DOPT: D-learning with Off-Policy Target toward Sample Efficiency and Fast Convergence Control

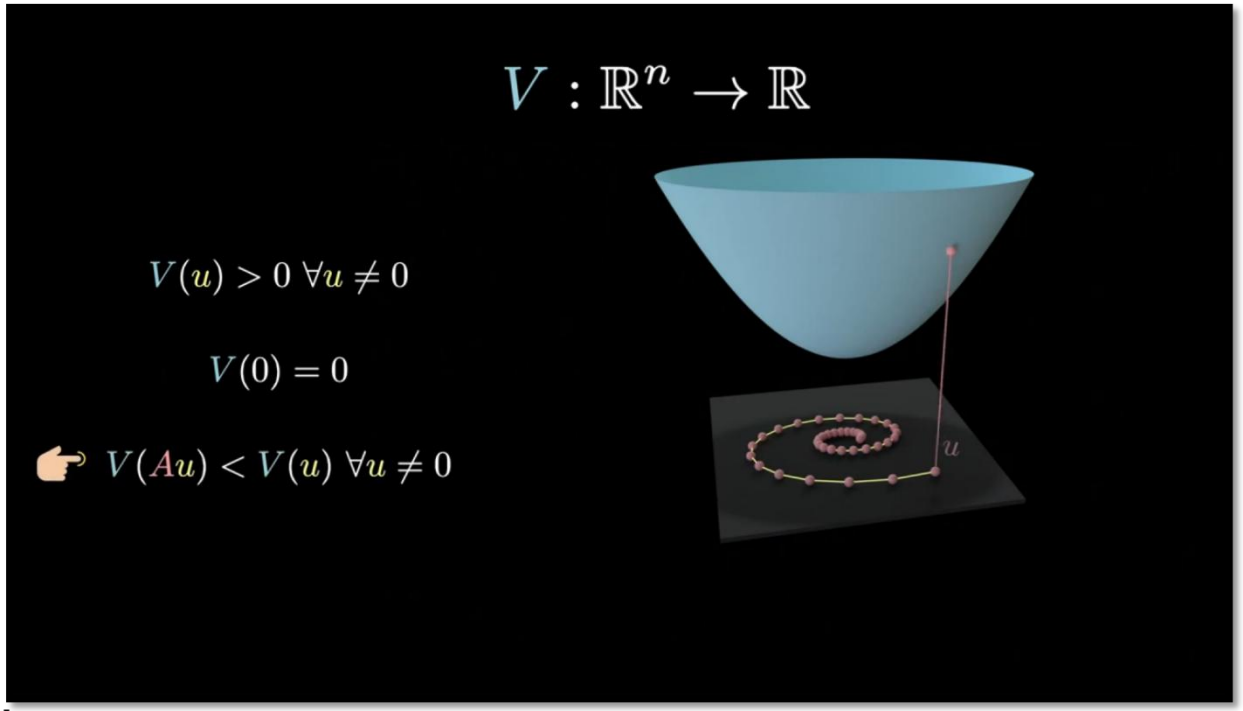
Zhaolong Shen and Quan Quan*

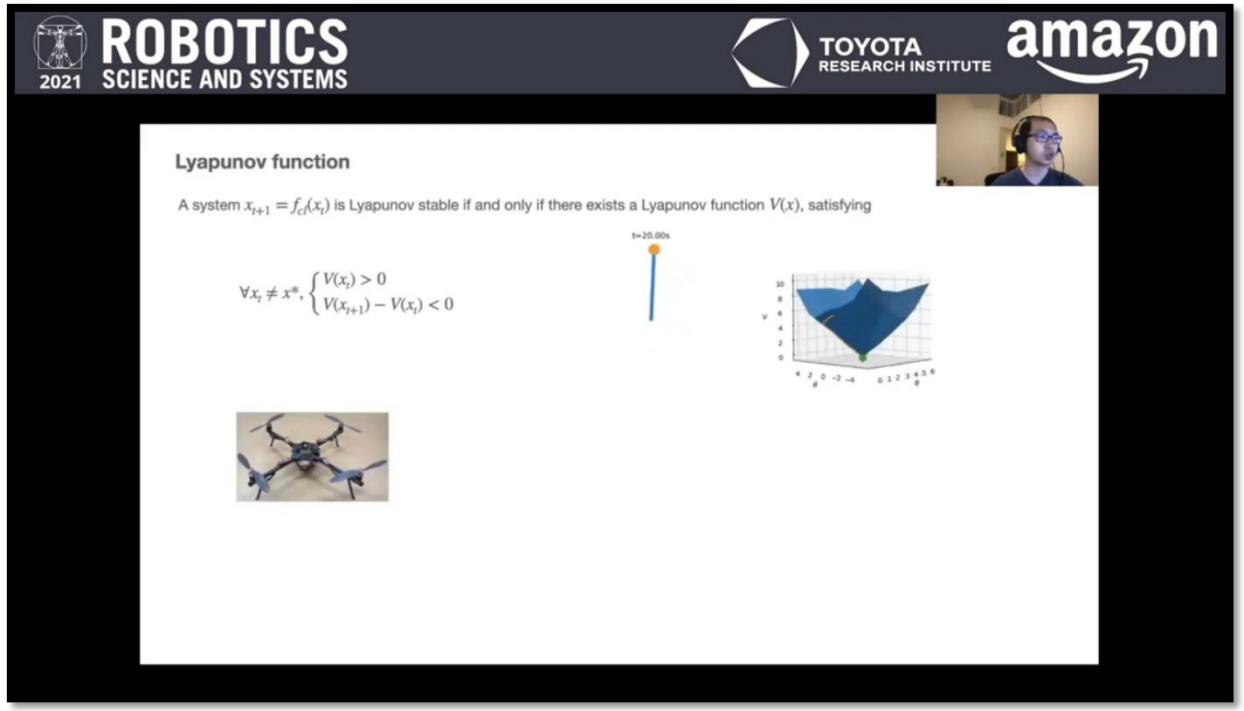
BEIHANG
UNIVERSITY

RELIABLE FLIGHT
CONTROL GROUP

Background

Learning-based Lyapunov Control





Hongkai Dai et al. “Lyapunov-stable neural-networkcontrol”. In: Robotics: Science and Systems (RSS)XVII. 2021.

Pros:

- Provide **RoA estimation**.
- Provide **stability and safety guarantee**.

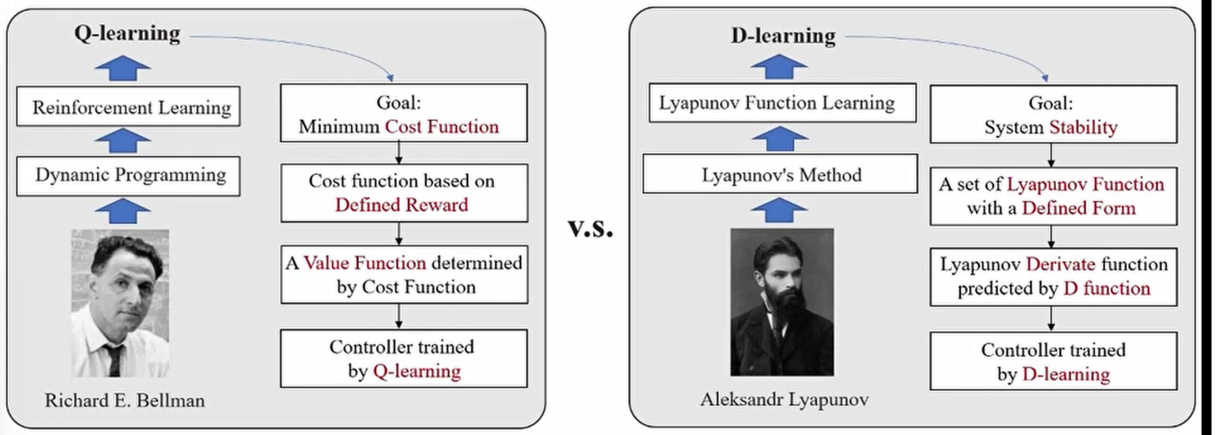
Cons:

- **Require system information**.
- Lacks **sample efficiency**.
- No improvement in **convergence speed**.

