Dual-Force: Enhanced Offline Diversity Maximization under Imitation Constraints

Pavel Kolev, Marin Vlastelica, Georg Martius

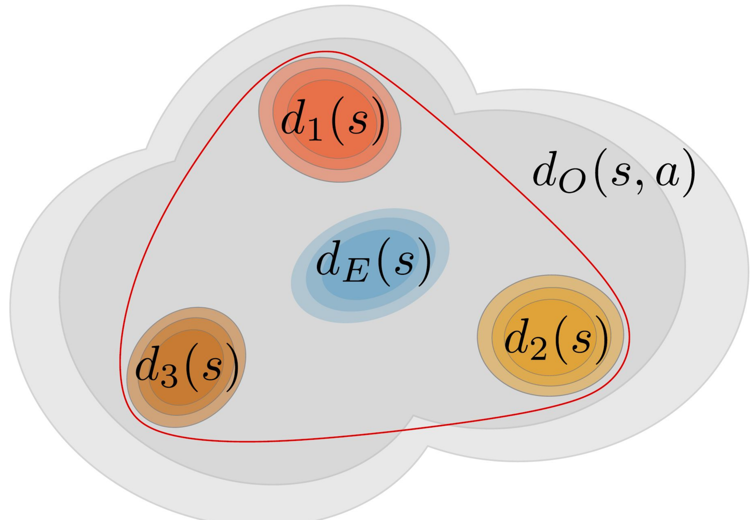
EBERHARD KARLS UNIVERSITÄT TÜBINGEN · Tübingen AI Center · VolkswagenStiftung

Diverse Offline Imitation

$$\max_{d_1,\dots,d_n} \text{Diversity}(d_1,\dots,d_n)$$

subject to

$$\text{D}_{\text{KL}}\left(d_i(S)||d_E(S)\right) \le \varepsilon \quad \forall i$$



Input:

