Epistemically-guided forward-backward exploration

Anonymous authors
Paper under double-blind review

Keywords: unsupervised RL, exploration, zero-shot, epistemic uncertainty, ensemble

Summary

The goal of zero-shot RL is to provide algorithms for recovering optimal policies for all possible reward functions given interaction data with the environment. Naturally, how well we can recover the optimal policies highly depends on the quality of the data used to learn them. Up until now, most algorithms leverage decoupled exploration policies for collecting data in order to learn a generalist representation of all optimal policies. A central argument to this paper is that the exploration policy should not be completely decoupled from the zero-shot algorithm and should try to minimize the uncertainty that the algorithm has of its representations. We frame the exploration problem for zero-shot RL as minimization of the epistemic uncertainty on the learned value functions, and realize this in the case of well familiar algorithm, forward-backward (FB) representations. Crucially, in several empirical settings, using an exploration policy that maximizes the cumulative epistemic uncertainty of the FB representation leads to significant improvements of the algorithm's sample complexity. This enables us to learn well-performing policies fast, with fewer amount of data than other exploration approaches.

Contribution(s)

1. This paper phrases the exploration problem for zero-shot RL as uncertainty minimization of a posterior over occupancy measures for a particular representation of an occupancy measure. The main difference to previous work is that, while previous work considers completely off-policy exploration algorithms to collect data, this paper considers the uncertainty of the model for data collection in an unsupervised RL setting.
Context: The representation for occupancy measure used is the FB-representation (Touati & Ollivier, 2021) which encodes all optimal policies. We use an ensemble method approximation to the posterior distribution. Crucially, because of non-uniqueness, the FB representation does not allow simple modeling of the posterior uncertainty over FB via ensemble disagreement – there is a necessity of having a single B representation in order to have an informative notion of uncertainty. Furthermore, the F-uncertainty is projected to the more practical uncertainty over Q-functions for particular latent policy conditioning z.
2. We introduce an efficient algorithm for exploration tailored to forward-backward (FB) representations which can be seen as a variant of *uncertainty sampling* (Lewis & Gale, 1994).
Context: The algorithm relies on sampling a posterior-mean greedy policy π_z which has highest uncertainty in the predictive posterior distribution for a particular state s and executing it in the environment. This exploration strategy, while simple and not considering correlation in uncertainty reduction across all policies $\pi_z, z \in \mathcal{Z}$, is a surprisingly efficient method for exploration in FB representations.
3. Experimental validation of proposed exploration on several continuous control environments from the DeepMind Control suite (Tassa et al., 2018) in the online learning setting, where we evaluate zero-shot performance on different reward functions within several environments.
Context: There is no notion of exploration in the unsupervised RL setting, therefore there is no need to balance the exploration-exploitation trade-off when collecting data. This setup is fundamentally different than single-task online learning, where typically we balance an intrinsic exploration signal or noise with the extrinsic task reward.

Epistemically-guided forward-backward exploration

Anonymous authors
Paper under double-blind review

Abstract

Zero-shot reinforcement learning is necessary for extracting optimal policies in absence of concrete rewards for fast adaptation to future problem settings. Forward-backward representations (FB) have emerged as a promising method for learning optimal policies in absence of rewards via a factorization of the policy occupancy measure. However, up until now, FB and many similar zero-shot reinforcement learning algorithms have been decoupled from the exploration problem, generally relying on other exploration algorithms for data collection. We argue that FB representations should fundamentally be used for exploration in order to learn more efficiently. With this goal in mind, we design exploration policies that arise naturally from the FB representation that minimize the posterior variance of the FB representation, hence minimizing its epistemic uncertainty. We empirically demonstrate that such principled exploration strategies improve sample complexity of the FB algorithm considerably in comparison to other exploration methods.

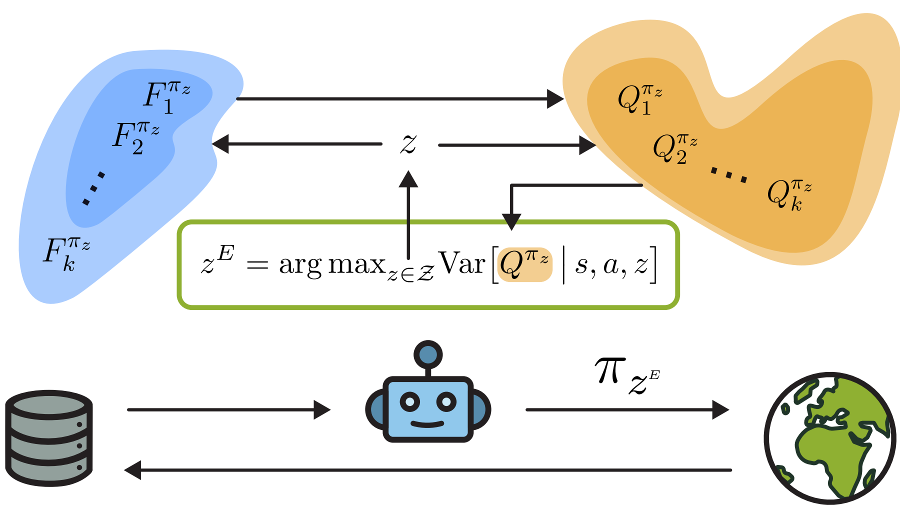

Figure 1: We condition an exploration policy on a reward embedding $z \in \mathcal{Z}$ maximizing the predictive variance of Q^{π_z}, and execute it for collecting data during learning. At inference time, we compute the reward embedding z based on reward evaluation of the dataset, aligned with Touati & Ollivier (2021).

1 Introduction

Reinforcement learning provides a framework to obtain optimal or near-optimal policies from sub-optimal data given a reward function. However, we cannot possibly enumerate all rewards which are of interest to solve in the future, and hence most RL approaches rely on fixed rewards for training, limiting the generalizability of the learnt policies to new tasks. Zero-shot RL aims to close this gap, by learning optimal policies for all possible reward functions. In this way, an agent may, with a minimal amount of extra computation, infer an optimal policy for any reward function given at test time.

There are several zero-shot RL methods that have been proposed to solve this problem. The earliest instantiation of such methods is that of the successor representation (SR) in the tabular setting (Dayan, 1993), which has subsequently been extended to the continuous setting with function approximations (Barreto et al., 2017). The caveat of SR is the need to assume a linear dependence between the reward and a feature map, which needs to be handcrafted in advance by the user. This approach cannot easily tackle generic rewards or goal-oriented RL. In the goal-oriented setting, for example, it would require introducing one feature per possible goal state, requiring infinitely many features in continuous spaces. Several frameworks have been proposed to learn this feature map efficiently (Hansen et al., 2019; Liu & Abbeel, 2021; Wu et al., 2018). More recent work has proposed forward-backward (FB) representations (Touati & Ollivier, 2021), which aims to factorize the occupancy distribution of the policies into a forward representation (F) of the current state and backward representation (B) of a target state. While the linearity of SR's allows us to infer the optimal policy by solving linear regression onto the sampled rewards, FB infers an optimal policy by Monte Carlo estimation of an integral, which, given a well-learned factorization of the occupancy distribution yields the optimal policy representation z for any given reward function. A critical part in both FB and SR frameworks is that of learning an accurate occupancy distribution (or successor measure) for all policies, which requires observing significant amount of environment state transitions.

