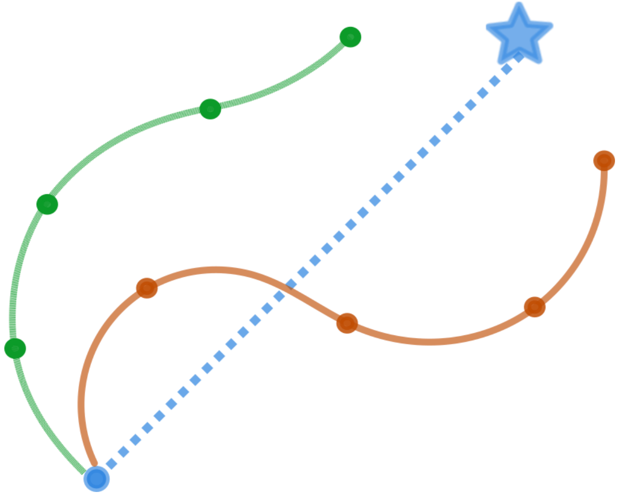
D-Shape: Demonstration Shaped Reinforcement Learning via Goal-Conditioning

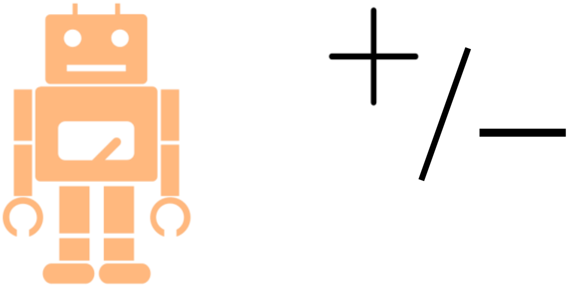


Caroline Wang[1], Garrett Warnell[1, 2], Peter Stone[1, 3]

[1]The University of Texas at Austin, [2] Army Research Laboratory, [3] Sony AI

Motivation

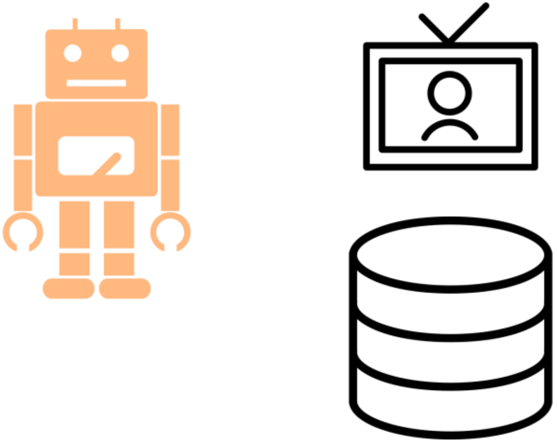
- Reinforcement learning (RL) can autonomously discover optimal behavior from a reward function



…But can be sample inefficient

Motivation

- Imitation learning (IL) methods can learn behaviors from demonstrations with high sample efficiency



…but usually assumes multiple, optimal, state-action demonstrations

Challenges of Combining RL and IL

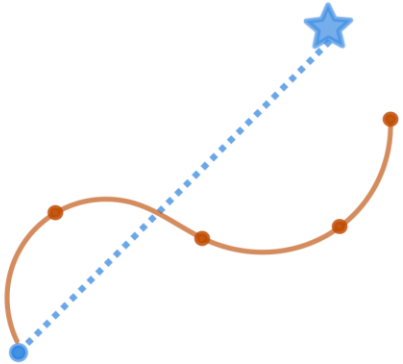
- IL objective: divergence minimization from demonstration distribution [1, 2]

- RL objective: cumulative task reward

Suboptimal demonstrations ⇒ Potential conflict between IL and RL objectives!

[1] Ghasemipour et al., A divergence minimization perspective on imitation learning methods, CoRL 2019.
[2] Ke et al., imitation learning as f-divergence minimization, WAFR 2020.

Can we *improve sample efficiency* of reinforcement learning with minimal demonstration knowledge,
while *preserving optimality guarantees*?



We assume access to a single, suboptimal, state-only demonstration trajectory.

Background

- Markov decision process $M = (S, A, P, r^{task}(s, a, s'), \gamma)$
  - Horizon H
  - Objective: $\mathrm{E}_\pi[\sum_{t=0}^{H-1} \gamma^t r^{task}]$

- Imitation from observation [1]: assumes access to state-only demonstrations

$$D^e = \{s_t^e\}_{t=1}^{H}$$

[1] Torabi et al., Recent advances in imitation from observation, IJCAI 2019.

Background

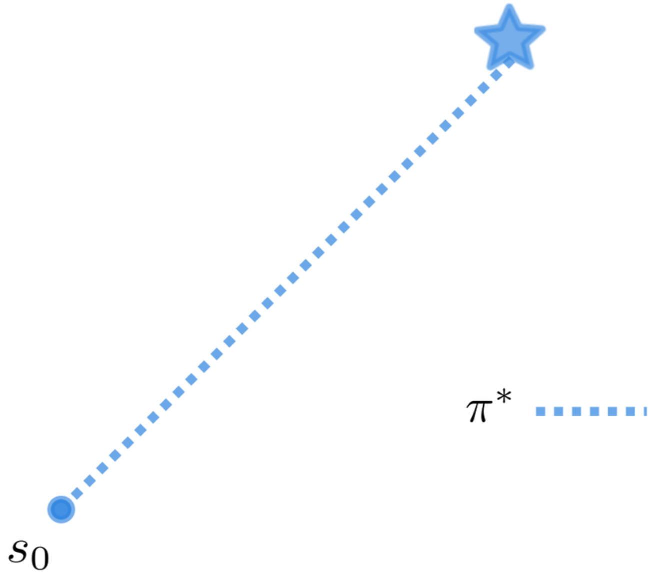
- Potential-based reward shaping (PBRS) [1]:
  - Learning is conducted in modified MDP, where $M = (S, A, P, R' := r^{task} + F, \gamma)$
  - Policy invariance $F(s, s') = \gamma\phi(s') - \phi(s).$

- Goal-conditioned RL (GCRL) [2, 3]:
  - Given a goal-reaching task, objective is to learn a goal-conditioned policy $\pi(\cdot | [s, g])$ that can reach any goal g drawn from goal set G
  - Reward function is typically sparsely informative
  - E.g. $r_t^g = \mathbb{1}_{s_t=g}$

