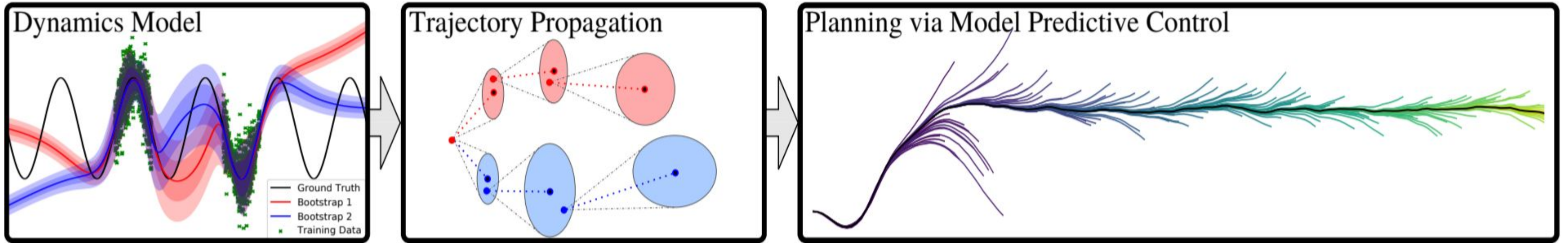
Methodology

Transfer risk knowledge from prior experience

- Safety-Critical Adaptation (SCA)
- Cautious Adaptation in RL (CARL)

Cautious Adaptation in RL (CARL)

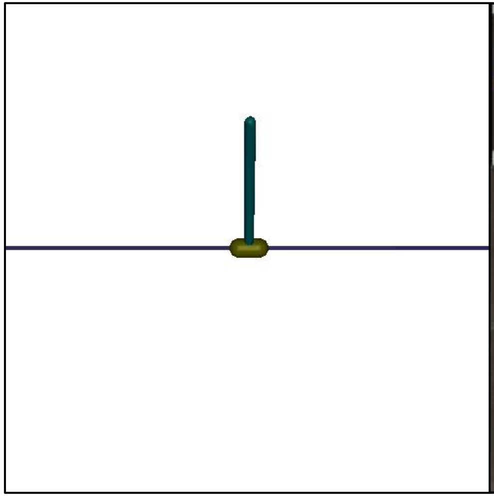
- Approach (Model-Based):
  - Pretraining: probabilistic models capture state transition uncertainty[1]



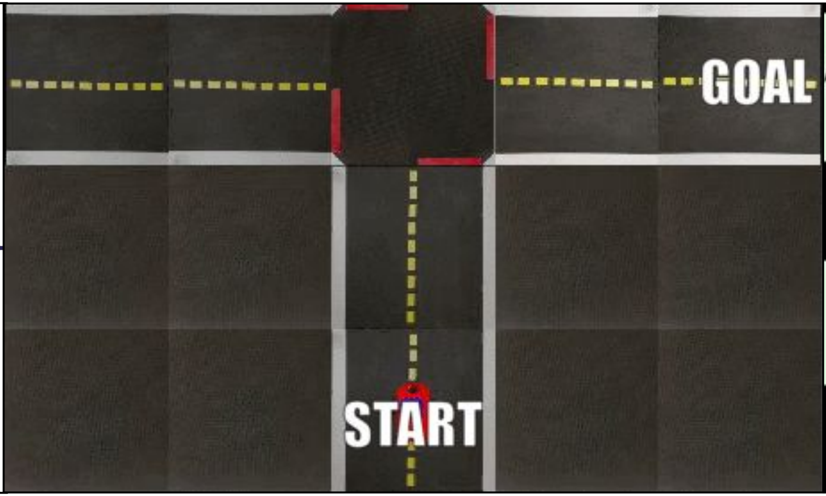
  - Adaptation: utilize uncertainty to safely adapt to new environment (planning cost function modification)

[1]PETS (Chua et al., 2018)

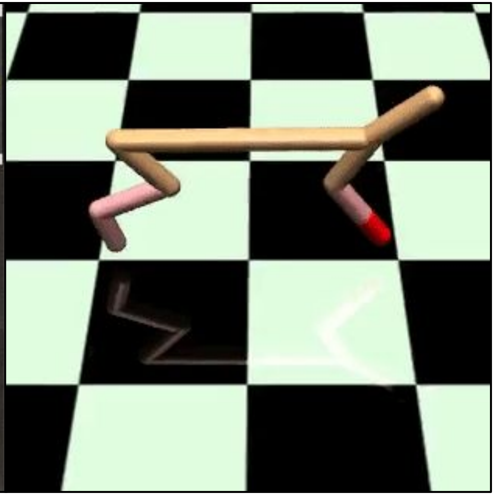
Environments Tested



Cartpole
(varying pole lengths)



