Guarantees for Epsilon-Greedy Reinforcement Learning with Function Approximation

Christoph Dann [1] **Yishay Mansour** [1 2] **Mehryar Mohri** [1 3] **Ayush Sekhari** [4] **Karthik Sridharan** [4]

Abstract

Myopic exploration policies such as ε-greedy, softmax, or Gaussian noise fail to explore efficiently in some reinforcement learning tasks and yet, they perform well in many others. In fact, in practice, they are often selected as the top choices, due to their simplicity. But, for what tasks do such policies succeed? Can we give theoretical guarantees for their favorable performance? These crucial questions have been scarcely investigated, despite the prominent practical importance of these policies. This paper presents a theoretical analysis of such policies and provides the first regret and sample-complexity bounds for reinforcement learning with myopic exploration. Our results apply to value-function-based algorithms in episodic MDPs with bounded Bellman Eluder dimension. We propose a new complexity measure called myopic exploration gap, denoted by α, that captures a structural property of the MDP, the exploration policy expl and the given value function class $\mathcal{F}$. We show that the sample-complexity of myopic exploration scales quadratically with the inverse of this quantity, $1/\alpha^2$. We further demonstrate through concrete examples that myopic exploration gap is indeed favorable in several tasks where myopic exploration succeeds, due to the corresponding dynamics and reward structure.

1. Introduction

Remarkable empirical and theoretical advances have been made in scaling up Reinforcement Learning (RL) approaches to complex problems with rich observations. On the theory side, a suite of algorithmic and analytical tools have been developed for provably sample-efficient RL in MDPs with general structural assumptions (Jiang et al., 2017; Sun et al., 2018; Wang et al., 2020; Jin et al., 2020; Foster et al., 2020; Dann et al., 2021b; Du et al., 2021). The key innovation in these works is to enable strategic exploration in combination with general function approximation, leading to strong worst-case guarantees. On the empirical side, agents with large neural networks have been trained, for example, to successfully play Atari games (Mnih et al., 2015), beat world-class champions in Go (Silver et al., 2017) or control real-world robots (Kalashnikov et al., 2018). However, perhaps surprisingly, many of these empirically successful approaches, do not rely on strategic exploration but instead perform simple myopic exploration. Myopic approaches simply perturb the actions prescribed by the current estimate of the optimal policy, for example by taking a uniformly random action with probability ε (called ε-greedy exploration).

While myopic exploration is known to have exponential sample complexity in the worst case (Osband et al., 2019), it still remains a popular choice in practice. The reason for this is because myopic exploration is easy to implement and works well in a range of problems, including many of practical interest (Mnih et al., 2015; Kalashnikov et al., 2018). On the other hand, it is unknown how to implement existing strategic exploration approaches computationally efficiently with general function approximation. They either require solving intricate non-convex optimization problems (Jiang et al., 2017; Jin et al., 2021; Du et al., 2021) or sampling from intricate distributions (Zhang, 2021).

Despite its empirical importance, theoretical analysis of myopic exploration is still quite rudimentary, and performance guarantees for sample complexity or regret bounds are largely unavailable. In order to bridge this gap between theory and practice, in this work we investigate:

> *In which problems does myopic exploration lead to sample-efficient learning, and can we provide a theoretical guarantee for myopic reinforcement learning algorithms such as ε-greedy?*

We address this question by providing a framework for analyzing value-function based RL with myopic exploration. Importantly, the algorithm we analyze is easy to implement

[1]Google Research [2]Tel Aviv University [3]Courant Institute of Mathematical Sciences [4]Cornell University. Correspondence to: Christoph Dann <cdann@cdann.net>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

since it only requires minimizing standard square loss on the value function class for which many practical approaches exist, even on complex neural networks (Mnih et al., 2015; Silver et al., 2016; 2017; Rakhlin & Sridharan, 2015).

Our main contributions are threefold:

- We propose a new complexity measure called myopic exploration gap $\alpha(f, \mathcal{F})$ that captures how easy it is for a given myopic exploration strategy to identify that a candidate value function $f \in \mathcal{F}$ is sub-optimal. This complexity measure is large for problems with favorable transition dynamics and rewards, where we expect myopic exploration to perform well.

- We derive a sample complexity upper bound for RL with myopic exploration. We show that for any sub-optimal $\mathcal{F}' \subset \mathcal{F}$, the algorithm uses policies corresponding to $\mathcal{F}'$ for at most $\tilde{O}\big(\frac{H^2 d}{\alpha(\mathcal{F}', \mathcal{F})^2}\big)$ episodes, where H is the episode length, d is the Bellman Eluder dimension and $\alpha(\mathcal{F}', \mathcal{F}) = \min_{f \in \mathcal{F}'} \alpha(f, \mathcal{F})$. Using this sample complexity bound, we provide the first regret bound for ε-greedy RL in MDPs.

- We prove an almost matching sample complexity lower bound of $\Omega(d/\alpha(\mathcal{F}', \mathcal{F})^2)$ for ε-greedy RL, hence showing that the dependency on Bellman Eluder dimension or the myopic exploration gap in our upper bound cannot be improved further.

2. Preliminaries

Episodic MDP. We consider RL in episodic Markov Decision Processes (MDPs), each denoted by $\mathcal{M} = (\mathcal{X}, \mathcal{A}, H, P, R)$, where $\mathcal{X}$ is the observation (or state) space, $\mathcal{A}$ is the action space, $H \in \mathbb{N}$ is the number of time steps in each episode, $P = (P_h)_{h \in [H]}$ is the collection of transition kernels and $R = (R_h)_{h \in [H]}$ is the collection of immediate reward distributions. An episode is a sequence of states, actions and rewards $(x_1, a_1, r_1, \ldots, x_H, a_H, r_H, x_{H+1})$ with a fixed initial state $x_1 = x_{\text{init}} \in \mathcal{X}$ and terminal state $x_{H+1} = x_{\text{end}} \in \mathcal{X}$. All states (except x_1) are sampled from the transition kernels $x_{h+1} \sim P_h(\cdot | x_h, a_h) = P_h(x_h, a_h)$ and rewards are generated as $r_h \sim R_h(x_h, a_h)$. We assume that P and R are s.t. $r_h \geq 0$ and the sum of all rewards in an episode is bounded as $\sum_{h=1}^{H} r_h \leq 1$ for any action sequence almost surely.

Policies and value functions. The agent interacts with the environment in multiple episodes indexed by t. It chooses actions according to a *policy* $\pi = (\pi_h \colon \mathcal{X} \to \Delta_{\mathcal{A}})_{h \in [H]}$ that maps states to distributions over actions. The space of all such policies is denoted by Π. We denote the distribution over an episode induced by following a policy π by $\mathbb{P}_\pi$ and the expectation w.r.t. this law by $\mathbb{E}_\pi$. We define the Q-function and value-function of a policy π at step h as

$$Q_h^\pi(x, a) = \mathbb{E}_\pi[r_h + V_{h+1}^\pi(x_{h+1}) \mid x_h = x, a_h = a],$$
$$V_h^\pi(x) = \mathbb{E}_\pi[Q_h^\pi(x_h, a_h) | x_h = x],$$

respectively. The optimal Q- and value functions are denoted by $Q_h^\star$ and $V_h^\star$. The occupancy measure of π at time h is denoted as $\mu_h^\pi(x, a) = \mathbb{P}_\pi(x_h = x, a_h = a)$. For any function $f \colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, the *Bellman operator* at step h is

$$(\mathcal{T}_h f)(x, a) = \mathbb{E}\big[r_h + \max_{a' \in \mathcal{A}} f(x_{h+1}, a') \,\big|\, x_h = x, a_h = a\big].$$

Value function class. We assume the learner has access to a Q-function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$ where $\mathcal{F}_h \subseteq (\mathcal{X} \times \mathcal{A} \to [0, 1])$, and for convention we set $\mathcal{F}_{H+1} = \{f_{H+1} = 0\}$. We denote by $\pi^f = \{\pi_h^f\}_{h \in [H]}$ the greedy policy w.r.t. $f \in \mathcal{F}$, i.e., $\pi_h^f(x) = \arg\max_{a \in \mathcal{A}} f_h(x, a)$. The set of greedy policies of $\mathcal{F}$ is denoted by $\Pi_{\mathcal{F}}$. The Bellman residual and squared Bellman error of $f \in \mathcal{F}$ at time h are

$$\mathcal{E}_h f = f_h - \mathcal{T}_h f_{h+1} \quad \text{and} \quad \mathcal{E}_h^2 f = (f_h - \mathcal{T}_h f_{h+1})^2.$$

We further adopt the following assumption common in RL theory with function approximation (cf. Jin et al., 2021; Wang et al., 2020; Du et al., 2021; Dann et al., 2021b):

Assumption 1 (Realizability and Completeness)**.** $\mathcal{F}$ is realizable and complete under the Bellman operator, that is, $Q_h^\star \in \mathcal{F}_h$ for all $h \in [H]$ and for every $h \in [H]$ and $f_{h+1} \in \mathcal{F}_{h+1}$ there is a $f_h \in \mathcal{F}_h$ such that $f_h = \mathcal{T}_h f_{h+1}$.

Bellman Eluder dimension. To capture the complexity of the MDP and the function class we use the notion of Bellman Eluder dimension $\dim_{\text{BE}}(\mathcal{F}, \Pi_{\mathcal{F}}, \mu)$ introduced by Jin et al. (2021). This complexity measure is a distributional version of the Eluder dimension applied to the class of Bellman residuals $\mathcal{E}_h \mathcal{F}$. The Bellman Eluder dimension matches the natural complexity measure in many special cases, e.g., number of states and actions (SA) in tabular MDPs or the intrinsic dimension d in linear MDPs. The formal definition and the main properties of the Bellman Eluder dimension we use in our work, can be found in Appendix E.

We denote by $N_{\mathcal{F}_h}(\lambda)$ the ℓ_∞ covering number of $\mathcal{F}_h$ w.r.t. radius λ, i.e., the size of the smallest $\mathcal{Z} \subset \mathcal{F}_h$ so that for any $f \in \mathcal{F}$ there is a $f' \in \mathcal{Z}$ with $\max_{h \in [H]} \|f_h - f'_h\|_\infty \leq \lambda$ and we use $\bar{N}_{\mathcal{F}}(\lambda) = \sum_{h=1}^{H} N_{\mathcal{F}_h}(\lambda)$.

3. RL with Myopic Exploration

Our work analyzes value-function based algorithms with myopic exploration that follow the template in Algorithm 1. The algorithm receives as input a value function class $\mathcal{F}$ as well as an exploration mapping expl that maps each function

in $\mathcal{F}$ to an exploration policy Π. Before each episode, the algorithm computes a Q-function estimate $\widehat{f}_t$ in a dynamic programming fashion by least squares regression using all data observed so far (finite horizon fitted Q iteration). The sampling policy $\widetilde{\pi}_t$ is then determined using the exploration mapping expl. After sampling one episode with $\widetilde{\pi}_t$, the algorithm adds all observations to the datasets and computes a new value function.

This algorithm template encompasses many common exploration heuristics by setting the exploration mapping expl appropriately. These include:[1]

Exploration strategy	$[\mathsf{expl}(f)]_h(a\|x) \propto$
ε-greedy	$(1-\varepsilon)\pi_h^f(a\|x) + \varepsilon/A$
softmax	$\exp(\beta f_h(x,a))$
Gaussian noise	$\exp\left(-\frac{1}{2\sigma^2}\left(a - \pi_h^f(x)\right)^2\right)$

Note that expl only takes the point estimate for the Q-function $\widehat{f}_t$ as input and no measure of uncertainty. Thus, approaches like upper-confidence bounds or Thompson sampling are beyond the scope of our work. Furthermore, to keep the exposition clean, we restrict ourselves to mappings expl that are independent of the number of episodes k, but we allow the mapping to depend on the total number of episodes T. Getting anytime guarantees with variable exploration policies (e.g. ε-greedy with decreasing ε) is an interesting direction for future work.

One important feature of Algorithm 1 is that it only requires solving least-squares regression problems, for which numerous efficient methods exist, even when $\mathcal{F}$ are large neural networks. This is in stark contrast to all known algorithms for general function approximation with worst-case sample-efficiency guarantees. Those require either solving an intricate non-convex optimization problem to ensure global optimism (Jiang et al., 2017; Dann et al., 2018; Jin et al., 2021; Du et al., 2021) or sampling from a posterior distribution without closed form (Dann et al., 2021b). The lack of computationally and statistically worst-case efficient algorithms is a strong motivation for studying the sample-efficiency of the simple and computationally tractable approach in Algorithm 1.

We assume that the regression problem in line 5 is solved exactly for ease of presentation but our analysis can be easily extended to allow small optimization error in each episode.

[1] Our examples are Markovian policies but we also support non-Markovian exploration policies that inject temporally correlated noise, e.g., the Ornstein-Uhlenbeck noise used by Lillicrap et al. (2015) or ε-greedy with options (Dabney et al., 2020).

Algorithm 1: RL with myopic exploration

input : function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_{H+1}$
input : myopic exploration mapping $\mathsf{expl}\colon \mathcal{F} \to \Pi$
1 Initialize datasets $\mathcal{D}_h \leftarrow \varnothing$ for all $h \in [H]$;
2 **for** *episode* $t = 1, 2, \ldots$ **do**
3 Set $\widehat{f}_{t,H+1} = 0$;
4 **for** $h = H, \ldots, 1$ **do**
5 Fit Q-function with least-squares regression
 $\widehat{f}_{t,h} \leftarrow \arg\min_{f \in \mathcal{F}_h} \mathcal{L}_h(f, \widehat{f}_{t,h+1})$ where
6 $$\mathcal{L}_h(f, f') = \sum_{(x,a,r,x') \in \mathcal{D}_h} \left(f(x,a) - r - \max_{a'} f'(x', a')\right)^2$$
7 Set myopic exploration policy $\widetilde{\pi}_t \leftarrow \mathsf{expl}(\widehat{f}_t)$;
8 Sample one episode $\{x_1, a_1, r_1, \ldots x_{H+1}\}$ with $\widetilde{\pi}_t$;
9 Add observations to datasets
 $\mathcal{D}_h \leftarrow \mathcal{D}_h \cup \{(x_h, a_h, r_h, x_{h+1})\}$ for all $h \in [H]$;

4. Myopic Exploration Gap

In this section, we introduce a new complexity measure called *myopic exploration gap* in order to formalize the sample-efficiency of myopic exploration. One important feature of this quantity is that it depends both on the dynamics and the reward structure of the MDP. This enables us to give tighter guarantees in the presence of favorable rewards, a key property absent in existing analyses of myopic exploration.

