How Does Goal Relabeling Improve Sample Efficiency?

Sirui Zheng[1] **Chenjia Bai**[2] **Zhuoran Yang**[3] **Zhaoran Wang**[1]

Abstract

Hindsight experience replay and goal relabeling are successful in reinforcement learning (RL) since they enable agents to learn from failures. Despite their successes, we lack a theoretical understanding, such as (i) why hindsight experience replay improves sample efficiency and (ii) how to design a relabeling method that achieves sample efficiency. To this end, we construct an example to show the information-theoretical improvement in sample efficiency achieved by goal relabeling. Our example reveals that goal relabeling can enhance sample efficiency and exploit the rich information in observations through better hypothesis elimination. Based on these insights, we develop an RL algorithm called `GOALIVE`. To analyze the sample complexity of `GOALIVE`, we introduce a complexity measure, the *goal-conditioned Bellman-Eluder (GOAL-BE) dimension*, which characterizes the sample complexity of goal-conditioned RL problems. Compared to the Bellman-Eluder dimension, the goal-conditioned version offers an exponential improvement in the best case. To the best of our knowledge, our work provides the first characterization of the theoretical improvement in sample efficiency achieved by goal relabeling.

1. Introduction

Numerous RL problems encountered in real-world scenarios involve sparse rewards and vast state spaces (Silver et al., 2017; Savinov et al., 2018; Riedmiller et al., 2018), making them challenging to solve due to the lack of meaningful feedback. A widely used technique to address these challenges is hindsight experience replay and goal relabeling (Andrychowicz et al., 2017). Such algorithms typically employ a goal-dependent reward. In each iteration, the agent updates the value function concerning a relabeled goal rather than the target goal. Numerous empirical studies have demonstrated the effectiveness of hindsight experience replay and goal relabeling in various scenarios (Andrychowicz et al., 2017; Pong et al., 2018; Fang et al., 2019; Colas et al., 2019; Li et al., 2020; Pitis et al., 2020; Zhang et al., 2020).

Despite its empirical success, the theoretical understanding of hindsight experience replay and goal relabeling is still elusive. Empirical studies have employed various methods for these techniques; however, these methods lack theoretical guarantees (Andrychowicz et al., 2017; Pong et al., 2018; Nair et al., 2018; Pitis et al., 2020). Additionally, the design of provably efficient algorithms incorporating goal relabeling remains unclear, as existing research on efficient exploration (Jiang et al., 2017; Cai et al., 2019; Jin et al., 2020b) does not incorporate goal relabeling. While prior research has produced sample-efficient algorithms for multitask RL and reward-free RL (Jin et al., 2020a; Lu et al., 2022; Cheng et al., 2022), which share similarities with goal-conditioned RL, these results do not fully capture the benefits of goal relabeling.

Consequently, we aim to enhance the theoretical understanding of goal relabeling by addressing the following questions: First, does hindsight experience replay and goal relabeling provably improve sample efficiency? Second, if the answer is positive, how does goal relabeling contribute to this improvement? Third, can we design a principled goal-conditioned algorithm that provably achieves better sample efficiency for problems with sparse rewards and large state spaces? Our contributions address these questions and can be summarized as follows.

- For the first problem, we provide an affirmative answer by constructive proof. We show in §3 that there exist MDPs for which model-free algorithms using goal-conditioned value functions achieve polynomial sample complexity. In contrast, all model-free methods, given a value function class without multiple goals, incur sample complexity exponential in the horizon. In other word, we show that **using**

[1] Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208, USA [2]Shanghai Artificial Intelligence Laboratory [3] Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA.. Correspondence to: Sirui Zheng <siruizheng2025@u.northwestern.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

goal-conditioned value functions exponentially improves the sample complexity in the best case. To the best of our knowledge, this is the first result illustrating the information-theoretical improvement achieved by using goal-conditioned value functions in model-free algorithms.

- For the second question, we show in §3.2 that we can better utilize the information in the observation by using goal relabeling in the algorithm. By incorporating goal-conditioned value functions into model-free algorithms, we can evaluate multiples Bellman errors with the rewards concerning multiple goals. Such a procedure allows us to eliminate more hypotheses using the value function of the **same state-action pair** with respect to **different goals**, which is more sample efficient.
- For the third question, we propose a novel algorithm called *GOAl-conditioned optimism Led Iterative Value-function Elimination* (`GOALIVE`) in §4 using the above intuition. By using goal relabeling, the agent can obtain rich feedback from the observation even when the reward with respect to the target goal is sparse. Additionally, we employed general function approximation to handle large state space. We show that the relabeling method employed in existing algorithms (Andrychowicz et al., 2017; Li et al., 2020; Pitis et al., 2020) can be regarded as a modification of our algorithm.
- To analyze the algorithm we proposed, we introduce a new complexity measure called *GOAL-conditioned Bellman-Eluder (GOAL-BE) dimension* in §5.1, which characterizes the complexity of goal-conditioned RL problems. We **demonstrate that the sample complexity of `GOALIVE` can be upper-bounded by the GOAL-BE dimension.** We prove that the proposed complexity measure is not larger than the original Bellman-Eluder dimension, which shows that our definition is more general. We also show in §5 that **GOAL-BE dimension can be exponentially smaller than the original one in the best case.**

To the best of our knowledge, this is the first study to develop a provably efficient goal-conditioned algorithm and explain the improvement achieved by goal relabeling.

1.1. Related Work

Our work is closely related to the line of research on provably efficient exploration in the function approximation setting (Jiang et al., 2017; Jin et al., 2020b; Cai et al., 2019; Du et al., 2021; Uehara et al., 2021; Zhang et al., 2022; Liu et al., 2024). This line of research typically considers exploration for a single task and is not readily applicable to scenarios involving multiple goals (Jiang et al., 2017; Jin et al., 2020b; Cai et al., 2019). Sun et al. (2019) observed that model-free algorithms can be inefficient since they fail to exploit the information in observations. Our work shows that using goal-conditioned value functions improves the sample efficiency of model-free algorithms.

The study of efficient exploration in multitask settings (Lu et al., 2022; Cheng et al., 2022) is more closely related to our work, as it considers multiple goals. However, these studies cannot explain the benefits of using multitask methods for solving a single task. The research on reward-free exploration (Jin et al., 2020a; Wang et al., 2020) also aligns with our work, as both aim to develop exploration strategies that do not rely on the original reward. Nonetheless, these studies employ model-based algorithms, whereas model-free algorithms are more prevalent in empirical results.

Our work is inspired by empirical findings in hindsight experience replay and goal relabeling (Andrychowicz et al., 2017; Pong et al., 2018; Fang et al., 2019; Colas et al., 2019; Li et al., 2020; Pitis et al., 2020; Zhang et al., 2020). We provide a theoretical explanation for the improvement in sample efficiency achieved through goal relabeling.

Notations. We define $[n] = \{1, \ldots, n\}$ when n is an integer. For a set $\mathcal{F}$, we denote by $|\mathcal{F}|$ the cardinality of $\mathcal{F}$.

2. Preliminary

In this paper, we focus on goal-conditioned reinforcement learning, where the goal serves as a parameter of the reward function. For example, the reward function can be designed as an indicator function that takes the goal as input and assigns non-zero values only when the final state matches the specified goal. More specifically, we consider an episodic MDP $\mathcal{V}^* = (\mathcal{S}, \mathcal{A}, \mathcal{G}, H, \mathcal{P}^*, r^*)$ with a state space $\mathcal{S} \in \mathbb{R}^d$, an action space $\mathcal{A}$, a goal space $\mathcal{G}$, a horizon H, transition kernels $\mathcal{P}^* = \{\mathcal{P}_h^*\}_{h=1}^H$, and known reward functions $r^* = \{r_h^*\}_{h=1}^H$. We assume that the reward functions are bounded and deterministic, that is, $\|r_h^*\|_\infty \in [0, 1]$ for all $h \in [H]$. The agent iteratively interacts with the environment as follows. At the beginning of each episode, the agent determines a policy $\pi = \{\pi_h\}_{h=1}^H$, where $\pi_h : \mathcal{S} \times \mathcal{G} \to \Delta(\mathcal{A})$ for any $h \in [H]$. Without loss of generality, we assume that the initial state is fixed to $s_{\text{init}} \in \mathcal{S}$ across all episodes. At the h-th step, the agent receives a state s_h and takes an action a_h following $a_h \sim \pi_h(\cdot \mid s_h, g^*)$, where g^* is the target goal. Subsequently, the agent receives a reward $r_h^*(s_h, a_h, g^*)$ and the next state following $s_{h+1} \sim \mathcal{P}_{h+1}^*(\cdot \mid s_h, a_h)$. The episode ends after the agent receives the last state s_{H+1}. For a given policy $\pi = \{\pi_h\}_{h=1}^H$, we define the value function V_h^π and the Q-function Q_h^π for any $h \in [H]$ and $g \in \mathcal{G}$ as

$$V_h^\pi(s,g) := \mathbb{E}_{\pi,\mathcal{P}^*}\Big[\sum_{i=h}^{H} r_i^*(s_i,a_i,g)\,\Big|\, s_h = s\Big], \tag{1}$$

$$Q_h^\pi(s,a,g) := \mathbb{E}_{\pi,\mathcal{P}^*}\Big[\sum_{i=h}^{H} r_i^*(s_i,a_i,g)\,\Big|\, s_h = s, a_h = a\Big].$$

Here the expectation $\mathbb{E}_{\pi,\mathcal{P}^*}[\cdot]$ in (1) is taken with respect to $s_{i+1} \sim \mathcal{P}_i^*(\cdot \,|\, s_i, a_i)$ and $a_i \sim \pi_i(\cdot \,|\, s_i, g)$ for $i \in \{h, h+1, \ldots, H\}$. For convenience, we define $V_{H+1}^\pi(s,g) = 0$ for any state $s \in \mathcal{S}$ and policy π. We then define the goal-conditioned Bellman operator $\mathcal{T}_h$ by

$$\mathcal{T}_h f(s,a,g) := r_h^*(s,a,g) + \mathbb{E}_{s' \sim \mathcal{P}_h}\big[\max_{a'} f(s',a',g)\big],$$

we then have $Q_h^*(s,a,g) = \mathcal{T}_h Q_{h+1}^*(s,a,g)$. For simplicity, we define the expected total reward $J(\pi,g)$ as $J(\pi,g) = V_1^\pi(s_{\text{init}},g)$. The objective of goal-conditioned RL is to find a policy π^* that maximizes the expected total reward with regard to the target goal g^*. Specifically, for the episodic MDP $\mathcal{V}^* = (\mathcal{S},\mathcal{A},\mathcal{G},H,\mathcal{P}^*,r^*)$, we define the optimal Q-function Q_h^* and the optimal value function V_h^* as $Q_h^*(s,a,g) = \max_\pi Q_h^\pi(s,a,g)$ and $V_h^*(s,g) = \max_\pi V_h^\pi(s,g)$ for any $(s,a,g) \in \mathcal{S} \times \mathcal{A} \times \mathcal{G}$. Correspondingly, we define the optimal goal-conditioned policy π^* by the policy that satisfies $Q_h^*(s,a,g) = Q_h^{\pi^*}(s,a,g)$ for any $(s,a,g) \in \mathcal{S} \times \mathcal{A} \times \mathcal{G}$.

In this paper, we consider RL with value function approximation. Formally, the learner is given a hypothesis class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$, where $\mathcal{F}_h \subset (\mathcal{S} \times \mathcal{A} \times \mathcal{G} \mapsto [0, H-h+1])$ is the set of approximators of Q_h^* defined in (1). We also define $f_{H+1} \equiv 0$ for convenience. For a hypothesis $f = \{f_h\}_{h=1}^H$, we define the goal-conditioned average Bellman error by

$$\begin{aligned}\mathcal{E}(f,\pi,h,g) &: \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (2)\\ &= \mathbb{E}_\pi\big[f_h(s,a,g) - r_h^*(s,a,g) - \max_{a' \in \mathcal{A}} f_{h+1}(s',a',g)\big]\\ \mathcal{E}(f,\pi,h) &:= \max_{g \in \mathcal{G}} |\mathcal{E}(f,\pi,h,g)|\,.\end{aligned}$$

By Bellman equation, we have $\mathcal{E}(Q_h^*,\pi,h) = 0$ for all policy π and $h \in [H]$. In this paper, an important tool we used in the analysis is the Bellman-Eluder dimension, which characterizes the complexity of an RL problem when implementing model-free algorithms. In the following, we introduce the definition of the Bellman-Eluder dimension. We begin by explaining the concept of ϵ-independence between distributions.

