Diverse Offline Imitation via Fenchel Duality

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

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]University of Tübingen, Tübingen, Germany

Abstract

There has been significant recent progress in the area of unsupervised skill discovery, with various works proposing mutual information based objectives, as a source of intrinsic motivation. Prior works predominantly focused on designing algorithms that require online access to the environment. In contrast, we develop an *offline* skill discovery algorithm. Our problem formulation considers the maximization of a mutual information objective constrained by a KL-divergence. More precisely, the constraints ensure that the state occupancy of each skill remains close to the state occupancy of an expert, within the support of an offline dataset with good state-action coverage. Our main contribution is to connect Fenchel duality, reinforcement learning and unsupervised skill discovery, and to give a simple offline algorithm for learning diverse skills that are aligned with an expert.[1]

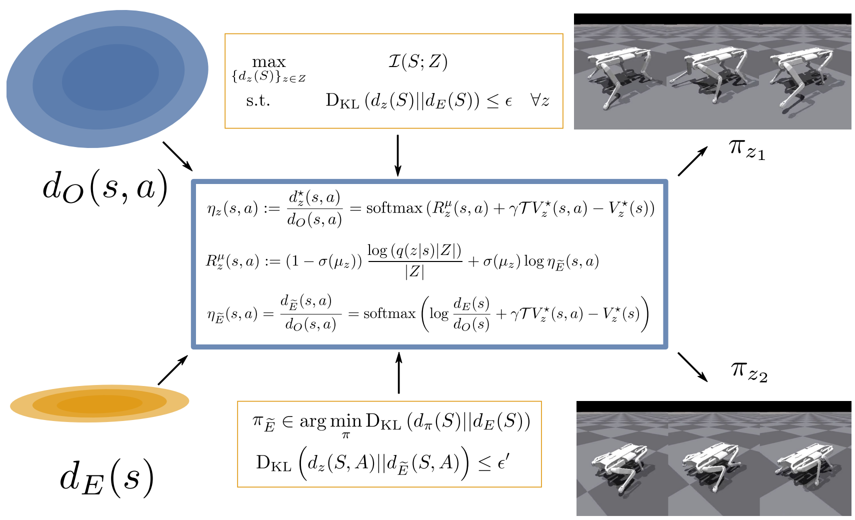


Figure 1: Diverse Offline Imitation (DOI) maximizes a variational lower bound on the mutual information between skills z and states s subject to a KL-divergence constraint to limit the deviation of the state occupancy $d_z(s)$ of each skill z from that of an expert $d_E(s)$. This requires offline datasets $\mathcal{D}_E$ sampled from $d_E(s)$ and $\mathcal{D}_O$ sampled from state-action occupancy $d_O(s,a)$ of various policies, to compute specific importance weights $\eta_z(s,a)$ of learned skill and $\eta_{\widetilde{E}}(s,a)$ expert.

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

16th European Workshop on Reinforcement Learning (EWRL 2023).

1 Introduction

Unsupervised skill discovery has received considerable attention in sequential decision-making. Meaningful skill extraction has shown to greatly increase learning efficiency in downstream tasks [Precup, 2000, Sharma et al., 2020b] and deal with miss-specified reward functions, or imperfect expert demonstrations [Ma et al., 2022]. On a more high level, intelligent behavior that improves upon non-native demonstrations emerges from diversity in autonomous agents.

Consequently, many works exist dealing with unsupervised skill discovery through various mechanisms, mostly in the online setting. A subset of these approaches attempt to maximize the mutual information between a skill-conditioning variable z and the skill-conditioned trajectory [Eysenbach et al., 2019, Gregor et al., 2017, Sharma et al., 2020a]. Multiple works have attempted to extract skills through the means of successor-feature formalism [Barreto et al., 2016, Dayan, 1993], which captures versatile behaviors by optimizing a family of reward functions parameterized by a skill variable z and a state representation [Barreto et al., 2016, Hansen et al., 2020, Barreto et al., 2018].

In the age of data abundance, focus is turning towards offline learning, which leverages collected experience from various sources. Offline reinforcement learning algorithms allow efficiency in terms of re-use of data and not requiring online samples, this makes them inherently more scalable and sample-efficient than the online counterpart. However, offline algorithms tend to suffer from the off-policy evaluation problem, which has been well studied in previous work [Levine et al., 2020, Prudencio et al., 2022]. In most approaches, the off-policy evaluation is tackled by remaining close to the data distribution [Wang et al., 2020, Kumar et al., 2020].

As we shall see, the Lagrange dual formulation of an optimization problem with mutual information objective and KL-divergence constraints can be reduced to solving a sequence of problems each of which has an inner maximization problem that admits a closed form solution, leveraging the Fenchel conjugate. In line with existing approaches [Nachum et al., 2019a,b, Kim et al., 2022, Zhang et al., 2020, Dai et al., 2020], the Fenchel duality allows us to compute state occupancy importance weights and in turn to train off-policy a skill discriminator, a skill-conditioned policy and an estimator of the KL-divergence constraint.

In this work, we consider the setting in which expert demonstrations without rewards are available, and the goal is to extract versatile near-optimal skill-conditioned policies by leveraging diverse offline data without access to expert actions. This setting is of particular interest in cases where expert data is expensive to obtain, an argument that makes offline learning particularly appealing [Fu et al., 2020]. However, as pointed out by others [Ma et al., 2022, Li et al., 2023], demonstration data oftentimes does not match the policy action space or we do not have access to the actions performed, which is mostly the case when we have demonstrations from an agent with a fundamentally different action space, such as a human.

