Data Driven Reward Initialization for Preference based Reinforcement Learning

Mudit Verma, Subbarao Kambhampati
School of Computing & AI, Arizona State University
{muditverma, rao} @ asu.edu

R2HCAI, AAAI 2023

Objectives

Preference Based Reinforcement Learning has gained immense popularity for several tacit tasks however even State of the art algorithms have high variance in agent performance.

We investigate reasons for this high variance and tie it to high bias during reward initialization.

We show that this bias occurs with popular initialization techniques like Xavier-Uniform and Orthonormal and present a simple data driven initialization method to mitigate this.

PbRL Objective

- **Approximating the Human Reward Model**

> **Objective:** Compute probability for the human preferring trajectory τ_1 over τ_2:

$$P_\psi[\tau_0 \succ \tau_1] = \frac{\exp(\sum_t R_h(s_t^0, a_t^0))}{\sum_{i \in \{0,1\}} \exp(\sum_t R_h(s_t^i, a_t^i))}$$

> **Objective:** The human reward model can then by learnt by minimizing the cross-entropy between the predictions made by the supervised learner and the ground truth human labels as follows:

$$\mathcal{L}_{CE} = - \underset{(\tau_0, \tau_1, y) \sim \mathcal{D}}{\mathbb{E}} [y(0) \log P_\psi[\tau_0 \succ \tau_1] + y(1) \log P_\psi[\tau_1 \succ \tau_0]]$$

Motivation & Method

- **Issue of Reward Initialization**

> Reward functions already suffer from the issue of degenracy. While this can be a source of high variance in the obtained policy, we find that initial bias of reward functions can also have unintended impacts the agent performance.

> We find that data-independent initialization techqniues are limited to ensuring that only the parameters of the reward model are uniform, while the co-domain of the reward model is still “patchy” and biased.

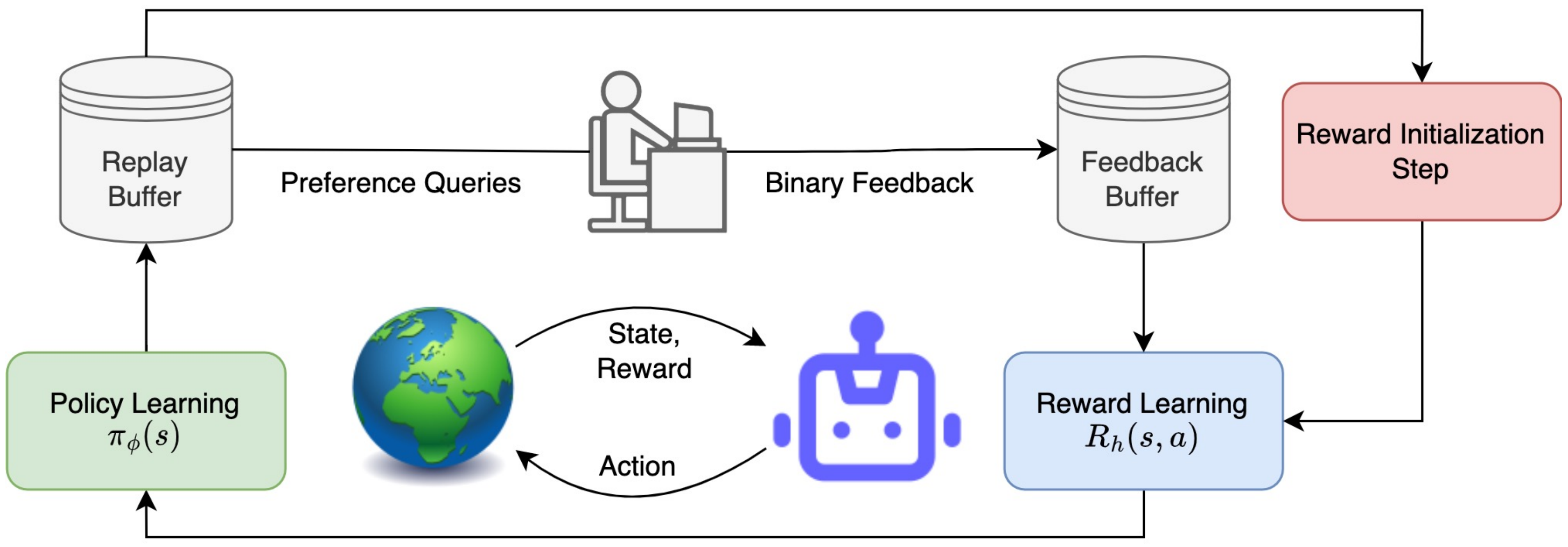
- **Proposed Data Driven Reward Initialization**

> Pre-training of the Reinforcement Learning agent is common in PbRL setups like PEBBLE and SURF.

