Offline Diversity Maximization Under Imitation Constraints

Marin Vlastelica, Jin Cheng, Georg Martius, Pavel Kolev

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS · EBERHARD KARLS UNIVERSITÄT TÜBINGEN · VolkswagenStiftung · ETH zürich

Overview

Diverse
- Robust solutions
- Multiple options

[DIAYN, DADS, DOMINO]

online setting

Offline
- Large datasets
- Safe learning

[AWAC, BC, CRR, IQL, CQL]

not diverse (single expert)

Imitation
- No reward engineering
- Human demonstrations

[GAIL, SMODICE]

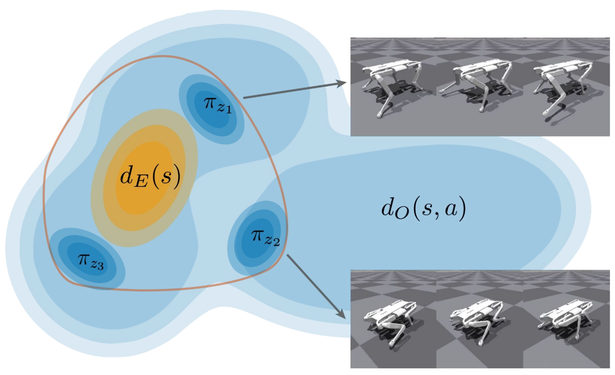
not diverse (single expert)

Propose: principled algorithm for **Diverse** **Offline** **Imitation** (DOI) learning

Diverse Offline Imitation

$$\max_{\{d_z(S)\}_{z\in Z}} \mathcal{I}(S;Z)$$

subject to
$\mathrm{D}_{\mathrm{KL}}(d_z(S)||d_E(S)) \leq \varepsilon \quad \forall z$



Input: $\mathcal{D}_E \sim d_E(s)$ $\mathcal{D}_O \sim d_O(s,a)$
state-only expert dataset
state-action behavior dataset

Def. State-action occupancy

$$d^\pi(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathrm{Pr}\left[s_t = s, a_t = a \,|\, s_0 \sim \rho_0, a_t \sim \pi(\ |\ s_t), s_{t+1} \sim \mathcal{P}(\ |\ s_t, a_t)\right]$$

Relaxed Problem Formulation

$$\max_{\{d_z(S)\}_{z\in Z}} \mathcal{I}(S;Z)$$

Mutual Info. variational lower bound

$$\mathcal{I}(S;Z) \geq \sum_z \mathbb{E}_{d_z(s)}\left[\frac{\log\left(|Z|q(z|s)\right)}{|Z|}\right]$$

$q(z|s)$ train a skill-discriminator

subject to
$\mathrm{D}_{\mathrm{KL}}(d_z(S)||d_E(S)) \leq \varepsilon \quad \forall z$

SMODICE expert (offline)

$$d_{\tilde{E}}(S,A) \approx \arg\min_{d(s,a)} \mathrm{D}_{\mathrm{KL}}(d(S)||d_E(S))$$

subject to
$\mathrm{D}_{\mathrm{KL}}(d_z(S,A)||d_{\tilde{E}}(S,A)) \leq \varepsilon \quad \forall z$

Algorithmic Approach

Diversity / Imitation

$$\max_{\substack{d_z(s,a)\\ q(z|s)}} \min_{\lambda>0} \sum_{z\in Z} \mathbb{E}_{d_z(s)}\left[\frac{\log\left(|Z|q(z|s)\right)}{|Z|}\right] + \sum_{z\in Z}\lambda_z\left[\epsilon - \mathrm{D}_{\mathrm{KL}}\left(d_z(S,A)||d_{\tilde{E}}(S,A)\right)\right]$$

Regularized RL Problem

$$\max_{\substack{d_z(s,a)\\ q(z|s)}} \min_{\lambda>0} \sum_z \lambda_z\left\{\epsilon + \mathbb{E}_{d_z(s,a)}\left[R_z^\lambda(s,a)\right] - \mathrm{D}_{\mathrm{KL}}\left(d_z(S,A)||d_O(S,A)\right)\right\}$$

DICE (offline)

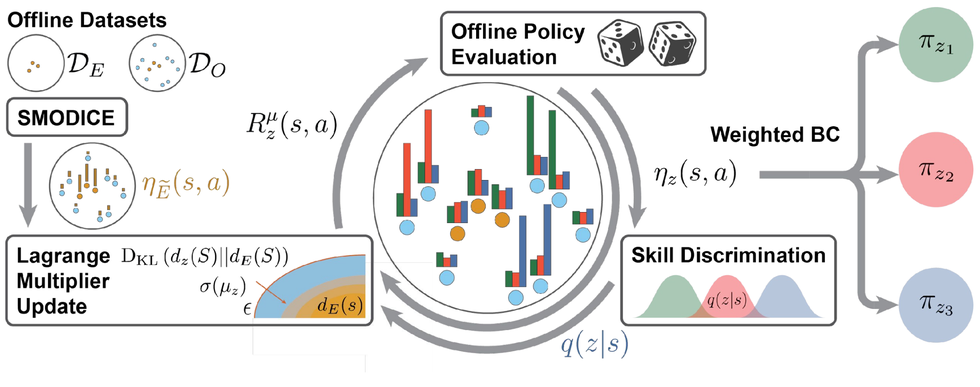
$$\eta_z(s,a) = \frac{d_z(s,a)}{d_O(s,a)}$$

$$R_z^\lambda(s,a) := \underbrace{\frac{1}{\lambda_z}}_{\text{Constraint Violation}} \underbrace{\frac{\log\left(q(z|s)|Z|\right)}{|Z|}}_{\text{Skill Diversity}} + \underbrace{\log \eta_{\tilde{E}}(s,a)}_{\text{Expert Imitation}}$$