Background

D-learning based Control

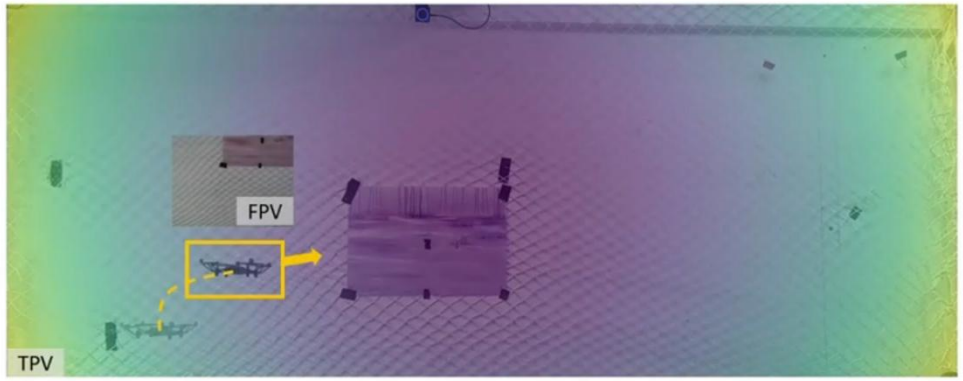
What is D-learning?



D-learning is a Lyapunov Function Learning method without requiring any knowledge of the system dynamics, which parallels to Q-learning.

Real Flight Experiments

- Controller Trained by Control with Patterns Based on D-learning



TPV: Third-Person View FPV: First-Person View

At the initial position, based on the feedback of the feature $\mathbf{s}$ extracted from the current image $\mathbf{I}$ and the desired feature $\mathbf{s}^*$ from the desired image $\mathbf{I}^*$, the controller outputs the velocity $\mathbf{v}$ that makes the Lyapunov function decrease fast.

Quan Quan, Kai–Yuan Cai, and Chenyu Wang. “Control with patterns: a D-learning method”. In: Conference on Robot Learning (CoRL) Accepted. PMLR.2024.

Pros:

- Provide **stability guarantee**.
- Optimize toward **fast convergence**.
- **Model-free.**

Cons:

- **Complex Optimization**.
- Lacks **sample efficiency**.
- **Instable** during Training.

Background

Two Parallel Evolutions

RL Community

DDPG

Q-Learning

Reinforcement Learning

Dynamic Programming

Richard E. Bellman

- Optimization goal is **maximizing rewards**.
- Controller trained by **Q-function *Q(s,a)***.
- Using **historical data** to achieve **sample efficiency**.
- Using **target networks** to **smooth the training process**.

v.s.

Our Method

DOPT

D-Learning

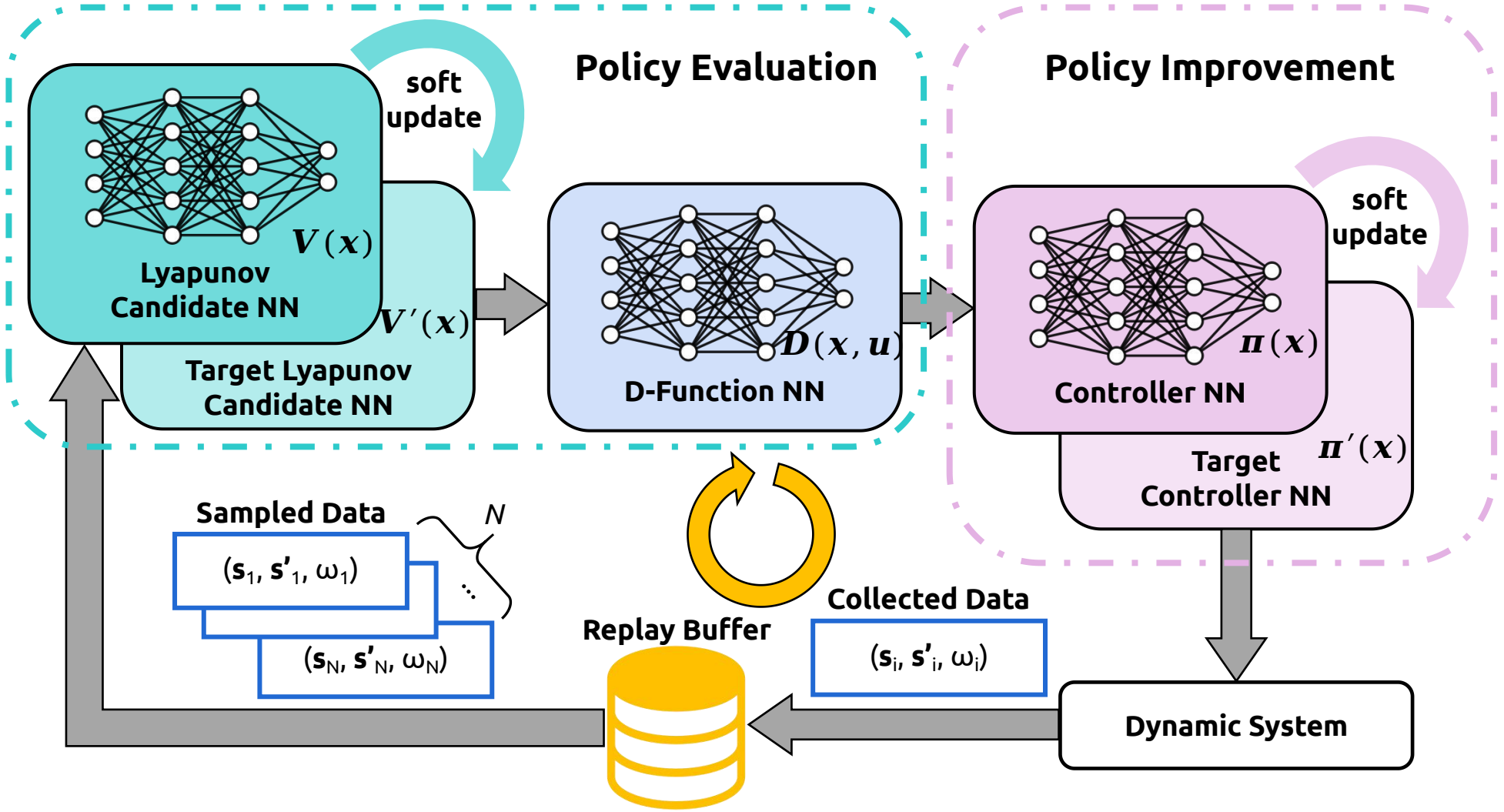
Learning-based Lyaounov

Lyapunov's Theory

Aleksandr Lyapunov

- Optimization goal is **stabilizing the system**.
- Leanring **Lyapunov function *V(x)* and the derivative *D(x,u)***.
- Controller trained by **D-function**.
- Using **historical data** to achieve **sample efficiency**.
- Using **target networks** to **smooth the training process**.

What is DOPT?



