D-Shape combines **reinforcement** and **imitation learning** for sample-efficient learning from a **single** demonstration with **optimality** guarantees.

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
Contact information: caroline.l.wang@utexas.edu; garrett.a.warnell.civ@army.mi; pstone@cs.utexas.edu

Overview

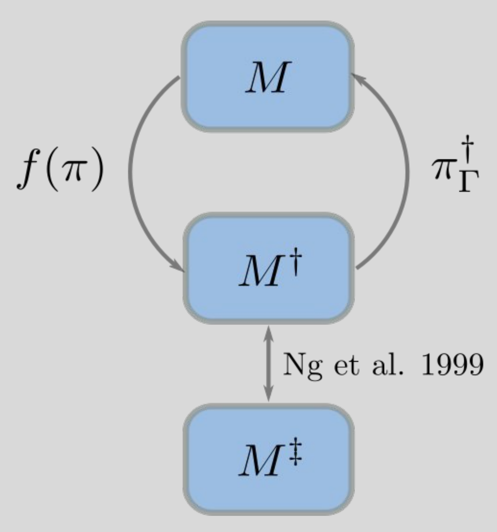
- Reinforcement learning (RL) discovers optimal behavior from a reward function but is sample inefficient
- Imitation learning (IL) learns behaviors from demonstration efficiently but usually requires multiple, optimal, state-action demonstrations
- Combining RL and IL is challenging due to conflicting objectives: cumulative task reward vs minimizing divergence from demonstration distribution
- D-Shape…
  - Only requires a **single**, **suboptimal**, **state-only** demonstration trajectory
  - **Improves sample efficiency** over RL alone
  - Preserves the **optimal policy**

Background

- Potential-based reward shaping (PBRS)
  - A method to alter the reward function such that the optimal policy is preserved (policy invariance)
- Goal-conditioned RL (GCRL)
  - Given a goal-reaching task, objective is to learn a goal-conditioned policy that can reach any goal g drawn from a goal-set G
  - Reward function is sparse-indicator for goal

Preserving the Optimal Policy

- Theorem: An optimal goal-conditioned policy learned by D-Shape can be optimally executed with any sequence of goals
- Key idea: view D-Shape as composition of modifications to base MDP (M), goal relabelling (M†), and PBRS (M‡)



Experimental Setting

- Goal-based s x s gridworld, $s \in [10, 20, 30]$, goal G
- Baselines [1, 2, 3]
  - Q-learning [1]
  - SBS [2]
  - RIDM [3]
  - RL + Manhattan distance
- Demonstrations: varying degrees of suboptimality

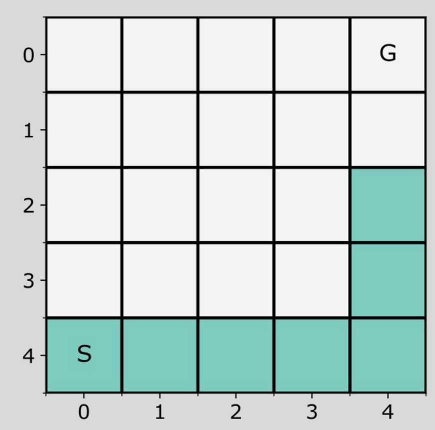


Figure 1. Example suboptimal demonstration on 5x5 gridworld.

[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.

Experimental Results

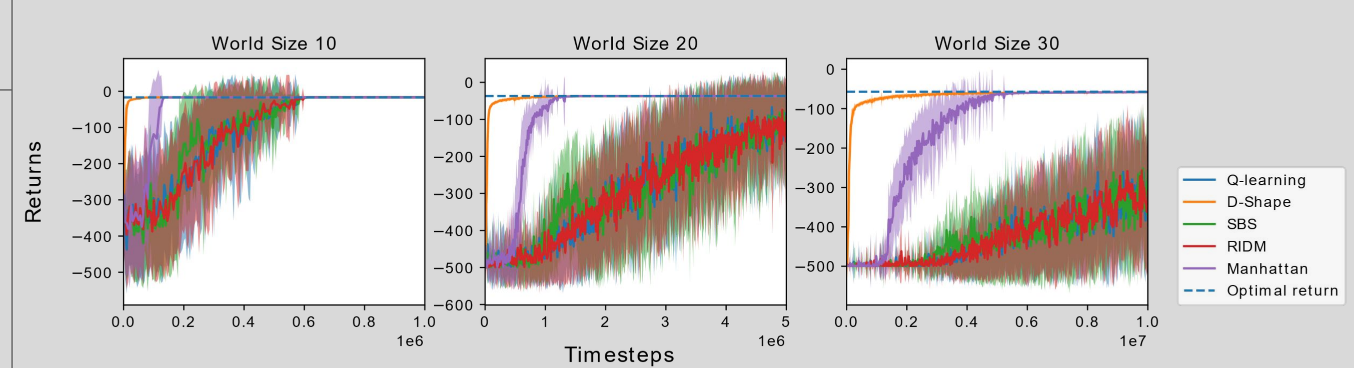


Figure 2. D-Shape improves sample efficiency over baselines.

