Offline Diversity Maximization Under Imitation Constraints

Marin Vlastelica[1,2], **Jin Cheng**[1,], **Georg Martius**[1,2], and **Pavel Kolev**[1,2]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]University of Tübingen, Tübingen, Germany
[3]ETH Zurich, Switzerland

Abstract

There has been significant recent progress in the area of unsupervised skill discovery, utilizing various information-theoretic objectives as measures of diversity. Despite these advances, challenges remain: current methods require significant online interaction, fail to leverage vast amounts of available task-agnostic data and typically lack a quantitative measure of skill utility. We address these challenges by proposing a principled offline algorithm for unsupervised skill discovery that, in addition to maximizing diversity, ensures that each learned skill imitates state-only expert demonstrations to a certain degree. Our main analytical contribution is to connect Fenchel duality, reinforcement learning, and unsupervised skill discovery to maximize a mutual information objective subject to KL-divergence state occupancy constraints. Furthermore, we demonstrate the effectiveness of our method on the standard offline benchmark D4RL and on a custom offline dataset collected from a 12-DoF quadruped robot for which the policies trained in simulation transfer well to the real robotic system.[1]

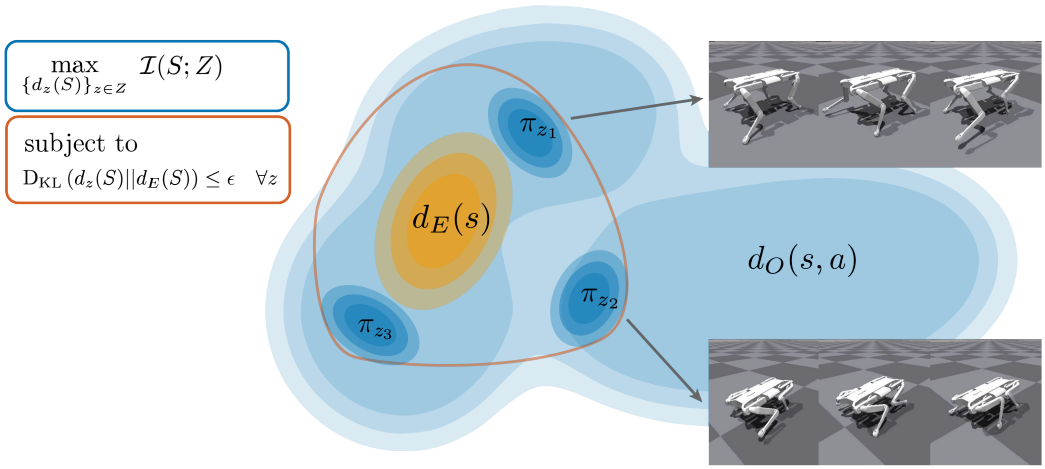


Figure 1: Diverse Offline Imitation (DOI) maximizes a variational lower bound on the mutual information between latent skills z and states s visited by associated skill-conditioned policies π_z, subject to a KL-divergence constraint to limit the deviation of the state occupancy $d_z(s)$ of each latent skill z from that of an expert $d_E(s)$.

[1]Project website with videos: https://tinyurl.com/diversity-via-duality

1 Introduction

Recent advancements in reinforcement learning (RL) have included substantial progress in unsupervised skill discovery, aiming to empower autonomous agents with the capability to acquire a diverse set of skills directly from their environment, without relying on predefined human-engineered rewards or demonstrations. These methods have the potential to revolutionize the way RL agents learn to solve complex tasks. The growing interest in unsupervised skill discovery has led to various approaches, typically rooted in information-theoretic concepts, including empowerment (Klyubin et al., 2005; Mohamed & Jimenez Rezende, 2015; Eysenbach et al., 2019), information bottleneck (Tishby et al., 1999; Goyal et al., 2019; Kim et al., 2021a) and information gain (Houthooft et al., 2016; Strouse et al., 2022; Park & Levine, 2023). Despite these advancements, there remains a significant challenge. Current methods demand substantial online interaction with the environment, making exploration in high-dimensional state-action spaces inefficient. Although Zahavy et al. (2022) introduced constraints to enhance skill performance and narrow the exploration space by incentivizing diverse skills to meet a certain utility measure, their approach does not eliminate the need for considerable online interaction with the environment. Meanwhile, there have been significant recent advances in large-scale data collection (Rob, 2020; Walke et al., 2023; Brohan et al., 2023) and in the development of scalable and sample-efficient offline RL algorithms that leverage diverse behaviors of pre-collected experience. However, these approaches struggle with well-known challenges, including off-policy evaluation and the out-of-distribution problem, which have been studied extensively in previous work (Levine et al., 2020; Prudencio et al., 2022).

In this work, we address the aforementioned challenges by introducing a novel problem formulation and complementing it with the first principled *offline* RL algorithm for unsupervised skill discovery that, in addition to maximizing diversity, ensures that each learned skill imitates state-only expert demonstrations to a certain degree. More specifically, we consider a problem formulation with two datasets: a large one with diverse state-action demonstrations and another much smaller one with state-only expert demonstrations. This setting is particularly valuable in robotics scenarios where expert demonstrations are limited and the domain of the expert may be different from that of the agent, such as in human demonstrations. Another potential application is to enhance the realism of computer games by creating an immersive experience of interacting with non-player characters, each behaving in a slightly different style, while all partially imitating the behavior of a human expert.

We formulate the problem as a Constrained Markov Decision Process (CMDP) (Altman, 1999; Szepesvári, 2020) that seeks to maximize diversity through a mutual information objective, subject to Kullback-Leibler (KL) divergence state occupancy constraints ensuring that each skill imitates state expert demonstrations to a certain degree. The resulting CMDP has convex objective and constraints, making the optimization problem intractable. We adopt a tractable relaxation approach consisting of an alternating scheme that maximizes a variational lower bound on mutual information, and to handle the constraints it applies Lagrange relaxation. Our method, Diverse Offline Imitation (DOI), overcomes the off-policy evaluation by leveraging the Fenchel-Rockafellar duality in RL (Nachum & Dai, 2020; Kim et al., 2022; Ma et al., 2022) to connect a dual optimal value solution (computed using offline samples) with primal optimal state-action occupancy ratios. These ratios serve as importance weights for offline training of a skill-conditioned policy, skill-discriminator, KL-divergence estimators, and Lagrange multipliers. We demonstrate the effectiveness of our method on the standard offline benchmark D4RL (Fu et al., 2020) and on a custom offline dataset collected from a 12-DoF quadruped robot Solo12 (Léziart et al., 2021). In addition, we show that DOI trained on simulation data transfers well to a real robot system.

2 Related Work

In the context of skill discovery Achiam et al. (2018) and Campos et al. (2020) showed that methods like DIAYN (Eysenbach et al., 2019) can struggle to learn large numbers of skills and have a poor coverage of the state space. Strouse et al. (2022) observed that when a novel state is visited, the discriminator lacks sufficient training data to accurately classify skills, which results in a low intrinsic reward for exploration. They address this by introducing an information gain objective (involving an ensemble of discriminators) as a bonus term. Kim et al. (2021b) gave a skill discovery approach based on an information bottleneck that leads to disentangled and interpretable skill representations. Park et al. (2022; 2023) proposed a Lipschitz-constrained skill

discovery method based on a distance-maximizing and controllability-aware distance function to overcome the bias toward static skills and to allow the agent to learn complex and far-reaching behaviors. Sharma et al. (2020) developed a method that simultaneously discovers predictable skills and learns their dynamics. In a follow-up work, Park & Levine (2023) addresses the problem of errors in predictive models by learning a transformed MDP, whose action space contains only easy to model and predictable actions. These works provide RL algorithms for unsupervised skill discovery that require *online* interaction with the environment and do not impose utility measures on the learned skills. In contrast, DOI gives a principled *offline* algorithm for maximizing diversity under imitation constraints.

A large body of research has focused on successor features (Dayan, 1993; Barreto et al., 2016), a powerful technique in RL for transfer of knowledge across tasks by capturing environmental dynamics, particularly promising for skill discovery when coupled with variational intrinsic motivation (Gregor et al., 2017; Barreto et al., 2018; Hansen et al., 2020) to enhance feature controllability, generalization, and task inference. In contrast to our work, these approaches do not impose performance constraints on the learned skills. Zahavy et al. (2022) cast the task of learning diverse skills, each achieving a near-optimal performance with respect to a given reward, into a constrained MDP setting with a physics-inspired diversity objective based on a minimum ℓ_2 distance between the successor features of distinct skills. However, this approach requires significant *online* interaction with the environment to learn the skills.

Numerous practical algorithms for offline RL have been proposed (Levine et al., 2020; Prudencio et al., 2022), including methods based on advantage-weighted behavioral cloning (Nair et al., 2020; Wang et al., 2020), conservative strategies to stay close to the original data distribution (Kumar et al., 2020; Cheng et al., 2022) and using only on-data samples (Kostrikov et al., 2022; Xu et al., 2023). While these methods excel at learning a policy that maximizes a fixed reward, they are not directly applicable in our setting, which has a non-stationary reward that depends on: i) the log-likelihood of a skill discriminator, and ii) Lagrange multipliers. In addition, these techniques cannot be used to i) train a skill discriminator and ii) estimate a KL divergence offline.

Naive importance sampling approaches for off-policy estimation are known to suffer from unbounded variance in the infinite horizon setting, a problem known in the literature as "the curse of horizon". Liu et al. (2018); Mousavi et al. (2020) addressed this challenge by providing theoretical foundations and a principled off-policy algorithm, using a backward Bellman operator, that avoids exploding variance by applying importance sampling to state-visitation distributions, and by providing practical solutions in Reproducing Kernel Hilbert Spaces. An alternative research direction in off-policy estimation, referred to as "Distribution Correction Estimation (DICE)", has introduced innovative techniques, with Nachum et al. (2019a) mitigating variance with importance sampling, Nachum et al. (2019b) enabling policy gradient from off-policy data without importance weighting, Kim et al. (2022) stabilizing offline imitation learning with imperfect demonstrations, Zhang et al. (2020) improving density ratio estimation, Dai et al. (2020) providing high-confidence off-policy evaluation. Subsequently, Xu et al. (2021) applied this approach to offline RL and demonstrated its effectiveness in continuous control tasks. Our work uses a DICE-based off-policy approach similar to OptiDICE (Lee et al., 2021; 2022) for estimating importance ratios, while considering a constrained formulation with a mutual information objective and KL-divergence imitation constraints.

3 Preliminaries

We utilize the framework of Markov decision processes (MDPs) (Puterman, 2014), where an MDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \rho_0, \gamma)$ denoting the state space, action space, reward mapping $\mathcal{R} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$, stochastic transition kernel $\mathcal{P}(s'|s, a)$, initial state distribution $\rho_0(s)$ and discount factor γ. A policy $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$ defines a probability distribution over the action space $\mathcal{A}$ conditioned on the state, where $\Delta(\cdot)$ stands for the probability simplex.

Given a policy π, the corresponding state-action occupancy measure is defined by

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

and its associated state occupancy $d^\pi(s)$ is given by marginalizing over the action space $\sum_{a\in\mathcal{A}} d^\pi(s,a)$.

In the skill discovery setting, $z \sim p(Z)$ denotes a fixed latent skill on which we condition a policy $\pi_z : S \times Z \mapsto \Delta(\mathcal{A})$. We will treat $p(Z)$ as a categorical distribution over a discrete set Z of $|Z|$ many distinct indicator vectors in $\mathbb{R}^{|Z|}$. The skill-conditioned policy π_z induces a state occupancy denoted by $d_z(s) := d^{\pi_z}(s)$, and when it is clear from the context we will refer to $d_z(s)$ as a "skill".

We consider an offline setting with access to the following datasets: i) $\mathcal{D}_E$ sampled from an expert state occupancy $d_E(S)$; and ii) $\mathcal{D}_O$ sampled from a state-action occupancy $d_O(S,A)$ generated by a mixture of behaviors. Similar to Ma et al. (2022), our analysis makes the following assumption, which requires that the offline state occupancy d_O sufficiently covers the expert's state occupancy d_E, a prerequisite for successful imitation learning. Although this assumption is not required in practice, it ensures well-defined state occupancy measures (i.e., avoiding division by zero).

Assumption 3.1 (Expert coverage)**.** *We assume that* $d_E(s) > 0$ *implies* $d_O(s) > 0$.

4 Method

Given an expert and a coverage dataset as above, we aim to solve *offline* the constrained optimization problem

$$\max_{\{d_z(S)\}_{z\in Z}} \quad \mathcal{I}(S;Z) \tag{1}$$

$$\text{subject to} \quad \mathrm{D_{KL}}\left(d_z(S)||d_E(S)\right) \le \epsilon \quad \forall z, \tag{2}$$

where $\mathcal{I}(S;Z)$ denotes the mutual information between states and skills. The identity $\mathcal{I}(S;Z) = \mathbb{E}_{p(z)}\mathrm{KL}(d_z(S)||\mathbb{E}_{z'}d_{z'}(S))$ shows an important geometric perspective that maximizing mutual information is equivalent to finding a set of $|Z|$ skills whose state occupancies $d_z(S)$ correspond as points on a probability simplex such that these points are positioned on the boundary of an ellipsoid and the pairwise distance between each point and the ellipsoid center is maximized (Zahavy et al., 2021; Eysenbach et al., 2022).

Henceforth, we shall make use of color coding to highlight the diversity signal in blue and the imitation signal in orange. The preceding problem formulation and our algorithmic framework can be easily extended to capture: i) objectives in (1) that combine conditional mutual information (c.f. DADS in (Sharma et al., 2020)) and information gain (c.f. DISDAIN in (Strouse et al., 2022)); and ii) general f-divergence constraints in (2), see Nachum & Dai (2020); Ma et al. (2022). We leave the study of these variants for future work.

Since maximizing the mutual information is generally intractable, in line with previous work (Eysenbach et al., 2019) we assume that the latent skills are sampled uniformly at random, i.e., $p(z) = \frac{1}{|Z|}$, and as a trackable surrogate we consider instead the following variational lower bound

$$\mathcal{I}\left(S;Z\right) \ge \mathbb{E}_{p(z),d_z(s)}\left[\log q(z|s)\right] + \mathcal{H}\left(p(z)\right) = \sum_z \mathbb{E}_{d_z(s)}\left[\frac{\log\left(|Z|q(z|s)\right)}{|Z|}\right]. \tag{3}$$

Here with $q(z|s)$ we denote a skill-discriminator tasked with distinguishing between latent skills.

Ma et al. (2022) proposed an offline algorithm (SMODICE) that on input an expert dataset $\mathcal{D}_E \sim d_E(S)$ and a coverage dataset $\mathcal{D}_O \sim d_O(S,A)$ such that $\mathcal{D}_E \subset \mathrm{States}[\mathcal{D}_O]$, trains a policy $\pi_{\widetilde{E}}$ which optimizes the problem

$$\min_\pi \mathrm{D_{KL}}\left(d^\pi(S)||d_E(S)\right), \tag{4}$$

and outputs the associated expert ratios $\eta_{\widetilde{E}}(s,a) = d_{\widetilde{E}}(s,a)/d_O(s,a)$ for every state-action pair $(s,a) \in \mathcal{D}_O$, where $d_{\widetilde{E}}(s,a)$ denotes the state-action occupancy induced by the recovered expert policy $\pi_{\widetilde{E}}$.