state expert dataset $\mathcal{D}_E \sim d_E(s)$

state-action behavior dataset $\mathcal{D}_O \sim d_O(s,a)$

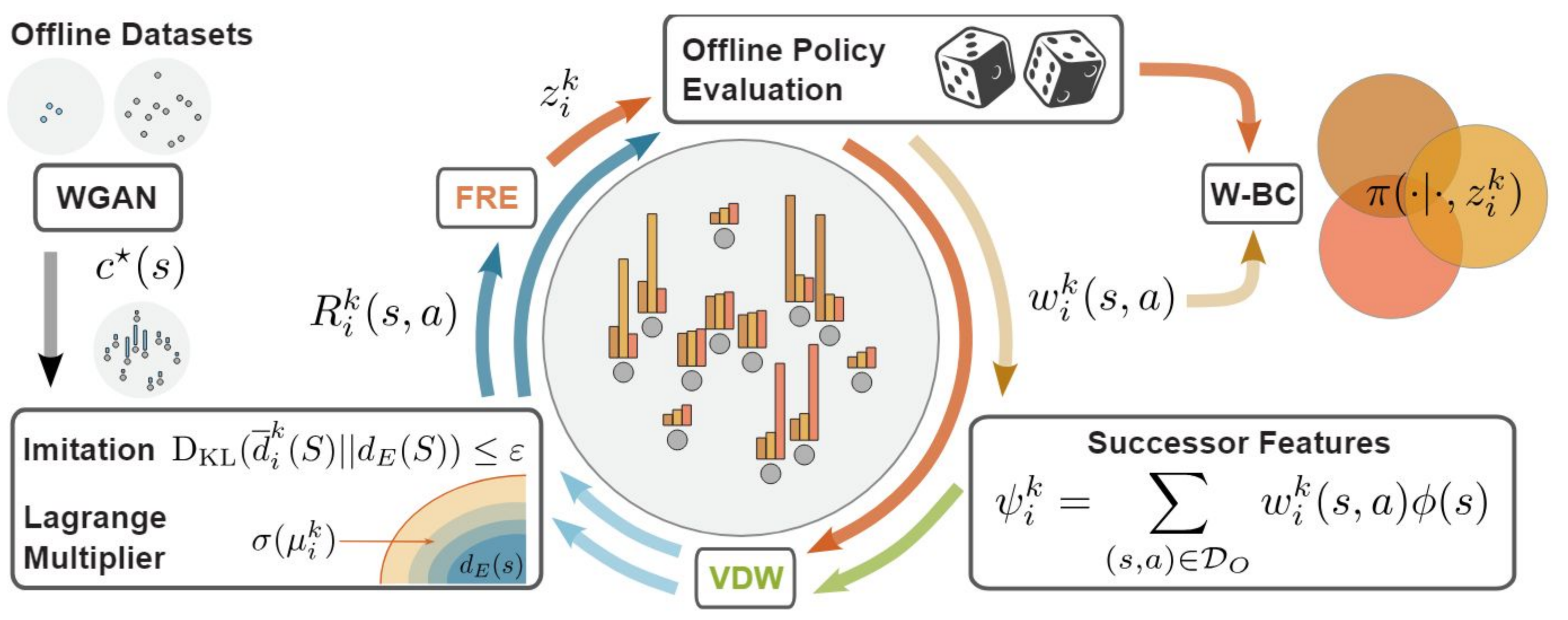
Def. State-action occupancy

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

Main Contribution

	Prior Work Limitations [DOI]	Dual-Force (Our method)
Diversity Objective	Requires learning a skill-discriminator i) hard to train it offline ii) InfoGain helps but quickly vanishes	Van der Waals [VDW] + Successor Features No need to learn a skill-discriminator Provides strong diversity signal
Non-Stationary Rewards	[DICE] assumes stationary reward, violating it makes Value training unstable	Handles non-stationary rewards by conditioning Value function on FRE embedding
Dependance on num_skills	Scales linearly with the num_skills Learning large set of skills is prohibitive	Independent of the num_skills All observed skills during training are invocable

Dual-Force



Fundamental Techniques

Van der Waals Force [VDW]

$$\max_{d_1,\dots,d_n} 0.5\sum_{i=1}^{n}\ell_i^2 - 0.2(\ell_i^5/\ell_0^3) \qquad \ell_i := \|\psi_i - \psi_{j_i^\star}\|_2$$

$$j_i^\star := \arg\min_{j\neq i} \|\psi_i - \psi_j\|_2$$

Functional Reward Encoding [FRE]

$$\text{FRE}_{\text{enc}}(\{(s, r(s))\}_{s\in S}) \to z_r$$

$$\text{FRE}_{\text{dec}}(s, z_r) \approx r(s)$$

Algorithmic Approach

Constraints

$$\text{D}_{\text{KL}}\left(d_i(S)||d_E(S)\right) \le \varepsilon \xrightarrow{\text{relaxation}} -\mathbb{E}_{d_i(s)}\left[\log\frac{d_E(s)}{d_O(s)}\right] + \text{D}_{\text{KL}}\left(d_i(S,A)||d_O(S,A)\right) \le \varepsilon$$

Problem Formulation

Diversity

Imitation

$$\min_{\lambda_i\ge 0}\max_{d_i}\mathbb{E}_{d_i(s,a)}\boxed{[\beta_i^k(s,a)]} + \lambda_i\left[\mathbb{E}_{d_i(s,a)}\left[\log\frac{d_E(s)}{d_O(s)}\right] - \text{D}_{\text{KL}}\left(d_i(S,A)||d_O(S,A)\right)\right]$$

Regularized RL Problem

$$\max_{d_i}\mathbb{E}_{d_i(s,a)}\boxed{[R_i^k(s,a)]} - \text{D}_{\text{KL}}\left(d_i(S,A)||d_O(S,A)\right)$$