Up until now, this problem has been tackled by using exploration policies that are decoupled from the zero-shot algorithm (Touati & Ollivier, 2021; Touati et al., 2022), mostly involving exploration policies trained with an intrinsic exploration reward (Eysenbach et al., 2018; Burda et al., 2018; Lee et al., 2019; Liu & Abbeel, 2021; Pathak et al., 2017; 2019). Relevant to this work, Chen et al. (2017) proposed ensemble disagreement on the Q-value as an intrinsic reward for efficient exploration. Alternatively, ensemble disagreement has been utilized in dynamics models for guiding exploration (Pathak et al., 2017). In fact, subsequently, many works have successfully used this type of approach for exploration, connecting it with the notion of "epistemic uncertainty" (Vlastelica et al., 2021; Sukhija et al., 2023; Sancaktar et al., 2022). While these methods yield successful exploration in some settings, a major disadvantage is that the exploration bonus doesn't depend on the rewards, so the exploration may focus on irrelevant aspects of the environment unrelated to the task (Chen et al., 2017).

A key question of this work is *how should we best interact with the environment to learn all optimal policies in the unsupervised RL setting sample efficiently*? We aim to collect samples that are most informative about the occupancy measure of optimal policies encoded by a zero-shot RL algorithm, in other words, we want to minimize the uncertainty over the occupancy measures. To this end, for modeling the occupancy measures we utilize the learned FB factorization of occupancies (Touati & Ollivier, 2021) which also has a representation space of optimal policies. Inspired by Lakshminarayanan et al. (2017), we model the posterior predictive uncertainty over the F representation by utilizing an ensemble of F representations. Consequently, the disagreement of the ensemble is a measure of uncertainty over F. Because of the mechanics of the FB representations, this naturally translates to the predictive uncertainty over the value function $Q^{\pi_z}(s, a)$ for particular policy π_z parametrized by reward embedding z, which is a more useful notion of uncertainty. Motivated by insights from *Bayesian experimental design*, we introduce an exploration algorithm that samples policies that are greedy w.r.t. to the mean of the Q^{π_z}-posterior, but have highest uncertainty. This can be seen as a variant of *uncertainty sampling* (Lewis & Gale, 1994). Our empirical evaluation indicates that utilizing this notion of uncertainty significantly improves the sample complexity of FB with a suprisingly simple exploration algorithm.

In summary, in this work we provide an epistemic-uncertainty-guided method for efficiently learning forward-backward representations that (i) exhibits zero-shot generalization in unsupervised RL, (ii) leads to sample efficiency gains compared to other exploration alternatives and (iii) compares favorably to current FB methods when evaluated on several benchmarks.

2 Related Work

Unsupervised Reinforcement Learning. Zero-shot (unsupervised) reinforcement learning frameworks can be traced back to the concept of a successor representation (Dayan, 1993), which relies on inferring the discounted occupancy measure of all policies. A direct extension of this are successor features (Barreto et al., 2017), where a feature map is a assumed which linearizes the reward w.r.t. a representation z, the main caveat being that the map needs to be a-priori specified. Consequently, many extensions exist to learn the feature map (Hansen et al., 2019; Laskin et al., 2022). Orthogonally, several works attempt to infer diverse skills in an online (Eysenbach et al., 2018) or offline fashion, mostly optimizing for a mutual-information objective. In contrast to the former, forward-backward representations assume a factorization of the occupancy measure, where z encodes an optimal value function for a specific reward. These can be traced back to Blier et al. (2021), and subsequent works have shown their effectiveness in deep RL benchmarks (Touati & Ollivier, 2021; Touati et al., 2022; Pirotta et al., 2024; Tirinzoni et al., 2025), also dealing with the offline estimation problem of the FB (Jeen et al., 2023). In contrast to successor features, there has been no proper analysis of exploration for learning FB representations more efficiently. Our work aims to fill this gap.

Exploration in Reinforcement Learning. Lee et al. (2019) attempt to solve the exploration problem by inferring the state marginal distribution of the policy and trying to match it to a user-defined target distribution. Osband et al. (2016) propose ensembles of Q values for exploration by uniformly sampling a Q function and subsequently following a policy associated with it for exploration. Several works have extended the classic upper confidence bound (UCB) exploration strategy to deep RL via ensemble methods (Chen et al., 2017; Lee et al., 2021), with Lee et al. (2021) additionally proposing to account for the error in Q-targets by down-weighting based on ensemble disagreement. Sukhija et al. (2024) utilize an ϵ-greedy policy with picking a Boltzmann policy with a mutual-information term for the dynamics. Metelli et al. (2019) propagate uncertainty over Q-values by constructing a TD update by Wasserstein barycenters V; they propose several variants for inferring a policy (mean estimation, particle sampling). Our work fits into the realm of ensemble-based exploration techniques, however in the context of zero-shot RL.

Deep Bayesian Inference. The problem of exploration is closely related to active learning (Chaloner & Verdinelli, 1995; Settles, 2009), also known as experimental design in the statistics literature. Active learning methods that yield strong theoretical generally query data points based on information-theoretic criteria (Krause et al., 2008; Settles, 2009; Hanneke, 2014). These methods have recently generalized to deep learning. Since exact Bayesian inference is computationally intractable for neural networks, a variety of approximations have been developed (Mackay, 1992; Neal, 2012). Gal et al. (2017); Chen et al. (2017) propose more computationally efficient methods than Bayesian neural networks, such as Monte Carlo dropout as an approximation of the posterior of the model parameters (Gal et al., 2017) or closer to our work, ensemble of neural networks (Osband et al., 2016; Chen et al., 2017; Lakshminarayanan et al., 2017) for predictive uncertainty quantification. Several recent works further leverage such uncertainty estimates for active fine-tuning of vision or action models (Hübotter et al., 2024; Bagatella et al., 2024).

3 Background

In this paper we will utilize the standard notion of a reward-free Markov Decision Process which is defined by a tuple $\mathcal{M} = (\mathcal{S}, \rho_0, \mathcal{A}, \mathcal{P}, \gamma)$, with state space $\mathcal{S}$, initial state distribution ρ_0, action space $\mathcal{A}$, transition kernel $\mathcal{P}$ and discount factor γ.

For the MDP $\mathcal{M}$, a policy $\pi : \mathcal{S} \to \mathcal{A}$ induces the successor measure M^π (Blier et al., 2021) for any initial state-action pair (s_0, a_0):

$$M^\pi(s_0, a_0, X) := \sum_{t \geq 0} \gamma^t P((s_{t+1}, a_{t+1}) \in X \mid s_0, a_0, \pi) \quad \forall X \subset \mathcal{S} \times \mathcal{A}. \tag{1}$$

Given M^{π} and a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we may write the value function of the policy π for reward r simply as $Q_r^{\pi}(s,a) = \sum_{s',a'} M^{\pi}(s,a,s',a')r(s',a')$.

