SAMPLE EFFICIENT MYOPIC EXPLORATION THROUGH MULTITASK REINFORCEMENT LEARNING WITH DIVERSE TASKS

Ziping Xu[*,1], **Zifan Xu**[2], **Runxuan Jiang**[3], **Peter Stone**[2,4] **& Ambuj Tewari**[3]
[1]Harvard University, [2]The University of Texas at Austin, [3]University of Michigan, [4]Sony AI
[*]This work was done while Z.X. was a Ph.D. student at the University of Michigan.

ABSTRACT

Multitask Reinforcement Learning (MTRL) approaches have gained increasing attention for its wide applications in many important Reinforcement Learning (RL) tasks. However, while recent advancements in MTRL theory have focused on the improved statistical efficiency by assuming a shared structure across tasks, exploration–a crucial aspect of RL–has been largely overlooked. This paper addresses this gap by showing that when an agent is trained on a sufficiently *diverse* set of tasks, a generic policy-sharing algorithm with myopic exploration design like ϵ-greedy that are inefficient in general can be sample-efficient for MTRL. To the best of our knowledge, this is the first theoretical demonstration of the "exploration benefits" of MTRL. It may also shed light on the enigmatic success of the wide applications of myopic exploration in practice. To validate the role of diversity, we conduct experiments on synthetic robotic control environments, where the diverse task set aligns with the task selection by automatic curriculum learning, which is empirically shown to improve sample-efficiency.

1 INTRODUCTION

Reinforcement Learning often involves solving multitask problems. For instance, robotic control agents are trained to simultaneously solve multiple goals in multi-goal environments (Andreas et al., 2017; Andrychowicz et al., 2017). In mobile health applications, RL is employed to personalize sequences of treatments, treating each patient as a distinct task (Yom-Tov et al., 2017; Forman et al., 2019; Liao et al., 2020; Ghosh et al., 2023). Many algorithms (Andreas et al., 2017; Andrychowicz et al., 2017; Hessel et al., 2019; Yang et al., 2020) have been designed to jointly learn from multiple tasks. These show significant improvement over those that learn each task individually. To provide explanations for such improvement, recent advancements in Multitask Reinforcement Learning (MTRL) theory study the improved statistical efficiency in estimating unknown parameters by assuming a shared structure across tasks (Brunskill & Li, 2013; Calandriello et al., 2014; Uehara et al., 2021; Xu et al., 2021; Zhang & Wang, 2021; Lu et al., 2021; Agarwal et al., 2022; Cheng et al., 2022; Yang et al., 2022a). Similar setups originate from Multitask Supervised Learning, where it has been shown that learning from multiple tasks reduces the generalization error by a factor of $1/\sqrt{N}$ compared to single-task learning with N being the total number of tasks (Maurer et al., 2016; Du et al., 2020). Nevertheless, these studies overlook an essential aspect of RL, namely exploration.

To understand how learning from multiple tasks, as opposed to single-task learning, could potentially benefit exploration design, we consider a generic MTRL scenario, where an algorithm interacts with a task set $\mathcal{M}$ in rounds T. *In each round, the algorithm chooses an exploratory policy* π *that is used to collect one episode worth of data in a task* $M \in \mathcal{M}$ *of its own choice. A sample-efficient algorithm should output a near-optimal policy for each task in a polynomial number of rounds.*

Exploration design plays an important role in achieving sample-efficient learning. Previous sample-efficient algorithms for single-task learning ($|\mathcal{M}| = 1$) heavily rely on strategic design on the exploratory policies, such as Optimism in Face of Uncertainty (OFU) (Auer et al., 2008; Bartlett & Tewari, 2009; Dann et al., 2017) and Posterior Sampling (Russo & Van Roy, 2014; Osband & Van Roy, 2017). Strategic design is criticized for either being restricted to environments with strong structural assumptions, or involving intractable computation oracle, such as non-convex optimization (Jiang, 2018; Jin et al., 2021a). For instance, GOLF (Jin et al., 2021a), a model-free general function approximation online learning algorithm, involves finding the optimal Q-value function f in a function class $\mathcal{F}$. It requires the following two intractable computation oracles in (1).

$$\mathcal{F}^t = \{f \in \mathcal{F} : f \text{ has low empirical Bellman error}\} \text{ and } f^{(t)} = \underset{f \in \mathcal{F}^t}{\arg\max}\, f(s_1, \pi(s_1 \mid f)). \quad (1)$$

In contrast, myopic exploration design like ϵ-greedy that injects random noise to a current greedy policy is easy to implement and performs well in a wide range of applications (Mnih et al., 2015; Kalashnikov et al., 2018), while it is shown to have exponential sample complexity in the worst case for single-task learning (Osband et al., 2019). Throughout the paper, we ask the main question:

Can algorithms with myopic exploration design be sample-efficient for MTRL?

In this paper, we address the question by showing that *a simple algorithm that explores one task with ϵ-greedy policies from other tasks can be sample-efficient if the task set $\mathcal{M}$ is adequately diverse.* Our results may shed some light on the longstanding mystery that ϵ-greedy is successful in practice, while being shown sample-inefficient in theory. We argue that in an MTRL setting, ϵ-greedy policies may no longer behave myopically, as they explore myopically around the optimal policies from other tasks. When the task set is adequately diverse, this exploration may provide sufficient coverage. It is worth-noting that this may partially explain the success of curriculum learning in RL (Narvekar et al., 2020), that is the optimal policy of one task may easily produce a good exploration for the next task, thus reducing the complexity of exploration. This connection will be formally discussed later.

To summarize our contributions, *this work provides a general framework that converts the task of strategic exploration design into the task of constructing diverse task set in an MTRL setting*, where myopic exploration like ϵ-greedy can also be sample-efficient. We discuss a sufficient diversity condition under a general value function approximation setting (Jin et al., 2021a; Dann et al., 2022). We show that the condition guarantees polynomial sample-complexity bound by running the aforementioned algorithm with myopic exploration design. We further discuss how to satisfy the diversity condition in different case studies, including tabular cases, linear cases and linear quadratic regulator cases. In the end, we validate our theory with experiments on synthetic robotic control environments, where we observe that a diverse task set is similar to the task selection of the state-of-the-art automatic curriculum learning algorithm, which has been empirically shown to improve sample efficiency.

2 Problem Setup

The following are notations that will be used throughout the paper.

Notation. For a positive integer H, we denote $[H] \coloneqq \{1, \ldots, H\}$. For a discrete set $\mathcal{A}$, we denote $\Delta_{\mathcal{A}}$ by the set of distributions over $\mathcal{A}$. We use $\mathcal{O}$ and Ω to denote the asymptotic upper and lower bound notations and use $\tilde{\mathcal{O}}$ and $\tilde{\Omega}$ to hide the logarithmic dependence. Let $\{\mathbf{1}_i\}_{i \in [d]}$ be the standard basis that spans $\mathbb{R}^d$. We let $N_{\mathcal{F}}(\rho)$ denote the ℓ_∞ covering number of a function class $\mathcal{F}$ at scale ρ. For a class $\mathcal{F}$, we denote the N-times Cartesian product of $\mathcal{F}$ by $(\mathcal{F})^{\otimes N}$.

2.1 Proposed Multitask Learning Scenario

Throughout the paper, we consider each task as an episodic MDP denoted by $M = (\mathcal{S}, \mathcal{A}, H, P_M, R_M)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $H \in \mathbb{N}$ represents the horizon length in each episode, $P_M = (P_{h,M})_{h \in [H]}$ is the collection of transition kernels, and $R_M = (R_{h,M})_{h \in [H]}$ is the collection of immediate reward functions. Each $P_{h,M} : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ and each $R_{h,M} : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$. Note that we consider a scenario in which all the tasks share the same state space, action space, and horizon length.

An agent interacts with an MDP M in the following way: starting with a fixed initial state s_1, at each step $h \in [H]$, the agent decides an action a_h and the environment samples the next state $s_{h+1} \sim P_{h,M}(\cdot \mid s_h, a_h)$ and next reward $r_h = R_{h,M}(s_h, a_h)$. An episode is a sequence of states, actions, and rewards $(s_1, a_1, r_1, \ldots, s_H, a_h, r_H, s_{H+1})$. In general, we assume that the sum of r_h is upper bounded by 1 for any action sequence almost surely. The goal of an agent is to maximize the cumulative reward $\sum_{h=1}^{H} r_h$ by optimizing their actions.

The agent chooses actions based on *Markovian policies* denoted by $\pi = (\pi_h)_{h \in [H]}$ and each π_h is a mapping $\mathcal{S} \mapsto \Delta_{\mathcal{A}}$, where $\Delta_{\mathcal{A}}$ is the set of all distributions over $\mathcal{A}$. Let Π denote the space of all such policies. For a finite action space, we let $\pi_h(a \mid s)$ denote the probability of selecting action a given state s at the step h. In case of the infinite action space, we slightly abuse the notation by letting $\pi_h(\cdot \mid s)$ denote the density function.

Proposed multitask learning scenario and objective. We consider the following multitask RL learning scenario. An algorithm interacts with a set of tasks $\mathcal{M}$ sequentially for T rounds. At the each round t, the algorithm chooses an exploratory policy, which is used to collect data for one episode in a task $M \in \mathcal{M}$ of its own choice. At the end of T rounds, the algorithm outputs a set of policies $\{\pi_M\}_{M \in \mathcal{M}}$. The goal of an algorithm is to learn a near-optimal policy π_M for each task $M \in \mathcal{M}$. The sample complexity of an algorithm is defined as follows.

Definition 1 (MTRL Sample Complexity)**.** *An algorithm $\mathcal{L}$ is said to have sample-complexity of $\mathcal{C}^{\mathcal{L}}_{\mathcal{M}} : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{N}$ for a task set $\mathcal{M}$ if for any $\beta > 0, \delta \in (0,1)$, it outputs a β-optimal policy π_M for each MDP $M \in \mathcal{M}$ with probability at least $1-\delta$, by interacting with the task set for $\mathcal{C}^{\mathcal{L}}_{\mathcal{M}}(\beta, \delta)$ rounds. We omit the notations for $\mathcal{L}$ and $\mathcal{M}$ when they are clear from the context to ease presentation.*

A sample-efficient algorithm should have a sample-complexity polynomial in the parameters of interests. For the tabular case, where the state space and action space are finite, $\mathcal{C}(\beta, \delta)$ should be polynomial in $|\mathcal{S}|$, $|\mathcal{A}|$, $|\mathcal{M}|$, H, and $1/\beta$ for a sample-efficient learning. Current state-of-the-art algorithm (Zhang et al., 2021) on a single-task tabular MDP achieves sample-complexity of $\tilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}|/\beta^2)$[1]. This bound translates to a MTRL sample-complexity bound of $\tilde{\mathcal{O}}(|\mathcal{M}||\mathcal{S}||\mathcal{A}|/\beta^2)$ by running their algorithm individually for each $M \in \mathcal{M}$. However, their exploration design closely follows the principle of Optimism in Face of Uncertainty, which is normally criticized for over-exploring.

2.2 Value Function Approximation

We consider the setting where value functions are approximated by general function classes. Denote the value function of an MDP M with respect to a policy π by

$$Q^{\pi}_{h,M}(s,a) = \mathbb{E}^{M}_{\pi}\left[r_h + V^{\pi}_{h+1,M}(s_{h+1}) \mid s_h = s, a_h = a\right]$$
$$V^{\pi}_{h,M}(s) = \mathbb{E}^{M}_{\pi}\left[Q^{\pi}_{h,M}(s_h, a_h) \mid s_h = s\right],$$

where by $\mathbb{E}^{M}_{\pi}$, we take expectation over the randomness of trajectories sampled by policy π on MDP M and we let $V^{\pi}_{H+1,M}(s) \equiv 0$ for all $s \in \mathcal{S}$ and $\pi \in \Pi$. We denote the optimal policy for MDP M by π^*_M. The corresponding value functions are denoted by $V^*_{h,M}$ and $Q^*_{h,M}$, which is shown to satisfy Bellman Equation $\mathcal{T}^M_h Q^*_{h+1,M} = Q^*_{h,M}$, where for any $g : \mathcal{S} \times \mathcal{A} \mapsto [0,1]$, $\left(\mathcal{T}^M_h g\right)(s,a) = \mathbb{E}[r_h + \max_{a' \in \mathcal{A}} g(s_{h+1}, a') \mid s_h = s, a_h = a]$.

The agent has access to a collection of function classes $\mathcal{F} = (\mathcal{F}_h : \mathcal{S} \times \mathcal{A} \mapsto [0,H])_{h \in [H+1]}$. We assume that different tasks share the same set of function class. For each $f \in \mathcal{F}$, we denote by $f_h \in \mathcal{F}_h$ the h-th component of the function f. We let $\pi^f = \{\pi^f_h\}_{h \in [H]}$ be the greedy policy with $\pi^f_h(s) = \arg\max_{a \in \mathcal{A}} f_h(s,a)$. When it is clear from the context, we slightly abuse the notation and let $f \in (\mathcal{F})^{\otimes |\mathcal{M}|}$ be a joint function for all the tasks. We further let f_M denote the function for the task M and $f_{h,M}$ denote its h-th component.

Define Bellman error operator $\mathcal{E}^M_h$ such that $\mathcal{E}^M_h f = f_h - \mathcal{T}^M_h f_{h+1}$ for any $f \in \mathcal{F}$. The goal of the learning algorithm is to approximate $Q^*_{h,M}$ through the function class $\mathcal{F}_h$ by minimizing the empirical Bellman error for each step h and task M.

To provide theoretical guarantee on this practice, we make the following realizability and completeness assumptions. The two assumptions and their variants are commonly used in the literature (Dann et al., 2017; Jin et al., 2021a).

Assumption 1 (Realizability and Completeness)**.** *For any MDP M considered in this paper, we assume $\mathcal{F}$ is realizable and complete under the Bellman operator such that $Q^*_{h,M} \in \mathcal{F}_h$ for all $h \in [H]$ and for every $h \in [H]$, $f_{h+1} \in \mathcal{F}_{h+1}$ there is a $f_h \in \mathcal{F}_h$ such that $f_h = \mathcal{T}^M_h f_{h+1}$.*

[1] This bound is under the regime with $1/\beta \gg |\mathcal{S}|$

2.3 Myopic Exploration Design

As opposed to carefully designed exploration, myopic exploration injects random noise to the current greedy policy. For a given greedy policy π, we use $\mathrm{expl}(\pi)$ to denote the myopic exploration policy based on π. Depending on the action space, the function expl can take different forms. The most common choice for finite action spaces is ϵ-greedy, which mixes the greedy policy with a random action: $\mathrm{expl}(\pi_h)(a \mid s) = (1-\epsilon_h)\pi_h(a \mid s)+\epsilon_h/A$.[2] As it is our main study of exploration strategies, we let expl be ϵ-greedy function if not explicitly specified. For a continuous action space, we consider exploration with Gaussian noise: $\mathrm{expl}(\pi_h)(a \mid s) = (1-\epsilon_h)\pi_h(a \mid s)+\epsilon_h \exp(-a^2/2\sigma_h^2)/\sqrt{2\pi\sigma_h^2}$. Gaussian noise is useful for Linear Quadratic Regulator (LQR) setting (discussed in Appendix F.)

3 Multitask RL Algorithm with Policy-Sharing

In this section, we introduce a generic algorithm (Algorithm 1) for the proposed multitask RL scenario without any strategic exploration, whose theoretical properties will be studied throughout the paper. In a typical single-task learning, a myopically exploring agent samples trajectories by running its current greedy policy estimated from the historical data equipped with naive explorations like ϵ-greedy.

In light of the exploration benefits of MTRL, we study Algorithm 1 as a counterpart of the single-task learning scenario in the MTRL setting. Algorithm 1 maintains a dataset for each MDP separately and different tasks interact in the following way: in each round, Algorithm 1 explores every MDP with an exploratory policy that is the mixture (defined in Definition 2) of greedy policies of all the MDPs in the task set (Line 8). One way to interpret Algorithm 1 is that we share knowledge across tasks by policy sharing instead of parameter sharing or feature extractor sharing in the previous literature.

Definition 2 (Mixture Policy). *For a set of policies $\{\pi_i\}_{i=1}^N$, we denote $\mathrm{Mixture}(\{\pi_i\}_{i=1}^N)$ by the mixture of N policies, such that before the start of an episode, it samples $I \sim \mathrm{Unif}([N])$, then runs policy π_I for the rest of the episode.*

The greedy policy is obtained from an offline learning oracle $\mathcal{Q}$ (Line 4) that maps a dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$ to a function $f \in \mathcal{F}$, such that $\mathcal{Q}(\mathcal{D})$ is an approximate solution to the following minimization problem $\arg\min_{f\in\mathcal{F}} \sum_{i=1}^N \left(f_{h_i}(s_i, a_i) - r_i - \max_{a'\in\mathcal{A}} f_{h_i+1}(s_i', a')\right)^2$. In practice, one can run fitted Q-iteration for an approximate solution.