Duckietown
(varying car width)

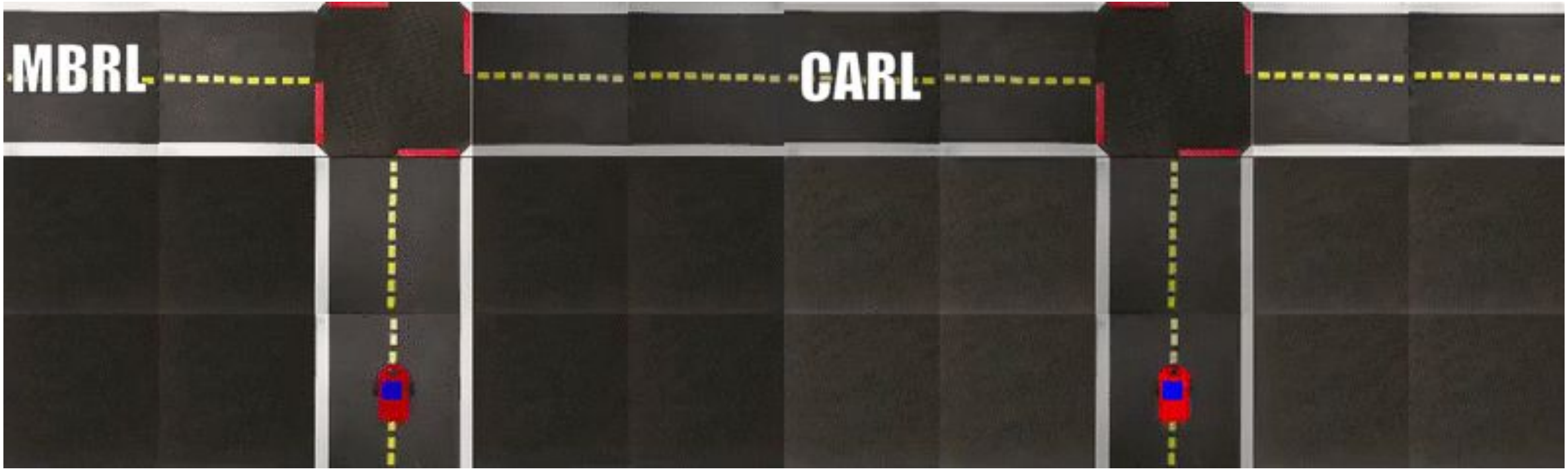


Half Cheetah
(varying disabled joint)

[1]Duckietown (Chevalier-Boisvert et al., 2018)

Results (Cartpole)

Results (Duckietown Driving)



Results (Half-Cheetah)

Short Summary

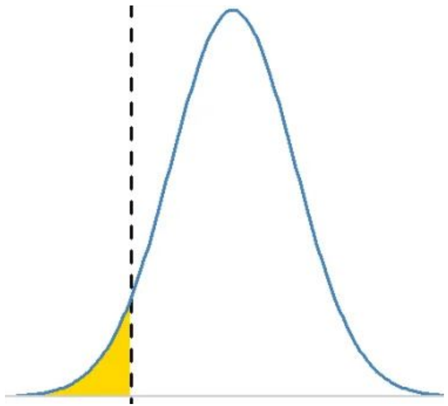
- Capture environment risk with prior experience
    - Probabilistic dynamics models
- Plan with risk in mind for safety-critical adaptation

Outline

- Discussion of related works
- Detailed discussion of CARL methodology
- Further analysis of results
  - Comparison to other methods
  - Average reward, # of catastrophic events

Related Work

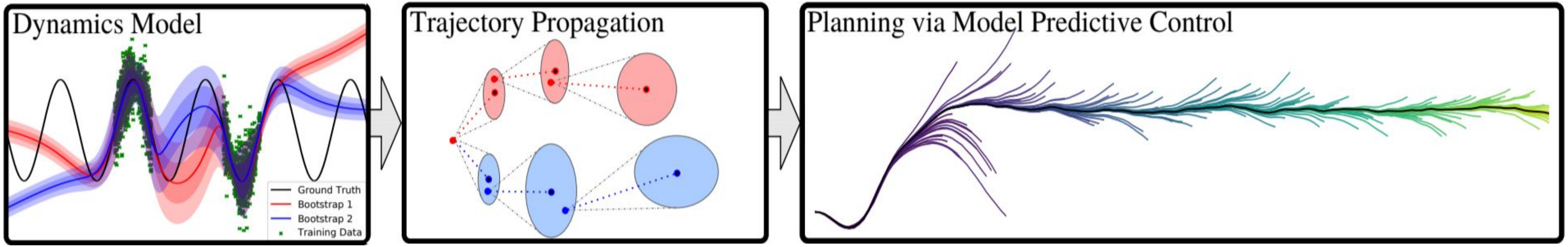
- Risk-Averse RL
  - Conditional Value at Risk