An important observation is that given the expert ratios $\eta_{\widetilde{E}}(s,a)$, the state constraints (2) can be relaxed to constraints with respect to the recovered expert state-action occupancy $d_{\widetilde{E}}(s,a)$. While in theory this relaxation restricts the imitation to the state-action occupancy of a specific expert, it also admits a simpler estimator (see Lemma 4.3) that is more stable to compute, yields faster runtime performance in practice, and simultaneously provides enough capacity for diversity by increasing the level ϵ. More specifically, for each

latent skill z we replace the state constraint (2) with the following state-action constraint

$$\mathrm{D}_{\mathrm{KL}}\left(d_z(S,A)||d_{\widetilde{E}}(S,A)\right) \leq \epsilon. \tag{5}$$

We focus on a reduction of CMDPs to MDPs using gradient-based techniques, known as Lagrangian methods (Borkar, 2005; Bhatnagar & Lakshmanan, 2012; Tessler et al., 2019). In contrast to prior work on CMDP, which has focused primarily on linear objectives and constraints, we consider the nonlinear setting with convex objectives and constraints. More specifically, we seek to maximize the right-hand side of eq. (3) subject to eq. (5). Solving this problem is equivalent to

$$\max_{\substack{d_z(s,a)\\ q(z|s)}} \min_{\lambda \geq 0} \sum_z \mathbb{E}_{d_z(s)}\left[\frac{\log\left(|Z| q(z|s)\right)}{|Z|}\right] + \sum_z \lambda_z \left[\epsilon - \mathrm{D}_{\mathrm{KL}}\left(d_z(S,A)||d_{\widetilde{E}}(S,A)\right)\right], \tag{6}$$

where with λ_z we denote the Lagrange multiplier corresponding to latent skill z.

4.1 Approximation Scheme

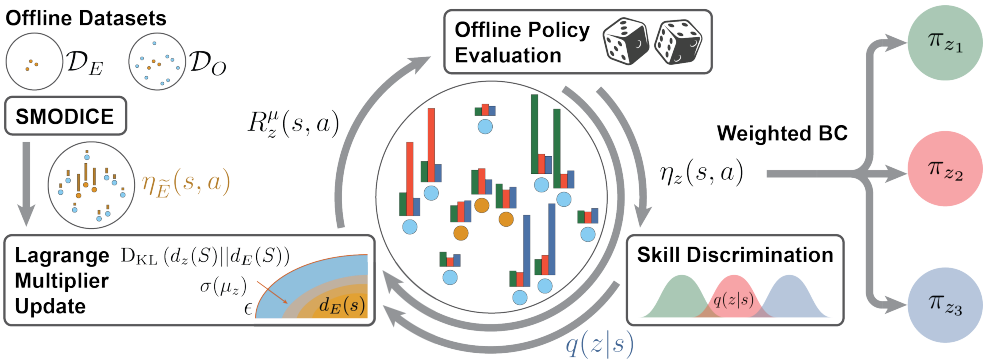


Figure 2: Illustration of Algorithm 1. We compute expert importance ratios $\eta_{\widetilde{E}}(s,a)$ by running SMODICE on the offline datasets $\mathcal{D}_E$ and $\mathcal{D}_O$. These expert ratios are then used in the alternating scheme described in Subsec. 4.1 to obtain the importance ratios $\eta_z(s,a)$ (with support in $\mathcal{D}_O$) for each skill z. Specifically, the skill-ratios $\eta_z(s,a)$ are computed by a DICE-like offline policy evaluation algorithm on input a reward $R_z^\mu(s,a)$ that balances skill diversity (skill-discriminator $q(z|s)$) and expert imitation (importance ratios $\eta_{\widetilde{E}}(s,a)$).

We use a popular heuristic, known in the literature as *alternating optimization*, to approximately compute a local optimum of Problem (6). More precisely, the method alternates between optimizing each model while holding all others fixed, and iteratively refines the solution until convergence is reached or a stopping criterion is met. Furthermore, as we can guarantee in practice that the Lagrange multipliers λ are always positive, we consider Problem (6) with $\lambda > 0$, that is

$$\max_{\substack{d_z(s,a)\\ q(z|s)}} \min_{\lambda > 0} \sum_z \lambda_z \Big\{\epsilon + \mathbb{E}_{d_z(s,a)}\left[R_z^\lambda(s,a)\right] - \mathrm{D}_{\mathrm{KL}}\left(d_z(S,A)||d_O(S,A)\right)\Big\}, \tag{7}$$

where

$$R_z^\lambda(s,a) := \underbrace{\frac{1}{\lambda_z}}_{\text{Constraint Violation}} \underbrace{\frac{\log\left(q(z|s)|Z|\right)}{|Z|}}_{\text{Skill Diversity}} + \underbrace{\log \eta_{\widetilde{E}}(s,a)}_{\text{Expert Imitation}}. \tag{8}$$

The reward in (8) is derived in Supp. C and relies on the following equality (see Supp. D.3) $\mathrm{D}_{\mathrm{KL}}(d_z(S,A)||d_{\widetilde{E}}(S,A)) = \mathrm{D}_{\mathrm{KL}}(d_z(S,A)||d_O(S,A)) - \mathbb{E}_{d_z(s,a)}[\log(d_{\widetilde{E}}(s,a)/d_O(s,a))]$ and the definition of $\eta_{\widetilde{E}}(s,a) = d_{\widetilde{E}}(s,a)/d_O(s,a)$.

Intuitively, the reward $R_z^\lambda(s,a)$ balances between diversity and KL-closeness to the expert state-action occupancy. The Lagrange multiplier λ_z scales down the log-likelihood of the skill-discriminator $q(z|s)$, effectively

reducing the diversity signal, when the state-action occupancy $d_z(S, A)$ violates the KL-divergence constraint (5), and vice versa. Each term in the reward (8) involves a separate optimization procedure, which will be described in the next section.

4.2 Approximation Phases

Using the alternating optimization scheme, Algorithm 1 decomposes into the following three optimization phases. In PHASE 1, we train a value function $V_z^\star$, ratios $\eta_z(s, a)$ and a skill-conditioned policy π_z. In PHASE 2, we train a skill-discriminator $q(z|s)$. Then in PHASE 3, we compute a KL constraint estimator ϕ_z and update accordingly the Lagrange multipliers λ_z. In addition, we perform a preprocessing phase to compute the expert ratios $\eta_{\widetilde{E}}(s, a)$ by invoking the SMODICE algorithm.

4.2.1 Phase 1

With fixed skill-discriminator $q(z|s)$ and Lagrange multipliers $\lambda > 0$, Problem (7) becomes

$$\max_{\{d_z(s,a)\}_{z\in Z}} \sum_z \lambda_z \Big\{ \mathbb{E}_{d_z(s,a)} \left[R_z^\lambda(s, a) \right] - \mathrm{D_{KL}} \left(d_z(S, A) || d_O(S, A) \right) \Big\}, \tag{9}$$

or equivalently for every skill z:

$$\begin{aligned} \max_{d_z(s,a)\geq 0} \quad & \mathbb{E}_{d_z(s,a)} \left[R_z^\lambda(s, a) \right] - \mathrm{D_{KL}} \left(d_z(S, A) || d_O(S, A) \right) \\ \text{subject to} \quad & \textstyle\sum_a d_z(s, a) = (1 - \gamma)\rho_0(s) + \gamma \mathcal{T} d(s) \qquad \forall s, \end{aligned} \tag{10}$$

where we denote with $\mathcal{T}$ the transition operator: $\mathcal{T}d(s') = \sum_{s,a} \mathcal{P}(s'|s, a)d(s, a)$.

Assumption 4.1 (Strict Feasibility)**.** *We assume there exists a solution such that the constraints* (10) *are satisfied and* $d(s, a) > 0$ *for all states-action pairs* $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Using Lagrange duality, Assum. 4.1 (which implies strong duality) and the Fenchel conjugate (see Supp. B), Nachum & Dai (2020, Sec. 6) and Ma et al. (2022, Theorem 2) showed that Problem 10 shares the same optimal value as the following optimization problem

$$V^\star = \arg\min_{V(s)} (1 - \gamma)\mathbb{E}_{s\sim\rho_0} \left[V(s) \right] + \log \mathbb{E}_{d_O(s,a)} \exp \left\{ R_z^\lambda(s, a) + \gamma \mathcal{T} V(s, a) - V(s) \right\}, \tag{11}$$

where $\mathcal{T}V(s, a) := \mathbb{E}_{\mathcal{P}(s'|s,a)} V(s')$. Moreover, the primal optimal solution is given by

$$\eta_z(s, a) := \frac{d_z^\star(s, a)}{d_O(s, a)} = \mathrm{softmax}_{d_O(s,a)} \left(R_z^\lambda(s, a) + \gamma \mathcal{T} V_z^\star(s, a) - V_z^\star(s) \right), \tag{12}$$

where $\mathrm{softmax}_{p(x)}(g(x)) = \exp\{g(x)\} \Big/ \mathbb{E}_{p(x')}[\exp\{g(x')\}]$. These ratios $\eta_z(s, a)$ are then used to design an offline importance-weighted sampling procedure that, for an arbitrary function f, satisfies

$$\mathbb{E}_{p(z)} \mathbb{E}_{d_z^\star(s,a)} [f(s, a, z)] = \mathbb{E}_{p(z)} \mathbb{E}_{d_O(s,a)} [\eta_z(s, a) f(s, a, z)]. \tag{13}$$

Afterwards, the optimal skill-conditioned policy $\pi_z^\star$ is trained offline using a weighted behavioral cloning, which is obtained by setting $f(s, a, z) = \log(\pi_z(a|s))$ and maximizing the RHS of eq. (13) over all skill-conditioned policies π_z. In practice, gradient descent is used for optimization.

4.2.2 Phase 2

We now give an offline procedure for training a skill-discriminator $q(z|s)$, which takes as input ratios $\eta_z(s, a)$ of a skill-conditioned policy $\pi_z^\star$. The proof is presented in Supp. D.2.

Lemma 4.2. *Given ratios* $\eta_z(s, a)$*, using eq.* (13) *applied with* $f(s, a, z) = \log(q(z|s))$*, we can compute offline an optimal skill-discriminator* $q^\star(z|s)$*. In particular, we optimize by gradient descent the following optimization problem* $\max_{q(z|s)} \mathbb{E}_{p(z)} \mathbb{E}_{d_O(s,a)} \left[\eta_z(s, a) \log \left(q(z|s) \right) \right]$.

The skill-conditioned policy $\pi_z^\star$ (PHASE 1) and the skill-discriminator $q^\star$ (PHASE 2), allow us to maximize *offline* the variational lower bound in eq. (3) and thus skill diversity. It remains to estimate possible constraint violations in eq. (5) and to update the Lagrange multipliers accordingly.

4.2.3 Phase 3

With fixed skill-discriminator $q^\star(z|s)$ and skill-conditioned policy $\pi_z^\star(s)$, Problem (7) reduces to $\min_{\lambda>0} \sum_z \lambda_z[\epsilon - \mathrm{D}_{\mathrm{KL}}(d_z^\star(S,A)||d_{\widetilde{E}}(S,A))]$. We optimize the Lagrange multipliers by gradient descent. To this end, we now give an offline estimator of the KL-divergence term. The proof is presented in Supp. D.3.

Lemma 4.3. *Given skill-conditioned policy ratios $\eta_z(s,a)$ and expert ratios $\eta_{\widetilde{E}}(s,a)$, using eq. (13) applied with $f(s,a,z) = \log(\eta_z(s,a)/\eta_{\widetilde{E}}(s,a))$, we can compute offline an estimator of $\mathrm{D}_{\mathrm{KL}}(d_z^\star(S,A)||d_{\widetilde{E}}(S,A))$ which is given by $\phi_z := \mathbb{E}_{d_O(s,a)}[\eta_z(s,a)\log(\eta_z(s,a)/\eta_{\widetilde{E}}(s,a))]$.*

We note that the ratios $\eta_z(s,a)$ and $\eta_{\widetilde{E}}(s,a)$ are computed only on state-action pairs within the offline dataset $\mathcal{D}_O$. Furthermore, in practice, we ensure that these ratios are strictly positive, so that the KL estimator ϕ_z is well defined and bounded.

5 Algorithm

Our optimization method consists of three phases, each of which optimizes a specific model and fixes the remaining ones. It is important to emphasize that in contrast to prior work, our problem formulation considers an optimization problem with constraints. Furthermore, the reward function in eq. (8) is non-stationary, since it depends on the bounded Lagrange multipliers that balance diversity ($\log q(z|s)$) and expert imitation ($\log \eta_{\widetilde{E}}(s,a)$). This has significant algorithmic implications, as it requires solving a sequence of standard RL problems, each of which admits offline policy evaluation.

To smooth the transition of the reward signal between successive iterations, we enforce a slow change of the Lagrange multipliers. More specifically, we use the technique of bounded Lagrange multipliers (Stooke et al., 2020; Zahavy et al., 2022), which applies a Sigmoid transformation $\lambda = \sigma(\mu)$ component-wise to unbounded variables $\mu \in \mathbb{R}^{|Z|}$, so that the effective reward is a convex combination of a diversity term and an expert imitation term. In practice, this transformation ensures that $\lambda > 0$. Hence, the reward for each latent skill z becomes

$$R_z^\mu(s,a) := (1-\sigma(\mu_z))\,\frac{\log\left(q^\star(z|s)|Z|\right)}{|Z|} + \sigma(\mu_z)\log \eta_{\widetilde{E}}(s,a). \tag{14}$$

We now present the resulting multi-phase optimization procedure in Algorithm 1. For the offline training of the policy (in Phase 1), the skill-discriminator (in Phase 2), and the estimation of the KL divergence value (in Phase 3), we use importance sampling eq. (13) and give the corresponding empirical estimators in Supp. E. Our practical implementation leverages the power of neural networks and deep learning techniques for accurate function approximation. More specifically, we train an expert policy $\pi_{\widetilde{E}}$, a skill-conditioned policy $\{\pi_z\}_{z\in Z}$ and a value function $\{V_z\}_{z\in Z}$. While practically convenient, this means that each phase of Algorithm 1 is only approximately solved. In particular, we do not solve the optimization problem to optimality in each phase, but rather perform a few gradient descent steps.

We have found that fitting the skill-discriminator $q(z|s)$ is prone to collapse to the uniform distribution. To alleviate this issue, in addition to the variational lower bound objective (3), we add the DISDAIN information gain term, proposed in (Strouse et al., 2022). This bonus term is an entropy-based disagreement penalty that estimates the epistemic uncertainty of the skill-discriminator, and is implemented in practice by an ensemble of randomly initialized skill-discriminators. Due to the high initial disagreement on unvisited states, this intrinsic reward provides a strong exploration signal and leads to the discovery of more diverse behaviors. Intuitively, for states with small epistemic uncertainty, the skill-discriminator (averaged over the ensemble members) should reliably discriminate between latent skills, thus making the intrinsic reward of the skill-discriminator's log-likelihood more accurate.

