Plan-Based Relaxed Reward Shaping for Goal-Directed Tasks

Ingmar Schubert[1], Ozgur S. Oguz[2,3], and Marc Toussaint[1,2]

[1]Learning and Intelligent Systems Group, TU Berlin, Germany

[2]Max Planck Institute for Intelligent Systems, Stuttgart, Germany

[3]University of Stuttgart, Germany

ICLR 2021

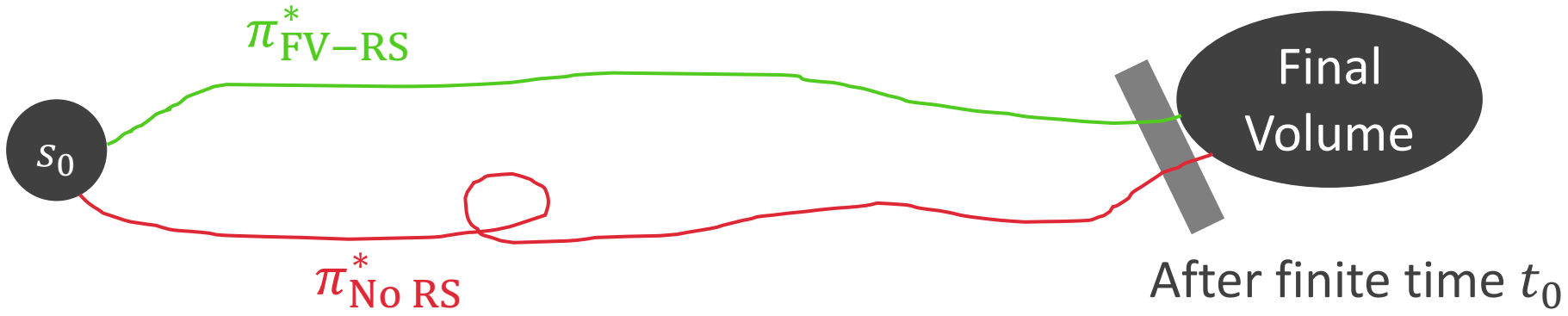
RL with sparse rewards is limited by exploration. This can be adressed by Reward Shaping.

$$R(s, a, s') \leftarrow R(s, a, s') + F(s, a, s')$$

Potential-Based Reward Shaping (PB-RS) [Ng et al., 1999]	Final-Volume-Preserving Reward Shaping (FV-RS, ours)
$F_{\text{PB}-\text{RS}}(s, a, s') = \gamma\Phi(s') - \Phi(s)$	Allows for more general $F_{\text{FV}-\text{RS}}$
Guarantees invariance of the optimal policy	Does not guarantee invariance of the optimal policy
$\pi^*_{\text{No RS}} \equiv \pi^*_{\text{PB}-\text{RS}}$	$\pi^*_{\text{No RS}} \not\equiv \pi^*_{\text{FV}-\text{RS}}$

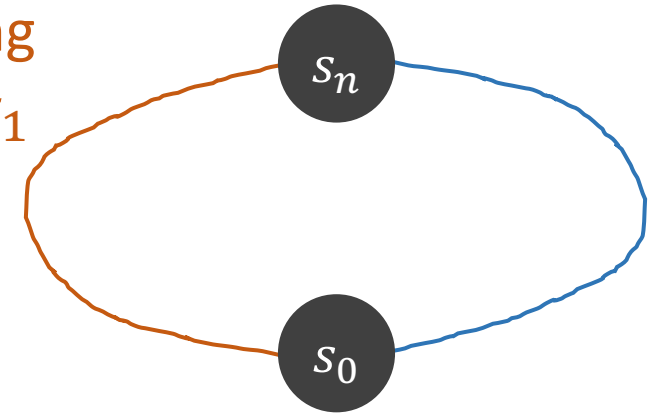
FV-RS relaxes the invariance guarantee of PB-RS to a guarantee of invariant long-term behavior

FV-RS (ours) does not leave the optimal policy unchanged.
Instead, FV-RS only leaves the long-term state of the MDP unchanged