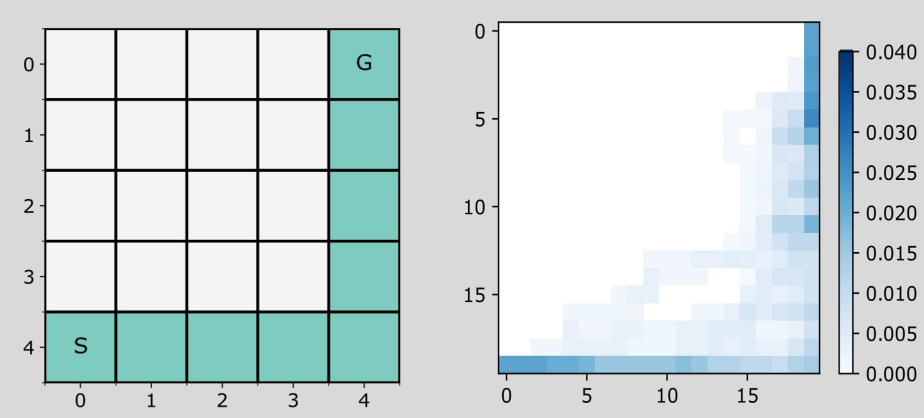


Figure 3. (Left) Optimal demonstration style. (Right) State visitation of D-Shape given optimal demonstration.

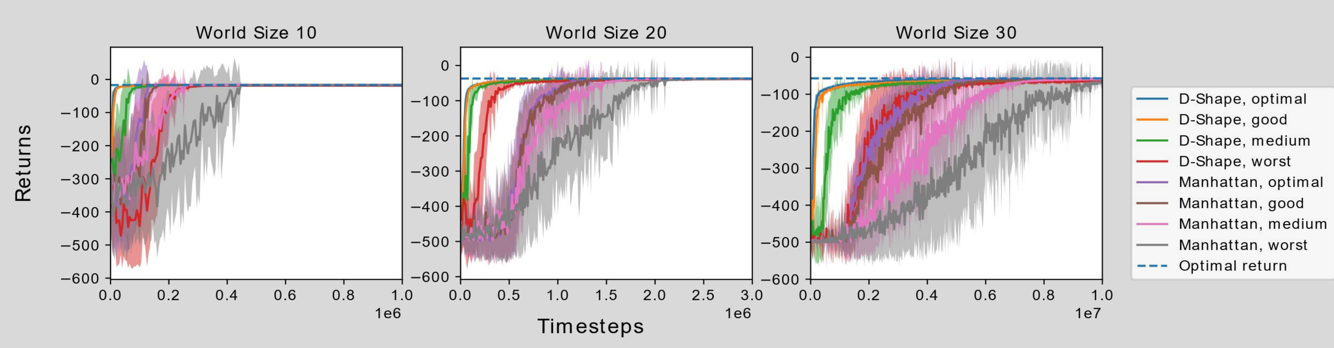
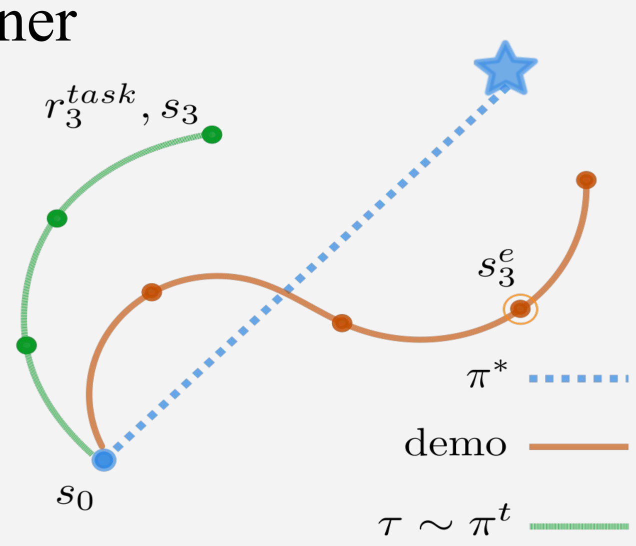


Figure 4: D-Shape converges to optimal return with high sample efficiency despite suboptimal demonstrations.

