Self-Supervised Sample-Efficient RL

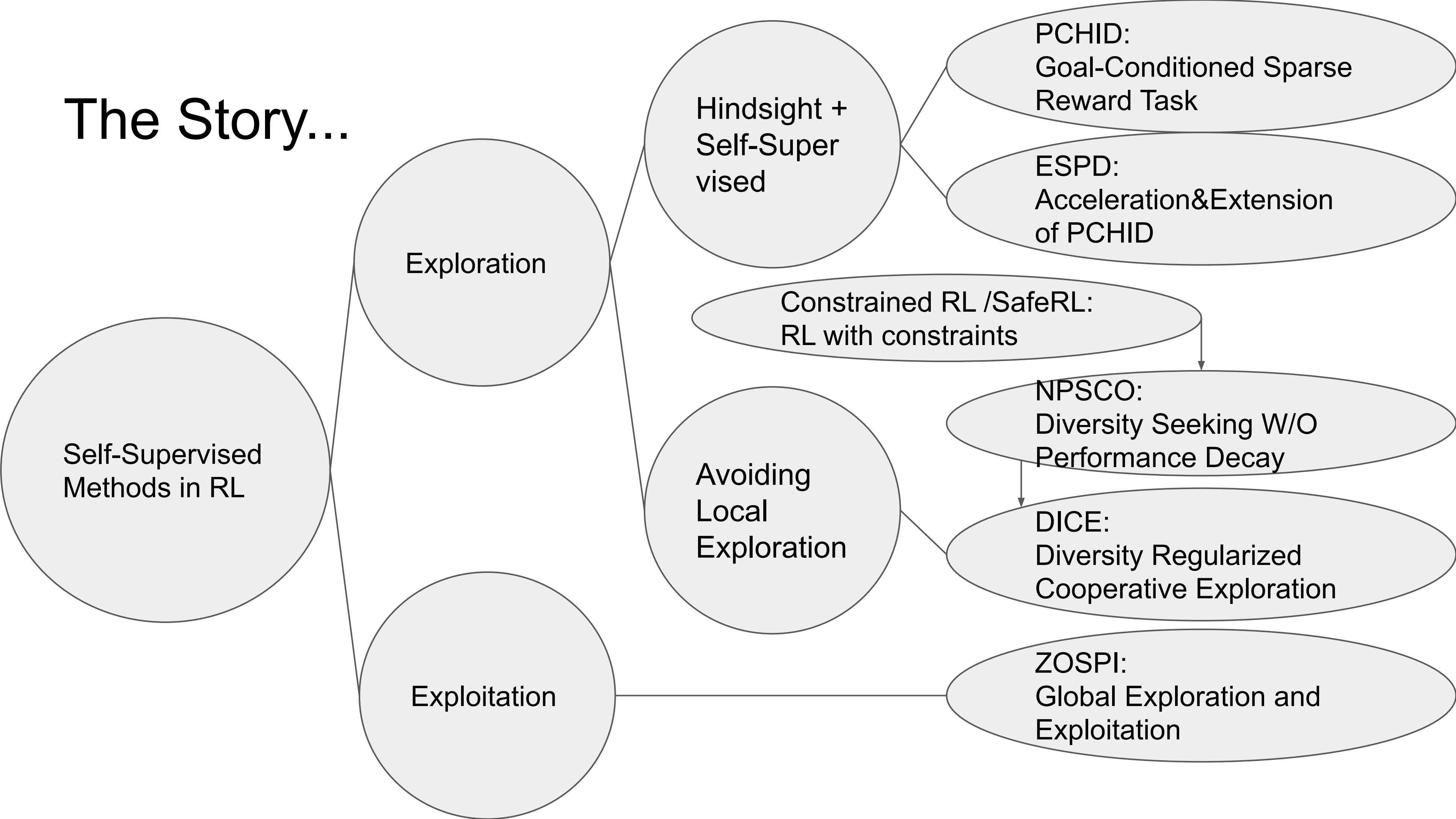
Hao Sun

Content

Self-Supervised RL

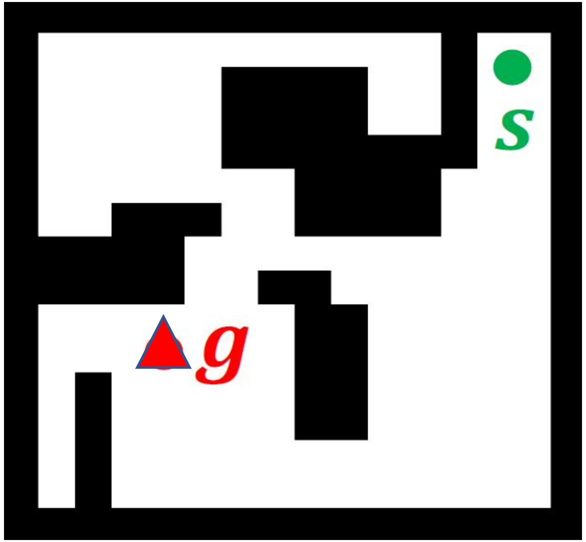
- PCHID and ESPD
- Exploration with Novelty Seeking
- Better Exploitation with Zeroth-Order Supervised Policy Improvement

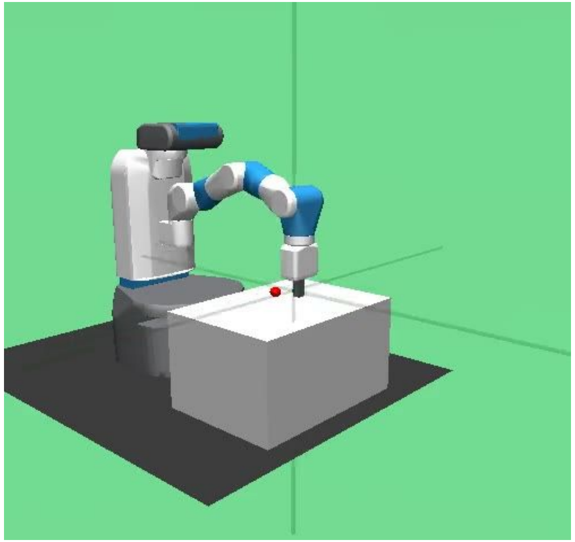
The Story...

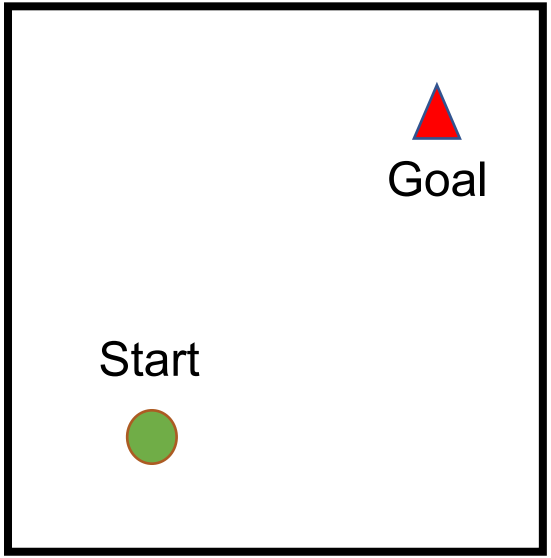

Self-Supervised Methods in RL
Exploration
Exploitation
Hindsight + Self-Supervised
Avoiding Local Exploration
PCHID: Goal-Conditioned Sparse Reward Task
ESPD: Acceleration&Extension of PCHID
Constrained RL /SafeRL: RL with constraints
NPSCO: Diversity Seeking W/O Performance Decay
DICE: Diversity Regularized Cooperative Exploration
ZOSPI: Global Exploration and Exploitation