Connection to Hindsight Experience Replay (HER) and multi-goal RL. We provide some justifications for the choice of the mixture policy. HER is a common practice (Andrychowicz et al., 2017) in the multi-goal RL setting (Andrychowicz et al., 2017; Chane-Sane et al., 2021; Liu et al., 2022), where the reward distribution is a function of goal-parameters. HER relabels the rewards in trajectories in the experience buffer such that they were as if sampled for a different task. Notably, this exploration strategy is akin to randomly selecting task and collecting a trajectory on using its own epsilon-greedy policy, followed by relabeling rewards to simulate a trajectory on, which is equivalent to Algorithm 1. Yang et al. (2022b); Zhumabekov et al. (2023) also designed an algorithm with ensemble policies, which is shown to improve generalizability. It is worth noting that the ensemble implementation of Algorithm 1 and that of Zhumabekov et al. (2023) differs. In Algorithm 1, policies are mixed trajectory-wise, with one policy randomly selected for the entire episode. In contrast, Zhumabekov et al. (2023) mixes policies on a step-wise basis for a continuous action space, selecting actions as a weighted average of actions chosen by different policies. This distinction is vital in showing the rigorous theoretical guarantee outlined in this paper.

Connection to curriculum learning. Curriculum learning is the approach that learns tasks in a specific order to improve the multitask learning performance (Bengio et al., 2009). Although Algorithm 1 does not explicitly implement curriculum learning by assigning preferences to tasks, improvement could be achieved through adaptive task selection that may reflect the benefits of curriculum learning. Intuitively, any curricula that selects tasks through an order of $M_1, \ldots, M_T$ is implicitly included in Algorithm 1 as it explores all the MDPs in each round with the mixture of all epsilon-greedy policies. This means that the sample-complexity of Algorithm 1 provides an upper bound on the sample complexity of underlying optimal task selection. A formal discussion is deferred to Appendix B.

[2] Note that we consider a more general setup, where the exploration probability ϵ can depend on h.

Algorithm 1 Generic Algorithm for MTRL with Policy-Sharing

```
**Input:** function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_{H+1}$, task set $\mathcal{M}$, exploration function expl
Initialize $\mathcal{D}_{0,M} \leftarrow \emptyset$ for all $M \in \mathcal{M}$
**for** round $t = 1, 2, \ldots, \lfloor T/|\mathcal{M}| \rfloor$ **do**
    Offline learning oracle outputs $\hat{f}_{t,M} \leftarrow \mathcal{Q}(\mathcal{D}_{t-1,M})$ for each $M$          ▷ Offline learning
    Set myopic exploration policy $\hat{\pi}_{t,M} \leftarrow \text{expl}(\pi^{\hat{f}_{t,M}})$ for each $M$
    Set $\hat{\pi}_t \leftarrow \text{Mixture}(\{\hat{\pi}_{t,M}\}_{M \in \mathcal{M}})$          ▷ Share policies
    **for** $M \in \mathcal{M}$ **do**
        Sample one episode $\tau_{t,M}$ on MDP $M$ with policy $\hat{\pi}_t$          ▷ Collect new trajectory
        Add $\tau_{t,M}$ to the dataset: $\mathcal{D}_{t,M} \leftarrow \mathcal{D}_{t-1,M} \cup \{\tau_{t,M}\}$
    **end for**
**end for**
**Return** $\hat{\pi}_M = \text{Mixture}(\{\hat{\pi}_{t,M}\}_{t \in \lfloor T/|\mathcal{M}| \rfloor})$ for each $M$
```

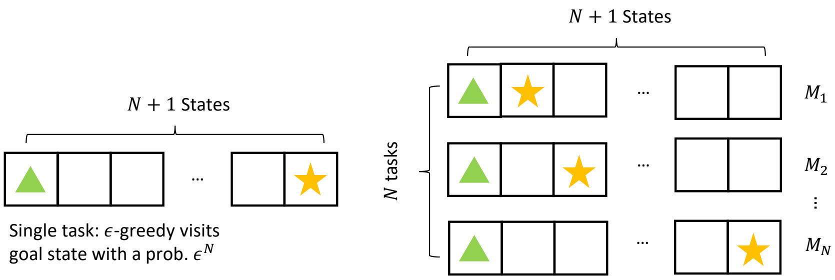

N + 1 States
Single task: ϵ-greedy visits goal state with a prob. ϵ^N
N + 1 States
N tasks
M_1
M_2
M_N

Figure 1: A diverse grid-world task set on a long hallway with $N+1$ states. From the left to the right, it represents a single-task and a multitask learning scenario, respectively. The triangles represent the starting state and the stars represent the goal states, where an agent receives a positive reward. The agent can choose to move forward or backward.

4 Generic Sample Complexity Guarantee

In this section, we rigorously define the diversity condition and provide a sample-complexity upper bound for Algorithm 1. We start with introducing an intuitive example on how diversity encourages exploration in a multitask setting.

Motivating example. Figure 1 introduces a motivating example of grid-world environment on a long hallway with $N+1$ states. Since this is a deterministic tabular environment, whenever a task collects an episode that visits its goal state, running an offline policy optimization algorithm with pessimism will output its optimal policy.

Left penal of Figure 1 is a single-task learning, where the goal state is N steps away from the initial state, making it exponentially hard to visit the goal state with ϵ-greedy exploration. Figure 1 on the right demonstrates a multitask learning scenario with N tasks, whose goal states "diversely" distribute along the hallway. A main message of this paper is the advantage of exploring one task by running the ϵ-greedy policies of other tasks. To see this, consider any current greedy policies $(\pi_1, \pi_2, \ldots, \pi_N)$. Let i be the first non-optimal policy, i.e. all $j < i$, π_j is optimal for M_j. Since π_{i-1} is optimal, by running an ϵ-greedy of π_{i-1} on MDP M_i, we have a probability of $\prod_{h=1}^{i-1}(1-\epsilon_h)\epsilon_i$ to visit the goal state of M_i, allowing it to improve its policy to optimal in the next round. Such improvement can only happen for N times and all policies will be optimal within polynomial (in N) number of rounds if we choose $\epsilon_h = 1/(h+1)$. Hence, myopic exploration with a diverse task set leads to sample-efficient learning. The rest of the section can be seen as generalizing this idea to function approximation.

4.1 Multitask Myopic Exploration Gap

Dann et al. (2022) proposed an assumption named Myopic Exploration Gap (MEG) that allows efficient myopic exploration for a single MDP under strong assumptions on the reward function, or on the mixing time. We extend this definition to the multitask learning setting. For the conciseness of the notation, we let $\text{expl}(f)$ denote the following mixture policy $\text{Mixture}(\{\text{expl}(\pi^{f_M})\}_{M \in \mathcal{M}})$

for a joint function $f \in (\mathcal{F})^{\otimes|\mathcal{M}|}$. Intuitively, a large myopic exploration gap implies that within all the policies that can be learned by the current exploratory policy, there exists one that can make significant improvement on the current greedy policy.

Definition 3 (Multitask Myopic Exploration Gap (Multitask MEG)). *For any $\mathcal{M}$, a function class $\mathcal{F}$, a joint function $f \in (\mathcal{F})^{\otimes|\mathcal{M}|}$, we say that f has $\alpha(f, \mathcal{M}, \mathcal{F})$-myopic exploration gap, where $\alpha(f, \mathcal{M}, \mathcal{F})$ is the value to the following maximization problem:*

$$\max_{M \in \mathcal{M}} \sup_{\tilde{f} \in \mathcal{F}, c \geq 1} \frac{1}{\sqrt{c}}(V_{1,M}^{\tilde{f}} - V_{1,M}^{f_M}), \text{ s.t. for all } f' \in \mathcal{F} \text{ and } h \in [H],$$

$$\mathbb{E}_{\pi^{\tilde{f}}}^{M}[(\mathcal{E}_h^M f')(s_h, a_h)]^2 \leq c\mathbb{E}_{\text{expl}(f)}^{M}[(\mathcal{E}_h^M f')(s_h, a_h)]^2$$

$$\mathbb{E}_{\pi^{f_M}}^{M}[(\mathcal{E}_h^M f')(s_h, a_h)]^2 \leq c\mathbb{E}_{\text{expl}(f)}^{M}[(\mathcal{E}_h^M f')(s_h, a_h)]^2.$$

Let $M(f, \mathcal{M}, \mathcal{F})$, $c(f, \mathcal{M}, \mathcal{F})$ be the corresponding $M \in \mathcal{M}$ and c that attains the maximization.

Design of myopic exploration gap. To illustrate the spirit of this definition, we specialize to the tabular case, where conditions in Definition 3 can be replaced by the concentrability (Jin et al., 2021b) condition: for all $s, a, h \in \mathcal{S} \times \mathcal{A} \times [H]$, we require

$$\mu_{h,M}^{\pi^{\tilde{f}}}(s, a) \leq c\mu_{h,M}^{\text{expl}(f)}(s, a) \text{ and } \mu_{h,M}^{\pi^{f_M}}(s, a) \leq c\mu_{h,M}^{\text{expl}(f)}(s, a), \tag{2}$$

where $\mu_{h,M}^{\pi}(s, a)$ is the occupancy measure, i.e. the probability of visiting (s, a) at the step h by running policy π on MDP M. The design of myopic exploration gap connects deeply to the theory of offline Reinforcement Learning. For a specific MDP M, Equation (2) defines a set of policies with concentrability assumption (Xie et al., 2021) that can be accurately evaluated through the offline data collected by the current behavior policy. As an extension to the Single-task MEG in Dann et al. (2022), Multitask MEG considers the maximum myopic exploration gap over a set of MDPs and the behavior policy is a mixture of all the greedy policies. Definition 3 reduces to the Single-task MEG when the task set $\mathcal{M}$ is a singleton.

4.2 Sample Complexity Guarantee

We propose Diversity in Definition 5, which relies on having a lower bounded Multitask MEG for any suboptimal policy. We then present Theorem 1 that shows an upper bound for sample complexity of Algorithm 1 by assuming diversity.

Definition 4 (Multitask Suboptimality). *For a multitask RL problem with MDP set $\mathcal{M}$ and value function class $\mathcal{F}$. Let $\mathcal{F}_\beta \subset (\mathcal{F})^{\otimes|\mathcal{M}|}$ be the β-suboptimal class, such that for any $f \in \mathcal{F}_\beta$, there exists f_M and π^{f_M} is β-suboptimal for MDP M, i.e. $V_{1,M}^{\pi^{f_M}} \leq \max_{\pi \in \Pi} V_{1,M}^{\pi} - \beta$.*

Definition 5 (Diverse Tasks). *For some function $\tilde{\alpha} : [0, 1] \mapsto \mathbb{R}$, and $\tilde{c} : [0, 1] \mapsto \mathbb{R}$, we say that a tasks set is $(\tilde{\alpha}, \tilde{c})$-diverse if any $f \in \mathcal{F}_\beta$ has multitask myopic exploration gap $\alpha(f, \mathcal{M}, \mathcal{F}) \geq \tilde{\alpha}(\beta)$ and $c(f, \mathcal{M}, \mathcal{F}) \leq \tilde{c}(\beta)$ for any constant $\beta > 0$.*

To simplify presentation, we state the result here assuming parametric growth of the Bellman eluder dimension and covering number, that is $d_{\text{BE}}(\mathcal{F}, \Pi_{\mathcal{F}}, \rho) \leq d_{\mathcal{F}}^{\text{BE}} \log(1/\rho)$ and $\log(N'(\mathcal{F})(\rho)) \leq d_{\mathcal{F}}^{\text{cover}} \log(1/\rho)$, where $\dim_{\text{BE}}(\mathcal{F}, \Pi_{\mathcal{F}}, \rho)$ is the Bellman-Eluder dimension of class $\mathcal{F}$ and $N'_{\mathcal{F}}(\rho) = \sum_{h=1}^{H-1} N_{\mathcal{F}_h}(\rho) N_{\mathcal{F}_{h+1}}(\rho)$. A formal definition of Bellman-Eluder dimension is deferred to Appendix H. This parametric growth rate holds for most of the regular classes including tabular and linear (Russo & Van Roy, 2013). Similar assumptions are made in Chen et al. (2022a).

Theorem 1 (Upper Bound for Sample Complexity). *Consider a multitask RL problem with MDP set $\mathcal{M}$ and value function class $\mathcal{F}$ such that $\mathcal{M}$ is $(\tilde{\alpha}, \tilde{c})$-diverse. Then Algorithm 1 with ϵ-greedy exploration function has a sample-complexity*

$$\mathcal{C}(\beta, \delta) = \tilde{\mathcal{O}}\left(|\mathcal{M}|^2 H^2 d_{\mathcal{F}}^{BE} d_{\mathcal{F}}^{cover} \frac{\ln \tilde{c}(\beta)}{\tilde{\alpha}^2(\beta)} \ln(1/\delta)\right).$$

4.3 Comparing Single-task and Multitask MEG

The sample complexity bound in Theorem 1 reduces to the single-task sample complexity in Dann et al. (2022) when $\mathcal{M}$ is a singleton. To showcase the potential benefits of multitask learning, we provide a comprehensive comparison between Single-task and Multitask MEG. We focus our discussion on $\alpha(f, \mathcal{M}, \mathcal{F})$ because $c(f, \mathcal{M}, \mathcal{F}) \leq ((\max_\pi V^\pi_{1,M(f,\mathcal{M},\mathcal{F})} - V^{\pi^{f_M}}_{1,M(f,\mathcal{M},\mathcal{F})})/\alpha(f, \mathcal{M}, \mathcal{F}))^2 \leq 1/\alpha^2(f, \mathcal{M}, \mathcal{F})$ and $c(f, \mathcal{M}, \mathcal{F})$ only impacts sample complexity bound through a logarithmic term.

We first show that Multitask MEG is lower bounded by Single-task MEG up to a factor of $1/\sqrt{|\mathcal{M}|}$.

Proposition 1. *Let $\mathcal{M}$ be any set of MDPs and $\mathcal{F}$ be any function class. We have that $\alpha(f, \mathcal{M}, \mathcal{F}) \geq \alpha(f_M, \{M\}, \mathcal{F})/\sqrt{|\mathcal{M}|}$ for all $f \in (\mathcal{F})^{\otimes|\mathcal{M}|}$ and $M \in \mathcal{M}$.*

Proposition 1 immediately implies that whenever all tasks in a task set can be learned with myopic exploration individually (see examples in Dann et al. (2022)), they can also be learned with myopic exploration through MTRL with Algorithm 1 with an extra factor of $|\mathcal{M}|^2$, which is still polynomial in all the related parameters.

We further argue that Single-task MEG can easily be exponentially small in H, in which case, myopic exploration fails.

Proposition 2. *Let M be a tabular sparse-reward MDP, i.e., $R_{h,M}(s, a) = 0$ at all (s, a, h) except for a goal tuple (s_t, a_t, h_t) and $R_{h_t,M}(s_t, a_t) = 1$. Recall $\mu^\pi_{h,M}(s, a) = \mathbb{E}^M_\pi[\mathbb{1}_{s_h=s,a_h=a}]$. Then*

$$\alpha(f, \{M\}, \mathcal{F}) \leq \max_{\tilde{\pi}}(\mu^{\tilde{\pi}}_{h_t}(s_t, a_t) - \mu^{\pi^f}_{h_t}(s_t, a_t))\sqrt{\mu^{\mathrm{expl}(\pi^f)}_{h_t}(s_t, a_t)/\mu^{\tilde{\pi}}_{h_t}(s_t, a_t)} \leq \sqrt{\mu^{\mathrm{expl}(\pi^f)}_{h_t}(s_t, a_t)}.$$

Sparse-reward MDP is widely studied in goal-conditioned RL, where the agent receives a non-zero reward only when it reaches a goal-state (Andrychowicz et al., 2017; Chane-Sane et al., 2021). Proposition 2 implies that a single-task MEG can easily be exponentially small in H as long as one can find a policy π such that its ϵ-greedy version, $\mathrm{expl}(\pi)$, visits the goal tuple (s_t, a_t, h_t) with a probability that is exponentially small in H. This is true when the environment requires the agent to execute a fixed sequence of actions to reach the goal tuple as it is the case in Figure 1. Indeed, in the environment described in the left panel of Figure 1, a policy that always moves left will have $\alpha(f, \{M\}, \mathcal{F}) \leq \sqrt{\mu^{\mathrm{expl}(\pi^f)}_{h_t}(s_t, a_t)} \leq \sqrt{\Pi^H_{h=1}(\epsilon_h/2)} \leq 2^{-H/2}$. This is also consistent with our previous discussion that ϵ-greedy requires $\Omega(2^H)$ number of episodes to learn the optimal policy in the worst case. As we will show later in Section 5.1, Multitask MEG for the tabular case can be lowered bounded by $\Omega(\sqrt{1/(|\mathcal{A}||\mathcal{M}|H)})$ for adequately diverse task set $\mathcal{M}$, leading to an exponential separation between Single-task and Multitask MEG.

5 Lower Bounding Myopic Exploration Gap

Following the generic result in Theorem 1, the key to the problem is to lower bound myopic exploration gap $\tilde{\alpha}(\beta)$. In this section, we lower bound $\tilde{\alpha}(\beta)$ for the linear MDP case. We defer an improved analysis for the tabular case and the Linear Quadratic Regulator cases to Appendix F.

