UCLA Computer Science

STAR AI RESEARCH LAB UCLA

# Off-Policy Deep Reinforcement Learning with Analogous Disentangled Exploration

**Anji Liu**[1], Yitao Liang[1], Guy Van den Broeck[1]
[1]Computer Science Department, UCLA

# Motivation: Why Reinforcement Learning

Is capable of solving large-scale and complex problems.

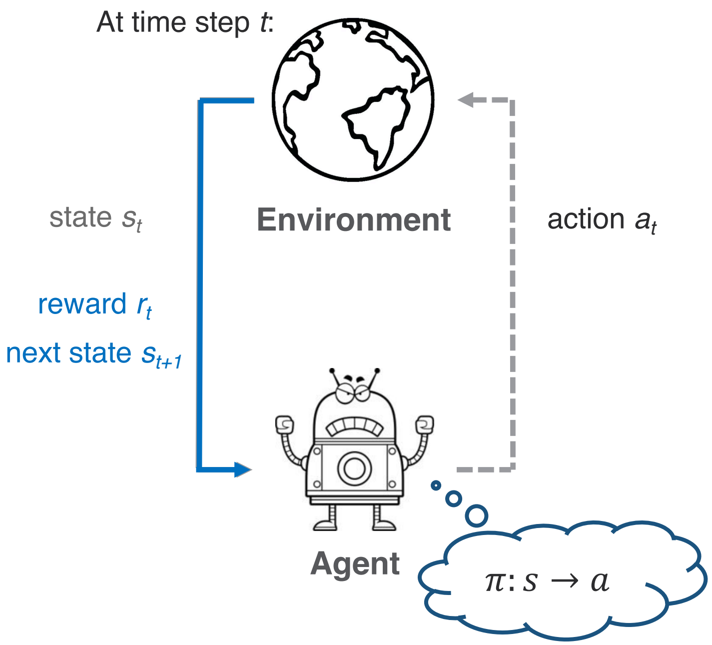
Learning through a trial-and-error process with little supervision.

**Video games**


**Autonomous driving**


**Go**

# Background: Markov Decision Process




## Policy

Determine which action to take given a state.

## Value function

Measures discounted long-term reward

$$Q(s,a) = \mathbb{E}_{a_t \sim \pi, s_{t+1} \sim \mathcal{P}, r_t \sim \mathcal{R}} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a \right]$$

## Goal of RL

Find a policy $\pi$ that maximizes the expected long-term reward.

$$\mathbb{E}_{s \sim \rho_0, a \sim \pi} \left[ Q(s,a) \right]$$

where $\rho_0$ is the initial state distribution.

# The Exploration-Exploitation Tradeoff

**Exploration** ⟷ **Exploitation**

Experiencing new states/rewards to potentially find better policies.

⇩

expressiveness, stochastic

Utilize existing knowledge to behave optimally.

⇩

optimality, deterministic

The necessity of two policies[1]: **target policy** $\pi$ and **behavior policy** $\mu$

[1]For simplicity, sometimes only one policy is explicitly constructed.

# Goals

## What we want to achieve?

Leverage the flexibility to explicitly design two policies to better balance the exploration-exploitation tradeoff.

## How we achieve it?

Analogous Disentangled Actor Critic

Restricting the disentangled behavior policy (**policy co-training**).

Restricting value updates (**critic bounding**).