PCHID and ESPD: Goal-Conditioned RL

Goal-Oriented Reward Sparse Tasks





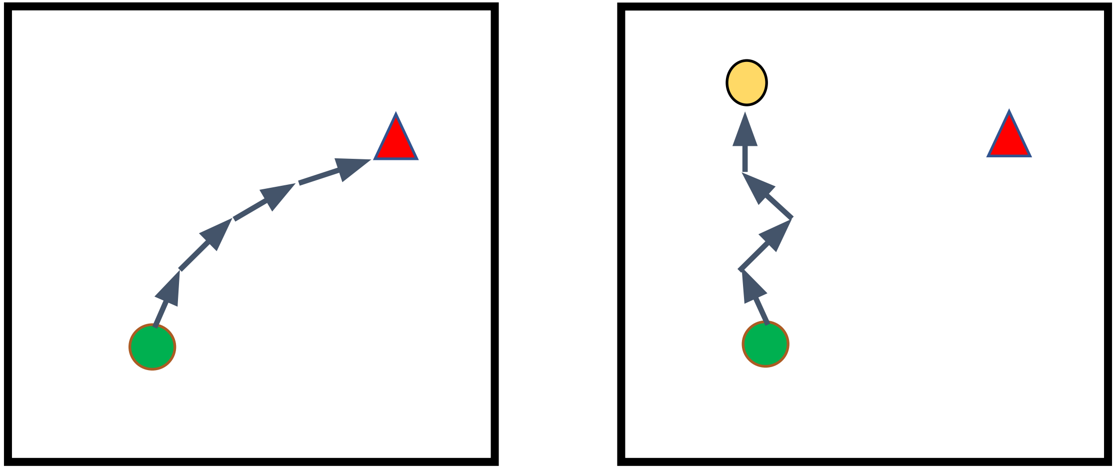


Inspirations from Human Learning

1. Learning from failures

[Hindsight Experience Replay, M Andrychowicz et al. 2017]