Linear MDPs have been an important case study for the theory of RL (Wang et al., 2019; Jin et al., 2020b; Chen et al., 2022b). It is a more general case than tabular MDP and has strong implication for Deep RL. In order to employ ϵ-greedy, we consider finite action space, while the state space can be infinite.

Definition 6 (Linear MDP Jin et al. (2020b)). *An MDP is called linear MDP if its transition probability and reward function admit the following form. $P_h(s' \mid s, a) = \langle \phi_h(s, a), \mu_h(s') \rangle$ for some known function $\phi_h : \mathcal{S} \times \mathcal{A} \mapsto (\mathbb{R}^+)^d$ and unknown function $\mu_h : \mathcal{S} \mapsto (\mathbb{R}^+)^d$. $R_h(s, a) = \langle \phi_h(s, a), \theta_h \rangle$ for unknown parameters θ_h [3]. Without loss of generality, we assume $\|\phi_h(s, a)\| \leq 1$ for all $s, a, h \in \mathcal{S} \times \mathcal{A} \times \mathcal{H}$ and $\max\{\|\mu_h(s)\|, \|\theta_h\|\} \leq \sqrt{d}$ for all $s, h \in \mathcal{S} \times [H]$.*

An important property of Linear MDPs is that the value function also takes the linear form and the linear function class defined below satisfies Assumption 1.

Proposition 3 (Proposition 2.3 (Jin et al., 2020b)). *For linear MDPs, we have for any policy π, $Q^\pi_{h,M}(s, a) = \langle \phi_h(s, a), w^\pi_{h,M} \rangle$, where $w^\pi_{h,M} = \theta_{h,M} + \int_{\mathcal{S}} V^\pi_{h+1,M}(s')\mu_h(s')ds' \in \mathbb{R}^d$. Therefore, we only need to consider $\mathcal{F}_h = \{(s, a) \mapsto \langle \phi_h(s, a), w \rangle : w \in \mathbb{R}^d, \|w\|_2 \leq 2\sqrt{d}\}$.*

Now we are ready to define a diverse set of MDPs for the linear MDP case.

[3] Note that we consider non-negative measure μ_h.

Definition 7 (Diverse MDPs for linear MDP case)**.** *We say $\mathcal{M}$ is a diverse set of MDPs for the linear MDP case, if they share the same feature extractor ϕ_h and the same measure μ_h (leading to the same transition probabilities) and for any $h \in [H]$, there exists a subset $\{M_{i,h}\}_{i\in[d]} \subset \mathcal{M}$, such that the reward parameter $\theta_{h,M_{i,h}} = \mathbf{1}_i$ and all the other $\theta_{h',M_{i,h}} = 0$ with $h' \neq h$, where $\mathbf{1}_i$ is the onehot vector with a positive entry at the dimension i.*

We need the assumption that the minimum eigenvalue of the covariance matrix is strictly lower bounded away from 0. The feature coverage assumption is commonly use in the literature that studies Linear MDPs (Agarwal et al., 2022). Suppose Assumption 2 hold, we have Theorem 2, which lower bounds the multitask myopic exploration gap. Combined with Theorem 1, we have a sample-complexity bound of $\tilde{\mathcal{O}}(|\mathcal{M}|^3 H^3 d^2 |\mathcal{A}|/(\beta^2 b_1^2))$ with $|\mathcal{M}| \geq d$.

Assumption 2 (Feature coverage)**.** *For any $\nu \in \mathbb{S}^{d-1}$ and $[\nu]_i > 0$ for all $i \in [d]$, there exists a policy π such that $\mathbb{E}_\pi[\nu^\top \phi_h(s_h, a_h)] \geq b_1$, for some constant $b_1 > 0$.*

Theorem 2. *Consider $\mathcal{M}$ to be a diverse set as in Definition 7. Suppose Assumption 2 holds and $\beta \leq b_1/2$, then we have for any $f \in \mathcal{F}_\beta$, $\alpha(f, \mathcal{F}, \mathcal{M}) = \Omega(\sqrt{\beta^2 b_1^2/(|\mathcal{A}||\mathcal{M}|H)}$ by setting $\epsilon_h = 1/(h+1)$.*

Proof of Theorem 2 critically relies on iteratively showing the following lemma, which can be interpreted as showing that the feature covariance matrix induced by the optimal policies is full rank.

Lemma 1. *Fix a step h and fix a $\beta < b_1/2$. Let $\{\pi_i\}_{i=1}^d$ be d policies such that π_i is a β-optimal policy for $M_{i,h}$ as in Definition 7. Let $\tilde{\pi} = \mathrm{Mixture}(\{\mathrm{expl}(\pi_i)\}_{i=1}^d)$. Then we have $\lambda_{\min}(\Phi_{h+1}^{\tilde{\pi}}) \geq \epsilon_h \prod_{h'=1}^{h-1}(1-\epsilon_{h'}) b_1^2/(2dA)$.*

5.1 Discussions on the Tabular Case

Diverse tasks in Definition 7, when specialized to the tabular case, corresponds to $S \times H$ sparse-reward MDPs. Interestingly, similar constructions are used in reward-free exploration (Jin et al., 2020a), which shows that by calling an online-learning oracle individually for all the sparse reward MDPs, one can generate a dataset that outputs a near-optimal policy for any given reward function. We want to point out the intrinsic connection between the two settings: our algorithm, instead of generating an offline dataset all at once, generates an offline dataset at each step h that is sufficient to learn a near-optimal policy for MDPs that corresponds to the step $h+1$.

Relaxing coverage assumption. Though feature coverage Assumption (Assumption 2) is handy for our proof as it guarantees that any β-optimal policy (with $\beta < b_1/2$) has a probability at least $b_1/2$ to visit their goal state, this assumption may not be reasonable for the tabular MDP case. Intuitively, without this assumption, a β-optimal policy can be an arbitrary policy and we can have at most S such policies in total leading to a cumulative error of $S\beta$. A naive solution is to request a $S^{-H}\beta$ accuracy at the first step, which leads to exponential sample-complexity. In Appendix G, we show that an improved analysis can be done for the tabular MDP without having to make the coverage assumption. However, an extra price of SH has to be paid.

6 Implications of Diversity in Deep RL

Though this is a paper focusing on the theoretical justification, we add some preliminary discussion on what diversity means in Deep RL. Thus far, we have provided concrete example for diverse task set in tabular and linear MDPs with lower bounded Multitask MEG. It is often the case that in Deep RL, a pretrained feature extractor is used to generate embeddings for Q-value function and then a linear mapping is applied to generate the final output (Bhateja et al., 2023; Hejna III & Sadigh, 2023). This manner is similar to the setup in linear MDPs. We have shown that to achieve diversity for linear MDPs, one important property is to have a full rank covariance matrix of the embeddings at each step h if the optimal policy is executed (Lemma 1). In this section, we conduct simple simulation studies on BipedalWalker (Portelas et al., 2020), a robotic control environment, to verify that whether a more spread spectrum of the covariance matrix of the embeddings would lead to better sample efficiency.

We found in the experiments that *the subset of tasks that leads to more spread spectrum is similar to the task selection of automatic curriculum learning, which has been empirically shown to learn more sample efficiently.* Due to the page limit we provide a high-level summary of the simulation study and the readers can find full details in Appendix I.

Experiment setup. The learning agent is embodied into a bipedal walker whose motors are controlled by continuous actions of torques. The objective of the agent is to move forward as far as possible, while crossing stumps with varying heights at regular intervals. An environment or task, denoted as $M_{p,q}$, is controlled by a parameter vector (p, q), where p and q denote the heights of the stumps and the spacings between the stumps, respectively. Intuitively, challenging environments have higher and denser stumps. The agent is trained by Proximal Policy Optimization (PPO) (Schulman et al., 2017) with a standard actor-critic framework (Konda & Tsitsiklis, 1999) and with Boltzmann exploration that regularizes entropy. Detailed training setup can be found in Appendix I.

Proper level of heights leads to more spread spectrum. We evaluate the extracted feature at the end of the training generated by near-optimal policies on 100 tasks with different parameter vectors (p, q). We then compute the covariance matrix of the features for each task, whose spectrum are shown in Figure 2 (a). We observe that the eigenvalues can be 10 times higher for environments with an appropriate height (1.0-2.3), compared to extremely high and low heights, while they are roughly the same at different levels of spacings. This indicates that choosing an appropriate height is the key to properly scheduling tasks.

Coincidence with automatic curriculum learning. In fact, the task selection is similar to the tasks selected by the state-of-the-art Automatic Curriculum Learning (ACL). We investigate the curricula generated by ALP-GMM (Portelas et al., 2020), a well-established curriculum learning algorithm, for training an agent in the BipedalWalker environment for 20 million timesteps. Figure 2 (b) gives the density plots of the ACL task sampler during the training process, which shows a significant preference over heights in the middle range, with little preference over spacing. Since ACL has been shown to significantly improve the learning performance in BipedalWalker, this is a strong evidence that the diversity notation discussed in this paper truly leads to a better sample complexity even when Deep RL is employed.

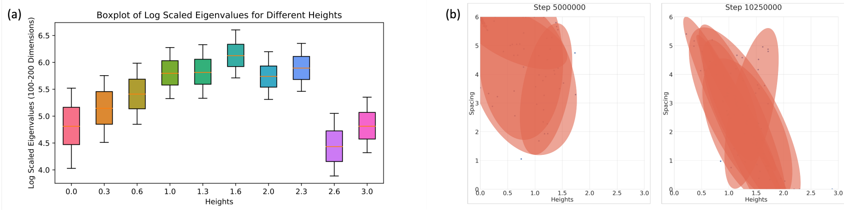


Figure 2: **(a)** Boxplot of the log-scaled eigenvalues of sample covariance matrices of the trained embeddings generated by the near optimal policies for different environments. **(b)** Task preference of automatically generated curriculum at 5M and 10M training steps respectively. The red regions are the regions where a task has a higher probability to be sampled.

7 Discussions

In this paper, we propose a new perspective to understand the sample efficiency of myopic exploration design through diverse multitask learning. We show that by learning a diverse set of tasks, a multitask RL algorithm with myopic exploration design can be sample-efficient. This paper is a promising first step towards understanding the exploration benefits of MTRL.

Towards diversity for general function classes. Despite the rich discussion on explicit form of diversity set for tabular and linear MDP cases, how to achieve diversity for any general function class is an open problem. Recalling our proof for the Linear MDP case, a sufficient condition is to include a set of MDPs for each step h, such that the state distribution generated by their optimal policies satisfy the concentrability assumptions. In other words, any MDP with positive reward only at the step $h + 1$ can be offline-learned through the dataset collected by these optimal policies. The diversity for general function classes poses the question on the number of tasks it takes to have sufficient coverage at the each step. We give a more detailed discussion of this topic in Appendix H.

Improving sample-complexity bound. Our sample complexity bound can be sub-optimal. For instance, Theorem 1 specialized to the tabular case has an upper sample complexity bound of $|\mathcal{M}|^2 S^3 H^5 A^2/\beta^2$, and the current optimal bound is $|\mathcal{M}|SA/\beta^2$ if tasks are learned independently. We conjecture that this gap may originate from two factors. First, the nature of the myopic exploration makes it less efficient because the exploration are conducted in a layered manner. Second, our algorithm collects trajectories for every MDP with the mixture of all the policy in each round, which may be improved if a curriculum is known beforehand.

8 Acknowledgement

Prof. Tewari acknowledges the support of NSF via grant IIS-2007055. Prof. Stone runs the Learning Agents Research Group (LARG) at UT Austin. LARG research is supported in part by NSF (FAIN-2019844, NRT-2125858), ONR (N00014-18-2243), ARO (E2061621), Bosch, Lockheed Martin, and UT Austin's Good Systems grand challenge. Prof. Stone also serves as the Executive Director of Sony AI America and receives financial compensation for this work. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

Part of this work is completed while Dr. Ziping Xu is a postdoc at Harvard University. He is supported by Prof. Susan A. Murphy through NIH/NIDA P50DA054039, NIH/NIBIB, OD P41EB028242 and NIH/NIDCR UH3DE028723.

References

Alekh Agarwal, Yuda Song, Wen Sun, Kaiwen Wang, Mengdi Wang, and Xuezhou Zhang. Provable benefits of representational transfer in reinforcement learning. *arXiv preprint arXiv:2205.14571*, 2022.

Jacob Andreas, Dan Klein, and Sergey Levine. Modular multitask reinforcement learning with policy sketches. In *International Conference on Machine Learning*, pp. 166–175. PMLR, 2017.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Peter Bartlett and Ambuj Tewari. Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Uncertainty in Artificial Intelligence: Proceedings of the 25th Conference*, pp. 35–42. AUAI Press, 2009.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Chethan Bhateja, Derek Guo, Dibya Ghosh, Anikait Singh, Manan Tomar, Quan Vuong, Yevgen Chebotar, Sergey Levine, and Aviral Kumar. Robotic offline rl from internet videos via value-function pre-training. *arXiv preprint arXiv:2309.13041*, 2023.

Emma Brunskill and Lihong Li. Sample complexity of multi-task reinforcement learning. *arXiv preprint arXiv:1309.6821*, 2013.

Daniele Calandriello, Alessandro Lazaric, and Marcello Restelli. Sparse multi-task reinforcement learning. *Advances in neural information processing systems*, 27, 2014.

Elliot Chane-Sane, Cordelia Schmid, and Ivan Laptev. Goal-conditioned reinforcement learning with imagined subgoals. In *International Conference on Machine Learning*, pp. 1430–1440. PMLR, 2021.

Jinglin Chen, Aditya Modi, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. On the statistical efficiency of reward-free exploration in non-linear rl. *arXiv preprint arXiv:2206.10770*, 2022a.

Liyu Chen, Rahul Jain, and Haipeng Luo. Improved no-regret algorithms for stochastic shortest path with linear mdp. In *International Conference on Machine Learning*, pp. 3204–3245. PMLR, 2022b.

Yuan Cheng, Songtao Feng, Jing Yang, Hong Zhang, and Yingbin Liang. Provable benefit of multitask representation learning in reinforcement learning. *arXiv preprint arXiv:2206.05900*, 2022.

Will Dabney, Georg Ostrovski, and André Barreto. Temporally-extended $\{epsilon\}$-greedy exploration. *arXiv preprint arXiv:2006.01782*, 2020.

Chris Dann, Yishay Mansour, Mehryar Mohri, Ayush Sekhari, and Karthik Sridharan. Guarantees for epsilon-greedy reinforcement learning with function approximation. In *International Conference on Machine Learning*, pp. 4666–4689. PMLR, 2022.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Milad Farjadnasab and Maryam Babazadeh. Model-free lqr design by q-function learning. *Automatica*, 137:110060, 2022.

Evan M Forman, Stephanie G Kerrigan, Meghan L Butryn, Adrienne S Juarascio, Stephanie M Manasse, Santiago Ontañón, Diane H Dallal, Rebecca J Crochiere, and Danielle Moskow. Can the artificial intelligence technique of reinforcement learning use continuously-monitored digital data to optimize treatment for weight loss? *Journal of behavioral medicine*, 42:276–290, 2019.

Susobhan Ghosh, Raphael Kim, Prasidh Chhabria, Raaz Dwivedi, Predrag Klasjna, Peng Liao, Kelly Zhang, and Susan Murphy. Did we personalize? assessing personalization by an online reinforcement learning algorithm using resampling. *arXiv preprint arXiv:2304.05365*, 2023.

Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *international conference on machine learning*, pp. 1311–1320. PMLR, 2017.

Donald Joseph Hejna III and Dorsa Sadigh. Few-shot preference learning for human-in-the-loop rl. In *Conference on Robot Learning*, pp. 2014–2025. PMLR, 2023.

Matteo Hessel, Hubert Soyer, Lasse Espeholt, Wojciech Czarnecki, Simon Schmitt, and Hado van Hasselt. Multi-task deep reinforcement learning with popart. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3796–3803, 2019.

Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander Hauptmann. Self-paced curriculum learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Nan Jiang. Pac reinforcement learning with an imperfect model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pp. 4870–4879. PMLR, 2020a.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020b.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021a.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021b.

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, pp. 651–673. PMLR, 2018.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Man Li, Jiahu Qin, Wei Xing Zheng, Yaonan Wang, and Yu Kang. Model-free design of stochastic lqr controller from a primal–dual optimization perspective. *Automatica*, 140:110253, 2022a.

Qiyang Li, Yuexiang Zhai, Yi Ma, and Sergey Levine. Understanding the complexity gains of single-task rl with a curriculum. *arXiv preprint arXiv:2212.12809*, 2022b.

Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22, 2020.

Minghuan Liu, Menghui Zhu, and Weinan Zhang. Goal-conditioned reinforcement learning: Problems and solutions. *arXiv preprint arXiv:2201.08299*, 2022.

Yao Liu and Emma Brunskill. When simple exploration is sample efficient: Identifying sufficient conditions for random exploration to yield pac rl algorithms. *arXiv preprint arXiv:1805.09045*, 2018.

Rui Lu, Gao Huang, and Simon S Du. On the power of multitask representation learning in linear mdp. *arXiv preprint arXiv:2106.08053*, 2021.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *The Journal of Machine Learning Research*, 21(1):7382–7431, 2020.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pp. 2701–2710. PMLR, 2017.