Algorithm 1 Diverse Offline Imitation (DOI)

Input: a state-only expert dataset $\mathcal{D}_E \sim d_E(S)$ and an offline dataset $\mathcal{D}_O \sim d_O(S, A)$ such that $\mathcal{D}_E \subset \text{States}[\mathcal{D}_O]$.
Pre-compute a state-discriminator $c^\star : \mathcal{S} \to (0, 1)$ via optimizing the following objective with the gradient penalty in (Gulrajani et al., 2017) $\min_c \mathbb{E}_{d_E(s)}[\log c(s)] + \mathbb{E}_{d_O(s)}[\log(1 - c(s))]$
Apply **Phase 1** with reward $R(s, a) = \log \frac{c^\star(s)}{1-c^\star(s)}$ to compute ratios $\eta_{\widetilde{E}}(s, a) = d_{\widetilde{E}}(s, a)/d_O(s, a)$ for all $s, a \in \mathcal{D}_O$

Repeat until convergence:
Phase 1. (Fixed Lagrange multipliers $\sigma(\mu)$ and skill-discriminator values $q^\star(z|s)$)
For each latent skill z:
compute a value function $V_z^\star$ optimizing eq. (11) with reward $R_z^\mu(s, a)$ in eq. (14)
compute ratios $\eta_z(s, a) = \text{softmax}_{d_O(s,a)} \left(R_z^\mu(s, a) + \gamma \mathcal{T} V_z^\star(s, a) - V_z^\star(s)\right)$ for all $s, a \in \mathcal{D}_O$
train a skill-conditioned policy $\pi_z^\star = \arg\max_{\pi_z} \mathbb{E}_{d_O(s,a)}[\eta_z(s, a) \log \pi_z(a|s)]$

Phase 2. (Fixed ratios $\eta_z(s, a)$ and bounded Lagrange multipliers $\sigma(\mu)$)
Train a skill-discriminator $q^\star = \arg\max_{q(\cdot|s)} \mathbb{E}_{p(z)} \mathbb{E}_{d_O(s,a)}[\eta_z(s, a) \log q(z|s)]$

Phase 3. (Fixed ratios $\eta_{\widetilde{E}}(s, a)$ and $\eta_z(s, a)$)
Compute for each latent skill z an estimator $\phi_z := \mathbb{E}_{d_O(s,a)}[\eta_z(s, a) \log(\eta_z(s, a)/\eta_{\widetilde{E}}(s, a))]$
Optimize the loss $\min_\mu \sum_z \sigma(\mu_z)(\epsilon - \phi_z)$

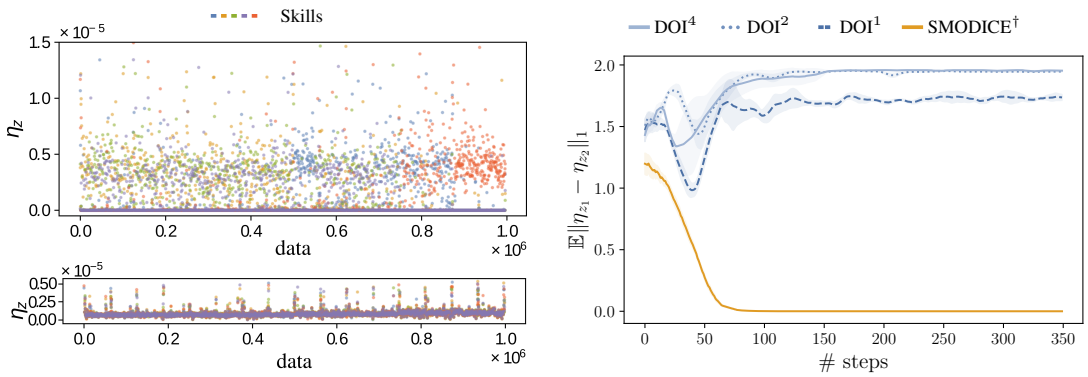
6 Experiments

For evaluation of our method we consider 12 degree-of-freedom quadruped robot SOLO12 (Grimminger et al., 2020), on a simple locomotion task in both the *simulation* and the *real* system. We complement this with an obstacle navigation task, in simulation, and demonstrate that some of the learned diverse skills robustly reach a target position while the expert fails. Furthermore, we provide evaluation on the ANT, WALKER2D, HALFCHEETAH and HOPPER environments from the standard D4RL benchmark (Fu et al., 2020).

6.1 Locomotion

Data collection. For the SOLO12 evaluation, we collected domain-randomized offline and expert data from simulation in the Isaac Gym (Makoviychuk et al., 2021), using pretrained policy checkpoints obtained by training the robot to track a certain speed of the base with the on-policy diversity maximization algorithm DOMiNiC (Cheng et al., 2024). We defer the data collection procedure to the Supp. G. The *expert dataset* was collected by using the best deterministic policy from the last checkpoint of the training procedure, which was trained to track forward velocity only without diversity objective. In contrast, the *offline dataset* was acquired by employing stochastic policies gathered from various checkpoints throughout the training of the expert, featuring multiple latent skills. More than half of the *offline dataset* was collected by a random Gaussian policy. In line with previous approaches by Kim et al. (2022) and Ma et al. (2022), our practical implementation aims to fulfill the expert coverage Assum. 3.1. To achieve this, we create the *coverage dataset* $\mathcal{D}_O$ by adding a small number of expert trajectories to the offline dataset, resulting in an *unlabeled* expert fraction of 1/160 in $\mathcal{D}_O$. We discard expert actions from the expert dataset to ensure that our algorithm does not have labeled access to them. The resulting *expert dataset* $\mathcal{D}_E$ is used to learn a state classifier $c(s)$. Then the SMODICE is executed to compute the importance ratios $\eta_{\widetilde{E}}(s, a)$, see Sec. 4. We trained the policy for 350 steps, where each step involves the stages described in Sec. 5. In each stage, we execute 200 epochs of batched training over the data.

Here with DOI$^\epsilon$ we denote an execution of Algorithm 1 with constraint threshold set to ϵ. We proceed by analyzing the learned DOI skills in three evaluation settings: i) over the fixed offline datasets; ii) a Monte Carlo on-policy evaluation in the simulator; and iii) the resulting clustering structure involving the offline and expert datasets, as well as the DOI skills and the SMODICE expert evaluated in the simulation.



(a) (b)

Figure 3: Data points separation by importance ratios $\eta_z(s, a)$, given different levels of ϵ in SOLO12. (a) Distribution of importance ratios $\eta_z(s, a)$ over the offline dataset $\mathcal{D}_O$ for distinct skills with DOI^4 ($\epsilon = 4$) (upper) and a skill-conditioned variant of SMODICE (lower). (b) Average ℓ_1 distance of ratios η_z belonging to distinct skills, depending on ϵ. The higher the value of ϵ, the greater the ℓ_1 distance. The shaded areas show the interval between the 0.25 and 0.75 quantiles, computed over 3 seeds.

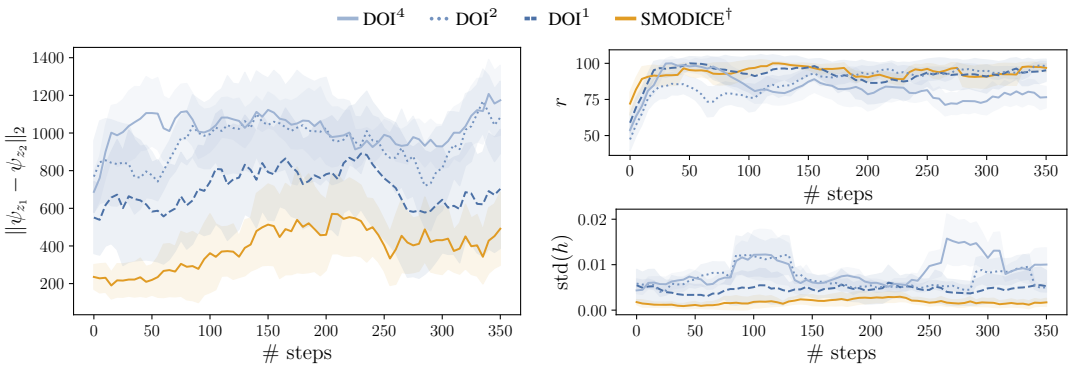
Importance ratios distance. In Figure 3, we measure the state-action occupancy $d_z(s, a)$ for each latent skill z through the proxy of importance ratios $\eta_z(s, a)$,[2] for different values of ϵ. As expected, a higher value of ϵ increases diversity, resulting in different importance ratios per skill for individual data points. This difference is then aggregated by computing an expected ℓ_1 distance between importance ratios of distinct skills, i.e., $\mathbb{E}\|\eta_{z_i} - \eta_{z_j}\|_1$, and is reported in Figure 3. We note that the looser the constraint (lighter color), the easier it is to diversify in the sense of η_z. Figure 3b shows the average ℓ_1 distance between skill importance vectors η_z over the dataset for $\epsilon \in \{0.0, 1.0, 2.0, 4.0\}$ (lighter color indicates higher ϵ). Moreover, the tighter the constraint (smaller ϵ), the smaller the difference between the distinct skill importance ratios.

To analyze the influence of the diversity objective on the learned skills, we consider as a baseline a skill-conditioned variant of (Ma et al., 2022), denoted SMODICE†, which does not have access to the skill discriminator $q(z|s)$. This is equivalent to DOI with fixed $\sigma(\mu_z) = 1$ in the reward eq. (14). We defer further experiments with fixed Lagrange multipliers to Supp. N. In Figure 3a, we observe diversification across the dataset assignment to skills when using DOI, whereas training an ensemble of skills with only expert imitation reward (i.e., $\sigma(\mu_z) = 1$) collapses to nearly the same importance ratios per skill per data point.

Successor features distance. We have further evaluated diversity on the Monte Carlo estimates of the expected successor features over the initial state, based on 30 policy rollouts per skill. The γ-discounted successor features (SFs) for state s are defined as $\psi_z(s) = \mathbb{E}_{d_z(s)}[\phi(s)]$, where $d_z(s)$ is the γ-discounted state occupancy for a skill policy π_z. With slight abuse of notation, we define $\psi_z = \mathbb{E}_{\rho_0(s)}[\psi_z(s)]$, the expected SFs over the initial state distribution. As a diversity metric, we take the expected ℓ_2 distance between the SFs of distinct skills, i.e., $\mathbb{E}\|\psi_{z_1} - \psi_{z_2}\|_2$. The results are presented in Figure 4 and are consistent with the proxy diversity metric. In particular, there is a correspondence between the offline data separation induced by the importance ratios η_z (see Figure 3a), and a higher distance between the expected SFs ψ_z (see Figure 4a). In terms of performance, DOI achieves a forward velocity comparable to the expert (see Figure 4a) while learning diverse skills with respect to base height h (see Figure 4b). We also observed that the multipliers $\sigma(\mu_z)$ are non-zero for all skills, indicating that the constraint is active. In addition, they stabilize at reasonable levels as training progresses, which we show in Supp. I for both the SOLO12 and ANT.

DOI skills form well-separated clusters. Here we conduct a controlled experiment with full trajectory information, which remains hidden to the DOI algorithm. In Figure 5, the Successor Features of each trajectory in the expert dataset are transformed by UMAP (McInnes et al., 2018) algorithm into 2D space. This

[2]For the computation of the skill-ratios $\eta_z(s, a)$, we choose a projection Π of the expert state (see Supp. K) that yields 3-dimensional planar and angular velocities of the robot's base in the base frame.



(a) (b)

Figure 4: (a) Average ℓ_2 distance between Monte Carlo estimates of successor features ψ_z of distinct skills; (b) return r as % of expert return and standard deviation of base height $\mathrm{std}_z(h)$. Both depend on ϵ for the SOLO12. The shaded areas show the interval between the 0.25 and 0.75 quantiles, computed over 3 seeds.

transformation is then used to map the SFs of each trajectory into 2D space for: i) the offline dataset, ii) the SMODICE expert evaluated in simulation, and iii) the learned DOI skills (red, green, blue, purple, cyan) also evaluated in simulation. The diversity of learned DOI skills is reflected in a well-separated cluster structure.

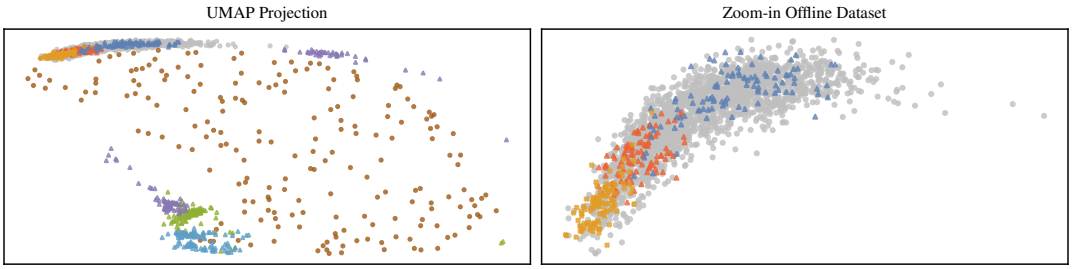


Figure 5: Successor Features projection onto 2D space using the UMAP algorithm.

6.2 Robust Obstacle Navigation

Data collection. Similarly to the locomotion task in Subsec. 6.1, both *expert dataset* and *offline dataset* were generated from pretrained policy checkpoints from training a robot to navigate in the terrain of obstacles with fixed time limit using the DOMiNiC (Cheng et al., 2024) algorithm. Unlike the previous task, the *expert dataset* was collected using the best deterministic skill-conditioned policy from the last checkpoint of the training procedure, which exhibits diverse strategies to navigate the obstacle terrain, including bypassing it from both sides or climbing over it. The *offline dataset* was acquired through rolling out stochastic policies gathered from multiple checkpoints with multiple skills. Both *expert dataset* and *offline dataset* were collected in a terrain of a single obstacle of a fixed height of 0.2 meters. Similar to Subsec. 6.1, we create the *coverage dataset* $\mathcal{D}_O$ by adding a small number of expert trajectories to the offline dataset. For details on collecting the dataset for the obstacle navigation task, we refer interested readers to the Supp. G.

Multi-modal expert limitations. Deriving a single policy by SMODICE from expert demonstrations, even in the setting when the dataset was collected from diverse expert strategies, may lack robustness to distribution shifts. This observation emphasizes the need for diverse policy extraction. To illustrate this with a concrete example, consider a scenario where a SOLO12 robot navigates around a single box obstacle to reach a target position behind it. The target position can be reached either from the sides (left or right) of the box or by climbing over it (the less robust path). In our experiments, the expert dataset $\mathcal{D}_E$ contains all of the above strategies. As shown in Figure 6, for boxes with a height of at least 0.3 meters, the SMODICE expert consistently positions itself in front of the box and thus fails to robustly reach the target position.