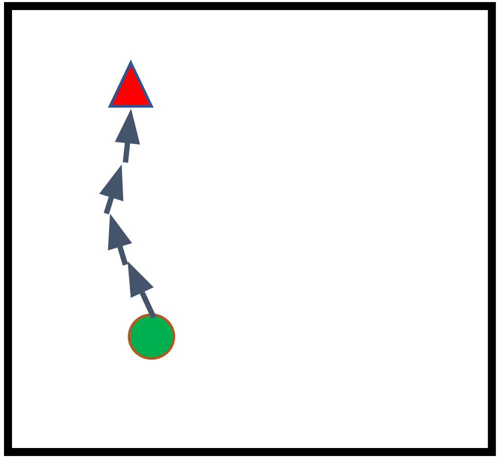
Aimed Achieved



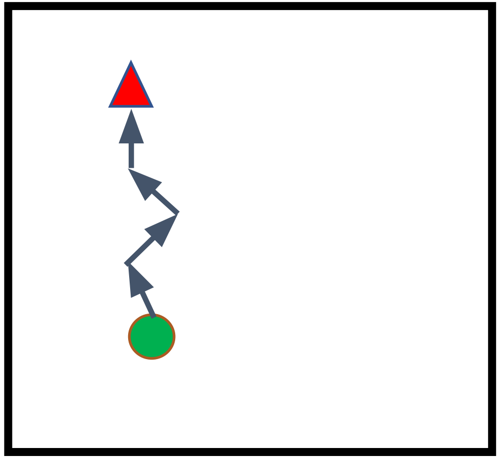
Inspirations from Human Learning

1. Learning from failures
[Hindsight Experience Replay, M Andrychowicz et al. 2017]

Aimed



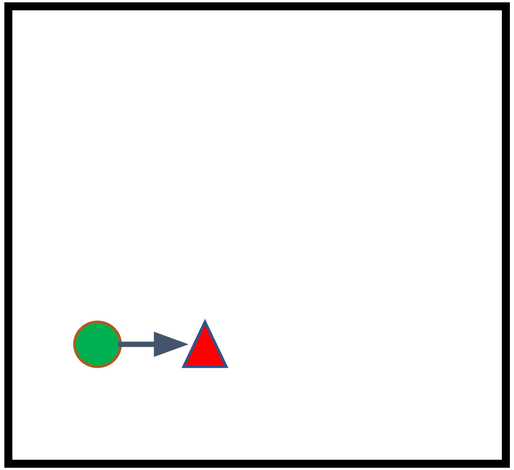
Achieved



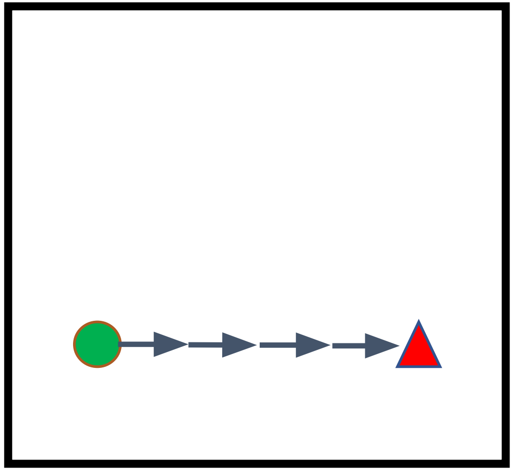
Inspirations from Human Learning

1. Learning from failures
2. Extrapolating Success

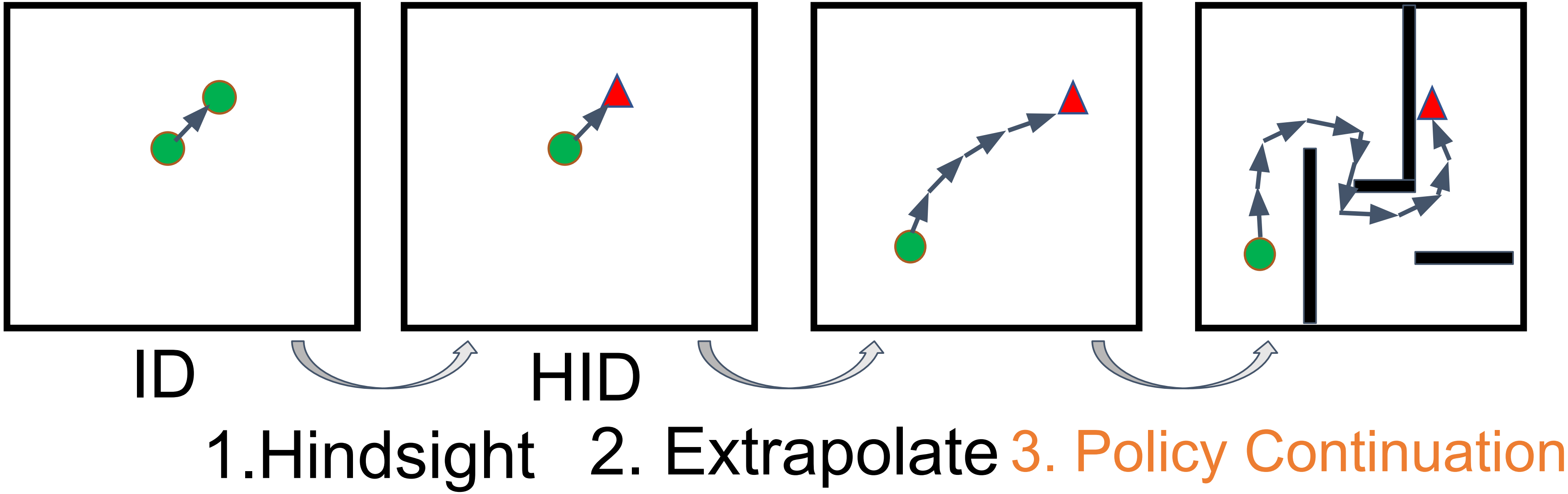
Learned



Extrapolate

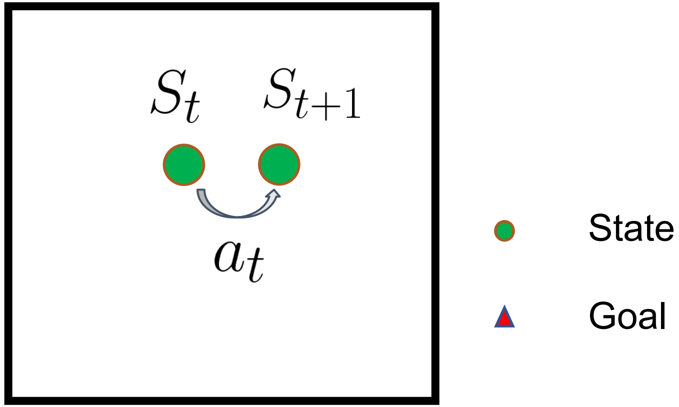


Our Proposed Method

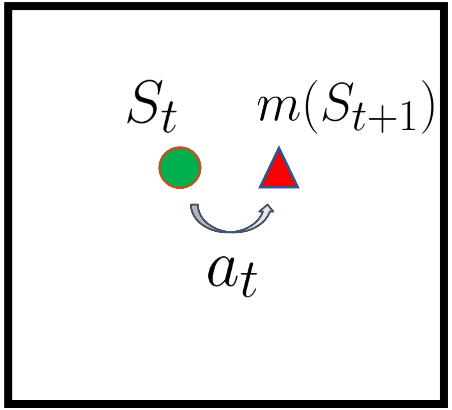


Equipe Inverse Dynamics with Hindsight

Inverse Dynamics:

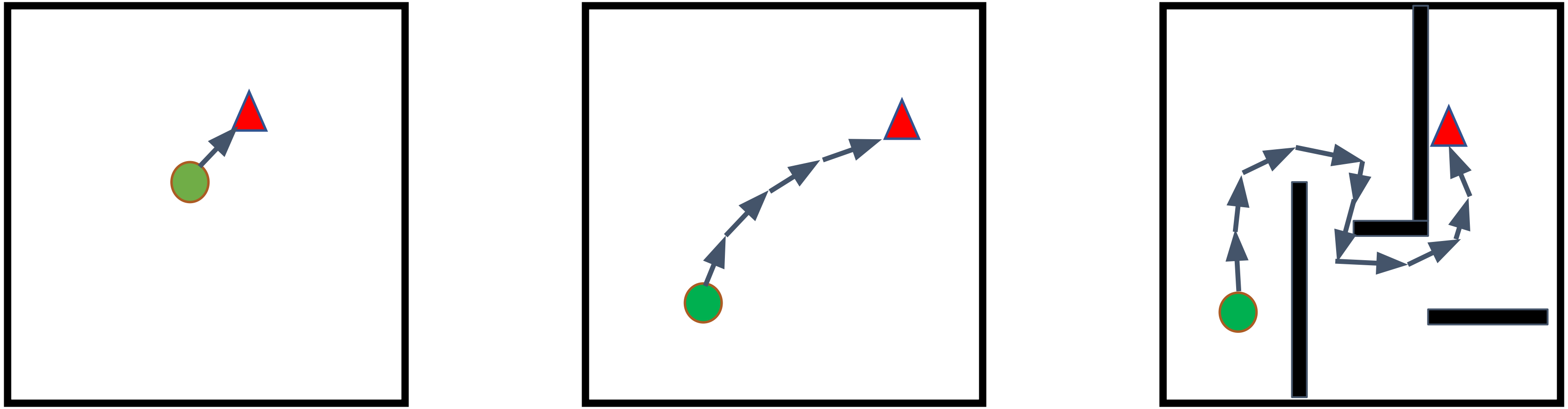


Hindsight Inverse Dynamics:



$$g = m(S_{goal})$$

1-step HID is Not Enough



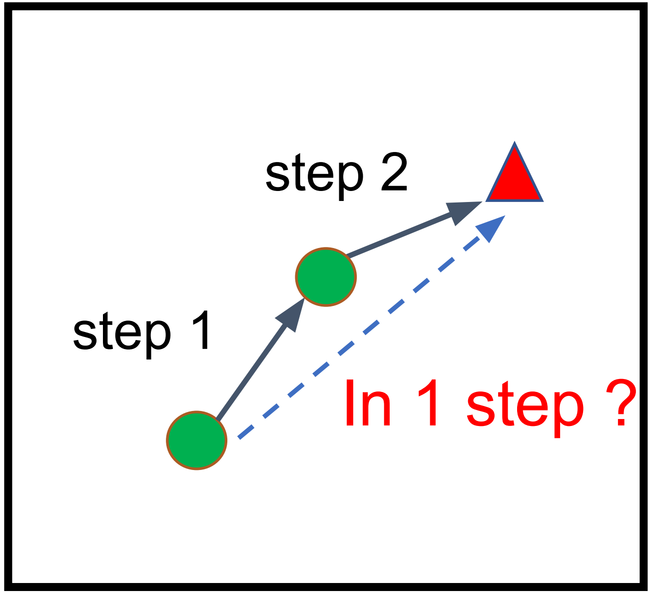
1-step HID

Linear Case

Non-linear Case

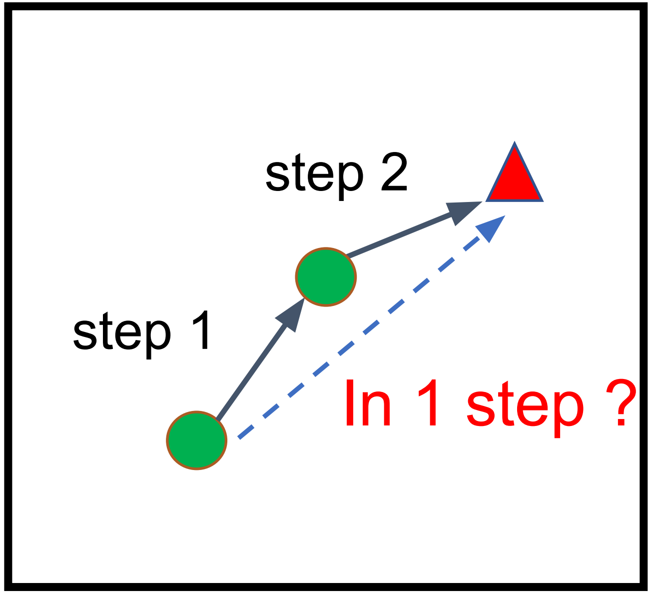
Multi-step Optimality?

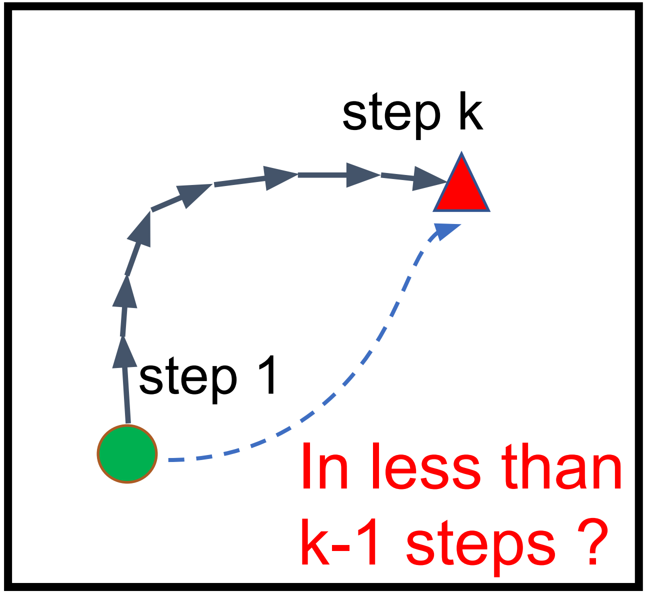
Policy Continuation: Test the optimality recursively



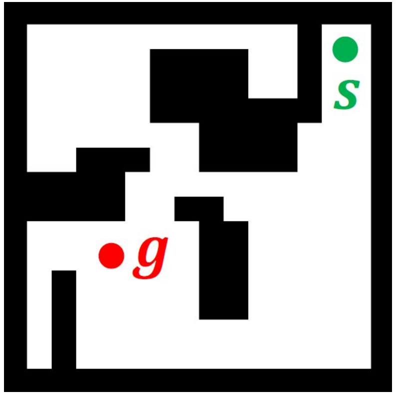
Multi-step Optimality?