D-Shape Walkthrough

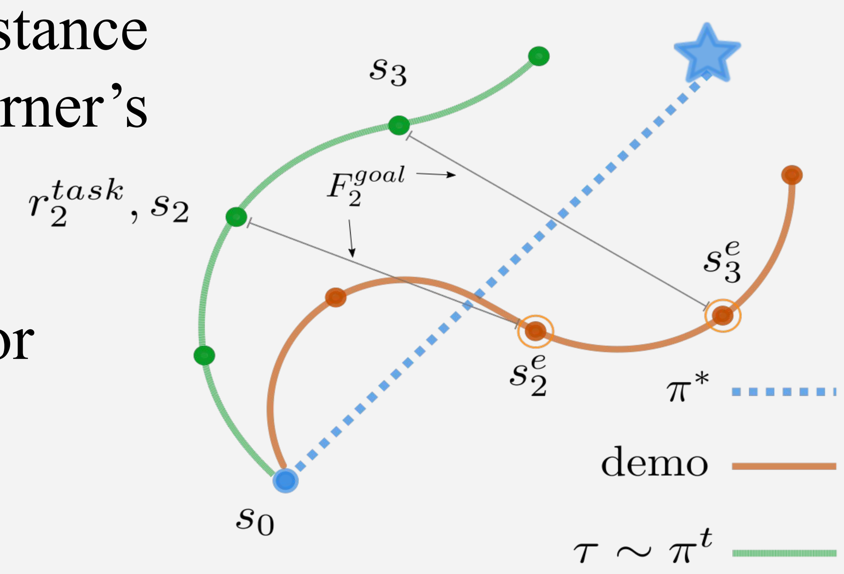
1. <u>**Key idea**</u>: shape exploration of reinforcement learner towards demonstration trajectory by **treating demonstration states as goals.**



2. D-Shape learner's state space consists of the **current state and next demonstrator state.**

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

3. **Goal-reaching potential reward** based on distance between learner's achieved state and demonstrator goal state.

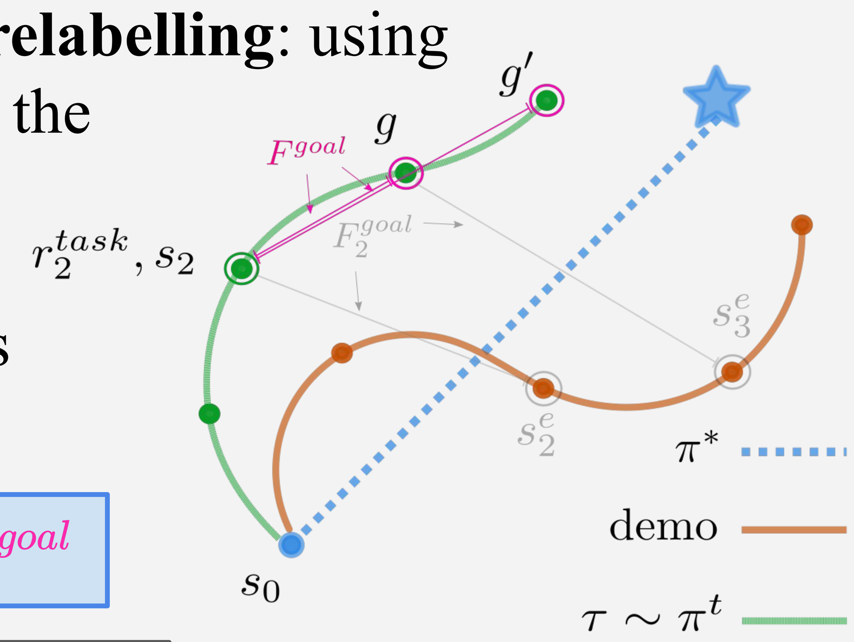


$$\boxed{r_t^{goal} = r_t^{task} + F_t^{goal}}$$

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

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

4. **Hindsight relabelling**: using states along the learner's achieved trajectory as goals.



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

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

Scan to view full paper

TEXAS
The University of Texas at Austin

DEVCOM
ARMY RESEARCH LABORATORY