The sample complexity and regret of any (reasonable) algorithm depends on a notion of gap or suboptimality, either of individual actions or entire policies, as shown by existing works on provably sample efficient RL (Simchowitz & Jamieson, 2019; Dann et al., 2021a; Wagenmaker et al., 2021). Intuitively, these algorithms need to test whether a candidate value function $f \in \mathcal{F}$ is optimal, and typically value functions that have large instantaneous regret $V^\star(x_1) - V^{\pi^f}(x_1)$ require fewer samples to be ruled out as being the optimal. This is also the case for myopic exploration based algorithms but the suboptimality gap alone is not sufficient to quantify their performance. This is because:

(a) Myopic exploration policies such as ε-greedy mostly collect samples in the "neighborhood" of the greedy policy. However, the optimal policy may visit high-reward state-action pairs that are outside of this neighborhood and thus have low probability of being visited under the exploration policy. In such cases, to discover the high reward states and to identify the gap, the agent thus needs to collect a very large number of episodes. Essentially, the effective size of the gap is reduced.

(b) The agent is not required to estimate the suboptimality

gap w.r.t. the optimal policy $\pi^\star$ in order to rule out a candidate policy π – it only needs to find a better policy π'. If there is a such a π' in the "neighborhood" of π, then the exploration policy may be more effective at identifying the optimality gap between π and π' even though it is smaller than the optimality gap between π and $\pi^\star$.

We address both of these issues in our definition of myopic exploration gap for $f \in \mathcal{F}$ by (1) considering the gap $V_1^{\pi'}(x_1) - V_1^{\pi^f}(x_1)$ between π^f and any other policy π', and (2) normalizing this gap by a scaling factor c. This factor, which we call *myopic exploration radius*, measures the effectiveness of the exploration policy $\mathsf{expl}(f)$ at collecting informative samples for identifying the gap.

Definition 1 (myopic exploration gap)**.** *Given a function class $\mathcal{F}$, exploration mapping $\mathsf{expl}\colon \mathcal{F} \to \Pi$, MDP $\mathcal{M}$ and policy class Π', we define the myopic exploration gap $\alpha(f, \mathcal{F}, \Pi', \mathsf{expl}, \mathcal{M})$ of $f \in \mathcal{F}$ as the value of*

$$\begin{aligned}&\sup_{\pi' \in \Pi', c \geq 1} \frac{1}{\sqrt{c}}(V_1^{\pi'}(x_1) - V_1^{\pi^f}(x_1)) \\ &\qquad \textit{such that for all } f' \in \mathcal{F} \textit{ and } h \in [H], \\ &\mathbb{E}_{\pi'}[(\mathcal{E}_h^2 f')(x_h, a_h)] \leq c\mathbb{E}_{\mathsf{expl}(f)}[(\mathcal{E}_h^2 f')(x_h, a_h)] \\ &\mathbb{E}_{\pi^f}[(\mathcal{E}_h^2 f')(x_h, a_h)] \leq c\mathbb{E}_{\mathsf{expl}(f)}[(\mathcal{E}_h^2 f')(x_h, a_h)].\end{aligned} \tag{1}$$

Furthermore, we define the myopic exploration radius $c(f, \mathcal{F}, \Pi', \mathsf{expl}, \mathcal{M})$ as the smallest value of c that attains the maximum in (1). To simplify the notation, we omit the dependence on expl and $\mathcal{M}$. When $\Pi' = \Pi_{\mathcal{F}}$ we also omit the dependence on Π' and use the notation $\alpha(f, \mathcal{F})$ and $c(f, \mathcal{F})$ respectively.

The constraints in (1) determine how small the normalization of the gap can be chosen. We compare the expected squared Bellman errors $\mathcal{E}_h^2 f'$ under different distributions. To see why, consider the following: least squares value iteration in Algorithm 1 minimizes the squared Bellman error on the dataset collected by exploration policies, which is closely related to the RHS in the constraints. Additionally, our analysis relates the LHS of the constraints to the gap—if $\mathcal{E}_h^2 f$ is small for all h under for both $\mathbb{E}_{\pi'}$ and $\mathbb{E}_{\pi^f}$, then the gap $V_1^{\pi'}(x_1) - V_1^{\pi^f}(x_1)$ cannot be large. Therefore, c measures how effective the exploration policy is at identifying the gap between π^f and π'.

Alternative Form in Tabular MDPs. The definition of myopic exploration gap in (1) only involves expectations w.r.t. policies induced by $\mathcal{F}$, which is desirable in the rich observation setting. However, in tabular MDPs where the number of states S and actions A is finite and $\mathcal{F}$ is the entire Q-function class with $N_{\mathcal{F}}(\lambda) \approx (A/\lambda)^{SAH}$, it may be more convenient to have constraints on individual state-action pairs instead. Since the class of squared Bellman errors $\mathcal{E}_h^2 \mathcal{F}$) for tabular representations is rich enough to include functions that are nonzero in a single state-action pair, we can write (1) with SAH constraints on the corresponding occupancy measures in tabular problems:

$$\begin{aligned}&\sup_{\pi' \in \Pi', c \geq 1} \frac{1}{\sqrt{c}}(V_1^{\pi'}(x_1) - V_1^{\pi^f}(x_1)) \\ &\text{such that for all } (x, a) \in \mathcal{X} \times \mathcal{A} \text{ and } h \in [H], \\ &\qquad \mu_h^{\pi'}(x, a) \leq c\mu_h^{\mathsf{expl}(f)}(x, a) \\ &\qquad \mu_h^{\pi^f}(x, a) \leq c\mu_h^{\mathsf{expl}(f)}(x, a).\end{aligned} \tag{2}$$

While determining the exact value of $\alpha(f, \mathcal{F})$ is challenging due to its intricate dependence on the MDP dynamics, chosen function class and the exploration policy, we can often provide meaningful and interpretable bounds.

5. When are Myopic Exploration Gaps Large?

We now discuss various structural properties of the transition dynamics and rewards under which the myopic exploration gap is large, and thus myopic exploration converges quickly.

5.1. Favorable Transition Dynamics

Structure in the transition dynamics can make myopic exploration approaches effective. To understand how effective is the myopic exploration gap for those cases, first note that $\alpha(f, \mathcal{F})$ can never be larger than the optimality gap of π^f (by definition). However, as shown in the following lemma, it is also lower bounded when the state-space distribution induced by $\mathsf{expl}(f)$ is close to that of any other policy.

Lemma 1. *When $Q^\star \in \mathcal{F}$, the myopic exploration gap is bounded for all $f \in \mathcal{F}$ as*

$$\left(\max_{\pi' \in \Pi'} \left\| \frac{\mathbb{P}_{\pi'}}{\mathbb{P}_{\mathsf{expl}(f)}} \right\|_\infty\right)^{-\frac{1}{2}} \leq \frac{\alpha(f, \mathcal{F})}{V_1^\star(x_1) - V_1^{\pi^f}(x_1)} \leq 1.$$

Furthermore, the myopic exploration radius is bounded as $c(f, \mathcal{F}) \leq \max_{\pi' \in \Pi'} \left\| \frac{\mathbb{P}_{\pi'}}{\mathbb{P}_{\mathsf{expl}(f)}} \right\|_\infty$.

Similar dependencies on the Radon-Nikodym derivative $\mathbb{P}_{\pi'}/\mathbb{P}_{\mathsf{expl}(f)}$ can also be found in the form of the concentrability coefficient in error bounds for fitted Q-iteration under a single behavior policy (Antos et al., 2008). The lower-bound in Lemma 1 can be used to bound the myopic exploration gap under various other assumptions, including the following worst-case lower-bound for ε-greedy.

Corollary 1. *For any MDP with finitely many actions with $|\mathcal{A}| = A$ and $f \in \mathcal{F}$, the myopic exploration gap of ε-greedy can be bounded as*

$$\alpha(f, \mathcal{F}) \geq \left(\frac{\varepsilon}{A}\right)^{\frac{H}{2}} (V_1^\star(x_1) - V_1^{\pi^f}(x_1))$$

and the exploration radius is bounded as $c(f, \mathcal{F}) \leq \left(\frac{\varepsilon}{A}\right)^H$.

We can similarly derive a lower bound on the myopic exploration gap for softmax-policies, where ε is replaced by $e^{-\beta}$, i.e., $\alpha(f, \mathcal{F}) \geq (Ae^{\beta})^{-H/2}(V_1^\star(x_1) - V_1^{\pi^f}(x_1))$.

The bound in Corollary 1 is exponentially small in H which is not surprising. It is well known that ε-greedy is ineffective in some problems and the construction in the proof of Theorem 2 will provide a formal example of this. However, in many problems, the gap can be much larger due to favorable dynamics.

5.1.1. Small Covering Length and Coverage

We first turn to insights from prior work and show that our framework captures the favorable cases that they identified. Liu & Brunskill (2018) investigated a related question to ours—when is myopic exploration sufficient for PAC learning with polynomial bounds in tabular MDPs. However, their work focuses on the infinite horizon setting and only considers explore-then-commit Q-learning where all the exploration is done using a uniformly random policy. They identify conditions under which the covering length of their exploration policy is polynomially small, a quantity that governs the sample-efficiency of Q-learning with a fixed exploration policy (Even-Dar & Mansour, 2003). Covering length is typically used in infinite-horizon MDPs but we can transfer the concept to our episodic setting as well:

Definition 2. *The covering length $L(\pi)$ of a policy $\pi \in \Pi$ is the number of episodes until all state-action pairs have been visited at all time steps with probability at least $1/2$.*

Lower bounding $L(\pi)$ can be thought of as a variant of the popular coupon collector's problem where each object has a different probability. Since the agent has to obtain at least one sample from each (x, a, h) triplet with probability $1/2$, and the probability to receive such a sample is $\mu_h^\pi(x, a)$ in each episode, the covering length must scale as

$$L(\pi) = \Omega\Big(\frac{1}{\min_{x,a,h} \mu_h^\pi(x, a)}\Big).$$

Intuitively, we expect learning to be efficient if $\min_{x,a,h} \mu_h^\pi(x, a)$ is large for the data collection policy π, since the agent is able to collect samples from the entire state-action space. Note that

$$\max_{\pi' \in \Pi'} \left\| \frac{\mathbb{P}_{\pi'}}{\mathbb{P}_{\mathsf{expl}(f)}} \right\|_\infty \leq \left\| \frac{1}{\mathbb{P}_{\mathsf{expl}(f)}} \right\|_\infty \leq \frac{1}{\min_{x,a,h} \mu_h^\pi(x, a)}.$$

Plugging the above bound in Lemma 1 we get that

$$\begin{aligned}\alpha(f, \mathcal{F}) &\geq \sqrt{\min_{x,a,h} \mu_h^\pi(x, a)}(V_1^\star(x_1) - V_1^{\pi^f}(x_1)) \\ &= \Omega(L(\mathsf{expl}(f))^{-1/2})(V_1^\star(x_1) - V_1^{\pi^f}(x_1)).\end{aligned}$$

This shows that myopic exploration gap is large when the covering length is small, and thus our definition recovers the prior results based on covering length.

5.1.2. Small Action Variation

One expects myopic exploration to be effective in MDPs where the transition dynamics have almost identical next-state distributions corresponding to taking different actions at any given state. In such MDPs, exploration is easy as all the policies will have similar occupancy measures.

Prior works (Jiang et al., 2016; Liu & Brunskill, 2018) have considered an additive notion of action variation that bounds the L_1 distance between the next-state distributions corresponding to different actions. However, for our applications, a multiplicative version of action variation is better suited.

Definition 3 (Multiplicative Action Variation)**.** *For any MDP with state space $\mathcal{X}$, action space $\mathcal{A}$ and transition dynamics P, define multiplicative action variation as the minimum δ_P such that for any x, a, a' and h,*

$$\left\| \frac{\mathrm{d}P_h(x, a)}{\mathrm{d}P_h(x, a')} \right\|_\infty \leq \delta_P.$$

The above definition implies that for any state x, actions a and a', and next state x', the transition probability $P_h(x' \mid x, a)$ is bounded by $\delta_P P_h(x' \mid x, a')$. Thus, by symmetry we always have that $\delta_P \geq 1$. The smaller the value of δ_P, the better we expect myopic exploration to perform.

Lemma 2. *Let $\mathcal{M}$ be any MDP with multiplicative action variation δ_P, and $Q^* \in \mathcal{F}$. Then, for any $f \in \mathcal{F}$, the myopic exploration gap of ε-greedy satisfies*

$$\alpha(f, \mathcal{F}) \geq \sqrt{\frac{\varepsilon}{A\delta_P^H}} \cdot (V_1^\star(x_1) - V_1^{\pi^f}(x_1)).$$

Further, $\forall f \in \mathcal{F}$, the exploration radius $c(f, \mathcal{F}) \leq A\delta_P^H/\varepsilon$.

A consequence of Lemma 2 is that when the multiplicative action variation is small, in particular when $\delta_H \leq 1 + 1/H$, we have $\alpha(f, \mathcal{F}) \geq \sqrt{\varepsilon/eA}(V_1^{\pi^\star}(x_1) - V_1^{\pi^f}(x_1))$ and thus myopic exploration converges quickly (see Theorem 1). An extreme case of this is when the next-state distributions do not depend on the chosen action at all and thus $\delta_P = 1$. This represents the contextual bandit setting.

Corollary 2 (MDPs with contextual bandit structure)**.** *Consider MDPs where actions do not affect the next-state distribution, i.e., $P_h(\cdot \mid x, a) = P_h(\cdot \mid x, a')$ for all $x \in \mathcal{X}, (a, a') \in \mathcal{A}^2, h \in [H]$. Then, for any $f \in \mathcal{F}$, the myopic exploration gap of ε-greedy satisfies*

$$\alpha(f, \mathcal{F}) \geq \sqrt{\frac{\varepsilon}{A}} \cdot (V_1^\star(x_1) - V_1^{\pi^f}(x_1)),$$

and the myopic exploration radius satisfies $c(f, \mathcal{F}) \leq A/\varepsilon$.