[DICE] offline

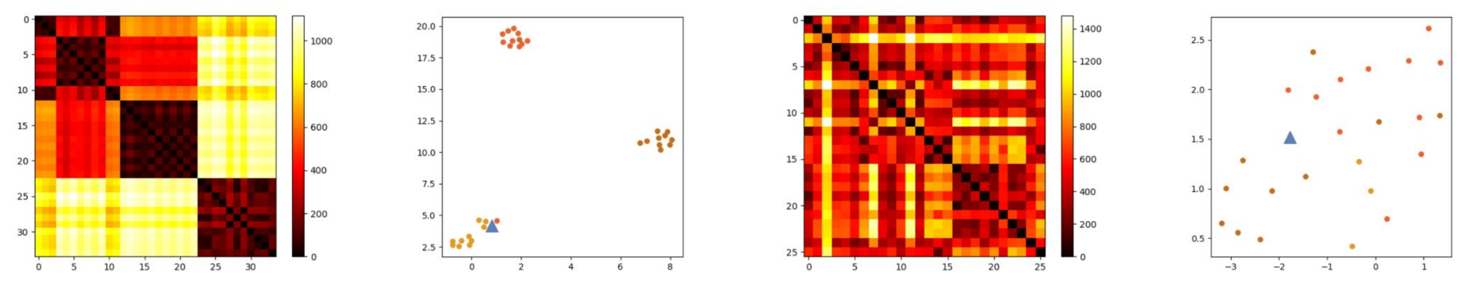
[WGAN]

$$\boxed{R_i^k(s,a)} := \underbrace{(1-\sigma(\mu_i))}_{\text{Constraint Satisfaction}}\underbrace{\boxed{\beta_i^k(s,a)}}_{\text{VdW-Diversity}} + \underbrace{\sigma(\mu_i)}_{\text{Constraint Violation}}\underbrace{\log\frac{c^\star(s)}{1-c^\star(s)}}_{\text{Expert-Imitation}}$$

Diversity $\boxed{\beta_i^k(s,a)} := (1-(\ell_i^k/\ell_0)^3)\langle\phi(s), \psi_i^k - \psi_{j_i^\star}^k\rangle$

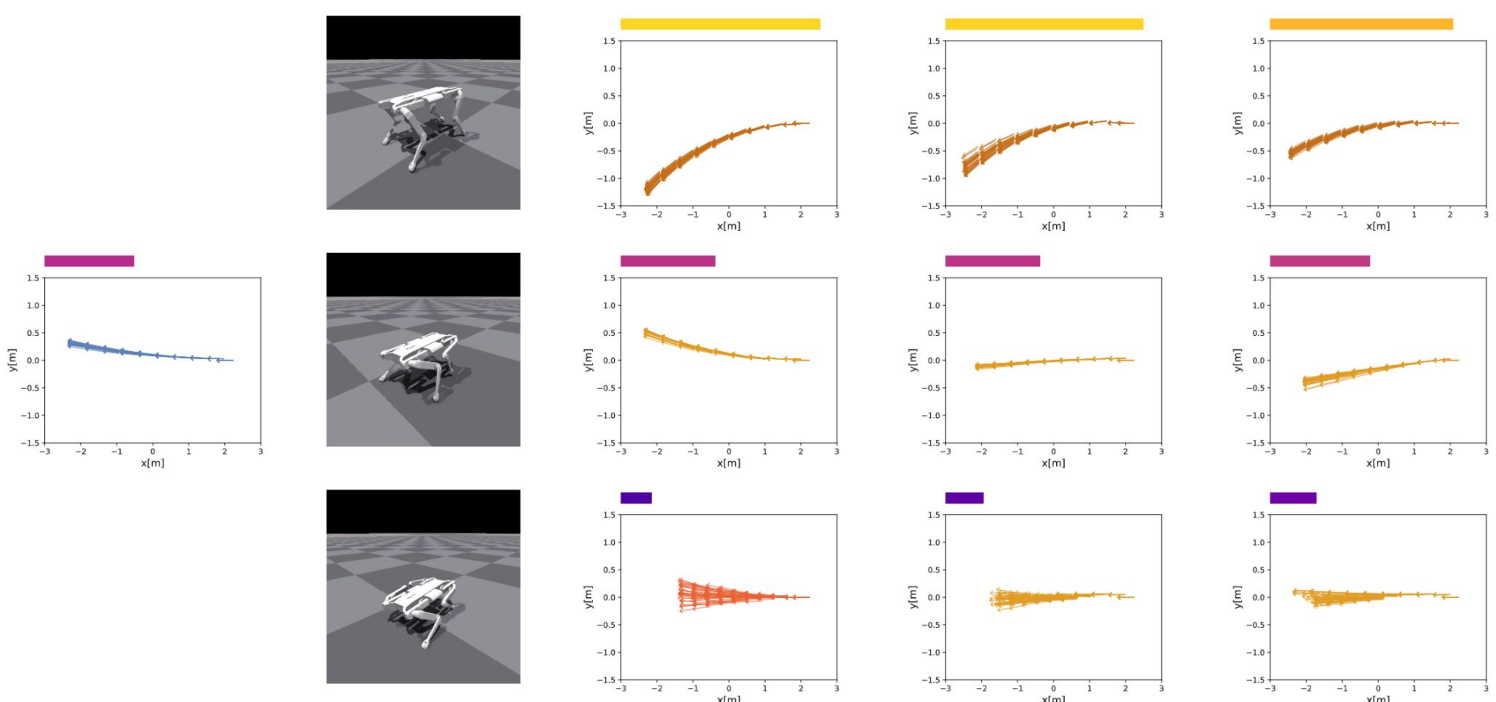
Experiments (Solo12)