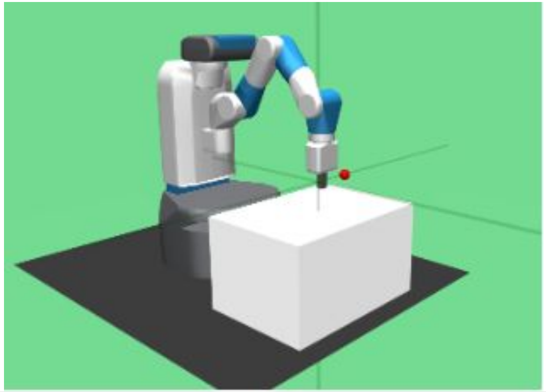
Policy Continuation: Test the optimality recursively

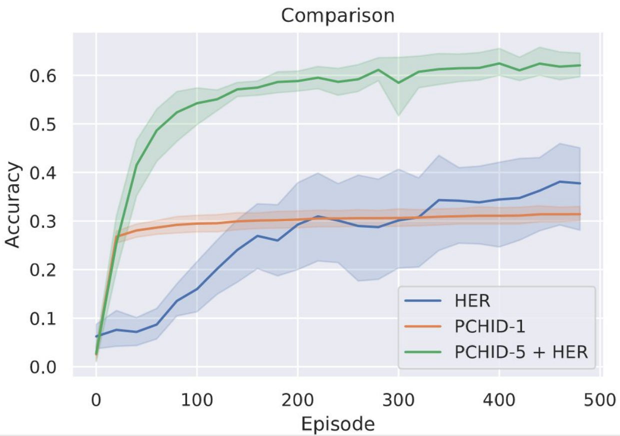


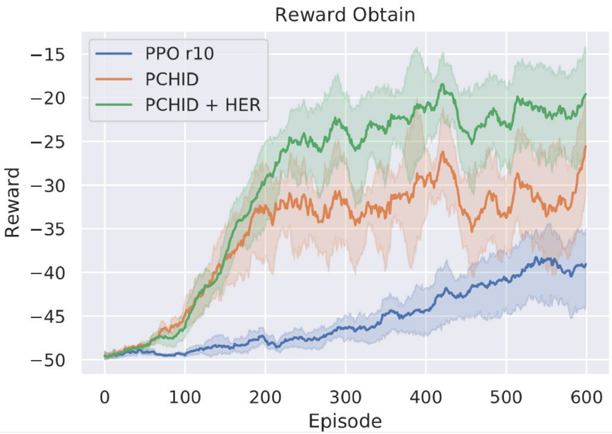


Experiments



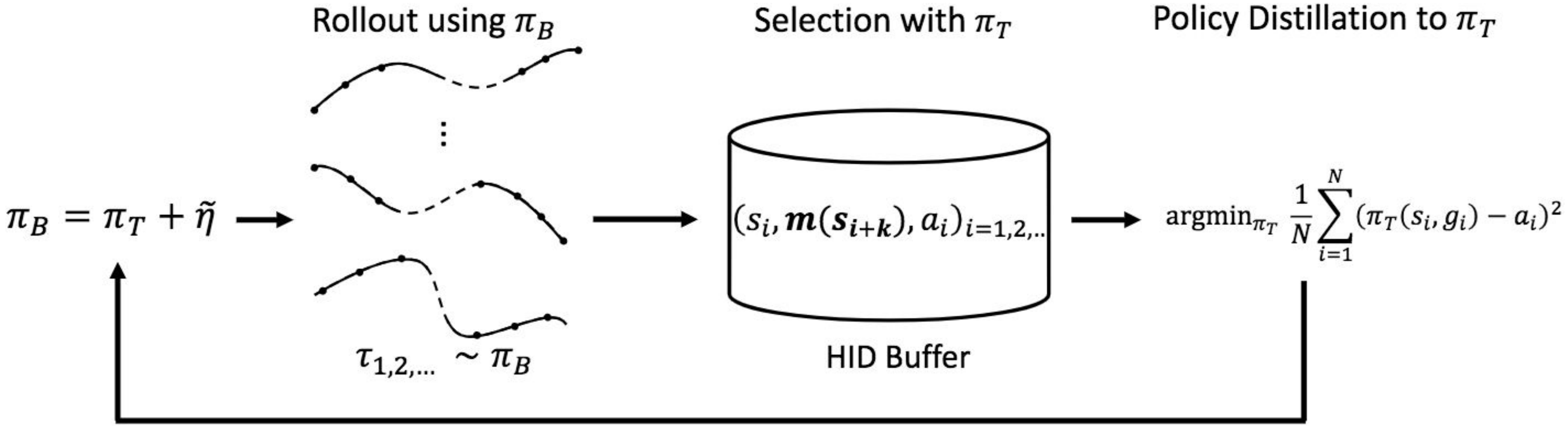






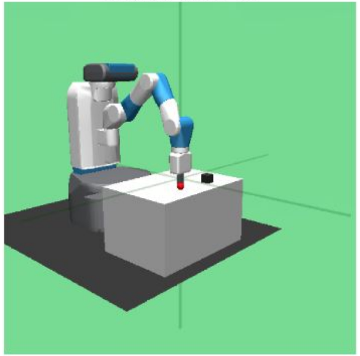
Evolutionary Stochastic Policy Distillation

- Accelerated PCHID

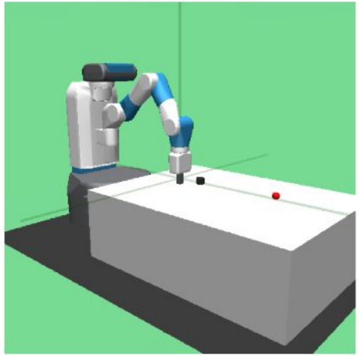


Experiments

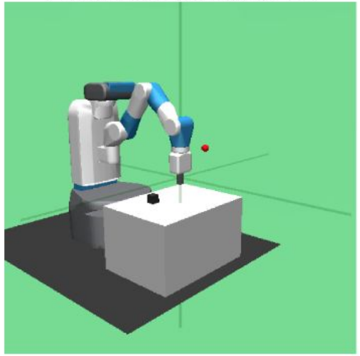
FetchPush-v1

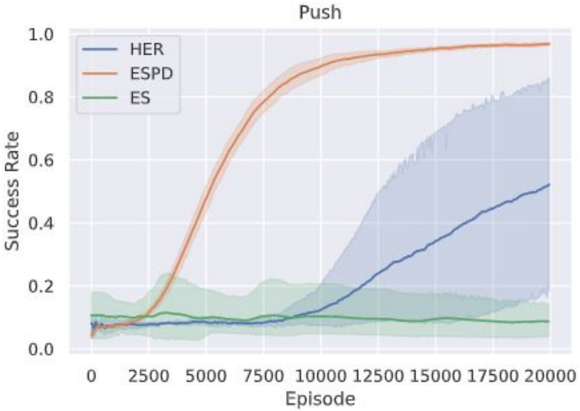


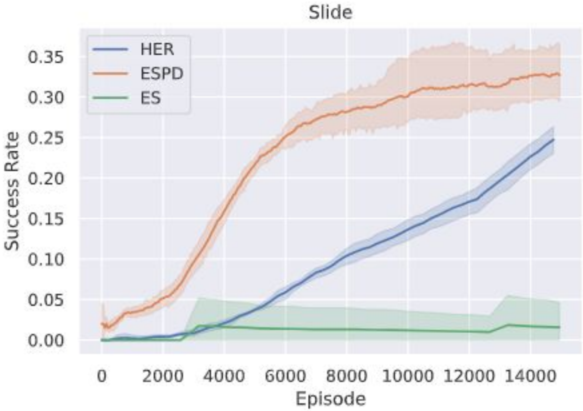
FetchSlide-v1

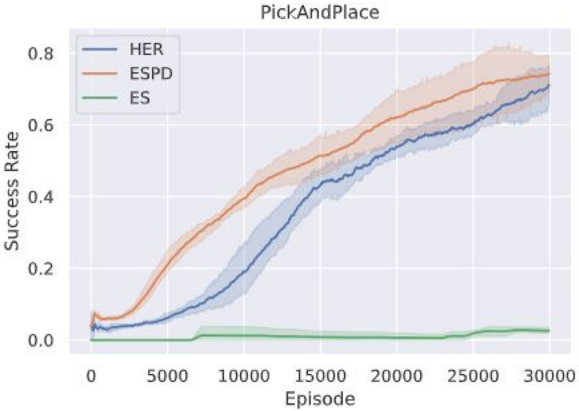


FetchPickAndPlace-v1









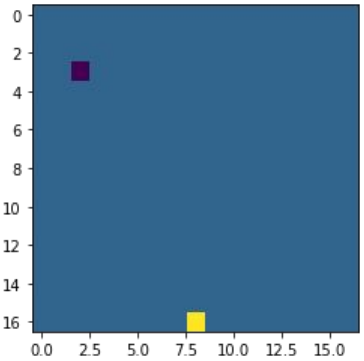
Global Exploration and Generalized Self-Supervised RL