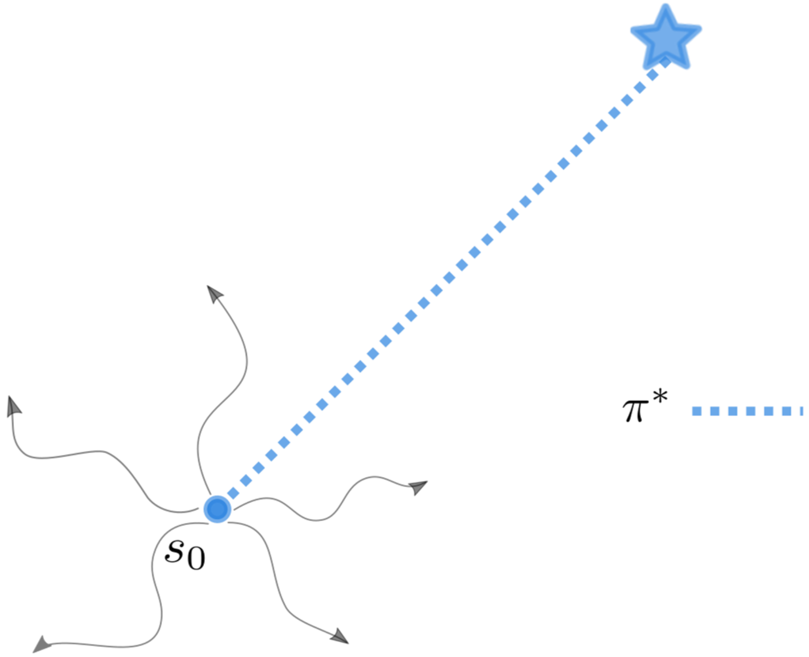
[1] Ng et al., Policy invariance under reward transformations, ICML 1999.
[2] Schaul et al., Universal value function approximators, ICML 2015.
[3] Kaelbling, Learning to achieve goals, IJCAI 1993.

D-Shape: Shaping reinforcement learning with a suboptimal demonstration trajectory

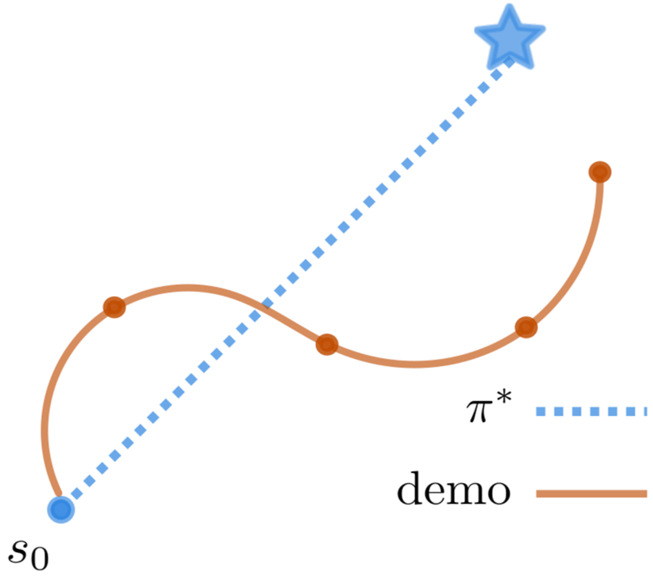
D-Shape: Shaping reinforcement learning with a suboptimal demonstration trajectory



D-Shape: Shaping reinforcement learning with a suboptimal demonstration trajectory



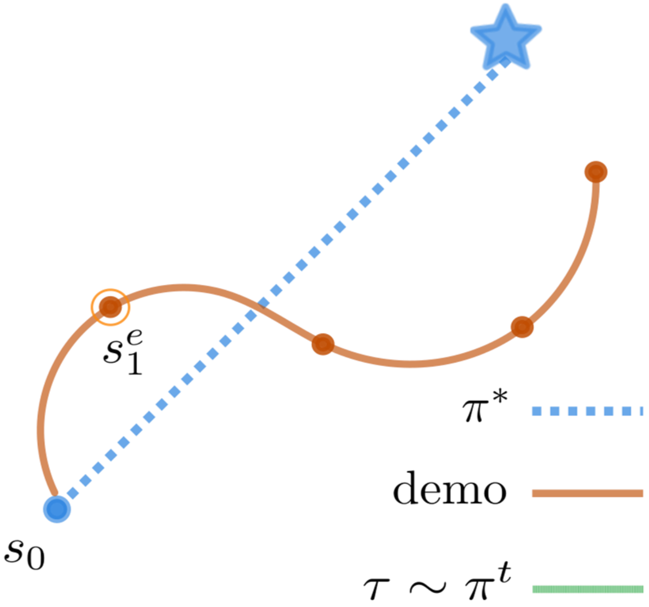
D-Shape: Shaping reinforcement learning with a suboptimal demonstration trajectory



D-Shape: Shaping reinforcement learning with a suboptimal demonstration trajectory

$$\cancel{\pi(s_t)}$$

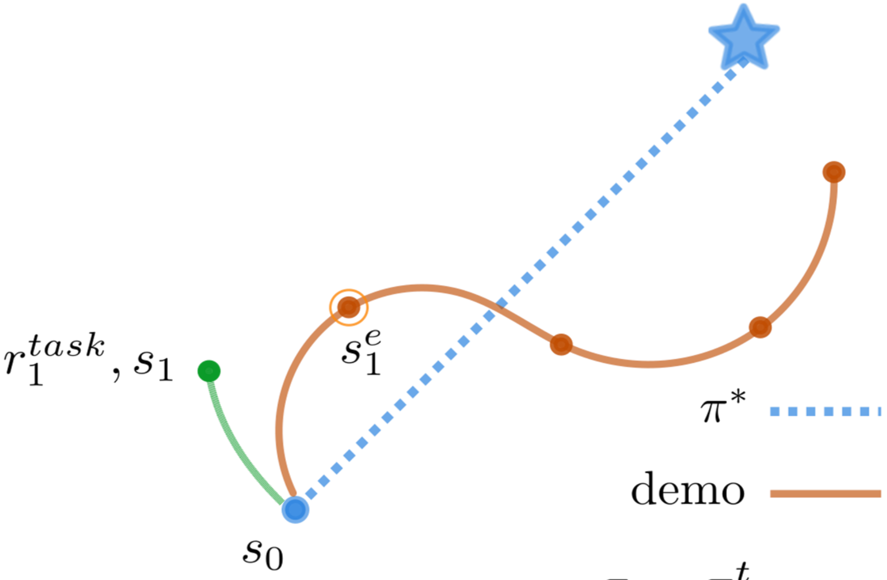
$$\pi([s_t, \boxed{s_t^e}])$$



D-Shape: Shaping reinforcement learning with a suboptimal demonstration trajectory