5.2. Favorable Rewards

In the previous section, we discussed several conditions under which the dynamics make an MDP conducive to efficient

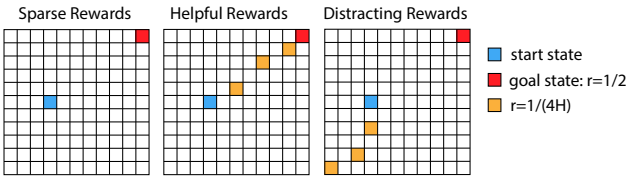

Figure 1. Three grid-world tasks with the same dynamics and optimal policy but different rewards. The optimal policy always moves from the start (blue) to the goal (red). Left: only reward of 1 at the goal. Middle: additional small rewards help myopic exploration. Right: additional small rewards hurt myopic exploration.

myopic exploration, independent of rewards. However the reward structure can also significantly impact the efficiency of myopic exploration. Empirically, this has been widely recognized (Mataric, 1994). The general rule of thumb is that dense-reward problems, where the agent receives a reward signal frequently, are easy for myopic exploration, and sparse-reward problems are difficult. Our myopic exploration gap makes this rule of thumb precise and quantifies when exactly dense rewards are helpful. We illustrate this for ε-greedy exploration in the following example.

5.2.1. Grid World Nagivation Example

Consider a grid-world navigation task with deterministic transitions. The agent is supposed to move from the start to the goal state. Figure 1 depicts this task with 3 different reward functions. The horizon is chosen so that the agent can reach the goal only without any detour ($H = 13$ here).

Sparse Rewards. In the version on the left, the agent only receives a non-zero reward when it reaches the goal for the first time. This is a sparse reward signal since the agent requires many time steps (roughly H) until it receives an informative reward. Exploration with ε-greedy thus requires $O(A^H)$ many samples in order to find an optimal policy. Our results confirm this, since there are many suboptimal greedy policies π^f for which the myopic exploration gap is

$$\alpha(f, \mathcal{F}) = \frac{1}{2}\left(\frac{\varepsilon}{A}\right)^{H/2}.$$

Helpful Dense Rewards. In the version in the middle, someone left breadcrumb every B steps along the shortest path to the goal. The agent receives a small reward of $1/2H$ each time it reaches a breadcrumb for the first time. These intermediate rewards are helpful for ε-greedy which performs well in this problem. Our theoretical results confirm this since any suboptimal greedy policy has a myopic exploration gap of at least

$$\alpha(f, \mathcal{F}) \geq \frac{1}{2H}\left(\frac{\varepsilon}{A}\right)^{B/2}. \tag{3}$$

This bound holds since any suboptimal policy misses out on at least one breadcrumb or the goal and it could reach that and increase its return by $\geq 1/2H$ by changing at most B actions. In this dense-reward setting with $B \ll H$, the corresponding myopic exploration gaps are much larger.

Distracting Dense Rewards. Dense rewards are not always helpful for myopic exploration. Consider the version on the right. Here, the breadcrumbs distract the agent from the goal. In particular, ε-greedy will learn to follow the breadcrumb trail and eventually get stuck in the bottom left corner. It would then require exponentially many episodes to discover the high reward at the goal state. We can show that the myopic exploration gaps are

$$\alpha(f, \mathcal{F}) \begin{cases} = \frac{1}{2}\left(\frac{\varepsilon}{A}\right)^{H/2} & \text{if } \pi^f \text{ reaches all breadcrumbs} \\ \geq \frac{1}{2H}\left(\frac{\varepsilon}{A}\right)^{B/2} & \text{otherwise} \end{cases}$$

for any suboptimal greedy policy π^f. Thus, there are value functions with suboptimal greedy policies that have an exponentially small gap, similar to the sparse reward setting.

5.2.2. Favorable Rewards Through Potential-Based Reward Shaping?

Potential-based reward shaping (Ng et al., 1999; Grzes, 2017) is a popular technique for finding a reward definition that preserve the optimal policies of a given reward function but may be easier to learn. Formally, we here call two average reward definitions $\bar{r} = \{\bar{r}_h\}_{h\in[H]}, \bar{r}' = \{\bar{r}'_h\}_{h\in[H]}$ with $\bar{r}_h, \bar{r}'_h \colon \mathcal{S} \times \mathcal{A} \to [0,1]$ a potential-based reward shaping of each other if $\bar{r}_h(s,a) - \bar{r}'_h(s,a) = \Phi_h(s) - \mathbb{E}[\Phi(s_{h+1}) \mid x_h = x, a_h = a]$ for all s, a, h where $\Phi = \{\Phi_h\}_{h\in[H]}$ is a state-based potential function with $\Phi_1(s_{\text{init}}) = \Phi_{H+1}(s_{\text{end}}) = 0$. One can show by a telescoping sum that the total return of any policies under $\bar{r}$ and $\bar{r}'_h$ is identical. As a result, since the myopic exploration gaps in the tabular setting only depend on the rewards through the return of policies (Equation 2), the gaps are identical under both reward functions. This suggests that the efficiency of ε-greedy exploration is not affected by potential-based reward shaping. Given the empirical success of potential-based reward shaping, this may seem surprising. However, as we illustrate with an example in Appendix A, this difference in empirical performance is due to optimistic or pessimistic initialization effects and not due to ε-greedy exploration.

6. Theoretical Guarantees

In this section, we present our main theoretical guarantees for Algorithm 1. At the heart of our analysis is a sample-complexity bound that controls the number of episodes in which Algorithm 1 selects a poor quality value function $\widehat{f}_t$, for which $\pi^{\widehat{f}_t}$ has large suboptimality. In particular, we

show that for any subset of function $\mathcal{F}' \subset \mathcal{F}$, the number of times for which $\widehat{f}_t \in \mathcal{F}'$ is selected scales inversely with the square of the smallest myopic exploration gap $\alpha(\mathcal{F}', \mathcal{F}) = \inf_{f \in \mathcal{F}'} \alpha(f, \mathcal{F})$.

Intuitively, when $\mathcal{F}'$ contains suboptimal policies and $\alpha(\mathcal{F}', \mathcal{F})$ is large, then for any $f \in \mathcal{F}'$ there exists a policy $\pi' \in \Pi_{\mathcal{F}}$ that achieves higher return than π_f, and the exploration policy $\mathsf{expl}(f)$ will quickly collect enough samples to allow Algorithm 1 to learn this difference. Thus, such an f would not be selected any further. The following theorem formalizes this intuition with the sample complexity bound.

Theorem 1 (Sample Complexity Upper Bound). *Let $\delta \in (0,1)$ and $T \in \mathbb{N}$, and suppose Algorithm 1 is run with a function class $\mathcal{F}$ that satisfies Assumption 1. Further, let $\mathcal{F}' \subseteq \mathcal{F}$ be any subset of value functions. Then, with probability at least $1-\delta$, the number of episodes within the first T episodes where $\widehat{f}_t \in \mathcal{F}'$ is selected is bounded by*

$$O\left(\frac{\ln c(\mathcal{F}', \mathcal{F})}{\alpha(\mathcal{F}', \mathcal{F})^2} H^2 d \ln\left(\frac{\bar{N}_{\mathcal{F}}(T^{-1}) \ln T}{\delta}\right)\right).$$

Here, $d = \dim_{\mathrm{BE}}(\mathcal{F}', \Pi_{\mathcal{F}}, 1/\sqrt{T})$ is the Bellman-Eluder dimension of $\mathcal{F}'$, and

$$\alpha(\mathcal{F}', \mathcal{F}) := \inf_{f \in \mathcal{F}'} \alpha(f, \mathcal{F}), \quad c(\mathcal{F}', \mathcal{F}) := \sup_{f \in \mathcal{F}'} c(f, \mathcal{F})$$

with $\alpha(f, \mathcal{F})$ and $c(f, \mathcal{F})$ defined in Definition 1.

We defer the complete proof of Theorem 1 to Appendix D and provide a brief sketch in Section 6.3. Note that, although Theorem 1 allows arbitrary subsets $\mathcal{F}' \subset \mathcal{F}$, the result is of interest only when $\mathcal{F}'$ contains value functions for which the corresponding greedy policies are suboptimal. When an optimal policy $\pi^\star \in \Pi_{\mathcal{F}'}$, we have that $\alpha(\mathcal{F}', \mathcal{F}) = 0$ and thus, the sample complexity bound is vacuous.

We next compare our result to the prior guarantees in RL with function approximation, specifically with the results of Jin et al. (2020). For any $\lambda \geq 0$, if we instantiate Theorem 1 with $\mathcal{F}'$ consisting of all the value functions that are not λ-optimal, i.e., $\mathcal{F}' = \mathcal{F}_\lambda = \{f \in \mathcal{F} \colon V^{\pi^f}(x_1) \leq V^\star(x_1) - \lambda\}$, we get that within the first

$$\widetilde{O}\left(\frac{\ln c(\mathcal{F}_\lambda, \mathcal{F})}{\alpha(\mathcal{F}_\lambda, \mathcal{F})^2} H^2 d \ln\left(\bar{N}_{\mathcal{F}}\left(\frac{\alpha(\mathcal{F}_\lambda, \mathcal{F})^2}{H^2 d \ln c(\mathcal{F}_\lambda, \mathcal{F})}\right)\right)\right) \quad (4)$$

episodes at least a constant fraction of the chosen greedy policies would be λ-optimal. Thus, terminating Algorithm 1 after collecting the above mentioned number of episodes, and returning the corresponding greedy policy of $\widehat{f}_t$ for a uniformly random choice of t ensures that the output policy is λ-suboptimal with at least a constant probability. Our sample complexity bound in (4) matches the sample complexity of the GOLF algorithm (Jin et al., 2020) in terms of its dependence on H and d. However, our bound replaces their λ dependency with the problem-dependent quantity $\alpha(\mathcal{F}_\lambda, \mathcal{F})/\sqrt{\ln c(\mathcal{F}_\lambda, \mathcal{F})}$ that captures the efficiency of the chosen exploration approach for learning (this dependence is tight as shown by the lower bound in Theorem 2). The comparison is a bit subtle; while our sample complexity bound is typically larger than that of Jin et al. (2020), the provided algorithm is computationally efficient with running time of the order of (4) whenever an efficient square loss regression oracle is available for the class $\mathcal{F}$. On the other hand, Jin et al. (2020) rely on optimistic planning which is typically computationally inefficient.

We can also compare our sample complexity bound to the result of Liu & Brunskill (2018) for pure random exploration. Liu & Brunskill (2018) bound the covering length L and rely on the results of Even-Dar & Mansour (2003) to turn that into a sample-complexity bound for Q-learning. Their sample-complexity bound scales as $\omega\left(\frac{L^3}{\varepsilon^2(1-\gamma)^2}\right)$ while our Theorem 1 in combination with Section 5.1.1 gives $\tilde{O}\left(\frac{LH^2 d \ln \bar{N}_{\mathcal{F}}}{\varepsilon^2}\right)$. A loose translation with $H \approx 1/(1-\gamma)$ and $d \ln \bar{N}_{\mathcal{F}} \lesssim S^2A^2 < L^2$ shows that our results are never worse in tabular MDPs while being much more general (e.g., apply to non-tabular MDPs and capture many other favorable cases).

6.1. Lower Bound

We next provide a lower bound that shows that an $S/\alpha(\mathcal{F}, \mathcal{F})^2$ dependency in tabular MDPs is unavoidable. This suggests that Bellman-Eluder dimension (or alternative notions of statistical complexity such as Bellman-rank, Eluder-dimension, decoupling coefficient, etc., which are all bounded by SA for tabular MDPs) alone are not sufficient to capture the performance of myopic exploration RL algorithms such as ε-greedy.

Theorem 2 (Sample Complexity Lower Bound). *Let $\bar{c} = 3\sqrt{3}/32$. For any given horizon $H \in \mathbb{N}$, number of states $S \geq 8$, number of actions $A \geq 2$, exploration parameter $\varepsilon \in (0,1)$ and $v \in [1, \bar{c}(\varepsilon/A)^{H/2}]$, there exists a tabular MDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, H, P, R)$ with $|\mathcal{A}| = A$ and $|\mathcal{X}| = S$ and a function class $\mathcal{F}$ such that:*

(a) *$\alpha(\mathcal{F}', \mathcal{F}) = \min_{f \in \mathcal{F}'} \alpha(f, \mathcal{F}) \in \left[v\bar{c}\sqrt{\frac{3\varepsilon}{A}}, v\right]$ where $\mathcal{F}' \subset \mathcal{F}$ denotes the set of all the value functions that are at least $1/16$ suboptimal, i.e., $V^\star(x_1) - V^{\pi^f}(x_1) > 1/16$ for any $f \in \mathcal{F}'$;*

(b) *the expected number of episodes for which Algorithm 1 with ε-greedy exploration does not select an $1/16$-optimal function $f \in \mathcal{F}'$ is*

$$\Omega\left(\frac{S}{\alpha(\mathcal{F}', \mathcal{F})^2}\right).$$

6.2. Regret Bound for ε-Greedy RL

Equipped with the sample complexity bound in Theorem 1, we can derive regret bounds for myopic exploration based RL. In the following, we derive regret bounds for ε-greedy algorithm. Note that the regret can be decomposed as

$$\begin{aligned}\mathrm{Reg}(T) &= \sum_{t=1}^{T}[V_1^\star(x_1) - V_1^{\widetilde{\pi}_t}(x_1)] \\ &= \sum_{t=1}^{T}[V_1^\star(x_1) - V_1^{\pi_t}(x_1)] + \sum_{t=1}^{T}[V_1^{\pi_t}(x_1) - V_1^{\widetilde{\pi}_t}(x_1)]\end{aligned}$$

where $\pi_t = \pi^{\widehat{f}_t}$ denotes the greedy policy at episode t. The second term in the above decomposition is bounded by εHT since the return of greedy and exploration policy can differ at most by εH in each episode. The first term denotes the regret of the corresponding greedy policies, which can be controlled by invoking Theorem 1 to bound the number of episodes for which the greedy policies are suboptimal. For favorable learning tasks in which every $f \in \mathcal{F}$ with a suboptimal greedy policy has a significant myopic exploration gap, we can simply invoke Theorem 1 with $\mathcal{F}' = \mathcal{F}^{\mathrm{sub}} = \{f \in \mathcal{F} \colon \pi^f \text{is not optimal}\}$. For illustration, consider the learning task in Figure 1 (middle) where gaps are large (cf. Equation 3). In this case,

$$\begin{aligned}\mathrm{Reg}(T) &= \varepsilon HT + \widetilde{O}\Big(\frac{H^2 d}{\alpha(\mathcal{F}^{\mathrm{sub}}, \mathcal{F})^2}\Big) \\ &= \varepsilon HT + \widetilde{O}\Big(\frac{SA^{1+B}H^4}{\varepsilon^B}\Big),\end{aligned}$$

Setting $\varepsilon \approx A\left(SH^3/T\right)^{1/(B+1)}$ thus yields the bound

$$\mathrm{Reg}(T) = \widetilde{O}\big(HAT^{\frac{B}{B+1}}(SH^3)^{\frac{1}{B+1}}\big)\,.$$

Clearly, as shown by the above example, the optimal choice of ε (and thus the regret) depends on how the myopic exploration gap scales with ε. We formalize this dependence in the following theorem.