Definition 2.1 (ϵ-independence between distributions). *Let $\mathcal{F}$ be a function class defined on $\mathcal{X}$, 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 $\mathcal{F}$ if there exists $f \in \mathcal{F}$ such that*

$$\sqrt{\sum_{i=1}^{n} \Big\{\mathbb{E}_{x \sim \mu_i}\big[f(x)\big]\Big\}^2} \le \epsilon,$$

but $|\mathbb{E}_{x \sim \nu}[f(x)]| > \epsilon$.

Intuitively, ν is independent of $\{\mu_1, \ldots, \mu_n\}$ means if that there exist a hypothesis, so that the value of the loss f is small at all distribution $\{\mu_i\}_{i=1}^n$, but the value of the loss f is rather big at ν. Jin et al. (2021a) propose a dimension based on the above notion of independence.

Definition 2.2 (Distributional Eluder dimension). *Let $\mathcal{F}$ be a function class defined on $\mathcal{X} \times \mathcal{G}$, and Π be a family of probability measures over $\mathcal{X}$. The distributional Eluder dimension $\dim_{\text{DE}}(\mathcal{F}, \Pi, \epsilon)$ is the length of the longest sequence $\{\rho_1, \ldots, \rho_n\} \subset \Pi$ such that there exists $\epsilon' \ge \epsilon$ where ρ_i is ϵ'-independent of $\{\rho_1, \ldots, \rho_{i-1}\}$ for all $i \in [n]$.*

Based on Definition 2.1 and 2.2, Jin et al. (2021a) proposed BE dimension, which characterizes the complexity of a RL problem using model-free algorithm with general function approximation in single-task settings.

Definition 2.3 (Bellman Eluder (BE) dimension). *Let $\overline{\mathcal{F}} = \overline{\mathcal{F}}_1 \times \cdots \times \overline{\mathcal{F}}_H$, where $\overline{\mathcal{F}}_h \subset (\mathcal{S} \times \mathcal{A} \mapsto [0, H-h+1])$. Let $(I - \mathcal{T}_h)\overline{\mathcal{F}} := \{f_h - \mathcal{T}_h f_{h+1} : f \in \overline{\mathcal{F}}\}$ be the set of Bellman residuals induced by $\overline{\mathcal{F}}$ at step h, and $\mathcal{D} = \{\mathcal{D}_h\}_{h=1}^H$ be a collection of H probability measure families over $\mathcal{S} \times \mathcal{A}$. The ϵ-Bellman-Eluder of $\overline{\mathcal{F}}$ with respect to $\mathcal{D}$ is defined as*

$$\dim_{\text{BE}}(\overline{\mathcal{F}}, \mathcal{D}, \epsilon) := \max_{h \in [H]} \dim_{\text{DE}}\big((I - \mathcal{T}_h)\overline{\mathcal{F}}, \mathcal{D}_h, \epsilon\big).$$

We remark that their definition is not directly applicable to goal-conditioned RL. The hypothesis class $\overline{\mathcal{F}}$ in Definition 2.3 does not take the goal g as input. Thus, it remains unclear how to incorporate the goal space within the given definition.

3. Mitigating Information Loss with Goal-conditioned Value Functions

In this section, we present an example demonstrating that goal-conditioned model-free algorithms can exponentially improve sample efficiency compared to model-free algorithms without multiple goals. To the best of our knowledge, this is the first result illustrating the information-theoretic improvement achieved by using a goal-conditioned value function. We introduce the example in §3.1 and provide the analysis in §3.2.

3.1. Example

We consider an episodic MDP with horizon $H > 3$ and set $d = H - 2$. The state space is $\mathcal{S} = \{1, 2\}^d$, and the action space is $\mathcal{A} = \{1, 2\}$. Our transition class contains 2^d transitions, each of which is uniquely indexed by a path of length d, $\mathbf{p} = \{p_1, p_2, \ldots, p_d\}$, with $p_i \in \{1, 2\}$. All the transitions are the same in the first $H - 1$ steps. For $h \leq H - 2$, when we use action a in the state $s_h = (s_{h,1}, \ldots, s_{h,h}, \ldots, s_{h,d})$ in the h-th step, we will transit to $s_{h+1} = (s_{h,1}, \ldots, a, \ldots, s_{h,d})$.

The transition in the $H - 1$-th step does not depend on the action a_{H-1}. In the transition indexed by $\mathbf{p} = \{p_1, p_2, \ldots, p_d\}$, we will transit to

$$s_H = (\mathbb{1}\{s_{H-1,1} = p_1\}, \ldots, \mathbb{1}\{s_{H-1,d} = p_d\})$$

when we select action a in $s_{H-1} = (s_{H-1,1}, \ldots, s_{H-1,d})$. The goal space is $\mathcal{G} = \{0, 1\}^d$, and goal-conditioned reward is

$$r_h^*(s, a, g) = 0 \ \text{ for } h \in [H-1], \ r_H^*(s, a, g) = \mathbb{1}\{s = g\}.$$

We fix the initial state as $s_1 = [1]^d$. We want to obtain the optimal policy with respect to the reward $\{r_h^*(s, a, g^*)\}_{h=1}^H$, where $g^* = [1]^d$. A graphical illustration of the setting is provided in Figure 1.

3.2. Sample Efficiency of Model-free Algorithms

In this subsection, we use the provided example to show that goal-conditioned model-free algorithms can be exponentially more sample-efficient compared to algorithms that do not incorporate goals. We first introduce the definition of model-free algorithms. We then introduce the value function class we use. Finally, we offer an analysis to show that goal-conditioned algorithms exponentially improve the sample efficiency in the given example.

Definition 3.1 (Goal-conditioned Model-free Algorithm)**.** *Given a function class $\mathcal{F} : (\mathcal{S} \times \mathcal{A} \times \mathcal{G}) \to \mathbb{R}$, define the original $\mathcal{F}$-profile $\Phi_{\mathcal{F},g^*} : \mathcal{S} \to \mathbb{R}^{|\mathcal{F}| \times |\mathcal{A}|}$ by $\Phi_{\mathcal{F},g^*}(s) := [f(s, a, g^*)]_{f \in \mathcal{F}, a \in \mathcal{A}}$. An algorithm is model-free using $\mathcal{F}$ without multiple goals if it accesses s exclusively through $\Phi_{\mathcal{F},g^*}(s)$ for all $s \in \mathcal{S}$ during its entire execution. We also define the goal-conditioned $\mathcal{F}$-profile $\Phi_{\mathcal{F},\mathcal{G}} : \mathcal{S} \to \mathbb{R}^{|\mathcal{F}| \times |\mathcal{A}| \times |\mathcal{G}|}$ by $\Phi_{\mathcal{F},\mathcal{G}}(s) := [f(s, a, g)]_{f \in \mathcal{F}, a \in \mathcal{A}, g \in \mathcal{G}}$. An algorithm is goal-conditioned model-free using $\mathcal{F}$ and $\mathcal{G}$ if it accesses s exclusively through $\Phi_{\mathcal{F},\mathcal{G}}(s)$ for all $s \in \mathcal{S}$ during its entire execution.*

Definition 3.1 is a modification of Definition 1 in Sun et al. (2019). In the definition, $\mathcal{G}$ is the space of goal, and $f(s, a, g)$ is the action-value of (s, a) when we are interested in the reward indexed by the goal g. We modified Definition 1 in Sun et al. (2019) to incorporate goal.

Value Function Class. We consider solving this MDP with model-free algorithms. In the following, we introduce the value function class we employ for solving this MDP.

For a hypothesis $\mathbf{p} = (p_1, \ldots, p_d) \in \{1, 2\}^d$, we define $Q_h^{\mathbf{p}}(s, a, g)$ as the optimal action value function of (s, a) with respect to the transition $\mathbf{p}$ and the reward indexed by g. It is obvious that $Q_H^{\mathbf{p}}(s, a, g) = \mathbb{1}\{s = g\}$. For $h < H$, $Q_h^{\mathbf{p}}(s, a, g)$ is defined as follows. For $s_h = (s_{h,1}, \ldots, s_{h,d})$ and $\mathbf{p} = (p_1, \ldots, p_d)$, we define

$$f_{h,\bar{d}}(s_h, a, \mathbf{p}) = \begin{cases} \mathbb{1}\{s_{h,\bar{d}} = p_{\bar{d}}\} & \text{when } \bar{d} < h, \\ \mathbb{1}\{a = p_{\bar{d}}\} & \text{when } \bar{d} = h, \end{cases}$$

for $\bar{d} \leq h$. By the definition of the setting, $f_{h,\bar{d}}(s_h, a, \mathbf{p})$ is the $\bar{d}$-th component of s_H in the transition indexed by $\mathbf{p}$ when we use a in the state s_H. Therefore, when we denote by $Q_h^{\mathbf{p}}$ the optimal value function in the MDP with the transition indexed by $\mathbf{p}$, we have

$$Q_h^{\mathbf{p}}(s_h, a, g) = \prod_{\bar{d}=1}^{h} \mathbb{1}\{f_{h,\bar{d}}(s_h, a, \mathbf{p}) = g_{\bar{d}}\},$$

where $g = (g_1, \ldots, g_d)$. Intuitively, $Q_h^{\mathbf{p}}$ represents the feasibility of reaching the goal state g from the state-action pair (s_h, a). We then define $\mathcal{F} = \mathcal{F}_1 \times \ldots \times \mathcal{F}_H$ and $\mathcal{F}_{g^*} = \mathcal{F}_{g^*,1} \times \ldots \times \mathcal{F}_{g^*,H}$, where

$$\begin{aligned} \mathcal{F}_h &:= \left\{Q_h^{\mathbf{p}}(\cdot, \cdot, \cdot) \mid \mathbf{p} \in \{1, 2\}^d\right\}, \\ \mathcal{F}_{g^*,h} &:= \left\{Q_h^{\mathbf{p}}(\cdot, \cdot, g^*) \mid \mathbf{p} \in \{1, 2\}^d\right\}. \end{aligned} \tag{3}$$

By definition, $\mathcal{F}$ contains goal-conditioned value functions, while $\mathcal{F}_{g^*}$ only contains value functions with respect to g^*. In the example above, we have the following lemma, which show that goal-conditioned value functions exponentially improve the sample efficiency of model-free algorithms.

Lemma 3.2 (Exponential Improvement of Goal-conditioned Model-free Algorithms)**.** *Fix $\delta, \epsilon \in (0, 1]$. In the given example, any model-free algorithm without multiple goals, which uses $\mathcal{F}_{g^*}$ in (3) as the value function class, using $o(2^H)$ trajectories outputs a policy $\widehat{\pi}$ with $V^{\widehat{\pi}}(s_1, g^*) \leq V^*(s_1, g^*) - 1/2$ with probability at least $1/3$. Meanwhile, with probability $1 - \delta$, there exists a goal-conditioned model-free algorithm using $\mathcal{F}$ in (3) as the value function class (Algorithm 1) outputs $\widehat{\pi}$ satisfying $V^{\widehat{\pi}}(s_1, g^*) \geq V^*(s_1, g^*) - \epsilon$ using at most $\mathrm{poly}(H, 1/\epsilon, \log(1/\delta))$ trajectories for any transition in this family.*

Proof. See Appendix §A for a detailed proof. □

Interpretation. Our observation is that, goal-conditioned model-free algorithms are more effective at utilizing the rich information in data compared to original model-free algorithms. In goal-conditioned algorithm, we can obtain nonzero reward even when the reward with respect to the target goal is sparse, since we can use the reward with respect to different goals for hypothesis elimination. As a

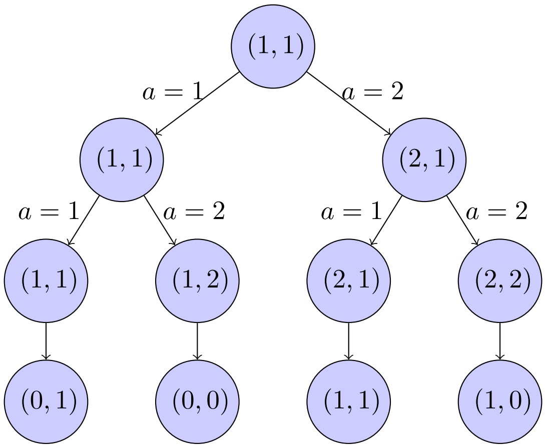

Figure 1. An example of the MDP construction in §3, with $d = 2$ and $H = 4$. All models are deterministic, and each model is uniquely indexed by a sequence of actions $\mathbf{p}$. We use $\mathbf{p}^*$ to denote the index of the true model. Here $\mathbf{p}^* = (2, 1)$. The last transition is designed such that the agent always lands in a state that contains "0" unless it follows path $\mathbf{p}^*$.

result, we can eliminate more hypotheses using Bellman errors with respect to different goals, which improves the sample efficiency. In Definition 3.1, a model-free algorithm without multiple goals accesses the state s only through the value concerning the target goal g^*. Therefore, it can only eliminate hypotheses using the Bellman error with respect to g^*, which can be zero for most of the hypotheses when the reward is sparse. By using the reward and value functions with respect to multiple goal, we obtain denser feedback, leading to more effective hypothesis elimination.