- Model Based RL for Safety
  - Explicit safety constraints
- Capturing Uncertainty
  - Meta-learning

[illegible]
Nagabandi et al (2018); Sæmundsson et al (2018); Finn et al (2017);
[illegible]

Model-based RL Preliminaries: PETS



Ensemble of Probabilistic Dynamics Models

Trajectory sampling for candidate action selection

Sequence with highest action score is executed

Action Score

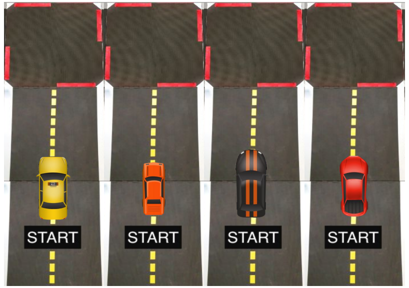
$$R(A) = \sum_i r_i / N$$

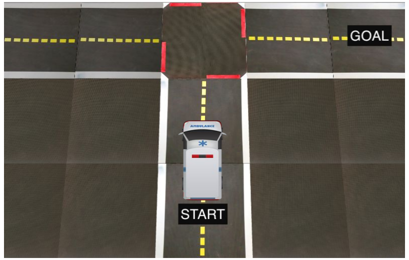
Over predicted trajectories with actions A

Reward for i'th trajectory

CARL for Safety Critical Adaptation

- PETS: Ensemble captures stochasticity in single environment
  - CARL: Captures uncertainty induced by variations across environments
- Pretraining: Train PETS
  - Randomly sample domain
  - Dynamics model captures uncertainty about state transitions, reward, and risk
- Adaptation: Unseen domain
  - Risk averse action selection





Risk Averse Action Selection

Case 1: Low Reward Risk-Aversion, *CARL (Reward)*

$$R(A) = \sum_i r_i / N \quad \Rightarrow \quad R_\gamma(A) = \sum_{i \in S_\gamma} r_i / |S_\gamma|$$

- Select actions that minimize worst-case outcomes
- S_γ: worst γ percentile of predicted trajectories

Risk Averse Action Selection

Case 2: Catastrophic State Risk-Aversion, *CARL (State)*

- Avoid catastrophic states directly
- Build state safety cost, *g(A)*
- Maximize: $R_\lambda(A) = R(A) - \lambda g(A)$
- Lagrangian relaxation of constraint minimizing probability of encountering states in a catastrophic set

