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

Caroline Wang¹, Garrett Warnell¹, ², Peter Stone¹, ³

¹The University of Texas at Austin, ²Army Research Laboratory, ³Sony 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 $\Rightarrow$ Potential conflict between IL and RL objectives!

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:  \( E_{\pi}[\sum_{t=0}^{H-1} \gamma^t r_{task}] \)

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

\[
D^e = \{s^e_t\}_{t=1}^H
\]

Background

- Potential-based reward shaping (PBRS) [1]:
  - Learning is conducted in modified MDP, where
    \[ M = (S, A, P, R' := r_{\text{task}} + F, \gamma) \]
  - 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. \( r^g_t = 1_{s_t=g} \)

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

\[
\pi(s_t)
\]

\[
\pi([s_t, s^e_t])
\]
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

$$\pi(s_t)$$
$$\pi(hs_t, se_t)$$

$$r^{task}_3, s_3$$
$$s^e_3$$
$$s_0$$

demo
$$\tau \sim \pi^t$$
D-Shape: Shaping reinforcement learning with a suboptimal demonstration trajectory
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

\[ r_{2}^{\text{goal}} = r_{2}^{\text{task}} + F_{2}^{\text{goal}} \]

\[ F_{2}^{\text{goal}} ([s_2, s_2^e], [s_3, s_3^e]) = \gamma d([s_3, s_3^e]) - d([s_2, s_2^e]) \]
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.

Experimental Setting

- Goal-based $s \times s$ 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. D-Shape improves sample efficiency

- World Size 10
- World Size 20
- World Size 30

Returns vs. Timesteps for different world sizes.
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
    - Plan based reward shaping w/demos: Brys et al. 2015; Suay et al. 2016; Wu et al. 2021.
  - Optimizing only the task reward:
    - 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)