$$\cancel{\pi(s_t)}$$

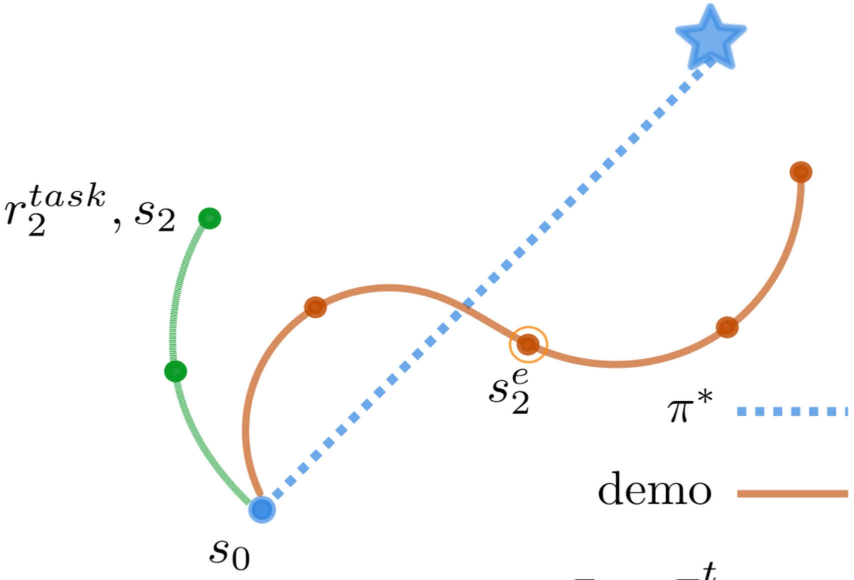
$$\pi([s_t, s_t^e])$$



D-Shape: Shaping reinforcement learning with a suboptimal demonstration trajectory

$$\cancel{\pi(s_t)}$$

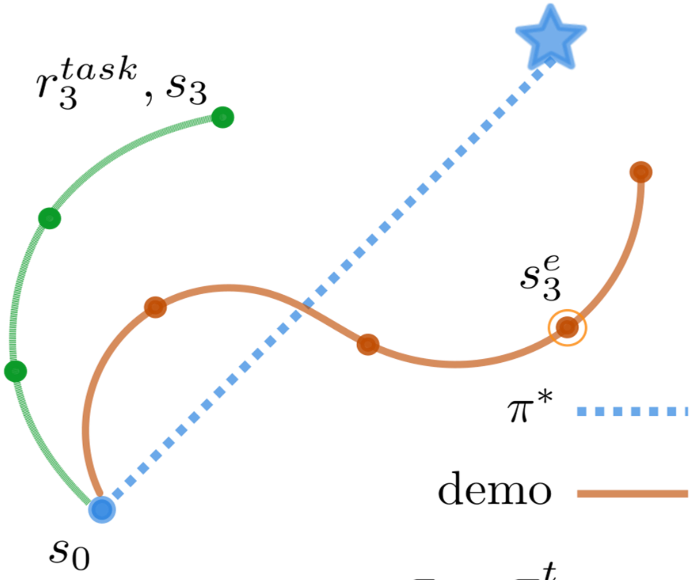
$$\pi([s_t, s_t^e])$$



D-Shape: Shaping reinforcement learning with a suboptimal demonstration trajectory

~~$\pi(s_t)$~~

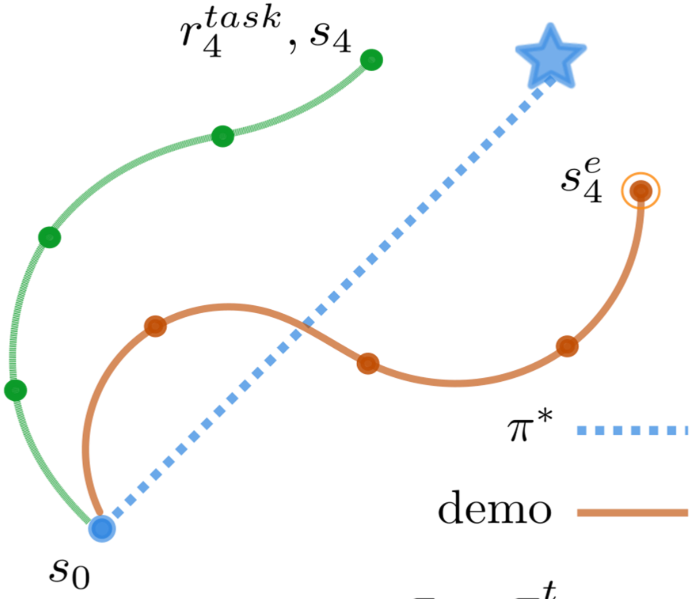
$\pi([s_t, \boxed{s_t^e}])$



D-Shape: Shaping reinforcement learning with a suboptimal demonstration trajectory

$$\cancel{\pi(s_t)}$$

$$\pi([s_t, \boxed{s_t^e}])$$

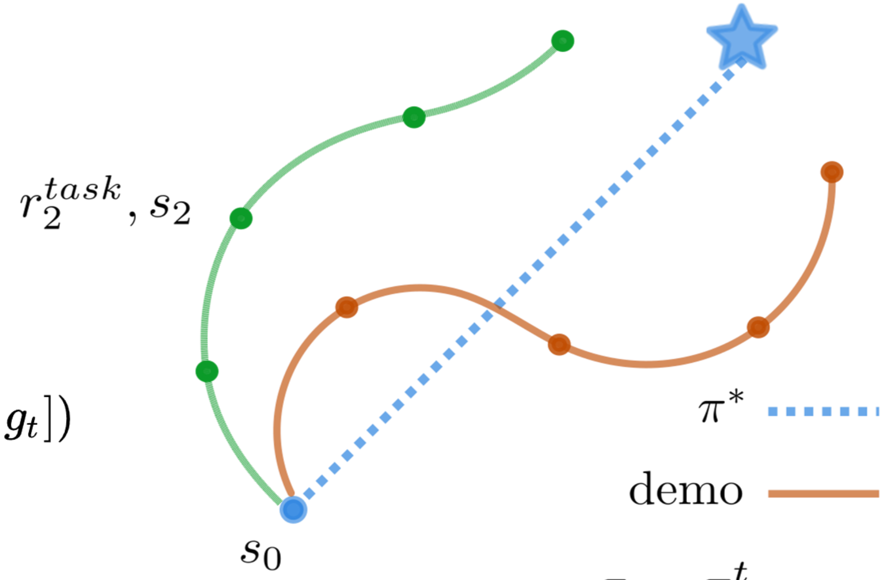


D-Shape: Shaping reinforcement learning with a suboptimal demonstration trajectory

$$r_t^{goal} = r_t^{task} + F_t^{goal}$$

$$F_t^{goal}([s_t, g_t, [s_{t+1}, g_{t+1}]]) = \gamma\phi([s_{t+1}, g_{t+1}]) - \phi([s_t, g_t])$$