Initialize a stabilizing $\pi'(x)$ and **Collect** data

Sample data from replay buffer and **train** the $V(x)$

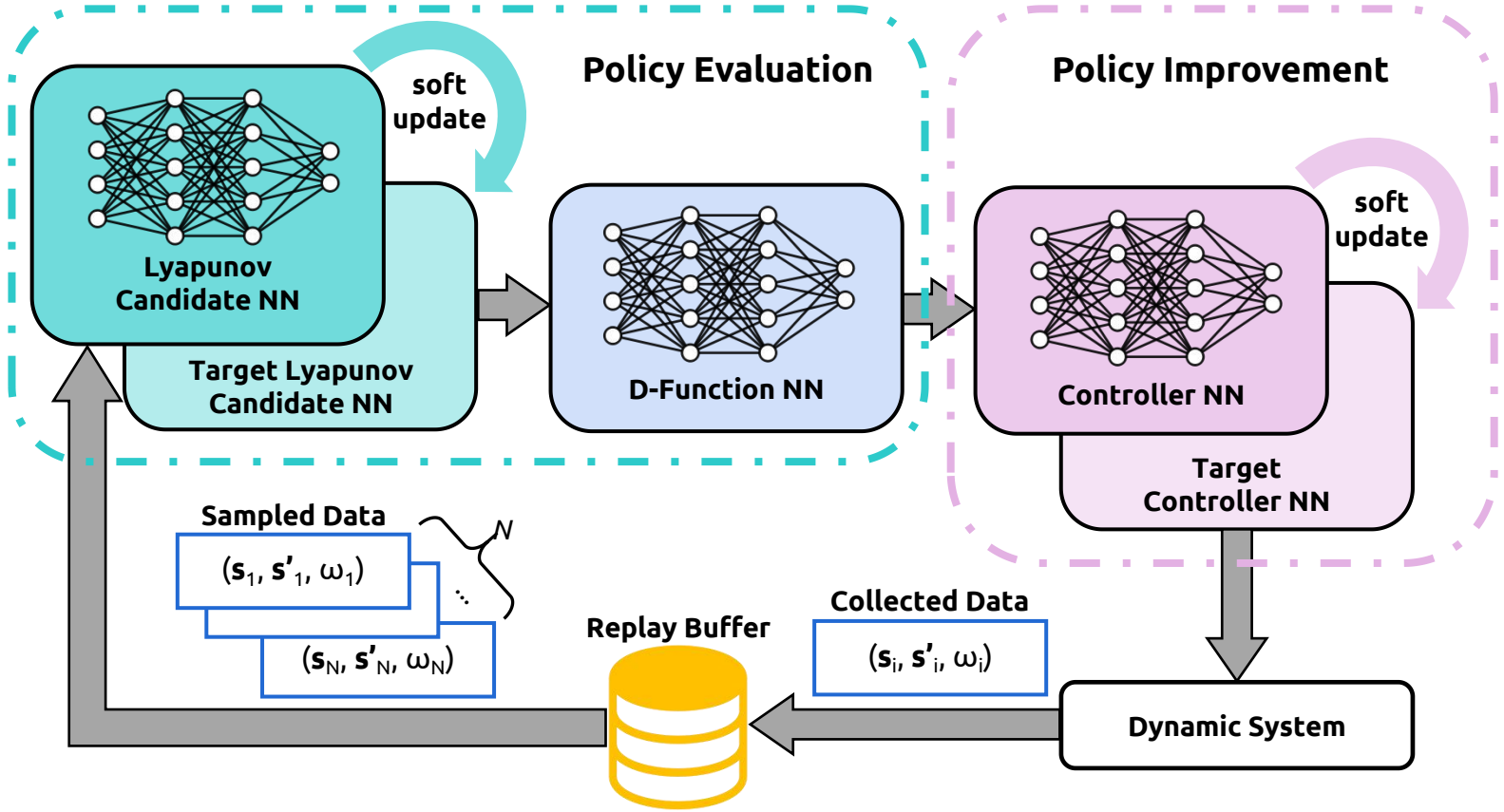
Soft update the target NN $V'(x)$

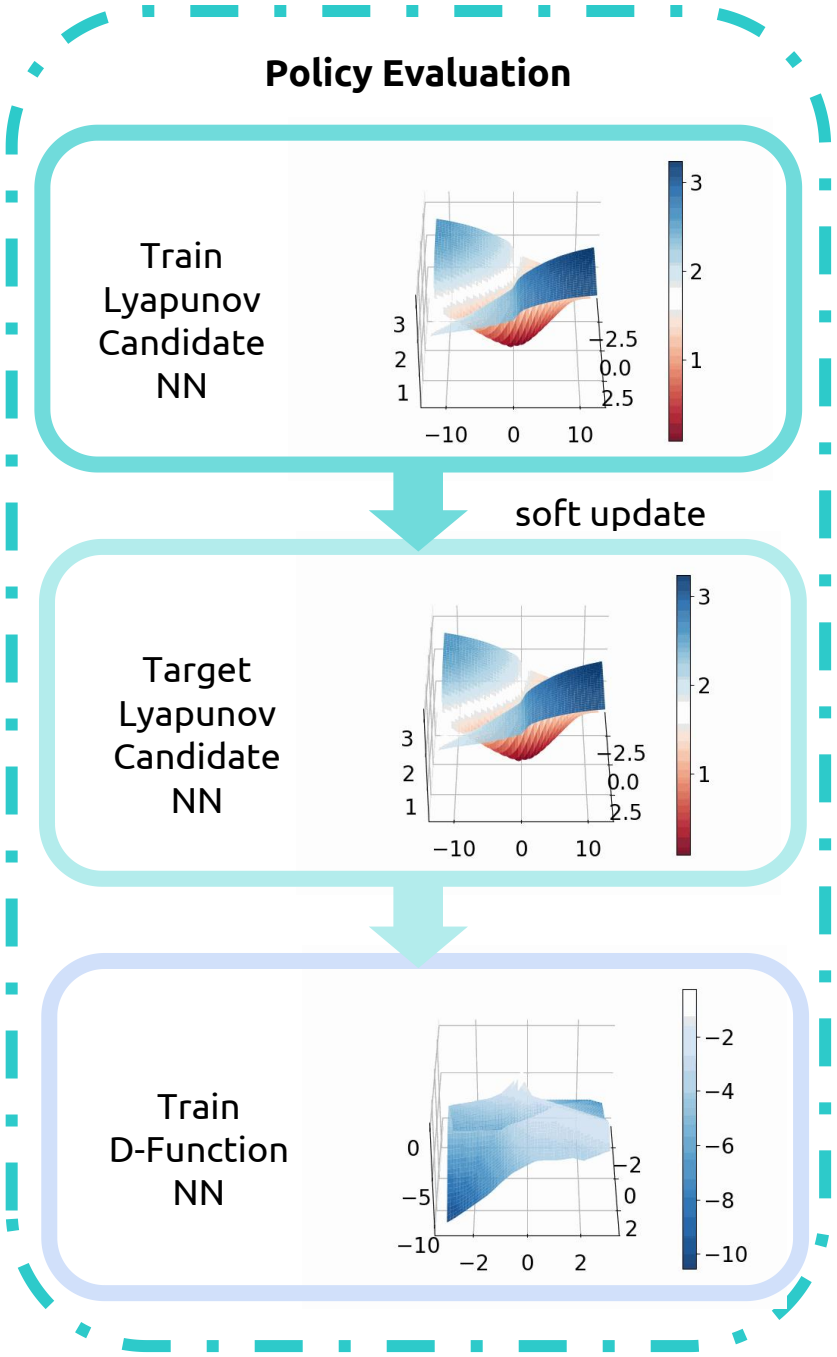
Train the $D(x,u)$ using the derivative of the $V'(x)$