Theorem 3 (Regret Bound of ε-Greedy)**.** *Let $T \in \mathbb{N}$ and suppose we run Algorithm 1 with ε-greedy exploration and a function class $\mathcal{F}$ that satisfies Assumption 1. Further, let there be a $h \in [1, H]$ such that for any $\lambda \in [0, 1]$, $\alpha(\mathcal{F}'_\lambda, \mathcal{F}) \geq \Omega((\varepsilon/A)^{h/2}\lambda)$ where $\mathcal{F}'_\lambda \subset \mathcal{F}$ denotes the set of all the value functions that are at least λ suboptimal. Then, with probability at least $1-\delta$, we have,*

$$\mathrm{Reg}(T) \leq \varepsilon HT + \widetilde{O}\Big(\sqrt{\frac{hA^h d H^3 T}{\varepsilon^h}\ln\frac{\bar{N}_{\mathcal{F}}(T^{-1})}{\delta}}\Big),$$

where $d = \dim_{\mathrm{BE}}(\mathcal{F}, \Pi_{\mathcal{F}}, 1/\sqrt{T})$ denotes the Bellman-Eluder dimension of $\mathcal{F}$. Furthermore, setting the exploration parameter $\varepsilon = \widetilde{\Theta}\big(\big(\frac{hHA^h d}{T}\big)^{\frac{1}{2+h}}\big)$, we get that

$$\mathrm{Reg}(T) \leq \widetilde{O}\Big(H^{\frac{h+3}{h+2}}T^{\frac{h+1}{h+2}}\big(hA^h d\ln\frac{\bar{N}_{\mathcal{F}}(T^{-1})}{\delta}\big)^{\frac{1}{h+2}}\Big).$$

For the contextual bandits problems, Corollary 2 implies that $\alpha(\mathcal{F}'_\lambda, \mathcal{F}) \geq \Omega((\varepsilon/A)^{1/2}\lambda)$ and thus $h = 1$. Plugging this in Theorem 3 gives us $\mathrm{Reg}(T) = \widetilde{O}(H^{4/3}A^{1/3}T^{2/3})$, which matches the optimal regret bound for ε-greedy for the contextual bandits problem in terms of dependence on T or A (Lattimore & Szepesvári, 2020). In the worst case for any RL problem, we always have that $\alpha(\mathcal{F}'_\lambda, \mathcal{F}) \geq \Omega((\varepsilon/A)^{H/2}\lambda)$ and for such problems Theorem 3 still results in a sub-linear regret bound of $\widetilde{O}(T^{(H+1)/(H+2)})$.

6.3. Proof Sketch of Theorem 1

We first partition $\mathcal{F}'$ into subsets $\{\mathcal{F}'_1, \dots \mathcal{F}'_{\ln c(\mathcal{F}', \mathcal{F})}\}$ such that the myopic exploration radius $c(f, \mathcal{F})$ is roughly the same for each $f \in \mathcal{F}'_i$. The proof follows by bounding the number of episodes $\mathcal{K}_i = \{t \in [T] \colon \widehat{f}_t \in \mathcal{F}'_i\}$ individually. In order do so, we consider the potential

$$\sum_{t \in \mathcal{K}_i} V^{\pi'_t}(x_1) - V^{\pi_t}(x_1), \tag{5}$$

where $\pi_t = \pi^{\widehat{f}_t}$ and $\pi'_t \in \Pi_{\mathcal{F}}$ denotes the policy that attains the maximum in (1) for $f = \widehat{f}_t$. We will bound this potential from above and below as

$$|\mathcal{K}_i|\alpha(\mathcal{F}'_i, \mathcal{F})\sqrt{c(\mathcal{F}'_i, \mathcal{F})} \leq (5) \leq \widetilde{O}\big(H\sqrt{c(\mathcal{F}'_i, \mathcal{F})|\mathcal{K}_i|d}\big)$$

which implies that $|\mathcal{K}_i| = \widetilde{O}(H^2 d/\alpha(\mathcal{F}'_i, \mathcal{F})^2)$, yielding the desired statement after summing over i. The lower bound follows immediately from the definition of α and c. For the upper bound, we first compose each value difference into expected Bellman errors using

Lemma 3. *Let $f = \{f_h\}_{h \in [H]}$ with $f_h \colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and π^f be the greedy policy of f. Then for any policy $\pi' \in \Pi$,*

$$\begin{aligned}&V_1^{\pi'}(x_1) - V_1^{\pi^f}(x_1) \leq \\ &\sum_{h=1}^{H}\mathbb{E}_{\pi^f}[(\mathcal{E}_h f)(x_h, a_h)] - \sum_{h=1}^{H}\mathbb{E}_{\pi'}[(\mathcal{E}_h f)(x_h, a_h)].\end{aligned}$$

This yields an upper-bound on (5) of

$$\sum_{h=1}^{H}\Big[\underbrace{\sum_{t \in \mathcal{K}_i}|\mathbb{E}_{\pi'_t}[(\mathcal{E}_h\widehat{f}_t)(x_h, a_h)]|}_{(A)} + \underbrace{\sum_{t \in \mathcal{K}_i}|\mathbb{E}_{\pi_t}[(\mathcal{E}_h\widehat{f}_t)(x_h, a_h)]|}_{(B)}\Big]$$

Now, both the terms (A) and (B) can be bound using the standard properties of Bellman-Eluder dimension as:

$$\max\{(B), (A)\} \lesssim \sqrt{d|\mathcal{K}_i|\max_{t \in \mathcal{K}_i}\sum_{\tau \in \mathcal{K}_i \cap [t-1]}(\mathbb{E}_{\pi_\tau}[(\mathcal{E}_h\widehat{f}_t)(x_h, a_h)])^2}.$$

The remaining sum can be controlled using Jensen's inequality and the definition of myopic exploration gap as

$$\sum_{\tau \in \mathcal{K}_i \cap [t-1]}\mathbb{E}_{\pi_\tau}[(\mathcal{E}_h\widehat{f}_t)(x_h, a_h)]^2 \leq \sum_{\tau \in \mathcal{K}_i \cap [t-1]}\mathbb{E}_{\pi_\tau}[(\mathcal{E}_h^2\widehat{f}_t)(x_h, a_h)]$$

$$\leq c(\mathcal{F}_i', \mathcal{F}) \sum_{\tau \in \mathcal{K}_i \cap [t-1]} \mathbb{E}_{\widetilde{\pi}_\tau}[(\mathcal{E}_h^2 \widehat{f}_t)(x_h, a_h)]$$
$$\lesssim c(\mathcal{F}_i', \mathcal{F}) \ln \frac{\bar{N}_{\mathcal{F}}(1/T) \ln(t)}{\delta},$$

where the last inequality follows from a uniform concentration bound for square loss minimization.

7. Related Work

The closest work to ours is Liu & Brunskill (2018) who provide conditions under which uniform exploration yields polynomial sample-complexity in infinite-horizon tabular MDPs, building on the Q-learning analysis of Even-Dar & Mansour (2003) (see Section 5.1.1). Many conditions that enable efficient myopic exploration also allow us to use a smaller horizon for planning in the MDP, a question studied by (Jiang et al., 2016) (e.g., small action variation). However, sufficient conditions for shallow planning are in general not sufficient for efficient myopic exploration and vice versa. One can for example construct cases where planning with horizon $H' = 1$ yields the optimal policy (i.e., $\pi^\star(x_h) = \arg\max_a r(x_h, a)$) for the original horizon H but there are distracting rewards just beyond the shallow planning horizon $H' + 1$ that would throw off algorithms with ε-greedy (see also the example in Appendix A).

Simchowitz & Foster (2020) and Mania et al. (2019) show that simple random noise explores optimally in linear quadratic regulators, but their analysis is tailored specifically to this setting.

There is a rich line of work on understanding the effect of reward functions and on designing suitable rewards, to speed up the rate of convergence of various RL algorithms and to make them more interpretable (Abel et al., 2021; Devidze et al., 2021; Hu et al., 2020; Mataric, 1994; Icarte et al., 2022; Marthi, 2007). While being extremely interesting, the problem of designing suitable reward functions to model the underlying objective is orthogonal to our focus in this paper, which is to understand rate of convergence for myopic exploration algorithms.

Potential based rewards shaping is a popular approach in practice (Ng et al., 1999) to speed up the rate of convergence of RL algorithms. In order to quantify the role of reward shaping, Laud & DeJong (2003) provide an algorithm for which they demonstrate, both theoretically and empirically, improvement in the rate of convergence from reward shaping. However, their algorithm is qualitatively very different from the myopic exploration style algorithms that we consider in this paper, and is in general not efficiently implementable for MDPs with large state spaces. For a more detailed discussion of potential-based reward shaping, see Section 5.2.2 and Appendix A.

Algorithm 1 determines the Q-function estimate by a finite-horizon version of fitted Q-iteration (Ernst et al., 2005). In the infinite horizon setting, this procedure has been analyzed by Munos & Szepesvári (2008); Antos et al. (2007). These works focus on characterizing error propagation of sampling and approximation error and simply assume sampling from a generative model or fixed behavior policy that explores sufficiently, i.e., has good state-action coverage. A recent line of work on offline RL aims to relax the coverage and additional structural assumptions, see e.g. Uehara & Sun (2021) and references therein for an overview. Our work instead has a different focus. We avoid approximation errors by Assumption 1 but aim to characterize the interplay between MDP structure and the different exploration policies and their impact on the efficiency of online RL.

8. Conclusion and Future Work

We view this work as a first step towards fine-grained theoretical guarantees for practical algorithms with myopic exploration. We provided a complexity measure that captures many favorable cases where such algorithms work well. We focused on general results that apply to problems with general function approximation, due to the importance of myopic exploration in such settings. One important direction for future work is an analysis of myopic exploration specialized to tabular MDPs. We believe that a finer characterization of the performance of ε-greedy algorithms in this setting is possible by making explicit assumptions on the initialization of function values in state-action pairs that have not been visited so far. We purposefully avoided such assumptions in our work to avoid conflating myopic exploration mechanisms with those of optimistic initializations which are known to be effective. It would further be interesting to compare the sample complexity of different myopic exploration strategies such as softmax-policies, additive noise or ε-greedy and perhaps develop new myopic strategies with improved bounds.

Acknowledgements

YM received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), the Israel Science Foundation (grant number 993/17), Tel Aviv University Center for AI and Data Science (TAD), and the Yandex Initiative for Machine Learning at Tel Aviv University. KS acknowledges support from NSF CAREER Award 1750575.

References

Abel, D., Dabney, W., Harutyunyan, A., Ho, M. K., Littman, M., Precup, D., and Singh, S. On the expressivity of markov reward. *Advances in Neural Information Processing Systems*, 34, 2021.

Antos, A., Munos, R., and Szepesvári, C. Fitted q-iteration in continuous action-space mdps. 2007.

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Dabney, W., Ostrovski, G., and Barreto, A. Temporally-extended epsilon-greedy exploration. *arXiv preprint arXiv:2006.01782*, 2020.

Dann, C., Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. On oracle-efficient pac rl with rich observations. *arXiv preprint arXiv:1803.00606*, 2018.

Dann, C., Marinov, T. V., Mohri, M., and Zimmert, J. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021a.

Dann, C., Mohri, M., Zhang, T., and Zimmert, J. A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021b.

Devidze, R., Radanovic, G., Kamalaruban, P., and Singla, A. Explicable reward design for reinforcement learning agents. *Advances in Neural Information Processing Systems*, 34, 2021.

Du, S. S., Kakade, S. M., Lee, J. D., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. *arXiv preprint arXiv:2103.10897*, 2021.

Ernst, D., Geurts, P., and Wehenkel, L. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

Even-Dar, E. and Mansour, Y. Learning rates for Q-learning. *Journal of machine learning Research*, 5(1), 2003.

Foster, D. J., Rakhlin, A., Simchi-Levi, D., and Xu, Y. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

Grzes, M. Reward shaping in episodic reinforcement learning. 2017.

Hu, Y., Wang, W., Jia, H., Wang, Y., Chen, Y., Hao, J., Wu, F., and Fan, C. Learning to utilize shaping rewards: A new approach of reward shaping. *Advances in Neural Information Processing Systems*, 33:15931–15941, 2020.

Icarte, R. T., Klassen, T. Q., Valenzano, R., and McIlraith, S. A. Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research*, 73:173–208, 2022.

Jiang, N., Singh, S. P., and Tewari, A. On structural properties of mdps that bound loss due to shallow planning. In *IJCAI*, pp. 1640–1647, 2016.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1704–1713. JMLR. org, 2017.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143, 2020.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman Eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815*, 2021.

Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*, 2018.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Laud, A. and DeJong, G. The influence of reward on the speed of reinforcement learning: An analysis of shaping. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 440–447, 2003.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

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

Mania, H., Tu, S., and Recht, B. Certainty equivalence is efficient for linear quadratic control. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 10154–10164, 2019.

Marthi, B. Automatic shaping and decomposition of reward functions. In *Proceedings of the 24th International Conference on Machine learning*, pp. 601–608, 2007.

Mataric, M. J. Reward functions for accelerated learning. In *Machine learning proceedings 1994*, pp. 181–189. Elsevier, 1994.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (5), 2008.

Ng, A. Y., Harada, D., and Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287, 1999.

Osband, I., Van Roy, B., Russo, D. J., Wen, Z., et al. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019.