Extracting robust policies. In Figure 6, we analyze return distributions and sampled trajectories for box heights of $\{0.3, 0.6\}$ meters. The SMODICE expert predominantly fails to reach the target position, due to a bias towards climbing over the box. In contrast, a DOI skill consistently chooses the left side, which leads robustly to the target position and achieves a superior return distribution. However, it is important to note that not all learned DOI skills are robust. Hence, a subsequent selection process is required. Further details about all learned DOI skills, their return distributions and sampled trajectories, different box heights, and additional experimental information are presented in Supp. M.

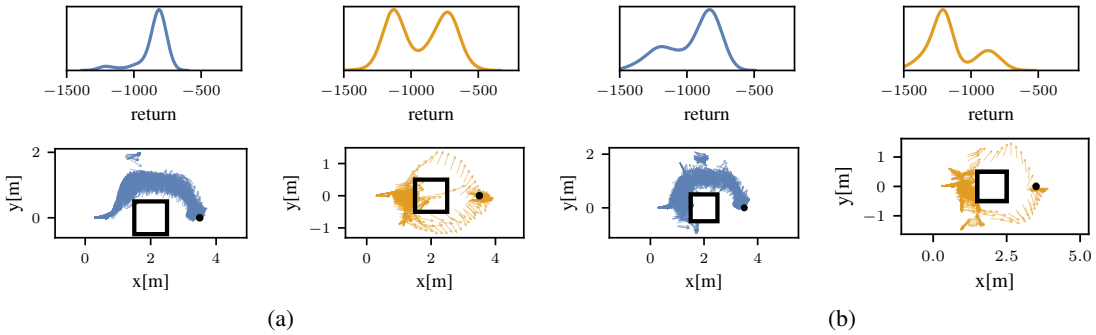


Figure 6: Return distributions and sampled trajectories of SMODICE and a DOI skill for terrains with box height (a) 0.3 and (b) 0.6. The heights of the boxes are out-of-distribution for the SMODICE, which tends to get stuck in front of the box due to a bias towards climbing over it. In contrast, the robust DOI skill takes a detour to the left side of the box.

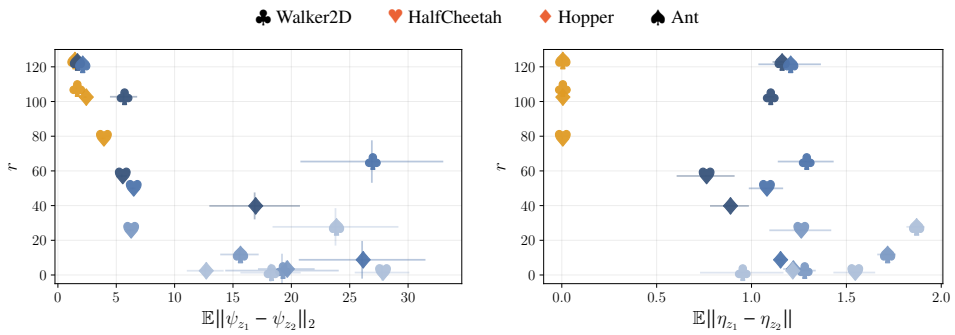
6.3 Standard D4RL Environments

We consider the case where we have offline data generated from a random policy mixed with a small amount of expert trajectories.[3] Figure 7 shows the results for both the expected ℓ_2 distance between the SFs or the importance ratios η_z of distinct skills. We normalize the state feature $\phi(s)$ when comparing SFs ψ_z across environments in Figure 7a. In most cases, we report a trade-off between the average skill return and the imitation level ϵ. The larger the imitation slack ϵ, the more diverse the skills become, but at the cost of lowering the average return, and vice versa. Nevertheless, in Figure 7a we show that ϵ retains some controllability over diversity. The WALKER2D is particularly sensitive to relaxation of the occupancy constraint with respect to performance. We hypothesize that this is due to the fact that the space of policies that achieve a stable gait is very restrictive, resulting in a significant loss of task return for even a small amount of skill diversity. In contrast, the ANT exhibits high stability, with several skills achieving close to expert performance in terms of r. These results are also consistent with SMODICE expert policies used for computing $\eta_{\widetilde{E}}(s, a)$ (see Supp. H).

7 Conclusion

We proposed DOI, a principled offline RL algorithm for unsupervised skill discovery that, in addition to maximizing diversity, ensures that each learned skill imitates state-only expert demonstrations to a certain degree. Our main analytical contribution is to connect Fenchel duality, reinforcement learning, and unsupervised skill discovery to maximize a mutual information objective subject to KL-divergence state occupancy constraints. We have shown that DOI can diversify offline policies for a 12-DoF quadruped robot (in simulation and in reality) and for several environments from the standard D4RL benchmark in terms of both ℓ_2 distance of expected successor features and ℓ_1 distance of importance ratios, which is visible from the data separation induced by $\eta_z(s, a)$ among skills. The importance ratio distance, computed offline, is a robust indicator of diversity, which aligns with the online Monte Carlo diversity metric of expected successor features. The resulting skill diversity naturally entails a trade-off in task performance. We can control the amount of diversity via an imitation level ϵ, which ensures that distinct skills remain close to the expert in terms of state-action

[3]The same setting was considered by Ma et al. (2022).



(a) (b)

Figure 7: Results on D4RL environments with offline data collected from a random policy for $\epsilon = 0.0, 0.5, 1.0, 2.0, 4.0$. In figure (a) we observe the tradeoff between average skill return and average successor features distance over skills. In figure (b), we report the tradeoff w.r.t. average ℓ_1 distance of importance ratios η_z. The lines indicate the standard deviation computed over 3 seeds.

occupancy, which also indirectly controls task performance loss. A promising direction for future research is to impose constraints on the value function of each skill to ensure near-optimal task performance.

Limitations

Our approach, while promising, is not without limitations. The diversity objective, which is given by a variational lower bound on the mutual information between states and skills, necessitates the training of a skill-discriminator. This design choice, however, presents several practical challenges: i) the single-step policy and skill-discriminator update in the offline setting does not provide as accurate a policy estimate as sampling a Monte Carlo trajectory in the online setting (Eysenbach et al., 2019); ii) this inaccuracy, when combined with the non-stationary reward (bounded Lagrange multipliers and skill-discriminator), could result in a skill-discriminator that fails to accurately discriminate among skills; and iii) while the introduction of an additional information gain term, as in Strouse et al. (2022), can alleviate this issue, its effect could fade quickly and serve only as an initial diversity boost in the offline setting. Furthermore, the current paradigm is well-suited for a discrete number of skills, leaving open the important questions of extending our framework to infinitely many skills and addressing in a principled way the practical challenges surrounding the skill-discriminator training. We leave these important open questions for future work.

Acknowledgments

We acknowledge the support from the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039B). Georg Martius is a member of the Machine Learning Cluster of Excellence, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanys Excellence Strategy EXC number 2064/1 Project number 390727645. This work was supported by the ERC - 101045454 REAL-RL. Pavel Kolev was supported by the Cyber Valley Research Fund and the Volkswagen Stiftung (No 98 571). We would like to thank Nico Gürtler for helpful comments on an earlier version of this paper.

References

Robohive – a unified framework for robot learning. https://sites.google.com/view/robohive, 2020. URL https://sites.google.com/view/robohive.

Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. Variational option discovery algorithms. *CoRR*, abs/1807.10299, 2018. URL http://arxiv.org/abs/1807.10299.

Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado Van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. *arXiv preprint arXiv:1606.05312*, 2016.

Andre Barreto, Diana Borsa, John Quan, Tom Schaul, David Silver, Matteo Hessel, Daniel Mankowitz, Augustin Zidek, and Remi Munos. Transfer in deep reinforcement learning using successor features and generalised policy improvement. In *International Conference on Machine Learning*, pp. 501–510. PMLR, 2018.

Shalabh Bhatnagar and K Lakshmanan. An online actor–critic algorithm with function approximation for constrained markov decision processes. *Journal of Optimization Theory and Applications*, 153(3):688–708, 2012.

Vivek S Borkar. An actor-critic algorithm for constrained markov decision processes. *Systems & control letters*, 54(3):207–213, 2005.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael S. Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: vision-language-action models transfer web knowledge to robotic control. *CoRR*, abs/2307.15818, 2023. doi: 10.48550/arXiv.2307.15818. URL https://doi.org/10.48550/arXiv.2307.15818.

Victor Campos, Alexander Trott, Caiming Xiong, Richard Socher, Xavier Giró-i-Nieto, and Jordi Torres. Explore, discover and learn: Unsupervised discovery of state-covering skills. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1317–1327. PMLR, 2020. URL http://proceedings.mlr.press/v119/campos20a.html.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3852–3878. PMLR, 2022.

Jin Cheng, Marin Vlastelica, Pavel Kolev, Chenhao Li, and Georg Martius. Learning diverse skills for local navigation under multi-constraint optimality. In *IEEE International Conference on Robotics and Automation, ICRA 2024, PACIFICO, Yokohama, May 13th to 17th, 2024*. IEEE, 2024.

Bo Dai, Niao He, Yunpeng Pan, Byron Boots, and Le Song. Learning from conditional distributions via dual embeddings. In *Artificial Intelligence and Statistics*, pp. 1458–1467. PMLR, 2017.

Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Coindice: Off-policy confidence interval estimation. *Advances in neural information processing systems*, 33:9398–9411, 2020.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993. doi: 10.1162/neco.1993.5.4.613.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=SJx63jRqFm.

Benjamin Eysenbach, Ruslan Salakhutdinov, and Sergey Levine. The information geometry of unsupervised reinforcement learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=3wU2UX0voE.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Anirudh Goyal, Riashat Islam, Daniel Strouse, Zafarali Ahmed, Hugo Larochelle, Matthew M. Botvinick, Yoshua Bengio, and Sergey Levine. Infobot: Transfer and exploration via the information bottleneck. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rJg8yhAqKm.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Skc-Fo4Yg.

F. Grimminger, A. Meduri, M. Khadiv, J. Viereck, M. Wüthrich, M. Naveau, V. Berenz, S. Heim, F. Widmaier, T. Flayols, J. Fiene, A. Badri-Spröwitz, and L. Righetti. An open torque-controlled modular robot architecture for legged locomotion research. *IEEE Robotics and Automation Letters*, 5(2):3650–3657, 2020.

Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5767–5777, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html.

Steven Hansen, Will Dabney, André Barreto, David Warde-Farley, Tom Van de Wiele, and Volodymyr Mnih. Fast task inference with variational intrinsic successor features. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=BJeAHkrYDS.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, 2016.

Geon-Hyeong Kim, Seokin Seo, Jongmin Lee, Wonseok Jeon, HyeongJoo Hwang, Hongseok Yang, and Kee-Eung Kim. DemoDICE: Offline imitation learning with supplementary imperfect demonstrations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=BrPdX1bDZkQ.

Jaekyeom Kim, Seohong Park, and Gunhee Kim. Unsupervised skill discovery with bottleneck option learning. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139, pp. 5572–5582. PMLR, 2021a. URL http://proceedings.mlr.press/v139/kim21j.html.

Jaekyeom Kim, Seohong Park, and Gunhee Kim. Unsupervised skill discovery with bottleneck option learning. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5572–5582. PMLR, 2021b. URL http://proceedings.mlr.press/v139/kim21j.html.

A.S. Klyubin, D. Polani, and C.L. Nehaniv. Empowerment: a universal agent-centric measure of control. In *IEEE Congress on Evolutionary Computation*, volume 1, pp. 128–135 Vol.1, 2005. URL https://ieeexplore.ieee.org/document/1554676.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=68n2s9ZJWF8.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pp. 6120–6130. PMLR, 2021.

Jongmin Lee, Cosmin Paduraru, Daniel J Mankowitz, Nicolas Heess, Doina Precup, Kee-Eung Kim, and Arthur Guez. Coptidice: Offline constrained reinforcement learning via stationary distribution correction estimation. In *International Conference on Learning Representations*, 2022.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *CoRR*, abs/2005.01643, 2020.

Pierre-Alexandre Léziart, Thomas Flayols, Felix Grimminger, Nicolas Mansard, and Philippe Souères. Implementation of a reactive walking controller for the new open-hardware quadruped solo-12. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5007–5013. IEEE, 2021.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 5361–5371, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/dda04f9d634145a9c68d5dfe53b21272-Abstract.html.

Yecheng Jason Ma, Andrew Shen, Dinesh Jayaraman, and Osbert Bastani. Versatile offline imitation from observations and examples via regularized state-occupancy matching. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14639–14663. PMLR, 2022. URL https://proceedings.mlr.press/v162/ma22a.html.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.*, 3(29):861, 2018. doi: 10.21105/JOSS.00861. URL https://doi.org/10.21105/joss.00861.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/e00406144c1e7e35240afed70f34166a-Abstract.html.

Ali Mousavi, Lihong Li, Qiang Liu, and Denny Zhou. Black-box off-policy estimation for infinite-horizon reinforcement learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=S1ltg1rFDS.

Ofir Nachum and Bo Dai. Reinforcement learning via fenchel-rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in Neural Information Processing Systems*, 32, 2019a.

Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience, 2019b.

Ashvin Nair, Murtaza Dalal, Abhishek Gupta, and Sergey Levine. Accelerating online reinforcement learning with offline datasets. *CoRR*, abs/2006.09359, 2020.

Seohong Park and Sergey Levine. Predictable MDP abstraction for unsupervised model-based RL. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27246–27268. PMLR, 2023. URL https://proceedings.mlr.press/v202/park23i.html.

Seohong Park, Jongwook Choi, Jaekyeom Kim, Honglak Lee, and Gunhee Kim. Lipschitz-constrained unsupervised skill discovery. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=BGvt0ghNgA.

Seohong Park, Kimin Lee, Youngwoon Lee, and Pieter Abbeel. Controllability-aware unsupervised skill discovery. *CoRR*, abs/2302.05103, 2023. doi: 10.48550/arXiv.2302.05103. URL https://doi.org/10.48550/arXiv.2302.05103.

Rafael Figueiredo Prudencio, Marcos R. O. A. Maximo, and Esther Luna Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *CoRR*, abs/2203.01387, 2022.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HJgLZR4KvH.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID lagrangian methods. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9133–9143. PMLR, 2020. URL http://proceedings.mlr.press/v119/stooke20a.html.

DJ Strouse, Kate Baumli, David Warde-Farley, Volodymyr Mnih, and Steven Stenberg Hansen. Learning more skills through optimistic exploration. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=cU8rknuhxc.

Csaba Szepesvári. Constrained mdps and the reward hypothesis. *Musings about machine learning and other things (blog)*, 2020. URL https://readingsml.blogspot.com/2020/03/constrained-mdps-and-reward-hypothesis.html.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkfrvsA9FX.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999. URL https://arxiv.org/abs/physics/0004057.

Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.

Ziyu Wang, Alexander Novikov, Konrad Zolna, Josh S Merel, Jost Tobias Springenberg, Scott E Reed, Bobak Shahriari, Noah Siegel, Caglar Gulcehre, Nicolas Heess, et al. Critic regularized regression. *Advances in Neural Information Processing Systems*, 33:7768–7778, 2020.