Given a representation space $Z = \mathbb{R}^d$ and a family of policies $(\pi_z)_{z\in\mathcal{Z}}$ parameterized by z, the FB representation looks for representations $F : \mathcal{S} \times \mathcal{A} \times Z \to Z$ and $B : \mathcal{S} \times \mathcal{A} \to Z$, such that the successor measure M^{π_z} in (1) factorizes as:

$$M^{\pi_z}(s_0,a_0,s,a) \approx \langle F(s_0,a_0,z), B(s,a)\rangle, \quad \pi_z(s) = \arg\max_{a\in\mathcal{A}} \langle F(s,a,z), z\rangle, \tag{2}$$

where z is the latent representation of the policy π_z of dimension d. Assuming (2) holds, then for any reward function r, the policy π_{z_r} where $z_r := \sum_{s,a\in\mathcal{S}\times\mathcal{A}} r(s,a)B(s,a)$ is optimal for r with optimal Q-function $Q_r^*(s,a) = \langle F(s,a,z_r), z_r\rangle$, i.e. the policy is guaranteed to be optimal for any reward function (Touati & Ollivier, 2021)[Theorem 2].

In practice, we choose a parametric model F_θ and B_ϕ for the F and B representations, as approximations to the true successor measure factorization. There are several off-the-shelf algorithms for learning M^{π_z} (Blier et al., 2021; Eysenbach et al., 2021), however the quality of the representation is tightly coupled with 1) the chosen factorization dimension d and 2) the approximation error which can be result of model miss-specification or lack of data. In this work we attempt to tackle the second issue, which can be handled via quantifying posterior uncertainty and utilizing it to guide exploration, as has been done in previous works (Osband et al., 2013; 2016; Chen et al., 2017).

3.1 Bayesian Reinforcement Learning

In the setting of Bayesian inference, ideally one would be able to formulate a prior distribution over the parameters of the FB representation $\Theta = (\theta, \phi)$ and subsequently, given evidence in form of data at the i-th iteration, compute the posterior distribution via Bayes' rule $p(\Theta|\mathcal{D}_i) = \frac{p(\Theta)p(\mathcal{D}_i|\Theta)}{p(\mathcal{D}_i)}$. This is intractable for high-dimensional Θ, since it requires computing the marginal $p(\mathcal{D}_i) = \int_\Theta p(\mathcal{D}_i|\Theta)p(\Theta)d\Theta$. Hence, many works have utilized various approximations to posterior distributions over neural networks (Blundell et al., 2015; Osband et al., 2016; Chen et al., 2017).

Beyond proper quantification of uncertainty over Θ, which is typically taken to be as variance or entropy of $\Theta \sim p(\Theta|\mathcal{D}_i)$ for continuous Θ, the uncertainty of the predictions is of crucial interest in optimal data collection, which is a fundamental question in *active learning* and *optimal experimental design*. In the field of *bandits* and *reinforcement learning*, this is also familiar under the term *exploration*. For RL in particular, one might want to compute a posterior over the unknown reward function r and transition kernel $\mathcal{P}$ of the MDP (Osband et al., 2013) or parameters of the Q-value function – this is often approximated as an ensemble of neural networks in deep reinforcement learning (Osband et al., 2016; Chen et al., 2017). Subsequently, this posterior is utilized in formulating an exploration strategy by a policy, a popular choice being a Upper-Confidence Bound (UCB) strategy by setting the policy to be $\pi(s) := \arg\max_{a\in\mathcal{A}} \bar{Q}(s,a) + \alpha\sqrt{\mathrm{Var}[Q(s,a)|s,a]}$ (Chen et al., 2017), encouraging more uncertain actions. This strategy stems from the well-known UCB algorithm in the bandit literature (Auer et al., 2002; Auer, 2002). Further strategies exist that have been utilized in the literature, such as Thompson sampling where a particle is sampled from the posterior and subsequently exploited (Osband et al., 2013; Thompson, 1933).

Remark 3.1. All of these methods focus on the exploitation-exploration tradeoff, which is ill-defined in the context of unsupervised reinforcement learning. This problem is fundamentally a *pure exploration problem.*

4 Posterior uncertainty in forward-backward representations

In the unsupervised RL setting, accurately estimating the successor measure for all policies is of crucial interest. Given our prior distribution over the parameters of the FB representation $\Theta = (\theta, \phi)$, we are tasked with updating the posterior distribution over the parameters as new evidence is collected.

Building on prior work that has successfully leveraged ensembles to approximate the posterior distribution over Q^* (Osband et al., 2016; Chen et al., 2017), we consider a similar approach for the FB representation. Crucially, Chen et al. (2017) suggest decoupled Q networks trained with standard 1-step TD error in order to approximate a posterior distribution given data $\mathcal{D}$. The FB representation entails a factorization of M into F and B, therefore naturally we might be tempted to construct a posterior over F and B. This however can cause issues, especially when utilizing ensemble methods, since the representation is non-unique (details can be found in Blier et al. (2021)). This is easy to see if we view the F and B functions as matrices, assuming a rotation matrix R, we have that $M = F^\top R R^{-1} B = \tilde{F}^\top \tilde{B}$, i.e. $\tilde{F}$ and $\tilde{B}$ encode the same set of occupancy measures, however with the representation space rotated. We alleviate this problem by fixing B and modeling the posterior distribution over F alone.

Following Chen et al. (2017), we adopt a naive posterior update over F: for the k-th ensemble member in $[0, \dots, K-1]$, we minimize the empirical forward-backward loss over a batch of b sampled transitions $(s_i, a_i, s_{i+1})_{i=0}^{b-1}$, independently sampled future states $(s_i')_{i=0}^{b-1}$, and reward embeddings z_i,

$$\ell(\theta_k, \phi) = \frac{1}{2b^2} \sum_{0 \le i,j < b-1} \Big(\big\langle F_{\theta_k}(s_i, a_i, z_i), B_\phi(s_j') \big\rangle - \gamma \sum_{a \in \mathcal{A}} \pi_{z_i}(a \mid s_{i+1}) \big\langle F_{\theta_k^-}(s_{i+1}, a, z_i), B_{\phi-}(s_j') \big\rangle \Big)^2 \tag{3}$$

$$- \frac{1}{b} \sum_{0 \le i < b} \big\langle F_{\theta_k}(s_i, a_i, z_i), B_\phi(s_i) \big\rangle, \tag{4}$$

where θ_k^- and ϕ^- denote the target networks for $F_{\theta_k^-}$ and B_{ϕ^-}, respectively. In practice, an additional orthonormality regularization on B is added as per Touati & Ollivier (2021) to normalize the covariance of B (otherwise one could for example scale F up and B down since only $F^T B$ is fixed).