Rakhlin, A. and Sridharan, K. Online nonparametric regression with general loss functions. *arXiv preprint arXiv:1501.06598*, 2015.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

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

Simchowitz, M. and Jamieson, K. G. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32:1153–1162, 2019.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. *arXiv preprint arXiv:1811.08540*, 2018.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Wagenmaker, A., Simchowitz, M., and Jamieson, K. Beyond no regret: Instance-dependent pac reinforcement learning. *arXiv preprint arXiv:2108.02717*, 2021.

Wang, R., Salakhutdinov, R. R., and Yang, L. Reinforcement learning with general value function approximation: Provably efficient approach via bounded Eluder dimension. *Advances in Neural Information Processing Systems*, 33, 2020.

Zhang, T. Feel-good thompson sampling for contextual bandits and reinforcement learning. *arXiv preprint arXiv:2110.00871*, 2021.

A. Additional Discussion of Myopic Exploration Gap

To provide further intuition behind the definition of myopic exploration gap in Definition 1, we discuss the following simple example. For any horizon $H \in \mathbb{N}$, consider an MDP with states $\mathcal{S} = [2^H - 1]$ organized in a binary tree. The agent starts at the root of the tree and each action in $\mathcal{A} = \{0, 1\}$ chooses one of the two branches to descent. There is no stochasticity in the transitions and the agent always ends up on one of the leaves of the tree in the final time step of the episode. The function class $\mathcal{F}$ corresponds to the set of all value functions for this tabular MDP. The goal of the agent is to reach a certain leaf state $s_\star$. We consider three different reward functions to formulate this objective (we provide an illustration of the three rewards for $H = 3$ in Figure 2):

Goal reward: The agent only receives a reward when it reaches the state $s_\star$, i.e.,

$$r(s, a) = \mathbb{1}\{s = s_\star\}$$

Let $s'_1, a'_1, s'_2, a'_2 \ldots, s'_{H-1}, a'_{H-1}, s_\star$ be the unique path that leads to $s_\star$, and let $f \in \mathcal{F}$ be any Q-function so that $\pi_f(s'_h) \neq a'_h$ for all $h \in [H]$. The myopic exploration gap of this function for ε-greedy with sufficiently small ε is

$$\alpha(f, \mathcal{F}) = \left(\frac{\varepsilon}{2}\right)^{\frac{H-1}{2}}.$$

This is true because only $\pi_\star$ achieves higher return than π_f but while $\pi_\star$ reaches $s_\star$ with probability 1, $\mathsf{expl}(f)$ only does so with probability $(\varepsilon/2)^{H-1}$ and, hence, $c = (\varepsilon/2)^{H-1}$.

Path reward: The agent receives a positive reward for any right action along the optimal path, i.e.,

$$r'(s, a) = \frac{1}{H}\mathbb{1}\{\exists i \in [H] \colon (s, a) = (s'_i, a'_i)\}.$$

Here, the myopic exploration gap of any $f \in \mathcal{F}$ with suboptimal greedy policy π_f is

$$\alpha'(f, \mathcal{F}) = \frac{\sqrt{\varepsilon}}{H\sqrt{2}}$$

because there is another $f' \in \mathcal{F}$ which is identical to $f \in \mathcal{F}$ except that the values of s'_h, the last state on the desired path taken by π_f, are so that $\pi_{f'}(s'_h) = a'_h$, i.e., $\pi_{f'}$ stays at least one time step longer on the optimal path. As we can see, the myopic exploration gap for this reward formulation is much more favorable. This is similar to the breadcrumb example in Figure 1. Indeed, ε-greedy exploration is much more effective in this formulation.

Potential-based shaping of goal reward: Potential-based reward shaping is a popular technique for designing rewards that may be beneficial to inprove the speed of learning, while retaining the policy preferences of given reward function (Ng et al., 1999). We here consider a reward function that give a reward of $+1$ if the agent takes the first right action and then a -1 whenever it takes a wrong action afterwards. Formally, this reward is

$$r''(s, a) = \begin{cases} +1 & \text{if } s = s'_1 \text{ and } a = a'_1 \\ -1 & \text{if } s = s'_h \text{ for } 1 < h < H \text{ and } a \neq a'_h \\ 0 & \text{otherwise} \end{cases}$$

The potential function $\Phi \colon \mathcal{S} \to \mathbb{R}$ that transforms r into r'' is $\Phi(s) = -\mathbb{1}\{\exists h \in [2, H] \colon s = s'_h\}$. One can verify easily that $r''(s, a) = r(s, a) + \Phi(s) - \mathbb{E}_{s' \sim P(s,a)}[\Phi(s')]$ and that the return of any policy under r and r'' is identical. Note that r'' is not in the range $[0, 1]$. One may apply a linear transformation to all reward functions to normalize r'' in the range $[0, 1]$, however in the following, we work with ranges $[-1, 1]$ for reward and value functions for the ease of exposition since this does not change the argument. Since we are in the tabular setting, we can use Equation 2 to compute the myopic exploration gap. Since the return of each policy under r and r'' is identical, all myopic exploration gaps under r and r'' are identical. This example illustrates the following general fact:

Fact 1. *Myopic exploration gaps are not affected by potential-based reward shaping in tabular Markov decision processes.*

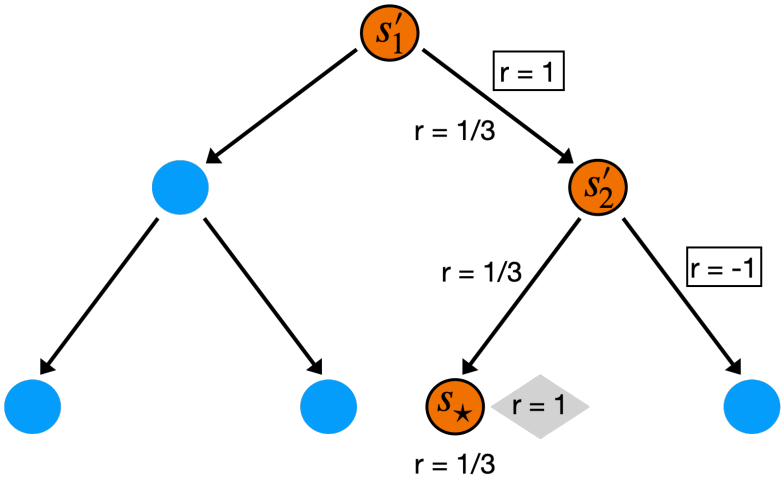

Figure 2. An illustration of different reward function for $H = 3$. Orange colored circles (circles that have a boundary) denote the optimal path $s'_1, s'_2, s_\star$. The *goal rewards* are specified in the shaded diamond. The *path rewards* are specified without any boundary. The *potential based shaped rewards* are specified in rectangular boxes. Corresponding to each reward type (given by shapes: diamond, none, boxes) any unspecified edges denote a reward of 0.

Here, this implies that the myopic exploration gap under r'' of any Q-function $f \in \mathcal{F}$ with $\pi_f(s'_h) \neq a'_h$ for all $h \in [H]$ is

$$\alpha''(f, \mathcal{F}) = \left(\frac{\varepsilon}{2}\right)^{\frac{H-1}{2}}.$$

This suggests that the sample complexity of ε-greedy exploration in this example is exponential in H. This may be surprising since the myopic greedy policy that maximizes only the immediate reward, i.e. takes actions $\arg\max_{a \in \mathcal{A}} r''(s, a)$, is optimal in this problem. This implies that shallow planning with horizon 1 is sufficient in this problem and may suggest that the sample complexity is *not* exponential in H. How can this conundrum be resolved and what is the sample-complexity of ε-greedy in this problem?

We will argue in the following that the efficiency of ε-greedy in this problem depends how we initialize the Q-function estimate of state-action pairs that have not been visited. There are initializations under which the sample-complexity is polynomial or exponential in H respecitvely. We therefore conclude that the benefit of potential-based reward shaping in this case is not due to exploration through ε-greedy but rather optimism or pessimism in the initializations. Note that our procedure in Algorithm 1 does not prescribe an initialization (all initializations are minimizers of $\mathcal{L}_h(f_{t,h+1})$).

First consider initializing the Q-function table with all entries to be equal to 0. In this case, the agent will try action a'_1 in state s'_1 after at most $\Omega(1/\varepsilon)$ episodes. Independent of which actions it took afterwards, the V-value estimate for s'_2 will be 0 and will always remain this value since everything is deterministic in this problem and 0 is the correct value. As a result the Q-function estimate for (s'_1, a'_1) is $+1$ and the greedy policy will take a'_1 in s'_1. Now, in any episode where the agent actually follows the greedy policy, it learns that one action that deviates from the optimal path is suboptimal. Hence, after $\mathrm{O}(H/\varepsilon)$ episodes, the algorithm will choose the optimal policy as its greedy policy and even learn the optimal value function. Hence, the sample-complexity of ε-greedy with this intialization is indeed polynomial in H. Interestingly, the ε-greedy with the same 0-initialization has exponential sample complexity for the original goal-based rewards r. This is because none of the Q-function estimates changes until the agent first reaches $s_\star$. Essentially, the same initialization is pessimistic under the original reward function r but optimistic under the shaped version r''.

Now consider initializing the Q-function table with all entries to be equal to -1. The agent will try action a'_1 in state s'_1 after at most $\Omega(1/\varepsilon)$ episodes. Unless it happens to exactly follow the optimal path, which only happens with probability $(1/2)^{H-1}$, the agent will learn to associate a Q-value of 0 for the initial action and a Q-value of -1 for all actions along the optimal path afterwards. Hence, it has no preference between the actions in any state of the optimal path (except for the first action) and would still need at least $(1/2)^{H-2}$ episodes to randomly follow the optimal path and discover the reward of 0 in the final state. Hence, the sample-complexity of ε-greedy with this initialization is indeed exponential in H, since there was no optimism in the initialization and ε-greedy was ineffective at exploring for this problem.

B. Proofs For Bounds on Myopic Exploration Gap in Section 5

Lemma 1. *When $Q^\star \in \mathcal{F}$, the myopic exploration gap is bounded for all $f \in \mathcal{F}$ as*

$$\left(\max_{\pi' \in \Pi'} \left\| \frac{\mathbb{P}_{\pi'}}{\mathbb{P}_{\mathsf{expl}(f)}} \right\|_\infty \right)^{-\frac{1}{2}} \leq \frac{\alpha(f, \mathcal{F})}{V_1^\star(x_1) - V_1^{\pi^f}(x_1)} \leq 1.$$

Furthermore, the myopic exploration radius is bounded as $c(f, \mathcal{F}) \leq \max_{\pi' \in \Pi'} \left\| \frac{\mathbb{P}_{\pi'}}{\mathbb{P}_{\mathsf{expl}(f)}} \right\|_\infty$.

Proof. For the upper-bound, note that the objective of (1) is bounded for all $c \geq 1$ and $\pi' \in \Pi_{\mathcal{F}} \subseteq \Pi$, as

$$\begin{aligned} \frac{1}{\sqrt{c}}(V_1^{\pi'}(x_1) - V_1^{\pi^f}(x_1)) &\leq V_1^{\pi'}(x_1) - V_1^{\pi^f}(x_1) \\ &\leq V_1^\star(x_1) - V_1^{\pi^f}(x_1). \end{aligned}$$

For the lower-bound, we show that $\pi' = \pi^\star$ and $c = \max_{\pi' \in \Pi'} \left\| \frac{\mathbb{P}_{\pi'}}{\mathbb{P}_{\mathsf{expl}(f)}} \right\|_\infty$ is a feasible solution for (1). Since $Q^\star \in \mathcal{F}$ by realizability, any optimal policy $\pi^\star \in \Pi_{\mathcal{F}} = \Pi'$ is feasible as the expected bellman error for f' corresponding to π^* is equal to 0. Further, for any $\pi' \in \Pi'$ and function $g \colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}^+$ we have

$$\mathbb{E}_{\pi'}[g(x_h, a_h)] \leq \mathbb{E}_{\mathsf{expl}(f)}[g(x_h, a_h)] \left\| \frac{\mathbb{P}_{\pi'}}{\mathbb{P}_{\mathsf{expl}(f)}} \right\|_\infty$$

which shows that the value of c is feasible. □

Corollary 1. *For any MDP with finitely many actions with $|\mathcal{A}| = A$ and $f \in \mathcal{F}$, the myopic exploration gap of ε-greedy can be bounded as*

$$\alpha(f, \mathcal{F}) \geq \left(\frac{\varepsilon}{A}\right)^{\frac{H}{2}} (V_1^\star(x_1) - V_1^{\pi^f}(x_1))$$

and the exploration radius is bounded as $c(f, \mathcal{F}) \leq \left(\frac{\varepsilon}{A}\right)^H$.

Proof. The likelihood ratio of an episode $\tau = (x_1, a_1, r_1, \ldots, x_H, a_H, r_H, x_{H+1})$ w.r.t. $\mathsf{expl}(f)$ and any other policy $\bar{\pi} \in \Pi$ satisfies

$$\frac{\mathbb{P}_{\bar{\pi}}(\tau)}{\mathbb{P}_{\mathsf{expl}(f)}(\tau)} \leq \prod_{h=1}^H \frac{\bar{\pi}(a_h|x_h)}{\mathsf{expl}(f)(a_h|x_h)} \leq \prod_{h=1}^H \frac{1}{\varepsilon/A} = \left(\frac{A}{\varepsilon}\right)^H.$$

The result then follows from Lemma 1. □

Lemma 2. *Let $\mathcal{M}$ be any MDP with multiplicative action variation δ_P, and $Q^* \in \mathcal{F}$. Then, for any $f \in \mathcal{F}$, the myopic exploration gap of ε-greedy satisfies*

$$\alpha(f, \mathcal{F}) \geq \sqrt{\frac{\varepsilon}{A\delta_P^H}} \cdot (V_1^\star(x_1) - V_1^{\pi^f}(x_1)).$$

Further, $\forall f \in \mathcal{F}$, the exploration radius $c(f, \mathcal{F}) \leq A\delta_P^H/\varepsilon$.

