SFP: State-free Priors for Exploration in Off-Policy Reinforcement Learning

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Marco Bagatella, Sammy Christen, Otmar Hilliges

Department of Computer Science, ETH Zürich, Switzerland

Advanced Interactive Technologies

Contributions

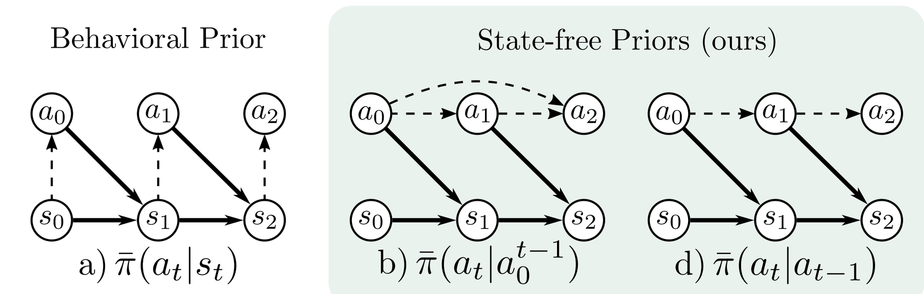
1. We propose **state-free priors** for guiding exploration in long-horizon, sparse rewards tasks.
2. We derive a **novel integration scheme** for priors into SAC [1].
3. We show how state-free priors can be **learned from few task-agnostic trajectories** and used to **improve exploration in weakly related tasks**.

In a nutshell: how can we **improve exploration** and accelerate downstream reinforcement learning from an offline dataset of **task-agnostic trajectories**?

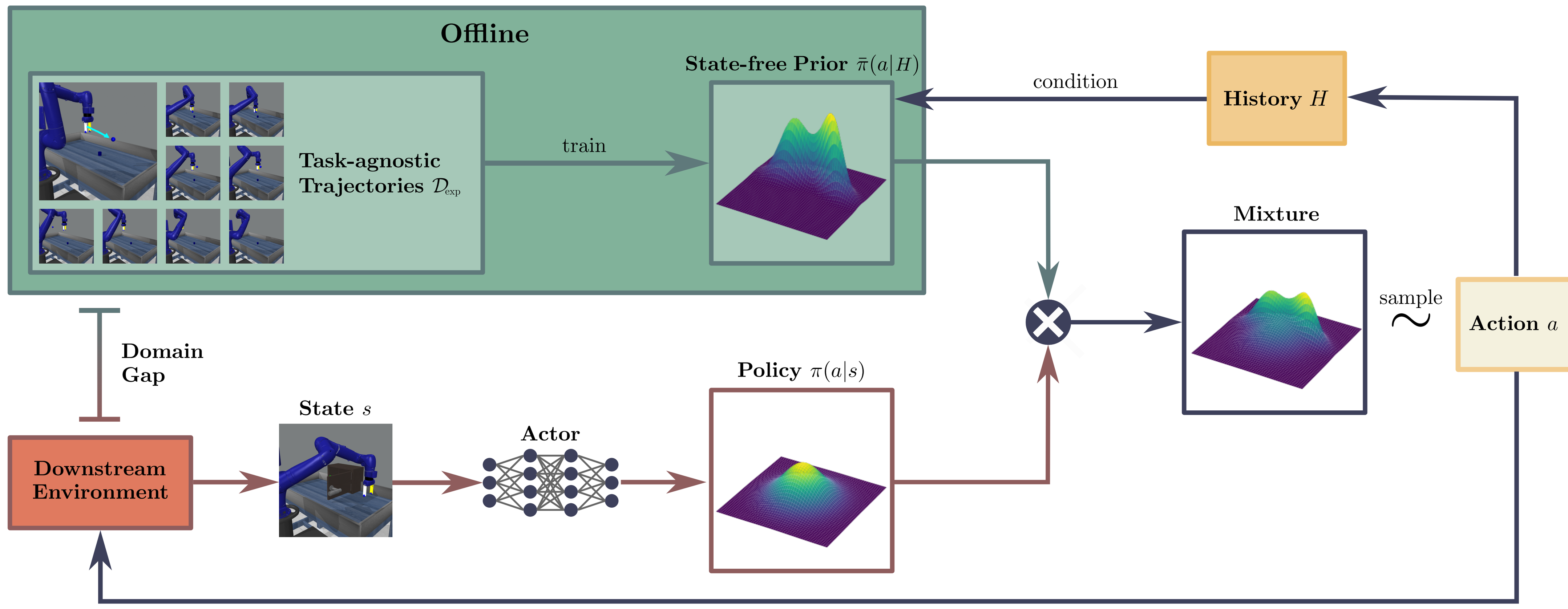
State-free Priors

- **Behavioral priors** $\bar{\pi}(a|s)$ can be trained from demonstrations and guide exploration, but struggle when deployed **on fundamentally different tasks** [2].
- Extracting non-Markovian patterns from demonstrations can be helpful in a **broader range of tasks**: we propose to focus on the **temporal structure** of demonstrations rather than on task-specific strategies.