Building upon the duality principles in optimization and reinforcement learning [Nachum and Dai, 2020], we design an *offline* algorithm which maximizes a *mutual information* objective subject to state occupancy KL-divergence constraints. More precisely, we exploit Fenchel duality to arrive at a principled importance-weighted offline training procedure for diverse skill discovery, while maintaining closeness in state occupancy to an expert. To the best of our knowledge, this is the first algorithm for unsupervised skill discovery that maximizes mutual information in the *offline* setting.

2 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. For simplicity, we consider infinite horizon (non-terminating) environments, which can be extended to finite horizon environments by considering an additional terminal state that loops in on itself continuously with zero reward.

Given a policy π, the associated state-action occupancy measure reads

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

and its 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, the set of skills is defined by Z, which we will treat as a finite set, and the skill-conditioned policy is given by $\pi_z : S \times Z \mapsto \Delta(\mathcal{A})$ with corresponding state occupancy $d_z(s) := d^{\pi_z}(s)$, for each skill $z \in Z$.

Throughout this work, we consider an offline setting with an access to the following two datasets: i) $\mathcal{D}_E$ is sampled from an expert state occupancy $d_E(S)$; and ii) $\mathcal{D}_O$ is sampled from a coverage distribution $d_O(S,A)$, possibly generated by a mixture of different behaviors. In addition, our analysis makes use of the following coverage assumption on state occupancies[2].

Assumption 2.1 (Expert coverage)**.** *We assume that* $d_E(s) > 0$ *implies* $d_O(s) > 0$.

2.1 Fenchel Conjugate

The Fenchel conjugate $f_\star$ of a function $f : \Omega \to \mathbb{R}$ is given by $f_\star(y) = \max_{x\in\Omega}\langle x, y\rangle - f(x)$, where $\langle\cdot,\cdot\rangle$ denotes the inner product defined on a probability space Ω. For any proper, convex and lower semi-continuous function f the following duality statement holds $f_{\star\star} = f$, that is $f(x) = \max_{y\in\Omega_\star}\langle x, y\rangle - f_\star(y)$, where $\Omega_\star$ denotes the domain of $f_\star$.

For any probability distributions $p, q \in \Delta(S)$ with $p(s) > 0$ implies $q(s) > 0$, we define for convex continuous functions f the family of f-divergences $\mathrm{D}_f(p||q) = \mathbb{E}_q[f(p(x)/q(x))]$. The Fenchel conjugate of an f divergence $\mathrm{D}_f(p||q)$ at a function $y(s) = p(s)/q(s)$ is, under mind conditions[3], given by $\mathrm{D}_{\star,f}(y) = \mathbb{E}_{q(s)}\left[f_\star(y(s))\right]$. Furthermore, its maximizer satisfies $p^\star(s) = q(s)f'_\star(y(s))$.

In the important special case when $f(x) = x\log(x)$, we obtain the well-known KL-divergence $\mathrm{D}_{\mathrm{KL}}(p||q) = \sum_s p(s)\log(p(s)/q(s))$. Moreover, 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 and Vandenberghe, 2004, Example 3.25] $\mathrm{D}_{\star,\mathrm{KL}}(y) = \log\mathbb{E}_{q(s)}[\exp y(s)]$ and any maximizer $p^\star$ satisfies $p^\star(s) = q(s)\mathrm{softmax}(y(s))$.

3 Method

In this work, given an expert and a coverage dataset as above, we seek to solve *offline* the following constrained optimization problem, which optimizes over all skill-conditioned policies $\{\pi_z\}_{z\in Z}$, i.e.,

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

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

where $\mathcal{I}(S;Z)$ denotes the mutual information between states and skills.

We note that 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., 2020b]) and information gain (c.f. DISDAIN in [Strouse et al., 2022]); and ii) general f-divergence constraints in (2), see Nachum and Dai [2020], Ma et al. [2022]. We leave the study of these variants for future work.

Since estimating the mutual information $\mathcal{I}(S;Z)$ is generally intractable, in line with previous work [Eysenbach et al., 2019] we assume that the 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

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

Here with $q(z|s)$ we denote a discriminator tasked with distinguishing between 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 $\mathcal{D}_E \subset \text{States}[\mathcal{D}_O]$, trains a policy $\pi_{\widetilde{E}}$ which

[2]Similarly to Kim et al. [2022] and Ma et al. [2022], in practice we ensure Assumption 2.1 by constructing i) the coverage dataset $\mathcal{D}_O$ to be the union of a mixture dataset $\mathcal{D}_M$ (generated by diverse policies of various expertise) and an expert dataset $\mathcal{D}_{E'}$, where both contain states and actions; and ii) the expert dataset $\mathcal{D}_E$, containing only states, is $\mathcal{D}_{E'}$ with filtered actions.

[3]f needs to satisfy certain regularity conditions [Dai et al., 2017]

optimizes the problem

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

and in addition outputs ratios $\eta_{\widetilde{E}}(s,a) = d_{\pi_{\widetilde{E}}}(s,a)/d_O(s,a)$ for every state-action pair $(s,a) \in \mathcal{D}_O$.

An important observation is that the state constraints (2) can be reduced to state-action constraints, by training independently an expert policy $\pi_{\widetilde{E}}$, using SMODICE. More precisely, for each skill z we replace the state constraint (2) with the following state-action constraint

$$\mathrm{D_{KL}}\left(d_z(S,A)||d_{\widetilde{E}}(S,A)\right) \le \epsilon, \tag{5}$$

where $d_{\widetilde{E}}(s,a)$ denotes the state-action occupancy $d_{\pi_{\widetilde{E}}}(s,a)$ induced by the expert policy $\pi_{\widetilde{E}}$.