Proof. First note that for any policy π, the occupancy measure for state x_h and action a_h is given by

$$\mu_h^\pi(x_h, a_h) = \sum_{x_1, \ldots, x_{h-1}} \left(\prod_{j=1}^{h-1} \sum_{a \in \mathcal{A}} \Pr(\pi(x_j) = a) P_j(x_{j+1} \mid x_j, a) \right) \Pr(\pi(x_h) = a_h), \tag{6}$$

where P denotes the transition dynamics corresponding to $\mathcal{M}$ and we used the fact that x_1 is fixed. Next, note that as a consequence of Definition 3, we have that for any x_j, x_{j+1}, a and a',

$$P_j(x_{j+1} \mid x_j, a) \leq \delta_P P_j(x_{j+1} \mid x_j, a').$$

Thus, we have that:

$$\begin{aligned} P_j(x_{j+1} \mid x_j, a) &\leq \sum_{a' \in \mathcal{A}} \Pr(\mathsf{expl}(f)(x_j) = a') P_j(x_{j+1} \mid x_j, a) \\ &\leq \delta_P \Big(\sum_{a' \in \mathcal{A}} \Pr(\mathsf{expl}(f)(x_j) = a') P_j(x_{j+1} \mid x_j, a') \Big). \end{aligned}$$

Plugging the above in (6), we get that:

$$\begin{aligned} \mu_h^\pi(x_h, a_h) &\leq \sum_{x_1, \ldots, x_{h-1}} \Big(\prod_{j=1}^{h-1} \sum_{a \in \mathcal{A}} \Pr(\pi(x_j) = a) \delta_P \Big(\sum_{a' \in \mathcal{A}} \Pr(\mathsf{expl}(f)(x_j) = a') P_j(x_{j+1} \mid x_j, a') \Big) \Big) \Pr(\pi(x_1) = a) \\ &= \sum_{x_1, \ldots, x_{h-1}} \Big(\prod_{j=1}^{h-1} \delta_P \Big(\sum_{a' \in \mathcal{A}} \Pr(\mathsf{expl}(f)(x_j) = a') P_j(x_{j+1} \mid x_j, a') \Big) \Big) \Pr(\pi(x_h) = a_h) \\ &= \delta_P^{h-1} \sum_{x_1, \ldots, x_{h-1}} \Big(\prod_{j=1}^{h-1} \sum_{a' \in \mathcal{A}} \Pr(\mathsf{expl}(f)(x_j) = a') P_j(x_{j+1} \mid x_j, a') \Big) \Pr(\pi(x_h) = a_h). \end{aligned}$$

Next, note that $\mathsf{expl}(f)$ is the ε-greedy policy and thus $\Pr(\mathsf{expl}(x_h) = a_h) \geq \varepsilon / A$. Using this fact in the above bound, we get that

$$\begin{aligned} \mu_h^\pi(x_h, a_h) &\leq \delta_P^{h-1} \sum_{x_1, \ldots, x_{h-1}} \Big(\prod_{j=1}^{h-1} \sum_{a' \in \mathcal{A}} \Pr(\mathsf{expl}(f)(x_j) = a') P_j(x_{j+1} \mid x_j, a') \Big) \frac{A \Pr(\mathsf{expl}(x_h) = a_h)}{\varepsilon} \\ &= \frac{A \delta_P^{h-1}}{\varepsilon} \sum_{x_1, \ldots, x_{h-1}} \Big(\prod_{j=1}^{h-1} \sum_{a' \in \mathcal{A}} \Pr(\mathsf{expl}(f)(x_j) = a') P_j(x_{j+1} \mid x_j, a') \Big) \Pr(\mathsf{expl}(x_h) = a_h) \\ &\leq \frac{A \delta_P^H}{\varepsilon} \mu_h^{\mathsf{expl}(f)}(x_h, a_h), \end{aligned} \tag{7}$$

where the inequality in the last line holds because $\delta_P \geq 1$ and by using the definition of $\mu_h^{\mathsf{expl}(f)}(x_h, a_h)$ from (6).

Observe that (7) holds for any policy π. Thus, using this relation for $\pi^\star$, we get that

$$\begin{aligned} \mathbb{E}_{\pi^\star}[(\mathcal{E}_h^2 f')(x_h, a_h)] &= \sum_{x_h, a_h} \mu_h^{\pi^\star}(x_h, a_h) \cdot (\mathcal{E}_h^2 f')(x_h, a_h) \\ &\leq \frac{A \delta_P^H}{\varepsilon} \mu_h^{\mathsf{expl}(f)}(x_h, a_h) \cdot (\mathcal{E}_h^2 f')(x_h, a_h) \\ &\leq \frac{A \delta_P^H}{\varepsilon} \mathbb{E}_{\mathsf{expl}(f)}[(\mathcal{E}_h^2 f')(x_h, a_h)]. \end{aligned} \tag{8}$$

A similar analysis reveals that

$$\mathbb{E}_{\pi^f}[(\mathcal{E}_h^2 f')(x_h, a_h)] \leq \frac{A \delta_P^H}{\varepsilon} \mathbb{E}_{\mathsf{expl}(f)}[(\mathcal{E}_h^2 f')(x_h, a_h)]. \tag{9}$$

The relations in (8) and (9) thus imply that $\pi^\star \in \Pi'$ satisfies the constraints in the definition of $\alpha(f, \mathcal{F})$ with $c = \frac{A \delta_P^H}{\varepsilon}$. We thus have that

$$\alpha(f, \mathcal{F}) \geq \frac{1}{\sqrt{c}} (V_1^{\pi^\star}(x_1) - V_1^{\pi^f}(x_1)) = \sqrt{\frac{\varepsilon}{A \delta_P^H}} \cdot (V_1^{\pi^\star}(x_1) - V_1^{\pi^f}(x_1)).$$

□

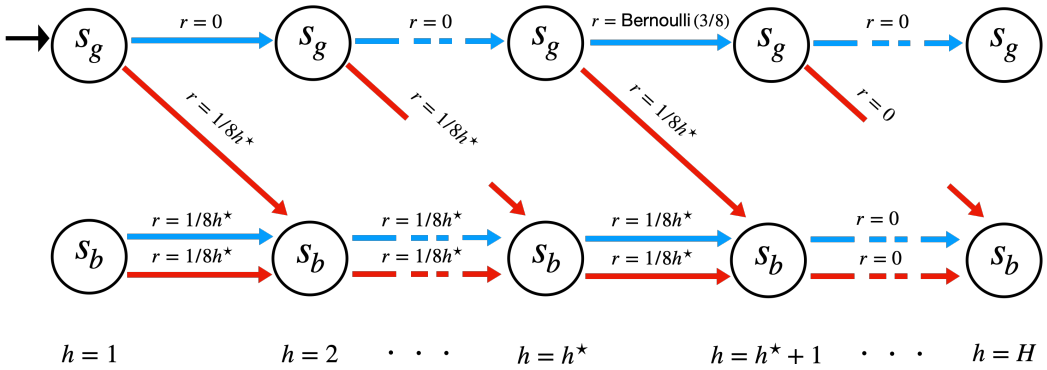

Figure 3. MDP constructon for the proof of Theorem 2. There are two states $\{s_g, s_b\}$ and A action. The black arrow denotes the starting state. The blue arrow denotes the deterministic transition when the agent takes action $a = 1$, and the red arrow denotes deterministic transition when the agent takes any other action in $[A] \setminus \{1\}$. As long as the agent picks actions 1, it stays on the good chain s_g but as soon as it chooses any other actions it transitions to the bad chain s_b where it will stay forever. The observed rewards are time dependent and are shown over the corresponding action arrows.

C. Proof of Sample Complexity Lower Bound

Theorem 2 (Sample Complexity Lower Bound)**.** *Let $\bar{c} = 3\sqrt{3}/32$. For any given horizon $H \in \mathbb{N}$, number of states $S \geq 8$, number of actions $A \geq 2$, exploration parameter $\varepsilon \in (0, 1)$ and $v \in [1, \bar{c}(\varepsilon/A)^{H/2}]$, there exists a tabular MDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, H, P, R)$ with $|\mathcal{A}| = A$ and $|\mathcal{X}| = S$ and a function class $\mathcal{F}$ such that:*

(a) $\alpha(\mathcal{F}', \mathcal{F}) = \min_{f \in \mathcal{F}'} \alpha(f, \mathcal{F}) \in \left[v\bar{c}\sqrt{\frac{3\varepsilon}{A}}, v\right]$ *where $\mathcal{F}' \subset \mathcal{F}$ denotes the set of all the value functions that are at least $1/16$ suboptimal, i.e., $V^\star(x_1) - V^{\pi^f}(x_1) > 1/16$ for any $f \in \mathcal{F}'$;*

(b) *the expected number of episodes for which Algorithm 1 with ε-greedy exploration does not select an $1/16$-optimal function $f \in \mathcal{F}'$ is*

$$\Omega\Big(\frac{S}{\alpha(\mathcal{F}', \mathcal{F})^2}\Big).$$

Proof. We first will show the statement for $S = 2$ and then extend the proof to $S > 2$.

Construction for $S = 2$. Let $A \in \mathbb{N}$, $H \in \mathbb{N}$ and ε be fixed and consider states $\mathcal{X} = \{s_g, s_b\}_{h \in [H+1]}$ and actions $\mathcal{A} = [A]$. The agent always starts in state $x_{\text{init}} = s_g$. The dynamics is deterministic, non-stationary and defined as

$$P_h(s_g|s_g, a = 1) = 1 \qquad P_h(s_b|s_g, a \neq 1) = 1 \qquad P_h(s_b|s_b, a \in \mathcal{A}) = 1$$

for $h \in [H]$. All other transitions have probability 0. Essentially, the state space forms two chains, and the agent always progresses on a chain. As long as the agent picks actions 1, it stays on the good chain s_g but as soon as it chooses any other actions it transitions to the bad chain s_b where it will stay forever. The function class $\mathcal{F} = \{\mathcal{F}_h\}_{h \in [H]}$ is the full tabular class, i.e., $\mathcal{F}_h = \mathcal{X} \times \mathcal{A} \to [0, H + 1 - h]$ for $h \in [H]$. Now, for the given value v, we define the reward distributions $R = \{R_h\}_{h \in [H]}$ as

$$R_h(x, a) = \begin{cases} 0 & \text{if } h > h^\star \\ \text{Bernoulli}\left(\frac{3}{8}\right) & \text{if } h = h^\star \text{ and } x = s_g \text{ and } a = 1 \\ 0 & \text{if } h < h^\star \text{ and } x = s_g \text{ and } a = 1 \\ \frac{1}{8h^\star} & \text{otherwise} \end{cases}$$

where $h^\star = \lceil 2 \log_{\varepsilon/A}(4v) \rceil \leq H$. We illustrate the transition dynamics and the corresponding reward function in Figure 3.

Value of myopic exploration gap. We now show that the smallest non-zero exploration gap is v up to a constant factor in this MDP. The value of a deterministic policy is only determined by how long it stays in s_g (by always taking action $a = 1$). For any $\pi \in \Pi_{\text{det}}$, let $L(\pi) \in [H]$ denote that length. We have

$$V^\pi(x_1) = \frac{3}{8}\mathbb{1}\{L(\pi) \geq h^\star\} + \frac{1}{8}\left(1 - \frac{L(\pi)}{h^\star}\right)\mathbb{1}\{L(\pi) < h^\star\}$$

The value of $\alpha(f, \mathcal{F}, \Pi_\mathcal{F}, \mathsf{expl}, \mathcal{M})$ can be lower-bounded for any $f \in \mathcal{F}$ with $L(\pi^f) < h^\star$ by considering $\pi' = \pi^\star$ in (1). This gives

$$\begin{aligned}
\alpha(f, \mathcal{F}, \Pi_\mathcal{F}, \mathsf{expl}, \mathcal{M}) &\geq \sqrt{\mathbb{P}_{\mathsf{expl}(f)}(a_1 = 1, \dots, a_{h^\star} = 1)}\left(\frac{3}{8} - \frac{1}{8}\left(1 - \frac{L(\pi^f)}{h^\star}\right)\right) \\
&= \left(\frac{\varepsilon}{A}\right)^{\frac{h^\star - L(\pi^f)}{2}}\left(\frac{1}{4} + \frac{1}{8}\frac{L(\pi^f)}{h^\star}\right) \\
&\geq \frac{1}{4}\left(\frac{\varepsilon}{A}\right)^{h^\star/2}.
\end{aligned}$$

Further, both inequalities are tight for any $f \in \mathcal{F}$ with $f_h(s_g, 1) < \max_{a \in \mathcal{A}\setminus\{1\}} f_h(s_g, a)$ for all $h \leq h^\star$, i.e., value functions for which the greedy policy would always choose to go to s_b in the first $h^\star$ steps. We therefore have shown that

$$\min_{f \in \mathcal{F}\colon \alpha(f, \mathcal{F}, \Pi_\mathcal{F}, \mathsf{expl}, \mathcal{M}) > 0} \alpha(f, \mathcal{F}, \Pi_\mathcal{F}, \mathsf{expl}, \mathcal{M}) = \frac{1}{4}\left(\frac{\varepsilon}{A}\right)^{h^\star/2} \in \left[v\frac{1}{4}\sqrt{\frac{\varepsilon}{A}}, v\right].$$

Performance of ε-greedy. Since the regression loss in Algorithm 1 accesses function values $\widehat{f}_{h+1}(x', a')$ at state-action pairs (x', a') for which the algorithm has never chosen a' in x', the behavior of ε-greedy depends on their default value. To avoid a bias towards optimistic or pessimistic intialization, we assume that the datasets $\mathcal{D}_h$ in Algorithm 1 are initialized with one sample transition (x, a, r, x') from each $(x, a) \in \mathcal{X} \times \mathcal{A}$. An alternative to this assumption is to simply define the value function class $\mathcal{F}$ given to the agent to be restricted to only those functions that match the optimal value as soon as an action $a \neq 1$ was taken, i.e.,

$$\mathcal{F}_h = \{f\colon \mathcal{X} \times \mathcal{A} \to [0, 1] \mid f_h(x, a) = Q^\star_h(x, a) \forall (x, a) \text{ with } x = s_b \text{ or } a \neq 1\}$$

In this case, the argument below applies as soon as the agent visits $(s_g, 1)$ at time $h^\star$ for the first time.