To illustrate the above idea, we revisit the example in §3.1. We consider the case where the true transition $\mathbf{p}^* = (2, 1)$ and the target goal $g^* = (1, 1)$. For the trajectory $\tau = (s_1, a_1, \cdots, s_s, a_4, s_5)$, we calculate the Bellman error and provide it in Table 1. As we show in Table 1, since the reward with respect to the target goal g^* is sparse, the Bellman error of most of the hypotheses is zero if we only consider the target goal g^*. Therefore, we can only eliminate one hypothesis using the Bellman error with respect to g^*. Meanwhile, we can obtain nonzero reward and eliminate all the wrong hypotheses using the reward and Bellman error with respect to $g = (1, 0)$. Therefore, we can achieve exponential improvement in sample efficiency using goal relabeling.

Rationale for Focusing Our Analysis on Model-Free Algorithms. In the above analysis, we focus on model-free algorithms as they are widely used in existing works that apply goal relabeling (Andrychowicz et al., 2017; Li et al., 2020; Pitis et al., 2020). Sun et al. (2019) show that model-based algorithms are more sample efficient as they leverage more supervision. However, model-based algorithms necessitate an additional, often complex, planning phase to derive an optimal policy. This phase, particularly in environments with complex dynamics, introduces substantial computational costs and a heightened risk of error accumulation. Errors in estimating dynamics during the planning phase can magnify, potentially resulting in significantly suboptimal policies.

In contrast, goal-conditioned model-free algorithms avoid the complexities of modeling the dynamics and subsequent planning, choosing instead to directly approximate the value function. This direct approach reduces the risk of error propagation inherent in planning, providing a potentially more robust solution in scenarios where accurately modeling the dynamics is challenging. Consequently, even though the goal-conditioned value function's theoretical complexity may match or surpass that of the dynamics, the practical advantages of operational efficiency, robustness, and applicability make goal-conditioned model-free methods a compelling choice in complex environments. Therefore, model-free algorithms are more popular in real-life application, making it crucial to theoretically analyze the improvements achieved by goal relabeling in model-free algorithms.

4. Algorithm: Goal-conditioned Optimism Led Iterative Value-function Elimination

In §3, we construct an example and show that using goal-conditioned value function can exponentially improve the sample efficiency in the best case. It is natural to ask, how to design an algorithm that provably achieves the improved sample efficiency? In §3.2 and Table 1, we show that goal-conditioned value functions improve the sample efficiency since it make the feedback denser and allows better hypothesis elimination. Based on the idea above, we propose an algorithm called ***GOA**l-conditioned optimism **L**ed **I**terative **V**alue function **E**limination* (`GOALIVE`), which is a modification of Algorithm `OLIVE` proposed in Jiang et al. (2017).

At a high level, our algorithm eliminate hypothesis $Q_h \in \mathcal{F}_h$ using the average Bellman error and the data we have collected. After elimination, we choose the exploration policy π_t by selecting the most optimistic hypothesis concerning the target goal g^* and the corresponding greedy

		Bellman error w.r.t the goal g				Elimination	
		$g=(0,0)$	$g=(0,1)$	$g=(1,0)$	$g=(1,1)$	All goals	Only g^*
Hypothesis of $\mathbf{p}$	$\mathbf{p}=(1,1)$	1	0	-1	0	Yes	No
	$\mathbf{p}=(1,2)$	0	1	-1	0	Yes	No
	$\mathbf{p}=(2,1)$	0	0	0	0	No	No
	$\mathbf{p}=(2,2)$	0	0	-1	1	Yes	Yes

Table 1. Comparison between using multiple goals and only using the target goal g^* for hypothesis elimination. In this table, we use the blue color to highlight the true hypothesis $\mathbf{p}^*$ and the target goal g^*. This table present the Bellman error $Q_3^{\mathbf{p}}(s_3, a_3, g) - r_3^*(s_3, a_3, g) - Q_4^{\mathbf{p}}(s_4, a_4, g)$, where $s_3 = (2,2)$ and $s_4 = (1,0)$. The table demonstrates that all incorrect hypotheses can be eliminated after goal relabeling with the given data, whereas only one hypothesis can be eliminated using the Bellman error with respect to the target goal g^*.

policy.

Algorithm 1 ***GOA**l-conditioned optimism **L**ed **I**terative **V**alue function **E**limination* (GOALIVE)

```
1:  Input: Hypothesis class $\mathcal{F}$, elimination thresholds $\zeta_{\text{act}}$ and $\zeta_{\text{elim}}$, numbers of iterations $n_{\text{act}}$ and $n_{\text{elim}}$.
2:  Initialize: $\mathcal{B}^0 \leftarrow \mathcal{F}$, $\mathcal{D}_h^t \leftarrow \emptyset$ for all $h$ and $t$.
3:  for iteration $t = 1, 2, \dots$ do
4:      Choose policy $\pi^t = \pi_{f^t}$, where $f^t = \operatorname{argmax}_{f \in \mathcal{B}^{t-1}} f(x_1, \pi_f(x_1), g^*)$.
5:      Execute $\pi^t$ for $n_{\text{act}}$ episodes and update $\mathcal{D}_h^t$ to include the new $(s_h, a_h, s_{h+1})$ tuples.
6:      Estimate $\widehat{\mathcal{E}}(f^t, \pi^t, h, g^*)$ for all $h \in [H]$, where
            $\widehat{\mathcal{E}}(f, \pi^t, h, g) = \frac{1}{|\mathcal{D}_h^t|} \sum_{(s,a,s') \in \mathcal{D}_h^t} \Delta_1[s, a, s', g, f_h, r_h, f_{h+1}]$,
        where $\Delta_1$ is defined in (4).
7:      if $\sum_{h=1}^{H} \widehat{\mathcal{E}}(f^t, \pi^t, h, g^*) \le H\zeta_{\text{act}}$ then
8:          Terminate and output $\pi^t$.
9:      end if
10:     Pick any $h \in [H]$ for which $\widehat{\mathcal{E}}(f^t, \pi^t, h, g^*) \ge \zeta_{\text{act}}$ and set $\mathcal{D}_h^t = \emptyset$.
11:     Execute $\pi^t$ for $n_{\text{elim}}$ episodes and update $\mathcal{D}_h^t$ to include the new $(s_h, a_h, s_{h+1})$ tuples.
12:     Estimate $\widehat{\mathcal{E}}(f, \pi^t, h)$ for all $f \in \mathcal{F}$, where
            $\widehat{\mathcal{E}}(f, \pi^t, h) = \frac{1}{|\mathcal{D}_h^t|} \left| \max_{g \in \mathcal{G}} \sum_{(s,a,s') \in \mathcal{D}_h^t} \Delta_1[s, a, s', g, f_h, r_h, f_{h+1}] \right|$.
13:     Update $\mathcal{B}^t = \left\{ f \in \mathcal{B}^{t-1} : \left| \widehat{\mathcal{E}}(f, \pi^t, h) \right| \le \zeta_{\text{elim}} \right\}$.
14: end for
```

The pseudocode of GOALIVE is given in Algorithm 1. In each episode, our algorithm performs three main steps:

- Optimistic planing: we compute the most optimistic hypothesis Q^k with regard to the target goal g^*, and choose π^k to be its greedy policy in Line 3. We then use π^k to interact with the environment in Line 4.
- Computing Bellman error: We evaluate the average Bellman error of f^k under π^k in Line 5. We activate the elimination phase if the Bellman error is large, and otherwise returned π^k in Line 7.
- Eliminating hypothesis with large Bellman error: pick a step $t \in [H]$ where the estimated Bellman error exceeds the activation threshold ζ_{act}; eliminate all hypotheses in the candidate set whose Bellman error at step t exceeds the elimination threshold ζ_{elim}. **We highlight that we evaluate the average Bellman error with regard to all the goals in the goal space, instead of only on g^*, which allows better hypothesis elimination.**

Connection with Existing Algorithms. The main difference between our algorithm and the original OLIVE proposed in Jiang et al. (2017) is that, we use goal relabeling when evaluating Bellman error. More specifically, for hypothesis f^t, step h, and the policy π^t, we use

$$g_h^t = \operatorname*{argmax}_{g \in \mathcal{G}} \sum_{(s,a,s') \in \mathcal{D}_h} \Delta_1[s, a, s', g, f_h, r_h, f_{h+1}], \quad (4)$$

where $\Delta_1[s, a, s', g, f_h, r_h, f_{h+1}]$

$$= \left(f_h(s, a, g) - r_h(s, a, g) - \max_{a' \in \mathcal{A}} f_{h+1}(s', a', g) \right),$$

for hypothesis elimination instead of only using the target goal g^*. Intuitively, goal relabeling allows us to eliminate more hypotheses in the elimination phase and improve the sample efficiency compared to algorithms that do not use multiple goals. **The idea of using Bellman error for goal relabeling has also appeared in previous work (Zhang et al., 2020)**. Our analysis provides theoretical justification for this relabeling method. We remark that we use average Bellman error to address the stochastic dynamic.

In general, Equation (4) is difficult to compute, as it requires iterating over the goal space. Existing algorithms (Andrychowicz et al., 2017; Li et al., 2020; Pitis et al., 2020) use the state space as the goal space. They view the states in the data as achieved goals, and minimize Bellman error using achieved goals instead of all possible goals. However, this approach is not sample efficient since they might eliminate more hypotheses by evaluating the Bellman error on multiple goals. **Therefore, their algorithms can be viewed as modifications of Algorithm 1, which favor computational efficiency over sample efficiency.** We further note that, in some special cases, the achieved goal is the same as the goal that maximizes the Bellman error, which is formalized in the following lemma.

Lemma 4.1 (Equivalence of Relabeling Method)**.** *We say a hypothesis class is deterministic, if for any $(f, s, a, g) \in \mathcal{F} \times \mathcal{S} \times \mathcal{A} \times \mathcal{G}$, there exists a unique $s' \in \mathcal{S}$, such that $f_h(s, a, g) = r^*_h(s, a, g) + \max_{a' \in \mathcal{A}} f_{h+1}(s', a', g)$.*

Suppose we have $r^*_H(s, a, g) = \mathbb{1}\{s = g\}$ and $r^*_{H-1}(s, a, g) = 0$ for all $(s, a, g) \in \mathcal{S} \times \mathcal{A} \times \mathcal{G}$. Then for a trajectory $(s_1, a_1, \dots, a_H, s_{H+1})$, we have

$$s_H \in \operatorname*{argmax}_{g \in \mathcal{G}} \Delta_2[s_{H-1}, a_{H-1}, g, f_{H-1}, r_h, f_{h+1}, P^*_h]$$

where $\Delta_2[s, a, g, f, r, f', P]$

$$= \left| f(s, a, g) - r(s, a, g) - \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in \mathcal{A}} f'(s', a', g) \right|$$

for any $f \in \mathcal{F}$ when $\mathcal{F}$ and P^*_{H-1} are deterministic. That is, the achieved goal is the goal that maximize the Bellman error.