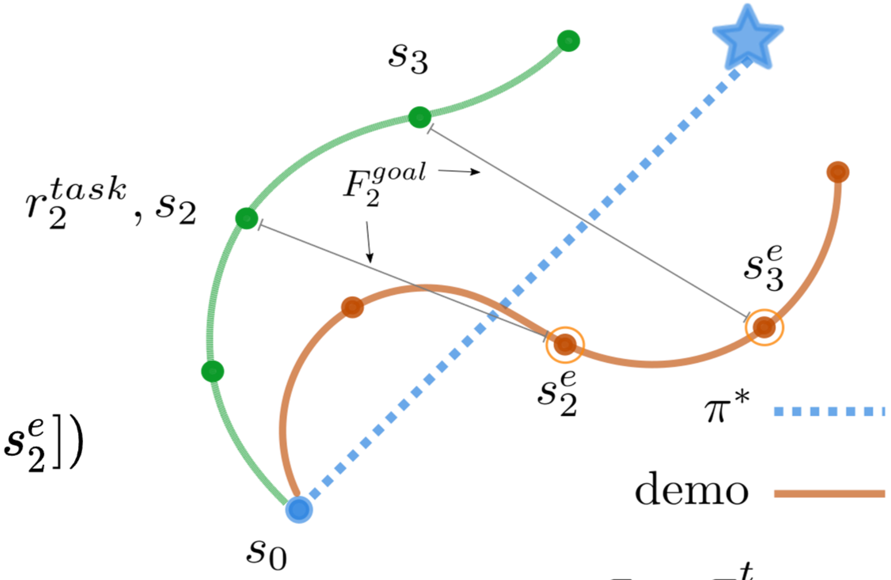
$$\phi([s_t, g_t]) = d(s_t, g_t)$$



D-Shape: Shaping reinforcement learning with a suboptimal demonstration trajectory

$$r_2^{goal} = r_2^{task} + F_2^{goal}$$

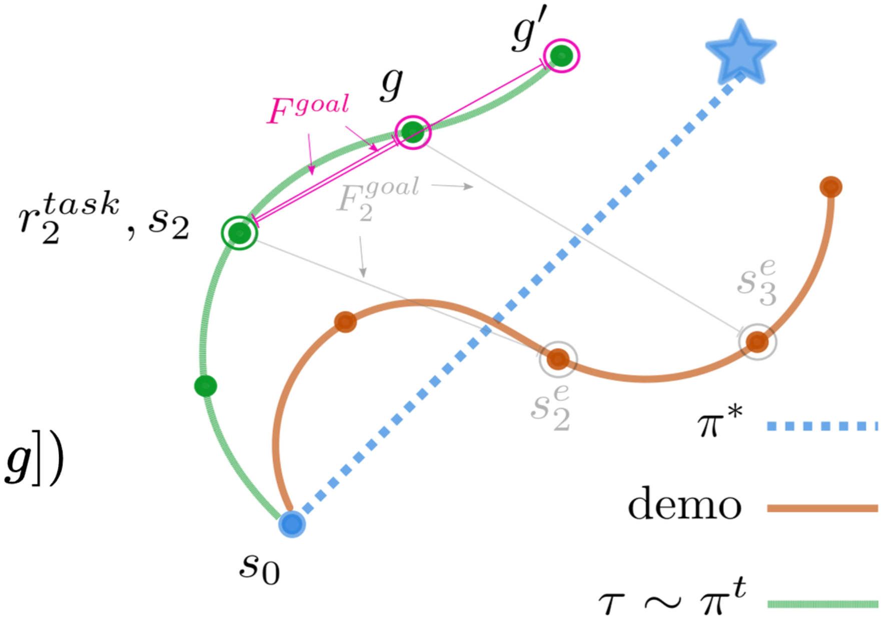
$$F_2^{goal}([s_2, s_2^e], [s_3, s_3^e])$$
$$= \gamma d([s_3, s_3^e]) - d([s_2, s_2^e])$$



D-Shape: Shaping reinforcement learning with a suboptimal demonstration trajectory

$$r_2^{goal} = r_2^{task} + F^{goal}$$

$$F^{goal}([s_2, g], [s_3, g']) = \gamma d([s_3, g']) - d([s_2, g])$$



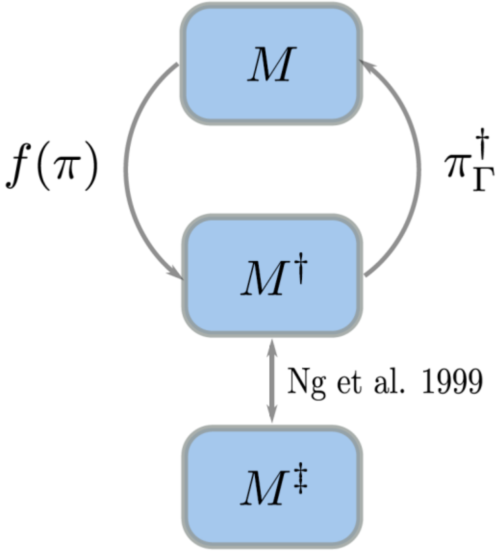
D-Shape: Shaping reinforcement learning with a suboptimal demonstration trajectory

Method Summary

- Demonstration states as goals
- Goal-reaching potential reward
- Goal relabelling with achieved states (Hindsight Experience Replay) [1]

Policy invariance guarantee

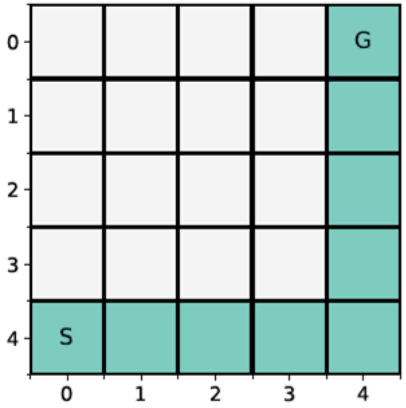
Theorem 1: An optimal goal-conditioned policy learned by D-Shape can be optimally executed with any sequence of goals.

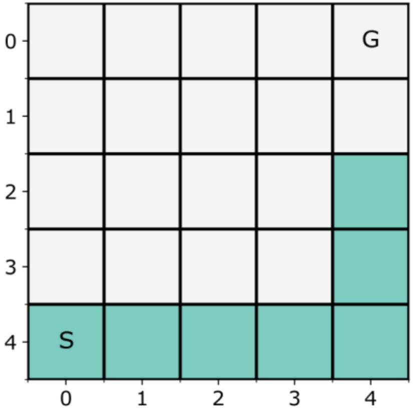


[1] Andrychowicz et al, Hindsight experience replay, Neurips 2017.

