Diverse Offline Imitation via Fenchel Duality

Anonymous Author(s)
Affiliation
Address
email

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-Rockafellar duality, reinforcement learning and unsupervised skill discovery, and to give a simple offline algorithm for learning diverse skills that are aligned with an expert.

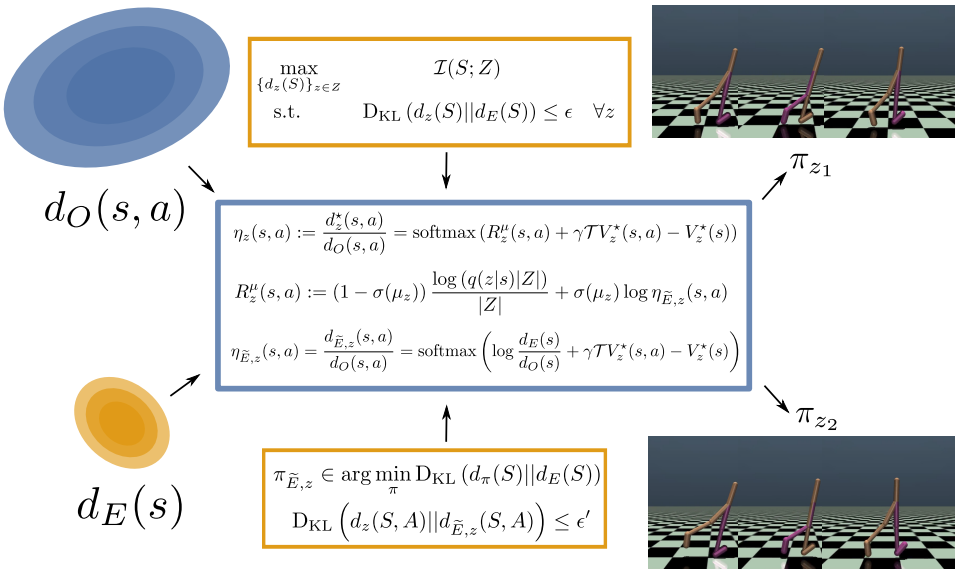

Figure 1: DOI maximizes a variational lower bound on the mutual information between skill z and states s subject to an KL-divergence constraint. This requires offline datasets $\mathcal{D}_E$ and $\mathcal{D}_O$ sampled uniformly at random from state $d_E(s)$ and state-action $d_O(s,a)$ occupancies, respectively, in order to compute specific importance weights $\eta_z(s,a)$ of learned skill and $\eta_{\widetilde{E},z}(s,a)$ pretrained expert.

Submitted to 16th European Workshop on Reinforcement Learning (EWRL 2023). Do not distribute.

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., 2020a) 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., 2020b). In the age of data abundance, focus is turning towards offline learning which leverages collected experience from various sources. In the offline setting, 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 described through z and a state representation (Barreto et al., 2016; Hansen et al., 2020; Barreto et al., 2018). 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-discussed in previous work (Levine et al., 2018). In most approaches, the off-policiness is accounted for by trying to remain 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. Similar to existing approaches (Nachum et al., 2019a,b; Kim et al., 2022; Zhang et al., 2020; Dai et al., 2020), the Fenchel-Rockafellar duality allow 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. Due to our constraint optimization formulation, we also compute constraint violation estimators and train the corresponding Lagrange multipliers.

In this work, we consider the setting in which access to expert demonstrations without rewards is 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 (Li et al., 2023; Ma et al., 2022), demonstration data oftentimes does not match the policy action space or we don't 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 Lagrange duality and Fenchel conjugate 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. We further draw connections to a broader framework where the target distribution in the KL-divergence constraint can be substituted by any state distribution with full support.