Ian Osband, Benjamin Van Roy, Daniel J Russo, Zheng Wen, et al. Deep exploration via randomized value functions. *J. Mach. Learn. Res.*, 20(124):1–62, 2019.

Anastasia Pentina, Viktoriia Sharmanska, and Christoph H Lampert. Curriculum learning of multiple tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5492–5500, 2015.

Rémy Portelas, Cédric Colas, Katja Hofmann, and Pierre-Yves Oudeyer. Teacher algorithms for curriculum learning of deep rl in continuously parameterized environments. In *Conference on Robot Learning*, pp. 835–853. PMLR, 2020.

Clément Romac, Rémy Portelas, Katja Hofmann, and Pierre-Yves Oudeyer. Teachmyagent: a benchmark for automatic curriculum learning in deep rl. In *International Conference on Machine Learning*, pp. 9052–9063. PMLR, 2021.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pp. 8937–8948. PMLR, 2020.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

Ruosong Wang, Simon S Du, Lin Yang, and Russ R Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *Advances in neural information processing systems*, 33:17816–17826, 2020.

Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2021.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Ziping Xu and Ambuj Tewari. Representation learning beyond linear prediction functions. *Advances in Neural Information Processing Systems*, 34:4792–4804, 2021.

Ziping Xu and Ambuj Tewari. On the statistical benefits of curriculum learning. In *International Conference on Machine Learning*, pp. 24663–24682. PMLR, 2022.

Ziping Xu, Amirhossein Meisami, and Ambuj Tewari. Decision making problems with funnel structure: A multi-task learning approach with application to email marketing campaigns. In *International Conference on Artificial Intelligence and Statistics*, pp. 127–135. PMLR, 2021.

Jiaqi Yang, Qi Lei, Jason D Lee, and Simon S Du. Nearly minimax algorithms for linear bandits with shared representation. *arXiv preprint arXiv:2203.15664*, 2022a.

Ruihan Yang, Huazhe Xu, Yi Wu, and Xiaolong Wang. Multi-task reinforcement learning with soft modularization. *Advances in Neural Information Processing Systems*, 33:4767–4777, 2020.

Zhengyu Yang, Kan Ren, Xufang Luo, Minghuan Liu, Weiqing Liu, Jiang Bian, Weinan Zhang, and Dongsheng Li. Towards applicable reinforcement learning: Improving the generalization and sample efficiency with policy ensemble. *arXiv preprint arXiv:2205.09284*, 2022b.

Elad Yom-Tov, Guy Feraru, Mark Kozdoba, Shie Mannor, Moshe Tennenholtz, and Irit Hochberg. Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. *Journal of medical Internet research*, 19(10):e338, 2017.

Chicheng Zhang and Zhi Wang. Provably efficient multi-task reinforcement learning with model transfer. *Advances in Neural Information Processing Systems*, 34:19771–19783, 2021.

Zihan Zhang, Xiangyang Ji, and Simon Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pp. 4528–4531. PMLR, 2021.

Abilmansur Zhumabekov, Daniel May, Tianyu Zhang, Aakash Krishna GS, Omid Ardakanian, and Matthew E Taylor. Ensembling diverse policies improves generalizability of reinforcement learning algorithms in continuous control tasks. 2023.

A Related Works

Multitask RL. Many recent theoretical works have contributed to understanding the benefits of MTRL (Agarwal et al., 2022; Brunskill & Li, 2013; Calandriello et al., 2014; Cheng et al., 2022; Lu et al., 2021; Uehara et al., 2021; Yang et al., 2022a; Zhang & Wang, 2021) by exploiting the shared structures across tasks. An earlier line of works (Brunskill & Li, 2013) assumes that tasks are clustered and the algorithm adaptively learns the identity of each task, which allows it to pool observations. For linear Markov Decision Process (MDP) settings (Jin et al., 2020b), Lu et al. (Lu et al., 2021) shows a bound on the sub-optimality of the learned policy by assuming a full-rank least-square value iteration weight matrix from source tasks. Agarwal et al. (Agarwal et al., 2022) makes a different assumption that the target transition probability is a linear combination of the source ones, and the feature extractor is shared by all the tasks. Our work differs from all these works as we focus on the reduced complexity of exploration design.

Curriculum learning. Curriculum learning refers to adaptively selecting tasks in a specific order to improve the learning performance (Bengio et al., 2009) under a multitask learning setting. Numerous studies have demonstrated improved performance in different applications (Jiang et al., 2015; Pentina et al., 2015; Graves et al., 2017; Wang et al., 2021). However, theoretical understanding of curriculum learning remains limited. Xu & Tewari (2022) study the statistical benefits of curriculum learning under Supervised Learning setting. For RL, Li et al. (2022b) makes a first step towards the understanding of sample complexity gains of curriculum learning without an explicit exploration bonus, which is a similar statement as we make in this paper. However, their results are under strong assumptions, such as prior knowledge on the curriculum and a specific contextual RL setting with Lipschitz reward functions. This work can be seen as a more comprehensive framework of such benefits, where we discuss general MDPs with function approximation.

Myopic exploration. Myopic exploration, characterized by its ease of implementation and effectiveness in many problems (Kalashnikov et al., 2018; Mnih et al., 2015), is the most commonly used exploration strategy. Many theory works (Dabney et al., 2020; Dann et al., 2022; Liu & Brunskill, 2018; Simchowitz & Foster, 2020) have discussed the conditions, under which myopic exploration is efficient. However, all these studies consider a single MDP and require strong conditions on the underlying environment. Our paper closely follows Dann et al. (2022) where they define Myopic Exploration Gap. An MDP with low Myopic Exploration Gap can be efficiently learned by exploration exploration.

B A Formal Discussion on Curriculum Learning

We formally discuss that how our theory could provide a potential explanation on the success of curriculum learning in RL (Narvekar et al., 2020). Although Algorithm 1 does not explicitly implement curriculum learning by ordering tasks, we argue that if any curriculum learn leads to polynomial sample complexity $\mathcal{C}$, then Algorithm 1 has $|\mathcal{M}|^2\mathcal{C}$ sample complexity. We denote a curricula by $((M_i, T_i))_{i=1}^T$ and an online algorithm that learns through the curricula interacts with M_i for T_i rounds by rolling out trajectories with the estimated optimal policy of M_{i-1} with epsilon greedy. This curricula is implicitly included in Algorithm 1 with $|\mathcal{M}|\sum_i T_i$ rounds. To see this, let us say in phase i, the algorithm has mastered all tasks $M_1, \ldots, M_{i-1}$. Then by running Algorithm 1 $|\mathcal{M}|^2 T_i$ rounds, we will roll out T_i trajectories on M_i using the exploratory policy from M_{i-1} on average, which reflects the schedules from curricula. This means that the sample-complexity of Algorithm 1 provides an upper bound on the sample complexity of underlying optimal curricula and in this way our theory provides some insights on the success of the curriculum learning.

C Comparing Single-task MEG and Multitask MEG

Proposition 1. *Let $\mathcal{M}$ be any set of MDPs and $\mathcal{F}$ be any function class. We have that $\alpha(f, \mathcal{M}, \mathcal{F}) \geq \alpha(f_M, \{M\}, \mathcal{F})/\sqrt{|\mathcal{M}|}$ for all $f \in (\mathcal{F})^{\otimes|\mathcal{M}|}$ and $M \in \mathcal{M}$.*

Proof. The proof is straightforward from the definition of multitask MEG. For any MDP M and any $f \in \mathcal{F}^{\otimes|\mathcal{M}|}$, $\alpha(f_M, \{M\}, \mathcal{F})$ is the value to the following optimization problem

$$\sup_{\tilde{f}\in\mathcal{F}, c\geq 1} \frac{1}{\sqrt{c}}(V_{1,M}^{\tilde{f}} - V_{1,M}^{f_M}), \text{ s.t. for all } f^{'} \in \mathcal{F} \text{ and } h \in [H],$$

$$\mathbb{E}_{\pi^{\tilde{f}}}^{M}[(\mathcal{E}_h^M f')(s_h, a_h)]^2 \leq c\mathbb{E}_{\text{expl}(f)}^{M}[(\mathcal{E}_h^M f')(s_h, a_h)]^2$$
$$\mathbb{E}_{\pi^{f_M}}^{M}[(\mathcal{E}_h^M f')(s_h, a_h)]^2 \leq c\mathbb{E}_{\text{expl}(f)}^{M}[(\mathcal{E}_h^M f')(s_h, a_h)]^2.$$

By choosing c in Definition 3 by $c|\mathcal{M}|$, and f' by the same f' that attains the maximization in Single-task MEG, we have $\alpha(f, \mathcal{M}, \mathcal{F}) \geq \alpha(f_M, \{M\}, \mathcal{F})/\sqrt{|\mathcal{M}|}$ □

D Efficient Myopic Exploration for Deterministic MDP with known Curriculum

In light of the intrinsic connection between Algorithm 1 and curriculum learning. We present an interesting results for curriculum learning showing that any deterministic MDP can be efficiently learned through myopic exploration when a proper curriculum is given.

Proposition 4. *For any deterministic MDP M, with sparse reward, there exists a sequence of deterministic MDPs $M_1, M_2, \ldots, M_H$, such that the following learning process returns a optimal policy for M:*

1. *Initialize π_0 by a random policy.*
2. *For $t = 1, \ldots, n$, follow π_{t-1} with an ϵ-greedy exploration to collect $4At\log(H/\delta)$ trajectories denoted by $\mathcal{D}_t$. Compute the optimal policy π_t from the model learned by $\mathcal{D}_t$.*
3. *Output π_H.*

The above procedure will end in $O(AH^2\log(H/\delta))$ episodes and with a probability at least $1-\delta$, π_H is the optimal policy for M.

Proof. We construct the sequence in the following manner. Let the optimal policy for an MDP M be π_M^*. Let the trajectory induced by π_M^* be $\{s_0^*, a_0^*, \ldots, s_H^*, a_H^*\}$. The MDP M receives a positive reward only when it reaches s_H^*. Without loss of generality, we assume that M is initialized at a fixed state s_0. We choose M_n such that

$$R_{M_i}(s,a) = \mathbb{1}(P_{M_n}(s,a) = s_i^*).$$

Furthermore, we set

$$P_{M_i}(s_i^*|s_i^*, a) = 1 \ \forall a \in \mathcal{A}$$

and

$$P_{M_i} = P_M$$

otherwise.

This ensures that any policy that reaches s_i^* on the i'th step is an optimal policy.

We first provide an upper bound on the expected number of episodes for finding an optimal policy using the above algorithm for M_i.

Fix $2 \leq i \leq H$. Let $\epsilon = \frac{1}{i}$. Define $k = |A|$. Then

Then the probability for reaching optimal reward for M_i is less than or equal to

$$(1-\frac{1}{i})^{i-1}(\frac{1}{ki})$$

So the expected number of episodes to reach this optimal reward (and thus find an optimal policy) is

$$\frac{1}{(1-\frac{1}{i})^{i-1}(\frac{1}{ki})} = (i-1)k(\frac{i}{i-1})^i \leq 4k(i-1)$$

since $i \geq 2$ and $(\frac{i}{i-1})^i$ is decreasing. By Chebyshev's inequality, a successful visit can be found in $4k(i-1)\log(H/\delta)$ with a probability at least $1-\delta/H$.

The expected total number of episodes for the all the MDP's is therefore

$$\sum_{j=2}^{H} 4k(j-1)\log(H/\delta) \leq \frac{H}{2}(4kH)\log(H/\delta)$$

which is $O(kH^2\log(H/\delta))$. □

E Generic Upper Bound for Sample Complexity

In this section, we prove the generic upper bound on sample complexity in Theorem 1. We first prove Lemma 2, which holds under the same condition of Theorem 1.

Lemma 2. *Consider a multitask RL problem with MDP set $\mathcal{M}$ and value function class $\mathcal{F}$ such that $\mathcal{M}$ is $(\tilde{\alpha}, \tilde{c})$-diverse. Then Algorithm 1 running T rounds with exploration function* expl *satisfies that with a probability at least $1-\delta$, the total number of rounds, where there exists an MDP M, such that $\pi^{\hat{f}_{t,M}}$ is β-suboptimal for M, can be upper bounded by*

$$\mathcal{O}\left(|\mathcal{M}|H^2 d_{BE}(\mathcal{F}, \Pi_{\mathcal{F}}, 1/\sqrt{T})\frac{\ln \tilde{c}(\beta)}{\tilde{\alpha}(\beta)^2}\ln\left(\frac{N'_{\mathcal{F}}\left(T^{-1}\right)\ln T}{\delta}\right)\right),$$

where $\dim_{BE}(\mathcal{F}, \Pi_{\mathcal{F}}, 1/\sqrt{T})$ is the Bellman-Eluder dimension of class $\mathcal{F}$ and $N'_{\mathcal{F}}(\rho) = \sum_{h=1}^{H-1} N_{\mathcal{F}_h}(\rho)N_{\mathcal{F}_{h+1}}(\rho)$.

Proof. Let us partition $\mathcal{F}_\beta$ into $\mathcal{F}_\beta = \{\mathcal{F}_{M,i}\}_{M\in\mathcal{M}, i\in[i_{\max}]}$ such that

$$\mathcal{F}_{M,i} := \{f \in \mathcal{F}_\beta : c(f, \mathcal{M}, \mathcal{F}) \in [e^{i-1}, e^i] \text{ and } M(f, \mathcal{M}, \mathcal{F}) = M\}.$$

Furthermore, denote $(\hat{f}_{t,M})_{M\in\mathcal{M}}$ by $\hat{f}_t$. We define $\mathcal{K}_{M,i,t} = \{\tau \in [t], \hat{f}_\tau \in \mathcal{F}_{M,i}\}$. The proof in Dann et al. (2022) can be seen as bounding the sum of $\mathcal{K}_{M,i,t}$ for a specific M, while apply the same bound for each M, which leads to an extra $|\mathcal{M}|$ factor.

Lemma 3. *Under the same condition in Theorem 1 and the above definition, we have*

$$|\mathcal{K}_{M,i,T}| \leq \mathcal{O}\left(\frac{H^2 d_{BE}(\mathcal{F}, \Pi_{\mathcal{F}}, 1/\sqrt{T})}{\tilde{\alpha}(\beta)^2}\ln\frac{N'_{\mathcal{F}}(1/T)\ln(T)}{\delta}\right).$$

Proof. In the following proof, we fix an MDP M and without further specification, the policies or rewards are with respect to the specific M. We study all the steps $t \in \mathcal{K}_{M,i,T}$.

For each $t \in \mathcal{K}_{M,i,T}$,

1. Recall that $\hat{\pi}_t$ is the mixture of exploration policy for all the MDPs: $\text{Mixture}(\{\text{expl}(\hat{\pi}_{t,M'})\}_{M'\in\mathcal{M}})$;
2. Define π'_t as the improved policy that attains the maximum in the multitask myopic exploration gap for $\hat{f}_t$ in Definition 3.

Note that π'_t is a policy for M since $t \in \mathcal{K}_{M,i,t}$. A key step in our proof is to upper bound the difference between the value of the current policy and the value of π'_t. By Lemma 4, The total difference in return between the greedy policies and the improved policies can be bounded by

$$\sum_{t\in\mathcal{K}_{M,i,T}}(V_{1,M}^{\pi'_t}(s_1) - V_{1,M}^{\hat{\pi}_{t,M}}(s_1)) \leq \sum_{t\in\mathcal{K}_{M,i,T}}\sum_{h=1}^{H}\mathbb{E}^M_{\hat{\pi}_{t,M}}[(\mathcal{E}_h^M \hat{f}_{t,M})(s_h,a_h)] - \sum_{t\in\mathcal{K}_{M,i,T}}\sum_{h=1}^{H}\mathbb{E}^M_{\pi'_t}[(\mathcal{E}_h^M \hat{f}_{t,M})(s_h,a_h)], \tag{3}$$

where the exportation is taken over the randomness of the trajectory sampled for MDP M.

