RUNE: Reward Uncertainty for Exploration in Preference-based Reinforcement Learning

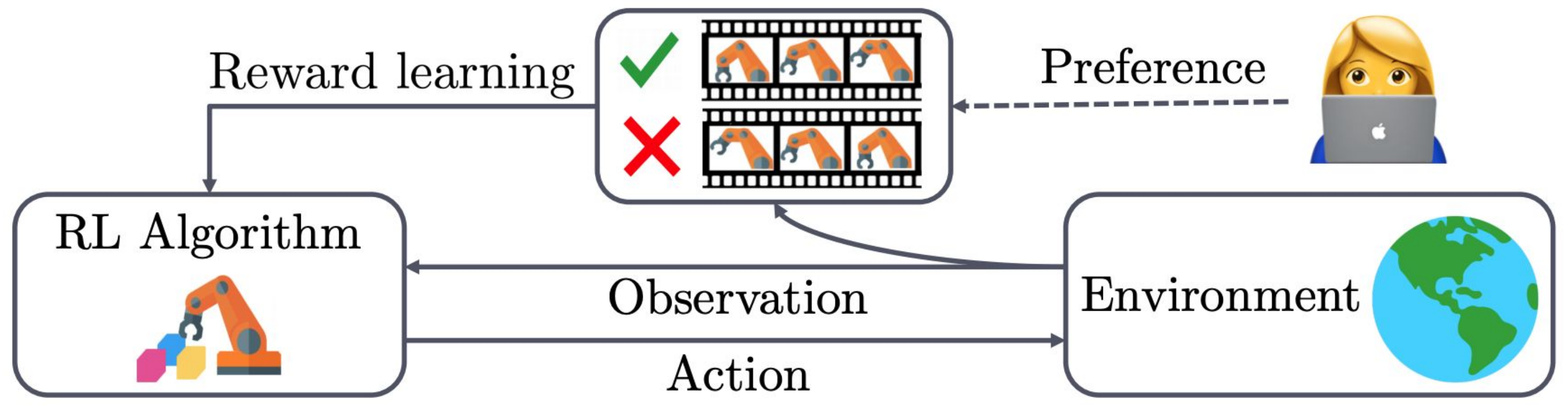
Xinran Liang, Katherine Shu,
Kimin Lee*, Pieter Abbeel*

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH
ICLR
ICLR 2022

Introduction

Background: Preference-based RL

Reward engineering is challenging for many complex tasks in real world [1]. Preference-based RL provides an alternative to resolve this challenge [2]. Human teacher provides preferences between the two behaviors. RL agent utilizes human feedback to learn desired behaviors.



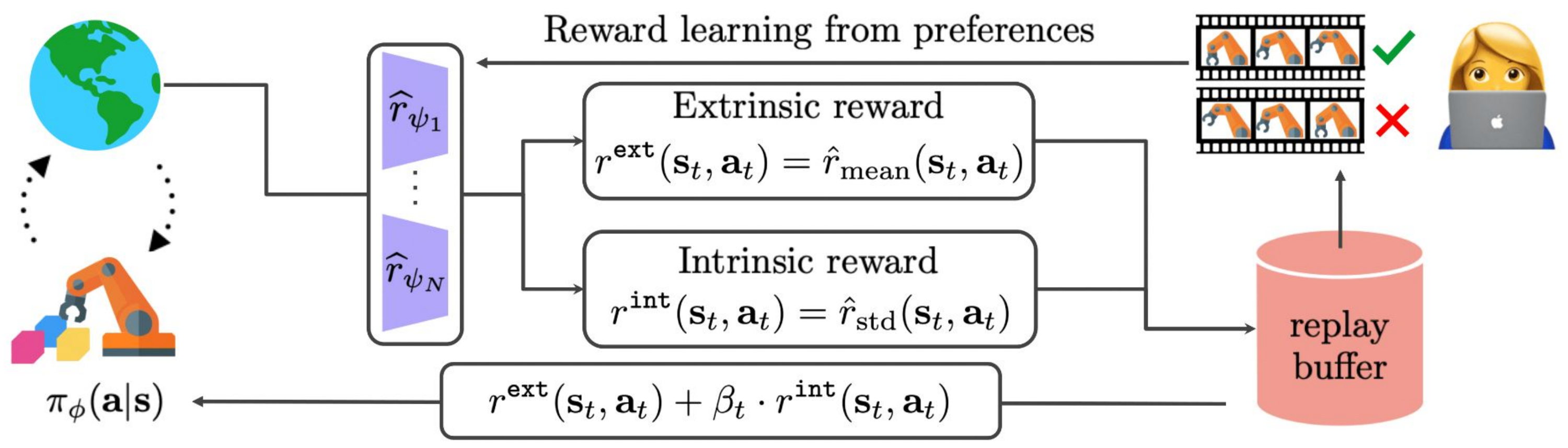
Main research question: How can we improve sample- and feedback-efficiency of preference-based RL?