Risk Averse Action Selection

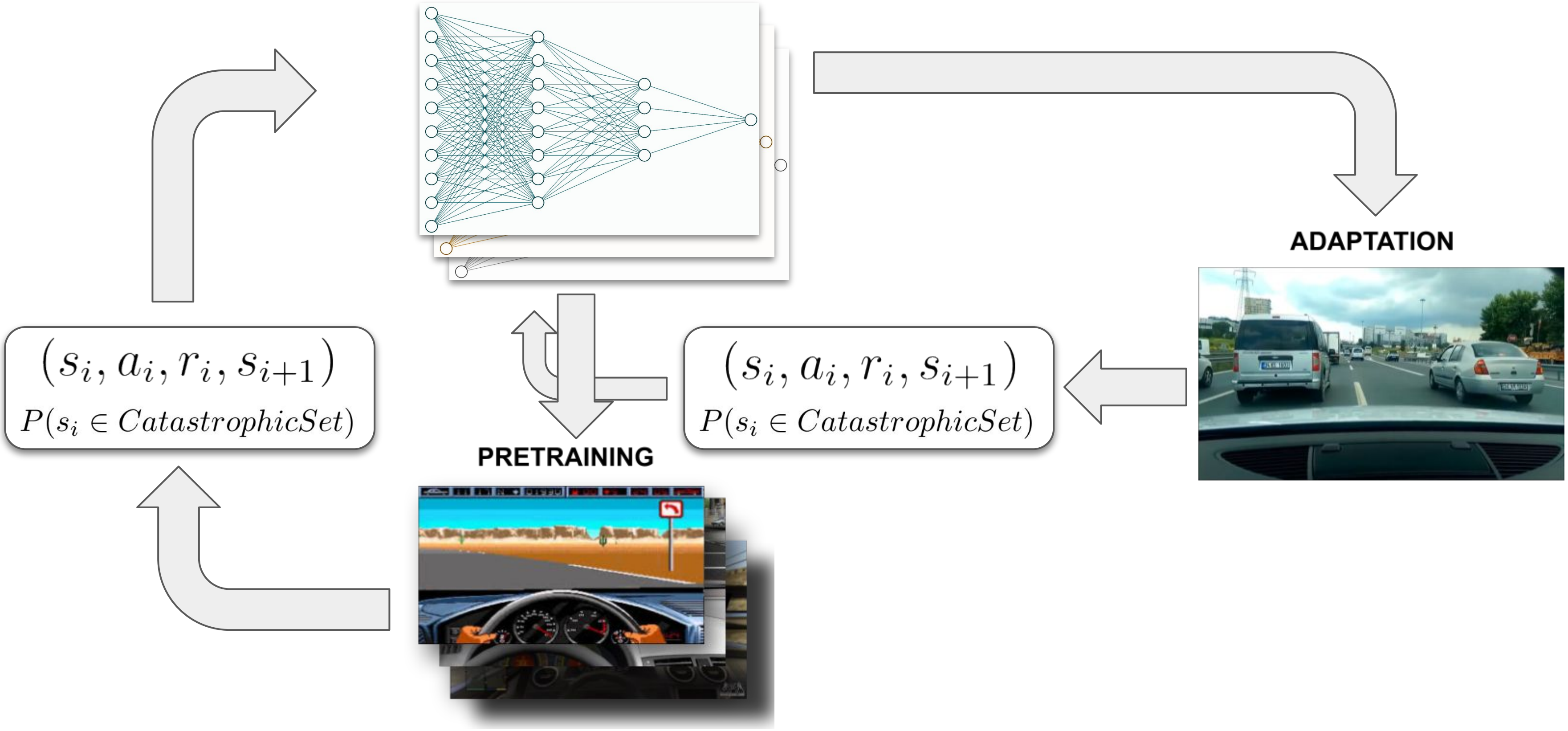
Case 2: Catastrophic State Risk-Aversion

$$g(A) = \sum_{i=1}^{H} P\left(s_i \in \text{ CatastrophicSet }\right)$$

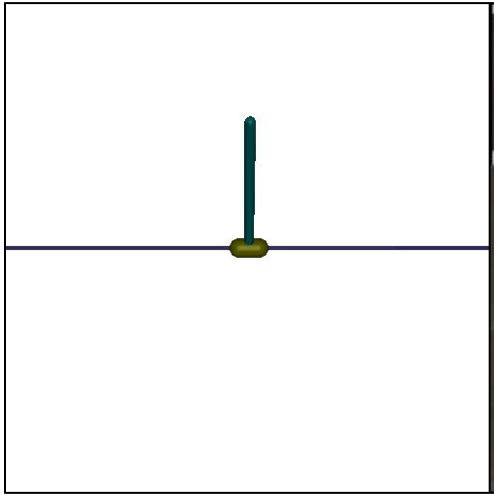
$$P\left(s_i \in \text{ CatastrophicSet }\right) = \frac{\sum_{j=1}^{|E|} 1[c_j(s_{i-1}, a_{i-1}) > \beta]}{|E|}$$