Under the completeness assumption in Assumption 1, by Lemma 5 we show that with a probability $1-\delta$ for all $(h,t) \in [H] \times [T]$,

$$\sum_{\tau=1}^{t-1} \mathbb{E}_{\hat{\pi}_\tau}^{M} \left[(\mathcal{E}_h f_{t,M})(s_h,a_h)\right]^2 \le 3\frac{t-1}{T} + 176 \ln \frac{6N'_{\mathcal{F}}(1/T)\ln(2t)}{\delta}.$$

We consider only the event, where this condition holds. Since $c(\hat{f}_t, \mathcal{M}, \mathcal{F}) \le e^i$ for all $t \in \mathcal{K}_{M,i,T}$, by Definition 3 we bound

$$\begin{aligned}
&\sum_{\tau \in \mathcal{K}_{M,i,t-1}} \mathbb{E}_{\pi'_\tau}^{M} \left[\left(\mathcal{E}_h^M \hat{f}_{t,M}\right)(s_h,a_h)\right]^2 \\
&\le \sum_{\tau \in [t-1]} \mathbb{E}_{\pi'_\tau}^{M} \left[\left(\mathcal{E}_h^M \hat{f}_{t,M}\right)(s_h,a_h)\right]^2 \\
&\le e^i \sum_{\tau \in [t-1]} \mathbb{E}_{\hat{\pi}_\tau}^{M} \left[\left(\mathcal{E}_h^M \hat{f}_{t,M}\right)(s_h,a_h)\right]^2 \\
&\le 179 e^i \ln \frac{6N'_{\mathcal{F}}(1/T)\ln(2t)}{\delta}.
\end{aligned}$$

Combined with the distributional Eluder dimension machinery in Lemma 7, this implies that

$$\sum_{t \in \mathcal{K}_{M,i,T}} \left|\mathbb{E}_{\pi'_t}^{M}\left[\left(\mathcal{E}_h^M \hat{f}_{t,M}\right)(s_h,a_h)\right]\right| \le \mathcal{O}\left(\sqrt{e^i d_{BE}(\mathcal{F}, \Pi_{\mathcal{F}}, 1/\sqrt{T}) \ln \frac{N'_{\mathcal{F}}(1/T)\ln(T)}{\delta} \left|\mathcal{K}_{M,i,T}\right|} \right.$$
$$\left. + \min\left\{\left|\mathcal{K}_{M,i,T}\right|, d_{BE}(\mathcal{F}, \Pi_{\mathcal{F}}, 1/\sqrt{T})\right\}\right),$$

Note that we can derive the same upper-bound for $\sum_{t \in \mathcal{K}_{M,i,T}} \left|\mathbb{E}_{\pi_t}^{M}\left[\left(\mathcal{E}_h^M \hat{f}_{t,M}\right)(s_h,a_h)\right]\right|$. Then plugging the above two bounds into Equation (3), we obtain

$$\sum_{t \in \mathcal{K}_{M,i,T}} \left(V_{1,M}^{\pi'_t}(s_1) - V_{1,M}^{\hat{\pi}_{t,M}}(s_1)\right) \le \mathcal{O}\left(\sqrt{e^i H^2 d_{BE}(\mathcal{F}, \Pi_{\mathcal{F}}, 1/\sqrt{T}) \ln \frac{N'_{\mathcal{F}}(1/T)\ln(T)}{\delta} \left|\mathcal{K}_{M,i,T}\right|} + H d\left(\mathcal{F}'_i\right)\right).$$

By the definition of myopic exploration gap, we lower bound the LHS by

$$\sum_{t \in \mathcal{K}_{M,i,T}} \left(V_{1,M}^{\pi'_t}(s_1) - V_{1,M}^{\hat{\pi}_{t,M}}(s_1)\right) \ge \left|\mathcal{K}_{M,i,T}\right| \sqrt{e^{i-1}} \alpha_\beta.$$

Combining both bounds and rearranging yields

$$\left|\mathcal{K}_{M,i,T}\right| \le \mathcal{O}\left(\frac{H^2 d_{BE}(\mathcal{F}, \Pi_{\mathcal{F}}, 1/\sqrt{T})}{\alpha_\beta^2} \ln \frac{N'_{\mathcal{F}}(1/T)\ln(T)}{\delta}\right).$$

□

Summing over $M \in \mathcal{M}$ and $i \le i_{max} < \ln \tilde{c}(\beta)$, we conclude Lemma 2. □

To convert Lemma 2 into a sample complexity bound in Theorem 1, we show that for all M, $\hat{\pi}_M = \text{Mixture}(\{\pi^{\hat{f}_{t,M}}\})$ is β-optimal for M.

$$\max_\pi V_{1,M}^{\pi} - V_{1,M}^{\hat{\pi}_M}(s_1) = \mathcal{O}\left(\frac{\beta |\mathcal{M}| H^2 d_{BE}(\mathcal{F}, \Pi_{\mathcal{F}}, 1/\sqrt{T}) \frac{\ln \tilde{c}(\beta)}{\tilde{\alpha}(\beta)^2} \ln\left(\frac{N'_{\mathcal{F}}\left(T^{-1}\right)\ln T}{\delta}\right)}{T}\right).$$

To have the above suboptimality controlled at the level β, we will need

$$\mathcal{O}\left(\frac{\beta|\mathcal{M}|H^2 d_{BE}(\mathcal{F}, \Pi_{\mathcal{F}}, 1/\sqrt{T}) \frac{\ln \tilde{c}(\beta)}{\tilde{\alpha}(\beta)^2} \ln\left(\frac{N'_{\mathcal{F}}\left(T^{-1}\right) \ln T}{\delta}\right)}{T}\right) = \beta.$$

Assume that $d_{BE}(\mathcal{F}, \Pi_{\mathcal{F}}, \rho) = \mathcal{O}(d_{\mathcal{F}}^{\text{BE}} \log(1/\rho))$ and $\log(N'(\mathcal{F})(\rho)) = \mathcal{O}(d_{\mathcal{F}}^{\text{cover}} \log(1/\rho))$ by ignoring other factors, which holds for most regular function classes including tabular and linear classes (Russo & Van Roy, 2013), we have

$$T = \mathcal{O}\left(|\mathcal{M}|H^2 d_{\mathcal{F}}^{BE} d_{\mathcal{F}}^{\text{cover}} \frac{\ln \tilde{c}(\beta)}{\tilde{\alpha}(\beta)^2} \ln \frac{1}{\delta} \ln(1/\rho)\right),$$

where $\rho^{-1} = \mathcal{O}\left(|\mathcal{M}|H^2 d_{\mathcal{F}}^{\text{BE}} d_{\mathcal{F}}^{\text{cover}} \frac{\ln \tilde{c}(\beta)}{\tilde{\alpha}(\beta)^2} \ln(1/\delta)\right)$. The final bound has an extra $|\mathcal{M}|$ dependence because we execute a policy for each MDP in a round.

E.1 Technical Lemmas

Lemma 4 (**Lemma 3** (Dann et al., 2022)). *For any MDP M, let $f = \{f_h\}_{h\in[H]}$ with $f_h : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}$ and π^f is the greedy policy of f. Then for any policy π',*

$$V_1^{\pi'}(s_1) - V_1^{\pi^f}(s_1) \leq \sum_{h=1}^{H} \mathbb{E}_{\pi^f}^{M}\left[\left(\mathcal{E}_h f\right)(s_h, a_h)\right] - \sum_{h=1}^{H} \mathbb{E}_{\pi'}^{M}\left[\left(\mathcal{E}_h f\right)(s_h, a_h)\right].$$

Lemma 5 (Modified from **Lemma 4** (Dann et al., 2022)). *Consider a sequence of policies $(\pi_t)_{t\in\mathbb{N}}$. At step τ, we collect one episode using $\hat{\pi}_\tau$ and define $\hat{f}_\tau$ as the fitted Q-learning estimator up to step t over the function class $\mathcal{F} = \{\mathcal{F}\}_{h\in[H]}$. Let $\rho \in \mathbb{R}^+$ and $\delta \in (0,1)$. If $\mathcal{F}$ satisfies Assumption 1, then with a probability at least $1-\delta$, for all $h \in [H]$ and $t \in \mathbb{N}$,*

$$\sum_{\tau=1}^{t-1} \mathbb{E}_{\hat{\pi}_\tau}^{M}[(\mathcal{E}_h \hat{f}_t)(s_h, a_h)]^2 \leq 3\rho t + 176 \ln \frac{6N'_{\mathcal{F}}(\rho)\ln(2t)}{\delta},$$

where $N'_{\mathcal{F}}(\rho) = \sum_{h=1}^{H} N_{\mathcal{F}_h}(\rho) N_{\mathcal{F}_{h+1}}(\rho)$ is the sum of ℓ_∞ covering number of $\mathcal{F}_h \times \mathcal{F}_{h+1}$ w.r.t. radius $\rho > 0$.

Proof. The only difference between our statement and the statement in Dann et al. (2022) is that they consider $\hat{\pi}_\tau = \text{expl}(\hat{f}_\tau)$, while this statement holds for any data-collecting policy $\hat{\pi}_\tau$. To show this, we go through the complete proof here.

Consider a fixed $t \in \mathbb{N}$, $h \in [H]$ and $f = \{f_h, f_{h+1}\}$ with $f_h \in \mathcal{F}_h$, $f_{h+1} \in \mathcal{F}_{h+1}$. Let $(x_{t,h}, a_{t,h}, r_{t,h})_{t\in\mathbb{N}, h\in[H]}$ be the collected trajectory in $[t]$. Then

$$\begin{aligned}
Y_{t,h}(f) &= \left(f_h\left(x_{t,h}, a_{t,h}\right) - r_{t,h} - \max_{a'} f_{h+1}\left(x_{t,h+1}, a'\right)\right)^2 - \left(\left(\mathcal{T}_h f_{h+1}\right)\left(x_{t,h}, a_{t,h}\right) - r_{t,h} - \max_{a'} f_{h+1}\left(x_{t,h+1}, a'\right)\right)^2 \\
&= \left(f_h\left(x_{t,h}, a_{t,h}\right) - \left(\mathcal{T}_h f_{h+1}\right)\left(x_{t,h}, a_{t,h}\right)\right) \\
&\quad \times \left(f_h\left(x_{t,h}, a_{t,h}\right) + \left(\mathcal{T}_h f_{h+1}\right)\left(x_{t,h}, a_{t,h}\right) - 2r_{t,h} - 2\max_{a'} f_{h+1}\left(x_{t,h+1}, a'\right)\right).
\end{aligned}$$

Let $\mathfrak{F}_t$ be the σ-algebra under which all the random variables in the first $t-1$ episodes are measurable. Note that $|Y_{t,h}(f)| \leq 4$ almost surely and the conditional expectation of $Y_{y,h}(f)$ can be written as

$$\mathbb{E}\left[Y_{t,h}(f) \mid \mathfrak{F}_t\right] = \mathbb{E}\left[\mathbb{E}\left[Y_{t,h}(f) \mid \mathfrak{F}_t, x_{t,h}, a_{t,h}\right] \mid \mathfrak{F}_t\right] = \mathbb{E}_{\pi_t}[\left(f_h - \mathcal{T}_h f_{h+1}\right)\left(x_h, a_h\right)^2].$$

The variance can be bounded by

$$\operatorname{Var}\left[Y_{t,h}(f) \mid \mathfrak{F}_t\right] \leq \mathbb{E}\left[Y_{t,h}(f)^2 \mid \mathfrak{F}_t\right] \leq 16\mathbb{E}\left[\left(f_h - \mathcal{T}_h f_{h+1}\right)\left(x_{t,h}, a_{t,h}\right)^2 \mid \mathfrak{F}_t\right] = 16\mathbb{E}\left[Y_{t,h}(f) \mid \mathfrak{F}_t\right],$$

where we used the fact that $|f_h(x_{t,h}, a_{t,h}) + (\mathcal{T}_h f_{h+1})(x_{t,h}, a_{t,h}) - 2r_{t,h} - 2\max_{a'} f_{h+1}(x_{h+1}, a')| \leq 4$ almost surely. Applying Lemma 6 to the random variable $Y_{t,h}(f)$, we have that with probability at least $1-\delta$, for all $t \in \mathbb{N}$,

$$\begin{aligned}
\sum_{i=1}^{t} \mathbb{E}\left[Y_{i,h}(f) \mid \mathfrak{F}_i\right] &\leq 2A_t \sqrt{\sum_{i=1}^{t} \operatorname{Var}\left[Y_{i,h}(f) \mid \mathfrak{F}_i\right] + 12A_t^2} + \sum_{i=1}^{t} Y_{i,h}(f) \\
&\leq 8A_t \sqrt{\sum_{i=1}^{t} \mathbb{E}\left[Y_{i,h}(f) \mid \mathfrak{F}_i\right] + 12A_t^2} + \sum_{i=1}^{t} Y_{i,h}(f),
\end{aligned}$$

where $A_t = \sqrt{2\ln\ln(2t) + \ln(6/\delta)}$. Using AM-GM inequality and rearranging terms in the above we have

$$\sum_{i=1}^{t} \mathbb{E}\left[Y_{i,h}(f) \mid \mathfrak{F}_i\right] \leq 2\sum_{i=1}^{t} Y_{i,h}(f) + 88A_t^2 \leq 2\sum_{i=1}^{t} Y_{i,h}(f) + 176 \ln \frac{6\ln(2t)}{\delta}.$$

Let $\mathcal{Z}_{\rho,h}$ be a ρ-cover of $\mathcal{F}_h \times \mathcal{F}_{h+1}$. Now taking a union bound over all $\phi_h \in \mathcal{Z}_{\rho,h}$ and $h \in [H]$, we obtain that with probability at least $1-\delta$ for all ϕ_h and $h \in [H]$

$$\sum_{i=1}^{t} \mathbb{E}\left[Y_{i,h}\left(\phi_h\right) \mid \mathfrak{F}_i\right] \leq 2\sum_{i=1}^{t} Y_{i,h}\left(\phi_h\right) + 176 \ln \frac{6N'_{\mathcal{F}}(\rho)\ln(2t)}{\delta}.$$

This implies that with probability at least $1-\delta$ for all $f = \{f_h, f_{h+1}\} \in \mathcal{F}_h \times \mathcal{F}_{h+1}$ and $h \in [H]$,

$$\sum_{i=1}^{t} \mathbb{E}\left[Y_{i,h}(f) \mid \mathfrak{F}_i\right] \leq 2\sum_{i=1}^{t} Y_{i,h}(f) + 3\rho(t-1) + 176 \ln \frac{6N'_{\mathcal{F}}(\rho)\ln(2t)}{\delta}.$$

Let $\hat{f}_{t,h}$ be the h-th component of the function $\hat{f}_t$. The above inequality holds in particular for $f = \{\hat{f}_{t,h}, \hat{f}_{t,h+1}\}$ for all $t \in \mathbb{N}$. Finally, we have

$$\begin{aligned}
\sum_{i=1}^{t-1} Y_{i,h}\left(\hat{f}_t\right) &= \sum_{i=1}^{t-1} \left(\hat{f}_{t,h}\left(s_{i,h}, a_{i,h}\right) - r_{i,h} - \max_{a'} \hat{f}_{t,h+1}\left(s_{i,h+1}, a'\right)\right)^2 \\
&\quad - \sum_{i=1}^{t-1} \left(\left(\mathcal{T}_h \hat{f}_{t,h+1}\right)\left(s_{i,h}, a_{i,h}\right) - r_{i,h} - \max_{a'} \hat{f}_{t,h+1}\left(s_{i,h+1}, a'\right)\right)^2 \\
&= \inf_{f' \in \mathcal{F}_h} \sum_{i=1}^{t-1} \left(f'\left(s_{i,h}, a_{i,h}\right) - r_{i,h} - \max_{a'} \hat{f}_{t,h+1}\left(s_{i,h+1}, a'\right)\right)^2 \\
&\quad - \sum_{i=1}^{t-1} \left(\left(\mathcal{T}_h \hat{f}_{t,h+1}\right)\left(s_{i,h}, a_{i,h}\right) - r_{i,h} - \max_{a'} \hat{f}_{t,h+1}\left(s_{i,h+1}, a'\right)\right)^2 \\
&\leq 0,
\end{aligned}$$

where the last inequality follows from the completeness in Assumption 1. □

Lemma 6 (Time-Uniform Freedman Inequality). *Suppose $\{X_t\}_{t=1}^{\infty}$ is a martingale difference sequence with $|X_t| \leq b$. Let*

$$\operatorname{Var}_{\ell}\left(X_{\ell}\right) = \operatorname{Var}\left(X_{\ell} \mid X_1, \cdots, X_{\ell-1}\right).$$

Let $V_t = \sum_{\ell=1}^{t} \operatorname{Var}_{\ell}(X_{\ell})$ be the sum of conditional variances of X_t. Then we have that for any $\delta' \in (0,1)$ and $t \in \mathbb{N}$

$$\mathbb{P}\left(\sum_{\ell=1}^{t} X_{\ell} > 2\sqrt{V_t}A_t + 3bA_t^2\right) \leq \delta'$$

where $A_t = \sqrt{2\ln\ln(2(\max(V_t/b^2, 1))) + \ln(6/\delta')}$.

Lemma 7 (Lemma 41 (Jin et al., 2021a)). *Given a function class Φ defined on $\mathcal{X}$ with $|\phi(x)| \leq C$ for all $(\phi, x) \in \Phi \times \mathcal{X}$ and a family of probability measures Π over $\mathcal{X}$. Suppose sequences $\{\phi_i\}_{i\in[K]} \subset \Phi$ and $\{\mu_i\}_{i\in[K]} \subset \Pi$ satisfy for all $k \in [K]$ that $\sum_{i=1}^{k-1}(\mathbb{E}_{\mu_i}[\phi_k])^2 \leq \beta$. Then for all $k \in [K]$ and* $w > 0$,

$$\sum_{t=1}^{k} |\mathbb{E}_{\mu_t}[\phi_t]| \leq O\left(\sqrt{\dim_{DE}(\Phi, \Pi, \omega)\beta k} + \min\{k, \dim_{DE}(\Phi, \Pi, \omega)\}\, C + k\omega\right).$$

Proof of Theorem 1. Denote $B = \Theta\left(|\mathcal{M}|^2 H^2 d_{\mathrm{BE}} \frac{\ln \tilde{c}(\beta)}{\tilde{\alpha}(\beta)^2 \beta} \ln\left(\frac{\bar{N}_{\mathcal{F}}(T^{-1}) \ln T}{\delta}\right)\right)$. The following Corollary transform Lemma 2 to Theorem 1, whose proof directly follows by taking $T = B/\beta$. Since at most B rounds are suboptimal according to Lemma 2, the mixing of all T policies are β-optimal. This leads to a sample complexity

$$\mathcal{C}(\tilde{\alpha}, \tilde{c}) = \Theta\left(|\mathcal{M}|^2 H^2 d_{\mathrm{BE}} \frac{\ln \tilde{c}(\beta)}{\tilde{\alpha}(\beta)^2 \beta} \ln\left(\frac{\bar{N}_{\mathcal{F}}\left(T^{-1}\right) \ln T}{\delta}\right)\right)$$

F Omitted Proofs for Case Studies

F.1 Linear MDP case

Note that in this section, we use $\mathbb{E}_\pi$ for the expectation over transition w.r.t a policy π.