Improve the $\pi(x)$ by decreasing the $D(x,\pi(x))$

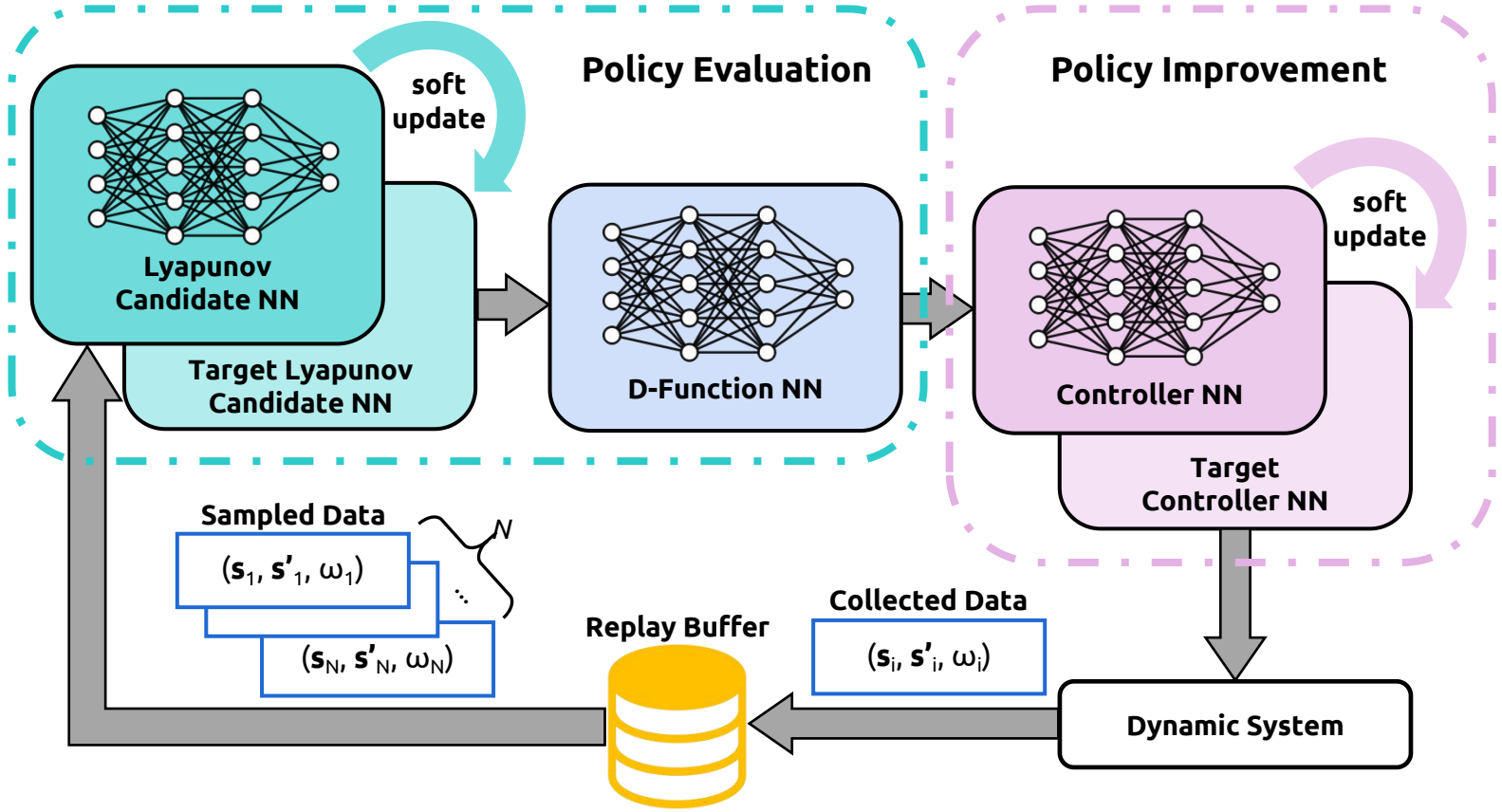
Soft update the target NN $\pi'(x)$

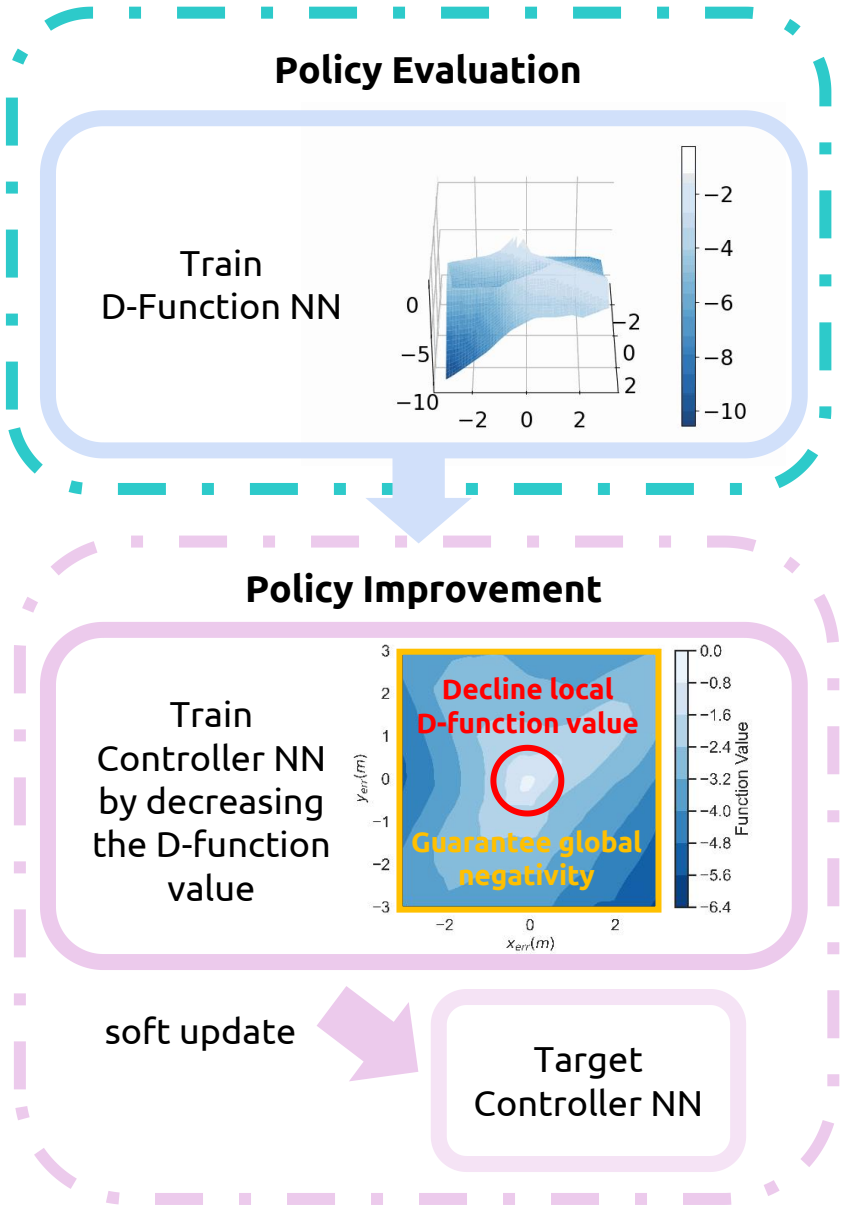
What is DOPT?


Policy Evaluation
soft update
Lyapunov Candidate NN
Target Lyapunov Candidate NN
D-Function NN
Policy Improvement
soft update
Controller NN
Target Controller NN
Sampled Data
(s_1, s'_1, ω_1)
(s_N, s'_N, ω_N)
N
Replay Buffer
Collected Data
(s_i, s'_i, ω_i)
Dynamic System


Policy Evaluation
Train Lyapunov Candidate NN
soft update
Target Lyapunov Candidate NN
Train D-Function NN

What is DOPT?