Equipped with a model to approximate the posterior distribution over forward representations, we are left with determining a strategy for collecting evidence to maximally reduce uncertainty of the posterior distribution, which is a challenging problem in deep learning. To design such an algorithm, we take inspiration from Bayesian experiment design (MacKay, 1992; Chaloner & Verdinelli, 1995). We shall adopt a well-known active learning heuristic – *uncertainty sampling* (Lewis & Gale, 1994), which queries data points with the highest predictive uncertainty, but still provably minimizes posterior uncertainty under a homoscedastic, independent Gaussian noise model.

Aligned with previous work that showed that disagreement in ensemble methods can be effectively used for quantifying predictive uncertainty (Lakshminarayanan et al., 2017), for a given query point $\boldsymbol{x} = (s, a, z)$, we model our distribution over F as a uniformly weighted mixture of $\{F_k\}_{k=1}^K$ of Gaussian distributions i.e.

$$p(F \mid \boldsymbol{x}; \theta, \mathcal{D}) \approx \frac{1}{K} \sum_{k}^{K} \mathcal{N}(F; \mu_{\theta_k}(\boldsymbol{x}), \Sigma_{\theta_k}(\boldsymbol{x})), \tag{5}$$

where $\boldsymbol{x} = (s, a, z)$ for ease of reading and $\mu_{\theta_k}, \Sigma_{\theta_k}$ are the predicted mean and covariance by ensemble member k.

In the limiting case of $\Sigma_{\theta_k} \to 0 \ \forall k$, this posterior distribution becomes a mixture of Dirac delta functions, with the corresponding covariance being

$$\mathrm{Cov}[F \mid \boldsymbol{x}; \theta, \mathcal{D}] = \frac{1}{K} \sum_{k}^{K} (F_k(\boldsymbol{x}) - \bar{F}(\boldsymbol{x})(F_k(\boldsymbol{x}) - \bar{F}(\boldsymbol{x}))^\top. \tag{6}$$

where $F_k := \mu_{\theta_k}$ and $\bar{F} := \frac{1}{K} \sum_k \mu_{\theta_k}$. While in previous work the variance of point estimates has been used in place of epistemic uncertainty exploration (Lakshminarayanan et al., 2017), here we have a matrix quantity, the covariance of the F-representations. One viable option is to measure

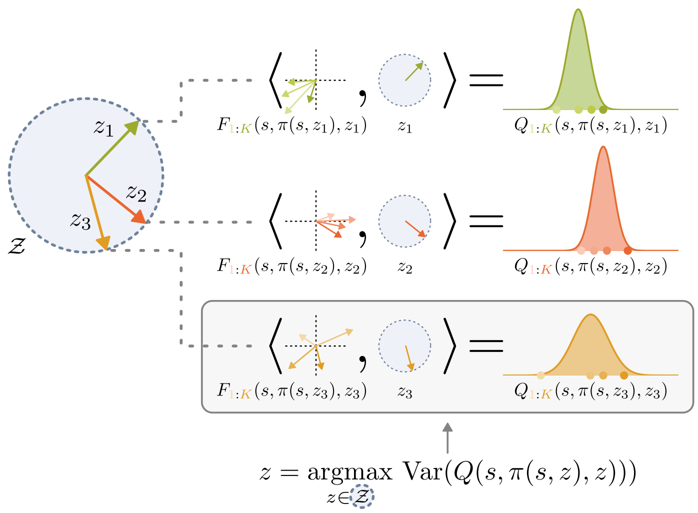

Figure 2: Epistemically guided FB exploration (FBEE). During exploration we uniformly sample reward embeddings from a hypersphere (left), and take samples over our posterior distribution F as represented by the K ensemble members $F_{1:K}$ ($K = 4$ in the figure) (middle-left). Then we project our F-posterior to a Q-posterior via $Q^{\pi_z} = \langle F(s, \pi_z(s), z), z \rangle$ (middle-right) and compute the Q-predictive uncertainty for all sampled z's (left) via ensemble disagreement. We finally explore with the reward embedding z^E that has maximum Q-predictive uncertainty.

the volume of Eq. (6), by computing $\operatorname{Det}(\operatorname{Cov}[F \mid s, a, z, \mathcal{D}])$, the trace or maximum eigenvalue. There is however an argument against using Eq. (6) to quantify uncertainty to guide data collection. The primary object of interest for us is Q^{π_z} for extracting greedy policies π_z that are optimal w.r.t. some reward. We may utilize the relationship $Q^{\pi_z} = \langle F(s, a, z), z \rangle$, to project our F-posterior to a Q-posterior, to arrive to the Q-predictive uncertainty for the query sample (s, a, z).

$$\operatorname{Var}[Q^{\pi_z}(s, a) \mid \mathcal{D}] = \frac{1}{K} \sum_{i=0}^{K} \langle F_k(s, a, z) - \bar{F}(s, a, z), z \rangle^2. \tag{7}$$

This corresponds to a Gaussian approximation to predictive posteriors on Q^{π_z}. In case of a Gaussian posterior, we have that its entropy is monotonic w.r.t. the variance, i.e. for the case of Eq. (7) the predictive variance can be seen as a measure of information for the input query $\boldsymbol{x}$. It is worth noting that because of the non-trivial dependence between F^{π_z} and z, it is unclear how the predictive uncertainty of F^{π_z} will affect the uncertainty on Q^{π_z}, which might be lower or higher after the projection with z. This has the important consequence that minimizing the uncertainty on one versus the other may lead to significantly different algorithmic behaviors.

5 Epistemic exploration for FB representations

While we have a notion of posterior uncertainty phrased as the variance of the empirical predictive Q-posterior distribution in Eq. (7), it is still unclear how one can best formulate an exploration policy for collecting data to improve the FB representation. To design such an algorithm, we take inspiration from *Bayesian experimental design* (Chaloner & Verdinelli, 1995; MacKay, 1992). A natural objective for active exploration is maximizing mutual information between F and observed transition data $\mathcal{D}_i$, which quantifies the reduction in entropy of F conditioned on the observations. In certain settings the predictive posterior variance is shown to be proportional to information gain (MacKay, 1992), hence it is a reasonable "guide" for exploration.

In general, we are seeking to define an exploration policy π^E which is going to extend $\mathcal{D}_{1:n-1}$ to $\mathcal{D}_{1:n}$ such that the collected data $\mathcal{D}_n$ provides the most amount of information about F^{π_z} for all

$\{\pi_z\}_{z \in \mathcal{Z}}$. To this end, we take the approach of selecting a π_z given s, a that we are most uncertain about in terms of predictive variance, which may be seen as a variant of *uncertainty sampling*,

$$\pi^E = \arg\max_{\pi_z} \mathrm{Var}\big[\mathbb{E}_{a \sim \pi_z(s)}[Q^{\pi_z}(s,a)] \,|\, s, a, z\big] \quad \text{s.t.} \quad z \in \mathcal{Z}, \tag{8}$$

where we make use of the posterior predictive variance in Eq. (7), which captures the uncertainty of the future return of π_z. Although the exploration policy in $Eq.$ (8) is a greedy policy $w.r.t.$ $\langle \bar{F}^{\pi_z}(s,a), z \rangle$, we can still expect that executing π_z reduces the uncertainty over Q^{π_z}. Moreover, the uncertainty of different Q-posteriors depends on z in a non-trivial way via F^{π_z}, hence a reduction in uncertainty in Q^{π_z} is likely to reduce uncertainty across multiple $z \in \mathcal{Z}$. This is loosely motivated by the "information never hurts" principle, which is a consequence of monotonicity of entropy $\mathcal{H}[X \,|\, Y] \leq \mathcal{H}[X]$ in light of new evidence Y. We provide pseudocode of our algorithm in Algorithm 1 and a visual schematic in Fig. 2.