Lemma 8. *Let $\mathcal{F}$ be the function class in Proposition 3. For any policy π such that $\lambda_{min}(\Phi_h^\pi) \geq \underline{\lambda}$, then for any policy π' and $f' \in \mathcal{F}$, $\mathbb{E}_{\pi'}\left[\left(\mathcal{E}_h^2 f'\right)(s_h, a_h)\right] \leq \mathbb{E}_\pi\left[\left(\mathcal{E}_h^2 f'\right)(s_h, a_h)\right]/\underline{\lambda}$.*

Proof. Recall that $\Phi_h^\pi := \mathbb{E}_\pi \phi_h(s_h, a_h)\phi_h(s_h, a_h)^\top$.

We derive the Bellman error term using the fact that f' is a linear function and the transitions admit the linear function as well. For any policy π, we have

$$\begin{aligned}
&\mathbb{E}_\pi[(\mathcal{E}_h^2 f')(s_h, a_h)] \\
&= \mathbb{E}_\pi\left[\left(f_h'(s_h, a_h) - \phi_h(s_h, a_h)^\top \theta_h - \max_{a'} \mathbb{E}_{s_{h+1}}[f_{h+1}'(s_{h+1}, a') \mid s_h, a_h]\right)^2\right] \\
&= \mathbb{E}_\pi\left[\left(\phi_h(s_h, a_h)^\top w_h - \phi_h(s_h, a_h)^\top \theta_h - \max_{a'} \mathbb{E}_{s_{h+1}}[\phi_{h+1}(s_{h+1}, a')^\top w_{h+1} \mid s_h, a_h]\right)^2\right] \\
&= \mathbb{E}_\pi\left[\left(\phi_h(s_h, a_h)^\top w_h - \phi_h(s_h, a_h)^\top \theta_h - \phi_h(s_h, a_h)^\top \int_{s'} \phi_{h+1}(s', \pi_{h+1}^{f'}(s'))^\top w_{h+1}\mu_h(s')ds'\right)^2\right] \\
&= \mathbb{E}_\pi\left[\left(\phi_h(s_h, a_h)^\top (w_h - \theta_h - w_{h+1}')\right)^2\right] \\
&= (w_h - \theta_h - w_{h+1}')^\top \mathbb{E}_\pi\left[\phi_h(s_h, a_h)\phi_h(s_h, a_h)^\top\right](w_h - \theta_h - w_{h+1}')
\end{aligned}$$

where $w_{h+1}' = \int_{s'} \phi_{h+1}(s', \pi_{h+1}^{f'}(s'))^\top w_{h+1}\mu_h(s')ds'$. Since by the assumption in Definition 6 that $\|\phi_h(s, a)\| \leq 1$ for any s, a, we have $\Phi_h^{\pi'} \prec I$. The result follow by the condition that $\lambda_{\min}(\Phi_h^\pi) \geq \underline{\lambda}$. □

Lemma 1. *Fix a step h. Let $\{M_{i,h}\}_{i\in[d]}$ be the d MDPs such that $\theta_{h,M_{i,h}} = e_i$ as in Definition 6. Let $\{\pi_i\}_{i=1}^d$ be d policies such that π_i is a β-optimal policy for $M_{i,h}$ with $\beta < b_1/2$. Let $\tilde{\pi} = \mathrm{Mixture}(\{\mathrm{expl}(\pi_i)\}_{i=1}^d)$. Then for any $\nu \in \mathbb{S}^{d-1}$, we have $\lambda_{\min}(\Phi_{h+1}^{\tilde{\pi}}) \geq \epsilon_h \prod_{h'=1}^{h-1}(1 - \epsilon_{h'})b_1^2/(2dA)$.*

Proof. Let π be any stationary policy and recall that Π is the set of all the stationary policies. We denote $A_h^\pi(s') \sim \pi_h(s')$ by the random variable for the action sampled at the step h using policy π given the state is s'. Let $\phi_h^\pi := \mathbb{E}_\pi \phi_h(s_h, a_h)$.

We further define

$$a^{\nu}_{h+1}(s) := \arg\max_{a\in\mathcal{A}}[\nu^{\top}\phi_{h+1}(s,a)\phi_{h+1}(s,a)^{\top}\nu].$$

Lower bound the following quadratic term for any unit vector $\nu \in \mathbb{R}^d$,

$$\begin{aligned}
&\max_{\pi\in\Pi}\nu^{\top}\Phi^{\pi}_{h+1}\nu \\
&= \max_{\pi\in\Pi}\mathbb{E}_{\pi}\left[\int_{s'}\nu^{\top}\phi_{h+1}(s', A^{\pi}_{h+1}(s'))\phi_{h+1}(s', A^{\pi}_{h+1}(s'))^{\top}\nu\mu_h(s')^{\top}\phi_h(s_h,a_h)ds'\right] \\
&= \max_{\pi}\mathbb{E}_{\pi}[\phi_h(s_h,a_h)^{\top}]\left(\int_{s'}\nu^{\top}\phi_{h+1}(s', a^{\nu}_{h+1}(s'))\phi_{h+1}(s', a^{\nu}_{h+1}(s'))^{\top}\nu\mu_h(s')ds'\right) \\
&= \max_{\pi\in\Pi}(\phi^{\pi}_h)^{\top}w^{\nu}_{h+1}.
\end{aligned}$$

where we let $w^{\nu}_{h+1} := \int_{s'}\nu^{\top}\phi_{h+1}(s', a^{\nu}_{h+1}(s'))\phi_{h+1}(s', a^{\nu}_{h+1}(s'))^{\top}\nu\mu_h(s')^{\top}ds'$.

By Assumption 2, we have $\max_{\pi\in\Pi}\mathbb{E}_{\pi}[\phi_h(s_h,a_h)^{\top}]w^{\nu}_{h+1} \geq b_1^2$.

For the mixture policy $\tilde{\pi}$ defined in our lemma,

$$\begin{aligned}
\nu^{\top}\Phi^{\tilde{\pi}}_{h+1}\nu &= \frac{1}{d}\sum_{i=1}^{d}\mathbb{E}_{\text{expl}(\pi_i)}[\nu^{\top}\phi_{h+1}(s_{h+1},a_{h+1})\phi_{h+1}(s_{h+1},a_{h+1})^{\top}\nu] \\
&\geq \frac{\epsilon_h\prod_{h'=1}^{h-1}(1-\epsilon_{h'})}{Ad}\sum_{i=1}^{d}(\phi^{\pi_i}_h)^{\top}w^{\nu}_{h+1}. \qquad (4)
\end{aligned}$$

Since π_i is a $b_1/2$-optimal policy for MDP $M_{i,h}$ and again by Assumption 2, we have

$$\theta^{\top}_{h,M_{i,h}}\phi^{\pi_i}_h \geq \frac{1}{2}\max_{\pi\in\Pi}\theta^{\top}_{h,M_{i,h}}\phi^{\pi}_h. \qquad (5)$$

For any vector $\nu \in \mathbb{R}^d$, let $[\nu]_i$ be the i-th dimension of the vector. Note that $\theta_{h,M_{i,h}} = e_i$, (5) indicates $[\phi^{\pi_i}_h]_i \geq \frac{1}{2}\max_{\pi}[\phi^{\pi}_h]_i$.

Combining the inequality (5) with (4), we have

$$\begin{aligned}
\nu^{\top}\Phi^{\tilde{\pi}}_{h+1}\nu &= \frac{\epsilon_h\prod_{h'=1}^{h-1}(1-\epsilon_{h'})}{dA}\sum_{i=1}^{d}\sum_{j=1}^{d}[\phi^{\pi_i}_h]_j[w^{\nu}_{h+1}]_j \\
&\geq \frac{\epsilon_h\prod_{h'=1}^{h-1}(1-\epsilon_{h'})}{dA}\sum_{i=1}^{d}[\phi^{\pi_i}_h]_i[w^{\nu}_{h+1}]_i \\
&\geq \frac{\epsilon_h\prod_{h'=1}^{h-1}(1-\epsilon_{h'})}{dA}\sum_{i=1}^{d}\max_{\pi}[\phi^{\pi}_h]_i[w^{\nu}_{h+1}]_i \\
&\geq \frac{\epsilon_h\prod_{h'=1}^{h-1}(1-\epsilon_{h'})}{2dA}\max_{\pi}(\phi^{\pi}_h)^{\top}w^{\nu}_{h+1} \\
&\geq \frac{\epsilon_h\prod_{h'=1}^{h-1}(1-\epsilon_{h'})b_1^2}{2dA}
\end{aligned}$$

□

F.2 Proof of Theorem 2

Theorem 2. *Consider $\mathcal{M}$ defined in Definition 7. With Assumption 2 holding and $\beta \leq b_1/2$, for any $f \in \mathcal{F}_{\beta}$, we have lower bound $\alpha(f,\mathcal{F},\mathcal{M}) \geq \sqrt{e\beta^2 b_1^2/(2A|\mathcal{M}|H)}$ by setting $\epsilon_h = 1/h$.*

Proof. Let h' be the smallest h, such that there exists $M_{i,h}$, $\pi^{f_{M_{i,h}}}$ is β-suboptimal. Let (i',h') be the index of the MDP that has the suboptimal policy. We show that $M_{i',h'}$ has lower bounded myopic exploration gap.

By definition, f is β-optimal for any MDP $M_{i,h'-1}$. By Lemma 1, letting $\tilde{\pi} = \text{expl}(f, \epsilon_{h'})$, we have

$$\nu^\top \Phi_{h'+1}^{\tilde{\pi}} \nu \geq \frac{\epsilon_{h'} \prod_{h''=1}^{h'-1}(1-\epsilon_{h''}) b_1^2}{2A|\mathcal{M}|}.$$

By Lemma 8, we have that the optimal value function f^* for MDP $M_{i',h'}$ satisfies that for any f'

$$\mathbb{E}^M_{\pi_{f^*}} \left[\left(\mathcal{E}_h^2 f'\right)(s_h, a_h)\right] \leq \frac{2A|\mathcal{M}|}{\epsilon_{h'} \prod_{h''=1}^{h'-1}(1-\epsilon_{h''}) b_1^2} \mathbb{E}^M_{\tilde{\pi}} \left[\left(\mathcal{E}_h^2 f'\right)(s_h, a_h)\right].$$

Thus, by Definition 3, the myopic exploration gap for f is lower bounded by

$$\beta \frac{1}{\sqrt{c}} = \beta \sqrt{\frac{\epsilon_{h'} \prod_{h''=1}^{h'-1}(1-\epsilon_{h''}) b_1^2}{2A|\mathcal{M}|}} \geq \sqrt{\frac{\beta^2 b_1^2}{2A|\mathcal{M}|eH}},$$

if we choose $\epsilon_h = 1/(h+1)$. □

F.3 Linear Quadratic Regulator

To demonstrate the generalizability of the proposed framework, we introduce another interesting setting called Linear Quadratic Regulator (LQR). LQR takes continuous state space $\mathbb{R}^{d_s}$ and action space $\mathbb{R}^{d_a}$. In the LQR system, the state $s_h \in \mathbb{R}^{d_s}$ evolves according to the following transition: $s_{h+1} = A_h s_h + B_h a_h$, where $A_h \in \mathbb{R}^{d_s \times d_s}$, $B_h \in \mathbb{R}^{d_s \times d_a}$ are unknown system matrices that are shared by all the MDPs. We denote $s_h = (s_h, a_h)$ as the state-action vector. The reward function for an MDP M takes a known quadratic form $r_{h,M}(s,a) = s^\top R^s_{h,M} s + a^\top R^a_{h,M} a$, where $R^s_{h,M} \in \mathbb{R}^{d_s \times d_s}$ and $R^a_{h,M} \in \mathbb{R}^{d_a \times d_a}$ [4].

Note that LQR is more commonly studied for the infinite-horizon setting, where stabilizing the system is a primary concern of the problem. We consider the finite-horizon setting, which alleviates the difficulties on stabilization so that we can focus our discussion on exploration. Finite-horizon LQR also allows us to remain consistent notations with the rest of the paper. A related work (Simchowitz & Foster, 2020) states that naive exploration is optimal for online LQR with a condition that the system injects a random noise onto the state observation with a full rank covariance matrix $\Sigma \succ 0$. Though this is a common assumption in LQR literature, one may notice that the analog of this assumption in the tabular MDP is that any state and action pair has a non-zero probability of visiting any other state, which makes naive exploration sample-efficient trivially. In this section, we consider a deterministic system, where naive exploration does not perform well in general.

Properties of LQR. It can be shown that the optimal actions are linear transformations of the current state (Farjadnasab & Babazadeh, 2022; Li et al., 2022a).

The optimal linear response is characterized by the discrete-time Riccati equation given by

$$P_{h,M} = A_h^\top (P_{h+1,M} - P_{h+1,M} \bar{R}_{h+1,M}^{-1} B_h^\top P_{h+1,M}) A_h + R^s_{h,M},$$

where $\bar{R}_{h+1,M} = R^a_h + B_h^\top P_{h+1,M} A_h$ and $P_{H+1} = \mathbf{0}$. Assume that the solution for the above equation is $\{P^*_{h,M}\}_{h \in [H+1]}$, then the optimal control actions takes the form

$$a_h = F^*_{h,M} s_h, \text{ where } F^*_{h,M} = -(R^s_{h,M} + B_h^\top P^*_{h,M} B_h)^{-1} B^\top P^*_{h,M} A_h.$$

and optimal value function takes the quadratic form: $V^*_{h,M}(s) = s^\top P^*_{h,M} s$ and

$$Q^*_{h,M}(x) = x^\top \begin{bmatrix} R^s_{h,M} + A_h^\top P^*_{h+1,M} A_h & A_h^\top P^*_{h+1,M} B_h \\ B_h^\top P^*_{h+1,M} A_h & R^a_{h,M} + B_h^\top P^*_{h+1,M} B_h \end{bmatrix} x.$$

This observation allows us to consider the following function approximation

$$\mathcal{F} = (\mathcal{F}_h)_{h \in [H+1]}, \text{ where each } \mathcal{F}_h = \{x \mapsto x^\top G_h x : G_h \in \mathbb{R}^{(d_s+d_a)\times(d_s+d_a)}\}.$$

The quadratic function class satisfies Bellman realiazability and completeness assumptions.

[4] Note that LQR system often consider a cost function and the goal of the agent is to minimize the cumulative cost with $R^s_{h,M}$ being semi-positive definite. We formulation this as a reward maximization problem for consistency. Thus, we consider $R^s_{h,M} \prec \mathbf{0}$

Definition 8 (Diverse LQR Task Set). *Inspired by the task construction in linear MDP case, we construct the diverse LQR set by $\mathcal{M} = \{M_{i,h}\}_{i\in[d_s],h\in[H]}$ such that these MDPs all share the same transition matrices A_h and B_h and each $M_{i,h}$ has $R^s_{h',M_{i,h}} = \mathbb{1}[h' = h]e_ie_i^\top$ and $R^a_{h',M_{i,h}} = -I$.*

Assumption 3 (Regularity parameters). *Given the task set in Definition 8, we define some constants that appears on our bound. Let $\pi^*_{i,h}$ be the optimal policy for $M_{i,h}$. Let*

$$b_4 = \max_{i,h} \mathbb{E}_{\pi^*_{i,h}} \max_{h'} \|s_{h'}\|_2, \textit{ and } b_5 = \max_{i,h} \mathbb{E}_{\pi^*_{i,h}} \max_{h'} \|a_{h'}\|_2.$$

These regularity assumption is reasonable because the optimal actions are linear transformations of states and we consider a finite-horizon MDP, with F_h^* having upper bounded eigenvalues.

Similarly to the linear MDP case, we assume that the system satisfies some visibility assumption.