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 defines a probability distribution over actions conditioned on the state $\pi : \mathcal{S} \mapsto \mathcal{A}$. 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 π, the state-action occupancy distribution reads

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

The 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, we define the set of skills Z, which we will treat as a finite set and analogously the skill-conditioned policies $\pi_z : S \times Z \mapsto \Delta(A)$ that each induce their respective state-action occupancy distributions $d_z(s,a)$. Of particular interest to us are going to be the expert state occupancy $d_E(s)$ and the offline data occupancy $d_O(s)$.

Given two probability distributions p and q, we define for convex continuous functions f the family of f-divergences $\mathrm{D}_f(p||q) = \mathbb{E}_q[f(p(x)/q(x))]$. In the special case when $f(x) = x\log(x)$, we obtain the well-known KL-divergence. The use of f divergences as dissimilarity measures in the domain of state occupancies offers significant advantages. This is mainly due to the availability of closed-form solutions of Fenchel conjugates for certain f-divergences, which are widely used in convex analysis (Boyd and Vandenberghe, 2004).

For any probability distributions $p, q \in \triangle^S$ such that $p(s) > 0$ implies $q(s) > 0$, by letting $y(s) = \frac{p(s)}{q(s)}$, the Fenchel conjugate D_f of a function f, satisfying certain regularity conditions (Dai et al., 2017), is defined by

$$\mathrm{D}_{\star,f}(y) = \mathbb{E}_{q(s)}\left[f_\star(y(s))\right]. \tag{1}$$

Furthermore, a maximizer of equation (1), which we will use extensively in our analysis, satisfies

$$p^\star(s) = q(s) f'_\star(y(s)). \tag{2}$$

3 Method

Although the framework studied in (Nachum and Dai, 2020; Ma et al., 2022) encompasses the general f-divergence setting, we focus here specifically on the KL-divergence for clarity and to simplify our presentation. In line with this previous work, we use the following standard assumption.

Assumption 3.1. *We assume (expert coverage) $d_E(s) > 0$ implies $d_O(s) > 0$; and (strict feasibility) there exists at least one $d(s,a)$ such that constraints (12) are satisfied and $d(s) > 0$ for all $s \in \mathcal{S}$.*

Through this work, we denote by π_z a skill-conditioned policy and with $d_z(s) := d^{\pi_z}(s)$ the corresponding state occupancy. For a fixed discrete set of skills Z, we seek to solve the following constrained optimization problem, which optimizes over all skill-conditioned policies $\{\pi_z\}_{z\in Z}$,

$$\max_{\{d_z(S)\}_{z\in Z}} \quad \mathcal{I}(S;Z) \tag{3}$$

$$\text{s.t.} \quad \mathrm{D}_{\mathrm{KL}}\left(d_z(S)||d_E(S)\right) \le \epsilon \quad \forall z, \tag{4}$$

where $\mathcal{I}(S;Z)$ denotes the mutual information between states and skills, and D_{KL} is the KL-divergence.

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

$$\mathcal{I}(S;Z) \ge \mathbb{E}_{p(z),d_z(s)}\left[\log q(z|s)\right] + \mathcal{H}\left(p(z)\right) = \sum_z \mathbb{E}_{d_z(s)}\left[\frac{\log\left(|Z| q(z|s)\right)}{|Z|}\right]. \tag{5}$$

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 approximately solves

$$\min_\pi \mathrm{D}_{\mathrm{KL}}\left(d_\pi(S)||d_E(S)\right), \tag{6}$$

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 (4) can be reduced to state-action constraints, by training independently an ensemble of expert policies $\{\pi_{\widetilde{E},z}\}_{z\in Z}$, using SMODICE. More precisely, for each skill z we replace the state constraint (4) with the following state-action constraint

$$\mathrm{D_{KL}}\left(d_z(S,A)||d_{\widetilde{E},z}(S,A)\right) \leq \epsilon, \tag{7}$$