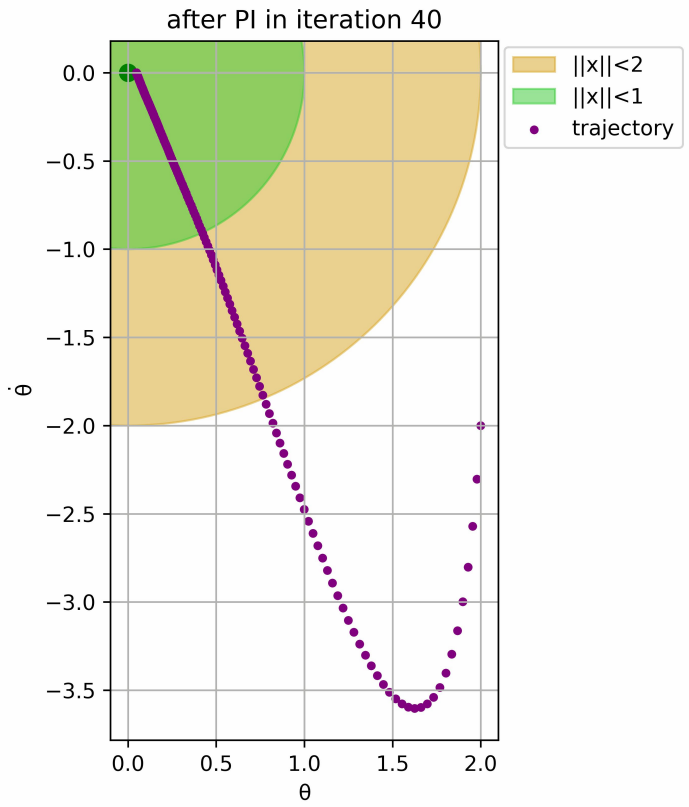

Policy Evaluation
soft update
Lyapunov Candidate NN
Target Lyapunov Candidate NN
D-Function NN
Policy Improvement
soft update
Controller NN
Target Controller NN
Sampled Data
(s_1, s'_1, ω_1)
(s_N, s'_N, ω_N)
N
Replay Buffer
Collected Data
(s_i, s'_i, ω_i)
Dynamic System


Policy Evaluation
Train D-Function NN
Policy Improvement
Train Controller NN by decreasing the D-function value
Decline local D-function value
Guarantee global negativity
Function Value
soft update
Target Controller NN