We now consider the Lagrange relaxation of Problem (1) with i) state constraints (2) replaced by state-action constraints (5) and ii) mutual information objective substituted by the variational lower bound in (3), namely

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

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

3.1 Approximation Scheme

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

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

where

$$R_z^\lambda(s,a) := \frac{\log\left(|Z|q(z|s)\right)}{\lambda_z|Z|} + \log \eta_{\widetilde{E}}(s,a). \tag{8}$$

The reward in (8) is derived in Supplementary A and relies on the following equality (see Supplementary B.4) $\mathrm{D_{KL}}(d_z(S,A)||d_{\widetilde{E}}(S,A)) = \mathrm{D_{KL}}(d_z(S,A)||d_O(S,A)) - \mathbb{E}_{d_z(s,a)}[\log \frac{d_{\widetilde{E}}(s,a)}{d_O(s,a)}]$ 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 discriminator, 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 shortly.

3.2 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 train accordingly the Lagrange multipliers λ_z. In addition, we perform a preprocessing phase to compute expert ratios $\eta_{\widetilde{E}}(s,a)$ with respect to a fixed reward $R(s,a) = \log\frac{d_E(s)}{d_O(s)}$ which ensures KL closeness to the expert state occupancy $d_E(S)$, i.e., optimizing Problem (4).

3.2.1 Phase 1

With fixed skill-discriminator $q(z|s)$ and Lagrange multipliers λ, Problem (6) 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:

$$\begin{aligned}\max_{d_z(s,a)\geq 0} \quad & \mathbb{E}_{d_z(s,a)}\left[R^{\lambda}_z(s,a)\right] - \mathrm{D}_{\mathrm{KL}}\left(d_z(S,A)||d_O(S,A)\right) \\ \text{subject to} \quad & \textstyle\sum_a d_z(s,a) = (1-\gamma)\mu_0(s) + \gamma\mathcal{T}d(s) \qquad \forall s,\end{aligned} \tag{10}$$

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 3.1 (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, Assumption 3.1 (which implies strong duality) and Fenchel conjugate, Nachum and Dai [2020, Section 6] and Ma et al. [2022, Theorem 2] showed that Problem 10 shares the same optimal value as the following optimization problem

$$V^\star = \arg\min_{V(s)}(1-\gamma)\mathbb{E}_{s\sim\mu_0}\left[V(s)\right] + \log\mathbb{E}_{d_O(s,a)}\exp\left\{R^{\lambda}_z(s,a) + \gamma\mathcal{T}V(s,a) - V(s)\right\}, \tag{11}$$

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

$$\eta_z(s,a) := \frac{d^\star_z(s,a)}{d_O(s,a)} = \mathrm{softmax}\left(R^{\lambda}_z(s,a) + \gamma\mathcal{T}V^\star_z(s,a) - V^\star_z(s)\right). \tag{12}$$

These ratios allow us to train a skill-conditioned policy π_z by importance-weighted behavior cloning.

Lemma 3.2. *Given a fixed skill-discriminator $q(z|s)$, Lagrange multipliers $\lambda \in \mathbb{R}^{|Z|}_{>0}$ and (primal) optimal ratios $\eta_z(s,a)$, using weighted behavioral cloning, we can train offline an optimal skill conditioned policy π_z. In particular, we optimize by gradient descent the following optimization problem $\max_\pi \mathbb{E}_{p(z)}\mathbb{E}_{d_O(s,a)}\left[\eta_z(s,a)\log\pi_z(a|s)\right]$.*

Given a fixed discriminator $q(z|s)$, we obtain by Lemma 3.2 an optimal policy $\pi^\star_z$. In the next phase, we show how to train off-policy the skill discriminator.

3.2.2 Phase 2

In the following Lemma we give an offline procedure for training an optimal discriminator $q(z|s)$ with respect to the learned policy $\pi^\star_z$. We present the proof in Supplementary B.3.

Lemma 3.3. *Given a fixed skill-conditioned policy $\pi^\star_z$, Lagrange dual variable $\lambda \in \mathbb{R}^{|Z|}_{>0}$ and (primal) optimal ratios $\eta_z(s,a)$, using weighted behavioral cloning, we can train offline an optimal skill-discriminator $q(z|s)$. In particular, we optimize by gradient descent the following optimization problem $\max_{q(z|s)}\mathbb{E}_{p(z)}\mathbb{E}_{d_O(s,a)}\left[\eta_z(s,a)\log\left(q(z|s)\right)\right]$.*

The key insight in Lemma 3.3 is that once we have a skill-conditioned policy π_z, we can train off-policy an optimal discriminator $q(z|s)$ with respect to state-action occupancy $d^\star_z(s,a)$, while sampling from the offline distribution $d_O(s,a)$ and reweighting accordingly by the ratios (12).

In the next phase, we show how to compute offline an estimator of the state-action KL constraint (5).

3.2.3 Phase 3

Here, we fix the discriminator $q(z|s)$ and the skill-conditioned policy $\pi^\star_z(s)$. Then, Problem (6) reduces to

$$\min_{\lambda\geq 0}\sum_z \lambda_z\left[\epsilon - \mathrm{D}_{\mathrm{KL}}\left(d^\star_z(S,A)||d_{\widetilde{E}}(S,A)\right)\right] \tag{13}$$

In the offline setting, it is important to note that direct computation of expectations with respect to the occupancy $d^\star_z(S,A)$ is not feasible. Nevertheless, we show next that combining the expert (4) and (primal) optimal (12) ratios are sufficient to design an off-policy estimator of the KL state-action constraint. More specifically, we give the following self-normalizing importance sampling procedure.