Proof. See Appendix §B.1 for a detailed proof. □

Lemma 4.1 shows that in the example in §3, the relabeling method in Algorithm 1 is the same with the relabeling method in Andrychowicz et al. (2017), which theoretically justifies their method. This equivalence also suggests that the empirical performance observed in standard goal-conditioned RL benchmarks for existing relabeling methods could serve as an indirect measure of our algorithm's efficacy.

Repeating sampling using the same policy. In Algorithm 1, we sample multiple trajectories using the policy π^k. This procedure differs from GOLF proposed by Jin et al. (2021a), and we only have the upper bound of sample complexity instead of regret. However, this procedure is necessary since the maximum operator cannot be interchanged with the expectation operator. More specifically, the loss we want to evaluate is

$$\max_{g \in \mathcal{G}} \left| E_{(s,a,s') \sim \mu}[\Delta_1(s, a, s', g, Q_h, r_h, Q_{h+1})] \right|$$

where Δ_1 is defined in (4), and we can only evaluate

$$E_{(s,a,s') \sim \mu}[\max_{g \in \mathcal{G}} |\Delta_1(s, a, s', g, Q_h, r_h, Q_{h+1})|]$$

if we have only one trajectory to each policy. The latter one can be large even for the true hypothesis, and cannot be used for hypothesis elimination.

5. Analysis of GOALIVE

In this section, we present the analysis of GOALIVE. We first introduce the complexity measure we use. We then present the sample complexity of GOALIVE.

5.1. Complexity Measure: Goal-conditioned Bellman-Eluder Dimension

In this section, we propose a new complexity measure called *goal-conditioned Bellman-Eluder dimension*, which quantifies the gain of using goal in model-free algorithms. To enable general function approximation, we modify the Bellman-Eluder dimension, which was proposed by Jin et al. (2021a) to characterize the complexity of an RL problem with general function approximation. We start by developing a new goal-conditioned version of distributional Eluder dimension.

Definition 5.1 (Goal-conditioned ϵ-independence between Distributions)**.** *Let $\mathcal{F}$ be a function class defined on $\mathcal{X} \times \mathcal{G}$, , and $\nu, \mu_1, \dots, \mu_n$ be probability measures over $\mathcal{X}$. We say ν is goal-conditioned ϵ-independent of $\{\mu_1, \mu_2, \dots, \mu_n\}$ with respect to $\mathcal{F}$ and g^* if there exists $f \in \mathcal{F}$ and $g \in \mathcal{G}$ such that*

$$\sqrt{\sum_{i=1}^{n} \max_{g \in \mathcal{G}} \left| \mathbb{E}_{x \sim \mu_i} \Big[f(x, g) \Big] \right|^2} \leq \epsilon,$$

but $|\mathbb{E}_{x \sim \nu}[f(x, g^*)]| > \epsilon$.

Definition 5.2 (Goal-conditioned Distributional Eluder Dimension)**.** *Let $\mathcal{F}$ be a function class defined on $\mathcal{X} \times \mathcal{G}$, and Π be a family of probability measures over $\mathcal{X}$. The goal-conditioned distributional Eluder dimension $\dim_{\mathrm{GOAL-DE}}(\mathcal{F}, \Pi, \epsilon)$ is the length of the longest sequence $\{\rho_1, \dots, \rho_n\} \subset \Pi$ such that there exists $\epsilon' \geq \epsilon$ where ρ_i is ϵ'-independent of $\{\rho_1, \dots, \rho_{i-1}\}$ for all $i \in [n]$.*

When the space of goal only contains the goal we concern about, that is, $\mathcal{G} = \{g^*\}$, Definition 5.2 degenerates to the original definition of distributional Eluder dimension in Jin et al. (2021a). In Definition 5.1, we can consider $\left|\mathbb{E}_{x\sim\mu_i}\left[f(x,g)\right]\right|$ as the loss of hypothesis f concerning the goal g evaluated on the distribution μ_i. Intuitively, ν is independent of $\{\mu_1,\ldots,\mu_n\}$ means if that there exist a hypothesis, where the loss of f is small with respect to all $g \in \mathcal{G}$ and $\{\mu_i\}_{i=1}^n$, but the loss is considerably larger on ν and the target goal g^*. Since we can choose any goals for hypothesis elimination, the loss of the hypothesis that have not been eliminated should be small with respect to all $g \in \mathcal{G}$. Therefore, we incorporate multiple goals in Definition 5.1 by taking a maximum over $g \in \mathcal{G}$. Equipped with Definitions 5.1 and 5.2, we are ready to present the definition of the *goal-conditioned Bellman-Eluder (GOAL-BE) dimension*.

Definition 5.3 (GOAL-conditioned Bellman-Eluder (GOAL-BE) Dimension)**.** *Let $(I-\mathcal{T}_h)\mathcal{F} := \{f_h - \mathcal{T}_h f_{h+1} : f \in \mathcal{F}\}$ 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 $\mathcal{X} \times \mathcal{U}$. The ϵ-goal-conditioned Bellman-Eluder of $\mathcal{F}$ with respect to Π is defined as*

$$\dim_{\mathrm{GOAL-BE}}(\mathcal{F},\Pi,\epsilon) := \max_{h\in[H]} \dim_{\mathrm{GOAL-DE}}\big((I-\mathcal{T}_h)\mathcal{F},\Pi_h,\epsilon\big).$$

Similar with Jin et al. (2021a), GOAL-BE dimension also depends on the choice of the distribution class Π. To simplify the presentation, we only consider $\mathcal{D}_{\mathcal{F}} := \{\mathcal{D}_{\mathcal{F},h}\}_{h=1}^H$, where $\mathcal{D}_{\mathcal{F},h}$ denotes the set of all probability measures over $\mathcal{S} \times \mathcal{A}$ at the $h-$th step, which can be generated by executing the greedy policy π_f with regard to the target goal g^* induced by any $f \in \mathcal{F}$, i.e., $\pi_{f,h}(\cdot) = \mathrm{argmax}_{a\in\mathcal{A}}\, f_h(\cdot,a,g^*)$ for all $h \in [H]$.

Comparison with Original Bellman-Eluder Dimension. Definition 5.3 can be viewed as a modification of the original Bellman-Eluder dimension proposed by Jin et al. (2021a). In fact, Definition 5.3 is more general than the Bellman-Eluder dimension. Definition 5.3 reduces to the original definition of the Bellman-Eluder dimension in Jin et al. (2021a) when $\mathcal{G} = \{g^*\}$. Intuitively, since we have multiple loss for the same hypothesis, we are able to eliminate more hypothesis with the same data, which improves the sample efficiency and reduces the complexity of the original problem. This intuition is formalized in the following two lemmas. Lemma 5.4 shows that GOAL-BE dimension is smaller than the original BE dimension. Lemma 5.5 shows that in the example in §3, GOAL-BE dimension is exponentially smaller than the original BE dimension. Lemmas 5.4 and 5.5 show that using goal-conditioned value functions provably reduces the complexity of a RL problem.

Lemma 5.4 (Strictly Improvement over Original Bellman-Eluder Dimension)**.** *We define*

$$\mathcal{F}_{g^*} = \mathcal{F}_{g^*,1}\times\ldots\times\mathcal{F}_{g^*,H}, \quad \mathcal{F}_g = \mathcal{F}_{g,1}\times\ldots\times\mathcal{F}_{g,H},$$

where $\mathcal{F}_{h,g^*} = \{Q_h(\cdot,\cdot,g^*) \mid Q \in \mathcal{F}_h\}$ and $\mathcal{F}_{h,g} = \{Q_h(\cdot,\cdot,g) \mid Q \in \mathcal{F}_h, g \in \mathcal{G}\}$. We then have

$$\begin{aligned} &\dim_{\mathrm{BE}}(\mathcal{F}_g,\Pi,\epsilon) \geq \dim_{\mathrm{BE}}(\mathcal{F}_{g^*},\Pi,\epsilon) \\ \textit{and } &\dim_{\mathrm{BE}}(\mathcal{F}_{g^*},\Pi,\epsilon) \geq \dim_{\mathrm{GOAL-BE}}(\mathcal{F},\Pi,\epsilon). \end{aligned} \tag{5}$$

Proof. See Appendix §B.2 for a detailed proof. □

Lemma 5.5 (Exponential Improvement in the Best Case)**.** *In the example in §3, we have*

$$\begin{aligned} \dim_{\mathrm{BE}}(\mathcal{F}_{g^*},\mathcal{D}_{\mathcal{F}},\epsilon) &\geq 2^d - 1, \\ \dim_{\mathrm{GOAL-BE}}(\mathcal{F},\mathcal{D}_{\mathcal{F}},\epsilon) &= 1, \end{aligned}$$

where $\mathcal{F}_{g^*} = \mathcal{F}_{g^*,1}\times\ldots\times\mathcal{F}_{g^*,H}$, $\mathcal{F} = \mathcal{F}_1\times\ldots\times\mathcal{F}_H$. Here $\mathcal{F}_{g^*,h}$ and $\mathcal{F}_h$ are defined in (3).

Proof. See Appendix §B.3 for a detailed proof. □

5.2. Sample Complexity of GOALIVE

We first present the realizability assumption on the hypothesis class $\mathcal{F}$, which is commonly used in previous work (Jin et al., 2021a; Sun et al., 2019).

Assumption 5.6 (Realizability)**.** *We assume that $Q_h^\star \in \mathcal{F}_h$ for all $h \in [H]$.*

Realizability requires that the hypothesis class is well-specified and is able to characterizes the connection between different goals. We also introduce the definition of the covering number, which is widely used in previous work (Jin et al., 2021a;b; Liu et al., 2022).

Definition 5.7 (Covering Number of Hypothesis Class)**.** *The ϵ-covering number of a set $\mathcal{V}$ under metric ρ, denoted as $\mathcal{N}(\mathcal{V},\epsilon,\rho)$, is the minimum integer n such that there exists a subset $\mathcal{V}_0 \subset$ with $|\mathcal{V}_0| = n$, and for any $x \in \mathcal{V}$, there exists $y \in \mathcal{V}_0$ such that $\rho(x,y) \leq \epsilon$. For two hypotheses $f = \{f_h\}_{h=1}^H$, $f' = \{f'_h\}_{h=1}^H$, we define*

$$\begin{aligned} &\|f - f'\|_\infty \\ &\quad = \max_{(h,s,a,g)\in[H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{G}} |f_h(s,a,g) - f'_h(s,a,g)|. \end{aligned} \tag{6}$$

For $\mathcal{F} = \mathcal{F}_1\times\ldots\times\mathcal{F}_H$, we define $\mathcal{N}(\mathcal{F},\epsilon) = \max_{h\in[H]}\mathcal{N}(\mathcal{F}_h,\epsilon,\|\cdot\|_\infty)$.

Definition 5.8 (Covering Number of Goal Space). *We say $\mathcal{G}_0$ is an ϵ-covering of the goal space $\mathcal{G}$ if for any $g \in \mathcal{G}$, there exists $g_0 \in \mathcal{G}_0$, such that*

$$|r_h(s,a,g) - r_h(s,a,g_0)| \leq \epsilon,$$
$$|f_h(s,a,g) - f_h(s,a,g_0)| \leq \epsilon,$$

hold for all $f \in \mathcal{F}$, $h \in [H]$, and $(x,u) \in \mathcal{X} \times \mathcal{U}$. Let $\mathcal{G}(\epsilon)$ be the ϵ-covering of $\mathcal{G}$ with the smallest cardinal, and let $\mathcal{N}(\mathcal{G},\epsilon) = |\mathcal{G}(\epsilon)|$.

When $|\mathcal{G}| < \infty$, we have $\mathcal{N}(\mathcal{G},\epsilon) \leq |\mathcal{G}|$. Equipped with the assumption and definitions above, we are ready to present the sample complexity of GOALIVE.