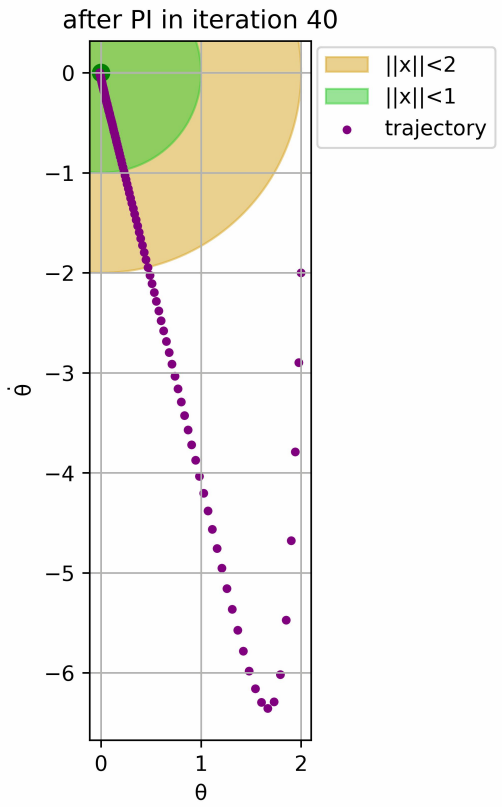
Experiments: Inverted Pendulum

Experiments: Inverted Pendulum

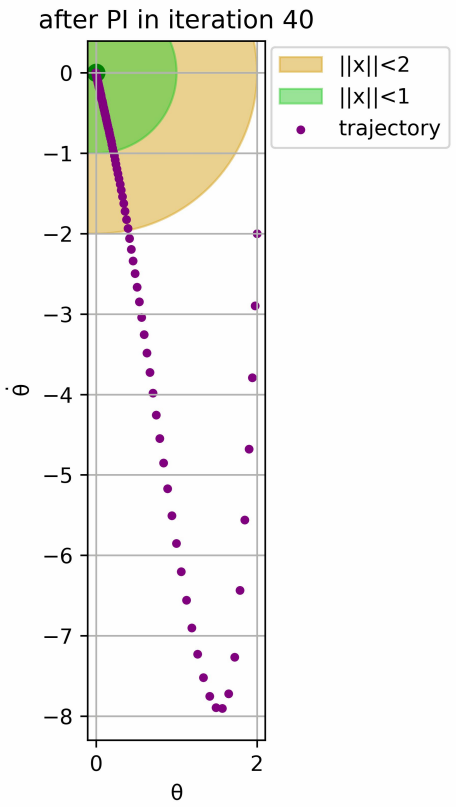
DDPG



D-learning

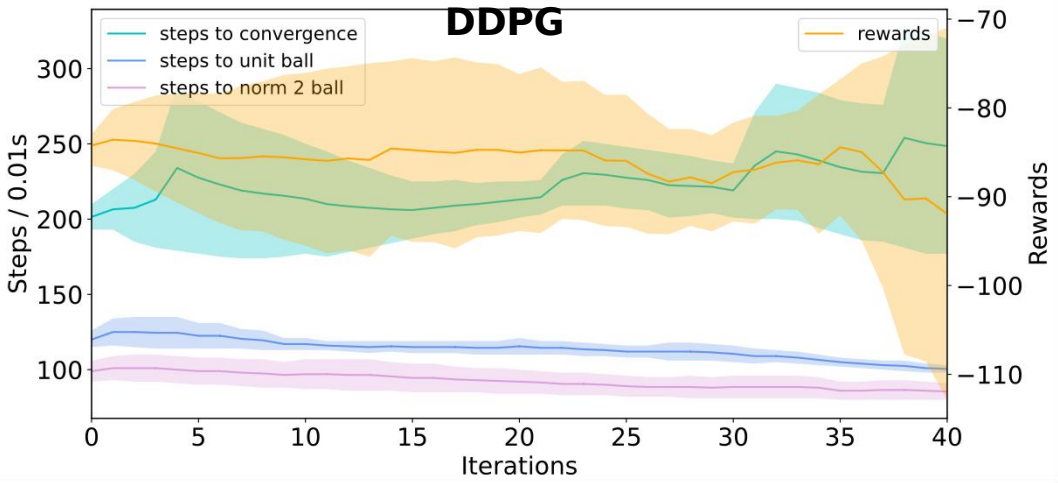


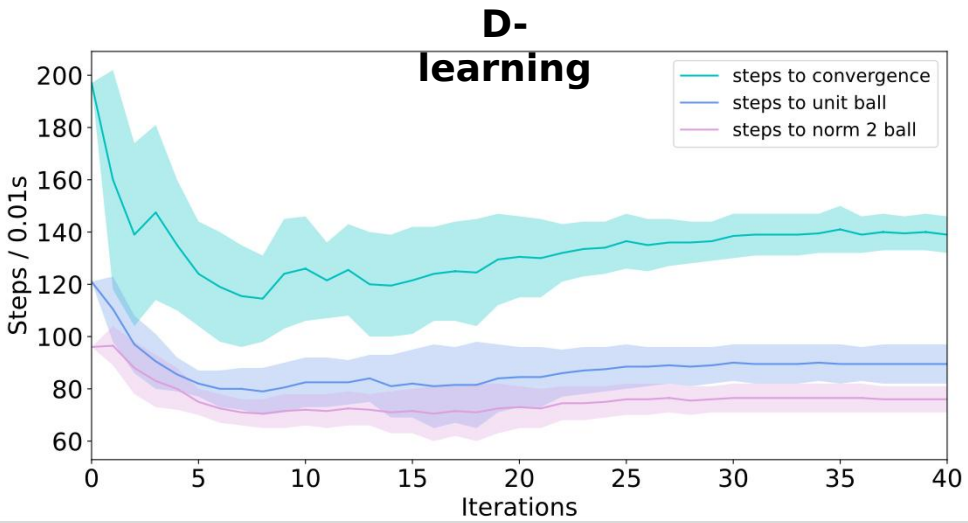
DOPT

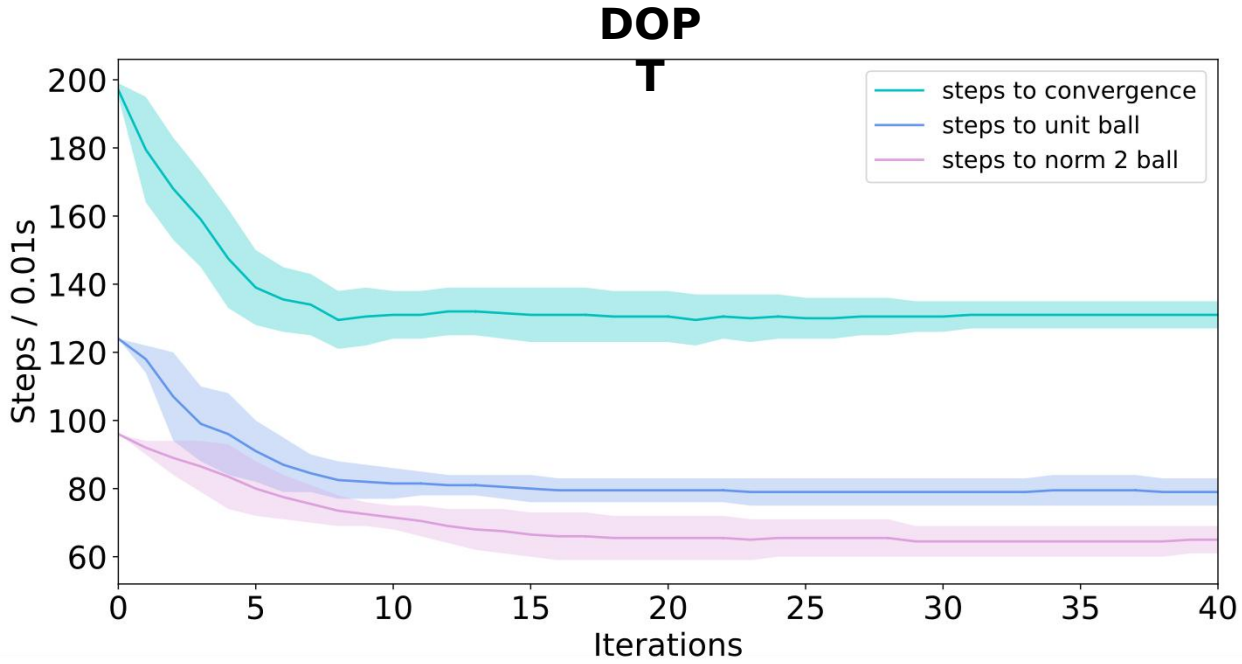


Training Process

Experiments: Inverted Pendulum

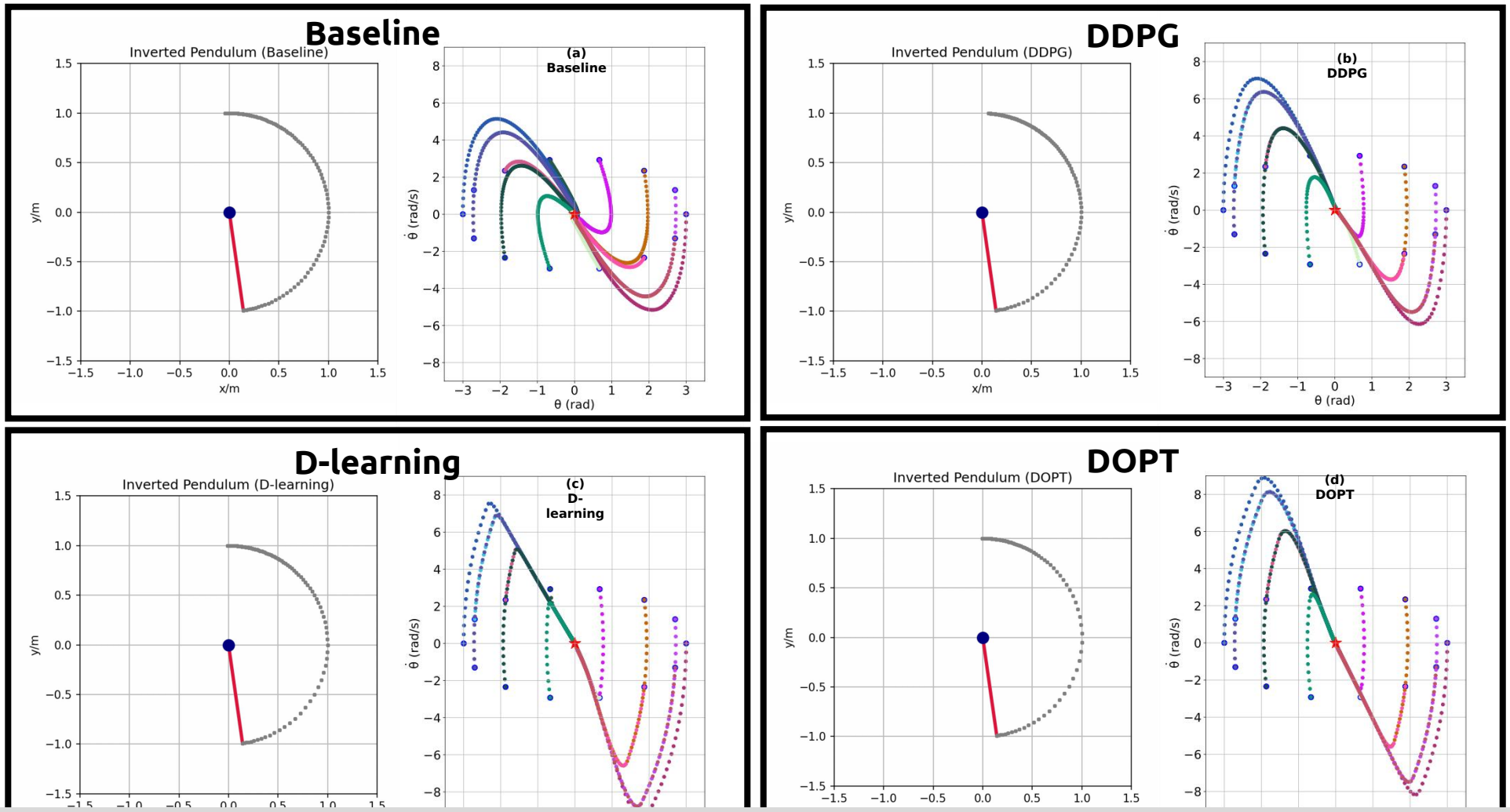






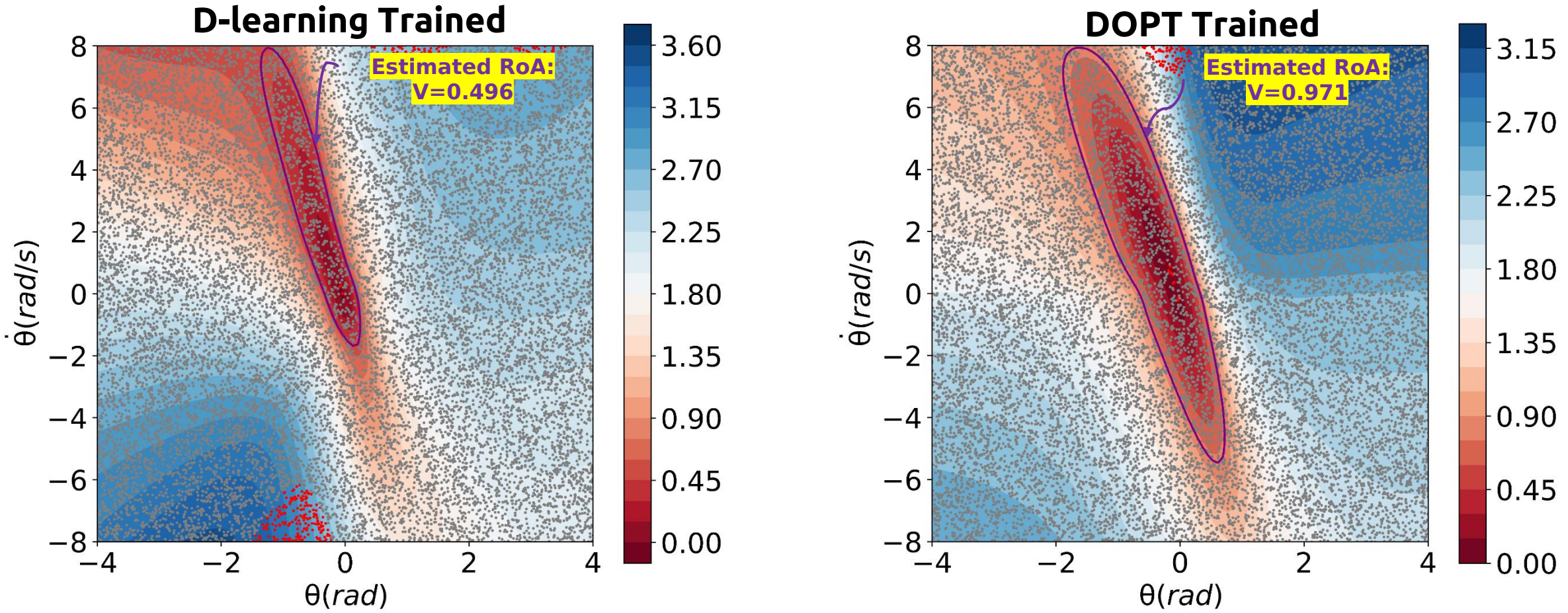
DOPT presents **steady and smooth training process**!