SMODICE expert (offline)

$$\eta_{\tilde{E}}(s,a) = \frac{d_{\tilde{E}}(s,a)}{d_O(s,a)}$$

Alternating Optimization Scheme



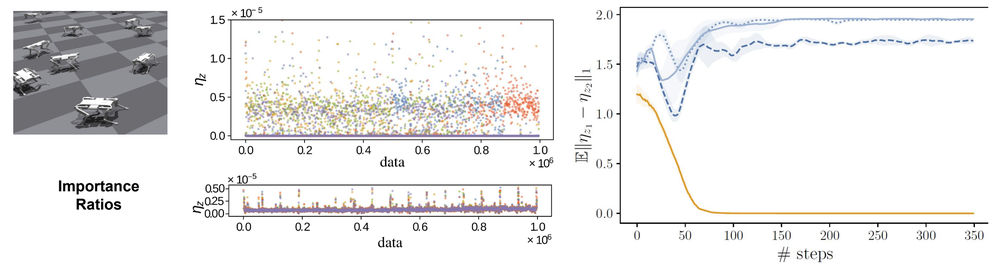
Experiments (SOLO12)

I. Locomotion Task (SIM)

Offline Evaluation

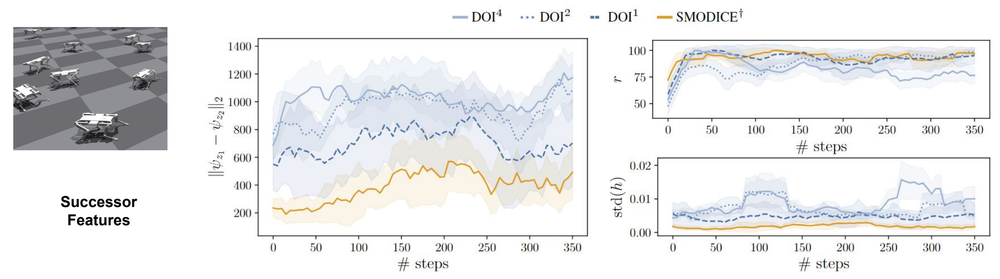
1) DOI skills well-separate data.

2) Constraint level ε controls ratio distance.



Online Evaluation

3) Relaxed constraints yield increased diversity, albeit at the expense of performance loss.



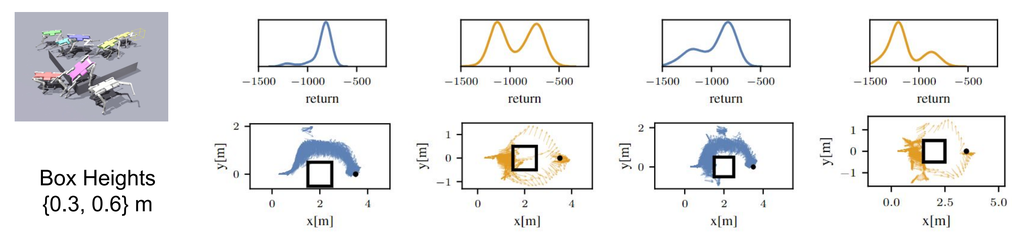
I. Locomotion Task (REAL)

4) DOI skills trained in **SIM** (*with domain randomization)* are successfully deployed in the **Real System**.

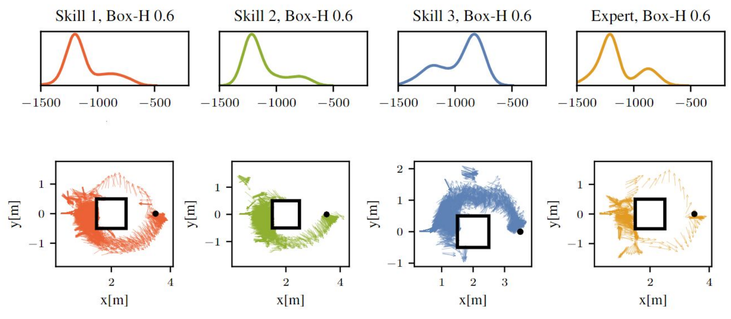
II. OBSTACLE NAVIGATION TASK (SIM)

Online Evaluation

5) SMODICE expert struggles with out-of-distribution (higher) box heights, while a robust DOI skill successfully navigates by detouring (to the left side).



Limitation: **6)** Not all learned DOI skills are robust. Selection is required.



Contact: {mvlastelica, pkolev}@uni-tuebingen.de

Paper · Project Website · Autonomous Learning

References

[SMODICE] Y. Ma, A. Shen, D. Jayaraman, O. Bastani "*Versatile Offline Imitation from Observations and Examples via Regularized State-Occupancy Matching*", ICML 2022
[DICE] O. Nachum, B. Dai "*Reinforcement learning via fenchel-rockafellar duality*", arXiv 2020
[DIAYN] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, Sergey Levine "*Diversity is All You Need: Learning Skills without a Reward Function*", ICLR 2019
[DOMINO] T. Zahavy, Y. Schroecker, F. Behbahani, K. Baumli, S. Flennerhag, S. Hou, S. Singh "*Discovering Policies with DOMiNO: Diversity Optimization Maintaining Near Optimality*", ICLR 2023
[OFFLINE] R. F. Prudencio, M. R. O. A. Maximo, and E. L. Colombini. "*A survey on offline reinforcement learning: Taxonomy, review, and open problems*", arXiv 2022