Haoran Xu, Xianyuan Zhan, Jianxiong Li, and Honglei Yin. Offline reinforcement learning with soft behavior regularization. *CoRR*, abs/2110.07395, 2021. URL https://arxiv.org/abs/2110.07395.

Haoran Xu, Li Jiang, Jianxiong Li, Zhuoran Yang, Zhaoran Wang, Wai Kin Victor Chan, and Xianyuan Zhan. Offline RL with no OOD actions: In-sample learning via implicit value regularization. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=ueYYgo2pSSU.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 25746–25759, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/d7e4cdde82a894b8f633e6d61a01ef15-Abstract.html.

Tom Zahavy, Yannick Schroecker, Feryal M. P. Behbahani, Kate Baumli, Sebastian Flennerhag, Shaobo Hou, and Satinder Singh. Discovering policies with domino: Diversity optimization maintaining near optimality. *CoRR*, abs/2205.13521, 2022. doi: 10.48550/arXiv.2205.13521. URL https://doi.org/10.48550/arXiv.2205.13521.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. Gradientdice: Rethinking generalized offline estimation of stationary values. In *International Conference on Machine Learning*, pp. 11194–11203. PMLR, 2020.

Supplementary for Diverse Offline Imitation Learning

A Reproducibility

For implementation of DOI we have used the PyTorch autograd framework. For the SOLO12 training we made use of Isaac Gym for data collection and evaluation of the learned skill policies. For the D4RL experiments we evaluated the policies using the Mujoco v2.1 rigid body simulator. The training of the skill policies with evaluation and pre-training of the SMODICE expert ratios takes about 4 hours on an NVIDIA GeForce RTX 4080 graphics card with a batch size of 512. We plan on opensourcing the code and the SOLO12 data post conference acceptance. The SOLO12 robot has been developed as part of the Open Dynamic Robot Initiative (Grimminger et al., 2020), and a full assembly kit is available at a cheap price in order to reproduce the real system experiments from Supp. J.

B Fenchel Conjugate

The Fenchel conjugate $f_\star$ of a function $f : \Omega \to \mathbb{R}$ is given by

$$f_\star(y) = \sup_{x \in \Omega} \langle x, y \rangle - f(x), \tag{S1}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product defined on a space Ω. For any proper, convex and lower semi-continuous function f the following duality statement holds $f_{\star\star} = f$, that is

$$f(x) = \sup_{y \in \Omega_\star} \langle x, y \rangle - f_\star(y), \tag{S2}$$

where $\Omega_\star$ denotes the domain of $f_\star$. For any probability distributions $p, q \in \Delta(S)$ with $p(s) > 0$ implying $q(s) > 0$, we define for convex continuous functions f the family of f-divergences

$$\mathrm{D}_f(p||q) = \mathbb{E}_q \left[f\left(\frac{p(x)}{q(x)} \right) \right]. \tag{S3}$$

The Fenchel conjugate of an f divergence $\mathrm{D}_f(p||q)$ at a function $y(s) = p(s)/q(s)$ is, under certain conditions[4], given by

$$\mathrm{D}_{\star,f}(y) = \mathbb{E}_{q(s)} \left[f_\star(y(s)) \right]. \tag{S4}$$

Furthermore, its maximizer satisfies

$$p^\star(s) = q(s) f'_\star(y(s)). \tag{S5}$$

In the important special case where $f(x) = x \log(x)$, we obtain the well-known Kullback-Leibler (KL) divergence

$$\mathrm{D}_{\mathrm{KL}}(p||q) = \sum_s p(s) \log \frac{p(s)}{q(s)}. \tag{S6}$$

The Fenchel conjugate $\mathrm{D}_{\star,\mathrm{KL}}$ of the KL-divergence at a function $y(s) = p(s)/q(s)$ has a closed-form (Boyd & Vandenberghe, 2004, Example 3.25)

$$\mathrm{D}_{\star,\mathrm{KL}}(y) = \log \mathbb{E}_{q(s)}[\exp y(s)], \tag{S7}$$

and its maximizer $p^\star$ satisfies

$$p^\star(s) = q(s) \mathrm{softmax}_q(y(s)), \quad \text{where} \quad \mathrm{softmax}_q(y(s)) = \frac{\exp y(s)}{\mathbb{E}_{q(s')}[\exp y(s')]} \tag{S8}$$

[4] f needs to satisfy certain regularity conditions (Dai et al., 2017)

C Lagrange Relaxation

The Lagrange relaxation is given by

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

By combining Lem. D.4 and the definition of $\eta_{\widetilde{E}}(s,a) = d_{\widetilde{E}}(s,a)/d_O(s,a)$, we have

$$\mathrm{D_{KL}} \left(d_z(S,A) || d_{\widetilde{E}}(S,A) \right) = \mathrm{D_{KL}} \left(d_z(S,A) || d_O(S,A) \right) - \mathbb{E}_{d_z(s,a)} \left[\log \eta_{\widetilde{E}}(s,a) \right]$$

and thus

$$\max_{d_z(s,a),q(z|s)} \min_{\lambda>0} \sum_z \lambda_z \left[\epsilon + \mathbb{E}_{d_z(s,a)} \left[R_z^\lambda(s,a) \right] - \mathrm{D_{KL}} \left(d_z(S,A) || d_O(S,A) \right) \right], \tag{S9}$$

where the reward is given by

$$R_z^\lambda(s,a) := \frac{\log\left(|Z| q(z|s)\right)}{\lambda_z |Z|} + \log \eta_{\widetilde{E}}(s,a).$$

D Algorithmic Phases

D.1 Value Function Training

With fixed skill-discriminator $q(z|s)$ and Lagrange multipliers $\lambda > 0$, the Problem S9 becomes:

$$\max_{\{d_z(s,a)\}_{z\in Z}} \sum_z \lambda_z \left\{ \mathbb{E}_{d_z(s,a)} \left[R_z^\lambda(s,a) \right] - \mathrm{D_{KL}} \left(d_z(s,a) || d_O(s,a) \right) \right\}$$

or equivalently for every skill z:

$$\begin{aligned} \max_{d_z(s,a)\geq 0} \quad & \mathbb{E}_{d_z(s,a)} \left[R_z^\lambda(s,a) \right] - \mathrm{D_{KL}} \left(d_z(S,A) || d_O(S,A) \right) \\ \text{s.t.} \quad & \textstyle\sum_a d_z(s,a) = (1-\gamma)\rho_0(s) + \gamma \mathcal{T} d(s) \quad \forall s. \end{aligned} \tag{S10}$$

We note that the preceding problem formulation involves state-action occupancy.

The strict feasibility in Assumption 4.1 implies strong duality, and thus Problem (S10) shares the same optimal value as the following dual minimization problem (for details see (Nachum & Dai, 2020, Section 6) and (Ma et al., 2022, Theorem 2)):

$$\begin{aligned} V^\star = \quad & \arg\min_{V(s)} (1-\gamma) \mathbb{E}_{s\sim\rho_0} \left[V(s) \right] \\ & + \log \mathbb{E}_{d_O(s,a)} \exp \left\{ R_z^\lambda(s,a) + \gamma \mathcal{T} V(s,a) - V(s) \right\}, \end{aligned} \tag{S11}$$

where

$$\mathcal{T} V(s,a) = \mathbb{E}_{\mathcal{P}(s'|s,a)} V(s').$$

Moreover, the optimal primal solution reads

$$\frac{d_z^\star(s,a)}{d_O(s,a)} = \mathrm{softmax}_{d_O(s,a)} \left(R_z^\lambda(s,a) + \gamma \mathcal{T} V_z^\star(s,a) - V_z^\star(s) \right). \tag{S12}$$

D.2 Skill Discriminator Training

With fixed skill-conditioned policy $\pi_z^\star$ and Lagrange multipliers $\lambda > 0$, the Problem S9 becomes

$$\max_{q(z|s)} \sum_z \left\{ \mathbb{E}_{d_z(s,a)} \left[R_z^\lambda(s,a) \right] - \mathrm{D_{KL}} \left(d_z(S,A) || d_O(S,A) \right) \right\}$$

and reduces to

$$\max_{q(z|s)} \mathbb{E}_{p(z)} \mathbb{E}_{d_z(s,a)} \log q(z|s).$$

Lemma D.1. *Given ratios $\eta_z(s,a)$, using weighted-importance sampling, we can train offline an optimal skill-discriminator $q(z|s)$. In particular, we optimize by gradient descent the following optimization problem*

$$\max_{q(z|s)} \mathbb{E}_{p(z)}\mathbb{E}_{d_O(s,a)}\left[\eta_z(s,a)\log q(z|s)\right].$$

Proof. The statement follows by combining Lem. D.2 and Lem. E.1. □

Lemma D.2 (Discriminator Gradient)**.** *It holds that*

$$\nabla_\phi \mathbb{E}_{p(s)}\left[\mathrm{D_{KL}}\left(p(Z|s)||q_\phi(Z|s)\right)\right] = -\mathbb{E}_{p(z)}\mathbb{E}_{p(s|z)}\left[\nabla_\phi \log q_\phi(z|s)\right].$$

Proof. Observe that

$$\begin{aligned}\nabla_\phi \mathrm{D_{KL}}\left(p(Z|s)||q(Z|s)\right) &= \nabla_\phi \mathbb{E}_{p(z|s)}\log\frac{p(z|s)}{q_\phi(z|s)}\\ &= -\mathbb{E}_{p(z|s)}\nabla_\phi \log q_\phi(z|s),\end{aligned}$$

where the second equality follows by

$$\nabla_\phi \log\frac{p(z|s)}{q_\phi(z|s)} = -\frac{q_\phi(z|s)}{p(z|s)}p(z|s)\frac{\nabla_\phi q_\phi(z|s)}{[q_\phi(z|s)]^2} = -\frac{\nabla_\phi q_\phi(z|s)}{q_\phi(z|s)} = -\nabla_\phi \log q_\phi(z|s).$$

□

D.3 KL-divergence Constraint Violation

Lemma D.3 (State-Action KL Estimator)**.** *Suppose we are given offline datasets $\mathcal{D}_O(S,A)\sim d_O$, $\mathcal{D}_E(S)\sim d_E$ and optimal ratios $\eta_z(s,a)=\frac{d_z(s,a)}{d_O(s,a)}$ and $\eta_{\widetilde{E}}(s,a)=\frac{d_{\widetilde{E}}(s,a)}{d_O(s,a)}$ for all $(s,a)\in\mathcal{D}_O$, where the state-action occupancy $d_{\widetilde{E}}$ is induced by a policy $\pi_{\widetilde{E}}$ agreeing on the state occupancy of an expert π_E, i.e.*

$$\pi_{\widetilde{E}} \in \arg\min_\pi \mathrm{D_{KL}}\left(d_\pi(S)||d_E(S)\right).$$

Then, we can compute offline *an estimator of* $\mathrm{D_{KL}}\left(d_z(S,A)||d_{\widetilde{E}}(S,A)\right)$ *which is given by*

$$\phi_z = \mathbb{E}_{d_O(s,a)}\left[\eta_z(s,a)\log\frac{\eta_z(s,a)}{\eta_{\widetilde{E}}(s,a)}\right].$$

Proof. By Lemma D.4 we have

$$\mathrm{D_{KL}}\left(d_z(S,A)||d_{\widetilde{E}}(S,A)\right) = \mathrm{D_{KL}}\left(d_z(S,A)||d_O(S,A)\right) - \mathbb{E}_{d_z(s,a)}\left[\log\frac{d_{\widetilde{E}}(s,a)}{d_O(s,a)}\right].$$

For the first term, we have

$$\begin{aligned}\mathrm{D_{KL}}\left(d_z(S,A)||d_O(S,A)\right) &= \mathbb{E}_{d_z(s,a)}\log\frac{d_z(s,a)}{d_O(s,a)}\\ &= \mathbb{E}_{d_O(s,a)}\left[\eta_z(s,a)\log\eta_z(s,a)\right].\end{aligned}$$

The second term reduces to

$$\mathbb{E}_{d_z(s,a)}\left[\log\frac{d_{\widetilde{E}}(s,a)}{d_O(s,a)}\right] = \mathbb{E}_{d_O(s,a)}\left[\eta_z(s,a)\log\eta_{\widetilde{E}}(s,a)\right].$$

□

Lemma D.4 (Structural)**.** *Suppose* $0 < \eta_z(s,a), \eta_{\widetilde{E}}(s,a) < \infty$ *for all* $(s,a) \in \mathcal{D}_O$*. Then, we have*

$$\mathrm{D_{KL}}\left(d_z(S,A)||d_{\widetilde{E}}(S,A)\right) = \mathrm{D_{KL}}\left(d_z(S,A)||d_O(S,A)\right) - \mathbb{E}_{d_z(s,a)}\left[\log \frac{d_{\widetilde{E}}(s,a)}{d_O(s,a)}\right].$$

Proof. By definition of KL-divergence, we have

$$\begin{aligned}\mathrm{D_{KL}}\left(d_z(S,A)||d_{\widetilde{E}}(S,A)\right) &= \mathbb{E}_{d_z(s,a)}\left[\log\left(\frac{d_z(s,a)}{d_O(s,a)} \cdot \frac{d_O(s,a)}{d_{\widetilde{E}}(s,a)}\right)\right] \\ &= \mathrm{D_{KL}}\left(d_z(S,A)||d_O(S,A)\right) - \mathbb{E}_{d_Z(s,a)}\left[\log \frac{d_{\widetilde{E}}(s,a)}{d_O(s,a)}\right].\end{aligned}$$

□

E Importance Sampling

Lemma E.1 (Importance Sampling)**.** *Given ratios* $\eta_z(s,a)$*, it holds for any function* $f(s,a)$ *that*

$$\mathbb{E}_{d_z^\star(s,a)}\left[f(s,a)\right] = \mathbb{E}_{d_O(s,a)}\left[\eta_z(s,a) f(s,a)\right].$$

In particular, for any function $g(s)$ *we have*

$$\mathbb{E}_{d_z^\star(s)}\left[g(s)\right] = \mathbb{E}_{d_O(s,a)}\left[\eta_z(s,a) g(s)\right].$$

Proof. The first conclusion follows by definition of $\eta_z(s,a) = d_z(s,a)/d_O(s,a)$, whereas the second uses

$$\mathbb{E}_{d_z^\star(s)}\left[g(s)\right] = \mathbb{E}_{d_z^\star(s,a)\pi_z^\star(a|s)}\left[g(s)\right] = \mathbb{E}_{d_z^\star(s,a)}\left[g(s)\right] = \mathbb{E}_{d_O(s,a)}\left[\eta_z(s,a) g(s)\right].$$

□

E.1 Empirical Estimators

Recall that the primal optimal solution satisfies

$$\eta_z(s,a) := \frac{d_z^\star(s,a)}{d_O(s,a)} = \mathrm{softmax}_{d_O(s,a)}\left(R_z^\lambda(s,a) + \gamma \mathcal{T} V_z^\star(s,a) - V_z^\star(s)\right),$$

where

$$\mathrm{softmax}_{p(x)}(g(x)) = \frac{\exp\{g(x)\}}{\mathbb{E}_{p(x')}[\exp\{g(x')\}]}. \tag{S13}$$

In the rest of this section, we denote the above TD-error term by

$$\delta_z(s,a) = R_z^\mu(s,a) + \gamma \mathcal{T} V_z^\star(s,a) - V_z^\star(s).$$

By assumption, the offline dataset $\mathcal{D}_O$ is sampled u.a.r. from a state-action occupancy distribution $d_O(s,a)$. Let $\{w_z(s,a)\}_{(s,a)\in\mathcal{D}_O}$ be a discrete probability distribution, computed by a softmax, over the offline dataset $\mathcal{D}_O$, namely

$$w_z(s,a) = \mathrm{softmax}_{\mathcal{D}_O}(\delta_z(s,a)) = \frac{\exp\{\delta_z(s,a)\}}{\sum_{(s',a')\in\mathcal{D}_O}\exp\{\delta_z(s',a')\}}.$$

We are now ready to state the main result of this section.