DOPT and D-learning can **efficiently accelerate the convergence**!

Experiments: Inverted Pendulum



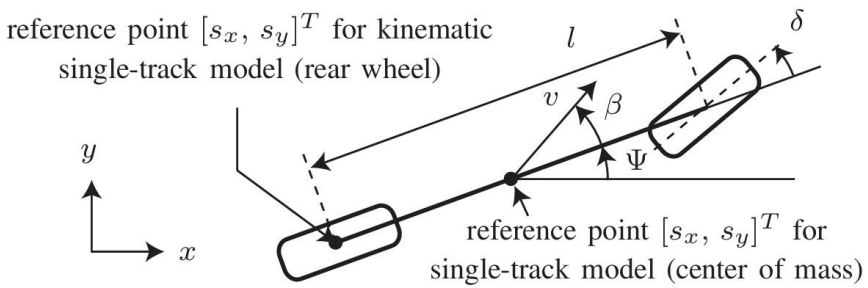
DOPT-trained controller achieves **fastest convergence** and **zero steady-state error**!

Experiments: Inverted Pendulum

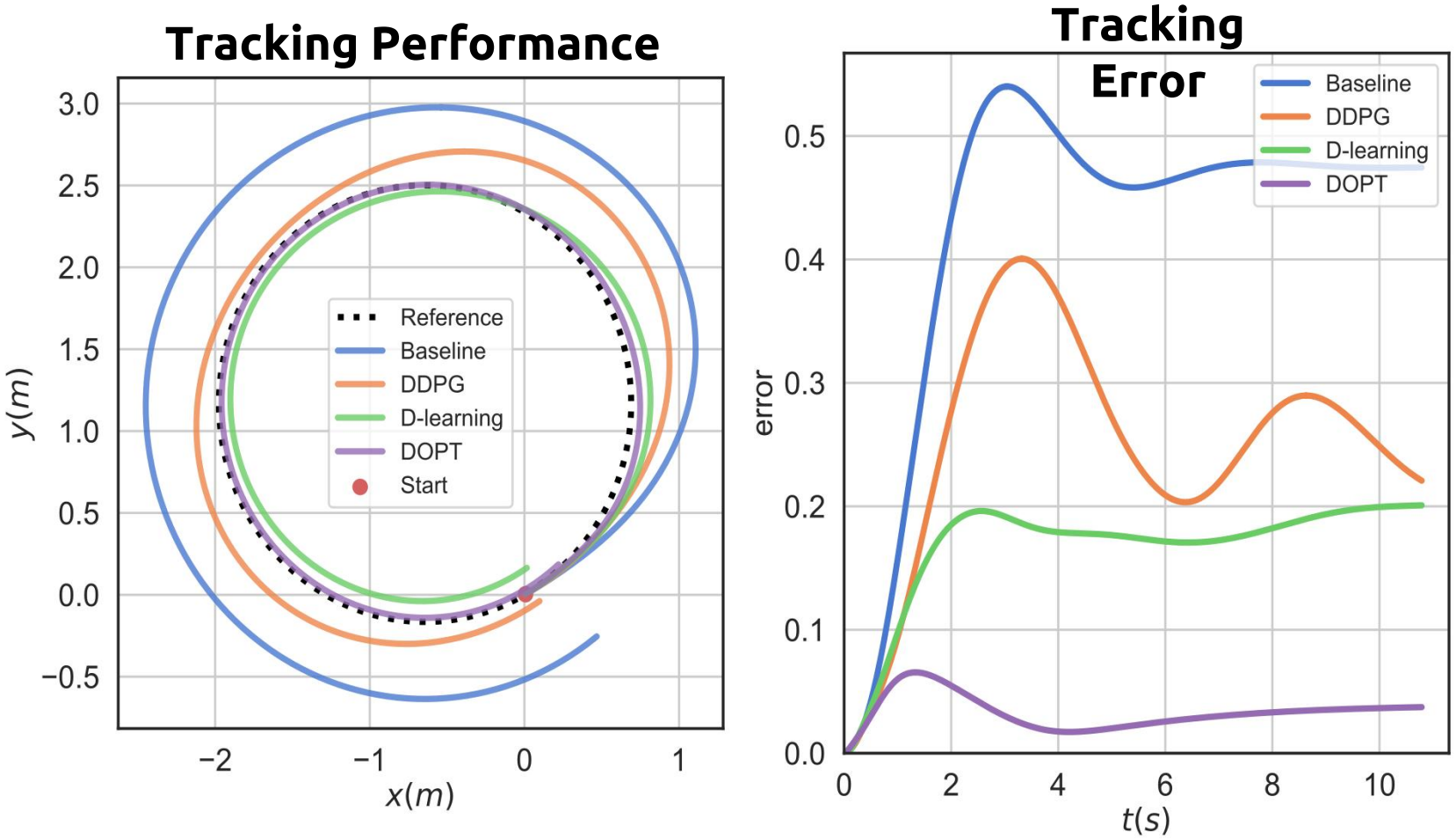


Both Lyapunov candidates satisfy **almost Lyapunov conditions**!
DOPT can capture **more expressive Lyapunov candidate**,
and provide **expanded estimated RoA**!