> We make use of the pretraining phase and the randomly collected trajectories to collect “diverse” state-action samples.

> We can then simply ensure that the reward model predicts some pre-defnined constant over all these states.

$$\mathcal{L}_P = \sum_{\tau_j \sim D_\tau} \sum_{i=1}^{i=|\tau_j|} \left\| R^h_\psi(s_i) - \epsilon \right\|_2$$

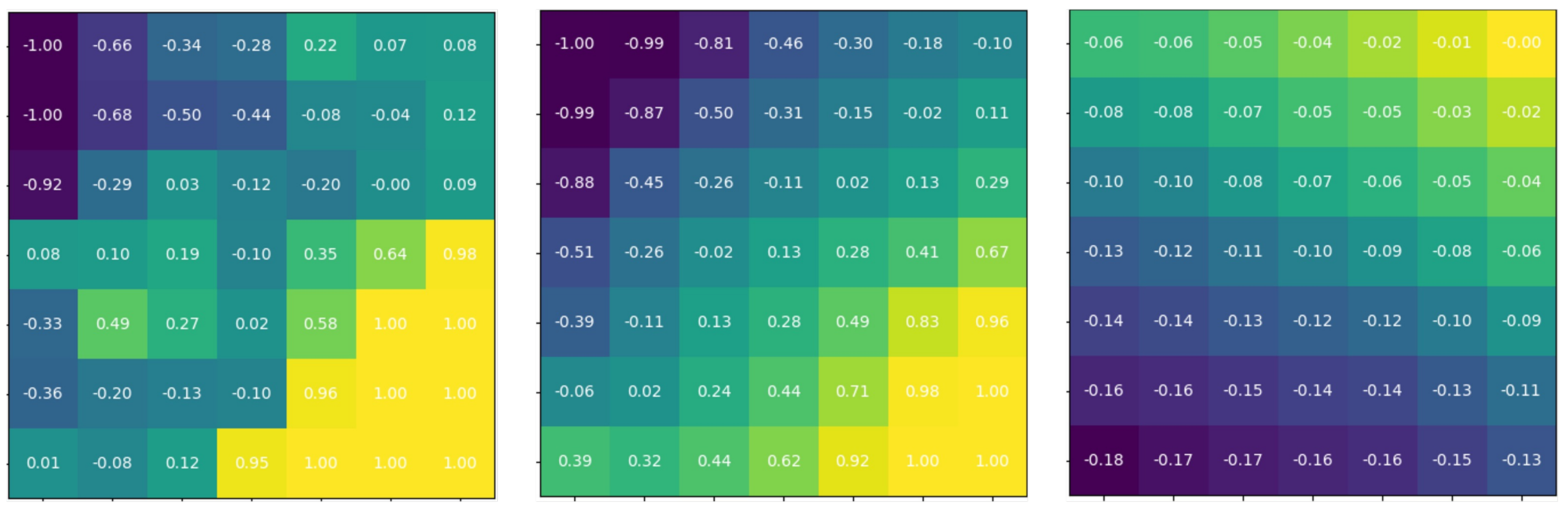

Replay Buffer
Preference Queries
Binary Feedback
Feedback Buffer
Reward Initialization Step
Policy Learning $\pi_\phi(s)$
State, Reward
Action
Reward Learning $R_h(s, a)$

Kambhampati's research is supported by the J.P. Morgan Faculty Research Award, ONR grants N00014-16-1-2892, N00014-18-1-2442, N00014-18-1-2840, N00014-9-1-2119, AFOSR grant FA9550-18-1-0067 and DARPA SAIL-ON grant W911NF19-2-0006.

Experiments & Results

- **Investigate via Empirical Evaluation**

> **Question 1:** Empirically verify that seed of the experiment impacting reward initialization yields substantially different reward models.