Human feedbacks are usually expensive and time-consuming to collect.

Main method: RUNE

Overview

In this paper, we present a human-guided exploration method that is based on uncertainty in teacher preferences. Our intuition is that disagreement in the ensemble of learned reward functions measures the level of uncertainty from human preferences.


Reward learning from preferences
$\widehat{r}_{\psi_1}$
$\widehat{r}_{\psi_N}$
Extrinsic reward
$r^{\text{ext}}(\mathbf{s}_t, \mathbf{a}_t) = \hat{r}_{\text{mean}}(\mathbf{s}_t, \mathbf{a}_t)$
Intrinsic reward
$r^{\text{int}}(\mathbf{s}_t, \mathbf{a}_t) = \hat{r}_{\text{std}}(\mathbf{s}_t, \mathbf{a}_t)$
✓
✗
replay buffer
$r^{\text{ext}}(\mathbf{s}_t, \mathbf{a}_t) + \beta_t \cdot r^{\text{int}}(\mathbf{s}_t, \mathbf{a}_t)$
$\pi_\phi(\mathbf{a}|\mathbf{s})$

- We learn an ensemble of reward functions from human preferences.
- We use average predictions of the ensemble as extrinsic rewards from the task environment.
- We use standard deviation in predictions of the ensemble as intrinsic rewards to drive exploration.
- In off-policy RL training, we maximize weighted sum of extrinsic and intrinsic reward.

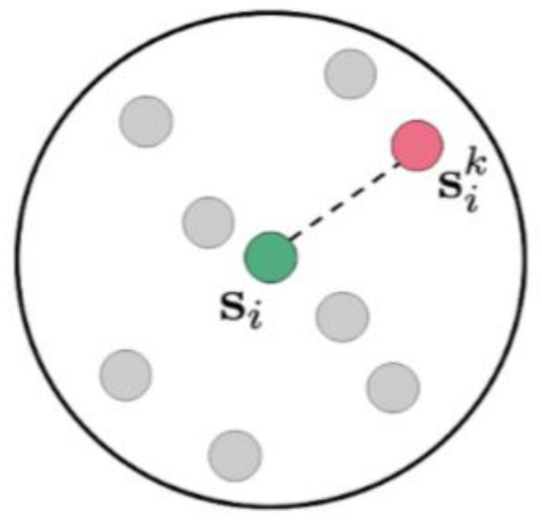
Experimental results

Experimental setups

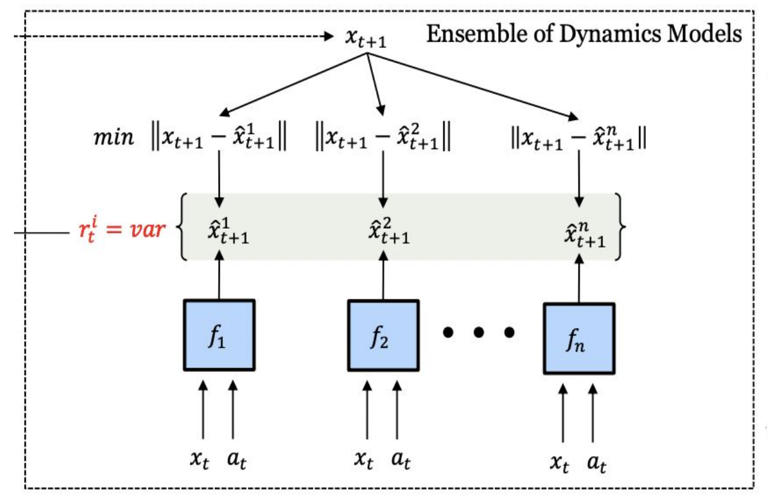
We consider robot manipulation tasks in Meta-World benchmark. For baseline method for comparison, we use preference-based RL algorithm PEBBLE [3]. We train an ensemble of 3 reward functions in all experiments.