# Combating the Exploration-Exploitation Dilemma

**Exploration** **Exploitation**

### Expressive behavior policies

e.g. maximum entropy RL objective

$$\mu(a \mid s) \propto \exp(-E(s,a))$$

Potential failure on exploitation

### Exploitative target policies

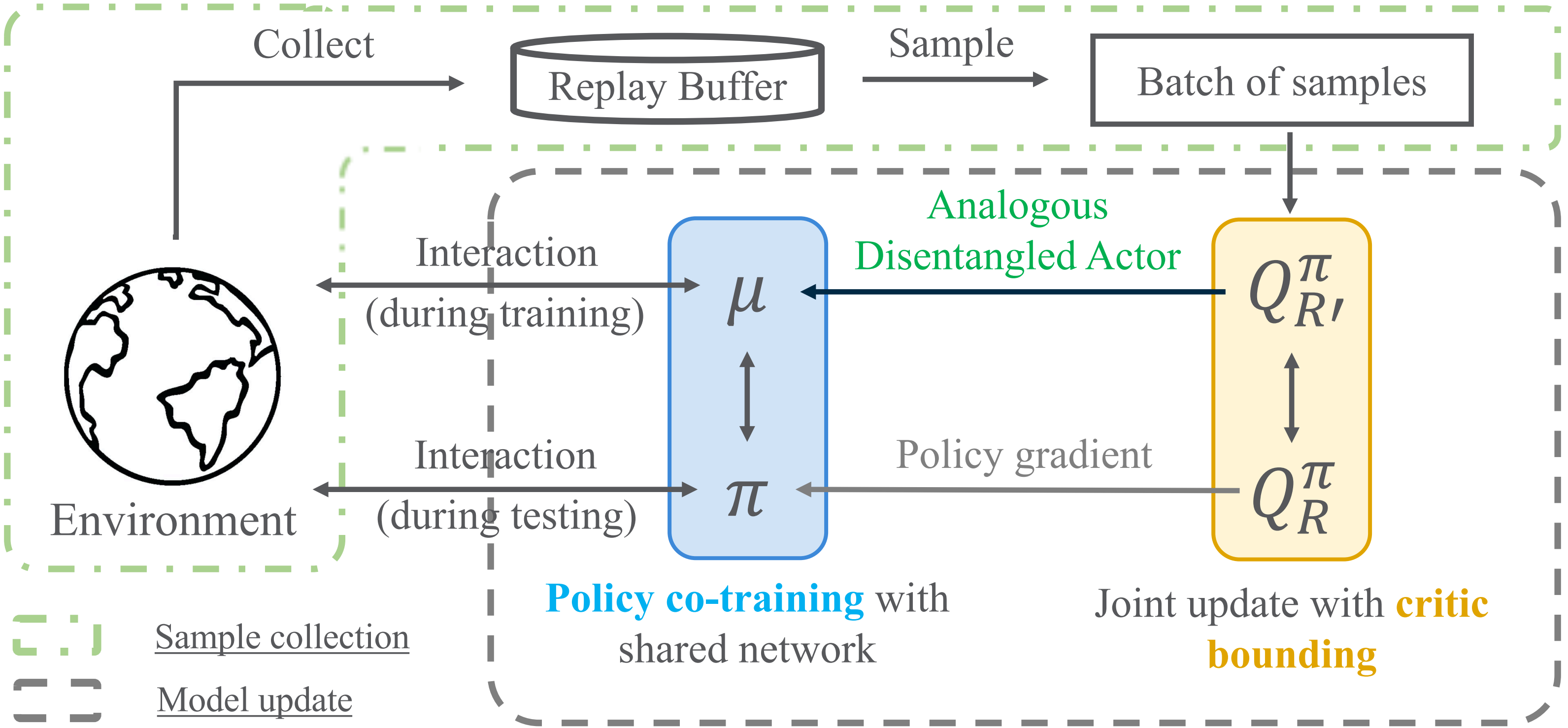
Always optimal w.r.t. current knowledge

$$\pi(a \mid s) = \arg\max_a Q(s,a)$$

Usually requires uninformative noise-based exploration.

Achieving **better tradeoff** through the flexibility to **design separated policies**?

# Analogous Disentangled Actor-Critic (ADAC)



# Stabilizing Policy Updates by Policy Co-training

## Motivation

**Expressiveness**
of the behavior policy

⇗ Improve efficiency of exploration ☺

⇘ Increased distance between target policy and behavior policy

⇩

**Stability problems** during learning ☹

Obtain expressive while stable and optimal behavior policy.