Lemma 3.4. *Given the (primal) optimal ratios $\eta_z(s,a)$ and the classifier $c^\star$, optimizing over the Lagrange multipliers reduces Problem (13) to $\min_{\lambda>0}\sum_z\lambda_z(\epsilon - \phi_z)$, where $\phi_z := \mathbb{E}_{d_O(s,a)}\left[\eta_z(s,a)\log\frac{\eta_z(s,a)}{\eta_{\widetilde{E}}(s,a)}\right]$.*

We give the proof of Lemma 3.4 in Supplementary B.4. Using Lemma 3.2, Lemma 3.3, and Lemma 3.4, we design our multi-phase algorithm and present it in Section 4. In practice, we do not solve the previous three optimization problems to optimality, but rather compute a few gradient descent steps per pass of the offline dataset $\mathcal{D}_O$.

4 Algorithm

The proposed optimization method consists of three phases, each of which optimizes a specific model and fixes the remaining ones. An important difference in comparison to SMODICE, is that our problem formulation considers an optimization problem with constraints. In particular, our reward function is non-stationary, as it depends on the Lagrange dual variables (constraint violation) and the intrinsic motivation signal (log likelihood of the discriminator). This has serious optimization repercussions, as it involves solving a sequence of standard RL problems, each of which can be solved offline by SMODICE.

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

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

We present the resulting multi-phase optimization procedure in Algorithm 1. For a practical implementation, we leverage 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.

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. 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 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, depending on the task at hand.

5 Related Work

In the context of skill discovery, Achiam et al. [2018], Campos et al. [2020] showed that methods like DIAYN 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. [2021] 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. [2020b] developed a method that simultaneously discovers predictable skills and learns their dynamics. In a follow-up work, Park and 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. Hansen et al. [2020] combine behavioral mutual information maximization with successor features, and show that BMI can effectively learn the features needed for constructing reward functions in the successor feature framework. Zahavy et al. [2022] cast the task of learning diverse skills, achieving near-optimal performance w.r.t. on an extrinsic linear reward, into a constrained MDP setting with physics-inspired concave intrinsic reward.

Algorithm 1 Diverse Offline Imitation (DOI)

Pre-compute a discriminator $c^\star : \mathcal{S} \to (0, 1)$ via optimizing the following objective with the gradient penalty in [Gulrajani et al., 2017]

$$\min_c \mathbb{E}_{d_E(s)}[\log c(s)] + \mathbb{E}_{d_O(s)}[\log(1 - c(s))]$$

Use **Phase 1** from below to precompute the following optimal ratios w.r.t. reward $R(s, a) = \log \frac{c^\star(s)}{1-c^\star(s)}$

$$\eta_{\widetilde{E}}(s, a) := \frac{d_{\widetilde{E}}(s, a)}{d_O(s, a)} \qquad \forall s, a \in \mathcal{D}_O \quad \forall z \in Z$$

Repeat until convergence:

Phase 1. (Fixed Lagrange multipliers $\sigma(\mu)$ and discriminator values $q^\star(z|s)$)

For each skill z: **compute** a value function $V_z^\star$ optimizing Equation (11) and ratios

$$\eta_z(s, a) := \frac{d_z^\star(s, a)}{d_O(s, a)} = \operatorname{softmax}\left(R_z^\mu(s, a) + \gamma \mathcal{T} V_z^\star(s, a) - V_z^\star(s)\right) \qquad \forall s, a \in \mathcal{D}$$

Phase 2. (Fixed ratios $\eta_z(s, a)$ and Lagrange multipliers $\sigma(\mu)$) Train a discriminator

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

Train a skill-conditioned policy (used in evaluation)

$$\pi_z^\star = \arg\max_{\pi_z} \mathbb{E}_{d_O(s,a)} \left[\eta_z(s, a) \log \pi_z(a|s)\right] \quad \forall z \in Z$$

Phase 3. (Fixed ratios $\eta_{\widetilde{E}}(s, a)$ and $\eta_z(s, a)$) Compute for each skill z an estimator

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

Optimize the loss

$$\min_\mu \sum_z \sigma(\mu_z)(\epsilon - \phi_z)$$

The diversity is measured using the successor feature ℓ_2 distance between the state occupancies of different skills.

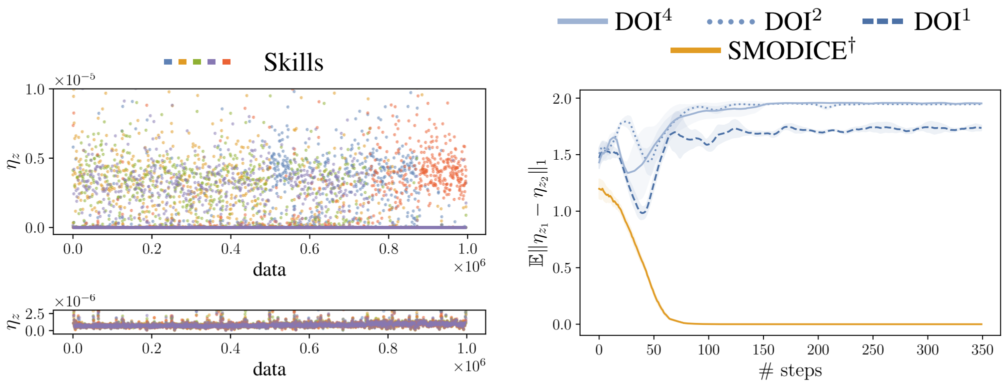
6 Experiments

We evaluate the proposed method for diverse offline imitation on a 12 degree-of-freedom quadruped robot, SOLO12 [Grimminger et al., 2020], on a simple locomotion task in both *simulation* and the *real* system. For this we had collected random and expert data from simulation in the IsaacGym [Makoviychuk et al., 2021]. The datasets are collected using the saved checkpoints obtained by training the robot to track certain velocity of the base using an on-policy version of DOMINO [Zahavy et al., 2022]. We fix the forward velocity to 1 m/s and the turning velocity to zero for collecting both *offline dataset* and *expert dataset*. We defer the training procedure of the policies used for data collection to the Supplementary C.