Improving sample-efficiency

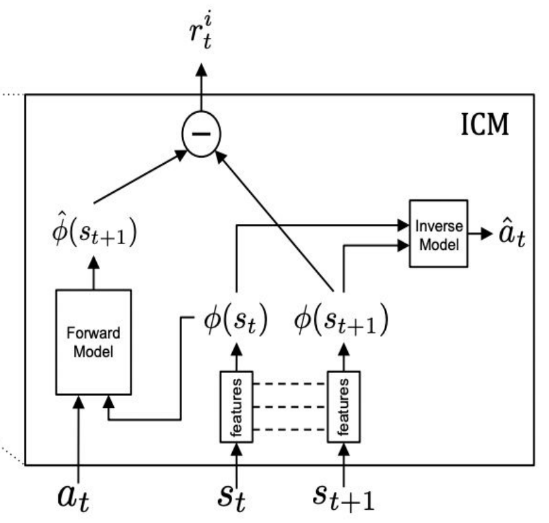
We compare learning curves of our proposed approach RUNE with following existing exploration method baselines:



StateEnt [4]

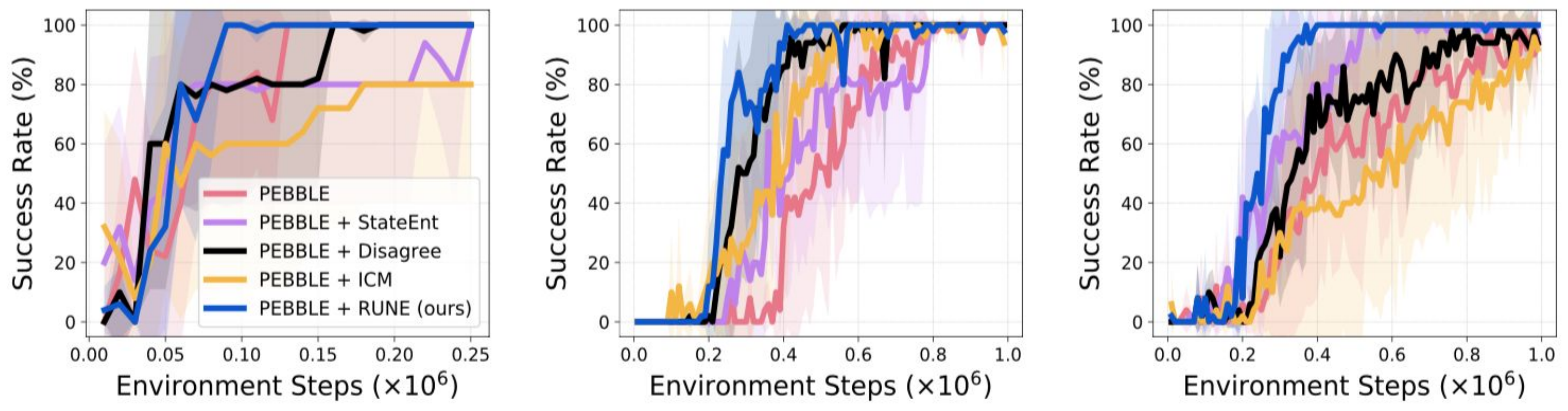

x_{t+1}
Ensemble of Dynamics Models
$min \; \|x_{t+1} - \hat{x}^1_{t+1}\|$
$\|x_{t+1} - \hat{x}^2_{t+1}\|$
$\|x_{t+1} - \hat{x}^n_{t+1}\|$
$r_t^i = var$
$\hat{x}^1_{t+1}$
$\hat{x}^2_{t+1}$
$\hat{x}^n_{t+1}$
f_1
f_2
f_n
$x_t \; a_t$

Disagreement [5]



ICM [6]

RUNE achieves consistently better sample-efficiency than PEBBLE baseline and other exploration baselines, using same total number of feedback queries during training.