-1.00	-0.66	-0.34	-0.28	0.22	0.07	0.08
-1.00	-0.68	-0.50	-0.44	-0.08	-0.04	0.12
-0.92	-0.29	0.03	-0.12	-0.20	-0.00	0.09
0.08	0.10	0.19	-0.10	0.35	0.64	0.98
-0.33	0.49	0.27	0.02	0.58	1.00	1.00
-0.36	-0.20	-0.13	-0.10	0.96	1.00	1.00
0.01	-0.08	0.12	0.95	1.00	1.00	1.00

-1.00	-0.99	-0.81	-0.46	-0.30	-0.18	-0.10
-0.99	-0.87	-0.50	-0.31	-0.15	-0.02	0.11
-0.88	-0.45	-0.26	-0.11	0.02	0.13	0.29
-0.51	-0.26	-0.02	0.13	0.28	0.41	0.67
-0.39	-0.11	0.13	0.28	0.49	0.83	0.96
-0.06	0.02	0.24	0.44	0.71	0.98	1.00
0.39	0.32	0.44	0.62	0.92	1.00	1.00

-0.06	-0.06	-0.05	-0.04	-0.02	-0.01	-0.00
-0.08	-0.08	-0.07	-0.05	-0.05	-0.03	-0.02
-0.10	-0.10	-0.08	-0.07	-0.06	-0.05	-0.04
-0.13	-0.12	-0.11	-0.10	-0.09	-0.08	-0.06
-0.14	-0.14	-0.13	-0.12	-0.12	-0.10	-0.09
-0.16	-0.16	-0.15	-0.14	-0.14	-0.13	-0.11
-0.18	-0.17	-0.17	-0.16	-0.16	-0.15	-0.13

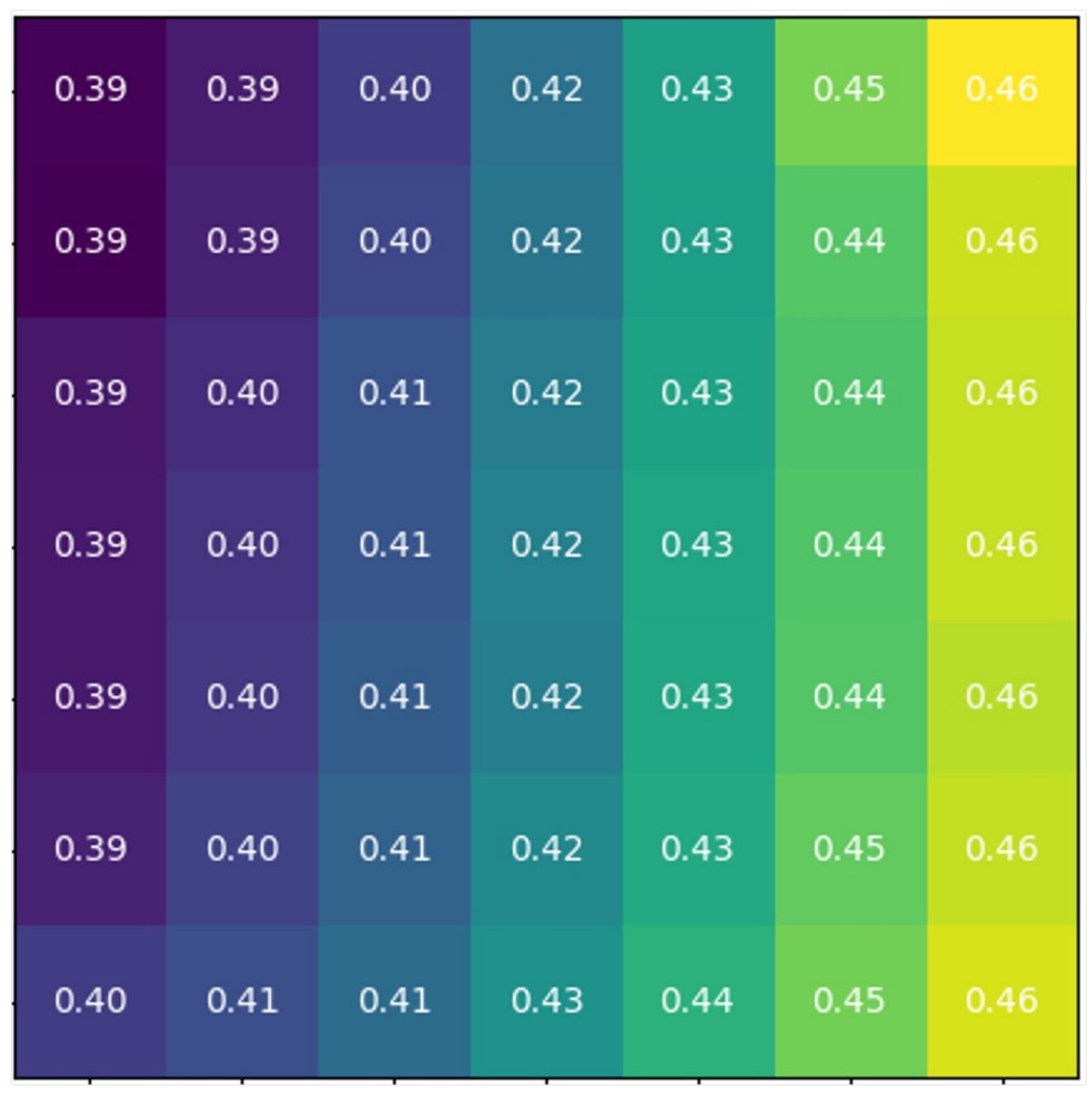
(a) Reward values for 7x7 Gridworld states over different seeds via Uniform Initialization