Note that the MDP $\mathcal{M}$ is deterministic with the exception of the reward at $(s_g, 1)$ at time $h^\star$. Therefore, only two possible intializations are possible. With probability at least $1/4$, the algorithm was initialized with $(s_g, 1, 0, s_b)$ at time $h^\star$. In this case, we have

$$\widehat{f}_h(s_g, a) = \begin{cases} \frac{h^\star - h}{8h^\star} & \text{if } a = 1 \\ \frac{h^\star - h + 1}{8h^\star} & \text{if } a \neq 1 \end{cases}$$

for all $h \leq h^\star$ and $\pi^{\widehat{f}}$ would always choose a wrong action. Unless the agent receives a new sample from state-action pair $(s_g, 1)$ at time $h^\star$, this estimate will also not change since all other observations are deterministic. The probability with which $\mathsf{expl}(\widehat{f})$ will receive such a sample in an episode is $(\varepsilon/A)^{h^\star}$ and thus, the agent will require $\Omega(1/v^2)$ samples in expectation before it can switch to a different function.

Extension to $S \geq 8$. Without loss of generality, we can assume that S is even, otherwise just choose $S \to S - 1$. We then create $n = S/2$ copies of the 2-state MDP described above. The initial state distribution is uniform over all copies of s_g.[2] The value of any deterministic policy is still only determined by how long it stays in $s_{g,i}$, each copy of s_g,

$$\mathbb{E}_\pi[V^\pi(x_1)] = \frac{3}{8} \cdot \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{L_i(\pi) \geq h^\star\} + \frac{1}{8} \cdot \frac{1}{n}\sum_{i=1}^n \left(1 - \frac{L(\pi)}{h^\star}\right)\mathbb{1}\{L(\pi) < h^\star\}$$

[2] If we desire a deterministic start state, we can just increase the horizon $H \leftarrow H + 1$ by one and have all actions transition uniformly to all copies in $h = 1$.

where $L_i(\pi)$ is the number of time steps the agent stays in $s_{g,i}$ for each copy i. The instantaneous regret of π is then

$$\begin{aligned}\mathbb{E}_\pi[V^\star(x_1) - V^\pi(x_1)] &= \frac{3}{8} - \frac{1}{8}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(1 + \frac{L(\pi)}{h^\star}\right)\mathbb{1}\{L(\pi) < h^\star\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{4} + \frac{L(\pi)}{8h^\star}\right)\mathbb{1}\{L(\pi) < h^\star\} \\ &\geq \frac{1}{4}\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{L(\pi) < h^\star\}.\end{aligned}$$

Thus, any policy π with instantaneous regret at least $1/16$ needs to behave suboptimally in at least $3n/4$ of the n copies. Let $\mathcal{F}' = \{f \in \mathcal{F} \colon \mathbb{E}[V^\star(x_1) - V^{\pi^f}(x_1)] > 1/16\}$ be all functions that have instantaneous regret at least $1/16$. We can lower bound the myopic exploration gap for any $f \in \mathcal{F}'$ by considering $\pi' = \pi^\star$ in (1). This gives

$$\begin{aligned}\alpha(f, \mathcal{F}, \Pi_\mathcal{F}, \mathsf{expl}, \mathcal{M}) &\geq \sqrt{\mathbb{P}_{\mathsf{expl}(f)}(a_1 = 1, \dots, a_{h^\star} = 1)}\frac{1}{4n}\sum_{i=1}^{n}\mathbb{1}\{L_i(\pi^f) < h^\star\} \\ &\geq \sqrt{\mathbb{P}_{\mathsf{expl}(f)}(a_1 = 1, \dots, a_{h^\star} = 1)}\frac{3}{16} \\ &\geq \sqrt{\frac{3n}{4n}\left(\frac{\varepsilon}{A}\right)^{h^\star}}\frac{3}{16} = \frac{3\sqrt{3}}{32}\left(\frac{\varepsilon}{A}\right)^{h^\star/2}\end{aligned}$$

Further, for $f \in \mathcal{F}$ that behave optimally in $1/4n$ copies and choose $a_1 \neq 1$ in all other copies, all inequalities are tight. Hence,

$$\min_{f \in \mathcal{F}'} \alpha(f, \mathcal{F}, \Pi_\mathcal{F}, \mathsf{expl}, \mathcal{M}) = \frac{3\sqrt{3}}{32}\left(\frac{\varepsilon}{A}\right)^{h^\star/2} \in \left[v\frac{3}{32}\sqrt{\frac{3\varepsilon}{A}}, v\right],$$

when we choose $h^\star = \lceil 2\log_{\varepsilon/A}(32v/(3\sqrt{3})\rceil \leq H$.

Using the same intialization of datasets $\mathcal{D}_h$ of Algorithm 1 as in the $S = 2$ case, each copy i of the MDP has probability $p = 3/8$ to be initialized with $(s_{g,i}, 1, 1, s_{g,i})$. By Hoeffding bound, the probability that at least $n/2$ copies are initialized in such way is at most $\mathbb{P}(\sum_{i=1}^n X_i - pn > n/8) \leq \exp(-2n/64)$. Thus, with probability at least $1 - \exp(-2n/64) \geq 1/5$, there are at least $n/2$ copies i which are initialized with reward 0 for state-action pair $(s_{g,i}, 1)$. As in the $S = 2$ case, for each i, the agent needs to collect another sample from this state-action pair which only happens with probability $(\varepsilon/A)^{h^\star}$ per copy. Therefore, the probability that the agent receives an informative sample for any of the suboptimal copies is bounded by $(\varepsilon/A)^{h^\star}$ and the agent needs to collect at least $n/8$ samples before the greedy policy can become $1/16$-optimal. Hence, the expected number of times until this happen is at least

$$\frac{1}{5}\cdot\frac{n}{8}\cdot(A/\varepsilon)^{h^\star} = \Omega\left(\frac{S}{\alpha(\mathcal{F}', \mathcal{F})^2}\right)$$

which completes the proof. □

D. Proofs for Regret and Sample Complexity Upper Bounds

Theorem 1 (Sample Complexity Upper Bound)**.** *Let $\delta \in (0,1)$ and $T \in \mathbb{N}$, and suppose Algorithm 1 is run with a function class $\mathcal{F}$ that satisfies Assumption 1. Further, let $\mathcal{F}' \subseteq \mathcal{F}$ be any subset of value functions. Then, with probability at least $1-\delta$, the number of episodes within the first T episodes where $\widehat{f}_t \in \mathcal{F}'$ is selected is bounded by*

$$O\left(\frac{\ln c(\mathcal{F}', \mathcal{F})}{\alpha(\mathcal{F}', \mathcal{F})^2}H^2 d \ln\left(\frac{\bar{N}_\mathcal{F}(T^{-1})\ln T}{\delta}\right)\right).$$

Here, $d = \dim_{\mathrm{BE}}(\mathcal{F}', \Pi_\mathcal{F}, 1/\sqrt{T})$ is the Bellman-Eluder dimension of $\mathcal{F}'$, and

$$\alpha(\mathcal{F}', \mathcal{F}) := \inf_{f \in \mathcal{F}'} \alpha(f, \mathcal{F}), \quad c(\mathcal{F}', \mathcal{F}) := \sup_{f \in \mathcal{F}'} c(f, \mathcal{F})$$

with $\alpha(f, \mathcal{F})$ and $c(f, \mathcal{F})$ defined in Definition 1.

Proof. We partition $\mathcal{F}'$ into $\mathcal{F}' = \mathcal{F}'_1 \cup \cdots \cup \mathcal{F}'_{i_{\max}}$ with $\mathcal{F}'_i = \{f \in \mathcal{F}' \colon c(f, \mathcal{F}, \Pi_\mathcal{F}, \mathsf{expl}, \mathcal{M}) \in [e^{i-1}, e^i]\}$ and $i_{\max} = \lceil \ln \sup_{f \in \mathcal{F}'} c(f, \mathcal{F}, \Pi_\mathcal{F}, \mathsf{expl}, \mathcal{M}) \rceil$ and denote by $\mathcal{K}_{i,T} = \{t \in [T] \colon \widehat{f}_k \in \mathcal{F}'_i\}$ the (random) set of episodes in $[T]$ where the Q-function estimate for the episode is in the i-th part. To keep the notation concise, we denote for each $t \in \mathbb{N}$

- π_t as the greedy policy of $\widehat{f}_t$, i.e., $\pi_t = \pi^{\widehat{f}_t}$
- $\widetilde{\pi}_t$ as the exploration policy in episode t, i.e., $\widetilde{\pi}_k = \mathsf{expl}(\widehat{f}_t)$
- π'_t as the improvement policy $\pi' \in \Pi_\mathcal{F}$ that attains the maximum in the myopic exploration gap definition for $\widehat{f}_t$ (defined in (1)).

The total difference in return between the greedy and improvement policies can be bounded using Lemma 3 as

$$\sum_{t \in \mathcal{K}_{i,T}} \left(V_1^{\pi'_t}(x_1) - V_1^{\pi_t}(x_1) \right) \leq \sum_{t \in \mathcal{K}_{i,T}} \sum_{h=1}^{H} \mathbb{E}_{\pi_t}[(\mathcal{E}_h \widehat{f}_t)(x_h, a_h)] - \sum_{t \in \mathcal{K}_{i,T}} \sum_{h=1}^{H} \mathbb{E}_{\pi'_t}[(\mathcal{E}_h \widehat{f}_t)(x_h, a_h)] \tag{10}$$

Using the completeness assumption in Assumption 1, we show in Lemma 4 that with probability at least $1 - \delta$ for all $(h, t) \in [H] \times \mathbb{N}$

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

where $N'_\mathcal{F}(1/T) = \sum_{h=1}^{H} N_{\mathcal{F}_h}(1/T) N_{\mathcal{F}_{h+1}}(1/T)$. In the following, we consider only the event where this condition holds. Leveraging the definition of $c(f, \mathcal{F}, \Pi_\mathcal{F}, \mathsf{expl}, \mathcal{M})$, we bound

$$\begin{aligned}
\sum_{\tau \in \mathcal{K}_{i,t-1}} \mathbb{E}_{\pi'_\tau}[(\mathcal{E}_h \widehat{f}_t)(x_h, a_h)]^2 &\leq \sum_{\tau \in \mathcal{K}_{i,t-1}} \mathbb{E}_{\pi'_\tau}[(\mathcal{E}_h^2 \widehat{f}_t)(x_h, a_h)] && \text{(Jensen's inequality)} \\
&\leq \sum_{\tau=1}^{t-1} \mathbb{E}_{\pi'_\tau}[(\mathcal{E}_h^2 \widehat{f}_t)(x_h, a_h)] \leq e^i \sum_{\tau=1}^{t-1} \mathbb{E}_{\widetilde{\pi}_\tau}[(\mathcal{E}_h^2 \widehat{f}_t)(x_h, a_h)] \\
&\leq 179 e^i \ln \frac{6 N'_\mathcal{F}(1/T) \ln(2t)}{\delta}.
\end{aligned}$$

Using the distributional Eluder dimension machinery in Lemma 5, this implies that

$$\sum_{t \in \mathcal{K}_{i,T}} |\mathbb{E}_{\pi'_t}[(\mathcal{E}_h \widehat{f}_t)(x_h, a_h)]| \leq O\left(\sqrt{e^i d(\mathcal{F}'_i) \ln \frac{N'_\mathcal{F}(1/T) \ln(T)}{\delta} |\mathcal{K}_{i,T}|} + \min\{|\mathcal{K}_{i,T}|, d(\mathcal{F}'_i)\} \right)$$

where $d(\mathcal{F}') = \dim_{BE}(\mathcal{F}', \Pi_{\mathcal{F}'}, T^{-1/2}) = \max_{h \in [H]} \dim_{DE}(\mathcal{F}'_h - \mathcal{K}_h \mathcal{F}'_h, \Pi_{\mathcal{F}'}, T^{-1/2})$ is the Bellman-Eluder dimension. Applying the arguments above verbatim, we can derive the same upper-bound for $\sum_{t \in \mathcal{K}_{i,T}} |\mathbb{E}_{\pi_t}[(\mathcal{E}_h^2 \widehat{f}_t)(x_h, a_h)]|$. Plugging the above two bounds in Equation 10, we obtain

$$\sum_{t \in \mathcal{K}_{i,T}} \left[V_1^{\pi'_t}(x_1) - V_1^{\pi_t}(x_1) \right] \leq O\left(\sqrt{e^i H^2 d(\mathcal{F}'_i) \ln \frac{N'_\mathcal{F}(1/T) \ln(T)}{\delta} |\mathcal{K}_{i,T}|} + H d(\mathcal{F}'_i) \right).$$

Using the myopic exploration gap in Definition 1, we lower-bound the LHS as

$$\sum_{t \in \mathcal{K}_{i,T}} \left[V_1^{\pi'_t}(x_1) - V_1^{\pi_t}(x_1) \right] \geq |\mathcal{K}_{i,T}| \sqrt{e^{i-1}} \inf_{f \in \mathcal{F}'_i} \alpha(f, \mathcal{F}, \Pi_\mathcal{F}, \mathsf{expl}, \mathcal{M})$$

Combining both bounds and rearranging yields

$$|\mathcal{K}_{i,T}| \leq O\left(\sqrt{\frac{H^2 d(\mathcal{F}'_i)}{\inf_{f \in \mathcal{F}'_i} \alpha(f, \mathcal{F}, \Pi_\mathcal{F}, \mathsf{expl}, \mathcal{M})^2} \ln \frac{N'_\mathcal{F}(1/T) \ln(T)}{\delta} |\mathcal{K}_{i,T}|} + \frac{H d(\mathcal{F}'_i)}{\inf_{f \in \mathcal{F}'_i} \alpha(f, \mathcal{F}, \Pi_\mathcal{F}, \mathsf{expl}, \mathcal{M})} \right).$$

We apply the AM-GM inequality and rearrange terms to arrive at

$$|\mathcal{K}_{i,T}| \le O\left(\frac{H^2 d(\mathcal{F}_i')}{\inf_{f\in\mathcal{F}_i'}\alpha(f,\mathcal{F},\Pi_\mathcal{F},\mathsf{expl},\mathcal{M})^2}\ln\frac{N'_\mathcal{F}(1/T)\ln(T)}{\delta}\right).$$

Taking a union bound over $i \in [i_{\max}]$ and summing the previous bound gives