Successor Features



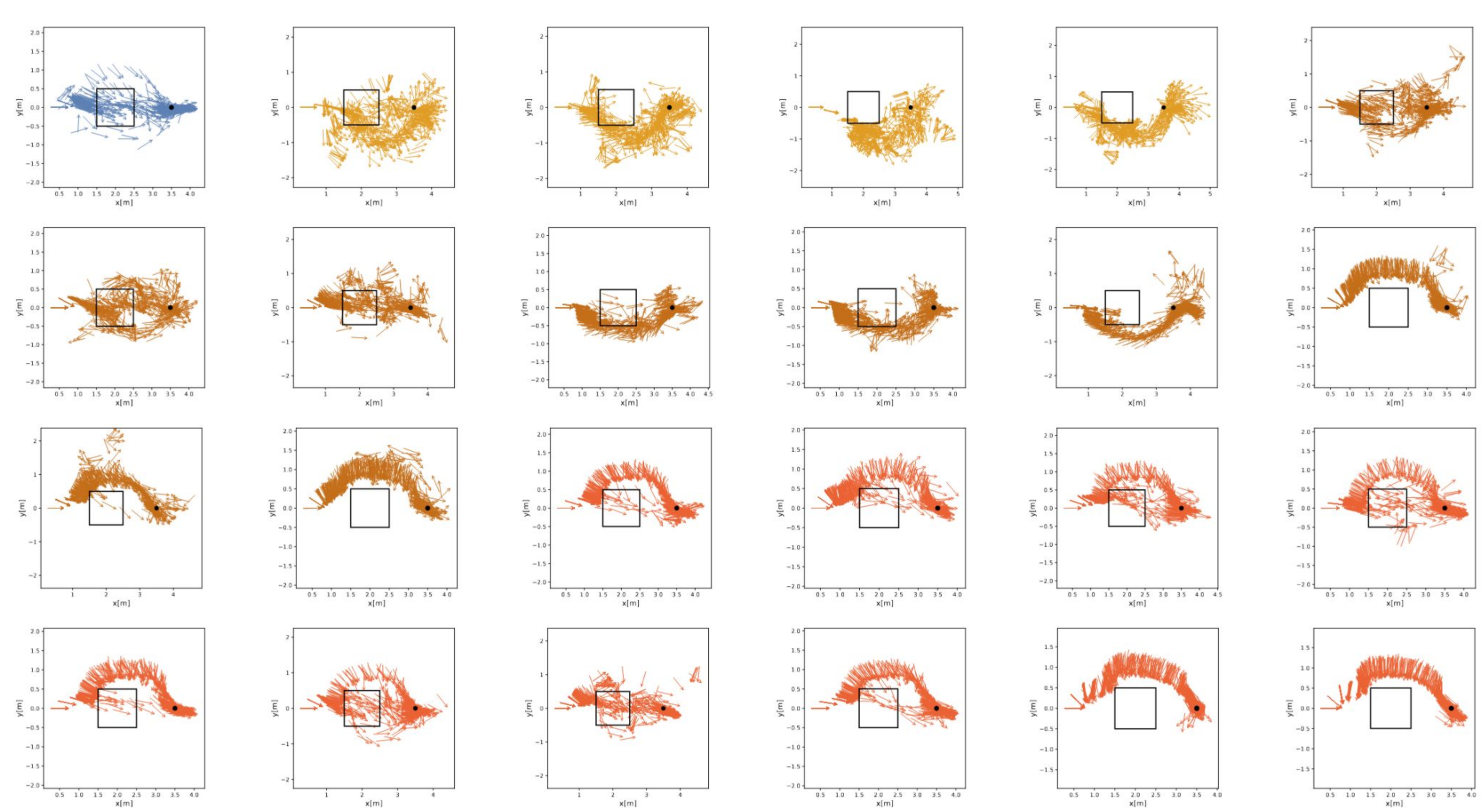
Locomotion Task

1) The learned skills find all base-height movements (high, middle, low) and have different angular velocity. The SMODICE-expert has middle base-height.



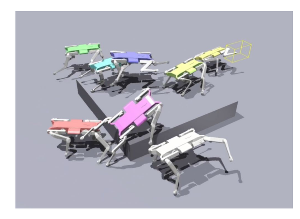
Obstacle Navigation Task

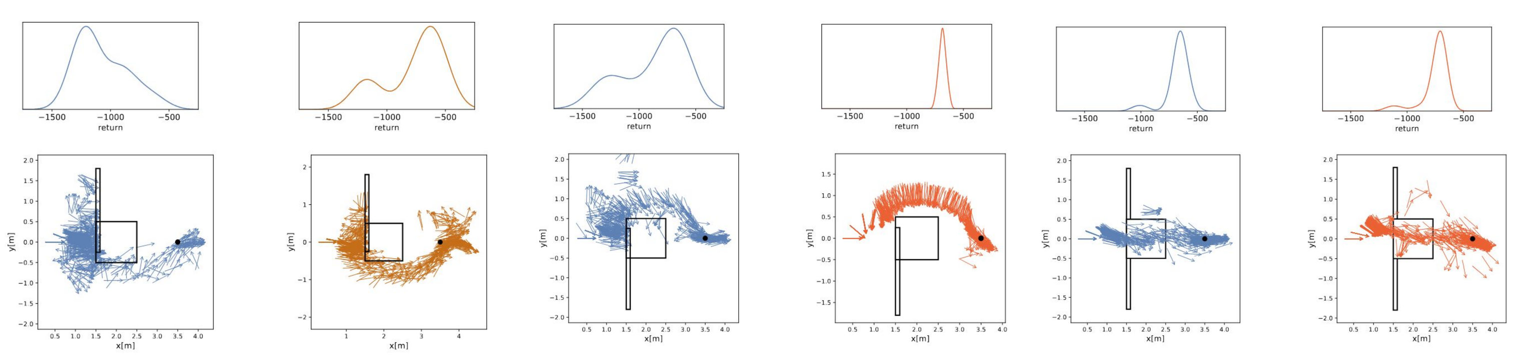
2) While the multi-modal SMODICE-expert prefers passing over the box, the set of learned skills capture all modalities: left, right, over, and mixed.



Robustness: Fence Obstacles

3) Learned skills outperform SMODICE-expert (left, right) or perform on par with it (both left and right).





Contact: pavel.kolev@uni-tuebingen.de

Paper · Project Website · Autonomous Learning

References

[DOI] M. Vlastelica, J. Cheng, G. Martius, P. Kolev: "*Offline Diversity Maximization Under Imitation Constraints*", RLC'24

[FRE] K. Frans, S. Park, P. Abbeel, S. Levine: "*Unsupervised Zero-Shot Reinforcement Learning via Functional Reward Encodings*", ICLM'24

[VDW] T. Zahavy, Y. Schroecker, F. Behbahani, K. Baumli, S. Flennerhag, S. Hou, S. Singh: "*Discovering Policies with DOMiNO: Diversity Optimization Maintaining Near Optimality*", ICLR'23

[SMODICE] Y. Ma, A. Shen, D. Jayaraman, O. Bastani: "*Versatile Offline Imitation from Observations and Examples via Regularized State-Occupancy Matching*", ICML'22

[DICE] O. Nachum, B. Dai: "*Reinforcement learning via fenchel-rockafellar duality*", arXiv'20

[WGAN] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin. A. Courville: "Improved Training of Wasserstein GANs", NeurIPS'17