Experimental Setting

- Goal-based *s x s* gridworld, $s \in [10, 20, 30]$
- Baselines:
  - Q-learning [1]
  - SBS [2]
  - RIDM [3]
  - RL+ Manhattan distance reward
- Demonstrations: optimal, suboptimal
- Desiderata:
  - sample efficiency
  - convergence to optimal returns



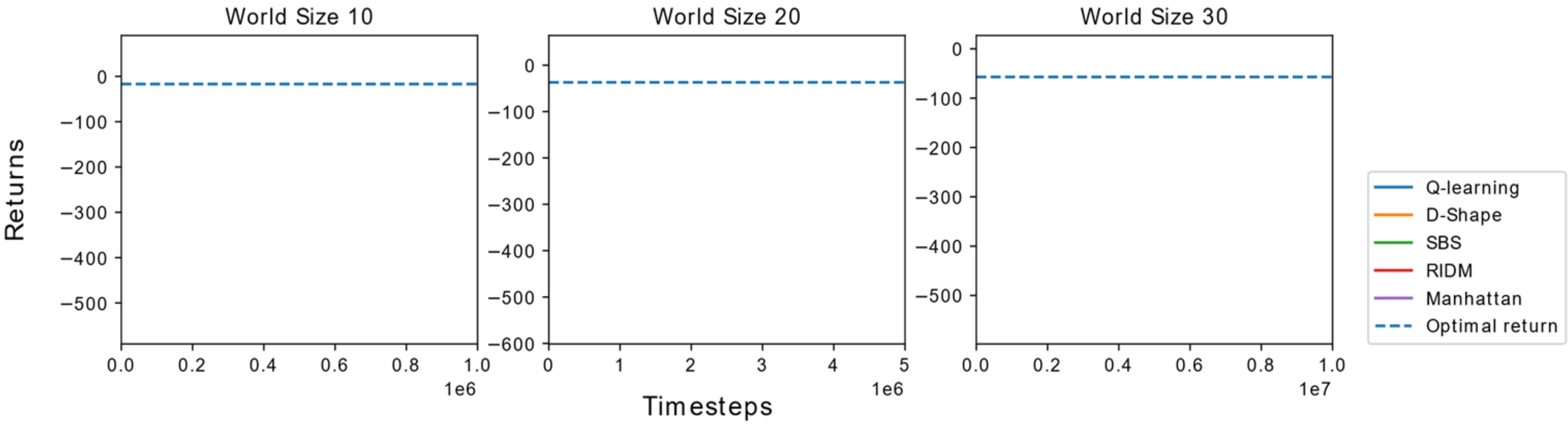


[1] Watkins, Learning from delayed rewards, PhD dissertation, 1989.

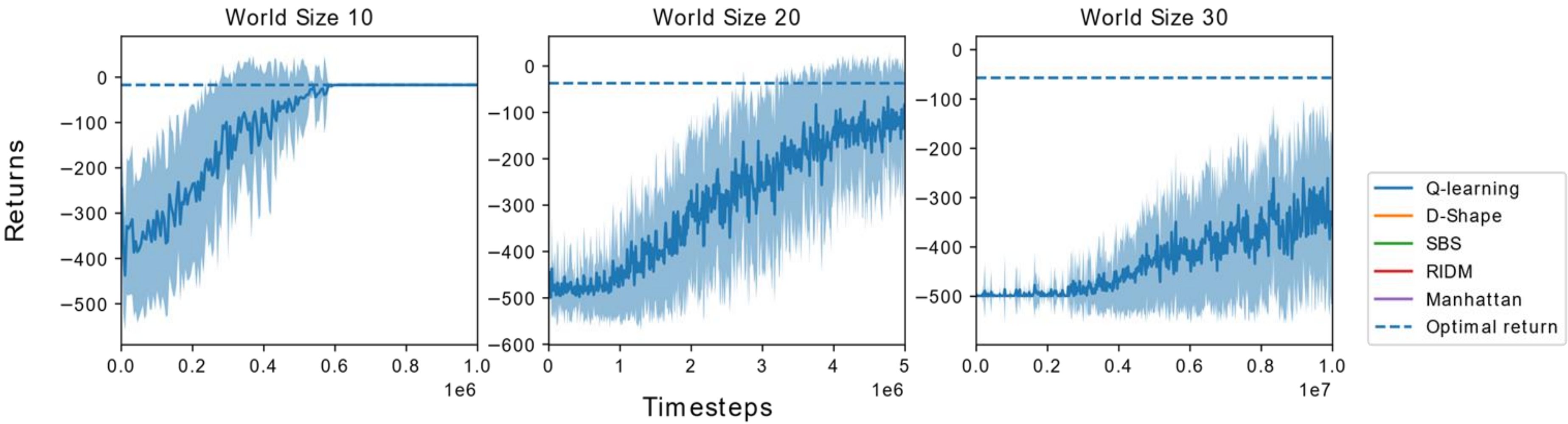
[2] Brys et al., Reinforcement learning from demonstration through shaping, IJCAI 2015.

[3] Pavse et al., RIDM: Reinforced inverse dynamics modelling for learning from a single observed demonstration, IROS 2020.