Lemma E.2 (KL-divergence Estimator)**.** *The following expression*

$$\sum_{(s,a)\in\mathcal{D}_O} w_z(s,a)[\log w_z(s,a) - \log w_{\widetilde{E}}(s,a)]$$

is an empirical estimator of the KL-divergence $\mathrm{D_{KL}}\left(d_z(S,A)||d_{\widetilde{E}}(S,A)\right)$.

Proof. We estimate the expectation $\mathbb{E}_{d_O(s,a)}\exp\{\delta(s,a)\}$ using an empirical estimate $\frac{1}{|\mathcal{D}_O|}\sum_{(s,a)\in\mathcal{D}_O}\exp\{\delta_z(s,a)\}$ over the offline-dataset $\mathcal{D}_O$. By definition of $\mathrm{softmax}_{d_O(s,a)}$, see eq. (S13), the following expression

$$\widetilde{\eta}_z(s,a) = \frac{\exp\{\delta_z(s,a)\}}{\frac{1}{|\mathcal{D}_O|}\sum_{(s',a')\in\mathcal{D}_O}\exp\{\delta_z(s',a')\}} = |\mathcal{D}_O| w_z(s,a)$$

is an empirical estimator of the importance weight $\eta_z(s,a)$. Similarly, $\widetilde{\eta}_{\widetilde{E}}(s,a) = |\mathcal{D}_O| w_{\widetilde{E}}(s,a)$ is an estimator of $\eta_{\widetilde{E}}(s,a)$. Then, the statement follows by combining Lemma Lem. D.3, the definition of importance ratios $\eta_z(s,a) = d_z(s,a)/d_O(s,a)$, $\eta_{\widetilde{E}}(s,a) = d_{\widetilde{E}}(s,a)/d_O(s,a)$ and

$$\begin{aligned}
D_{\mathrm{KL}}\left(d_z(S,A)||d_{\widetilde{E}}(S,A)\right) &= \mathbb{E}_{d_O(s,a)}\left[\eta_z(s,a)\log\frac{\eta_z(s,a)}{\eta_{\widetilde{E}}(s,a)}\right] \\
&\approx \frac{1}{|\mathcal{D}_O|}\sum_{(s,a)\in\mathcal{D}_O}\widetilde{\eta}_z(s,a)\log\frac{\widetilde{\eta}_z(s,a)}{\widetilde{\eta}_{\widetilde{E}}(s,a)} \\
&= \sum_{(s,a)\in\mathcal{D}_O} w_z(s,a)\log\left(\frac{w_z(s,a)}{w_{\widetilde{E}}(s,a)}\right).
\end{aligned}$$

□

Lemma E.3 (Off-Policy Expectation Estimator)**.** *For any function* $f(s,a)$ *the following expression*

$$\sum_{(s,a)\in\mathcal{D}_O} w_z(s,a) f(s,a)$$

is an empirical estimator of the expectation $\mathbb{E}_{d_z^\star(s,a)}[f(s,a)]$.

Proof. By combining Lem. E.1 and similar arguments as in the proof of Lem. E.2, we have

$$\begin{aligned}
\mathbb{E}_{d_z^\star(s,a)}[f(s,a)] &= \mathbb{E}_{d_O(s,a)}[\eta_z(s,a)f(s,a)] \\
&\approx \frac{1}{|\mathcal{D}_O|}\sum_{(s,a)\in\mathcal{D}_O}\widetilde{\eta}_z(s,a)f(s,a) \\
&= \sum_{(s,a)\in\mathcal{D}_O} w_z(s,a)f(s,a).
\end{aligned}$$

□

F Unconstrained Formulation

SMODICE (Ma et al., 2022) minimizes a KL-divergence between the policy state occupancy and the expert state occupancy, expressed as

$$\min_{d(S)} \mathrm{D}_{\mathrm{KL}}\left(d(S) || d_E(S)\right). \tag{S14}$$

A naive approach to extend the above problem formulation to the unsupervised skill discovery setting, is to consider an additional diversity term in the objective. In particular, adding a scaled mutual information term $\mathcal{I}(S;Z)$ and maximizing over a set of skill-conditioned state occupancies $\{d_z(S)\}_{z\in Z}$, namely

$$\max_{\{d_z(S)\}_{z\in Z}} \alpha\mathcal{I}(S;Z) - \sum_{z\in Z} \mathrm{D}_{\mathrm{KL}}\left(d_z(S) || d_E(S)\right). \tag{S15}$$

Here, the level of diversity is controlled by a hyperparameter α. However, α is arbitrary, and no constraint on closeness to the expert state occupancy is enforced. We proceed by using the variational lower bound in eq. (3) and assuming a categorical uniform distribution $p(z)$ over the set of latent skills Z, which consists of $|Z|$ distinct indicator vectors in $\mathbb{R}^{|Z|}$. This reduce the optimization problem to

$$\max_{d_z(s), q(z|s)} \sum_{z\in Z} \left\{ \alpha \mathbb{E}_{d_z(s)} \left[\frac{\log\left(q(z|s)|Z|\right)}{|Z|} \right] - \mathrm{D}_{\mathrm{KL}}\left(d_z(S) || d_E(S)\right) \right\}. \tag{S16}$$

Theorem F.1. *(Ma et al., 2022) Suppose Assum. 3.1 holds. Then, we have*

$$\mathrm{D}_{\mathrm{KL}}\left(d_z(S) \| d_E(S)\right) \leq \mathbb{E}_{d_z(s)} \left[\log \frac{d_O(s)}{d_E(s)} \right] + \mathrm{D}_{\mathrm{KL}}(d_z(S,A) \| d_O(S,A)).$$

By Thm. F.1 and linearity of the objective, Problem (S16) reduces to optimizing separately for each latent skill z the following optimization problem

$$\max_{d_z(s), q(z|s)} \mathbb{E}_{d_z(s)}\left[R_z^\alpha(s,a)\right] - \mathrm{D}_{\mathrm{KL}}(d_z(S,A) \| d_O(S,A)), \tag{S17}$$

where $R_z^\alpha(s,a)$ is defined as

$$R_z^\alpha(s,a) := \underbrace{\log \frac{d_E(s)}{d_O(s)}}_{\text{Expert Imitation}} + \underbrace{\alpha \frac{\log\left(q(z|s)|Z|\right)}{|Z|}}_{\text{Skill Diversity}}. \tag{S18}$$

The ratios $\frac{d_E(s)}{d_O(s)}$ can be computed by training a discriminator $c(s)$ tasked to distinguish between samples from $d_E(s)$ and $d_O(s)$. More specifically, since the optimal Bayesian discriminator satisfies $c^\star(s) = d_E(s)/(d_E(s) + d_O(s))$, in practice we can use an estimator $c(s)/(1-c(s)) \approx \frac{d_E(s)}{d_O(s)}$.

Similar to the DOI, we can apply the alternating optimization scheme, here with two phases:(i) fixed skill-discriminator (similarly to Subsec. 4.2.1); and (ii) fixed importance ratios and policy $\pi_z^\star$, where we train the skill-discriminator $q(z|s)$ (see Supp. D.2). For the first phase, we use the importance ratios $\eta_z(s,a)$ computed by optimizing the dual-value problem and then applying softmax to the corresponding TD error terms (see eq. (12) and Nachum & Dai (2020); Ma et al. (2022)).

G Solo-12 Dataset Collection

As shown in Figure S1, both *expert dataset* and *offline dataset* are collected in parallelized GPU-based environments in Isaac Gym (Makoviychuk et al., 2021). The policies from both locomotion task and obstacle navigation tasks with SOLO12 are trained using the DOMiNiC (Cheng et al., 2024) algorithm to exhibit diverse behaviors while maintaining a certain level of task completion. For details on the algorithm used to train the data collection policies, we refer interested readers to (Cheng et al., 2024).

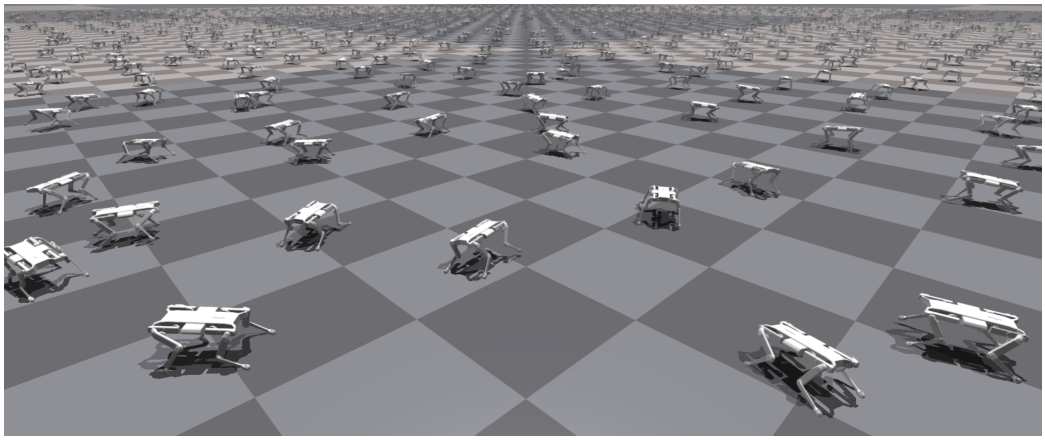


Figure S1: Solo-12 datasets are collected with 4000 environments in parallel using IsaacGym.

Locomotion task. The collecting policies are trained to track randomly sampled velocity commands on the flat ground. The state space consists of the linear and angular base velocity vectors, projected gravity vector, joint position, and velocity and commanded velocity. The actions contain the joint target angles, which will be taken by a PD controller to generate applied torque for each motor. During collecting, the policies are fed with a fixed forward velocity command of 1 m/s, and zeros for side velocity and yaw rate. As mentioned in Sec. 6, the policy used for collecting the *expert dataset* is the last and best checkpoint (iteration 2000) and trained without diversity objective, which exhibits a stable mid-height trotting gait pattern. The policies for collecting the *offline dataset* are different stochastic checkpoints throughout the training of the skill-conditioned policy. The intrinsic reward is designed to maximize the ℓ_2 distance of the successor features (Barreto et al., 2016) between distinct skills, where in this setting the feature space includes: the base height velocity, base roll and pitch velocities, and feet height velocities. The *offline dataset* is composed of 1/2 data from checkpoint 0, 1/4 data from checkpoint 50, 1/8 data from checkpoint 100, 1/16 data from checkpoint 500, 1/32 data from checkpoint 1500 and 1/32 data from checkpoint 2000. For each policy checkpoint, we collect data from the 5 corresponding skills, including the target skill. It is worth noting that more than half of the data from the *offline dataset* comes from the nearly random policies from the start of the training (checkpoints 0 and 50). Both datasets contain 4000 trajectories with an episode length of 250 steps, or 1 million transitions each.

Obstacle navigation task. The policies are trained to track the target position in a terrain of random obstacles of various heights of $\{0.0, 0.05, \ldots, 0.25\}$ meters within a fixed time horizon. The state space of the agent contains the linear and angular base velocity vectors, projected gravity vector, joint position and velocity, a surrounding height map of the robot and time information, while the actions remain the same as the locomotion task. During data collection, the policies are tasked with tracking the target 3.0 meter away in the front direction while confronting a 1.0×1.0 meter square obstacle of 0.2 meter height. The intrinsic reward for training the policy is designed to diversify the base velocity direction such that distinct skills exhibit diverse strategies. For the *expert dataset*, the used policy is the last and best checkpoint (iteration 2000) trained with diversity objective. The *expert dataset* is multi-modal in nature, as the dataset contains diverse strategies for navigating in front of the obstacle, either avoiding it from both sides or climbing it. On the other hand, the policies for collecting the *offline dataset* are the skill-conditioned checkpoints from iterations $\{0, 50, 100, 150, 200, 250, 500, 1000, 1500, 2000\}$. Both datasets contain 2000 trajectories with an episode length of 500 steps, or 1 million transitions each.

Sim-to-Real transfer. In addition, we use domain randomization during training and data collection, in order to tackle the sim-to-real transfer and to simulate more diverse environment interaction. Specifically, we randomize the friction coefficient between $[0.5, 1.5]$, additional base mass between $[-0.5, 0.5]$ kg, and simulate the observation noise and an actuator lag of 15 ms.

H SMODICE Expert Return

In table S1 we show the performance of the evaluated policies trained by SMODICE(Ma et al., 2022) on the WALKER2D and HALFCHEETAH. The results are consistent with the performance that we obtain with DOI in Figure 7. We also note here the importance of having expert state coverage in the offline data that is reflected in the performance of the policies.

Environment	dataset	N	r
halfcheetah	medium-expert	25	81.25
		50	80.47
		200	73.56
	medium-replay	25	29.28
		50	36.73
		200	60.67
	random	25	10.89
		50	27.71
		200	78.94
walker2d	medium-expert	25	3.98
		50	19.22
		200	4.10
	medium-replay	25	15.09
		50	3.60
		200	0.95
	random	25	52.62
		50	103.52
		200	108.20

Table S1: Expected return for SMODICE-learned expert policies in the WALKER2D and ANT environments for N expert trajectories mixed-in.

I Lagrange Multiplier Stability

In Figure S2 we observe the behavior of the Lagrange multipliers for different levels of ϵ for a specific skill z in the SOLO12 experiment. In case of $\epsilon \in \{1.0, 2.0\}$, the multipliers fluctuate around a specific level that strikes the balance between diversity and expert imitation. This can also be validated when observing the violation level in Figure S2b of the constraint given estimator ϕ_z, which is for $\epsilon \in \{1.0, 2.0\}$ around 0. On the other hand, if we introduce a strong constraint on the KL-divergence ($\epsilon = 0.0$), which is constantly violated, hence $\sigma(\mu_z) = 1$. Similarly, if the constraint is too weak, only diversity is optimized, in which case there is a significant degradation in performance (see figure Figure 4).

In Figure S3 we show the bounded lagrange multiplier values for three skills and the resulting violations for different ϵ levels for the ANT experiment. Again, the multiplier values fluctuate around appropriate levels ensuring the the violation of the constraint remains close to 0.