> **Question 2:** Whether the proposed data-driven reward initialization scheme improves / stabilises the performance?

> Empirical results on 7x7 and 15x15 gridworlds with backbone PEBBLE are shown in figures (b-d).

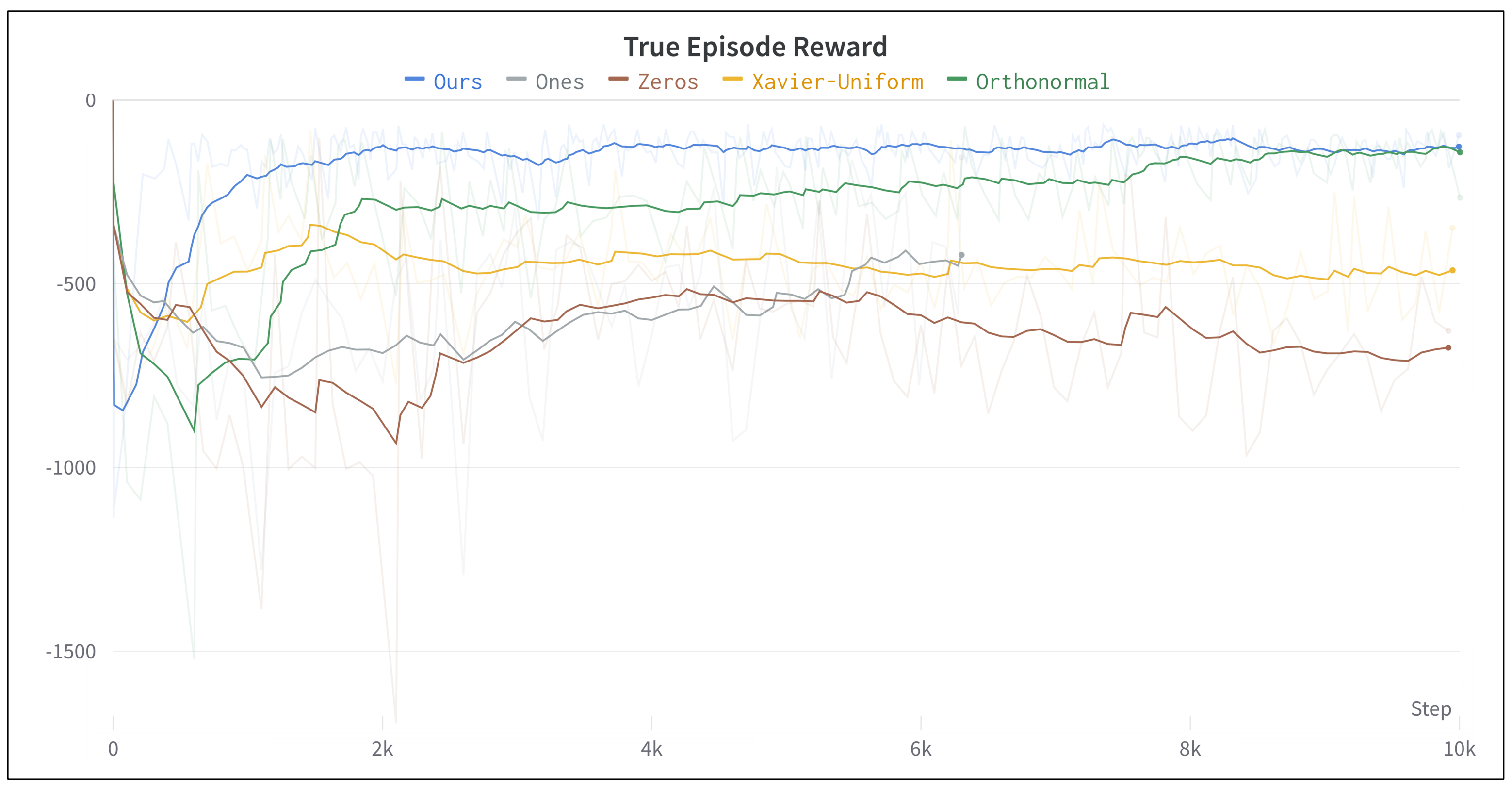
> Using the data-driven initialization has zero cost to the human in the loop, and requires zero additional trajectory sampling (for algorithms that have a pre-training phase).