Remark 5.1. While our definition of the policy in Eq. (8) is purely explorational, in the absence of a set of evaluation reward functions it is also reasonable, since there is no direct notion of "exploitation" in purely unsupervised RL.

Algorithm 1 FB Uncertainty Sampling (FBEE)

```
Input: K-ensemble of F_{θ_k} and F_{θ_k^-}, B_φ and B_{φ^-}.
while not converged do
    Pick π^E according to Eq. (8).
    Collect data D_i = Rollout(π^E)
    Add data to buffer D_{1:n} = D_{1:n-1} ∪ D_n.
    Fit {F_{θ_i}}_i^K, B_φ and policies π_z with D_{0:n}.
```

6 Experiments

Our experimental section is designed to provide an empirical answer to the following two questions: **(i)** Does FBEE exhibit similar zero-shot generalization in online unsupervised RL compared to the original FB method?, **(ii)** Does the epistemically guided exploration in FBEE lead to sample efficiency gains compared to other exploration alternatives?, **(iii)** What is the effect of exploring over reward embeddings z's compared to over actions? and **(iv)** How often should we update the chosen reward embedding z^E during an exploration episode?

Environments: We benchmark FBEE on 15 downstream tasks across 5 domains in the DeepMind Control Suite (DMC) (Tassa et al., 2018), see Fig. 3). Details on the domains and tasks can be found in Appendix A.1.

Baselines: We compare FBEE with several baselines for online unsupervised RL. The first baseline is FB (Touati & Ollivier, 2021), the original FB algorithm that conducts uninformed exploration by uniformly sampling random reward emedding z's. We also compare against a naive RANDOM policy that performs random exploration over the action space. We additionally compare against FB-RND (Touati et al., 2022), which decouples the exploration method from the learning of the FB representation by leveraging a pure exploration method, namely RND (Burda et al., 2018). We note that the exploration bonuses distilled by RND remain independent of any estimate of FB representations. Notably, in this setting, we can leverage precollected exploration datasets from the Unsupervised Reinforcement Learning Benchmark (Laskin et al., 2021), and hence the FB representation is trained fully offline. We also implement two variants of our algorithm: FBEE-POLICY explicitly learns an exploration policy $\pi_\theta : \mathcal{S} \to \mathcal{Z}$ by maximizing the objective in Eq. (8) through gradient descent, while FBEE-SAMPLING approximates the maximizer via zero-order optimization. Due to lack of space, we reserve results of FBEE-POLICY to the Appendix. Finally, we implement an ablation of our method FBEE-EPISODE to study the impact of how long to optimize for the most uncertain reward embedding z^E. With FBEE-EPISODE, we only compute z^E (via

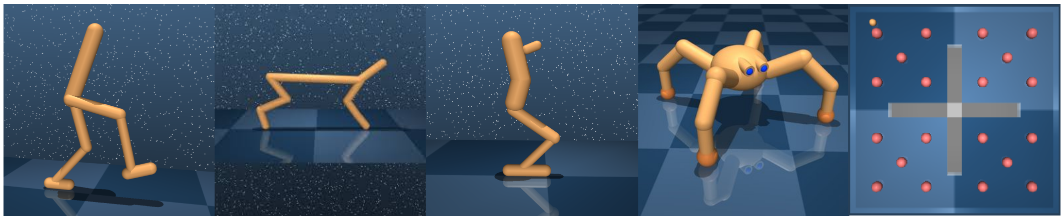

Figure 3: Environments used in our experiments. (Left to right): `Walker, Cheetah, Hopper, Quadruped, Point-mass maze`. In the `Point-mass maze` domain, we show an example of initial state (yellow point), which always starts in the top-left room, and the 20 test goals (red circles).

Eq. (8)) at the beginning of each training episode, whereas the default implementation optimizes for it every 100 interaction steps (10 times more frequently). We implement this ablation for both FB (FB-EPISODE) and our method FBEE -EPISODE.

Results We evaluate zero-shot performance of FBEE on 15 tasks across 5 domains in DMC every 100k exploration steps. At evaluation time, given a task reward function $r(s,a)$, the agents acts with the reward representation $z_R = \mathbb{E}_{(s,a)\sim\mathcal{D}}[r(s,a)B(s,a)]$ for 1000 environment steps. The reward function is bounded to [0,1], hence maximum return per task is of 1000. In practice, we compute the expectation by taking the average over relabeled samples from the current replay buffer. Zero-shot scores curves averaged across tasks for every domain are shown in Fig. 4. For zero-shot scores per each task, see Fig. 6.

As shown in Fig. 4, FBEE asymptotically achieves similar or better performance than the original FB method, hence answering our question i). Most importantly, we observe that in all the environments, FBEE exhibits significant sample efficiency gains compared to FB across all domain, empirically showcasing that FBEE achieves the most important goals of our work, which is that of driving efficient exploration, hence answering our question ii). We notice that in easier tasks such as `cheetah` the performance gap between FB and FBEE is reduced, showcasing that random exploration over reward embeddings is still a fairly good strategy. In these lines, we would like to notice that we find somewhat remarkable the general sample efficiency showcased by the naive FB exploration. We reserve to future work a deeper analysis on this finding. This naturally flows to answering our question iii) by which we empirically show that randomly exploring over reward embeddings leads to much sample efficiency than doing it at the action level. This can be observed by the low performance of the RANDOM among all domains.

Finally, we are left with question iv), evaluating the impact on the z^E frequency update during an exploration episode. We observe that for all methods the higher the frequency the better, although differences are only highly noticeable for the `hopper, maze` tasks. For FBEE, this could be caused by several reasons. Our posterior update over F differs from theoretically sound approaches (e.g., Metelli et al. (2019) suggested propagating uncertainty through a TD update involving Wasserstein barycenters), and can potentially incur in myopic behavior. In practice, however, practical instantiations of similar algorithms (Metelli et al., 2019) resort to the same approach as ours. Our hypothesis is that, as we update each of the ensemble members against its own target network, each member provides a temporally extended (and consistent) estimate of the value uncertainty via TD estimates, hence propagating uncertainty and alleviating myopic behavior. This was also observed by Osband et al. (2016). A second hypothesis would be that of our exploration strategy π^E not guaranteed of picking the z^E to maximize uncertainty in Q *over all z's*, but instead picking the z that greedily maximizes it. However, we empirically show that our method leads to significant sample efficiency gains compared to other exploration alternatives and we leave this analysis for future work.