$$R_\lambda(A) = R(A) - \lambda g(A)$$

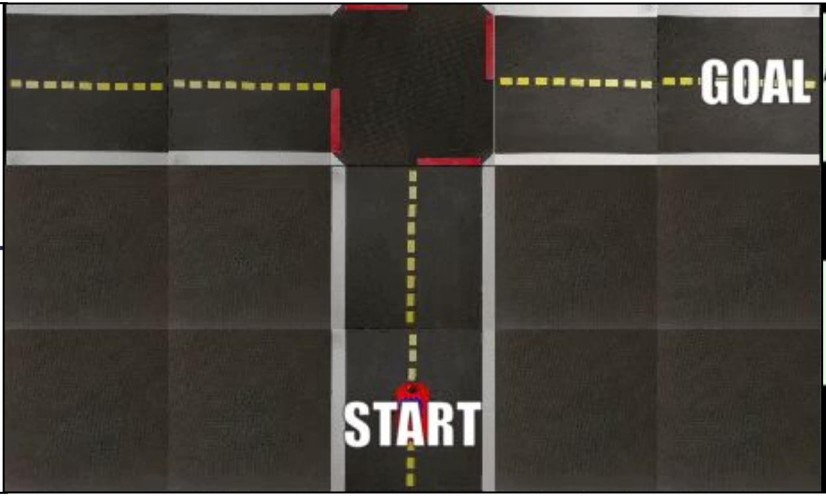
CARL System Overview



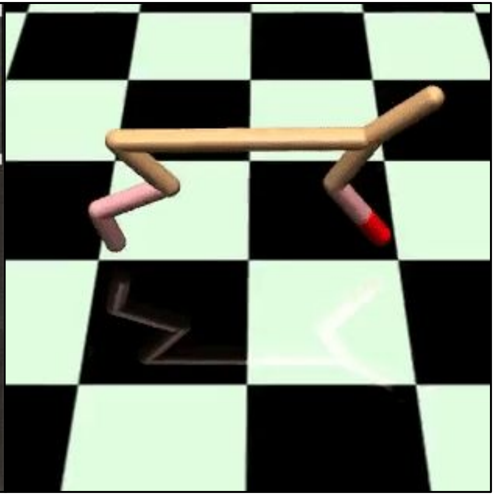
Environments Tested



Cartpole
(varying pole lengths)



Duckietown
(varying car width)



Half Cheetah
(varying disabled joint)

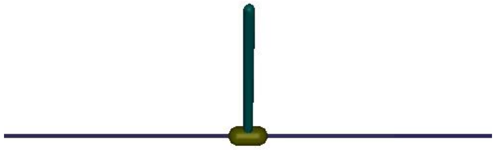
[1]Duckietown (Chevalier-Boisvert et al., 2018)