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{1}\left\{\widehat{f}_t\in\mathcal{F}'\right\} &= \sum_{i=1}^{i_{\max}}|\mathcal{K}_{i,T}| \le O\left(\sum_{i=1}^{i_{\max}}\left(\frac{d(\mathcal{F}_i')}{\inf_{f\in\mathcal{F}_i'}\alpha(f,\mathcal{F},\Pi_\mathcal{F},\mathsf{expl},\mathcal{M})^2}\right)H^2\ln\frac{N'_\mathcal{F}(1/T)\ln(T)}{\delta}\right)\\ &\le O\left(\frac{H^2 d(\mathcal{F}')}{\inf_{f\in\mathcal{F}'}\alpha(f,\mathcal{F},\Pi_\mathcal{F},\mathsf{expl},\mathcal{M})^2}\ln(\sup_{f\in\mathcal{F}'}c(f,\mathcal{F},\Pi_\mathcal{F},\mathsf{expl},\mathcal{M}))\ln\frac{N'_\mathcal{F}(1/T)\ln(T)}{\delta}\right).\end{aligned}$$

Finally, since $N'_\mathcal{F}(1/T) = \sum_{h=1}^{H} N_{\mathcal{F}_h}(1/T)N_{\mathcal{F}_{h+1}}(1/T) \le \left(\sum_{h=1}^{H}N_{\mathcal{F}_h}(1/T)\right)^2 = \bar{N}_\mathcal{F}(1/T)^2$, the statement follows. □

Lemma 3. *Let $f = \{f_h\}_{h\in[H]}$ with $f_h\colon \mathcal{X}\times\mathcal{A}\to\mathbb{R}$ and π^f be the greedy policy of f. Then for any policy $\pi'\in\Pi$,*

$$V_1^{\pi'}(x_1) - V_1^{\pi^f}(x_1) \le \sum_{h=1}^{H}\mathbb{E}_{\pi^f}[(\mathcal{E}_h f)(x_h,a_h)] - \sum_{h=1}^{H}\mathbb{E}_{\pi'}[(\mathcal{E}_h f)(x_h,a_h)].$$

Proof. For any function $g\colon\mathcal{X}\to\mathbb{R}$, we define $(\mathcal{B}_h g)(x,a) = \mathbb{E}\left[r_h + g(x_{h+1}) \mid x_h = x, a_h = a\right]$. First, we write the difference in value functions at any state $x\in\mathcal{X}$ and time $h\in[H]$ as

$$\begin{aligned}V_h^{\pi'}(x) - V_h^{\pi}(x) &= \mathbb{E}[Q_h^{\pi'}(x,a)\mid a\sim\pi_h'(x)] - Q_h^{\pi}(x,\pi_h(x))\\ &= \mathbb{E}[Q_h^{\pi'}(x,a) - f_h(x,a)\mid a\sim\pi_h'(x)] + f_h(x,\pi_h(x)) - Q_h^{\pi}(x,\pi_h(x))\\ &\quad + \underbrace{\mathbb{E}[f_h(x,a)\mid a\sim\pi_h'(x)] - f_h(x,\pi_h(x))}_{\le 0}\end{aligned} \tag{11}$$

where the last term is non-positive because π is the greedy policy of f. Let $g_h(x) = \max_{a\in\mathcal{A}} f_h(x,a)$ for all $x\in\mathcal{X}$ and write the difference $Q_h^{\pi'} - f_h$ as

$$\begin{aligned}Q_h^{\pi'} - f_h &= \mathcal{B}_h V_{h+1}^{\pi'} - f_h + \mathcal{B}_h g_{h+1} - \mathcal{B}_h g_{h+1} = \mathcal{B}_h(V_{h+1}^{\pi'} - g_{h+1}) - f_h + \mathcal{T}_h f_{h+1}\\ &= \mathcal{B}_h(V_{h+1}^{\pi'} - g_{h+1}) - (\mathcal{E}_h f)(x_h,a_h)\end{aligned}$$

where we used the linearity of $\mathcal{B}$ and the fact that $\mathcal{B}_h g_{h+1} = \mathcal{T}_h f_{h+1}$. For any $x\in\mathcal{X}$ we can further bound

$$V_{h+1}^{\pi'}(x) - g_{h+1}(x) \le \mathbb{E}[Q_{h+1}^{\pi'}(x,a) - f_{h+1}(x,a)\mid a\sim\pi_h'(x)]$$

and combining this with the previous identity, we have

$$\begin{aligned}\mathbb{E}_{\pi'}[Q_h^{\pi'}(x_h,a_h) - f_h(x_h,a_h)\mid x_h = x] &\le \mathbb{E}_{\pi'}[V_{h+1}^{\pi'}(x_{h+1}) - g_{h+1}(x_{h+1}) - (\mathcal{E}_h f)(x_h,a_h)\mid x_h = x]\\ &\le \mathbb{E}_{\pi'}[(Q_{h+1}^{\pi'}(x_{h+1},a_{h+1}) - f_{h+1}(x_{h+1},a_{h+1})) - (\mathcal{E}_h f)(x_h,a_h)\mid x_h = x].\end{aligned} \tag{12}$$

Similarly, we can write

$$\begin{aligned}\mathbb{E}_\pi[(f_h - Q_h^\pi)(x_h,a_h)\mid x_h = x] &= \mathbb{E}_\pi[(f_h - \mathcal{T}_h f_{h+1} + \mathcal{B}_h g_{h+1} - \mathcal{B}_h V_{h+1}^\pi)(x_h,a_h)\mid x_h = x]\\ &= \mathbb{E}_\pi[(\mathcal{E}_h f)(x_h,a_h) + g_{h+1}(x_{h+1}) - V_{h+1}^\pi(x_{h+1})\mid x_h = x]\\ &= \mathbb{E}_\pi[(\mathcal{E}_h f)(x_h,a_h) + (f_{h+1}(x_{h+1},a_{h+1}) - Q_{h+1}^\pi(x_{h+1},a_{h+1}))\mid x_h = x]\,.\end{aligned} \tag{13}$$

Applying Equation 12 and Equation 13 recursively to Equation 11, we arrive at the desired statement

$$V_1^{\pi'}(x_1) - V_1^{\pi}(x_1) \le \mathbb{E}_{\pi'}[Q_1^{\pi'}(x_1,a_1) - f_1(x_1,a_1)] + \mathbb{E}_\pi[(f_1 - Q_1^\pi)(x_1,a_1)]$$

$$\leq \sum_{h=1}^{H} \mathbb{E}_{\pi}[(\mathcal{E}_h f)(x_h, a_h)] - \sum_{h=1}^{H} \mathbb{E}_{\pi'}[(\mathcal{E}_h f)(x_h, a_h)].$$

□

Lemma 4. *Consider Algorithm 1 with a function class $\mathcal{F}$ that satisfies Assumption 1. Let $\rho \in \mathbb{R}^+$ and $\delta \in (0,1)$. Then with probability at least $1-\delta$ for all $h \in [H]$ and $t \in \mathbb{N}$*

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

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 proof closely follows the proof of Lemma 39 by Jin et al. (2021). We first 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

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

and 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 $Y_{t,h}(f)$ can be written as

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

The variance is bounded as

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

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

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

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

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

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 $h \in [H]$

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

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 $h \in [H]$,

$$\sum_{i=1}^{t} \mathbb{E}[Y_{i,h}(f) \mid \mathfrak{F}_i] \leq 2 \sum_{i=1}^{t} Y_{i,h}(f) + 3\rho(t-1) + 176 \ln \frac{6 N_{\mathcal{F}}(\rho) \ln(2t)}{\delta} \,.$$

This holds in particular for $f = \widehat{f}_t = \{\widehat{f}_{t,h}, \widehat{f}_{t,h+1}\}$ for all $t \in \mathbb{N}$. Finally, we have

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

where the final inequality follows from completeness in Assumption 1. Therefore, we have with probability at least $1 - \delta$ for all $t \in \mathbb{N}$

$$\sum_{i=1}^{t-1} \mathbb{E}_{\mathsf{expl}(\widehat{f}_i)}[(\widehat{f}_{t,h} - \mathcal{T}_h \widehat{f}_{t,h+1})(x_h, a_h)^2] \leq 3\rho(t-1) + 176 \ln \frac{6HN'_{\mathcal{F}}(\rho)\ln(2t)}{\delta} \,.$$

□

Theorem 3 (Regret Bound of ε-Greedy)**.** *Let $T \in \mathbb{N}$ and suppose we run Algorithm 1 with ε-greedy exploration and a function class $\mathcal{F}$ that satisfies Assumption 1. Further, let there be a $h \in [1, H]$ such that for any $\lambda \in [0, 1]$, $\alpha(\mathcal{F}'_\lambda, \mathcal{F}) \geq \Omega((\varepsilon/A)^{h/2}\lambda)$ where $\mathcal{F}'_\lambda \subset \mathcal{F}$ denotes the set of all the value functions that are at least λ suboptimal. Then, with probability at least $1 - \delta$, we have,*

$$\mathrm{Reg}(T) \leq \varepsilon HT + \widetilde{O}\Big(\sqrt{\frac{hA^h dH^3 T}{\varepsilon^h} \ln \frac{\bar{N}_{\mathcal{F}}(T^{-1})}{\delta}}\Big),$$

where $d = \dim_{\mathrm{BE}}(\mathcal{F}, \Pi_{\mathcal{F}}, 1/\sqrt{T})$ denotes the Bellman-Eluder dimension of $\mathcal{F}$. Furthermore, setting the exploration parameter $\varepsilon = \widetilde{\Theta}\big(\big(\frac{hHA^h d}{T}\big)^{\frac{1}{2+h}}\big)$, we get that

$$\mathrm{Reg}(T) \leq \widetilde{O}\Big(H^{\frac{h+3}{h+2}} T^{\frac{h+1}{h+2}} \big(hA^h d \ln \frac{\bar{N}_{\mathcal{F}}(T^{-1})}{\delta}\big)^{\frac{1}{h+2}}\Big).$$

Proof. We decompose the regret as

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^{T} \Big(V^\star(x_{t,1}) - V^{\mathsf{expl}(\widehat{f}_t)}(x_{t,1})\Big) \\
&= \sum_{t=1}^{T} (V^\star(x_{t,1}) - V^{\pi_t}(x_{t,1})) + \sum_{t=1}^{T} \Big(V^{\pi_t}(x_{t,1}) - V^{\mathsf{expl}(\widehat{f}_t)}(x_{t,1})\Big) \,,
\end{aligned}$$

and bound both terms individually. The excess regret due to exploration in the second term is bounded as

$$\sum_{t=1}^{T} \Big(V^{\pi_t}(x_{t,1}) - V^{\mathsf{expl}(\widehat{f}_t)}(x_{t,1})\Big) \leq TH\varepsilon.$$

Second, let $\mathcal{F}_i = \{f \in \mathcal{F} \colon V^\star(x_{t,1}) - V^{\pi^f}(x_{t,1}) \in [(1/2)^i, (1/2)^{i-1}]\}$ the value functions that incur regret $[(1/2)^i, (1/2)^{i-1}]$ per episode. Applying Theorem 1 above, we get

$$\sum_{t=1}^{T} (V^\star(x_{t,1}) - V^{\pi_t}(x_{t,1})) \leq T2^{-m} + \sum_{i=1}^{m} 2^{-i} O\left(\frac{H^2 d}{\alpha(\mathcal{F}_i, \mathcal{F})^2} \ln(c(\mathcal{F}_i, \mathcal{F})) \ln \frac{m\bar{N}_{\mathcal{F}_i}(1/T)\ln(T)}{\delta}\right) .$$

for any $m \in \mathbb{N}$. For ε-greedy, we can bound $\alpha(\mathcal{F}_i, \mathcal{F}) \leq (A/\varepsilon)^H$ and thus

$$\mathrm{Reg}(T) \leq TH\varepsilon + T2^{-m} + H^3 d \ln(A/\varepsilon) \ln \frac{m\bar{N}_{1/T}(\mathcal{F}) \ln(T)}{\delta} \sum_{i=1}^{m} 2^{-i} O\left(\frac{1}{\alpha(\mathcal{F}_i, \mathcal{F})^2}\right).$$

Assume $\alpha(\mathcal{F}_i, \mathcal{F}) \geq (\varepsilon/A)^{h/2} 2^{-i-1}$, which which always holds for $h = H$. Then

$$\sum_{i=1}^{m} 2^{-i} \frac{1}{\alpha(\mathcal{F}_i, \mathcal{F})^2} \leq \frac{A^h}{\varepsilon^h} \sum_{i=1}^{m} \frac{2^{-i}}{2^{-2i-2}} = \frac{A^h}{\varepsilon^h} \sum_{i=1}^{m} 2^{i+2} \leq \frac{4A^h m}{\varepsilon^h} 2^m$$

and setting m such that $2^m = \sqrt{\frac{T}{mdH^3 \ln(A/\varepsilon) \ln \frac{\bar{N}_{\mathcal{F}}(1/T) \ln(T)}{\delta}}} (\varepsilon/A)^{h/2}$ gives

$$\mathrm{Reg}(T) \leq TH\varepsilon + \tilde{O}\left(\sqrt{\frac{hdH^3 A^h T}{\varepsilon^h} \ln(A) \ln(T) \ln \frac{\bar{N}_{\mathcal{F}}(1/T)}{\delta}}\right).$$

Finally, denote by $L = \tilde{O}(\ln(A) \ln(T) \ln \frac{\bar{N}_{\mathcal{F}}(1/T)}{\delta})$ the log-terms above and assume the exploration parameter is chosen as

$$\varepsilon = \Theta\left(\left(\frac{hHdA^h L}{T}\right)^{\frac{1}{2+h}}\right).$$

Then the regret bound evaluates to

$$\mathrm{Reg}(T) \leq \tilde{O}\left(HT^{\frac{h+1}{h+2}} \left(HhA^h d \ln(A) \ln \frac{\bar{N}_{\mathcal{F}}(1/T)}{\delta}\right)^{\frac{1}{h+2}}\right).$$

□

E. Supporting Technical Results

We recall the following standard definitions.

Definition 4 (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 5 ((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 6 (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*

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

Lemma 5 (Lemma 41, Jin et al. (2021))**.** *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 $\{\phi_i\}_{i \in [K]} \subseteq \Phi$ and $\{\mu_i\}_{i \in [K]} \subseteq \Pi$ 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 $\omega > 0$*

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

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

$$\mathrm{Var}_{\ell}(X_{\ell}) = \mathbf{Var}(X_{\ell} | X_1, \cdots, X_{\ell-1})$$

Let $V_t = \sum_{\ell=1}^{t} \mathrm{Var}_\ell(X_\ell)$ *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 $A_t = \sqrt{2\ln\ln\left(2\left(\max\left(\frac{V_t}{b^2},1\right)\right)\right) + \ln\frac{6}{\delta'}}$.