The *expert dataset* was collected using the deterministic policy from the final checkpoint of the trained procedure with only the best skill of tracking the forward velocity. The *offline dataset* was collected using the stochastic policies collected from different checkpoints during training the expert with multiple skill latents. Also, enabling domain randomization helps to collect various data for both datasets and for better sim-to-real results. It is worth noting that that more than half of the *offline dataset* was collected using the first checkpoint, which represents a policy with random Gaussian actions only. To satisfy the expert coverage part in Assumption 2.1, a fraction of 1/160 of the *offline dataset* is expert data. The *expert dataset* is used to learn a state classifier, in order to compute the ratios $\eta_{\widetilde{E}}(s, a)$.

We trained the policy for 350 steps, where each training step involves multiple stages as described in Section 4. In each stage, we execute 200 epochs of batched training over the data. For the computation of the ratios $\eta_z(s, a)$, we choose a projection Π of the expert state, see Section 4, that yields 3-dimensional planar and angular velocities of the robot's base in the base frame.

We have further found that fitting the 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 discriminator, and is implemented in practice by an ensemble of randomly initialized discriminators. Due to the high initial disagreement on unvisited states, this intrinsic reward provides a strong exploration signal and leads to more diverse skill discovery. Intuitively, for states with small epistemic uncertainty, the discriminator (averaged over the ensemble members) should reliably discriminate between skills, and thus making the intrinsic reward of the discriminator's log-likelihood more accurate. We defer further experimental details to Supplementary D.



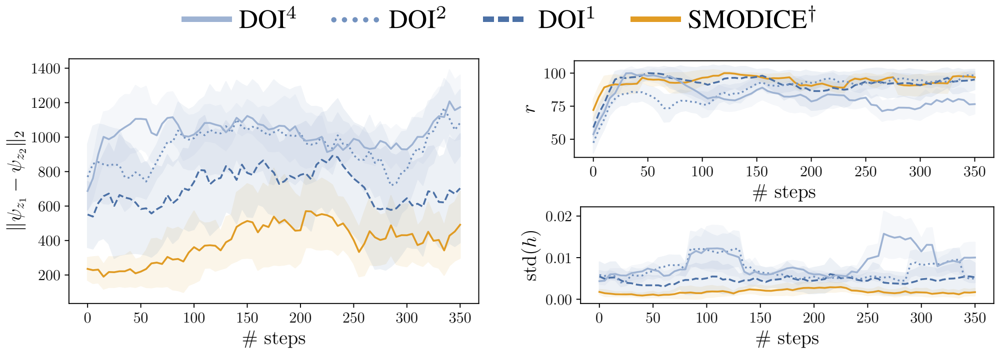
(a) (b)

Figure 2: Data and importance weight η_z separation given different levels of ε. (a) Distribution of importance weights $\eta_z(s, a)$ over dataset for different skills with DOI^4 ($\epsilon = 4$) (upper) and skill-conditioned SMODICE (lower). (b) Average ℓ_1 distance of η_z's belonging to different skills depending on ε. Higher levels of ϵ cause larger differences in attributed importance.

As a baseline, we consider a skill-conditioned SMODICE variant denoted as SMODICE† that does not have access to the discriminator. This is equivalent to setting $\epsilon = 0$ for DOI. In Figure 2 we measure how different are the data attributions with respect to the constraint levels ϵ. As expected, higher ϵ allow more flexibility, and therefore different data points obtain different importances for different skills. To identify this, we compute $\mathbb{E}\|\eta_{z_i} - \eta_{z_j}\|_1$ as a proxy metric for diversity and report it in Figure 2. We note that the looser the constraint (lighter color), the easier it is to 'diversify' in the sense of η_z. In Figure 2a we observe diversification across the dataset assignment to skills in case of using DOI, in contrast, simply training an ensemble of experts on the data corresponding to $\sigma(\mu_z) = 1$ collapses to nearly the same importance per skill per data point. Figure 2b shows the average ℓ_1 distance between skill importance vectors η_z over the data for $\epsilon \in \{0.0, 1.0, 2.0, 4.0\}$ (lighter color indicates higher ϵ). In all the figures, we denote with DOI$^\epsilon$ the different constraint levels. As expected, for a more conservative constraint, the data importances are more similar across skills.

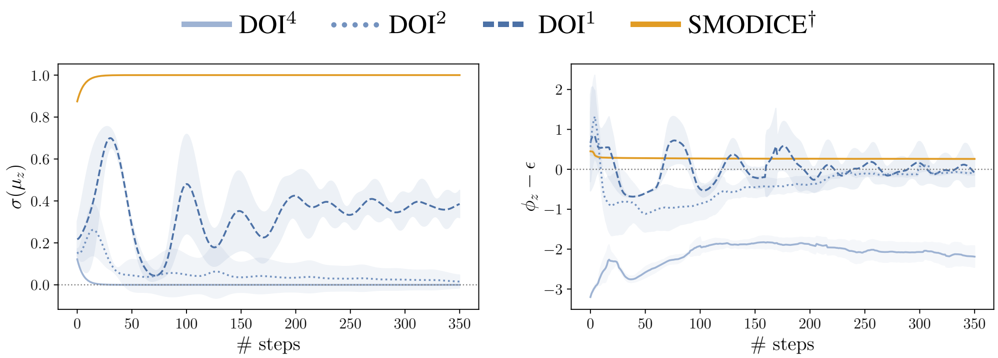
We have further evaluated diversity on the Monte Carlo estimates of the expected successor representation of the initial state, ψ_z. As a diversity metric, we take $\|\psi_{z_1} - \psi_{z_2}\|_2$. The results can be seen in Figure 3, and they nicely align with the proxy diversity metric, meaning that separation of data that is indicated by η_z also indicates higher distance amongst successor representations ψ_z. In terms of performance, DOI is able to achieve forward velocity comparable to the expert (see Figure 3a) while diversifying the behavior in terms of base height h (Figure 3b).