F-uncertainty versus Q-uncertainty. As we have argued in Section 4, F^{π_z}-uncertainty and Q^{π_z}-uncertainty may lead to different exploration behaviors. For purpose of demonstration, we analyze

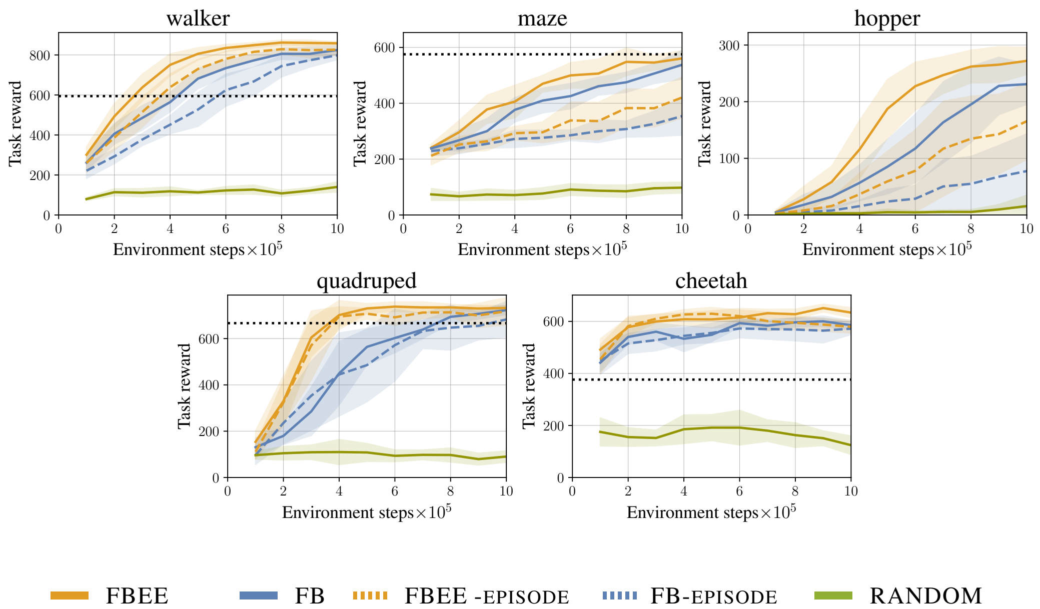

Figure 4: Zero-shot scores averaged over different downstream task as number of environment samples increases. Metrics are averaged over 30 evaluation episodes and 10 independent random seeds. Shaded area is 1-standard deviation. Topline is maximum score of FB-RND (offline method with precollected data). Note: RND buffer for the `Hopper` task is not available in URLB benchmark (Laskin et al., 2021).

the average uncertainty across state-action pairs for the `Maze` experiment. In Fig. 5 we observe how the uncertainty of F^{π_z} relates to the uncertainty of Q^{π_z} for different z samples in a particular FB checkpoint from training – although there is a slight positive correlation between the determinant of $\mathrm{CoVar}[F^{\pi_z} \mid s, a, z]$ and $\mathrm{Var}[Q^{\pi_z} \mid s, a, z]]$ in expectation. With a quite low R^2 score of 0.18, this signifies that there is no strong correlation signal. In fact, we observe instances where we have high Q^{π_z} uncertainty and low F^{π_z} uncertainty.

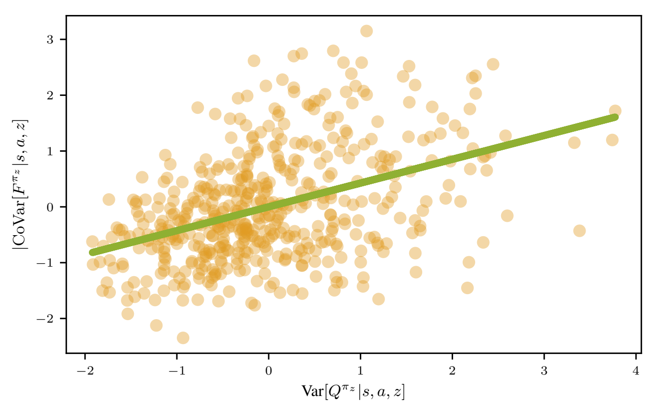

Figure 5: Regression scatter plot of the determinant of $\mathrm{CoVar}[F^{\pi_z} \mid s, a, z]$ and $\mathrm{Var}[Q^{\pi_z} \mid s, a, z]$ for a FB checkpoint in `Maze` experiment.

7 Conclusion

In this work we have proposed an epistemically-guided exploration framework for sample efficient learning of FB representations. We have done so by maintaining an ensemble approximation of the predictive posterior distribution over Q^{π_z}, and subsequently picking the least certain π_z in terms of variance of Q^{π_z}, which can be seen as an instance of *uncertainty sampling*. This is a *pure exploration* algorithm since the exploration-exploitation trade-off is non-existant in the zero-shot RL setting. In experiments, this is a suprisingly effective exploration strategy which outperforms other exploration algorithms on the DMC benchmark.

While this is an initial attempt at phrasing an exploration algorithm for zero-shot RL, many extensions are henceforth possible, such as extending this approach to further uncertainty-based exploration algorithms such as UCB or Thompson sampling. An efficient exploration algorithm necessarily needs to take into account how information is correlated across different $z \in \mathcal{Z}$ in order to maximally reduce it with least amount of data. Finally, a full Bayesian treatment of FB representations is still an open question, especially with the assumption of a full posterior over F and B, which is a difficult object because of the non-uniqueness of FB.

References

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

Marco Bagatella, Jonas Hübotter, Georg Martius, and Andreas Krause. Active fine-tuning of generalist policies. *arXiv preprint arXiv:2410.05026*, 2024.

André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Léonard Blier, Corentin Tallec, and Yann Ollivier. Learning successor states and goal-dependent values: A mathematical viewpoint. *CoRR*, abs/2101.07123, 2021. URL https://arxiv.org/abs/2101.07123.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

Kathryn Chaloner and Isabella Verdinelli. Bayesian Experimental Design: A Review. *Statistical Science*, 10(3):273 – 304, 1995. DOI: 10.1214/ss/1177009939. URL https://doi.org/10.1214/ss/1177009939.

Richard Y Chen, Szymon Sidor, Pieter Abbeel, and John Schulman. Ucb exploration via q-ensembles. *arXiv preprint arXiv:1706.01502*, 2017.

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

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

Benjamin Eysenbach, Ruslan Salakhutdinov, and Sergey Levine. C-learning: Learning to achieve goals via recursive classification. In *International Conference on Learning Representations*, 2021.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR, 2017.

Steve Hanneke. 2014. DOI: 10.1561/2200000037.

Steven Hansen, Will Dabney, Andre Barreto, Tom Van de Wiele, David Warde-Farley, and Volodymyr Mnih. Fast task inference with variational intrinsic successor features. *arXiv preprint arXiv:1906.05030*, 2019.