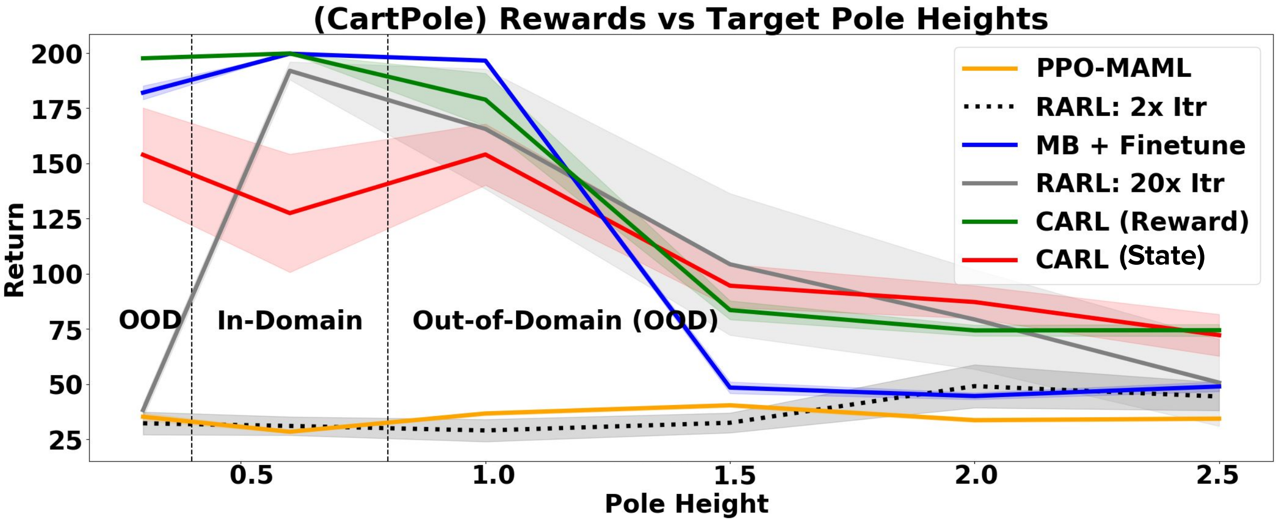
Experiment Setup

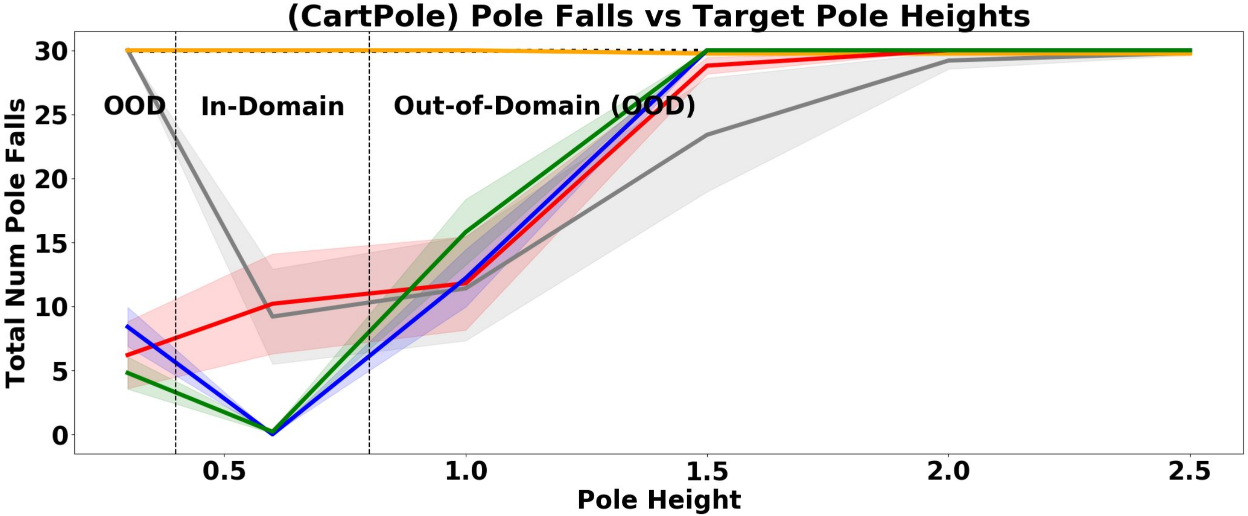
- MB + Finetune: PETS, finetune on test environment
- RARL: Robust Adversarial Reinforcement Learning[1]
- PPO-MAML: Model-Agnostic Meta Learning[2]
- CARL (Reward): Reward-based CARL
- CARL (State): State-based CARL

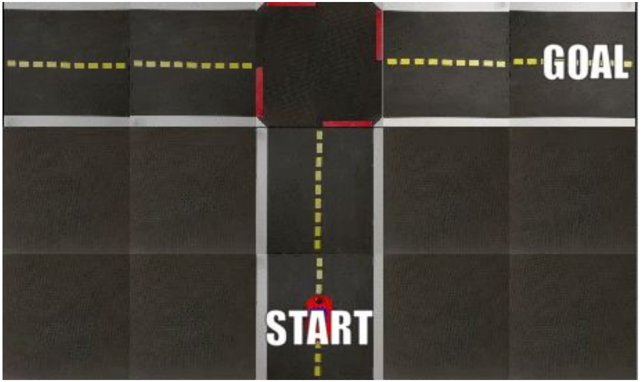
[1](Pinto et al., 2017)

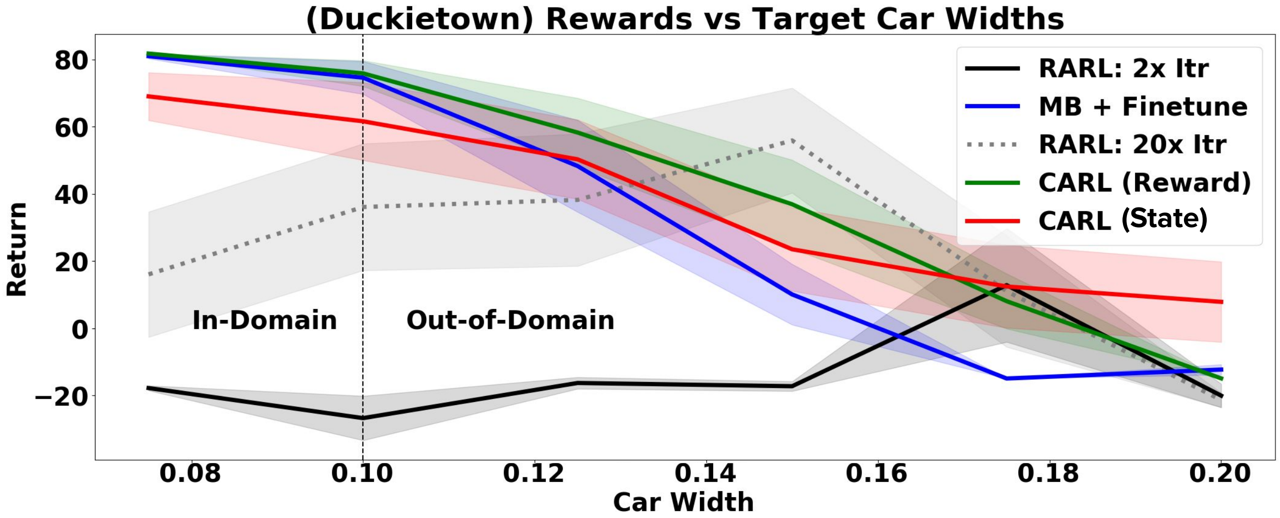
[2](Finn et al., 2017)