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

We now consider the Lagrange relaxation of Problem (3) with i) state constraints (4) replaced by state-action constraints (7) and ii) mutual information objective substituted by the variational lower bound in (5), namely

$$\max_{\substack{d_z(s,a)\\ q(z|s)}} \min_{\lambda\geq 0} \sum_z \mathbb{E}_{d_z(s)}\left[\frac{\log\left(|Z|q(z|s)\right)}{|Z|}\right] + \sum_z \lambda_z\left[\epsilon - \mathrm{D_{KL}}\left(d_z(S,A)||d_{\widetilde{E},z}(S,A)\right)\right], \tag{8}$$

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 (8). 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 (8) with $\lambda > 0$, that is

$$\max_{\substack{d_z(s,a)\\ q(z|s)}} \min_{\lambda>0} \sum_z \lambda_z\Big\{\epsilon + \mathbb{E}_{d_z(s,a)}\left[R_z^\lambda(s,a)\right] - \mathrm{D_{KL}}\left(d_z(S,A)||d_O(S,A)\right)\Big\}, \tag{9}$$

where

$$R_z^\lambda(s,a) := \frac{\log\left(|Z|q(z|s)\right)}{\lambda_z|Z|} + \log\eta_{\widetilde{E},z}(s,a). \tag{10}$$

The reward in (10) is derived in Appendix A and relies on the following equality (see Appendix B.3) $\mathrm{D_{KL}}(d_z(S,A)||d_{\widetilde{E},z}(S,A)) = \mathrm{D_{KL}}(d_z(S,A)||d_O(S,A)) - \mathbb{E}_{d_z(s,a)}[\log\frac{d_{\widetilde{E},z}(s,a)}{d_O(s,a)}]$ and the definition of $\eta_{\widetilde{E},z}(s,a) = d_{\widetilde{E},z}(s,a)/d_O(s,a)$.

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 (7), and vice versa. Each term in the reward (10) 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 $\eta_{\widetilde{E},z}(s,a)$ 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 (6).

3.2.1 Phase 1

With fixed skill-discriminator $q(z|s)$ and Lagrange multipliers λ, Problem (8) becomes

$$\max_{\{d_z(s,a)\}_{z\in Z}} \sum_z \lambda_z\Big\{\mathbb{E}_{d_z(s,a)}\left[R_z^\lambda(s,a)\right] - \mathrm{D_{KL}}\left(d_z(S,A)||d_O(S,A)\right)\Big\}, \tag{11}$$

or equivalently for every skill z:

$$\begin{aligned} \max_{d_z(s,a)\geq 0} \quad & \mathbb{E}_{d_z(s,a)}\left[R_z^\lambda(s,a)\right] - \mathrm{D_{KL}}\left(d_z(S,A)||d_O(S,A)\right) \\ \text{s.t.} \quad & \textstyle\sum_a d_z(s,a) = (1-\gamma)\mu_0(s) + \gamma\mathcal{T}d(s) \quad \forall s. \end{aligned} \tag{12}$$

Using the Lagrange duality and Fenchel conjugate, it can be shown (see Appendix B.1) that the following optimization problem shares the same optimal value as Problem 12

$$V^\star = \arg\min_{V(s)} (1-\gamma)\mathbb{E}_{s\sim\mu_0}\left[V(s)\right] + \log \mathbb{E}_{d_O(s,a)} \exp\left\{R_z^\lambda(s,a) + \gamma \mathcal{T}V(s,a) - V(s)\right\}, \quad (13)$$

where $\mathcal{T}V(s,a) := \mathbb{E}_{\mathcal{P}(s'|s,a)}V(s')$. Moreover, by combining Assumption 3.1, Fenchel-Rockafellar duality and equation (2), it follows that the optimal primal solution satisfies

$$\eta_z(s,a) := \frac{d_z^\star(s,a)}{d_O(s,a)} = \mathrm{softmax}\left(R_z^\lambda(s,a) + \gamma \mathcal{T}V_z^\star(s,a) - V_z^\star(s)\right). \quad (14)$$

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 $\lambda \in \mathbb{R}_{>0}^{|Z|}$ 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*

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

Proof (Sketch). We follow 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 as in (Goodfellow et al., 2014), 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 (12) with fixed rewards $R_z^\lambda(s,a)$, yields dual optimal value function $V^\star$. Using Fenchel-Rockafellar duality, see (14), 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. □

Given a fixed discriminator $q(z|s)$, we obtain by Lemma 3.2 an optimal policy $\pi_z^\star$. 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 $\pi_z^\star$. We present the proof in Appendix B.2.