Jonas Hübotter, Bhavya Sukhija, Lenart Treven, Yarden As, and Andreas Krause. Transductive active learning: Theory and applications. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Scott Jeen, Tom Bewley, and Jonathan M Cullen. Zero-shot reinforcement learning from low quality data. *arXiv preprint arXiv:2309.15178*, 2023.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(8):235–284, 2008. URL http://jmlr.org/papers/v9/krause08a.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Michael Laskin, Denis Yarats, Hao Liu, Kimin Lee, Albert Zhan, Kevin Lu, Catherine Cang, Lerrel Pinto, and Pieter Abbeel. Urlb: Unsupervised reinforcement learning benchmark. *arXiv preprint arXiv:2110.15191*, 2021.

Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. Cic: Contrastive intrinsic control for unsupervised skill discovery. *arXiv preprint arXiv:2202.00161*, 2022.

Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. In *International Conference on Machine Learning*, pp. 6131–6141. PMLR, 2021.

Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In Bruce W. Croft and C. J. van Rijsbergen (eds.), *SIGIR '94*, pp. 3–12, London, 1994. Springer London. ISBN 978-1-4471-2099-5.

Hao Liu and Pieter Abbeel. Aps: Active pretraining with successor features. In *International Conference on Machine Learning*, pp. 6736–6747. PMLR, 2021.

David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 07 1992. ISSN 0899-7667. DOI: 10.1162/neco.1992.4.4.590. URL https://doi.org/10.1162/neco.1992.4.4.590.

David John Cameron Mackay. *Bayesian methods for adaptive models*. California Institute of Technology, 1992.

Alberto Maria Metelli, Amarildo Likmeta, and Marcello Restelli. Propagating uncertainty in reinforcement learning via wasserstein barycenters. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/f83630579d055dc5843ae693e7cdafe0-Paper.pdf.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, pp. 2778–2787. PMLR, 2017.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International Conference on Machine Learning*, pp. 5062–5071. PMLR, 2019.

Matteo Pirotta, Andrea Tirinzoni, Ahmed Touati, Alessandro Lazaric, and Yann Ollivier. Fast imitation via behavior foundation models. In *The Twelfth International Conference on Learning Representations*, 2024.

Cansu Sancaktar, Sebastian Blaes, and Georg Martius. Curious exploration via structured world models yields zero-shot object manipulation. *Advances in Neural Information Processing Systems*, 35:24170–24183, 2022.

Burr Settles. Active learning literature survey, 2009.

Bhavya Sukhija, Lenart Treven, Cansu Sancaktar, Sebastian Blaes, Stelian Coros, and Andreas Krause. Optimistic Active Exploration of Dynamical Systems, October 2023. URL http://arxiv.org/abs/2306.12371. arXiv:2306.12371 [cs, eess].

Bhavya Sukhija, Stelian Coros, Andreas Krause, Pieter Abbeel, and Carmelo Sferrazza. Maxinforl: Boosting exploration in reinforcement learning through information gain maximization. *arXiv preprint arXiv:2412.12098*, 2024.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Andrea Tirinzoni, Ahmed Touati, Jesse Farebrother, Mateusz Guzek, Anssi Kanervisto, Yingchen Xu, Alessandro Lazaric, and Matteo Pirotta. Zero-shot whole-body humanoid control via behavioral foundation models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. *Advances in Neural Information Processing Systems*, 34:13–23, 2021.

Ahmed Touati, Jérémy Rapin, and Yann Ollivier. Does zero-shot reinforcement learning exist? *arXiv preprint arXiv:2209.14935*, 2022.

Marin Vlastelica, Sebastian Blaes, Cristina Pinneri, and Georg Martius. Risk-averse zero-order trajectory optimization. In *5th Annual Conference on Robot Learning*, 2021.

Yifan Wu, George Tucker, and Ofir Nachum. The laplacian in rl: Learning representations with efficient approximations. *arXiv preprint arXiv:1810.04586*, 2018.

A Appendix

A.1 Environments

All the environments are based on the *DeepMind Control Suite*(Tassa et al., 2018) and some adapted by (Touati et al., 2022).

- **Point-mass Maze**: a 2-dimensional continuous maze with four rooms. The states are 4-dimensional vectors encoding for positions and velocities of the point mass, and the actions are 2-dimensional vectors. Importantly, the initial position of the point-mass is always sampled from a uniform distribution over the spatial domain of the top-left room only. At test, we evaluate performance of agents on 20 goal-reaching tasks (5 goals in each room described by their (x,y) coordinates. See figure . This task is set as a goal-reaching task and hence we compute z_R at evaluation time by: $z_R = B(s)$.
- **Cheetah**: A 17 state-dimensional running planar biped consisting of positions and velocities of robot joints. Actions are 6-dimensional. We evaluate on 4 tasks `walk, run, walk backward, run backward`. Rewards are linearly proportional to the achieved velocity up to the desired task velocity.
- **Walker**: A 24 state-dimensional planar walker consisting of positions and velocities of robot joints. Actions are 6-dimensional. We evaluate on 4 tasks: `stand, run, flip`. In the `stand` task reward is a combination of terms encouraging an upright torso and some minimal torso height. The `walk` and `run` task rewards include a component linearly proportional to the achieved velocity up to the desired task velocity. `flip` includes a component encouraging angular momentum.
- **Hopper**: A 15-dimensional planar one-legged hopper. Actions are 4 dimensional. We evaluate on 5 tasks: `stand, hop, flip`. In the `stand` the reward encourages a minimal torso height. In the `hop, hop backward` tasks the rewards have an additional term that is linearly proportional to the achieved velocity up to the desired task velocity. In the `flip, flip backward` includes a component encouraging angular momentum.
- **Quadruped** a four-leg spider navigating in 3D space. States and actions are 78 and 12 dimensional respectively. We evaluate on 4 tasks: `stand, walk, run jump`. `stand` reward encourages an upright torso, `walk` and `run` have an additional term that is linearly proportional to the achieved velocity up to the desired task velocity. `jump` includes a term encouraging some minimal height of the center of mass.

A.2 Prior information on rewards

When dealing with high dimensionality environments, learning future probabilities for all states is very difficult and generally requires large d to accommodate for all possible rewards. In general, we are often interested in rewards that depend not on the full state but on a subset of it. If this is known in advance, the representation B can be trained on that part of the state only, with same theoretical guarantees (Appendix, Theorem 4 (Touati & Ollivier, 2021)). Hence, when knowing that the reward will be only a function of a subset of the state and action spaces G, we can leverage an environment-dependant feature map $\varphi : S \times A \to G$, and learn $B(g)$ instead of $B(s, a)$, where $g = \varphi(s, a)$. Importantly, rewards can be arbitrary functions of g. This was also suggested in (Touati & Ollivier, 2021). In what follows, we list the feature maps that were used for the different environments.