Theorem 5.9 (Sample complexity of GOALIVE.). *Under Assumption 5.6, if we choose*

$$\begin{aligned} &\zeta_{goal} = \epsilon/\big(600H\sqrt{d}\big), \qquad \zeta_{\text{act}} = 2\epsilon/H,\\ &\zeta_{elim} = \epsilon/\big(4H\sqrt{d}\big), \qquad n_{act} = 2592H^2\iota/\epsilon^2,\\ &\text{and } n_{elim} = 4608H^2 d\log^2(\mathcal{N}(\mathcal{F},\zeta_{\text{elim}}/8))\cdot\iota/\epsilon^2,\\ &\text{where} \quad d = \dim_{\text{GOAL}-\text{BE}}\big(\mathcal{F},\mathcal{D}_{\mathcal{F}},\epsilon/H\big),\\ &\qquad\qquad \iota = \log^2[\mathcal{N}(\mathcal{G},\zeta_{goal})Hd/(\delta\epsilon)], \end{aligned}$$

then with probability at least $1-\delta$, Algorithm 1 will output an $\mathcal{O}(\epsilon)$-optimal policy using at most $\mathcal{O}(H^3 d^2 \log^2[\mathcal{N}_{\mathcal{F}}(\zeta_{elim}/8)]\cdot\iota/\epsilon^2)$ episodes.

Proof. See Appendix §C for a detailed proof. □

Theorem 5.9 shows that RL problems with low GOAL-BE dimension can be solved efficiently by GOALIVE under the realizability assumption. The sample complexity of GOALIVE is $\widetilde{\mathcal{O}}(H^3d^2/\epsilon^2)$, which is polynomial in $1/\epsilon$ and the number of horizon. The dependency on the complexity measure is the same with Jin et al. (2021a). However, as we show in Lemma 5.4 and Lemma 5.5, GOAL-BE dimension is not larger than the original BE dimension, and can be exponentially smaller in the best case. The cost we pay for the improvement on the complexity measure is the log-covering number of the goal space. For the example in §3.1, we have $|\mathcal{G}| = 2^d$. Therefore, we have $\log\mathcal{N}(\mathcal{G},\zeta_{\text{goal}}) \leq H\log 2$, which is polynomial in the horizon H.

Comparison with previous work on multitask RL. The setting we study can be viewed as a special case of multitask RL. The difference is that we only concern about the suboptimality of the target task g^* instead of all task. Previous work (Lu et al., 2022) has also developed sample-efficient algorithms for multitask RL. Lu et al. (2022) designed a model-free algorithm for multitask RL and demonstrated its sample efficiency. However, the complexity measure they used is the original Eluder dimension, which cannot exploit the connection between goals. Moreover, they considered that the tasks are randomly generated, which is different from the empirical work of goal relabeling, where the algorithm can set the goal in the training phase. Therefore, their results cannot explain the success of goal relabeling.

Connection with reward-free exploration. Our algorithm is similar with reward-free exploration, in the sense that the exploration strategy does not rely on the original reward. Existing results in reward-free exploration use model-based algorithm (Jin et al., 2020a; Wang et al., 2020). However, model-based algorithms tend to have a larger asymptotic bias. In the case of complex problems, the dynamics cannot be learned perfectly, and the final policy can be highly suboptimal (Pong et al., 2018). Therefore, empirical work that uses goal relabeling often applies model-free algorithms, and their success and not be explained by the work that use model-based algorithms. In fact, our algorithm can be viewed as a model-free algorithm for reward-free exploration.

Comparison with Zhu & Zhang (2023). In the context of goal-conditioned RL, previous research, including the notable work by Zhu & Zhang (2023), has provided a theoretical analysis. However, their analysis focuses on offline RL and does not design algorithm for provably efficient exploration. Moreover, they assume that the dataset covers the optimal policy with respect to all $g \in \mathcal{G}$, whereas our analysis does not rely on such a stringent assumption.

Acknowledgement

Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma for their supports.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Pieter Abbeel, O., and Zaremba, W. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

Cheng, Y., Feng, S., Yang, J., Zhang, H., and Liang, Y. Provable benefit of multitask representation learning in reinforcement learning. *arXiv preprint arXiv:2206.05900*, 2022.

Colas, C., Fournier, P., Chetouani, M., Sigaud, O., and Oudeyer, P.-Y. Curious: intrinsically motivated modular multi-goal reinforcement learning. In *International conference on machine learning*, pp. 1331–1340. PMLR, 2019.

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.

Fang, M., Zhou, T., Du, Y., Han, L., and Zhang, Z. Curriculum-guided hindsight experience replay. *Advances in neural information processing systems*, 32, 2019.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pp. 4870–4879. PMLR, 2020a.

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. PMLR, 2020b.

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

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

Li, A., Pinto, L., and Abbeel, P. Generalized hindsight for reinforcement learning. *Advances in neural information processing systems*, 33:7754–7767, 2020.

Liu, Z., Zhang, Y., Fu, Z., Yang, Z., and Wang, Z. Learning from demonstration: Provably efficient adversarial policy imitation with linear function approximation. In *International conference on machine learning*, pp. 14094–14138. PMLR, 2022.

Liu, Z., Lu, M., Xiong, W., Zhong, H., Hu, H., Zhang, S., Zheng, S., Yang, Z., and Wang, Z. Maximize to explore: One objective function fusing estimation, planning, and exploration. *Advances in Neural Information Processing Systems*, 36, 2024.

Lu, R., Zhao, A., Du, S. S., and Huang, G. Provable general function class representation learning in multitask bandits and mdps. *arXiv preprint arXiv:2205.15701*, 2022.

Nair, A. V., Pong, V., Dalal, M., Bahl, S., Lin, S., and Levine, S. Visual reinforcement learning with imagined goals. *Advances in neural information processing systems*, 31, 2018.

Pitis, S., Chan, H., Zhao, S., Stadie, B., and Ba, J. Maximum entropy gain exploration for long horizon multi-goal reinforcement learning. In *International Conference on Machine Learning*, pp. 7750–7761. PMLR, 2020.

Pong, V., Gu, S., Dalal, M., and Levine, S. Temporal difference models: Model-free deep rl for model-based control. *arXiv preprint arXiv:1802.09081*, 2018.

Riedmiller, M., Hafner, R., Lampe, T., Neunert, M., Degrave, J., Wiele, T., Mnih, V., Heess, N., and Springenberg, J. T. Learning by playing solving sparse reward tasks from scratch. In *International conference on machine learning*, pp. 4344–4353. PMLR, 2018.

Savinov, N., Raichuk, A., Marinier, R., Vincent, D., Pollefeys, M., Lillicrap, T., and Gelly, S. Episodic curiosity through reachability. *arXiv preprint arXiv:1810.02274*, 2018.

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, 2017.

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. In *Conference on learning theory*, pp. 2898–2933. PMLR, 2019.

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

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

Zhang, T., Ren, T., Yang, M., Gonzalez, J., Schuurmans, D., and Dai, B. Making linear mdps practical via contrastive representation learning. In *International Conference on Machine Learning*, pp. 26447–26466. PMLR, 2022.

Zhang, Y., Abbeel, P., and Pinto, L. Automatic curriculum learning through value disagreement. *Advances in Neural Information Processing Systems*, 33:7648–7659, 2020.

Zhu, H. and Zhang, A. Provably efficient offline goal-conditioned reinforcement learning with general function approximation and single-policy concentrability. *arXiv preprint arXiv:2302.03770*, 2023.

List of Notation

In the sequel, we present a list of notations in the paper.

Notation	Explanation
$\mathcal{S}, \mathcal{A}, \mathcal{G}$	The state, the action, and the goal spaces, respectively.
h	The length of an episode.
t	The index of the iteration in Algorithm 1 (GOALIVE).
$\mathcal{E}$	The average Bellman error defined in (2).
Φ	The $\mathcal{F}$-profile we defined i Definition 3.1.
$\mathcal{F}$	The hypothesis class used in model-free algorithms.
g^*	The target goal we concern about.
$\mathrm{dim}_{\mathrm{BE}}, \mathrm{dim}_{\mathrm{GOAL-BE}}$	The original BE dimension, and the GOAL-BE dimension.
$\{\mathcal{T}_h\}_{h=1}^H$	The Bellman operator we defined in §2.

Structure of Appendix

We provide a detailed proof of Lemma 3.2 in Appendix A, and provide the proof of Lemmas in §4, §5.1 in Appendix §C. We provide the proof of auxiliary lemmas in §D.

Appendix

Table of Contents

A. Proof of Lemma 3.2

In this section, we first provide the proof of Lemma 3.2. We then provide more detail on the action-value function $Q_h^{\mathbf{p}}$.

Proof of Lemma 3.2. Our proof consists of two parts. In the first part, we show the inefficiency of model-free algorithms without multiple goals. In the second part, we show that there exists a goal-conditioned model-free algorithm that achieves provably efficient exploration.

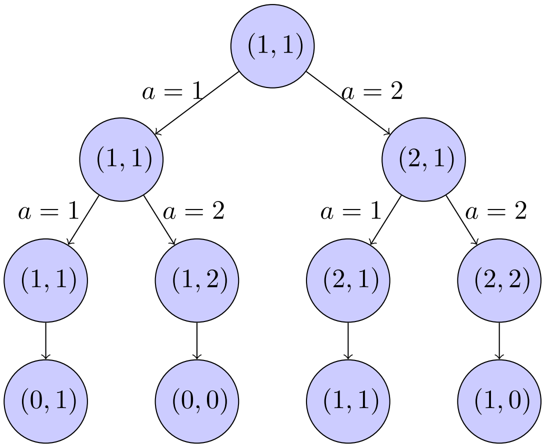

Figure 2. An example of the MDP construction in §3, with $d = 2$ and $H = 4$. All models are deterministic, and each model is uniquely indexed by a sequence of actions $\mathbf{p}$. We use $\mathbf{p}^*$ to denote the index of the true model. Here $\mathbf{p}^* = (2, 1)$. The last transition is designed such that the agent always lands in a state that contains "0" unless it follows path $\mathbf{p}^*$.

Inefficiency of Model-free Algorithms without Multiple Goals. Our proof of the inefficiency of model-free algorithms her is similar with the proof of Theorem 2 of Sun et al. (2019). We construct another class of non-factored models, such that (1) learning in this new class is intractable, and (2) the two families are indistinguishable to any model-free algorithm without multiple goals.

The new model class is obtained by transforming each $\mathcal{P}^{\mathbf{p}} \in \mathcal{M}$ into $\widetilde{\mathcal{P}}^{\mathbf{p}}$. $\widetilde{\mathcal{P}}^{\mathbf{p}}$ has the same state space and transitions as $\mathcal{P}^{\mathbf{p}}$, except for the transition from horizon $H-1$ to H. The transition in the H-th step still does not depend on the action a_H. In the transition indexed by $\mathbf{p} = \{p_1, p_2, \dots, p_d\}$, we will transit to

$$s_{H+1} = \begin{cases} [1]^d & \text{when } s_H = \mathbf{p}, \\ [0]^d & \text{else.} \end{cases}$$

The reward function is the same as in the original model class. This example is similar with the one in Sun et al. (2019). Solving this example is equivalent to solving a multi-armed bandit problem with one optimal arm among 2^{H-2} arms.

Therefore, the sample complexity of any algorithm in the new example is $\Omega(2^H)$.

To prove that the two model families are indistinguishable for model-free algorithms without multiple goals, which is defined in Definition 3.1, we show that the original $\mathcal{F}$-profiles, which is $\Phi_{\mathcal{F},g^*}$ in Definition 3.1, in $\mathcal{P}^{\mathbf{p}}$ are identical to those in $\widetilde{\mathcal{P}}^{\mathbf{p}}$. This implies that any model-free algorithm behaves in the same way in $\mathcal{P}^{\mathbf{p}}$ and $\widetilde{\mathcal{P}}^{\mathbf{p}}$, so that the sample complexity must be identical, and hence $\Omega(2^H)$. To continue, we introduce the definition of the optimal planning, which is an important tool to connect these two model families.