The environment:

A maze with a non-trivial positive reward located at different positions (at centers of different sides)

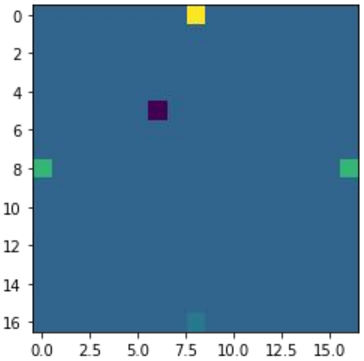
Each time the point will receive a negative reward of -0.1 if it has not reached the positive reward

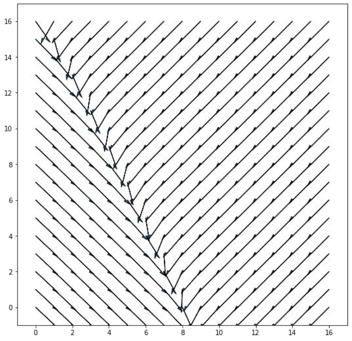
Finite time horizon: 2*N, where in the experiments I set N = 17, the scale of the gridworld



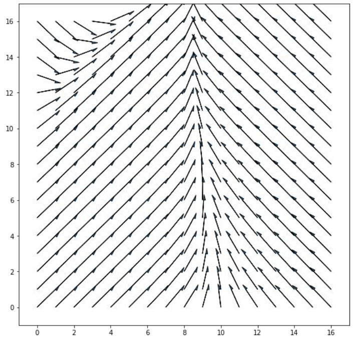
The environment can be easily extend to multiple reward cases.

I'm working with the multiple reward cases with a previous draft.

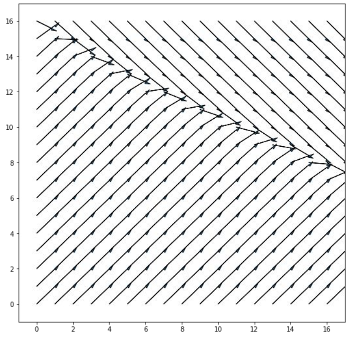




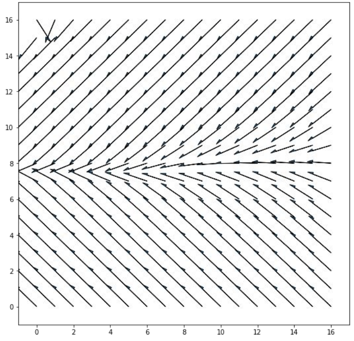
down : +10
200 epoches
~227 reward



up : +10
200 epoches
~237 reward



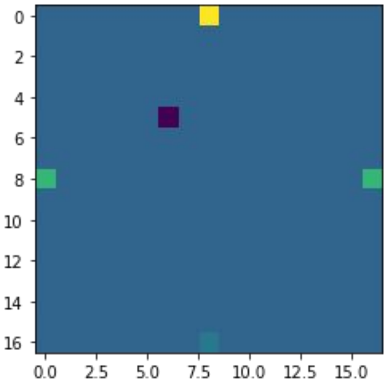
right : +10
200 epoches
~224 reward

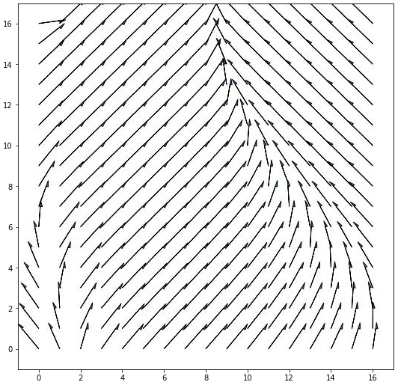


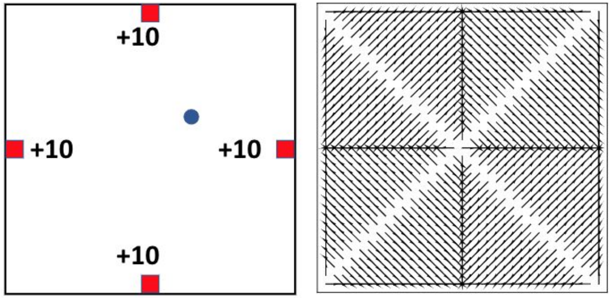
left : +10
200 epoches
~233 reward

4 reward maze, result with vanilla PPO

N=17,reward: left = 10,right = 10, up=10, down = 10

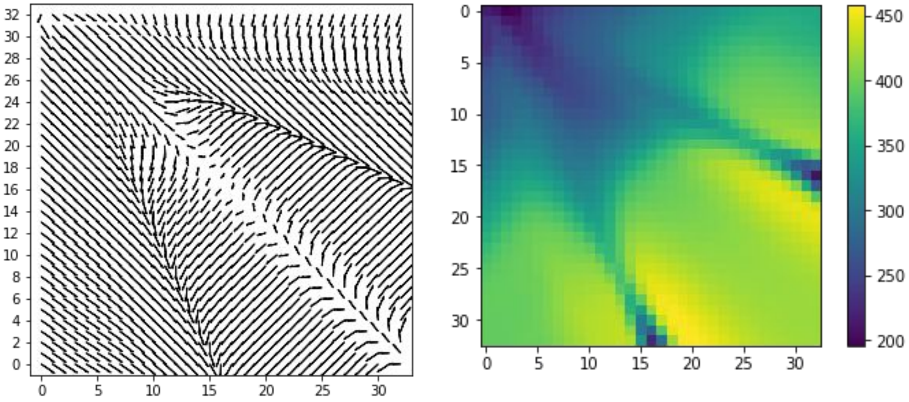




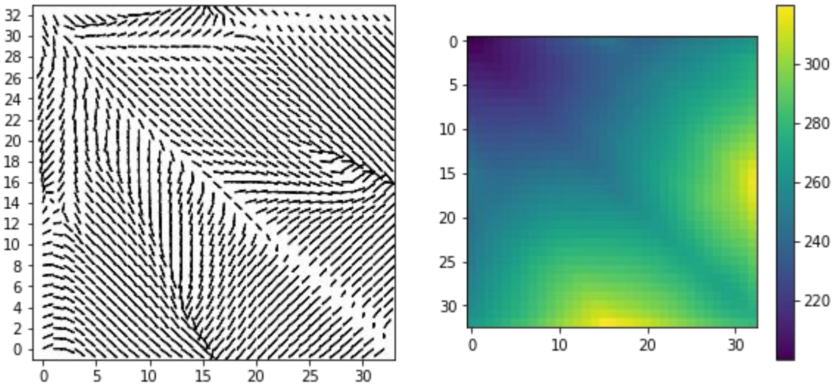


(a) The Four-Solution-Maze environment and the optimal solution

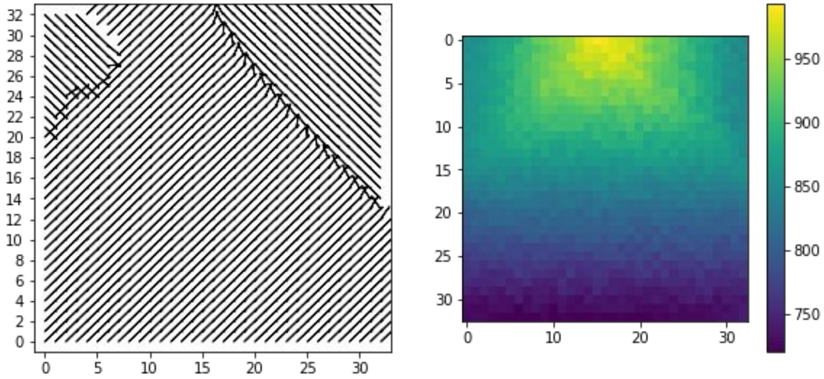
PPO: 30w timestep, reward = 435(477,80W)