In Figure 4 we observe the behavior of the Lagrange multipliers for different levels of ϵ for a specific skill z. 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 4b 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 $\sigma(\lambda_z) = 1$. Similarly, if the constraint is too weak, only diversity is optimized, in which case there is a significant degradation in performance (see figure Figure 3).



(a) (b)

Figure 3: Average ℓ_2 distance between Monte Carlo estimated successor representations ψ_z of different skills (a), return r as $\%$ of expert return and standard deviation of base height $\mathrm{std}_z(h)$ (b), depending on ϵ.



(a) (b)

Figure 4: 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.

7 Real Robot Experiments

We successfully deployed policies that exhibit diverse skills extracted from an offline dataset, while being able to track a certain velocity similar to an expert on real hardware. Our skill-conditioned policy exhibits different walking behaviors, each with a distinct base height. We provide below snapshots of these diverse behaviors.

More precisely, we evaluate the proposed method for diverse offline imitation on a 12 degree-of-freedom quadruped robot, SOLO12 [Grimminger et al., 2020], on a simple locomotion task in both *simulation* and the *real* system. For this we had collected random and expert data from simulation in the IsaacGym [Makoviychuk et al., 2021]. The datasets are collected using the saved checkpoints obtained by training the robot to track certain velocity of the base using an on-policy version of DOMINO [Zahavy et al., 2022]. We fix the forward velocity to 1 m/s and the turning velocity to zero for collecting both *offline dataset* and *expert dataset*. The training procedure of the policies used for data collection is given in Supplementary C.

Result. 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 heights. Snapshots of these diverse behaviors are presented in Figure 5.

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

8 Conclusion

We propose an offline optimization method for maximizing a diversity objective, formulated in terms of mutual information, which is constrained to have small KL-divergence with respect to a fixed target state distribution. Using the Fenchel duality, we derive a principled and practical reinforcement learning algorithm for offline unsupervised skill discovery, which we also validate through experiments in both simulation and on real hardware. We considered an ℓ_2 distance of expected successor representations across skills as our diversity metric. The experimental results confirm the expected behavior, i.e., a stronger constraint causes the policy to be closer to the to the expert and less diverse. Further, to validate the diversity, we show that the skill-conditioned policy clusters the state-action pairs in the offline dataset (using the skill-specific importance weights), in the case of non-zero Lagrange multipliers.

Acknowledgements

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, EXC number 2064/1 – Project number 390727645. Pavel Kolev was supported by the Cyber Valley Research Fund and the Volkswagen Stiftung (No 98 571).

References

J. Achiam, H. Edwards, D. Amodei, and P. Abbeel. Variational option discovery algorithms. *CoRR*, abs/1807.10299, 2018. URL http://arxiv.org/abs/1807.10299.

A. Barreto, W. Dabney, R. Munos, J. J. Hunt, T. Schaul, H. Van Hasselt, and D. Silver. Successor features for transfer in reinforcement learning. *arXiv preprint arXiv:1606.05312*, 2016.

A. Barreto, D. Borsa, J. Quan, T. Schaul, D. Silver, M. Hessel, D. Mankowitz, A. Zidek, and R. Munos. Transfer in deep reinforcement learning using successor features and generalised policy improvement. In *International Conference on Machine Learning*, pages 501–510. PMLR, 2018.

S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

V. Campos, A. Trott, C. Xiong, R. Socher, X. Giró-i-Nieto, and J. 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*, pages 1317–1327. PMLR, 2020. URL http://proceedings.mlr.press/v119/campos20a.html.

B. Dai, N. He, Y. Pan, B. Boots, and L. Song. Learning from conditional distributions via dual embeddings. In *Artificial Intelligence and Statistics*, pages 1458–1467. PMLR, 2017.

B. Dai, O. Nachum, Y. Chow, L. Li, C. Szepesvári, and D. Schuurmans. Coindice: Off-policy confidence interval estimation. *Advances in neural information processing systems*, 33:9398–9411, 2020.

P. 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.

B. Eysenbach, A. Gupta, J. Ibarz, and S. 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.

J. Fu, A. Kumar, O. Nachum, G. Tucker, and S. Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial networks. *CoRR*, abs/1406.2661, 2014. URL http://arxiv.org/abs/1406.2661.

K. Gregor, D. J. Rezende, and D. 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.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5767–5777, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html.

S. Hansen, W. Dabney, A. Barreto, D. Warde-Farley, T. V. de Wiele, and V. 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.

G.-H. Kim, S. Seo, J. Lee, W. Jeon, H. Hwang, H. Yang, and K.-E. Kim. DemoDICE: Offline imitation learning with supplementary imperfect demonstrations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=BrPdX1bDZkQ.

J. Kim, S. Park, and G. Kim. Unsupervised skill discovery with bottleneck option learning. In M. Meila and T. Zhang, editors, *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*, pages 5572–5582. PMLR, 2021. URL http://proceedings.mlr.press/v139/kim21j.html.

A. Kumar, A. Zhou, G. Tucker, and S. Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

S. Levine, A. Kumar, G. Tucker, and J. Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *CoRR*, abs/2005.01643, 2020.

C. Li, M. Vlastelica, S. Blaes, J. Frey, F. Grimminger, and G. Martius. Learning agile skills via adversarial imitation of rough partial demonstrations. In *Conference on Robot Learning*, pages 342–352. PMLR, 2023.