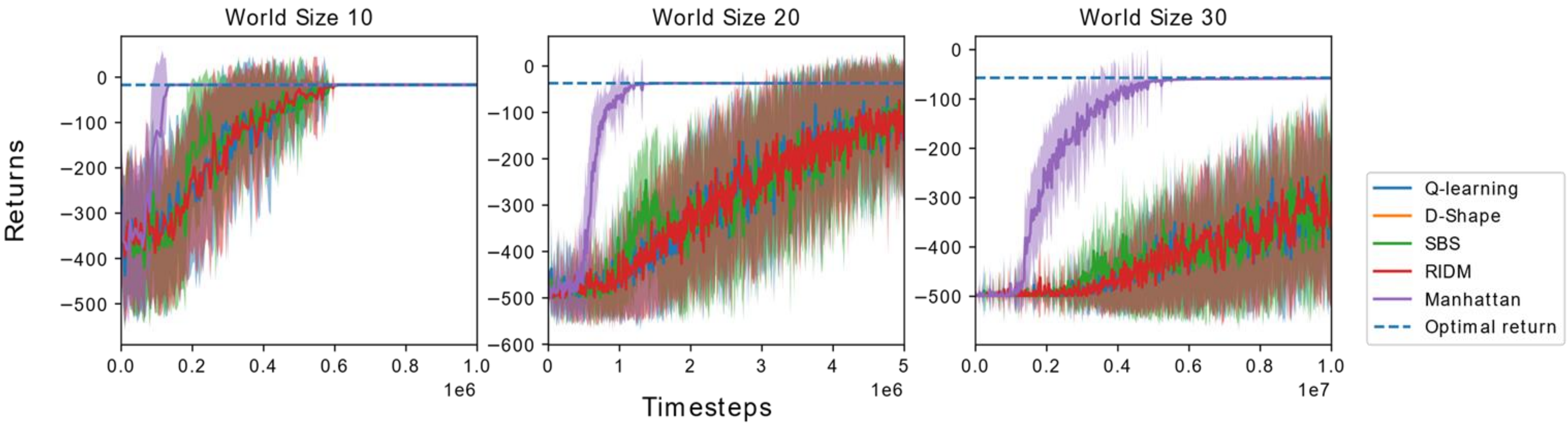
1. D-Shape improves sample efficiency



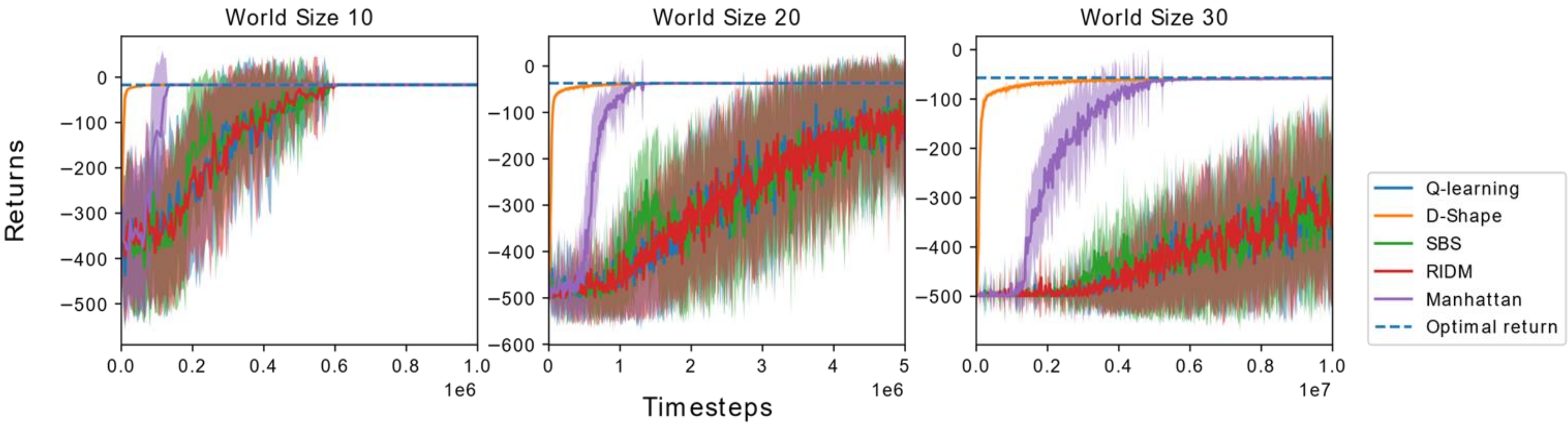
1. D-Shape improves sample efficiency



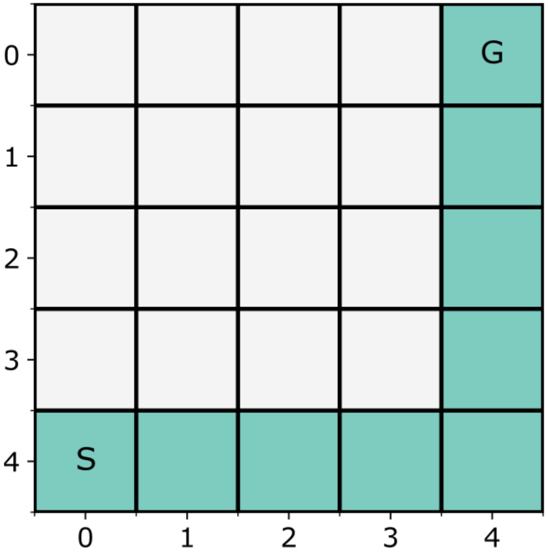
1. D-Shape improves sample efficiency

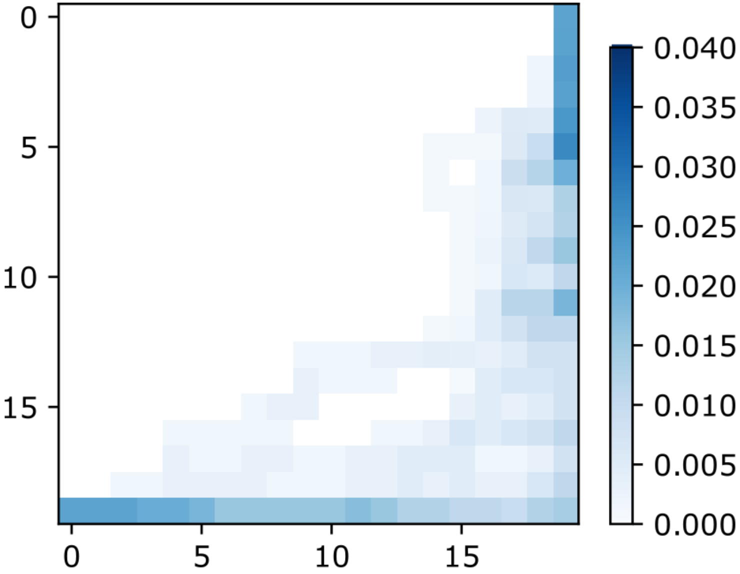


1. D-Shape improves sample efficiency



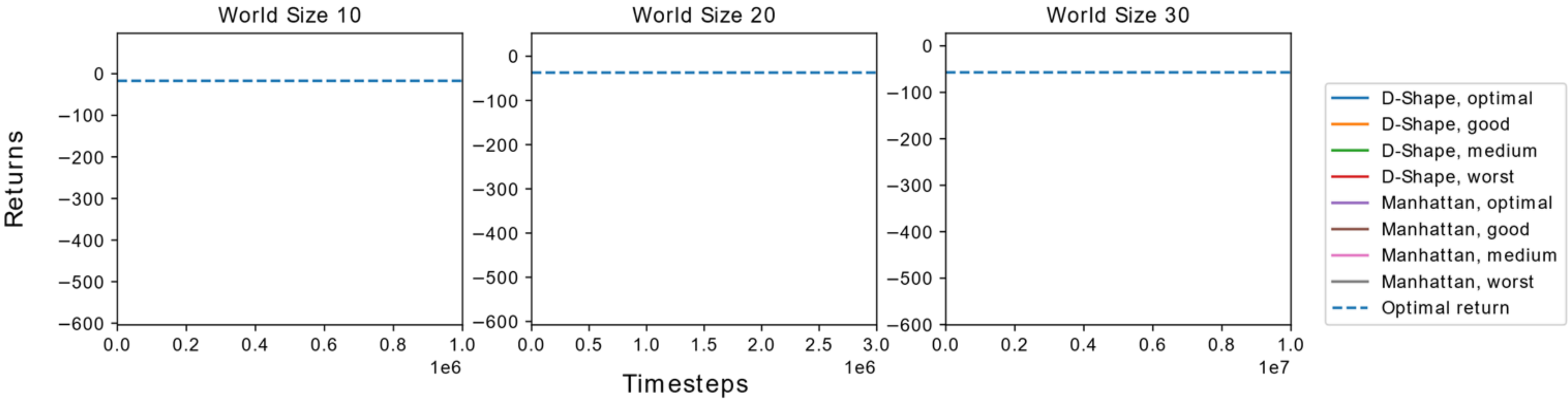
D-Shape State Visitation