A state-free prior is a **state-independent non-Markovian action distribution** modeling promising actions conditioned on past action: $\bar{\pi}(a_t|a_0^{t-1})$.



Integration in SAC



Actions are sampled from a **mixture** between the policy π and the prior $\bar{\pi}$:

$$a_t \sim (1-\lambda_t)\pi(\cdot|s_t) + \lambda_t\bar{\pi}(\cdot|s_t, H_t) \quad \text{with } 0 \le \lambda_t \le 1$$

Ideally, $\lambda_t \approx 1$ when exploration is needed.

- We learn a **mixing function** (i.e. $\lambda_t = \Lambda_\omega(s_t)$) and maximize the max-entropy objective w.r.t. the mixture $\tilde{\pi}$:

$$\underset{\pi_\phi, \Lambda_\omega}{argmax} \ \underset{\tau\sim\tilde{\pi}}{\mathbb{E}}\left[\sum_{t=0}^{\infty}\gamma^t\Big(\mathcal{R}(s_t,a_t) + \alpha\mathcal{H}(\pi_\phi(\cdot|s_t))\Big)\right].$$

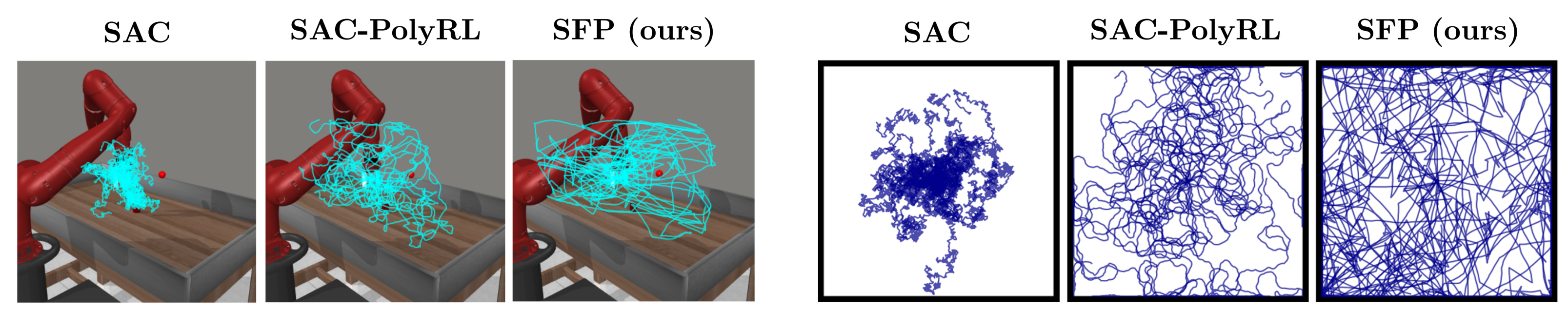
- We derive slight modifications to SAC's policy and value loss, and an objective for Λ_ω:

$$J_{\Lambda_\omega} = -\underset{(s)\sim\mathcal{D}}{\mathbb{E}}\left[\Lambda_\omega(s)\big(Q_\theta^{\tilde{\pi}}(s,\bar{a}) - Q_\theta^{\tilde{\pi}}(s,a)\big)\right] \quad \text{with } \bar{a}\sim\bar{\pi}(\cdot), a\sim\pi_\phi(\cdot|s).$$