> We compare against baselines : Ones, Zeros, Xavier-Uniform, Orthonormal

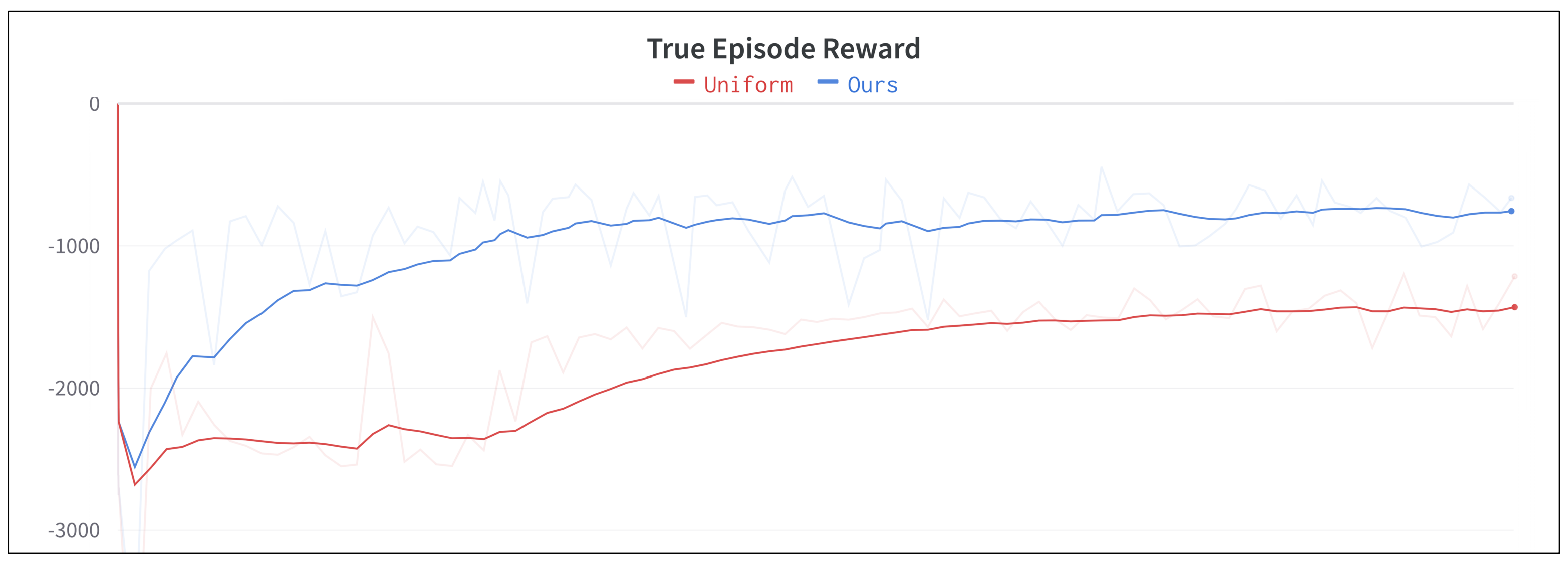


0.39	0.39	0.40	0.42	0.43	0.45	0.46
0.39	0.39	0.40	0.42	0.43	0.44	0.46
0.39	0.40	0.41	0.42	0.43	0.44	0.46
0.39	0.40	0.41	0.42	0.43	0.44	0.46
0.39	0.40	0.41	0.42	0.43	0.44	0.46
0.39	0.40	0.41	0.42	0.43	0.45	0.46
0.40	0.41	0.41	0.43	0.44	0.45	0.46

(b) Reward values for 7x7 Gridworld states after data-driven Reward Initialization



(c) Learning curves for 7x7 gridworld across various initialization schemes.



(d) Learning curves for 15x15 gridworld comparing Ours v/s Uniform.

Key Takeaways

1. Reward learning and Policy learning *is sensitive to Reward Initialization.*
2. Empirically, we show that several initialization schemes for the reward model results in high variance in agent performance.
3. We propose a pre-training data-driven reward initialization step that forces the reward-model co-domain to be uniform, minimizing any prior bias over trajectories.