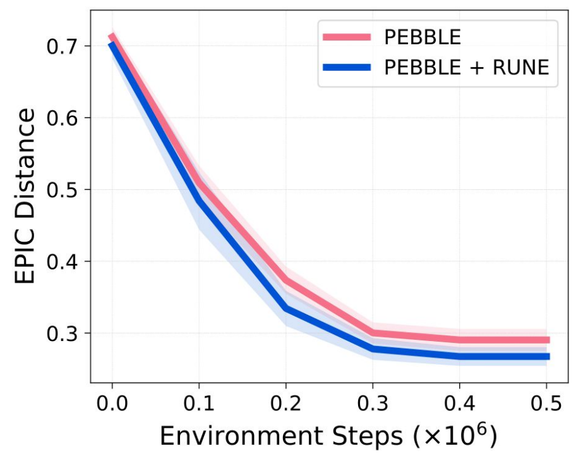


(a) Door Close (feedback = 1K) (b) Door Open (feedback = 5K) (c) Drawer Open (feedback = 10K)

Learning curves on robot manipulation tasks in Meta-World benchmarks as measured on the success rate (%). The solid and shaded regions represent means and standard deviations, respectively, across five runs.

Ablation study

Quality of learned reward functions



Compared to PEBBLE [3], RUNE achieves better reward learning, i.e. learned reward functions are closer to ground truth reward as measured by Equivalent-Policy Invariant Comparison (EPIC) distance [7].

Improving feedback-efficiency

We compare performance of RUNE and PEBBLE baseline using different budgets of total feedback queries during training. We use asymptotic success rate evaluated at the end of training as evaluation metrics.

Task	Feedback Queries	Method	Convergent Success Rate
Drawer Open	10000	PEBBLE	0.98 ± 0.08
		PEBBLE + RUNE	$\mathbf{1 \pm 0}$
	5000	PEBBLE	0.94 ± 0.08
		PEBBLE + RUNE	$\mathbf{0.99 \pm 0.02}$
Sweep Into	10000	PEBBLE	0.8 ± 0.4
		PEBBLE + RUNE	$\mathbf{1 \pm 0}$
	5000	PEBBLE	0.8 ± 0.08
		PEBBLE + RUNE	$\mathbf{0.9 \pm 0.14}$
Door Unlock	5000	PEBBLE	0.66 ± 0.42
		PEBBLE + RUNE	$\mathbf{0.8 \pm 0.4}$
	2500	PEBBLE	0.64 ± 0.45
		PEBBLE + RUNE	$\mathbf{0.8 \pm 0.4}$
Door Open	4000	PEBBLE	1 ± 0
		PEBBLE + RUNE	1 ± 0
	2000	PEBBLE	0.9 ± 0.2
		PEBBLE + RUNE	$\mathbf{1 \pm 0}$
Door Close	1000	PEBBLE	1 ± 0
		PEBBLE + RUNE	1 ± 0
	500	PEBBLE	0.8 ± 0.4
		PEBBLE + RUNE	$\mathbf{1 \pm 0}$
Window Close	1000	PEBBLE	0.94 ± 0.08
		PEBBLE + RUNE	$\mathbf{1 \pm 0}$
	500	PEBBLE	0.86 ± 0.28
		PEBBLE + RUNE	$\mathbf{0.99 \pm 0.02}$
Button Press	20000	PrefPPO	0.46 ± 0.20
		PrefPPO + RUNE	$\mathbf{0.64 \pm 0.18}$
	10000	PrefPPO	0.35 ± 0.31
		PrefPPO + RUNE	$\mathbf{0.51 \pm 0.27}$

RUNE consistently converges to better asymptotic success rate than PEBBLE and PrefPPO baseline [3] with different budgets of total feedback queries. For each experiment, we report means and standard deviations across five runs, respectively.

References

[1] Wu, Jeff, et al. "Recursively summarizing books with human feedback." *arXiv preprint arXiv:2109.10862* (2021).
[2] Christiano, Paul, et al. "Deep reinforcement learning from human preferences." *arXiv preprint arXiv:1706.03741* (2017).
[3] Lee, Kimin, et al. "B-Pref: Benchmarking Preference-Based Reinforcement Learning." *arXiv preprint arXiv:2111.03026* (2021).
[4] Liu, Hao, and Pieter Abbeel. "Behavior from the void: Unsupervised active pre-training." *arXiv preprint arXiv:2103.04551* (2021).
[5] Pathak, Deepak, Dhiraj Gandhi, and Abhinav Gupta. "Self-supervised exploration via disagreement." *International conference on machine learning*. PMLR, 2019.
[6] Pathak, Deepak, et al. "Curiosity-driven exploration by self-supervised prediction." *International conference on machine learning*. PMLR, 2017.
[7] Gleave, Adam, et al. "Quantifying differences in reward functions." *arXiv preprint arXiv:2006.13900* (2020).