Definition A.1 (Optimal Planning). *Let $\mathcal{M}$ be a class of transition. For a transition $\mathcal{P} = \{\mathcal{P}_h\}_{h=1}^H \in \mathcal{M}$, we denote by $\pi^{\mathcal{P}} = \{\pi_h^{\mathcal{P}}\}_{h=1}^H$, $V^{\mathcal{P}} = \{V_h^{\mathcal{P}}\}_{h=1}^H$, $Q^{\mathcal{P}} = \{Q_h^{\mathcal{P}}\}_{h=1}^H$ the optimal policy, value function, and action-value function when the transition of the MDP is $\mathcal{P}$. We denote by $\mathrm{OP}(\mathcal{P})$ the mapping that maps a transition $\mathcal{P}$ to its optimal policy and optimal Q function, that is $\mathrm{OP}(\mathcal{P}) \triangleq (\pi^{\mathcal{P}}, Q^{\mathcal{P}})$. We then define $\mathrm{OP}(\mathcal{M}) \triangleq \{(\pi^{\mathcal{P}}, Q^{\mathcal{P}}) \mid \mathcal{P} \in \mathcal{M}\}$.*

Let $\mathcal{M} = \{\mathcal{P}^{\mathbf{p}}\}_{\mathbf{p}\in\{1,2\}^d}$ and $\widetilde{\mathcal{M}} = \{\widetilde{\mathcal{P}}^{\mathbf{p}}\}_{\mathbf{p}\in\{-1,1\}^d}$. Let $\mathcal{F} = \{\mathcal{F}_h\}_{h=1}^H, \Pi = \{\Pi_h\}_{h=1}^H$ to be the Q class and policy classes from $\mathrm{OP}(\mathcal{M},\mathcal{G})$, $\widetilde{\mathcal{F}} = \{\widetilde{\mathcal{F}}_h\}_{h=1}^H$ and $\widetilde{\Pi} = \{\widetilde{\Pi}_h\}_{h=1}^H$ be the policy class from $\mathrm{OP}(\widetilde{\mathcal{M}},\mathcal{G})$. Since all MDPs of interest have fully deterministic dynamics, and non-zero rewards only occur at the last step, it suffices to show that for any sequence of actions $\mathbf{a} = (a_1, \dots, a_H)$, (1) the final reward has the same distribution for $P^{\mathbf{p}}$ and $\widetilde{P}^{\mathbf{p}}$, and (2) the original $\mathcal{F}$-profiles $[Q_h(s_h, a, g^*)]_{Q_h\in\mathcal{F}_h, a\in\mathcal{A}}$ and $[Q_h(s_h, a, g^*)]_{Q_h\in\widetilde{\mathcal{F}}_h, a\in\mathcal{A}}$ are equivalent at all states generated by taking $\mathbf{a}$ in $P^{\mathbf{p}}$ and $\widetilde{P}^{\mathbf{p}}$, respectively. The equivalence of the reward is obvious, In the following, we study the equivalence of the original $\mathcal{F}$-profiles.

In $P^{\mathbf{p}}$ and at level H, by the definition of the reward function, the original $\mathcal{F}$-profile is $[1]^{|\mathcal{F}_H|}$ for the state without “0” and $[0]^{|\mathcal{F}_H|}$ otherwise. Thus, when the action sequence $\mathbf{a} = \mathbf{p}$, the original $\mathcal{F}$-profile of the reached state is $[1]^{|\mathcal{Q}|}$. Otherwise, the original $\mathcal{F}$-profile of the reached state is $[0]^{|\mathcal{Q}|}$. Similarly, in $\widetilde{P}^{\mathbf{p}}$ the original $\mathcal{F}$-profile is $[0]^{|\widetilde{\mathcal{F}}_H|}$ if the state is $[0]^d$, and it is $[1]^{|\widetilde{\mathcal{F}}_H|}$ otherwise. The equivalence here is obvious since $|\widetilde{\mathcal{F}}_H| = |\mathcal{F}_H| = 2^{H-2}$.

For level $H-1$, no matter the true model path $\mathbf{p}$, the $Q^{\mathbf{p}'}$ associated with path $\mathbf{p}'$ has value $Q_{H-1}^{\mathbf{p}'}(\mathbf{a}, +1, g^*) = \mathbb{1}\{\mathbf{a} = \mathbf{p}'\}$ at state $\mathbf{a}$. Hence the Q-profile at $\mathbf{a}$ can be represented as $[\mathbb{1}\{\mathbf{a} = \mathbf{p}'\}]_{\mathbf{p}'\in\{-1,1\}^d}$, for both $P^{\mathbf{p}}$ and $\widetilde{P}^{\mathbf{p}}$. We remark that the original $\mathcal{F}$-profile does not depend on the true model $\mathbf{p}$ because all models agree on the dynamics before the last step. Similarly, for $h < H-1$ where each state has two actions $\{-1, 1\}$, we have:

$$Q^{\mathbf{p}'}(\mathbf{a}_{1:h-1}, 1, g^*) = \mathbb{1}\{\mathbf{a}_{1:h-1} \circ 1 = \mathbf{p}'_{1:h}\}, Q^{\mathbf{p}'}(\mathbf{a}_{1:h-1}, 2, g^*) = \mathbb{1}\{\mathbf{a}_{1:h-1} \circ 2 = \mathbf{p}'_{1:h}\}.$$

Hence, the original $\mathcal{F}$-profile can be represented as:

$$\Big[\big(\mathbb{1}\{\mathbf{a}_{1:h-1} \circ 1 = \mathbf{p}'_{1:h}\}, \mathbb{1}\{\mathbf{a}_{1:h-1} \circ 2 = \mathbf{p}'_{1:h}\}\big)\Big]_{\mathbf{p}'\in\{-1,1\}^d},$$

which is the same between $P^{\mathbf{p}}$ and $\widetilde{P}^{\mathbf{p}}$. Therefore, the model $P^{\mathbf{p}}$ and $\widetilde{P}^{\mathbf{p}}$ have exactly the same original $\mathcal{F}$-profile for all action sequences, implying that any model-free algorithm without multiple goals, must have the same behavior on both transition classes. Since the transition class $\widetilde{\mathcal{M}} = \{\widetilde{P}^{\mathbf{p}}\}_{\mathbf{p}}$ admits an information-theoretic sample complexity lower bound of $\Omega(2^H)$, the same lower bound applies to $\mathcal{M} = \{P^{\mathbf{p}}\}_{\mathbf{p}}$ for model-free algorithms without multiple goals.

Sample Efficiency of Goal-conditioned Model-free Algorithm This can be directly proved by combining Lemma 5.5 and Theorem 5.9. By Theorem 5.9, when we use Algorithm `GOALIVE`, we can obtain a 4ϵ-optimal policy with at most $15000H^3 d_0^2 \log^2(\mathcal{N}(\mathcal{F}, \zeta_{\text{elim}}/8)) \cdot \iota/\epsilon^2$ episodes with probability $1-\delta$. Here $\iota = \log^2[\mathcal{N}(\mathcal{G}, \zeta_{\text{goal}})Hd_0/(\delta\epsilon)]$ and $d_0 = \dim_{\mathrm{GOAL-BE}}\big(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon/H\big)$. Since both $\mathcal{F}$ and $\mathcal{G}$ is finite, we have $\mathcal{N}(\mathcal{G}, \zeta_{\text{goal}}) = \mathcal{N}(\mathcal{F}, \zeta_{\text{elim}}/8) = 2^{H-2}$. By Lemma 5.5, we have $d_0 = 1$. Therefore, the number of episodes is bounded from the above by $15000H^7\log^2[H/(\delta\epsilon)]/\epsilon^2$, which is $\mathrm{poly}(H, 1/\epsilon, \log(1/\delta))$. Thus, we conclude the proof of Lemma 3.2.

□

B. Proof of Lemmas in §5.1 ans §4

B.1. Proof of Lemma 4.1

Proof of Lemma 4.1. Since we have $Q_H(s,a,g) = \mathbb{1}_{s=g}$, for any $s \in \mathcal{S}$, there exists a unique $g \in \mathcal{G}$, such that $Q_H(s,a,g) = 1$. Since P_{H-1} is deterministic, there exists a unique $g \in \mathcal{G}$, such that

$$\mathcal{T}_{H-1} f_H(s_{H-1}, a_{H-1}, g') = \mathbb{1}_{g=g'}.$$

Since $\mathcal{F}$ is deterministic, there exists a unique $g \in \mathcal{G}$, such that $f_{H-1}(s_{H-1}, a_{H-1}, g') = \mathbb{1}_{g=g'}$. We also know that $f_H(s_H, a_H, g) = 1$ if and only is $g = s_H$. Therefore, we have

$$f_{H-1}(s_{H-1}, a_{H-1}, g') - r^*_{H-1}(s_{H-1}, a_{H-1}, g') - \max_{a' \in \mathcal{A}} f_H(s_H, a', g') \in \{0, 1\}$$

for any $g' \in \mathcal{G}$. Therefore, Lemma 4.1 holds if we have

$$f_{H-1}(s_{H-1}, a_{H-1}, g') - r^*_{H-1}(s_{H-1}, a_{H-1}, g') - \max_{a' \in \mathcal{A}} f_H(s_H, a', g') = 1$$

when $g' = s_H$. In the following, we consider the case that

$$f_{H-1}(s_{H-1}, a_{H-1}, g') - r^*_{H-1}(s_{H-1}, a_{H-1}, g') - \max_{a' \in \mathcal{A}} f_H(s_H, a', g') = 0$$

when $g' = s_H$. In this case, we have $f_{H-1}(s_{H-1}, a_{H-1}, g') = 0$ when $g' = s_H$. Therefore, we have $f_{H-1}(s_{H-1}, a_{H-1}, g') = \mathbb{1}_{g'=s_H}$, and

$$f_{H-1}(s_{H-1}, a_{H-1}, g) - r^*_{H-1}(s_{H-1}, a_{H-1}, g) - \max_{a' \in \mathcal{A}} f_H(s_H, a', g) = \mathbb{1}_{g=s_H} - \mathbb{1}_{g=s_H} = 0$$

for all $g \in \mathcal{G}$. Therefore, we conclude the proof of Lemma 4.1. □

B.2. Proof of Lemma 5.4

Proof of Lemma 5.4. By the definition of $\mathcal{F}_{h,g^*}$ and $\mathcal{F}_{h,g}$, we have $\mathcal{F}_{h,g^*} \subset \mathcal{F}_{h,g}$. By Definition 2.2, we have $\dim_{\mathrm{DE}}((\mathcal{I} - \mathcal{T}_h)\mathcal{F}_{h,g^*}, \Pi_h, \epsilon) \leq \dim_{\mathrm{DE}}((\mathcal{I} - \mathcal{T}_h)\mathcal{F}_{h,g}, \Pi_h, \epsilon)$. By Definition 2.3, we have $\dim_{\mathrm{BE}}(\mathcal{F}_{g^*}, \Pi, \epsilon) \leq \dim_{\mathrm{BE}}(\mathcal{F}_g, \Pi, \epsilon)$, which finish the first part of the proof.