# Stabilizing Policy Updates by Policy Co-training

## Key observation

classic policy optimization

$$\pi(a \mid s) = \arg\max_a Q(s, a)$$

Deterministic policy gradient

$\Downarrow$

maximum-entropy policy optimization

$$\mu(a \mid s) \propto \exp(Q(s, a))$$

Deterministic policy gradient +
entropy regularization

$\Downarrow$

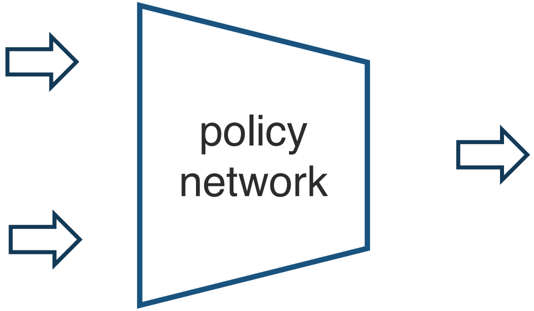
**Joint learning** of behavior/target policy in a **shared neural network**.

# Analogous Disentangled Behavior Policy

state $s$

target policy $\xi = [0, 0, \dots, 0]^T$

behavior policy $\xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$




action $a$

## Benefits

Increase the effectiveness of the behavior policy

- Reduce the distance between two policies and stabilize the learning process.
- Allowing expressive behavior policy for efficient exploration.

# Incorporating Intrinsic Reward via Critic Bounding

## Motivation

**Intrinsic reward**

Guided and efficient exploration ☺

Alter the environment-defined objective ☹

Letting intrinsic reward to **only affect the behavior policy**.

# Incorporating Intrinsic Reward via Critic Bounding

$Q^{\pi}_{\mathcal{R}}$ value function w.r.t. the target policy and the environment-defined reward

→ Adopted to improve the target policy.

$Q^{\pi}_{\mathcal{R}'}$ value function w.r.t. the target policy and the enhanced reward,
where $\mathcal{R}' = \mathcal{R} + \mathcal{R}^{\text{in}}$

→ Adopted to improve the behavior policy.

## Theoretical justification

- Bounded training stability.
- Bounded training effectiveness.

# Summary of Advantages

**Policy co-training**

- Binds the target policy and the behavior policy to stabilize the training process.
- Allowing expressive behavior policy.

**Critic bounding**

- Incorporate intrinsic reward for effective exploration.
- Has no effect on the optimality of the target policy.

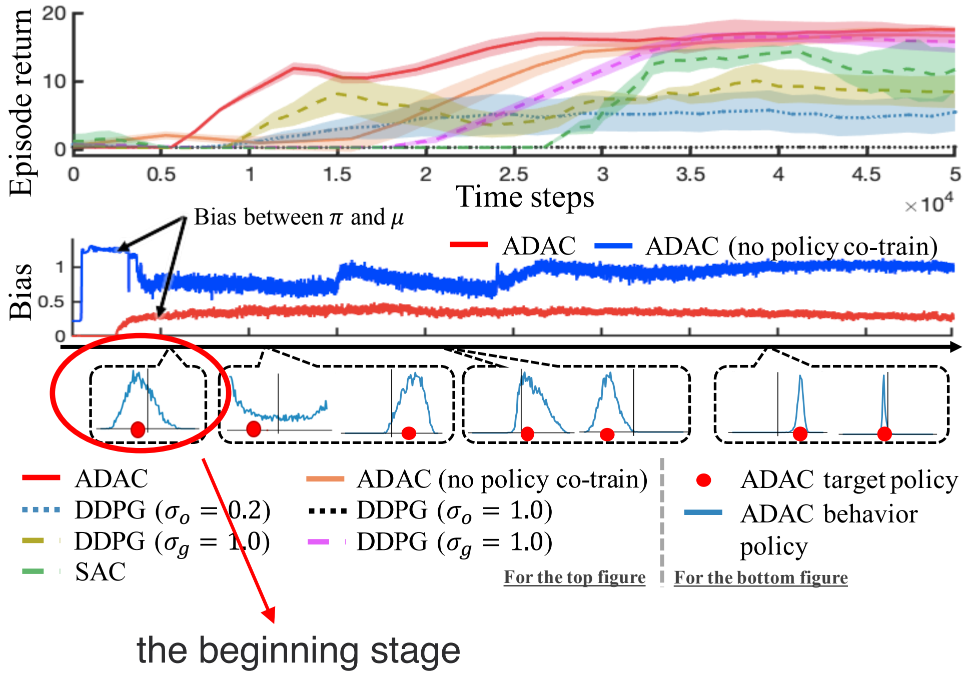
# Analysis of Analogous Disentangled Behavior Policy

## Key results

Behavior policy
- Act curiously at the beginning.
- Focus on potentially rewarding actions after obtaining preliminary understanding of the environment.