J Real Robot Deployment

For the locomotion task, we successfully deployed policies exhibiting diverse skills extracted from the *offline dataset* while being able to track a certain velocity similar to the expert on real hardware. Our skill-conditioned policy exhibits different walking behaviors with diverse base motions. Snapshots of these diverse behaviors can be seen in Figure S4.

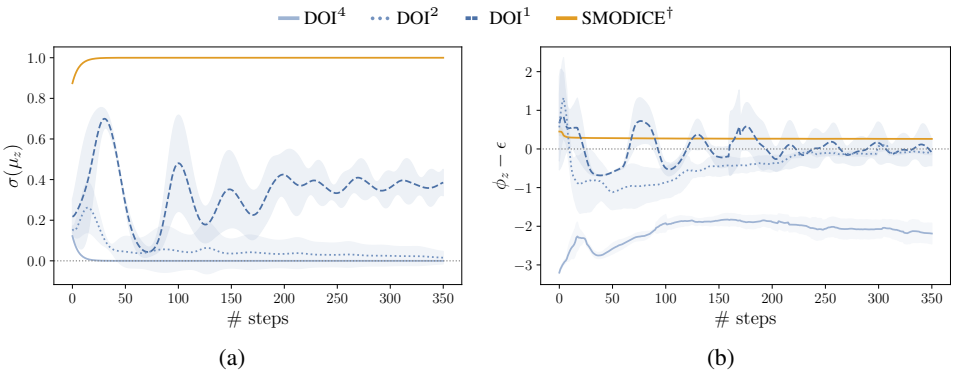


Figure S2: Behavior of Lagrange multipliers. (a) Evolution of $\sigma(\lambda_z)$ for one skill ($z = 1$ chosen arbitrarily), (b) violation of the constraint for different ϵ. Negative $\phi_z - \epsilon$ indicates no violation. Means and standard deviation across restarts. The shaded areas show the interval between the 0.25 and 0.75 quantiles, computed over 3 seeds.

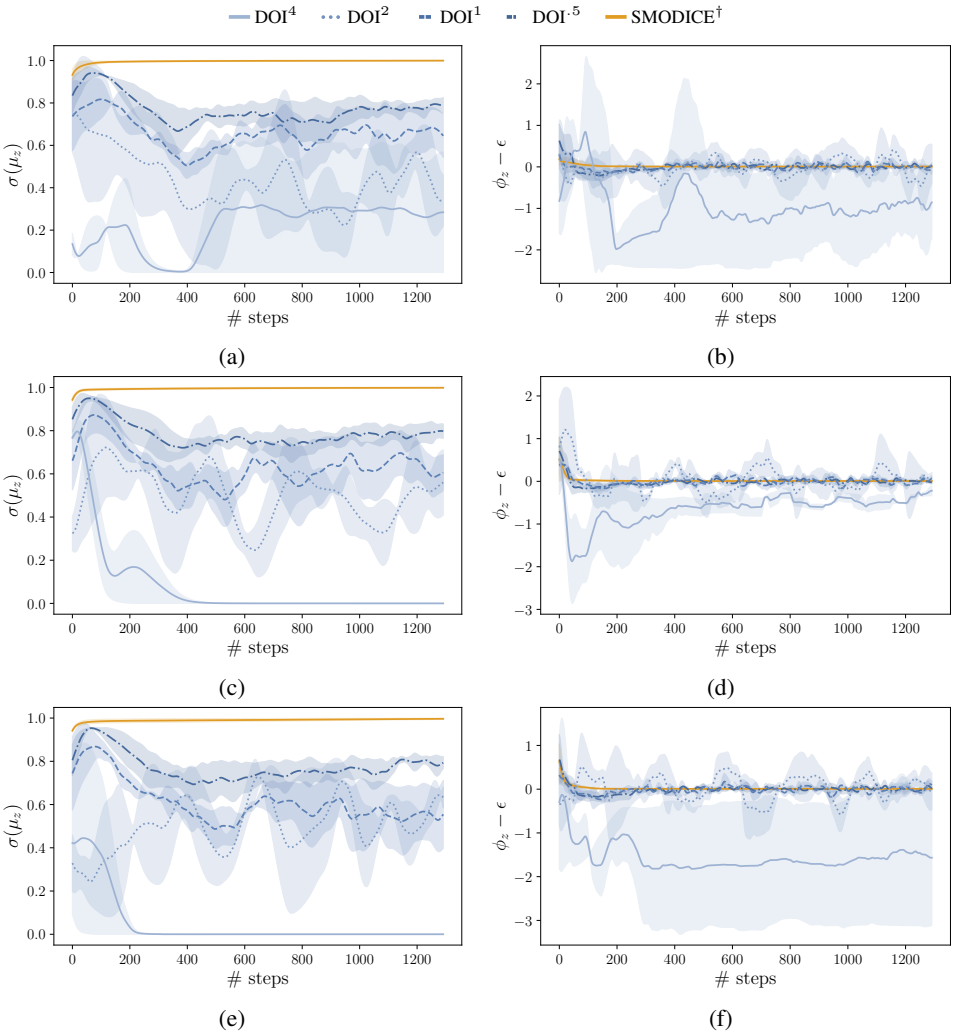
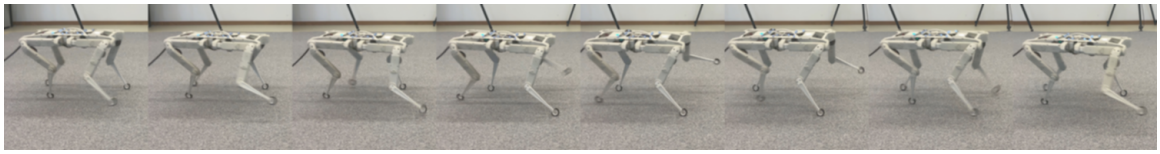


Figure S3: Behavior of Lagrange multipliers. (a) Evolution of $\sigma(\lambda_z)$ for one skill ($z = 1$ chosen arbitrarily), (b) violation of the constraint for different ϵ. Negative $\phi_z - \epsilon$ indicates no violation. Means and standard deviation across restarts. The shaded areas show the interval between the 0.25 and 0.75 quantiles, computed over 3 seeds.

(a) Trot locomotion with wave trunk motion and low base height.

(b) Trot locomotion with middle base height.



(c) Trot locomotion with high base height.

Figure S4: Snapshots of the trained policy exhibiting distinct skills on hardware. From above to bottom, the policy has low, middle and high base positions while moving forward.

K Observation Projection

Imitation learning is of particular interest when the agent's and the target expert policy's state spaces do not necessarily match, but overlap in certain parts, as is often the case when learning from demonstrations. Our framework naturally accounts for this. If we consider $\mathcal{S}'$ to be the state space of the expert and $\mathcal{S}$ the state space of the agent, we assume that there exists a simple projection mapping $\Pi : \mathcal{S}' \mapsto \mathcal{O}$, where $\mathcal{O} := \{o : o \subset s, s \in \mathcal{S}\}$ is the power set of observations, allowing us to potentially imitate beyond expert policies with the same state space as the agent. Note that the agent still observes its full state s, however the projected state $\Pi(s)$ is observed by the expert classifier and skill discriminator. The projection Π can be selected to specify which parts of the state we want to diversify and constrain in terms of occupancy, depending on the task at hand.

L Limitations

The DOI method also comes with certain caveats. Maximizing the mutual information, as a diversity objective, poses a hard optimization problem due to its convexity. Thus, designing alternative diversity objectives can be beneficial. Furthermore, closeness in state-action occupancy can be quite restrictive in terms of availability of diverse behaviors that satisfy the constraint. Replacing this with constraints on the return of the policy would allow more freedom to optimize diversity in cases where the optimal policy may be multimodal. The above challenges are promising directions for future work.

M Robust Obstacle Navigation

When the expert data is multi-modal, some modes might be more robust to distribution shift than others. However, using a uni-modal algorithm such as SMODICE, which tries to match the expert's state occupancy distribution, may not result in a robust policy. In contrast, each learned DOI skill recovers a particular mode, and as shown in this experiment, at least one DOI skill is robust against a distribution shift.

We consider the task of navigating across a box obstacle to a target position behind it, for the SOLO12 robot. For training the DOI skills, we choose the feature vector $\phi(s)$ with linear and angular velocity as input to the skill-discriminator $q(z|\phi(s))$. The agent used to collect the *expert dataset* can go over or around the box obstacle from the left or right side to reach the target position in the traversable obstacle terrain. The box has a height of 0.2 meters and a square size of 1.0×1.0 meters. As a result, the collected expert data is multi-modal and consists of trajectories over and from the sides of the box obstacle to the target position.

It is important to emphasize that the less direct route to the target position (left or right side of the box) is always the more robust choice, since the agent runs into the risk of slipping or falling while climbing the box. We evaluate the learned DOI skills and SMODICE expert on 6 different heights: $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$ meters. The $\{0.3, 0.4, 0.5, 0.6\}$ meters boxes are out-of-distribution and increasingly difficult to traverse from above the box. In Figure S5, we observe the trajectory distributions of the DOI skills and SMODICE expert collected in simulation. The arrows indicate the yaw angle of the robot at the trajectory points.

As we can see from the return distributions in Figure S7, the performance of the SMODICE expert is strongly affected by the height of the box, as it is biased towards climbing over the box (this also depends on the initial state of the agent), which becomes increasingly difficult and may not be feasible. This can be observed from the trajectory distribution shown in the right-most column of Figure S5; the trajectories of the SMODICE expert become increasingly concentrated in front of the box as its height increases. On the other hand, the three DOI skills (learned with a fixed Lagrange multiplier $\sigma(\mu) = 0.5$) recover diverse behaviors and robustly reach the goal. Here it is DOI-Skill 3, which is the most robust in reaching the target position and gives the highest return (see Figure S5 and Figure S7).

In Figure S5, each row corresponds to a box with a fixed height $H \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$ meters. Each of the first three columns is associated with a fixed DOI skill (red, green, and blue) and the last column represents the SMODICE expert. Each cell shows every 10th step of 60 randomly initialized trajectories, all computed in simulation. This experiment demonstrates that although SMODICE expert is multimodal, it gets stuck in front of the box and fails to robustly reach the target position already at a box height of 0.4 meters. In contrast, the DOI-Skill 3 robustly reaches the target position by bypassing the box from the left side. The fraction of randomly initialized trajectories stuck in front of the box is significantly smaller for the DOI-Skill 3 than the SMODICE expert. This is reflected in the return distribution shown in Figure S7, which has the same row and column structure as Figure S5.

N Additional Experiments

Instead of learning the Lagrange multipliers λ_z via KL estimators ϕ_z, we can also fix λ_z at a certain level, making it a hyperparameter. In our setting, this also works well, and we demonstrate a tradeoff between diversity and task reward optimization, see Figures S8 and S9. However, in this case we lose the possibility to enforce a certain constraint on the KL-divergence between the skill state-action occupancy and expert state-action occupancy.

We further provide results of applying DOI to different levels of expert trajectory mix-in to the *medium-replay* and *random* datasets of WALKER2D and HALFCHEETAH in tables S2 and S3.

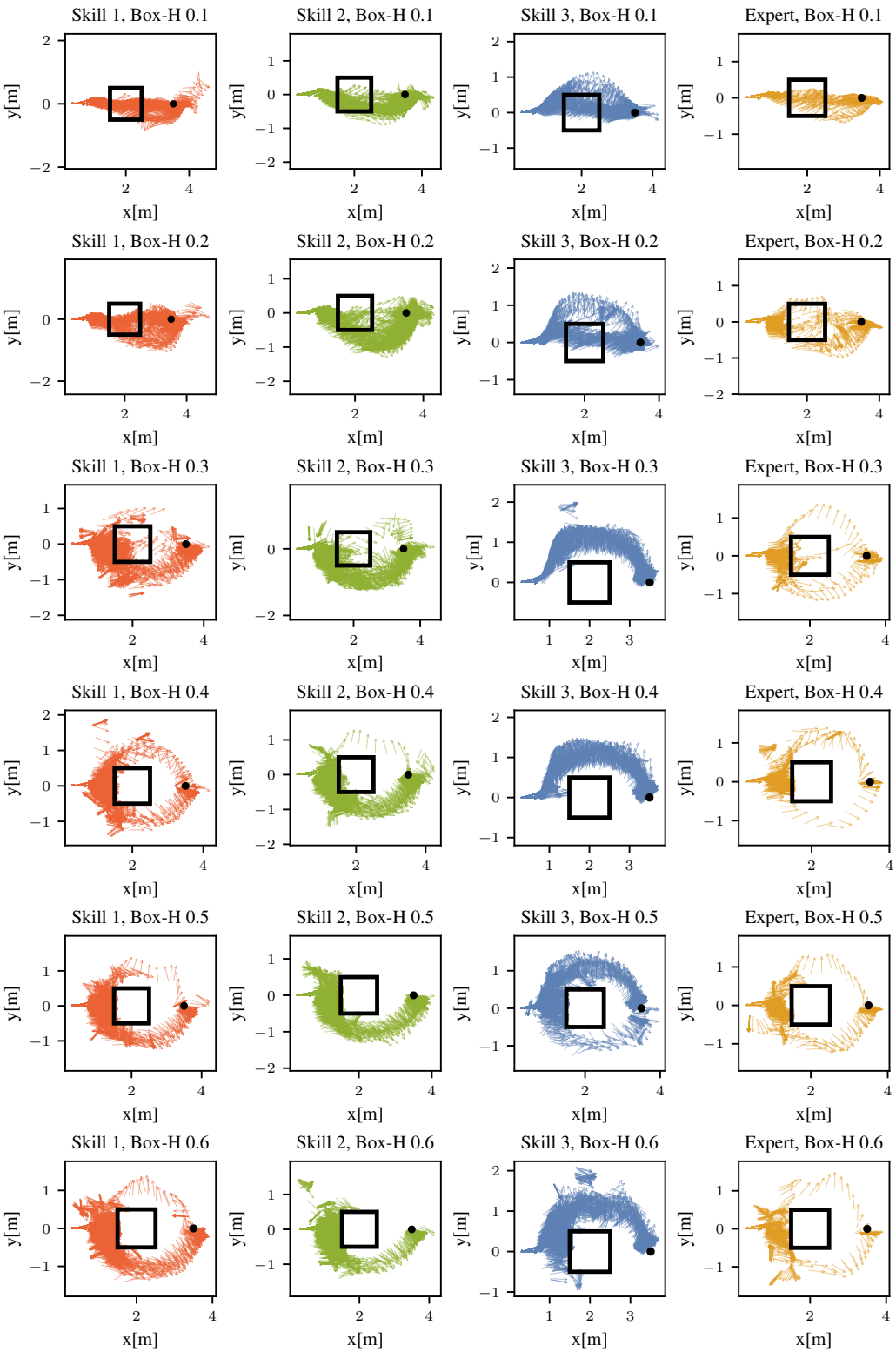


Figure S5: A performance benchmark of the DOI skills and the SMODICE expert on an obstacle navigation task, where the SOLO12 is initialized in front of a box and tries to reach a target position behind the box. The task consists of six levels of increasing difficulty depending on the height of the box.