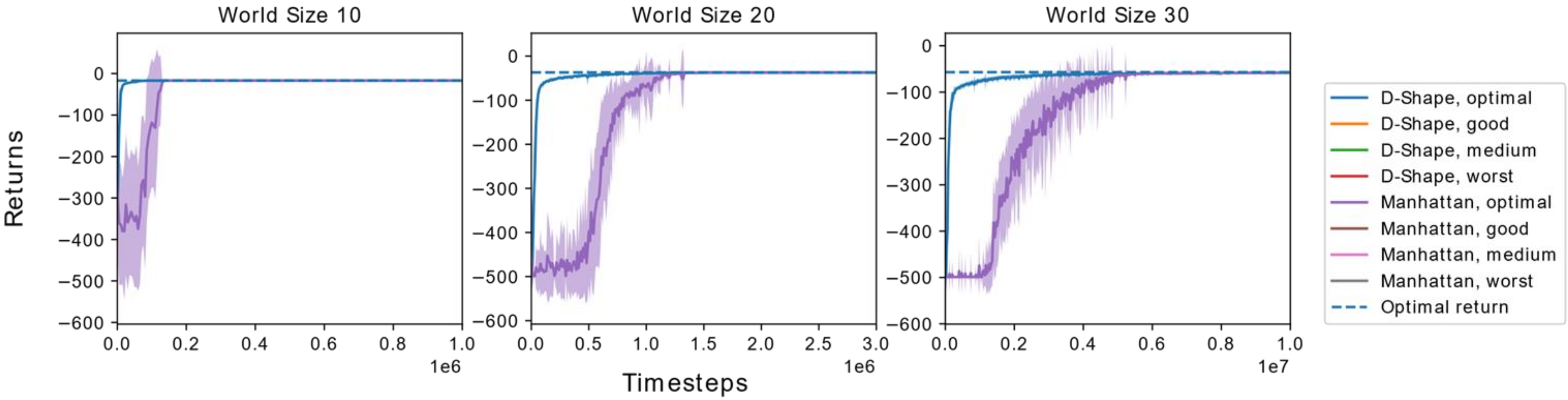
2. Learning with suboptimal demonstrations

Suboptimality Type I : demonstration trajectory goes to incorrect goal state



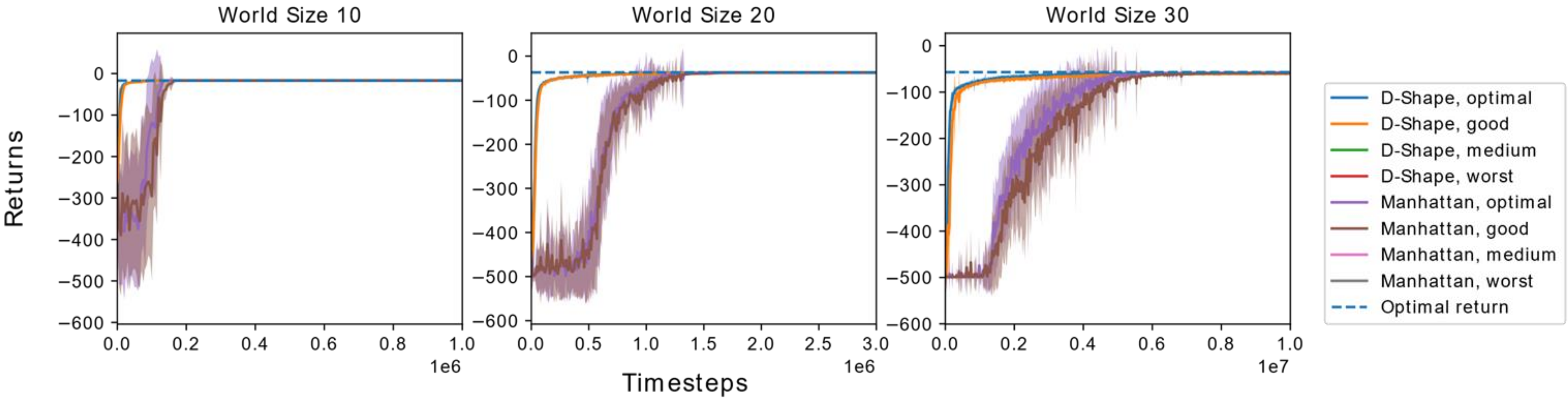
2. Learning with suboptimal demonstrations

Suboptimality Type I : demonstration trajectory goes to incorrect goal state



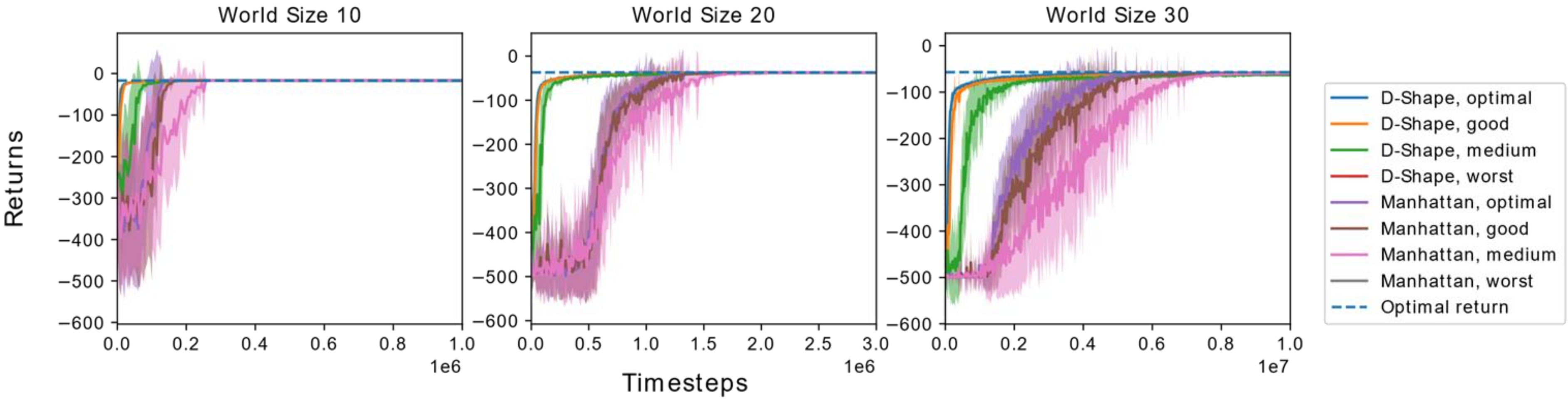
2. Learning with suboptimal demonstrations