Y. J. Ma, A. Shen, D. Jayaraman, and O. Bastani. Versatile offline imitation from observations and examples via regularized state-occupancy matching. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 14639–14663. PMLR, 2022. URL https://proceedings.mlr.press/v162/ma22a.html.

V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

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

O. Nachum, Y. Chow, B. Dai, and L. Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in Neural Information Processing Systems*, 32, 2019a.

O. Nachum, B. Dai, I. Kostrikov, Y. Chow, L. Li, and D. Schuurmans. Algaedice: Policy gradient from arbitrary experience, 2019b.

S. Park and S. Levine. Predictable MDP abstraction for unsupervised model-based RL. *CoRR*, abs/2302.03921, 2023. doi: 10.48550/arXiv.2302.03921. URL https://doi.org/10.48550/arXiv.2302.03921.

S. Park, J. Choi, J. Kim, H. Lee, and G. 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.

S. Park, K. Lee, Y. Lee, and P. 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.

D. Precup. *Temporal abstraction in reinforcement learning*. University of Massachusetts Amherst, 2000.

R. F. Prudencio, M. R. O. A. Maximo, and E. L. Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *CoRR*, abs/2203.01387, 2022.

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

A. Sharma, S. Gu, S. Levine, V. Kumar, and K. Hausman. Dynamics-aware unsupervised discovery of skills. In *International Conference on Learning Representations*, 2020a.

A. Sharma, S. Gu, S. Levine, V. Kumar, and K. 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, 2020b. URL https://openreview.net/forum?id=HJgLZR4KvH.

A. Stooke, J. Achiam, and P. 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*, pages 9133–9143. PMLR, 2020. URL http://proceedings.mlr.press/v119/stooke20a.html.

D. Strouse, K. Baumli, D. Warde-Farley, V. Mnih, and S. S. 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.

Z. Wang, A. Novikov, K. Zolna, J. S. Merel, J. T. Springenberg, S. E. Reed, B. Shahriari, N. Siegel, C. Gulcehre, N. Heess, et al. Critic regularized regression. *Advances in Neural Information Processing Systems*, 33:7768–7778, 2020.

T. Zahavy, Y. Schroecker, F. M. P. Behbahani, K. Baumli, S. Flennerhag, S. Hou, and S. 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.

S. Zhang, B. Liu, and S. Whiteson. Gradientdice: Rethinking generalized offline estimation of stationary values. In *International Conference on Machine Learning*, pages 11194–11203. PMLR, 2020.

Supplementary for Diverse Offline Imitation via Fenchel Duality

A Lagrange Relaxation

The Lagrange relaxation is given by

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

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

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

and thus

$$\max_{d_z(s,a),q(z|s)} \min_{\lambda>0} \sum_z \lambda_z \left[\epsilon + \mathbb{E}_{d_z(s,a)} \left[R_z^\lambda(s,a) \right] - \mathrm{D}_{\mathrm{KL}} \left(d_z(S,A) || d_O(S,A) \right) \right], \tag{S1}$$

where the reward is given by

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

B Algorithmic Phases

B.1 Value Function Training

With fixed skill-discriminator $q(z|s)$ and Lagrange multipliers $\lambda > 0$, the Problem S1 becomes:

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

or equivalently for every skill z:

$$\begin{array}{ll} \max\limits_{d_z(s,a)\geq 0} & \mathbb{E}_{d_z(s,a)} \left[R_z^\lambda(s,a) \right] - \mathrm{D}_{\mathrm{KL}} \left(d_z(S,A) || d_O(S,A) \right) \\ \text{s.t.} & \sum_a d_z(s,a) = (1-\gamma)\mu_0(s) + \gamma \mathcal{T} d(s) \quad \forall s. \end{array} \tag{S2}$$

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

The strict feasibility in Assumption 3.1 implies strong duality, and thus Problem (S2) shares the same optimal value as the following dual minimization problem (for details see [Nachum and Dai, 2020, Section 6] and [Ma et al., 2022, Theorem 2]):

$$\begin{array}{rl} V^\star = & \arg\min_{V(s)} (1-\gamma) \mathbb{E}_{s\sim\mu_0} \left[V(s) \right] \\ & + \log \mathbb{E}_{d^{\pi_O}(s,a)} \exp \left\{ R_z^\lambda(s,a) + \gamma \mathcal{T} V(s,a) - V(s) \right\}, \end{array} \tag{S3}$$

where

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

Moreover, the optimal primal solution reads

$$\frac{d_z^\star(s,a)}{d_O(s,a)} = \mathrm{softmax} \left(R_z^\lambda(s,a) + \gamma \mathcal{T} V_z^\star(s,a) - V_z^\star(s) \right). \tag{S4}$$

B.2 Weighted Behavioral Cloning

The proof of Lemma 3.2 follows the approach in Ma et al. [2022, SMODICE], which is briefly summarized below for completeness. We pretrain a state discriminator $c^\star(s)$, by optimizing an objective and a gradient penalty as in Goodfellow et al. [2014] and Gulrajani et al. [2017], that distinguishes between expert and offline states. The Bayes optimal classifier $c^\star$ satisfies $\frac{c^\star(s)}{1-c^\star(s)} = \frac{d_E(s,a)}{d_O(s,a)}$, and thus $\log \frac{c^\star(s)}{1-c^\star(s)} = \log \frac{d_E(s)}{d_O(s)}$. Solving Problem (10) with fixed rewards $R_z^\lambda(s,a)$, yields dual optimal value function $V^\star$. Using Fenchel duality, see (12), we compute (primal) optimal ratios $\eta_z(s,a)$ which we further use for training off-policy, via importance-weighted behavior cloning, the skill-conditioned policy π_z.