Lemma 3.3. *Given a fixed skill-conditioned policy $\pi_z^\star$, Lagrange dual variable $\lambda \in \mathbb{R}_{>0}^{|Z|}$ 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*

$$\max_{q(z|s)} \mathbb{E}_{p(z)}\mathbb{E}_{d_O(s,a)}\left[\eta_z(s,a)\log\left(q(z|s)\right)\right]. \quad (15)$$

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 $d_z^\star(s,a)$, while sampling from the offline distribution $d_O(s,a)$ and reweighting accordingly by the ratios (14).

In the next phase, we show how to compute offline an estimator of the state-action KL constraint (7).

3.2.3 Phase 3

Here, we fix the discriminator $q(z|s)$ and the skill-conditioned policy $\pi_z^\star(s)$. Then, Problem (8) reduces to

$$\min_{\lambda\geq 0} \sum_z \lambda_z \left[\epsilon - \mathrm{D_{KL}}\left(d_z^\star(S,A)||d_{\widetilde{E},z}(S,A)\right)\right] \quad (16)$$

In the offline setting, it is important to note that direct computation of expectations with respect to the occupancy $d_z^\star(S,A)$ is not feasible. Nevertheless, we show next that combining the expert (6) and (primal) optimal (14) ratios is 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 (16) to*

$$\min_{\lambda>0} \sum_z \lambda_z(\epsilon - \phi_z), \quad (17)$$

where we denoted by

$$\phi_z := \mathbb{E}_{d_O(s,a)}\left[\eta_z(s,a)\log\frac{\eta_z(s,a)}{\eta_{\widetilde{E},z}(s,a)}\right]. \tag{18}$$

Proof (Sketch). We show in Appendix B.3 (see Lemma B.5) that

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

□

Using Lemma 3.2, Lemmas 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 $\mathcal{D}_O$.

3.3 Target State Distributions

We have shown that Problem 3 can be relaxed to the formulation in Problem 8 which can be approximately solved by applying a three phase alternation optimization scheme. In principle, the expert state occupancy $d_E(S)$ in Problem 3 can be replaced by an arbitrary target state distribution. We now present an important special case, for which we establish a quantitative relationship between the ϵ violation threshold in the state KL constraint (4) and the corresponding cumulative return of the learned policy.

We consider a β-softmax target distribution defined through a state reward $r(s)$, namely

$$p^\beta(s) := \frac{\exp\{\beta r(s)\}}{\sum_{s'}\exp\{\beta r(s')\}}.$$

Further, the associated β-softmax maximum is defined by

$$r^\beta_{\text{smax}} := \frac{1}{\beta}\log\sum_{s'}\exp\{\beta r(s')\}.$$

Lemma 3.5. ***[Boltzmann]*** *For any $\beta, \delta > 0$ it holds that*

$$\mathrm{D_{KL}}\left(d_z(S)||p^\beta(S)\right) \le \delta r^\beta_{\text{smax}} \iff \mathbb{E}_{d_z(s)}\left[r(s)\right] + \frac{1}{\beta}\mathcal{H}(d_z(S)) \ge \left(1-\frac{\delta}{\beta}\right) r^\beta_{\text{smax}}. \tag{19}$$

We give the proof in Appendix C (see Lemma 3.5). This provides a natural connection to state-occupancy constraint and reward with entropy regularization. In particular, as β increases, the entropy regularization disappears, leaving only the return.