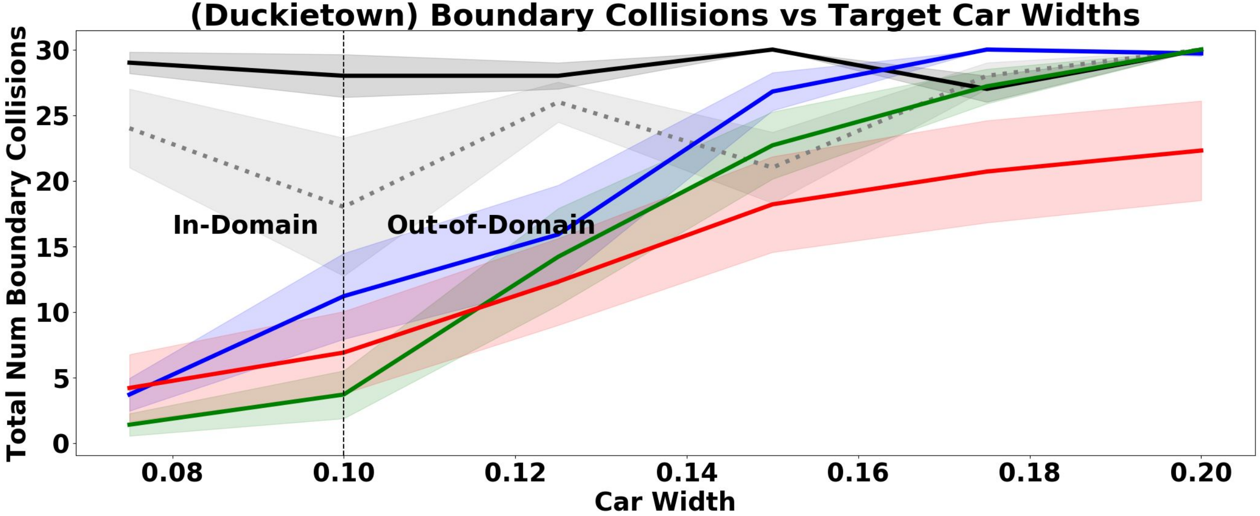


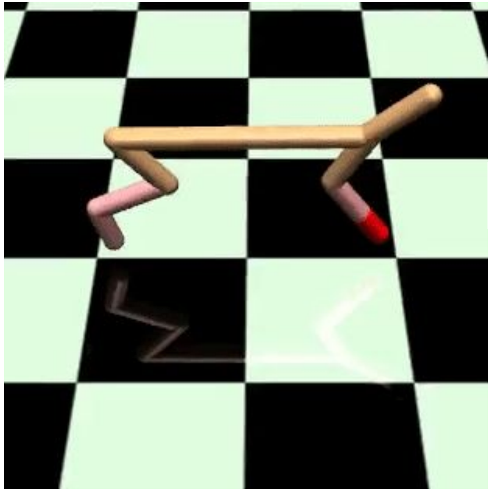

(CartPole) Rewards vs Target Pole Heights
PPO-MAML
RARL: 2x Itr
MB + Finetune
RARL: 20x Itr
CARL (Reward)
CARL (State)
Return
200
175
150
125
100
75
50
25
OOD
In-Domain
Out-of-Domain (OOD)
0.5
1.0
1.5
2.0
2.5
Pole Height