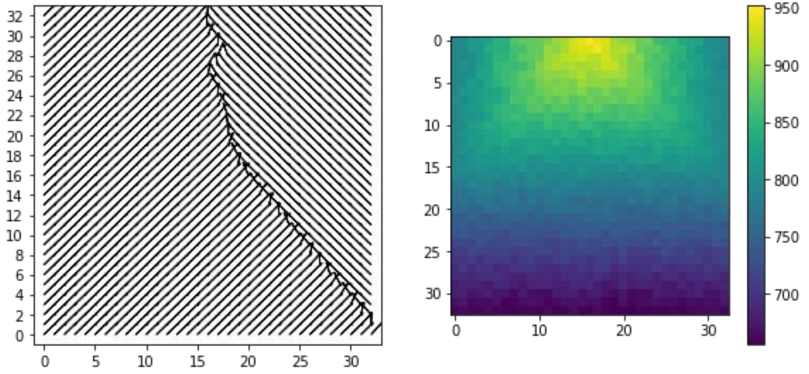
SAC: 30w timestep, reward = 419 (451, 120w)



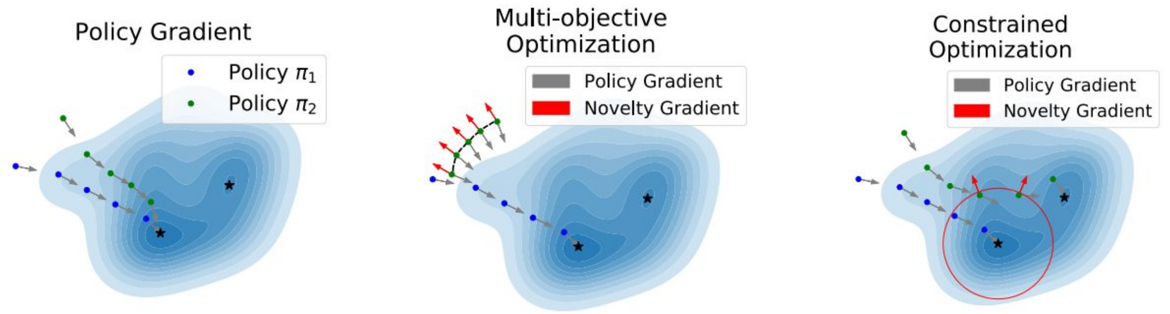
DDPG: 30w timestep, reward = 463

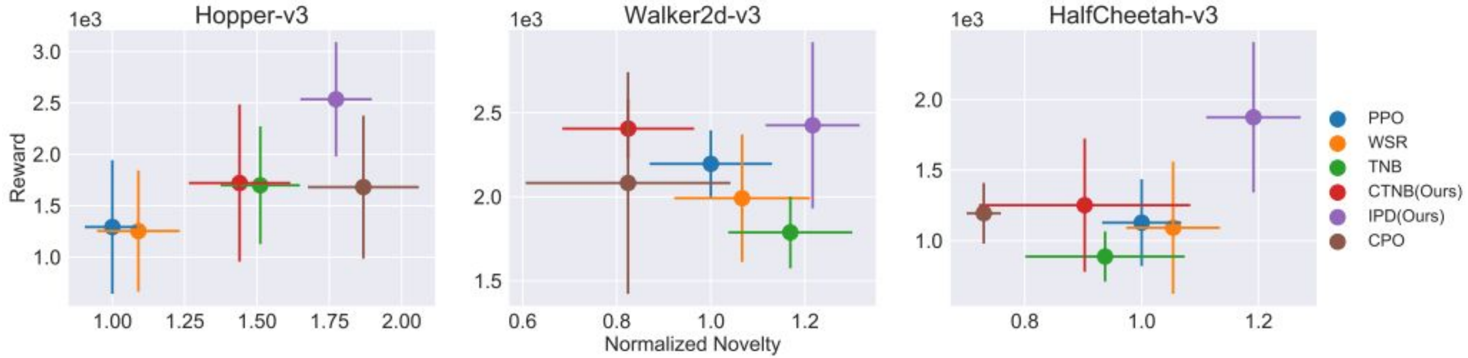


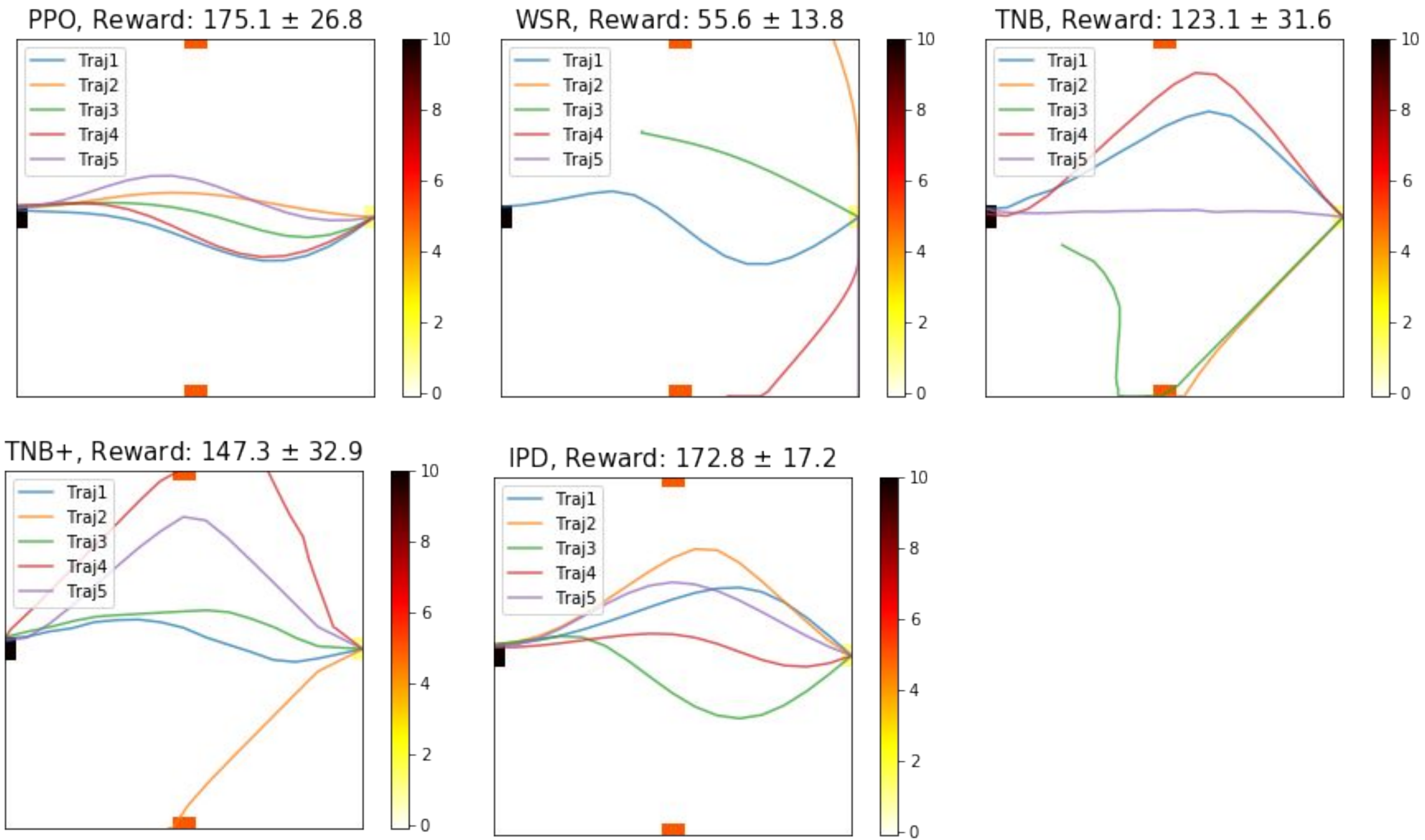
TD3: 30w timestep, reward = 493



Novelty Seeking Methods







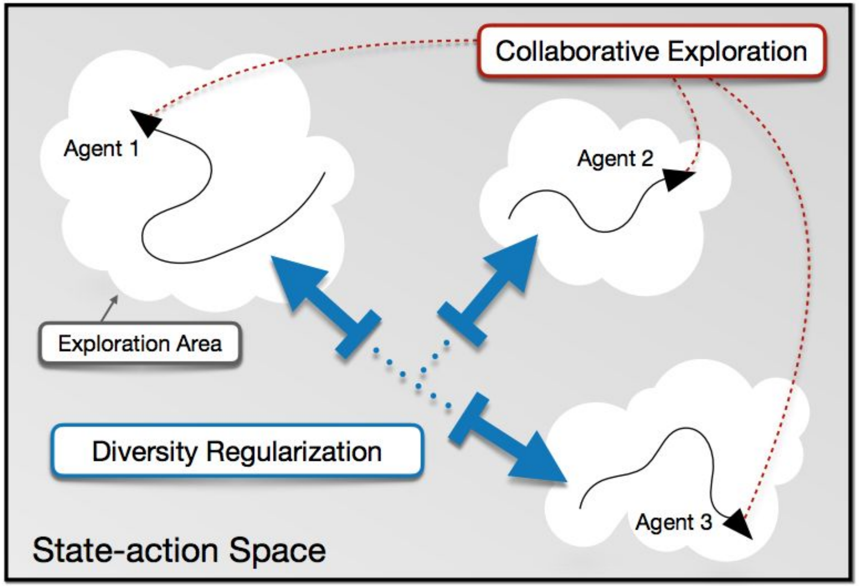
PPO, Reward: 175.1 ± 26.8
WSR, Reward: 55.6 ± 13.8
TNB, Reward: 123.1 ± 31.6
TNB+, Reward: 147.3 ± 32.9
IPD, Reward: 172.8 ± 17.2
Traj1
Traj2
Traj3
Traj4
Traj5