Assumption 4 (Coverage Assumption). *For any $\nu \in \mathbb{R}^{d_s-1}$, there exists a policy π with $\|a_h\|_2 \leq 1$ such that*

$$\max_\pi \mathbb{E}_\pi[s_h^\top \nu] \geq b_3, \textit{ for } b_3 > 1.$$

Theorem 3. *Given Assumption 3, 4 and the diverse LQR task set in Definition 8, we have that for any $f \in \mathcal{F}_\beta$ with $\beta \leq (b_3^2 - 1)b_5^2/2$,*

$$\alpha(f, \mathcal{F}, \mathcal{M}) = \Omega\left(\frac{\max\{b_4^2, b_5^2\}b_4^2}{d_sH\max\{(b_3^2-1)b_5^2, d_s\sigma^2\}(b_3^2-1)b_5^2}\right).$$

F.4 Proof of Theorem 3

Lemma 9. *Assume that we have a set of policies $\{\pi_i\}_{i\in[d]}$ such that the i-th policy is a $(b_3^2 - 1)b_5^2/2$-optimal policy for LQR with $R^s_{h,i} = e_ie_i^\top$ and $R^a_{h,i} = -I$. Let $\tilde{\pi} = \textit{Mixture}(\mathrm{expl}\{\pi_i\})$. Then we have*

$$\lambda_{\mathit{min}}(\mathbb{E}_{\tilde{\pi}} s_{h+1}s_{h+1}^\top) \geq \frac{d_s\max\{\underline{\lambda}, d\sigma^2\}}{2\max\{b_4^2, b_5^2\}\prod_{h'=1}^{h-1}(1-\epsilon_{h'})\epsilon_h}\underline{\lambda},$$

with $\underline{\lambda} = (b_3^2 - 1)b_5^2$.

Proof. We directly analyze the state covariance matrix at the step $h+1$. Let $\eta_h \sim \mathcal{N}(0, \sigma^2)$

$$\begin{aligned}\mathbb{E}_{\tilde{\pi}} s_{h+1}s_{h+1}^\top &= \mathbb{E}_{\tilde{\pi}}(A_hs_h + B_ha_h)(A_hs_h + B_ha_h)^\top \\ &\succeq \frac{\prod_{h'=1}^{h-1}(1-\epsilon_{h'})\epsilon_h}{d_s}\sum_{i=1}^{d_s}\left(\mathbb{E}_{\pi_i}(A_hs_h + B_h\eta_h)(A_hs_h + B_h\eta_h)^\top\right) \\ &= \frac{\prod_{h'=1}^{h-1}(1-\epsilon_{h'})\epsilon_h}{d_s}\sum_{i=1}^{d_s}\left(A_h\mathbb{E}_{\pi_i}s_hs_h^\top A_h^\top + B_h\mathbb{E}\eta_h\eta_h^\top B_h^\top\right) \qquad (6)\end{aligned}$$

To proceed, we show that $\sum_{i=1}^{d_s}\mathbb{E}_{\pi_i}s_hs_h^\top \succeq \underline{\lambda}I$.

From Assumption 4, we have $\mathbb{E}_{\pi_i^*}[s_h^\top e_ie_i^\top s_h - a_ha_h^\top] \succeq b_3^2b_5^2 - b_5^2$, and by the fact that π_i is a $(b_3^2-1)b_5^2/2$-optimal policy, we have

$$\mathbb{E}_{\pi_i^*}[s_h^\top e_ie_i^\top s_h - a_ha_h^\top] \succeq (b_3^2b_5^2 - b_5^2)/2.$$

Since $\mathbb{E}_{\pi_i}a_ha_h^\top \succeq 0$, we have $\mathbb{E}_{\pi_i}[s_h^\top e_ie_i^\top s_h] \succeq (b_3^2-1)b_5^2/2$. Therefore, $\sum_{i=1}^{d_s}\mathbb{E}_{\pi_i}s_hs_h^\top \succeq \underline{\lambda}I$ with $\underline{\lambda} = (b_3^2-1)b_5^2/2$.

Combined with (6), we have

$$\mathbb{E}_{\tilde{\pi}}s_{h+1}s_{h+1}^\top \succeq \frac{\prod_{h'=1}^{h-1}(1-\epsilon_{h'})\epsilon_h}{d_s}\left(\underline{\lambda}A_hA_h^\top + d_s\sigma^2B_hB_h^\top\right).$$

Apply Assumption 4 again, for each $\nu_i = e_i, i = 1, \ldots, d_s$, there exists some policy π_i' with $\|a_h\|_2 \leq b_5$, such that $\nu_i^\top\mathbb{E}_{\pi_i'}s_{h+1}s_{h+1}^\top\nu_i \geq b_3^2b_5^2 - b_5^2$. Therefore, we have that $\sum_{i=1}^{d_s}\mathbb{E}_{\pi_i'}s_{h+1}s_{h+1}^\top \succeq (b_3^2-1)b_5^2I$

The proof is completed by

$$\begin{aligned}\sum_{i=1}^{d_s} \mathbb{E}_{\pi_i'} s_{h+1} s_{h+1}^\top &\preceq 2 \sum_{i=1}^{d_s} \left(A_h \mathbb{E}_{\pi_i'} s_h s_h^\top A_h^\top + B_h \mathbb{E}_{\pi_i'} a_h a_h^\top B_h^\top \right) \\ &\preceq 2 \sum_{i=1}^{d_s} \left(b_4^2 A_h A_h^\top + b_5^2 B_h \mathbb{E}_{\pi_i'} B_h^\top \right) \\ &\preceq \frac{2 \max\{b_4^2, b_5^2\}}{\max\{\underline{\lambda}, d\sigma^2\}} \frac{\prod_{h'=1}^{h-1}(1-\epsilon_{h'})\epsilon_h}{d_s} \mathbb{E}_{\tilde{\pi}} s_{h+1} s_{h+1}^\top .\end{aligned}$$

To complete the proof of Theorem 3, we combine Lemma 10 and Lemma 9.

□

F.5 Supporting lemmas

Lemma 10 shows that having a full rank covariance matrix for the state s_h is a sufficient condition for bounded occupancy measure.

Lemma 10. *Let $\mathcal{F}$ be the function class described above. For any policy π and h such that*

$$\lambda_{min}(\mathbb{E}_\pi[s_h s_h^\top]) \geq \underline{\lambda},$$

we have for any π' such that $\max_h \|s_h\|_2 \leq b_4$, and for any $f' \in \mathcal{F}$,

$$\mathbb{E}_{\pi'}^M \left[\left(\mathcal{E}_h^2 f' \right)(s_h, a_h) \right] \leq \frac{b_4^2}{\underline{\lambda}^2} \mathbb{E}_\pi^M \left[\left(\mathcal{E}_h^2 f' \right)(s_h, a_h) \right].$$

Proof. Lemma 11 shows that the Bellman error also takes a quadratic form of s_h.

Lemma 11. *For any $f \in \mathcal{F}$, there exists some matrix $\tilde{G}_h$ such that $(\mathcal{E}_h f)(x) = x^\top \tilde{G}_h x$.*

To complete the proof of Lemma 10, let $w_h = s_h \otimes s_h$ be the Kronecker product between s_h and itself. By Lemma 11, we can write $(\mathcal{E}_h f)(s_h) = \mathrm{Vec}(\tilde{G}_h)^\top w_h$. Again, this is an analog of the linear form we had for thee linear MDP case. Thus, we can write $(\mathcal{E}_h^2 f)(s_h) = \mathrm{Vec}(\tilde{G}_h)^\top w_h w_h^\top \mathrm{Vec}(\tilde{G}_h)$.

By Lemma 12 and the fact that $\mathbb{E}_\pi(w_h w_h^\top) = \mathbb{E}_\pi(s_h s_h^\top) \otimes \mathbb{E}_\pi(s_h s_h^\top)$, we have $\lambda_{\text{min}}(\mathbb{E}_\pi w_h w_h^\top) \geq \underline{\lambda}^2$. For any other policy π', and using the fact that $\|w_h\| \leq b_4^2$, we have

$$\mathbb{E}_{\pi'}(\mathcal{E}_h^2 f)(s_h) = \mathbb{E}_{\pi'}[\mathrm{Vec}(\tilde{G}_h)^\top w_h w_h^\top \mathrm{Vec}(\tilde{G}_h)] \leq \frac{b_4^2}{\underline{\lambda}^2} \mathbb{E}_\pi[\mathrm{Vec}(\tilde{G}_h)^\top w_h w_h^\top \mathrm{Vec}(\tilde{G}_h)] \leq \frac{b_4^2}{\underline{\lambda}^2} \mathbb{E}_\pi(\mathcal{E}_h^2 f)(s_h).$$

□

Lemma 11. *For any $f \in \mathcal{F}$, there exists some matrix $\tilde{G}_h$ such that $(\mathcal{E}_h f)(x) = x^\top \tilde{G}_h x$.*

Proof. The Bellman error of the LQR can be written as

$$(\mathcal{E}_h f)(x) = \left(x^\top G_h x - s^\top R_h^s s - a^\top R_h^a a - \max_{a' \in \mathbb{R}^{d_a}} [(A_h s + B_h a)^\top, a'^\top] G_{h+1} \begin{bmatrix} A_h s + B_h a \\ a' \end{bmatrix} \right)$$

Note that the optimal a' can be written as some linear transformation of x. Thus we can write

$$\max_{a' \in \mathbb{R}^{d_a}} [(A_h s + B_h a)^\top, a'^\top] G_{h+1} \begin{bmatrix} A_h s + B_h a \\ a' \end{bmatrix} = x^\top G' x.$$

The whole equation can be written as a quadratic form as well. □

Lemma 12. *Let $A \in \mathbb{R}^{d_1 \times d_1}$ have eigenvalues $\{\lambda_i\}_{i \in [d]}$ and $B \in \mathbb{R}^{d_2 \times d_2}$ have eigenvalues $\{\mu_i\}_{i \in [d]}$. The eigenvalues of $A \otimes B$ are $\{\lambda_i \mu_j\}_{i \in [d_1], j \in d_2}$.*

G Relaxing Visibility Assumption

G.1 Tabular Case

A simple but interesting case to study is the tabular case, where the value function class is the class of any bounded functions, i.e. $\mathcal{F}_h = \{f : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]\}$. A commonly studied family of multitask RL is the MDPs that share the same transition probability, while they have different reward functions, this problem is studied in a related literature called reward-free exploration (Jin et al., 2020a; Wang et al., 2020; Chen et al., 2022a). Specifically, (Jin et al., 2020a) propose to learn $S \times H$ sparse reward MDPs separately and generates an offline dataset, with which one can learn a near-optimal policy for any potential reward function. With a similar flavor, we show that any superset of the $S \times H$ sparse reward MDPs has low myopic exploration gap. Though the tabular case is a special case of the linear MDP case, the lower bound we derive for the tabular case is slightly different, which we show in the following section.

We first give a formal definition on the sparse reward MDP.

Definition 9 (Sparse Reward MDPs). *Let $\mathcal{M}$ be a set of MDPs sharing the same transition probabilities. We say $\mathcal{M}$ contains all the sparse reward MDPs if for each $s, h \in \mathcal{S} \times [H]$, there exists some MDP $M_{s,h} \in \mathcal{M}$, such that $R_{h', M_{s,h}}(s', a') = \mathbb{1}(s = s', h = h')$ for all s', a', h'.*

To show a lower bound on the myopic exploration gap, we make a further assumption on the occupancy measure $\mu_h^\pi(s, a) := \Pr_\pi(s_h = s, a_h = a)$, the probability of visiting s, a at the step h by running policy π.

Assumption 5 (Lower bound on the largest achievable occupancy measure)**.** *For all $s, h \in \mathcal{S} \times [H]$, we assume that $\max_\pi \mu_h^\pi(s) \geq b_1$ for some constant b or $\max_\pi \mu_h^\pi(s) = 0$.*

Assumption 5 guarantees that any β-optimal policy (with $\beta < b_1$) is not a vacuous policy and it provides a lower bound on the corresponding occupancy measure. We will discuss later in Appendix G on how to remove this assumption with an extra $S \times H$ factor on the sample complexity bound.

Proposition 5. *Consider a set of sparse reward MDP as in Definition 9. Assume Assumption 5 is true. For any $\beta \leq b_1/2$ and $f \in \mathcal{F}_\beta$, we have $\alpha(f, \mathcal{F}, \mathcal{M}) \geq \bar{\alpha}$ for some constant $\bar{\alpha} = \sqrt{\beta^2/(2e|\mathcal{M}|AH)}$ by choosing $\epsilon_h = 1/h$.*

Proof. We prove this lemma in a layered manner. Let h' be the minimum step such that there exists some $M_{s,h'}$ is β-suboptimal. By definition, in the layer $h' - 1$, all the MDPs are β-suboptimal, in which case $\boldsymbol{\pi}_{M_{s,h'-1}}$ visits $(s, h' - 1)$ with a probability at least $b/2$. Now we show that the optimal policy $\pi^*_{M_{s,h'}}$ of a suboptimal MDP $M_{s,h'}$ has lower bounded occupancy ratio.

For a more concise notation, we let $M' = M_{s,h'}$. Note that

$$
\begin{aligned}
\mu_{h'}^{\pi^*_{M'}}(s) &= \sum_{s' \in \mathcal{S}} \mu_{h'-1}^{\pi^*_{M'}}(s') P_{h'-1}(s \mid s', \pi^*_{M'}(s')) \\
&\leq \sum_{s' \in \mathcal{S}} \max_{\pi \in \Pi} \mu_{h'-1}^{\pi}(s') P_{h'-1}(s \mid s', \pi^*_{M'}(s')) \\
&\quad \text{(By the fact that } \mu_{h'-1}^{\boldsymbol{\pi}_{M_{s',h'-1}}}(s') \text{ is } \beta\text{-optimal policy of } M_{s',h'-1}) \\
&\leq \sum_{s' \in \mathcal{S}} \frac{b_1}{b_1 - \beta} \mu_{h'-1}^{\boldsymbol{\pi}_{M_{s',h'-1}}}(s') P_{h'-1}(s \mid s', \pi^*_{M'}(s')) \\
&\leq \sum_{s' \in \mathcal{S}} \frac{b_1 |\mathcal{M}| A}{(b_1 - \beta)(1 - \epsilon)^{h'-1} \epsilon} \mu_{h'-1}^{\mathrm{expl}(\boldsymbol{\pi})}(s') P_{h'-1}(s \mid s', \mathrm{expl}(\boldsymbol{\pi})(s')) \\
&= \frac{b_1 |\mathcal{M}| A}{(b_1 - \beta)(1 - \epsilon)^{h'-1} \epsilon} \mu_{h'}^{\mathrm{expl}(\boldsymbol{\pi})}(s)
\end{aligned}
$$

The occupancy measure ratio can be upper bounded by $c = \frac{b_1 |\mathcal{M}| A}{(b_1 - \beta)(1 - \epsilon)^{h'-1} \epsilon}$. Then the myopic exploration gap can be lower bounded by

$$
\frac{\beta}{\sqrt{c}} = \sqrt{\frac{(b_1 - \beta)\beta^2 (1 - \epsilon)^{h'-1} \epsilon}{b_1 |\mathcal{M}| A}} \geq \sqrt{\frac{\beta^2 (1 - \epsilon)^{h'-1} \epsilon}{2 |\mathcal{M}| A}}.
$$

To proceed, we choose $\epsilon_h = 1/h$, which leads to $(1-\epsilon_h)^{h-1}\epsilon \geq 1/(eH)$. □

Plugging this into Theorem 1, we achieve a sample complexity bound of $\mathcal{O}(S^2AH^5/\beta^2)$, with $|\mathcal{M}| = SH$. This is not a near-optimal bound for reward-free exploration (a fair comparison in our setup). This is because the sample complexity bound in Theorem 1 is not tailored for tabular case.

G.2 Removing coverage assumption

Though Assumption 2 and Assumption 4 are relatively common in the literature, we have not seen an any like Assumption 5. In fact, Assumption 5 is not a necessary condition for sample-efficient myopic exploration as we will discuss in this section. The main technical invention is to construct a mirror transition probability that satisfies the conditions in Assumption 5. However, we will see that a inevitable price of an extra SH factor has to be paid.

To illustrate the obstacle of removing Assumption 5, recall that the proof of Proposition 5 relies on the fact that all β-optimal policies guarantee a non-zero probability of visiting the state corresponding to their sparse reward with $\beta < b_1/2$. Without Assumption 5, a β-optimal policy can be an arbitrary policy. At the step h, we have at most S such MDPs, which may accumulate an irreducible error of $S\beta$, which means that at the round $h+1$, we can only guarantee $S\beta$-optimal policies. An naive adaptation will require us to set the accuracy $\beta' = \beta/S^H$ in order to guarantee a β error in the last step. The following discussion reveals that the error does not accumulate in a multiplicative way.

Mirror MDP construction. It is helpful to consider a mirror transition probability modified from our original transition probability. We denote the original transition probability by $P = \{P_h\}_{h\in[H]}$. Consider a new MDP with transition $P' = \{P'_h\}_{h\in[H]}$ and state space $\mathcal{S}' = \mathcal{S} \cup \{s_0\}$, where s_0 is a dummy state. We initialize P' such that

$$P'_h(s' \mid s, a) = P_h(s' \mid s, a) \text{ for all } s', s, a, h, \text{ where } s', s \neq s_0, \text{ and } P'_h(s_0 \mid s_0, \cdot) = 1 \quad (7)$$

Starting from $h = 1$, we update P'_h by a forward induction according to Algorithm 2. The design principle is to direct the probability mass of visiting $(s, h+1)$ to $(s_0, h+1)$, whenever the maximal probability of visiting $(s, h+1)$ is less than β.