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 and depends on the Lagrange dual variables (constraint violation) and in addition to 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 $\lambda = \sigma(\mu)$ to unbounded variables $\mu \in \mathbb{R}^{|Z|}$, so that the effective reward is a convex combination of the diversity and constraint violation term. In practice, this transformation ensures that $\lambda > 0$. Hence, the reward for each skill z becomes

$$R^\mu_z(s,a) := (1-\sigma(\mu_z))\frac{\log\left(q(z|s)|Z|\right)}{|Z|} + \sigma(\mu_z)\log\eta_{\widetilde{E},z}(s,a). \tag{20}$$

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 the ensemble of expert policies $\{\pi_{\widetilde{E},z}\}_{z\in Z}$ and skill-conditioned policies $\{\pi_z\}_{z\in Z}$ using two instantiations of a skill-conditioned Value neural network. While convenient, this means that each phase of Algorithm 1 is only approximately solved.

Algorithm 1 DOI

Pre-compute:
Compute discriminator $c^\star : \mathcal{S} \to (0,1)$ via optimizing

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

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

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

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 (13) and ratios

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

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

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

Optimize the loss

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

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. (2020a) 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. The diversity is measured using the successor feature ℓ_2 distance between the state occupancies of different skills.

6 Discussion

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. Two important special cases for the target state distribution were considered: an expert state occupancy (imitating the expert behavior) and a Boltzmann distribution induced by a state reward (visiting beneficial states). Exploiting the Fenchel-Rockafellar duality allows us to derive a sound and practical reinforcement learning algorithm for offline unsupervised skill discovery. Preliminary empirical results are promising, but too early to present here. We are looking forward to presenting and discussing them during EWRL.

References

Achiam, J., Edwards, H., Amodei, D., and Abbeel, P. (2018). Variational option discovery algorithms. *CoRR*, abs/1807.10299.

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

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

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

Campos, V., Trott, A., Xiong, C., Socher, R., Giró-i-Nieto, X., and Torres, J. (2020). 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.

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

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

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624.

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. (2019). 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.

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

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

Gregor, K., Rezende, D. J., and Wierstra, D. (2017). Variational intrinsic control. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Hansen, S., Dabney, W., Barreto, A., Warde-Farley, D., de Wiele, T. V., and Mnih, V. (2020). 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.

Kim, G.-H., Seo, S., Lee, J., Jeon, W., Hwang, H., Yang, H., and Kim, K.-E. (2022). DemoDICE: Offline imitation learning with supplementary imperfect demonstrations. In *International Conference on Learning Representations*.

Kim, J., Park, S., and Kim, G. (2021). Unsupervised skill discovery with bottleneck option learning. In Meila, M. and Zhang, T., 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.

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

Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., and Quillen, D. (2018). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436.

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

Ma, Y. J., Shen, A., Jayaraman, D., and Bastani, O. (2022). Versatile offline imitation from observations and examples via regularized state-occupancy matching. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S., 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.

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

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

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

Park, S., Choi, J., Kim, J., Lee, H., and Kim, G. (2022). Lipschitz-constrained unsupervised skill discovery. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Park, S., Lee, K., Lee, Y., and Abbeel, P. (2023). Controllability-aware unsupervised skill discovery. *CoRR*, abs/2302.05103.

Park, S. and Levine, S. (2023). Predictable MDP abstraction for unsupervised model-based RL. *CoRR*, abs/2302.03921.

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

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

Sharma, A., Gu, S., Levine, S., Kumar, V., and Hausman, K. (2020a). Dynamics-aware unsupervised discovery of skills. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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

Stooke, A., Achiam, J., and Abbeel, P. (2020). 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.