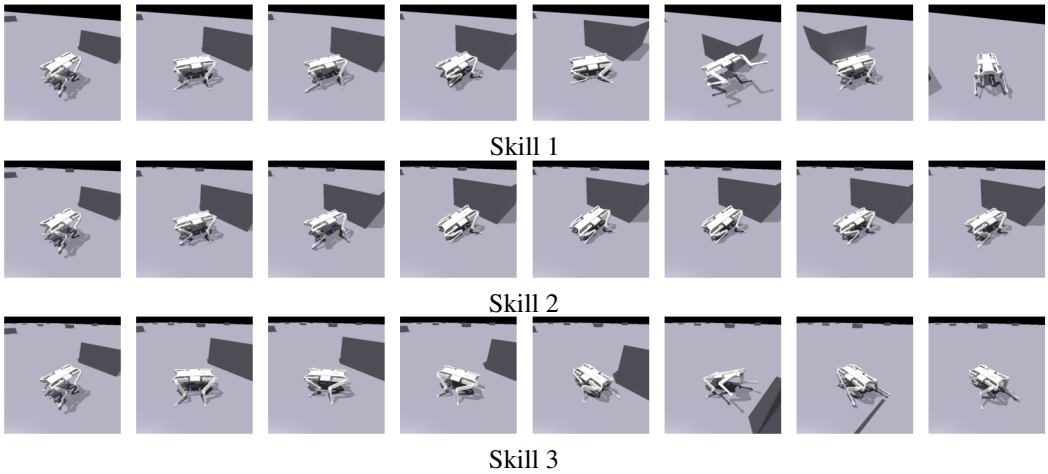


Figure S6: Frames from rollout videos of the learned DOI skills for the highest box task, skills 1 and 3 go from the side of the boxes to the goal, and skill 2 reimains in front of the box since it mostly tries to climb it.

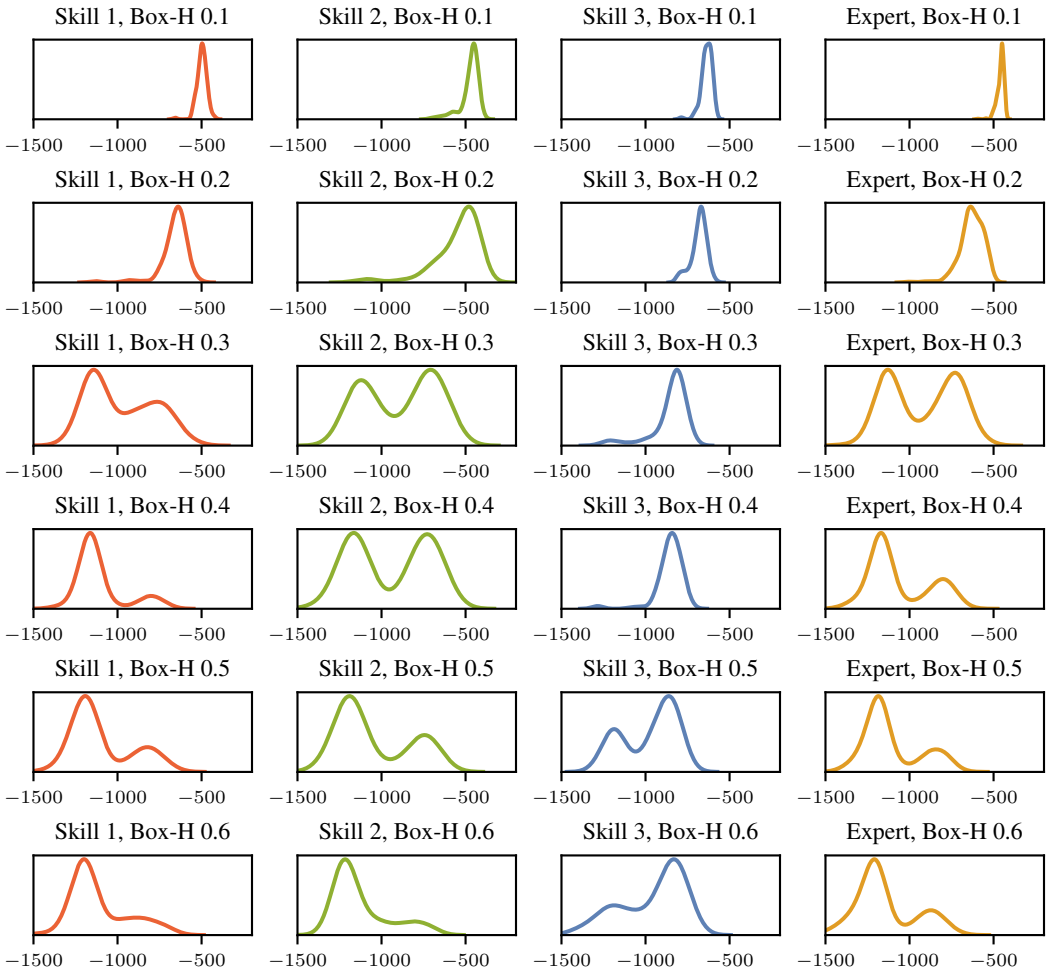
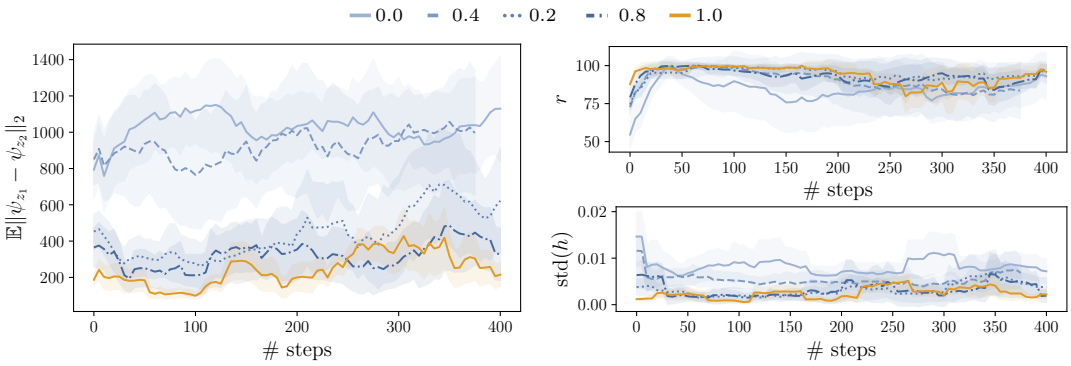
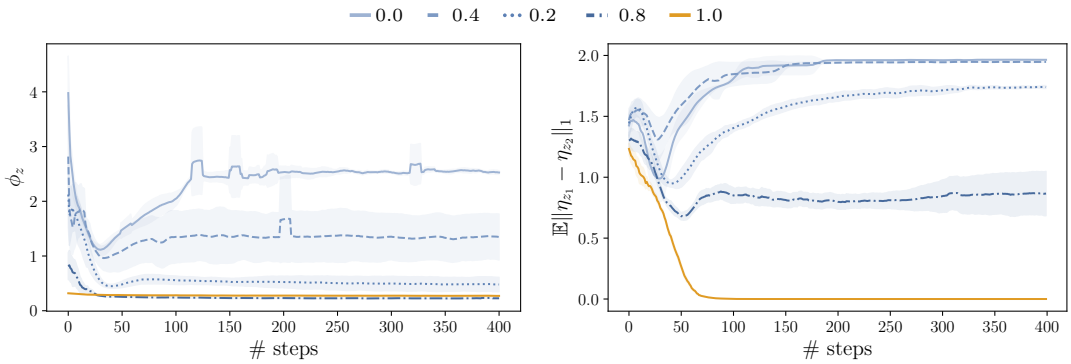


Figure S7: Return distributions for DOI skills and SMODICE, we see in particular that the SMODICE policy return distribution is greatly affected by increasing the height of the box.



(a) (b)

Figure S8: (a) Average ℓ_2 distance between Monte Carlo estimated successor representations ψ_z of distinct skills, (b) return r as % of expert return and standard deviation of base height $\mathrm{std}_z(h)$, depending on a fixed $\sigma(\lambda_z)$ (see legend). The shaded areas show the interval between the 0.25 and 0.75 quantiles, computed over 3 seeds.



(a) (b)

Figure S9: Divergence estimate and η_z distance for the case of fixed $\sigma(\lambda_z)$. (a) Value of divergence estimator ϕ_z for a specific skill over the course of training ($z = 1$ chosen arbitrarily), (b) average ℓ_1 distance of η_z's of skills. Means and standard deviation across restarts. The shaded areas show the interval between the 0.25 and 0.75 quantiles, computed over 3 seeds.

dataset	# expert mixin	ϵ	$\mathbb{E}\|\eta_{z_1} - \eta_{z_2}\|$	r	$\mathbb{E}\|\psi_{z_1} - \psi_{z_2}\|$
medium-replay	25	0.0	0.00 ± 0.00	46.00 ± 1.46	6.16 ± 0.30
		0.5	0.21 ± 0.08	0.33 ± 0.48	3.54 ± 2.14
		1.0	1.40 ± 0.05	2.33 ± 0.51	6.09 ± 2.40
		2.0	1.30 ± 0.03	0.64 ± 0.11	7.67 ± 4.27
		4.0	1.54 ± 0.08	2.30 ± 1.64	19.26 ± 2.29
	50	0.0	0.00 ± 0.00	54.29 ± 2.13	5.53 ± 0.14
		0.5	0.82 ± 0.28	31.31 ± 7.03	14.13 ± 1.86
		1.0	1.21 ± 0.15	4.33 ± 0.75	0.42 ± 0.05
		2.0	1.37 ± 0.03	1.61 ± 0.41	13.85 ± 2.50
		4.0	1.48 ± 0.12	1.11 ± 0.36	22.02 ± 1.33
	200	0.0	0.00 ± 0.00	98.33 ± 0.44	2.67 ± 0.26
		0.5	0.45 ± 0.11	74.59 ± 8.96	6.22 ± 1.17
		1.0	1.20 ± 0.09	2.52 ± 1.50	12.97 ± 4.33
		2.0	1.30 ± 0.03	2.07 ± 0.65	3.23 ± 2.02
		4.0	1.59 ± 0.06	1.43 ± 0.64	19.48 ± 1.43
random	25	0.0	0.00 ± 0.00	36.49 ± 11.54	15.70 ± 0.48
		0.5	0.93 ± 0.02	20.48 ± 7.90	16.81 ± 3.14
		1.0	1.30 ± 0.12	3.72 ± 1.38	8.16 ± 5.43
		2.0	1.45 ± 0.09	1.22 ± 0.32	20.47 ± 3.08
		4.0	1.27 ± 0.05	0.60 ± 0.26	20.60 ± 4.17
	50	0.0	0.00 ± 0.00	103.16 ± 0.69	3.32 ± 0.07
		0.5	1.03 ± 0.13	33.60 ± 6.64	18.27 ± 2.50
		1.0	1.37 ± 0.09	5.05 ± 2.66	20.16 ± 3.05
		2.0	1.46 ± 0.06	0.77 ± 0.29	10.46 ± 3.77
		4.0	1.23 ± 0.09	0.26 ± 0.11	14.33 ± 1.97
	200	0.0	0.00 ± 0.00	107.43 ± 0.26	1.84 ± 0.08
		0.5	1.29 ± 0.07	103.29 ± 1.38	6.75 ± 0.77
		1.0	1.26 ± 0.22	2.43 ± 0.30	7.30 ± 4.86
		2.0	1.46 ± 0.10	0.47 ± 0.15	15.39 ± 1.56
		4.0	1.29 ± 0.01	1.91 ± 0.57	19.66 ± 3.36

Table S2: WALKER2D metrics across random and medium-replay variants with varying number of mixed-in trajectories of the expert to satisfy the coverage assumption.

dataset	# expert mixin	ϵ	$\mathbb{E}\|\eta_{z_1} - \eta_{z_2}\|$	r	$\mathbb{E}\|\psi_{z_1} - \psi_{z_2}\|$
medium-replay	25	0.0	0.00 ± 0.00	37.64 ± 0.30	3.22 ± 0.06
		0.5	0.83 ± 0.12	36.95 ± 0.63	3.02 ± 0.10
		1.0	1.36 ± 0.09	24.30 ± 6.28	13.34 ± 4.84
		2.0	1.44 ± 0.06	6.73 ± 3.65	22.09 ± 8.15
		4.0	1.27 ± 0.09	2.68 ± 0.72	21.68 ± 1.87
	50	0.0	0.01 ± 0.01	45.40 ± 0.22	3.26 ± 0.27
		0.5	1.14 ± 0.02	42.89 ± 0.19	2.94 ± 0.12
		1.0	1.41 ± 0.12	37.28 ± 2.41	6.18 ± 1.21
		2.0	1.32 ± 0.11	8.60 ± 4.66	13.66 ± 1.97
		4.0	1.24 ± 0.16	1.72 ± 0.18	28.74 ± 7.84
	200	0.0	0.00 ± 0.00	73.60 ± 0.39	3.65 ± 0.09
		0.5	1.16 ± 0.08	69.91 ± 1.14	3.67 ± 0.10
		1.0	1.28 ± 0.13	23.74 ± 12.94	13.47 ± 1.73
		2.0	1.49 ± 0.10	15.52 ± 4.29	32.03 ± 0.56
		4.0	1.42 ± 0.07	2.16 ± 0.04	11.92 ± 2.28
random	25	0.0	0.00 ± 0.00	2.80 ± 0.36	5.55 ± 1.18
		0.5	1.12 ± 0.04	3.03 ± 0.28	4.30 ± 0.85
		1.0	1.14 ± 0.12	2.24 ± 0.09	10.45 ± 3.30
		2.0	1.24 ± 0.08	1.73 ± 0.33	25.01 ± 8.78
		4.0	1.44 ± 0.03	1.60 ± 0.30	35.08 ± 8.27
	50	0.0	0.00 ± 0.00	31.89 ± 1.14	9.97 ± 0.58
		0.5	1.14 ± 0.11	10.29 ± 3.13	17.90 ± 6.01
		1.0	1.42 ± 0.15	6.45 ± 2.95	23.30 ± 0.96
		2.0	1.41 ± 0.08	2.73 ± 0.43	23.91 ± 6.98
		4.0	1.68 ± 0.06	1.44 ± 0.27	35.07 ± 8.08
	200	0.0	0.00 ± 0.00	68.35 ± 1.25	5.20 ± 0.31
		0.5	1.30 ± 0.08	50.85 ± 17.30	9.80 ± 3.68
		1.0	1.21 ± 0.12	15.06 ± 5.58	29.57 ± 4.26
		2.0	1.03 ± 0.10	2.10 ± 1.99	10.84 ± 7.57
		4.0	1.20 ± 0.20	2.16 ± 0.05	16.90 ± 5.95

Table S3: HALFCHEETAH metrics across random and medium-replay variants with varying number of mixed-in trajectories of the expert to satisfy the coverage assumption.