Algorithm 2 Creating Mirror Transitions

```
Input: Original Transition $P$, threshold $\beta > 0$.
Initialize $P'$ according to (7)
for $h = 1, 2, \dots, H-1$ do
    for each $s \in \mathcal{S}$ such that $\max_\pi \mu'^{\pi}_{h+1}(s) \leq \beta$ do
        $P'_h(s_0 \mid \tilde{s}, \tilde{a}) \leftarrow P'_h(s_0 \mid \tilde{s}, \tilde{a}) + P'_h(s \mid \tilde{s}, \tilde{a})$ for each $\tilde{s}$, $\tilde{a}$.
        $P'_h(s \mid \tilde{s}, \tilde{a}) \leftarrow 0$ for each $\tilde{s}$, $\tilde{a}$.
    end for
end for
Return $P'$
```

By definition of P', we have two nice properties.

Proposition 6. *For any $h \in [H]$, $s \in \mathcal{S}$, we have $\max_\pi \mu'^{\pi}_h(s) = 0$ or $\max_\pi \mu'^{\pi}_h(s) > \beta$.*

Thus, P' nicely satisfies our Assumption 5. We also have that P' is not significantly different from P.

Proposition 7. *For any policy π, $\mu'^{\pi}_h(s) \geq \mu^{\pi}_h(s) - HS\beta$. Further more, any $(SH+1)\beta$-suboptimal policy for P is at least β-suboptimal for P' with respect to the same reward.*

Proof. We simply observe that $\max_\pi \mu'^{\pi}_h(s_0) \leq (h-1)S\beta$. This is true since at any round, we have at most S states with $\max_\pi \mu^{'\pi}(s) \leq \beta$, all the probability that goes to s will be deviated to s_0. Therefore, for any π

$$\mu'^{\pi}_{h+1}(s_0) \leq \mu'^{\pi}_h(s_0) + S\beta.$$

□

Therefore, any $(SH+1)\beta$-suboptimal policy for P has the myopic exploration gap of β-suboptimal policy for P'.

Theorem 4. *Consider a set of sparse reward MDP as in Definition 9. For any $\beta \in (0,1)$ and $f \in \mathcal{F}_\beta$, we have $\alpha(f, \mathcal{F}, \mathcal{M}) \geq \bar{\alpha}$ for some constant $\bar{\alpha} = \Omega(\sqrt{\beta^2/(|\mathcal{M}|AS^2H^3)})$ by choosing $\epsilon_h = 1/(h+1)$.*

H Connections to Diversity

Diversity has been an important consideration for the generalization performance of multitask learning. How to construct a diverse set, with which we can learn a model that generalizes to unseen task is studied in the literature of multitask supervised learning.

Previous works (Tripuraneni et al., 2020; Xu & Tewari, 2021) have studied the importance of diversity in multitask representation learning. They assume that the response variable is generated through mean function $f_t \circ h$, where h is the representation function shared by different tasks and f_t is the task-specific prediction function of a task indexed by t. They showed that diverse tasks can learn the shared representation that generalizes to unseen downstream tasks. More specifically, if $f_t \in \mathcal{F}$ is a discrete set, a diverse set needs to include all possible elements in $\mathcal{F}$. If $\mathcal{F}$ is the set of all bounded linear functions, we need d tasks to achieve a diverse set. Note that these results align with the results presented in the previous section. *Could there be any connection between the diversity in multitask representation learning and the efficient myopic exploration?*

Xu & Tewari (2021) showed that Eluder dimension is a measure for the hardness of being diverse. Here we introduce a generalized version called distributional Eluder dimension (Jin et al., 2021a).

Definition 10 (ε-independence between distributions)**.** *Let $\mathcal{G}$ be a class of functions defined on a space $\mathcal{X}$, and $\nu, \mu_1, \ldots, \mu_n$ be probability measures over $\mathcal{X}$. We say ν is ε-independent of $\{\mu_1, \mu_2, \ldots, \mu_n\}$ with respect to $\mathcal{G}$ if there exists $g \in \mathcal{G}$ such that $\sqrt{\sum_{i=1}^n (\mathbb{E}_{\mu_i}[g])^2} \leq \varepsilon$, but $|\mathbb{E}_\nu[g]| > \varepsilon$*

Definition 11 (Distributional Eluder (DE) dimension)**.** *Let $\mathcal{G}$ be a function class defined on $\mathcal{X}$, and Π be a family of probability measures over $\mathcal{X}$. The distributional Eluder dimension $\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \varepsilon)$ is the length of the longest sequence $\{\rho_1, \ldots, \rho_n\} \subset \Pi$ such that there exists $\varepsilon' \geq \varepsilon$ where ρ_i is ε'-independent of $\{\rho_1, \ldots, \rho_{i-1}\}$ for all $i \in [n]$.*

Definition 12 (Bellman Eluder (BE) dimension (Jin et al., 2021))**.** *Let $\mathcal{E}_h\mathcal{F}$ be the set of Bellman residuals induced by $\mathcal{F}$ at step h, and $\Pi = \{\Pi_h\}_{h=1}^H$ be a collection of H probability measure families over $\mathcal{X} \times \mathcal{A}$. The ε-Bellman Eluder dimension of $\mathcal{F}$ with respect to Π is defined as*

$$\dim_{\mathrm{BE}}(\mathcal{F}, \Pi, \varepsilon) := \max_{h \in [H]} \dim_{\mathrm{DE}}\left(\mathcal{E}_h\mathcal{F}, \Pi, \varepsilon\right)$$

Constructing a diverse set. For each $h \in [H]$, consider a sequence of functions $f_1, \ldots, f_d \in \mathcal{F}$, such that the induced policy $(\pi^{f_i})_{i \in [d]}$ generates probability measures $(\mu_{h+1}^{f_i})_{i \in [d]}$ at the step $h+1$. Let $(\mu_{h+1}^{f_i})_{i \in [d]}$ be ϵ-independence w.r.t the function class $\mathcal{E}_h\mathcal{F}$ between their predecessors. By the definition of BE dimension, we can only find at most $\dim_{\mathrm{DE}}(\mathcal{E}_h\mathcal{F}, \Pi, \varepsilon)$ of these functions. Then conditions in Definition 3 is satisfied with $c = 1/(dH)$.

Revisiting linear MDPs. The task set construction in 7 seems to be quite restricted as we require a set of standard basis. One might conjecture that a task set $M_{i,h}$ with full rank $[\theta_{1,h}, \ldots, \theta_{d,h}]$ will suffice. From what we discussed in the general case, we will need the occupancy measure generated by the optimal policies for these MDPs to be ϵ-independent and any other distribution is ϵ-dependent. This is generally not true even if the reward parameters are full rank. To see this, let us consider a tabular MDP case with two states $\{1, 2\}$, where at the step h, we have two tasks M_1, M_2, with $R_{h,M_1}(s,a) = \mathbb{1}[s=1]$ and $R_{h,M_2}(s,a) = 0.51\mathbb{1}[s=1] + 0.49\mathbb{1}[s=2]$. This gives $\theta_{h,M_1} = [1,0]$ and $\theta_{h,M_2} = [0.49, 0.51]$ as shown in Figure 3.

Construct the MDP such that the transition probability and action space any policy either visit state 1 or state 2 at the step h. Then the optimal policies for both tasks are the same policy which visits state 1 with probability one, even if the reward parameters $[\theta_{h,M_1}, \theta_{h,M_2}]$ are full-rank.

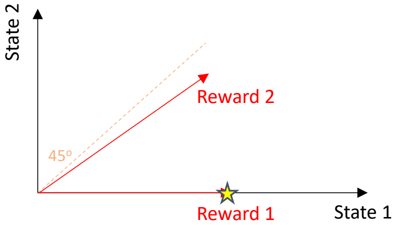


Figure 3: An illustration of why a full-rank set of reward parameters does not achieve diversity. The red arrows are two reward parameters and the star marks the generated state distributions of the optimal policies corresponding to the two rewards at the step h. Since both optimal policies only visit state 1, they may not provide a sufficient exploration for the next time step $h + 1$.

I IMPLICATIONS OF DIVERSITY ON ROBOTIC CONTROL ENVIRONMENTS

In this section, we conduct simulation studies on robotic control environments with practical interests. Since myopic exploration has been shown empirically efficient in many problems of interest, we focus on the other main topic–diversity. We investigate how our theory guides a diverse task set selection. More specifically, our prior analysis on Linear MDPs suggests that a diverse task set should prioritize tasks with full-rank feature covariance matrices. We ask whether tasks with a more spread spectrum of the feature covariance matrix lead to a better training task set. *Note that the goal of this experiment is not to show the practical interests of Algorithm 1. Instead, we are revealing interesting implications of the highly conceptual definition of diversity in problems with practical interests.*

Environment and training setup. We adopt the BipedalWalker environment from (Portelas et al., 2020). The learning agent is embodied into a bipedal walker whose motors are controllable with torque (i.e. continuous action space). The observation space consists of laser scan, head position, and joint positions. The objective of the agent is to move forward as far as possible, while crossing stumps with varying heights at regular intervals (see Figure 4 (a)). The agent receives positive rewards for moving forward and negative rewards for torque usage. An environment or task, denoted as $M_{p,q}$, is controlled by a parameter vector (p, q), where p and q denote the heights of the stumps and the spacings between the stumps, respectively. Intuitively, an environment with higher and denser stumps is more challenging to solve. We set the parameter ranges for p and q as $p \in [0, 3]$ and $q \in [0, 6]$ in this study.

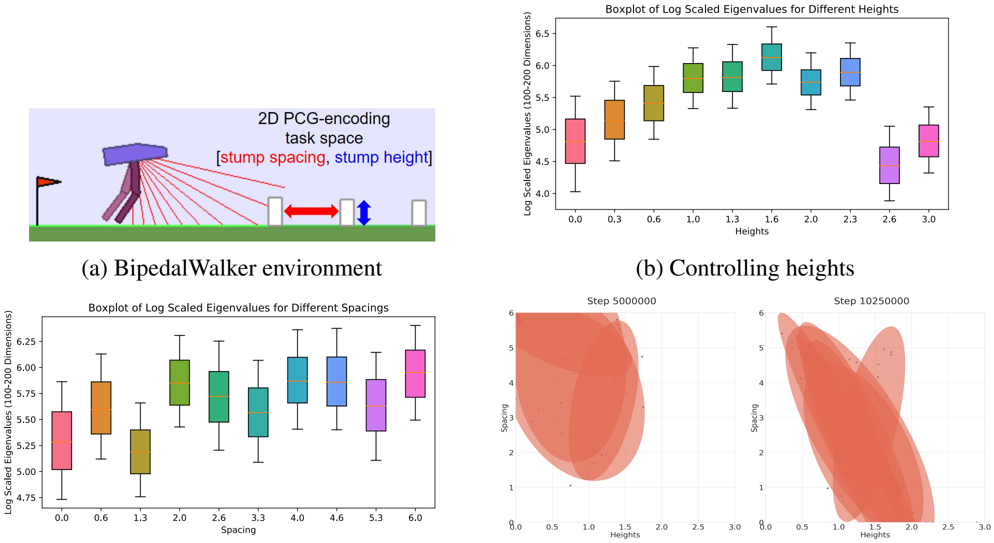
The agent is trained by Proximal Policy Optimization (PPO) (Schulman et al., 2017) with a standard actor-critic framework (Konda & Tsitsiklis, 1999) and with Boltzmann exploration that regularizes entropy. Note that Boltzmann exploration strategy is another example of myopic exploration, which is commonly used for continuous action space. Though it deviates from the ϵ-greedy strategy discussed in the theoretical framework, we remark that the theoretical guarantee outlined in this paper can be trivially extend to Boltzmann exploration. The architecture for the actor and critic feature extractors consists of two layers with 400 and 300 neurons, respectively, and Tanh (Rumelhart et al., 1986) as the activation function. Fully-connected layers are then used to compute the action and value. We keep the hyper-parameters for training the agent the same as Romac et al. (2021).

I.1 INVESTIGATING FEATURE COVARIANCE MATRIX

We denote by $\phi(s, a)$ the output of the feature extractor. We evaluate the extracted feature at the end of the training generated by near-optimal policies, denoted as π, on 100 tasks with different parameter vectors (p, q). We then compute the covariance matrix of the features for each task, denoted as $V_{p,q} = \mathbb{E}_{\pi}^{M_{p,q}} \sum_{h=1}^{H} \phi(s_h, a_h)\phi(s_h, a_h)^T$, whose spectrums are shown in Figure 4 (b) and (c).

By ignoring the extremely large and small eigenvalues on two ends, we focus on the largest 100-200 dimension, where we observe that the height of the stumps p has a larger impact on the distribution of eigenvalues. In Figure 4 (b), we show the boxplot of the log-scaled eigenvalues of 100-200 dimensions for environments with different heights, where we average spacings. We observe that the eigenvalues can be 10 times higher for environments with an appropriate height (1.0-2.3), compared to extremely high and low heights. However, the scales of eigenvalues are roughly the same if we control the spacings and take average over different heights as shown in Figure 4 (c). This indicates that choosing an appropriate height is the key to properly scheduling tasks.

In fact, the task selection coincidences with the tasks selected by the state-of-the-art Automatic Curriculum Learning (ACL). We investigate the curricula generated by ALP-GMM (Portelas et al., 2020), a well-established curriculum learning algorithm, for training an agent in the BipedalWalker environment for 20 million timesteps. Figure 4 (d) gives the density plots of the ACL task sampler during the training process, which shows a significant preference over heights in the middle range, with little preference over spacing.



(a) BipedalWalker environment

(b) Controlling heights

(c) Controlling spacings

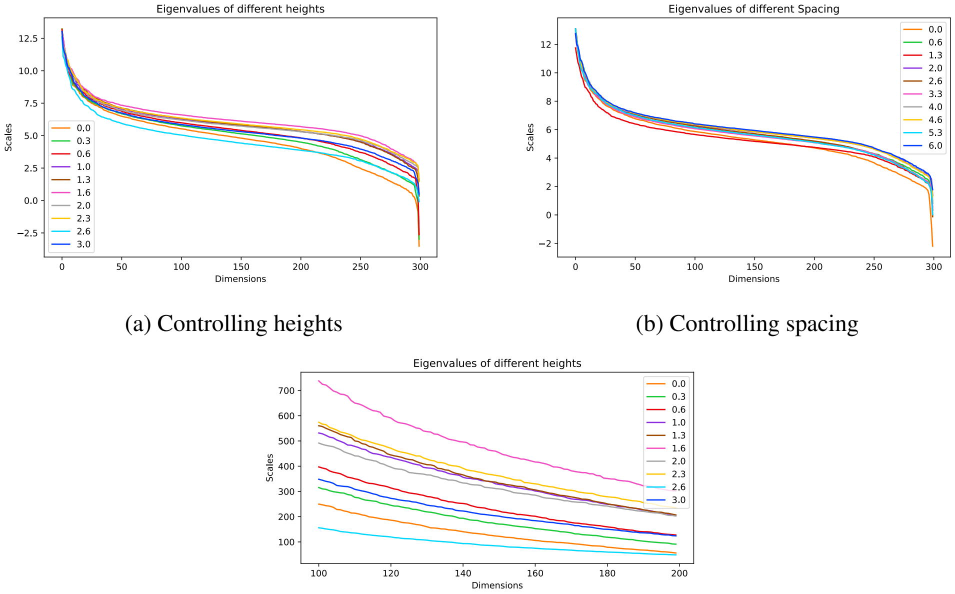
(d) Preference of automatic CL

Figure 4: **(a)** BipedalWalker Environment with different stump spacing and heights. **(b-c)** Boxplots of the log-scaled eigenvalues of sample covariance matrices of the trained embeddings generated by the near optimal policies for different environments. In (b), we take average over environments with the same height while in (c), over the same spacing. **(d)** Task preference of automatically generated curriculum at 5M and 10M training steps respectively. The red regions are the regions where a task has a higher probability to be sampled.

Training on different parameters. To further validate our finding, we train the same PPO agent with different means of the stump heights and see that how many tasks does the current agent master. As we argued in the theory, a diverse set of tasks provides good behavior policies for other tasks of interest. Therefore, we also test how many tasks it could further master if one use the current policy as behavior policy for fine-tuning on all tasks. The number of tasks the agent can master by learning on environments with heights ranging in [0.0, 0.3], [1.3, 1.6], [2.6, 3.0] are 28.1, 41.6, 11.5, respectively leading to a significant outperforming for diverse tasks ranging in [1.3, 1.6]. Table 1 gives a complete summary of the results.

Table 1: Training on different environment parameters. Each row represents a training scenario, where the first two columns are the range of sampled parameters. The mastered tasks are out of 121 evaluated tasks with the standard deviation calculated from ten independent runs.

Obstacle spacing	Stump height	Mastered task
[2, 4]	[0.0, 0.3]	28.1 ± 6.1
[2, 4]	[1.3, 1.6]	41.6 ± 9.8
[2, 4]	[2.6, 3.0]	11.5 ± 10.9



(a) Controlling heights

(b) Controlling spacing

(c) Controlling heights on the original scale

Figure 5: **(b-c)** Log-scaled eigenvalues of sample covariance matrices of the trained embeddings generated by the near optimal policies for different environments.