Strouse, D., Baumli, K., Warde-Farley, D., Mnih, V., and Hansen, S. S. (2022). Learning more skills through optimistic exploration. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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

Zahavy, T., Schroecker, Y., Behbahani, F. M. P., Baumli, K., Flennerhag, S., Hou, S., and Singh, S. (2022). Discovering policies with domino: Diversity optimization maintaining near optimality. *CoRR*, abs/2205.13521.

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

Supplementary for Diverse Offline Imitation Learning

A Lagrange Relaxation

The Lagrange relaxation is given by

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

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

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

and thus

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

where the reward is given by

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

B Algorithmic Phases

B.1 Value Function Training

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

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

or equivalently for every skill z:

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

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

Fact B.1. (Ma et al., 2022, Theorem 2) Problem (S2) shares the same optimal value as the following minimization problem:

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

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 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.3 Estimating State KL Constraint Violation

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

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

Proof. We have

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

□

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*

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

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 Connection between KL-divergence and Return

Given a state reward function r, for any $\delta, \beta > 0$ we define a β-softmax

$$r^\beta_{\mathrm{smax}} := \frac{1}{\beta}\log \sum_{s'} \exp\left\{\beta r(s')\right\}$$

and a state distribution

$$p^\beta(s) := \frac{\exp\left\{\beta r(s)\right\}}{\sum_{s'} \exp\left\{\beta r(s')\right\}}.$$

We now present a connection between state-based KL-divergence constraint and cumulative return.

Lemma C.1. *[Structural] For any $\beta, \delta > 0$ it holds that*

$$\begin{aligned}&\mathrm{D_{KL}}\left(d_z(S)||p^\beta(S)\right) \le \delta r^\beta_{\mathrm{smax}} \\ \iff\quad &\mathbb{E}_{d_z(s)}\left[r(s)\right] + \frac{1}{\beta}\mathcal{H}(d_z(S)) \ge \left(1-\frac{\delta}{\beta}\right) r^\beta_{\mathrm{smax}}.\end{aligned} \tag{S5}$$

In particular, the larger β is, the tighter inequality (S6) becomes (as p^β concentrates on highest reward states).

Proof. Observe that

$$\begin{aligned}\mathbb{E}_{d_z(s)}\left[\log p^\beta(S)\right] &= \mathbb{E}_{d_z(s)}\left[\log \frac{\exp\left\{\beta r(s)\right\}}{\sum_{s'}\exp\left\{\beta r(s')\right\}}\right] \\ &= -\beta r^\beta_{\mathrm{smax}} + \beta \mathbb{E}_{d_z(s)}\left[r(s)\right].\end{aligned}$$

Since

$$\mathrm{D_{KL}}\left(d_z(S)||p^\beta(S)\right) = -\mathcal{H}(d_z(S)) - \mathbb{E}_{d_z(s)}\left[\log p^\beta(s)\right]$$

we have

$$\begin{aligned}&\mathrm{D_{KL}}\left(d_z(S)||p^\beta(S)\right) \le \epsilon \\ \iff\quad &-\mathcal{H}(d_z(S)) + \beta r^\beta_{\mathrm{smax}} - \beta\mathbb{E}_{d_z(s)}\left[r(s)\right] \le \epsilon \\ \iff\quad &\mathbb{E}_{d_z(s)}\left[r(s)\right] + \frac{1}{\beta}\mathcal{H}(d_z(S)) \ge r^\beta_{\mathrm{smax}} - \frac{\epsilon}{\beta}.\end{aligned}$$

Thus, for $\epsilon = \delta r^\beta_{\mathrm{smax}}$ we have

$$\begin{aligned}&\mathrm{D_{KL}}\left(d_z(S)||p^\beta(S)\right) \le \delta r^\beta_{\mathrm{smax}} \\ \iff\quad &\mathbb{E}_{d_z(s)}\left[r(s)\right] + \frac{1}{\beta}\mathcal{H}(d_z(S)) \ge \left(1-\frac{\delta}{\beta}\right) r^\beta_{\mathrm{smax}}\end{aligned}$$

□