Suboptimality Type I : demonstration trajectory goes to incorrect goal state



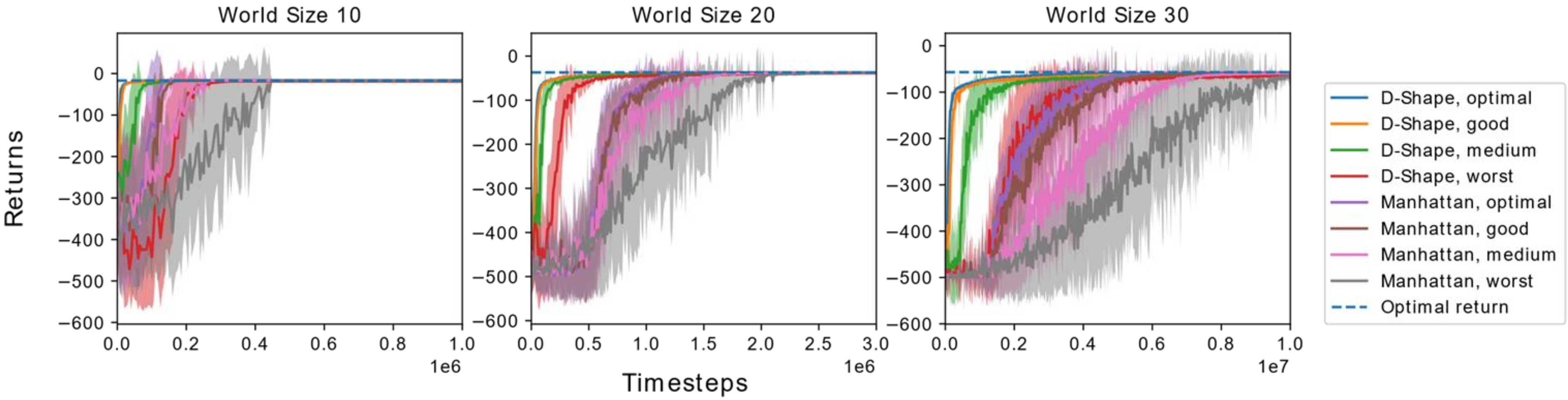
2. Learning with suboptimal demonstrations

Suboptimality Type I : demonstration trajectory goes to incorrect goal state



2. Learning with suboptimal demonstrations

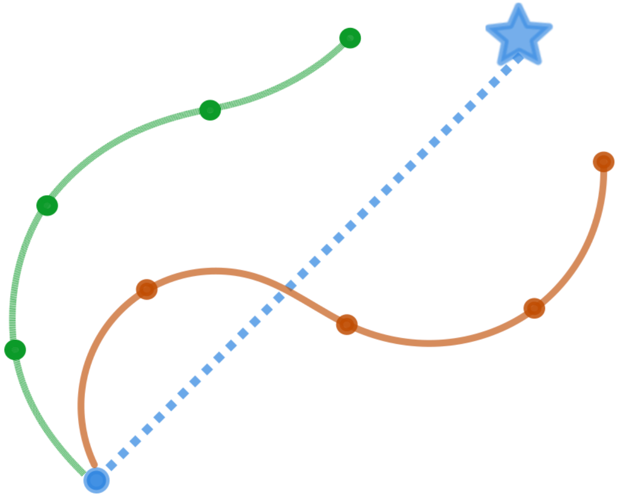
Suboptimality Type I : demonstration trajectory goes to incorrect goal state



Conclusions

- D-Shape accelerates reinforcement learning given access to a single state-only demonstration
- Future work:
    - Extending method to multiple demonstrations
    - Learned distance metrics for continuous state-action spaces
    - Exploring other GCRL techniques for RL + IL

Thanks for listening!



Caroline Wang
caroline.l.wang@utexas.edu

Peter Stone
pstone@cs.utexas.edu

Garrett Warnell
garrett.a.warnell.civ@army.mi

https://arxiv.org/abs/2210.14428

Related Works

- RL+IL
  - Constructing rewards with demonstrations
    - Annealing hybrid rewards: Ding et al. 2019; Zolna et al. 2019.
  - Plan based reward shaping w/demos: Brys et al. 2015; Suay et al. 2016; Wu et al. 2021.
  - Optimizing only the task reward:
    - State augmentation: Pavse et al. 2020; Paine et al. 2018.
    - Resetting: Salimans and Chen 2018; Ecoffet et al. 2021; Nair et al. 2018.
    - Initializing with demonstration information: Hester et al. 2018; Taylor et al. 2011.
- Accelerating goal-conditioned RL with demonstrations
  - Nair et al. 2018; Paul et al. 2019.

Citations (Related Work)

[1] Brys et al. 2015. Reinforcement Learning from Demonstration through Shaping. In IJCAI. IJCAI.
[2] Ding et al. 2019. Goal conditioned Imitation Learning. In NeurIPS.
[3] Ecoffet et al. 2021. First return, then explore. Nature 590 (2021)
[4] Hester et al. 2018. Deep Q-learning From Demonstrations. In AAAI.
[5] Nair et al. 2018. Overcoming Exploration in Reinforcement Learning with Demonstrations. In ICRA.
[6] Paine et al. 2018. One-Shot High Fidelity Imitation: Training Large-Scale Deep Nets with RL. ArXiv abs/1810.05017.
[7] Paul et al. 2019. Learning from Trajectories via Subgoal Discovery. In Neurips.
[8] Pavse et al. 2020. RIDM: Reinforced Inverse Dynamics Modeling for Learning from a Single Observed Demonstration. In IROS.
[9] Salimans and Chen 2018. Learning Montezuma’s Revenge from a Single Demonstration. In Workshop on Deep Reinforcement learning at NeurIPS.
[10] Suay et al. 2016. Learning from Demonstration for Shaping through Inverse Reinforcement Learning. In AAMAS.
[11] Taylor et al. 2011. Integrating reinforcement learning with human demonstrations of varying ability, In AAMAS.
[12] Wu et al. 2021. Shaping Rewards for Reinforcement Learning with Imperfect Demonstrations using Generative Models. In ICRA.
[13] Zolna et al. 2019. Reinforced Imitation in Heterogeneous Action Spaces, In Imitation Learning and its Challenges in Robotics Workshop at NeurIPS.