In the following part of the proof, we show that $\dim_{\mathrm{GOAL-BE}}(\mathcal{F}, \Pi, \epsilon) \leq \dim_{\mathrm{BE}}(\mathcal{F}_{g^*}, \Pi, \epsilon)$. It is sufficient to show that $\dim_{\mathrm{GOAL-DE}}((\mathcal{I} - \mathcal{T}_h)\mathcal{F}, \Pi_h, \epsilon) \leq \dim_{\mathrm{DE}}((\mathcal{I} - \mathcal{T}_h)\mathcal{F}_{g^*}, \Pi, \epsilon)$. We only need to show that if ν is goal-conditioned ϵ-independent with $\{\mu_1, \ldots, \mu_n\}$, we then have ν is ϵ-independent with $\{\mu_1, \ldots, \mu_n\}$. By the definition of goal-conditioned ϵ-independence, we have $|\mathbb{E}_{s,a\sim\nu} f_h(s,a,g^*) - r_h(s,a,g^*) - \mathcal{T}_h f_{h+1}(s,a,g^*)| \geq \epsilon$ and

$$\sqrt{\sum_{i=1}^n \max_{g \in \mathcal{G}} |\mathbb{E}_{s,a\sim\mu_i} f_h(s,a,g) - r_h(s,a,g) - \mathcal{T}_h f_{h+1}(s,a,g)|^2} \leq \epsilon$$

for some $f \in \mathcal{F}$. Therefore, we have

$$\sqrt{\sum_{i=1}^n |\mathbb{E}_{s,a\sim\mu_i} f_h(s,a,g^*) - r_h(s,a,g^*) - \mathcal{T}_h f_{h+1}(s,a,g^*)|^2} \leq \epsilon.$$

We conclude the proof by the definition of ϵ-independence. □

B.3. Proof of Lemma 5.5

Proof of Lemma 5.5. By the definition of $\mathcal{F}$, we know that

$$f_h(s,a,g) - r_h^*(s,a,g) - \mathcal{T}_h f_{h+1}(s,a,g) = 0$$

for any $(f,h,s,a,g) \in \mathcal{F} \times [H-2] \times \mathcal{S} \times \mathcal{AG}$. Therefore, we have

$$\dim_{\mathrm{DE}}((\mathcal{I} - \mathcal{T}_h)\mathcal{F}_{h,g^*}, \mathcal{D}_{\mathcal{F},h}, \epsilon) = \dim_{\mathrm{GOAL-DE}}((\mathcal{I} - \mathcal{T}_h)\mathcal{F}_h, \mathcal{D}_{\mathcal{F},h}, \epsilon) = 1$$

for $h \in [H-2]$. Therefore, we only need to consider the case where $h = H-1$. For a hypothesis $\mathbf{p} \in \{1,2\}^d$, we denote by $f^{\mathbf{p}} = \{f_h^{\mathbf{p}}\}_{h=1}^H \in \mathcal{F}$ the corresponding value. By the setting of the dynamic system, when we use the optimal policy with respect to g^* and $f^{\mathbf{p}}$, we will arrive the state $s_{H-1} = \mathbf{p}$. Therefore, we have $\mathcal{D}_{\mathcal{F},H-1} = \{\delta_{(s,a)} \mid s \in \{1,2\}^d, a \in \mathcal{A}\}$. For any $f^{\mathbf{p}} \in \mathcal{F}$, $\delta_{(s_{H-1},a)} \in \mathcal{D}_{\mathcal{F},H-1}$, we have

$$\mathbb{E}_{(s,a)\sim\mu} f_{H-1}^{\mathbf{p}}(s,a,g^*) = 0,\ r_{H-1}^*(s,a,g^*) = 0,\ \mathbb{E}_{(s,a)\sim\mu} \mathcal{T}_{H-1} f_H^{\mathbf{p}}(s,a,g^*) = 0$$

when $s_{H-1} \neq \mathbf{p}^*$ and $s_{H-1} \neq \mathbf{p}$. We also have

$$\mathbb{E}_{(s,a)\sim\mu} f_{H-1}^{\mathbf{p}}(s,a,g^*) = 1,\ r_{H-1}^*(s,a,g^*) = 0,\ \mathbb{E}_{(s,a)\sim\mu} \mathcal{T}_{H-1} f_H^{\mathbf{p}}(s,a,g^*) = 0$$

when $s_{H-1} \neq \mathbf{p}^*$ and $s_{H-1} = \mathbf{p}$. Since we have 2^d hypotheses in total, we consider the sequence $\mu_1 = \delta_{(s_{H-1}^1,a)}, \ldots, \mu_n = \delta_{(s_{H-1}^n,a)}$, such that $n = 2^d - 1$, $\mu_i \neq \delta_{(\mathbf{p}^*,a)}$ for $i \in [n]$, $\mu_i \neq \mu_j$ when $i \neq j$. We set $\mathbf{p}[i] = s_{H-1}^1$ and choose $f^i = \{f_h^i\}$ the corresponding value function. We can easily verify that

$$\mathbb{E}_{(s,a)\sim\mu_i}[f_{H-1}^j(s,a,g^*) - r_{H-1}^*(s,a,g^*) - \mathcal{T}_{H-1} f_H^j(s,a,g^*)] = 0$$

when $i < j$. Therefore, we have

$$\sum_{i=1}^{j-1} \left| \mathbb{E}_{(s,a)\sim\mu_i}[f_{H-1}^j(s,a,g^*) - r_{H-1}^*(s,a,g^*) - \mathcal{T}_{H-1} f_H^j(s,a,g^*)] \right|^2 = 0.$$

We also have

$$\mathbb{E}_{(s,a)\sim\mu_i}[f_{H-1}^i(s,a,g^*) - r_{H-1}^*(s,a,g^*) - \mathcal{T}_{H-1} f_H^i(s,a,g^*)] = 1.$$

Therefore, μ_i is ϵ-independent with $\{\mu_1, \cdots, \mu_{i-1}\}$. By the definition of distributional Eluder dimension, we have $\dim_{\mathrm{DE}}((\mathcal{I} - \mathcal{T}_{H-1})\mathcal{F}_{H-1,g^*}, \mathcal{D}_{\mathcal{F},H-1}, \epsilon) \geq 2^d - 1$ for $\epsilon < 1$.

For the second part of the proof, we first show that

$$\max_{g\in\mathcal{G}} \left| f_{H-1}(s,a,g) - r_{H-1}^*(s,a,g) - \mathbb{E}_{s'} \max_{a'} f_H(s',a',g) \right| < 1$$

if and only if $f_{H-1} = Q_{H-1}^*$, $f_H = Q_H^*$. First, in the given hypothesis class, we have $f_H = Q_H^*$ holds for all $f \in \mathcal{F}$. When we select $g = s_H$, we have

$$\left| f_{H-1}(s,a,g) - r_{H-1}^*(s,a,g) - \mathbb{E}_{s'} \max_{a'} f_H(s',a',g) \right| = \left| f_{H-1}(s,a,s_H) - 1 \right|.$$

When $f_{H-1} \neq Q_{H-1}$, we have $f_{H-1}(s,a,s_H) = 0$, which contradict with $|f_{H-1}(s,a,s_H) - 1| < 1$. Therefore, we have $f_{H-1} = Q_{H-1}^*$.

If ν is goal-conditioned ϵ-independent with μ, we have

$$\max_{g\in\mathcal{G}} \mathbb{E}_{s,a\sim\mu}\left|f_{H-1}(s,a,g) - r^*_{H-1}(s,a,g) - \mathbb{E}_{s'}\max_{a'} f_H(s',a',g)\right| < \epsilon \tag{7}$$

and

$$\mathbb{E}_{s,a\sim\nu}\left|f_{H-1}(s,a,g^*) - r^*_{H-1}(s,a,g^*) - \mathbb{E}_{s'}\max_{a'} f_H(s',a',g^*)\right| > \epsilon. \tag{8}$$

By (7), we have $f_{H-1} = Q^*_{H-1}$, $f_H = Q^*_H$, which implies

$$\mathbb{E}_{s,a\sim\nu}\left|f_{H-1}(s,a,g^*) - r^*_{H-1}(s,a,g^*) - \mathbb{E}_{s'}\max_{a'} f_H(s',a',g^*)\right| = 0,$$

which contradicts with (8). Therefore, for any $\mu,\nu\in\mathcal{D}_{\mathcal{F},H-1}$, ν is not goal-conditioned ϵ-independent with μ. Therefore, we have $\dim_{\mathrm{GOAL-DE}}((\mathcal{I}-\mathcal{T}_{H-1})\mathcal{F},\mathcal{D}_{\mathcal{F},H-1},\epsilon) = 1$, which concludes the proof. □

C. Proof of Theorem 5.9

Proof of Theorem 5.9. We denote by T the index of the last iteration of Algorithm 1. By Algorithm 1, we know that the elimination procedure is activated at the t-th iterations for $t\in[T-1]$, and is not activated in the T-th iteration. We also know that the output policy is π^T, which is the greedy policy with respect to f^T. We use h^t to denote the step that the elimination procedure is activated in the t-th iteration. The following lemma provides a characterization of $\{f_t\}_{t=1}^{T-1}$ and the elimination procedure.

Lemma C.1. *We have $\mathcal{E}(f^t,\pi^t,h^t,g^*) > \epsilon/H$ for all $t\in[T']$ with probability $1-\delta/8$. We also have $\mathcal{E}(f,\pi^t,h^t) < \epsilon/(H\sqrt{d})$ for all $t\in[T']$ when $f\in\mathcal{B}^t$ with probability $1-\delta/8$. Here $\mathcal{E}(f^t,\pi^t,h^t,g^*)$ and $\mathcal{E}(f,\pi^t,h^t)$ are defined in (2), and $T' = \min\{T-1, dH+1\}$.*

Proof. See Appendix §D.1 for a detailed proof. □

Lemma C.1 shows that, the average Bellman error of f^t with respect to π^t is large, and hypotheses with a large average Bellman error with respect to π^t will be eliminated with high probability. We denote by $\mathcal{E}_1$ the event in Lemma C.1. The following lemma shows that $T\le dH+1$ when $\mathcal{E}_1$ holds, which bounds the number of iterations in Algorithm 1.

Lemma C.2. *When we condition on $\mathcal{E}_1$, we have $T\le dH+1$.*

Proof. See Appendix §D.2 for a detailed proof. □

We also have the following lemma, which shows that the average Bellman error of f^T is small.

Lemma C.3. *When $T\le dH+1$, we have $\sum_{h=1}^{H}\mathcal{E}(f^T,\pi^T,h,g^*) < 4\epsilon$ holds with probability $1-\delta/4$.*

Proof. See Appendix §D.3 for a detailed proof. □

In addition, we have the following lemma, which show that the true hypothesis will not be eliminated with high probability.

Lemma C.4. *We define $Q^* = \{Q^*_h\}_{h=1}^H$, where Q^*_h is the optimal goal-conditioned action-value function we defined in §2. Then with probability $1-\delta/4$, we have $Q^*\in\mathcal{B}^t$ for all $t\in[T']$. Here $T' = \min\{T-1, dH+1\}$.*

Proof. See Appendix §D.4 for a detailed proof. □

We denote by $\mathcal{E}_3$ the event we defined in Lemma C.4. In the following part of the proof, we condition on Events $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$. In the following, we show that the suboptimality of π^T is small when we condition on $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$. When we condition on Event $\mathcal{E}_1$, We have $V_1^*(s_1, g^*) < f_1^T(s_1, a_1, g^*)$ by Line 4 of Algorithm 1. Therefore, we have

$$V_1^\star(x_1, g^*) - V_1^{\pi^T}(x_1, g^*) \le \max_a f_1^T(s_1, a, g^*) - V^{\pi^T}(s_1, g^*). \tag{9}$$

Since $f_{H+1}^T(s, a, g) = 0$ and $V_{H+1}^{\pi^T}(s, a, g) = 0$ for all $(s, a, g) \in \mathcal{S} \times \mathcal{A} \times \mathcal{G}$, we have

$$\max_a f_1^T(s_1, a, g^*) - V^{\pi^T}(s_1, g^*) = \sum_{h=1}^{H} \mathbb{E}_{\pi^T}\big[f_h^T(s_h, a_h, g^*) - r_h^*(s_h, a_h, g^*) - f_{h+1}^T(s_{h+1}, a_{h+1}, g^*)\big]$$
$$- \sum_{h=1}^{H} \mathbb{E}_{\pi^T}\big[Q_h^{\pi^T}(s_h, a_h, g^*) - r_h^*(s_h, a_h, g^*) - Q_{h+1}^{\pi^T}(s_{h+1}, a_{h+1}, g^*)\big]. \tag{10}$$

Here the expectation $\mathbb{E}_{\pi^T}[\cdot]$ is taken with respect to $s_{h+1} \sim \mathcal{P}_h^*(\cdot \mid s_h, a_h)$ and $a_h \sim \pi_h^T(\cdot \mid s_h, g^*)$. By Bellman equation, we have

$$\mathbb{E}_{\pi^T}\big[Q_h^{\pi^T}(s_h, a_h, g^*) - r_h^*(s_h, a_h, g^*) - Q_{h+1}^{\pi^T}(s_{h+1}, a_{h+1}, g^*)\big] = 0. \tag{11}$$

By (2), we have

$$\mathbb{E}_{\pi^T}\big[f_h^T(s_h, a_h, g^*) - r_h^*(s_h, a_h, g^*) - f_{h+1}^T(s_{h+1}, a_{h+1}, g^*)\big] = \mathcal{E}(f^T, \pi^T, h, g^*). \tag{12}$$

Combining (9), (10), (11), and (12), we have

$$V_1^\star(x_1, g^*) - V_1^{\pi^T}(x_1, g^*) \le \sum_{h=1}^{H} \mathcal{E}(f^T, \pi^T, h, g^*).$$

When we condition on $\mathcal{E}_2$ in Lemma C.3, we have $\sum_{h=1}^{H} \mathcal{E}(f^T, \pi^T, h, g^*) \le 4\epsilon$. Therefore, when we condition on $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$, the policy we return is $\mathcal{O}(\epsilon)$-optimal. By Lemmas C.1, C.3, and C.4, we have $P(\cap_{i=1}^3 \mathcal{E}_i) \ge 1 - \delta$. Therefore, with probability at least $1 - \delta$, Algorithm 1 will terminate in $T \le (dH + 1)$ iterations and output a 4ϵ-optimal policy using at most

$$(dH + 1)(n_{\text{act}} + n_{\text{elim}}) \le \frac{15000 H^3 d^2 \log^2(\mathcal{N}(\mathcal{F}, \zeta_{\text{elim}}/8)) \cdot \iota}{\epsilon^2}$$

episodes. Thus, we conclude the proof of Theorem 5.9. □

D. Proof of Lemmas in Appendix §C

D.1. Proof of Lemma C.1

Proof. First, we have the following lemma, which shows that in the activation phase, $\widehat{\mathcal{E}}$ in Line 6 of Algorithm 1 is a good estimator of $\mathcal{E}$ in (2).