0
2
4
6
8
10

Diversity Seeking for Non-Local Exploration

Every policy explores locally

But with **diversity regularization**,

they can explore **cooperatively**.



Diversity Seeking for Non-Local Exploration

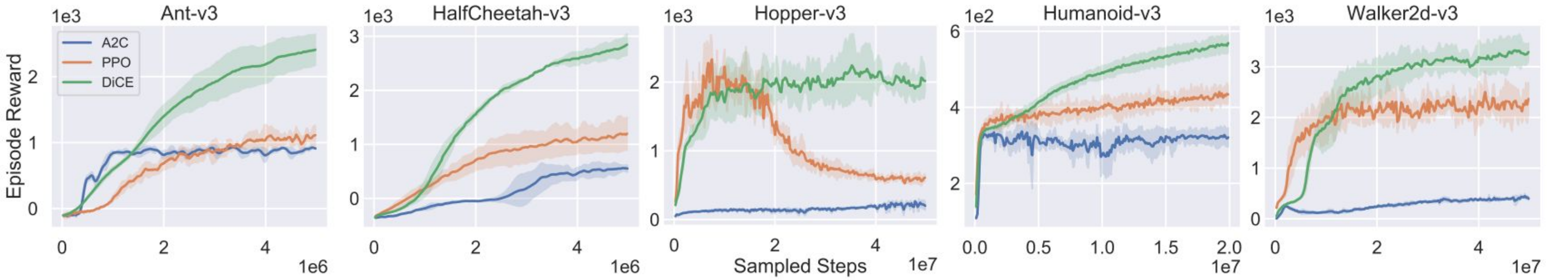


Figure 4: The learning curves of A2C, PPO, and on-policy DiCE in 5 MuJoCo environments.

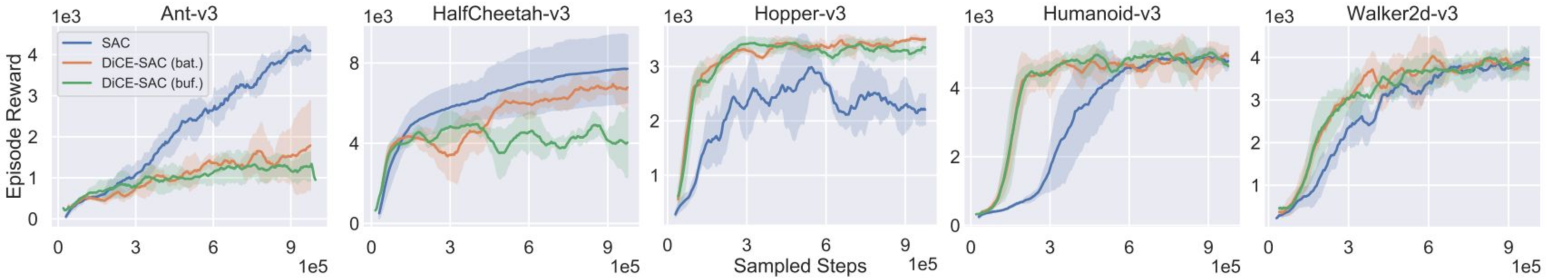
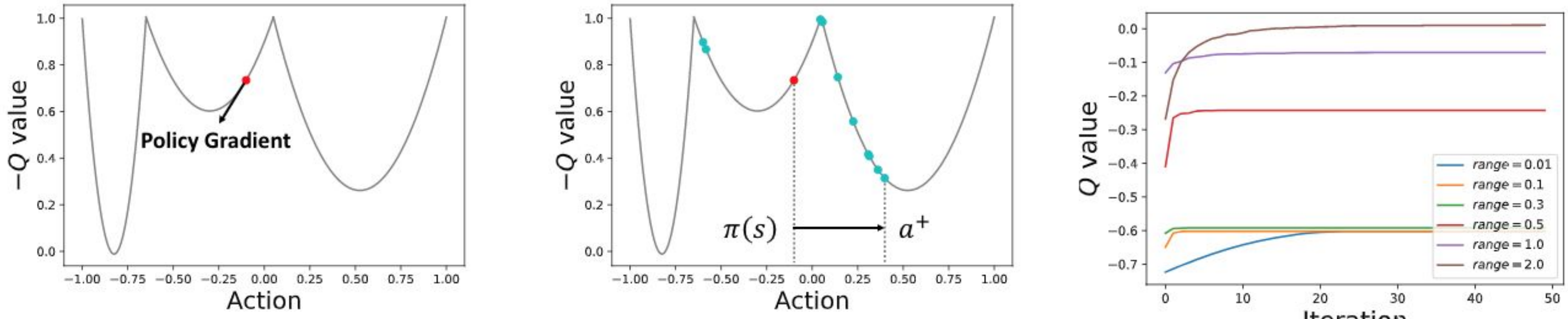


Figure 5: The learning curves of SAC and off-policy DiCE in 5 MuJoCo environments.