Target policy
- Remains optimal w.r.t. the current knowledge.

Target policy & behavior policy
- Bias between them significantly reduced.

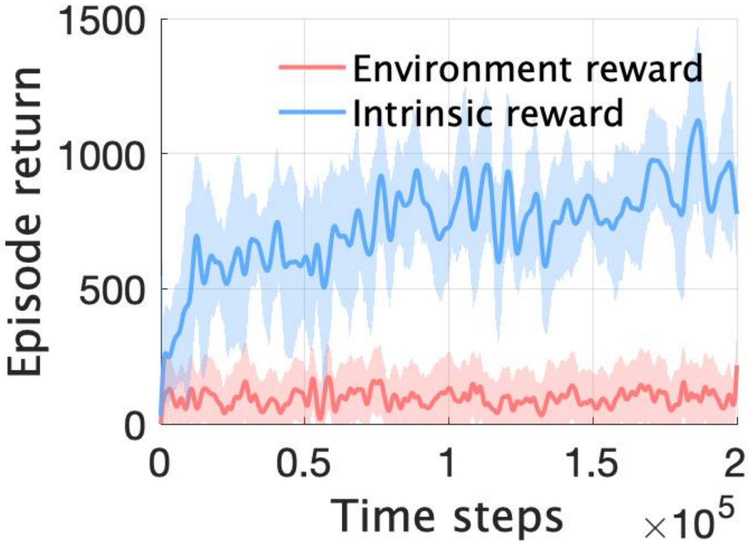


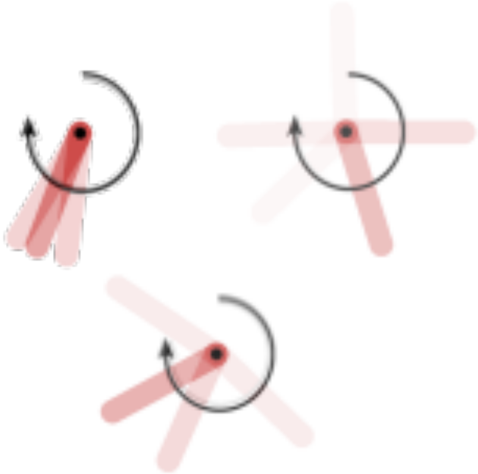
# Comparison with the State of the Art

Out-performs state-of-the-art methods in 10 out of 14 benchmark environments.
- ADAC consistently out-performs its base model, and retains the benefits of improvements developed by the base models.
- Comparison with SAC reveals the benefit brought by the disentangled structure.