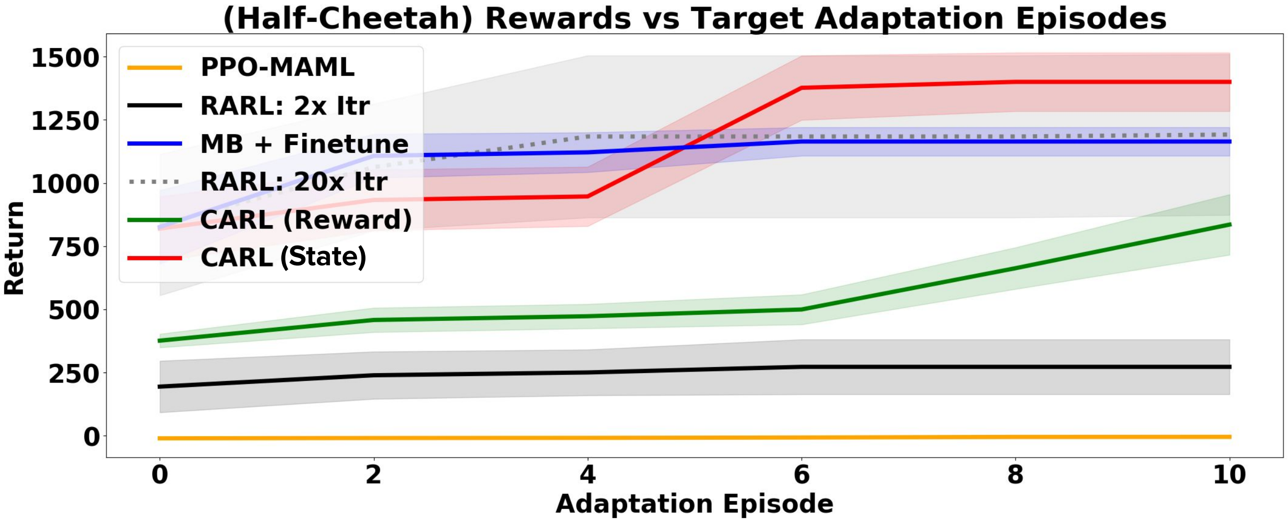

(CartPole) Pole Falls vs Target Pole Heights
Total Num Pole Falls
30
25
20
15
10
5
0
OOD
In-Domain
Out-of-Domain (OOD)
0.5
1.0
1.5
2.0
2.5
Pole Height

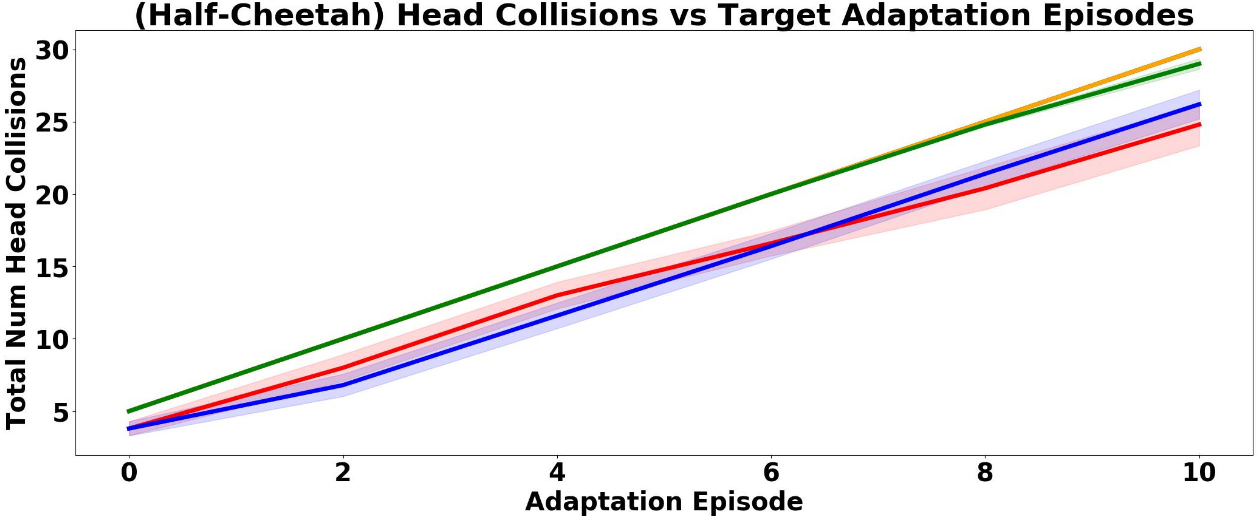

GOAL
START


(Duckietown) Rewards vs Target Car Widths
Return
Car Width
In-Domain
Out-of-Domain
RARL: 2x Itr
MB + Finetune
RARL: 20x Itr
CARL (Reward)
CARL (State)
80
60
40
20
0
−20
0.08
0.10
0.12
0.14
0.16
0.18
0.20


(Duckietown) Boundary Collisions vs Target Car Widths
Total Num Boundary Collisions
Car Width
In-Domain
Out-of-Domain
30
25
20
15
10
5
0
0.08
0.10
0.12
0.14
0.16
0.18
0.20




(Half-Cheetah) Rewards vs Target Adaptation Episodes
PPO-MAML
RARL: 2x Itr
MB + Finetune
RARL: 20x Itr
CARL (Reward)
CARL (State)
Return
Adaptation Episode
0
250
500
750
1000
1250
1500
0
2
4
6
8
10


(Half-Cheetah) Head Collisions vs Target Adaptation Episodes
Total Num Head Collisions
Adaptation Episode
5
10
15
20
25
30
0
2
4
6
8
10

Summary

- Safety-Critical Adaptation (SCA)
  - Train on sandbox environments, adapt to safety-critical environments
- CARL and CARL (Reward)
  - Capture source uncertainty, perform risk-averse planning

Thank you!