Zeroth-Order Method: RL without Policy Gradient

Intuition:

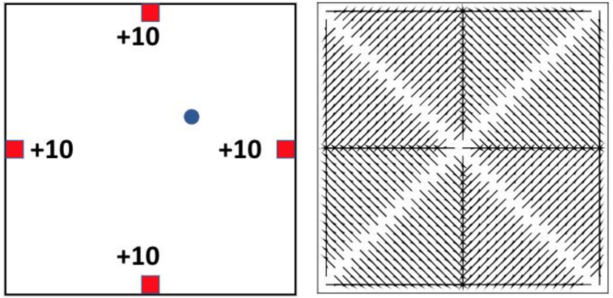


(a) Policy Gradient (b) Supervised Policy Improvement (c) Simulation

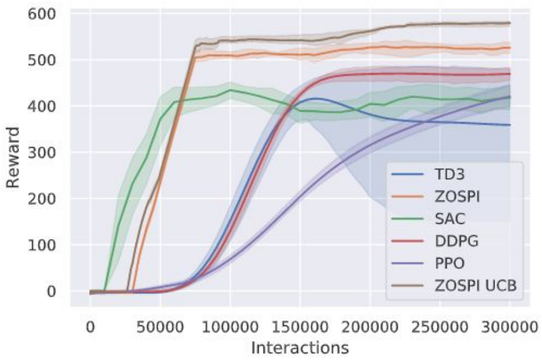
Toy Model Results

ZOSPI is able to explore all the rewards

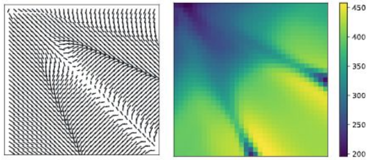
i.e., Global Exploration



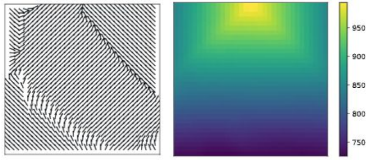
(a) The Four-Solution-Maze environment and the optimal solution



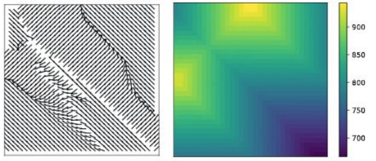
(b) Performance comparison



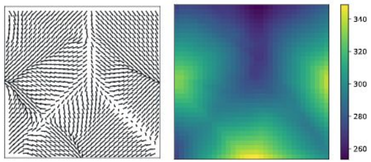
(c) PPO



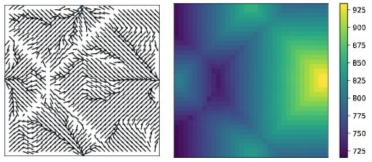
(d) DDPG



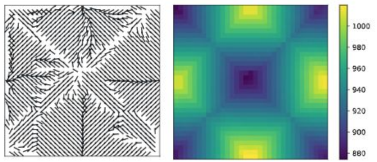
(e) TD3



(f) SAC

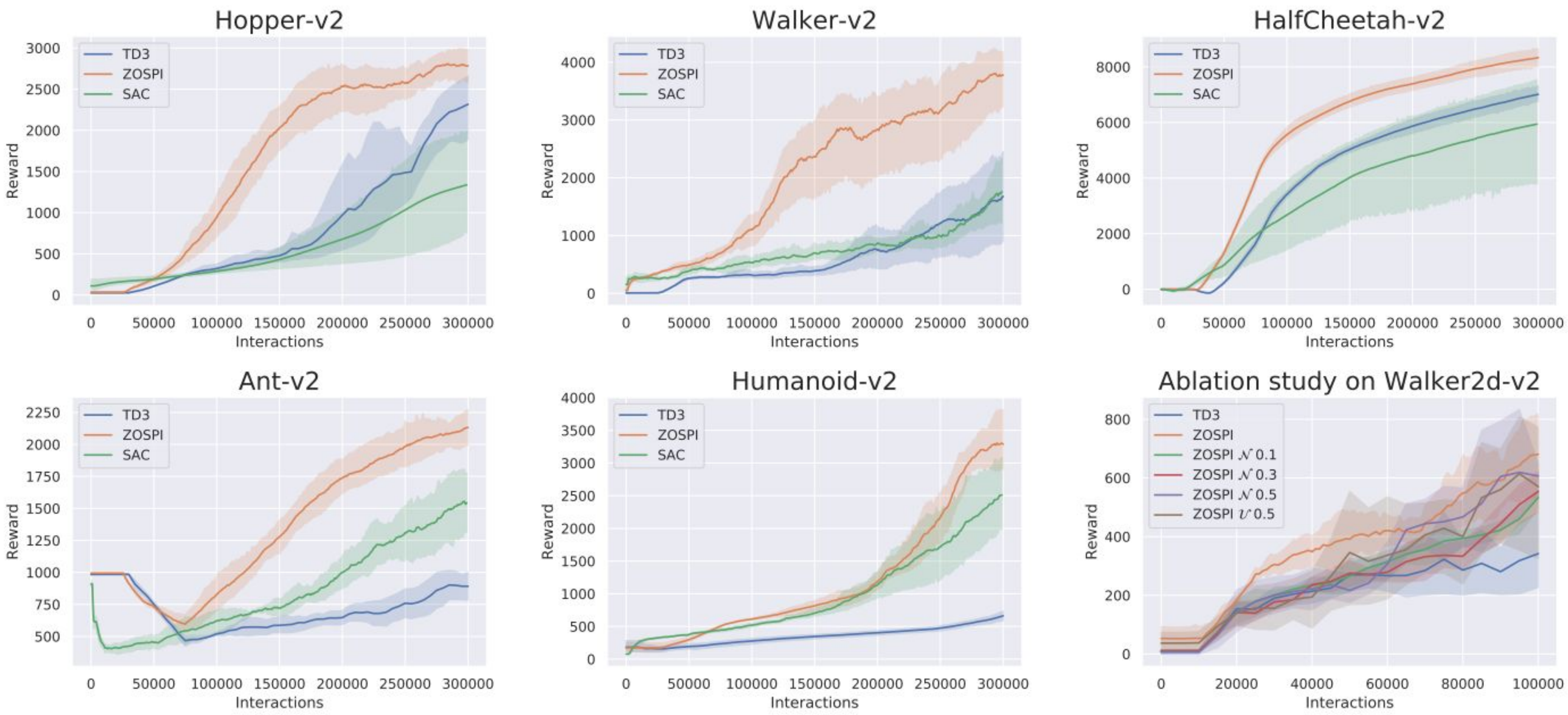


(g) ZOSPI



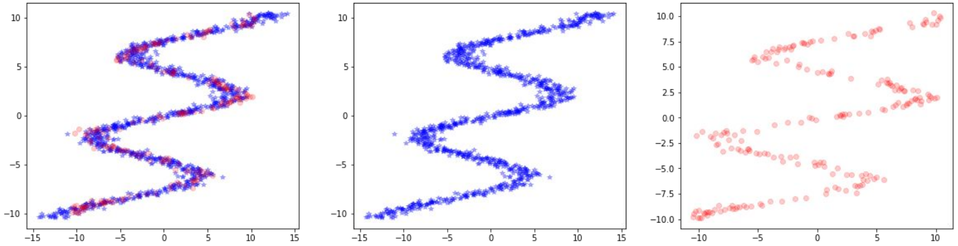
(h) ZOSPI-UCB

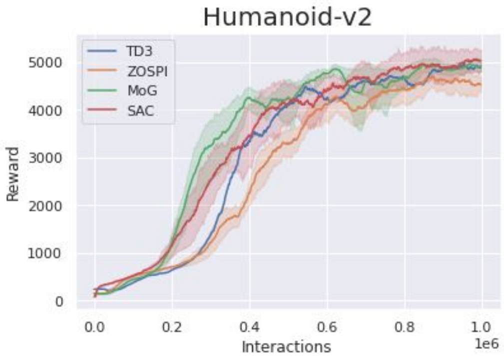
MuJoCo Benchmarks



Extensions Based on ZOSPI

- ZOSPI is learned with self-supervised learning (regression)
  - Multi-modal policy can be used (MDN policy)





  - Non-parametric model can be used (e.g., GP policy)