Experiments: Single-Track Car

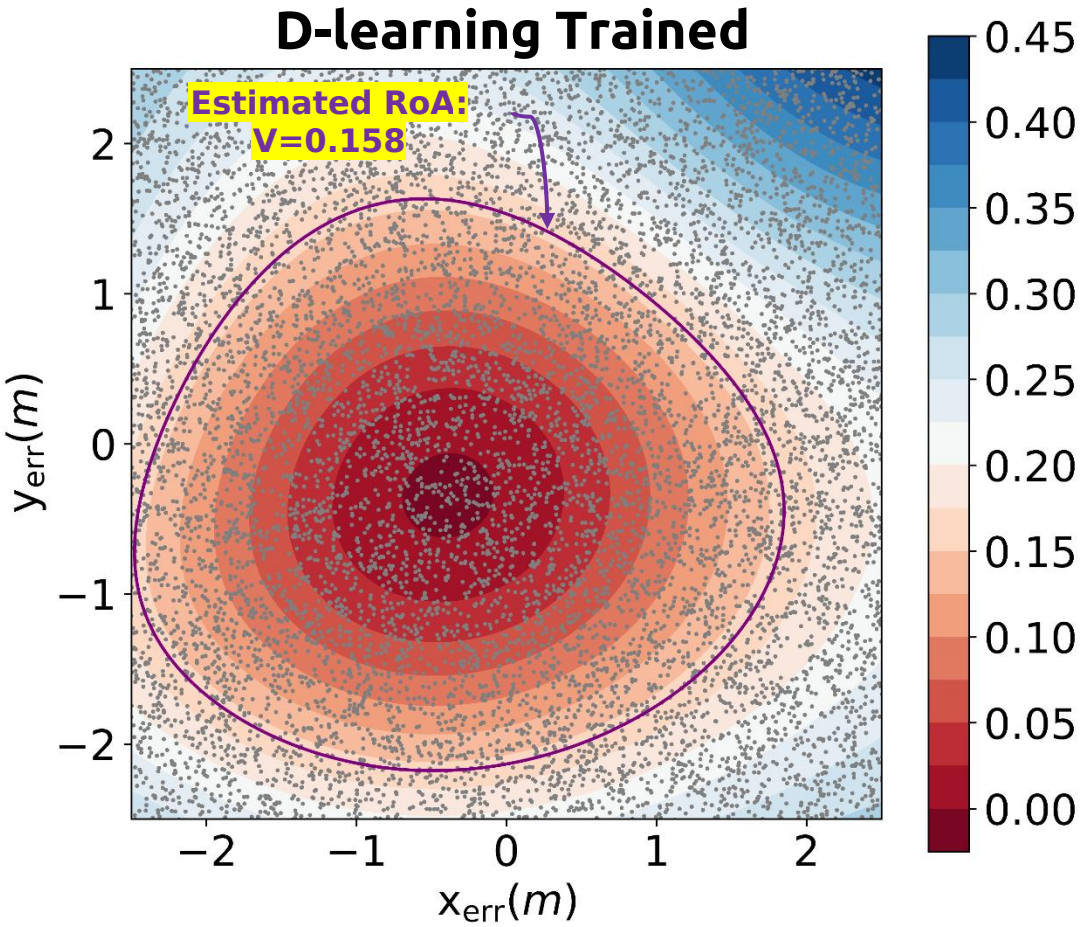


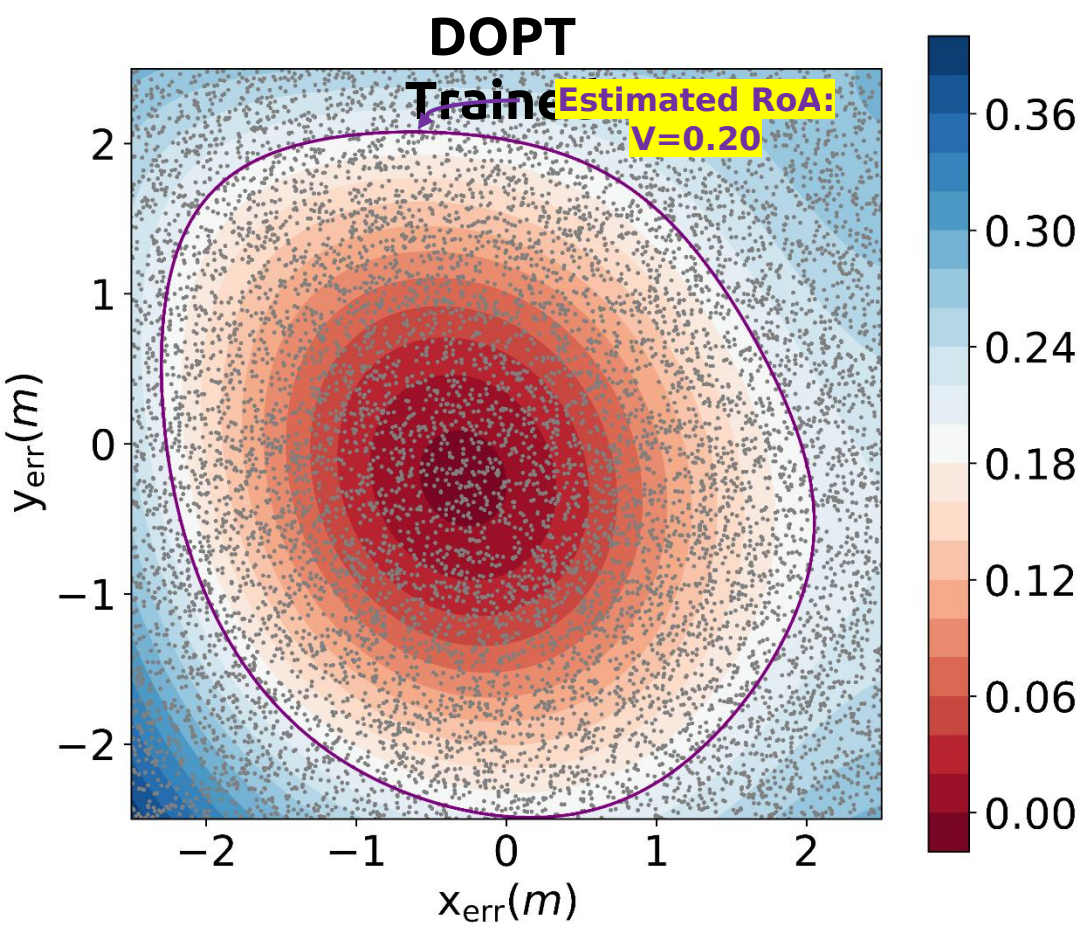
Experiments: Single-Track Car



DOPT-trained controller achieves the **lowest steady-state error** tracking on the slippery surface!

Experiments: Single-Track Car





Both Lyapunov candidates satisfy **almost Lyapunov conditions**!
DOPT can capture **more expressive Lyapunov candidate**,
and provide **expanded estimated RoA**!

Experiments: Sample Efficiency

TRAINING EFFICIENCY AND DATA UTILIZATION OF D-LEARNING AND DOPT

Approaches	Environment	Data volume (sample/training)	Training time	Iterations	Final PII	Indicator
D-learning	inverted pendulum	2000/2000	31min46s	40	-610.296	144 (convergence steps)
DOPT	inverted pendulum	1000/2000	28min21s	40	-624.529	136
DOPT	inverted pendulum	2000/4000	34min29s	40	-669.143	122
D-learning	single-track car	2000/2000	63min47.9s	20	-98.4	0.152(steady-state error)
DOPT	single-track car	1000/2000	52min1.6s	20	-110.5	0.124
DOPT	single-track car	2000/4000	69min48.9s	20	-113.7	0.085

DOPT demonstrates **superior sample efficiency** and achieves **better training results** under limited and equal sample volumes.

Thanks for Watching!

http://rfly.buaa.edu.cn