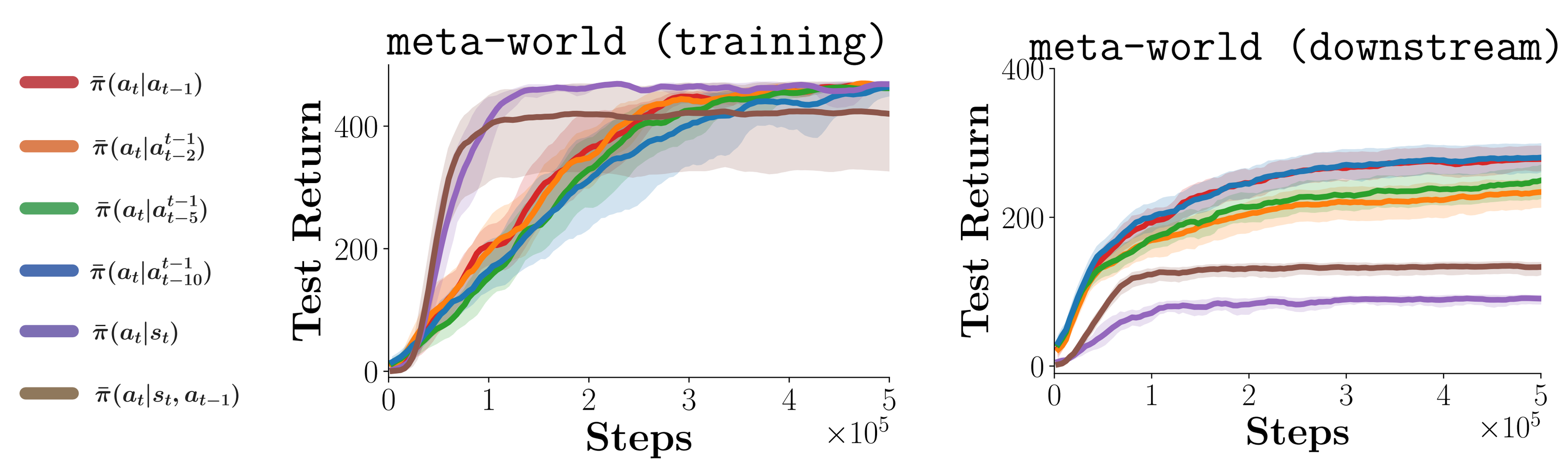
Experimental Setup

Experiments

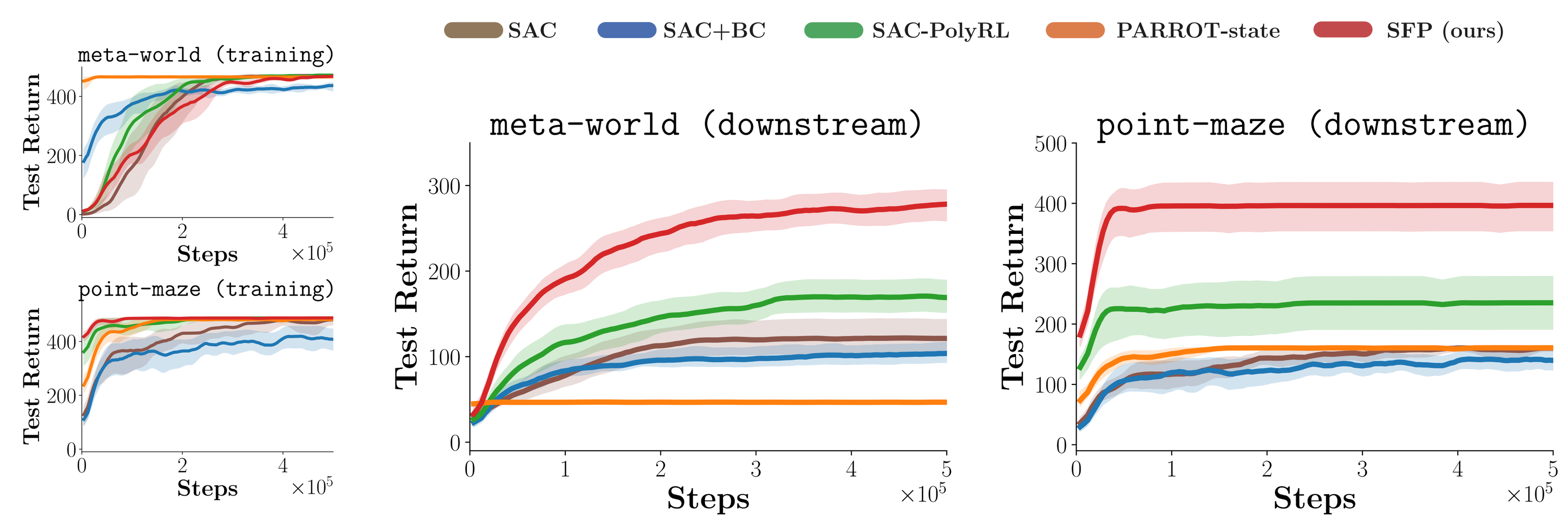
Sampled Trajectories



Conditioning Ablation



Transfer Learning



Correspondence

mbagatella@ethz.ch

Website

References

[1] Tuomas Haarnoja et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor". In: *International conference on machine learning*. 2018.

[2] Avi Singh et al. "Parrot: Data-Driven Behavioral Priors for Reinforcement Learning". In: *International Conference on Learning Representations*. 2021.

[3] Susan Amin et al. "Locally Persistent Exploration in Continuous Control Tasks with Sparse Rewards". In: *Proceedings of the 38th International Conference on Machine Learning*. 2021.