B.3 Skill Discriminator Training

With fixed skill-conditioned policy $\pi_z^\star$ and Lagrange multipliers $\lambda > 0$, the Problem S1 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 B.1. *Given a fixed skill-conditioned policy $\pi_z^\star$, Lagrange dual variable $\lambda > 0$ and ratios $\eta_z(s,a)$, using weighted behavioral cloning, 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 Lemma B.2 and Lemma B.3. □

Lemma B.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

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

□

Lemma B.3 (Importance Sampling)**.** *Given ratios $\eta_z(s,a)$, it holds for any function $f(s)$ that*

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

Proof. Observe that

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

□

B.4 Estimating State KL Constraint Violation

Lemma B.4 (Structural)**.** *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. We have

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

□

Lemma B.5 (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 $\eta_\pi(s,a) = \frac{d_\pi(s,a)}{d_O(s,a)}$ 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*

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

Proof. By Claim 1, we have

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

For the first term, we have

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

The second term reduces to

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

□

C Dataset Collection

Both *expert dataset* and *offline dataset* are collected using locomotion policies trained to track certain velocity in IsaacGym [Makoviychuk et al., 2021]. The policies are trained using an on-policy version of DOMINO [Zahavy et al., 2022] to exhibit diverse behaviors while maintaining a certain level of tracking. Even trained with randomly sampled velocity, the policies are fed with forward velocity of 1 m/s when collecting both datasets. Both datasets contain 4000 trajectories with an episode length of 250 steps, or 1 million transitions each.

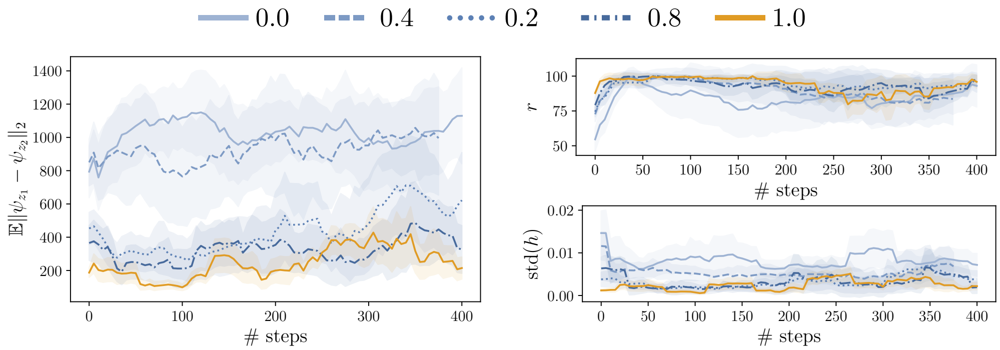
We summarize the main ideas of the training procedure, for details see [Zahavy et al., 2022]. Using DOMINO, we train policies that are conditioned on discrete skill latents and present different behaviors across different skills. Each skill-conditioned policy has a designated skill which is trained with only extrinsic reward and is maintained as the target in the constraint formulation in [Zahavy et al., 2022]. We use this target skill from the last training checkpoint (iteration 2000) as the expert of our method. For each skill-conditioned policy, all skills except the target, are trained to balance between extrinsic and intrinsic reward, so as to generate diverse behaviours while being aligned to some degree to the target skill, i.e., maintaining a certain level of tracking velocity. The intrinsic reward is designed to maximize the ℓ_2 distance of the successor features [Barreto et al., 2016] between different skills, where in our setting the feature space includes: the base height velocity, base roll and pitch velocities, and feet height velocities.

We collected the *offline dataset* using these skill-conditioned policy from different checkpoints during training. 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 (checkpoint 0 and 50).

Furthermore, in the data collection process, we use a deterministic policy for the *expert dataset*, while for the *offline dataset* we use a stochastic policy. Randomizing the action selection in the latter case, results in more diverse interactions with the environment. 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 $[0.5, 1.5]$ and additional base mass between $[-0.5, 0.5]$ kg, as well as simulate the observation noise and an actuator lag of 15 ms.

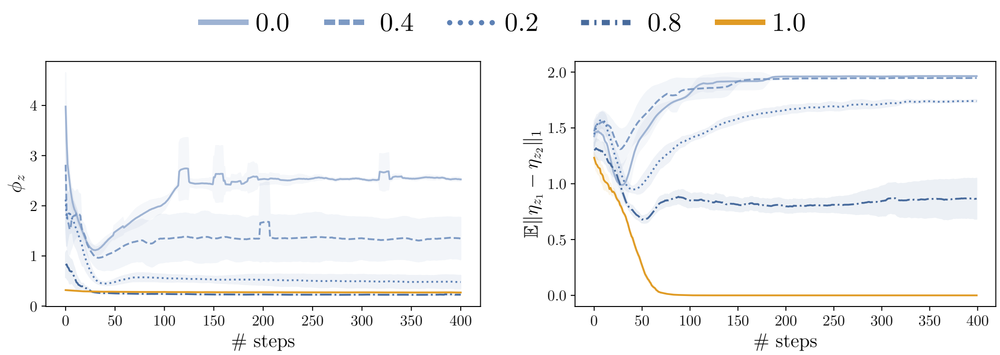
D Additional Experiments

Instead of learning the Lagrangian multipliers μ_z via KL estimators ϕ_z, we can also fix μ_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 S1 and S2.



(a) (b)

Figure S1: (a) Average ℓ_2 distance between Monte Carlo estimated successor representations ψ_z of different skills, (b) return r as % of expert return and standard deviation of base height $\mathrm{std}_z(h)$, depending on a fixed $\sigma(\mu_z)$ (see legend).



(a) (b)

Figure S2: Divergence estimate and η_z distance for the case of fixed $\sigma(\mu_z)$. (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.