- **Point-mass Maze**: $\phi(s, a) = [x, y]$.
- **Chetah**: $\phi(s, a)$ = $[v_x, L_y\ ]$ where v_x is the velocity along the x-axis in the robot frame and L_x is the angular momentum about x-axis.
- **Walker**: $\phi(s, a)$ = $[v_x, torso_z, torso_{z_w}\ ]$ where v_x is the horizontal velocity of the center of mass, $torso_z$ is the height of the torso and $torso_{z_w}$ is the projection from the z-axis of the torso to the z-axis of the world frame.

- **Hopper**: $\phi(s,a) = [v_x, torso_{z,foot}]$ where v_x is the horizontal velocity of the center of mass and $torso_{z,foot}$ is the height of the torso with respect to the foot.
- **Humanoid**: $\phi(s,a) = [torso_z, v, torso_{z_w}]$ where $torso_z$ is the height of the torso, v is the velocity of the center of mass in the local frame, and $torso_{z_w}$ is the projection from the z-axis of the torso to the z-axis of the world frame.
- **Quadruped** $\phi(s,a) = [v, torso_{z_w}]$ where v is the torso velocity vector in the local frame and $torso_{z_w}$ is the projection from the z-axis of the torso to the z-axis of the world frame.

B Hyperparameters

In Table 1 we summarize the hyperparameters used in our experiments. For a fair comparison, unless specified, we used the same parameters among all methods. Most of the parameters were adapted from (Touati et al., 2022).

Table 1: Hyperparameters.

Hyperparameter	**Value**
Optimizer	Adam (default hyperparameters)
Learning rate	10^{-4}
Batch size	256
Ratio gradient step/environment step	0.5
1 Z-dimension	50 (100 for maze)
Discount factor γ	0.98 (0.99 for maze)
Mix ratio for z sampling	0.3
Momentum coefficient for target networks update	0.99
Number of reward labels for task inference	10^4
Number of ensemble members	5
Frequency of z updates (training)	0.01

C Additional experiments

C.1 Zero-shot scores per task

We evaluate zero-shot performance of FBEE on 15 tasks across 5 domains in DMC every 100k exploration steps. At evaluation time, given a task reward function $r(s,a)$, the agents acts with the reward representation $z_R = \mathbb{E}[r(s,a)B(s,a)]$ for 1000 environment steps. The reward function is bounded to [0,1], hence maximum return per task is of 1000. In practice, we compute the expectation by taking the average over relabeled samples from the current replay buffer. Zero-shot scores across domains for all tasks is shown in Fig. 6. In this section we additionally show another ablation of our method, namely FBEE-POLICY which explicitly learns an exploration policy $\pi_\theta : \mathcal{S} \to \mathcal{Z}$ by maximizing the objective in Eq. (8) through gradient descent. In general we observe that it performs in par with FBEE-SAMPLING, and we attribute the mismatches in performance to not extensive hyperparameter finetuning.

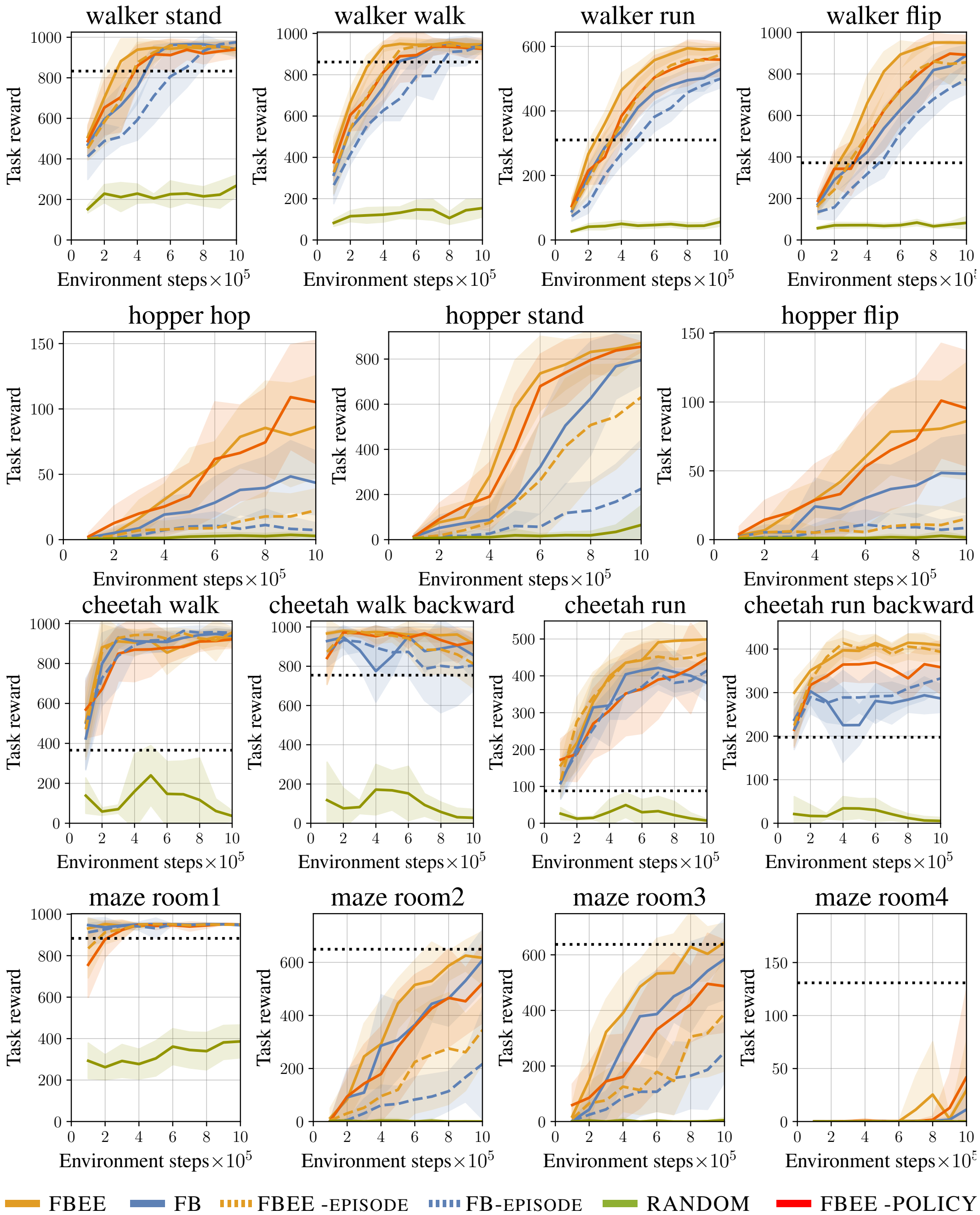

Figure 6: Zero-shot scores for different downstream task as number of environment samples increases. Metrics are averaged over 30 evaluation episodes and 10 independent random seeds. Shaded area is 1-standard deviation. Topline is maximum score of FB-RND (offline method with precollected data). Note: RND buffer for the `Hopper` task is not available in URLB benchmark (Laskin et al., 2021).