| Environment | ADAC (TD3) | ADAC (DDPG) | TD3 | DDPG | SAC | PPO |
|---|---|---|---|---|---|---|
| RoboschoolAnt | 2219±373 | 838.1*±97.1 | **2903**†±666 | 450.0±27.9 | **2726**±652 | 1280±71 |
| RoboschoolHopper | **2299**†±333 | 766.5*±10 | **2302**†±537 | 543.8±307 | 2089±657 | 1229±345 |
| RoboschoolHalfCheetah | 1578†±166 | **1711***±95 | 607.2±246.2 | 441.6±120.4 | 807.0±252.6 | 1225±184.2 |
| RoboschoolAtlasForwardWalk | **234.6**†±55.7 | 186.7*±37.9 | 190.6±50.1 | 52.63±26.2 | 126.0±47.1 | 107.6±29.4 |
| RoboschoolWalker2d | **1769**†±452 | **1564***±651 | 995.1±146.3 | 208.7±137.1 | 1021±263 | 578.9±231.3 |
| Ant | 3353±847 | 1226*±18 | 4034†±517 | 370.5±223 | **4291**±1498 | 1401±168 |
| Hopper | **3598**†± 374 | 374.5*±36.5 | 2845±609 | 38.93±0.88 | **3307**±825 | 1555±458 |
| HalfCheetah | 9392±199 | 2238*±40 | 10526†±2367 | 1009±49 | **11541**±2989 | 881.7±10.1 |
| Walker2d | **5122**†±1314 | 1291*±42 | 4630†±778 | 186.2±33.3 | 4067±1211 | 1146±368 |
| InvertedPendulum | **1000**†±0 | **1000***±0 | **1000**†±0 | **1000***±0 | **1000**±0 | 98.90±2.08 |
| InvertedDoublePendulum | **9359**†±0.17 | 9334*±1.39 | 7665±566 | 27.20±2.61 | 9353±2896 | 98.90±5.88 |
| BipedalWalker | **309.8**†±15.6 | -52.77*±1.94 | 288.4†±51.25 | -123.90±11.17 | **307.2**±57.92 | 266.9±28.52 |
| BipedalWalkerHardcore | **-10.76**†±27.70 | -98.52±3.21 | -57.97±21.08 | -50.05*±10.27 | -127.4±45.2 | -105.3±22.2 |
| LunarLanderContinuous | **290.0**†±50.9 | 85.67*±23.42 | **289.7**†±54.1 | -65.89±96.48 | 283.3±69.29 | 59.32±68.44 |

# Why intrinsic reward sometimes harms RL

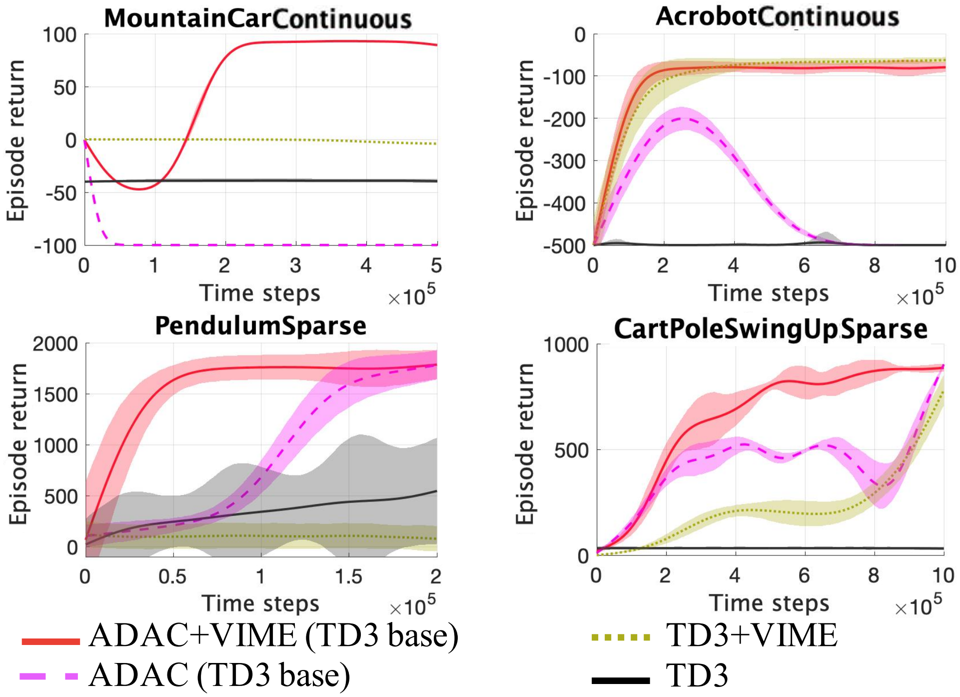


**Environment-defined reward** stays the same while **intrinsic reward keep growing**.

# ADAC with Intrinsic Reward

ADAC **out-performs baseline methods on challenging sparse-reward tasks** when using intrinsic reward.



# Thank You

Open source code: github.com/UCLA-StarAI/Analogous-Disentangled-Actor-Critic