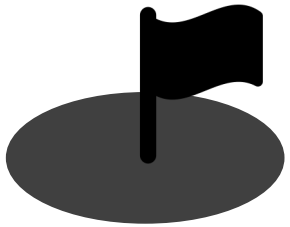


As a result, FV-RS can introduce information in a more direct way than PB-RS

Path when following policy π_1



Path when following policy π_2



Goal region with nonzero reward

- PB-RS: $V_{\pi_1}(s_0) = V_{\pi_2}(s_0) = \gamma^n \Phi(s_n) - \Phi(s_0)$
- FV-RS: $V_{\pi_1}(s_0) > V_{\pi_2}(s_0)$

With PB-RS, a policy's value only depends on the initial and final state.
With FV-RS, a policy's value can depend on all states along the trajectory.

Experiments

We compare in a plan-based setting:

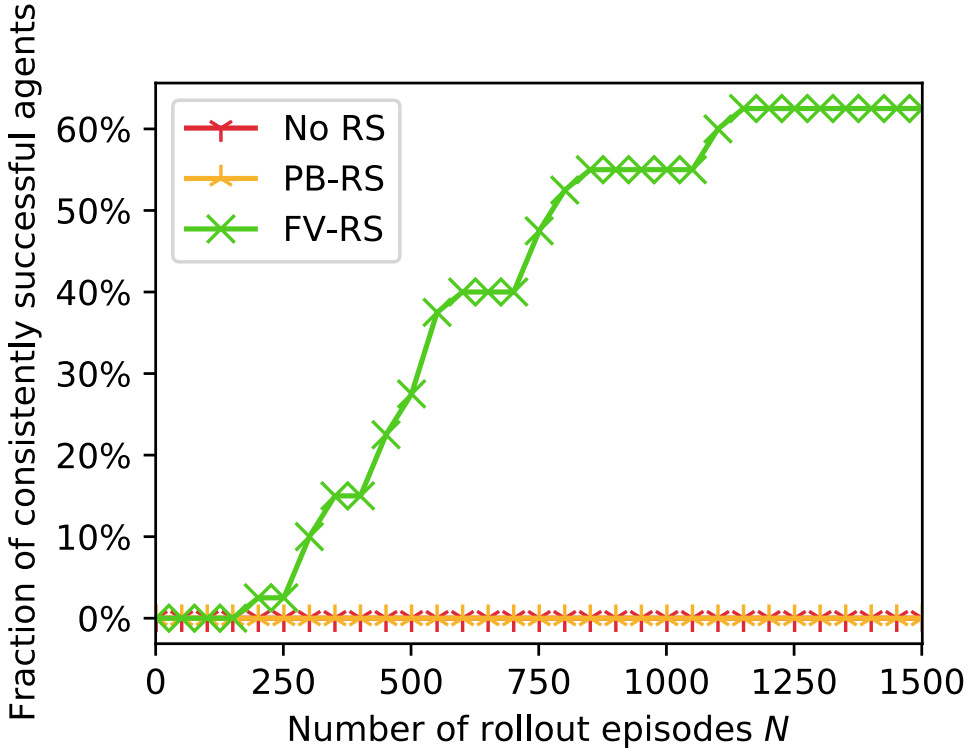
1. Potential-Based reward shaping (PB-RS)
2. Final-Volume-Preserving reward shaping (FV-RS, ours)

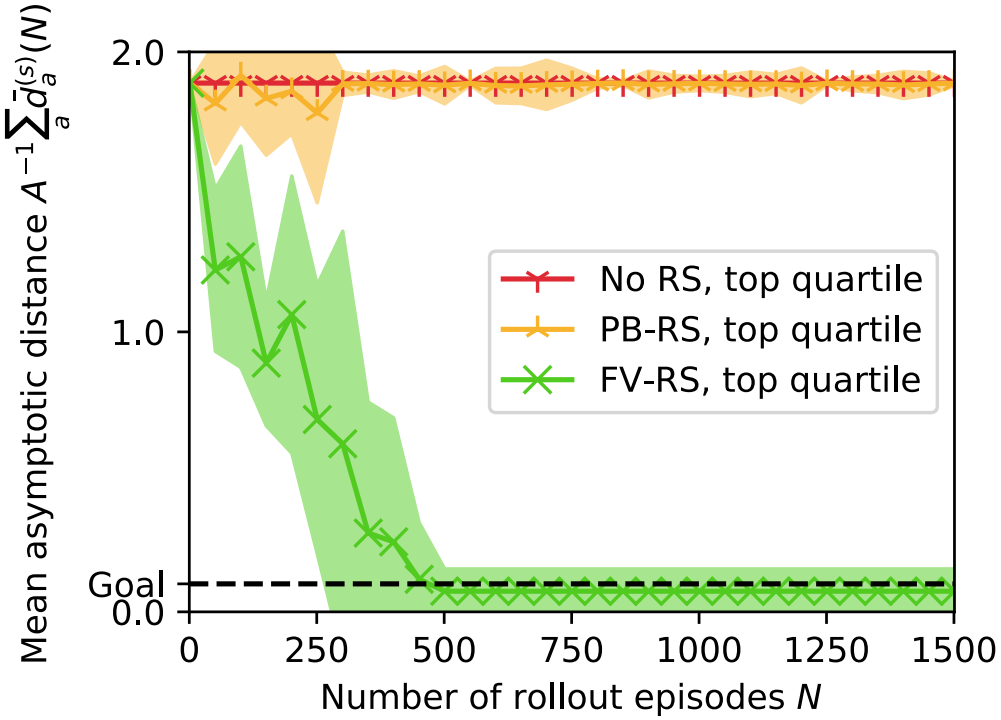
Using simulated robotic manipulation examples:

1. Pushing task (4 examples)
2. Pick-and-place task (2 examples)

Pushing Task: Example 1

FV-RS (ours) increases sample efficiency significantly over PB-RS





Improved efficiency of FV-RS is consistent…

- …across all examples
- …when using different RL algorithms (DDPG[1] and PPO[2])
- …when using different shaping functions for FV-RS and PB-RS

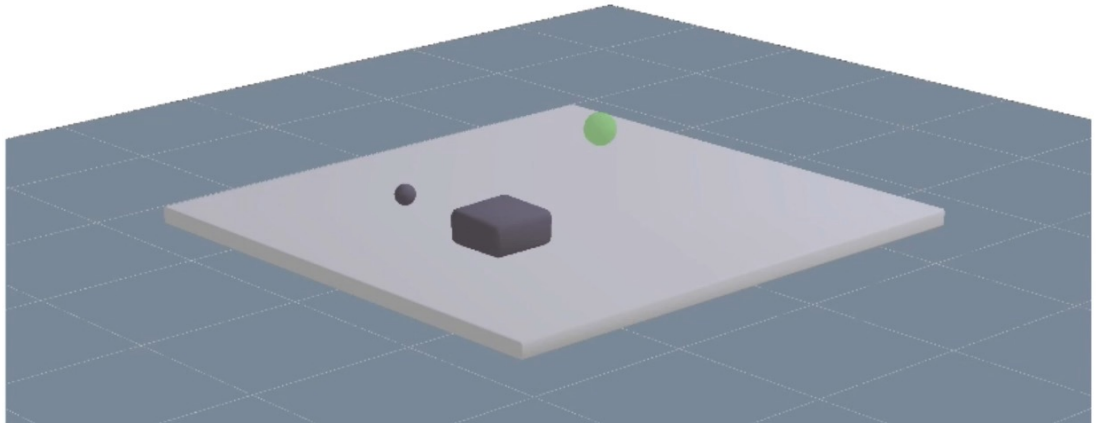
1 Lillicrap et al., 2015
2 Schulman et al., 2017

Summary

- We introduce FV-RS
- We propose to use FV-RS for plan-based reward shaping
- We demonstrate the increased sample efficiency of FV-RS over PB-RS in various plan-based robotic manipulation tasks

Join us at poster session 2 (May 3, 9 a.m. PDT)

Task 1:
Best agent using PB-RS



Task 1:
Best agent using FV-RS (ours)