Lemma D.1. *[Concentration in the Activation Phase] With probability $1 - \delta/8$, we have*

$$|\widehat{\mathcal{E}}(f^t, \pi^t, h, g^*) - \mathcal{E}(f^t, \pi^t, h, g^*)| \le \epsilon/(6H)$$

holds for all $(t, h) \in [dH + 1] \times [H]$.

Proof. See Appendix §D.5 for a detailed proof. □

Since the elimination phase is activated for $t < T$, we have $\widehat{\mathcal{E}}(f^t, \pi^t, h^t, g^*) \geq \zeta_{\text{act}}$. Therefore, we have

$$\mathcal{E}(f^t, \pi^t, h, g^*) \geq \zeta_{\text{act}} - \epsilon/(6H) \geq \epsilon/H$$

for $t \leq \min\{T-1, dH+1\}$ when we condition on the event in Lemma D.1. Thus, we conclude the proof of the first part of Lemma C.1. We also have the following lemma, which shows that in the elimination phase, $\widehat{\mathcal{E}}$ in Line 12 of Algorithm 1 is a good estimator of $\mathcal{E}$ in (2).

Lemma D.2. *[Concentration in the Elimination Phase] With probability* $1 - \delta/8$, *we have*

$$|\widehat{\mathcal{E}}(f, \pi^t, h) - \mathcal{E}(f, \pi^t, h)| \leq \epsilon/(4H\sqrt{d})$$

holds for all $(f, t, h) \in \mathcal{F} \times [dH+1] \times [H]$. *Here* $\widehat{\mathcal{E}}$ *is defined in Line 12 of Algorithm 1.*

Proof. See Appendix §D.6 for a detailed proof. □

By the definition of $\mathcal{B}^t$ in Line 13 of Algorithm 1, we have $\widehat{\mathcal{E}}(f, \pi^t, h^t) \leq \zeta_{\text{elim}}$ for $f \in \mathcal{B}^t$. Therefore, we have

$$\mathcal{E}(f^t, \pi^t, h) \leq \zeta_{\text{elim}} + \epsilon/(4H\sqrt{d}) \leq \epsilon/(H\sqrt{d})$$

for $t \leq \min\{T-1, dH+1\}$ when we condition on the event in Lemma D.1. Thus, we conclude the proof of the second part of Lemma C.1. □

D.2. Proof of Lemma C.2

Proof. We prove Lemma C.2 by contradiction. For the sake of contradiction, We assume that $T > dH+1$. When $T > dH+1$, there exists $h \in [H]$ and $t_1 < \cdots < t_{d+1} \leq dH+1$, such that the elimination phase is activated in the t_i-th iteration at the $h-$th stage for $i \in [d+1]$. When we condition on $\mathcal{E}_1$ in Lemma C.1, we have $\mathcal{E}(f^{t_i}, \pi^{t_i}, h, g^*) > \epsilon/H$ for $i \in [d+1]$. Since $f^{t_i} \in \mathcal{B}^{t_j}$ when $i > j$, we have $\mathcal{E}(f^{t_i}, \pi^{t_j}, h) \leq \epsilon/(H\sqrt{d})$ when $i > j$. Therefore, we have

$$\sqrt{\sum_{j=1}^{i-1} \big(\mathcal{E}(f^{t_i}, \pi^{t_j}, h)\big)^2} < \sqrt{d} \times \epsilon/(H\sqrt{d}) = \epsilon/H.$$

Therefore, the roll-in distribution of $\pi^{t_1}, \ldots, \pi^{t_{d+1}}$ at step h is a goal-conditioned ϵ/H-independent sequence of length $d+1$, which contradicts with the definition of GOAL-BE dimension. Therefore, we have $T \leq dH+1$ when we condition on $\mathcal{E}_1$, which conclude the proof of Lemma C.2. □

D.3. Proof of Lemma C.3

Proof. Since the elimination phase is not activated in the T-th iteration, we have $\widehat{\mathcal{E}}(f^T, \pi^T, h, g^*) \leq \zeta_{\text{act}}$. Therefore, when we condition on the event in Lemma D.1, we have

$$\mathcal{E}(f^T, \pi^T, h, g^*) \leq \zeta_{\text{act}} + \epsilon/(6H) \leq 4\epsilon/H.$$

Therefore, we have $\sum_{h=1}^{H} \mathcal{E}(f^T, \pi^T, h, g^*) \leq 4\epsilon$, which concludes the proof of Lemma C.3. □

D.4. Proof of Lemma C.4

Proof. By Bellman equation, we have $\mathcal{E}(Q^*, \pi, h) = 0$ for all policy π and $h \in [H]$. By Lemma D.2, we have $\widehat{\mathcal{E}}(Q^*, \pi^t, h) \leq \epsilon/(4H\sqrt{d}) = \zeta_{\text{elim}}$ holds for all $(t, h) \in [dH+1] \times [H]$. By Line 13 of Algorithm 1, we have $Q^* \in \mathcal{B}^t$ when $t \leq dH+1$. Thus, we conclude the proof of Lemma C.4. □

D.5. Proof of Lemma D.1

Proof. Consider a fixed $(t, h) \in [dH + 1] \times [H]$ pair. Since

$$\left| f_h(x, u, g^*) - r_h(x, u, g^*) - \max_{u' \in \mathcal{U}} f_{h+1}(x', u', g^*) \right| \leq 3,$$

we have

$$|\widehat{\mathcal{E}}(f^t, \pi^t, h, g^*) - \mathcal{E}(f^t, \pi^t, h, g^*)| \leq \frac{6\sqrt{2}}{\sqrt{n_{\text{act}}}} \log\bigl(8H(dH^2 + 1)/\delta\bigr)$$

holds with probability at least $1 - \delta/(8H(dH + 1))$ by Azuma-Hoefdding's inequality. Here $\widehat{\mathcal{E}}$ is defined in Line 6 of Algorithm 1. By taking a union bound, we have

$$|\widehat{\mathcal{E}}(f^t, \pi^t, h, g^*) - \mathcal{E}(f^t, \pi^t, h, g^*)| \leq \frac{6\sqrt{2}}{\sqrt{n_{\text{act}}}} \log\bigl(8H(dH^2 + 1)/\delta\bigr)$$

holds for all $(t, h) \in [dH + 1] \times [H]$ with probability at least $1 - \delta/8$. Since we set $n_{\text{act}} = 2592H^2\iota/\epsilon^2$, we have

$$|\widehat{\mathcal{E}}(f^t, \pi^t, h, g^*) - \mathcal{E}(f^t, \pi^t, h, g^*)| \leq \epsilon/(6H).$$

Thus, we conclude the proof of Lemma D.1. □

D.6. Proof of Lemma D.2

Proof. Let $\mathcal{G}(\zeta_{\text{goal}})$ be an ζ_{goal}-cover of $\mathcal{G}$ with cardinality $\mathcal{N}(\mathcal{G}, \zeta_{\text{goal}})$, and let $\mathcal{Z}$ be an $\zeta_{\text{elim}}/16$-cover of $\mathcal{F}$ with cardinality $\mathcal{N}(\mathcal{F}, \zeta_{\text{elim}}/8)$. We define

$$\begin{aligned}
&\widehat{\mathcal{E}}\bigl(\widehat{f}, \pi^t, h_t, \mathcal{G}(\zeta_{\text{goal}})\bigr) \\
&\quad = \frac{1}{|\mathcal{D}_h^t|} \left| \max_{g \in \mathcal{G}(\zeta_{\text{goal}})} \sum_{(s,a,s') \in \mathcal{D}_h^t} \left(f_h(s, a, g) - r_h(s, a, g) - \max_{a' \in \mathcal{A}} f_{h+1}(s', a', g) \right) \right|,
\end{aligned}$$

where $\mathcal{D}_h^t$ is the dataset in the elimination phase. By applying Azuma-Hoeffding's inequality to all $(t, \widehat{f}, g) \in [dH + 1] \times \mathcal{Z} \times \mathcal{G}(\epsilon)$ and taking a union bound, we have

$$\begin{aligned}
|\widehat{\mathcal{E}}(\widehat{f}, \pi^t, h_t, g) - \mathcal{E}(\widehat{f}, \pi^t, h_t, g)| &\leq \frac{6\sqrt{2}}{\sqrt{n_{\text{elim}}}} \log\bigl(8H(dH^2 + 1)\mathcal{N}(\mathcal{G}, \zeta_{\text{goal}})\mathcal{N}(\mathcal{F}, \zeta_{\text{elim}}/8)/\delta\bigr) \\
&\leq \epsilon/(9H\sqrt{d})
\end{aligned} \tag{13}$$

holds for all $(t, \widehat{f}, g) \in [dH + 1] \times \mathcal{Z} \times \mathcal{G}(\epsilon)$ with probability at least $1 - \delta/8$.

By the definition of the goal covering, we have

$$\begin{aligned}
\left|\widehat{\mathcal{E}}(f, \pi^t, h_t) - \mathcal{E}(f, \pi^t, h_t)\right| &\leq 6\zeta_{\text{goal}} + \left|\widehat{\mathcal{E}}(f, \pi^t, h_t, \mathcal{G}(\zeta_{\text{goal}})) - \mathcal{E}(f, \pi^t, h_t, \mathcal{G}(\zeta_{\text{goal}}))\right| \\
&\leq \epsilon/(100H) + \max_{g \in \mathcal{G}(\zeta_{\text{goal}})} \left|\widehat{\mathcal{E}}(f, \pi^t, h_t, g) - \mathcal{E}(f, \pi^t, h_t, g)\right|
\end{aligned} \tag{14}$$

holds for all $f \in \mathcal{F}$. Combining Equations (13) and (14), we have

$$\left|\widehat{\mathcal{E}}(\widehat{f}, \pi^t, h_t) - \mathcal{E}(\widehat{f}, \pi^t, h_t)\right| \leq 6\zeta_{\text{goal}} + \epsilon/(9H\sqrt{d}) \leq \epsilon/(8H\sqrt{d}) \tag{15}$$

holds for all $(t, \widehat{f}) \in [dH+1] \times \mathcal{Z}$ with probability at least $1-\delta/8$. We denote the event in Equation (15) as $\mathcal{E}_4$. By the definition of $\mathcal{E}$ in (2), we have

$$\begin{aligned}\left|\mathcal{E}(\widehat{f}, \pi, h) - \mathcal{E}(f, \pi, h)\right| &= \left|\max_{g\in\mathcal{G}} \mathcal{E}(f, \pi, h, g) - \max_{g\in\mathcal{G}} \mathcal{E}(\widehat{f}, \pi, h, g)\right| \\ &\le \max_{g\in\mathcal{G}} \left|\mathcal{E}(f, \pi, h, g) - \mathcal{E}(\widehat{f}, \pi, h, g)\right| \le 2\left\|f - \widehat{f}\right\|_\infty .\end{aligned}$$

Similarly, we have $\left|\widehat{\mathcal{E}}(\widehat{f}, \pi, h) - \widehat{\mathcal{E}}(f, \pi, h)\right| \le 2\left\|f - \widehat{f}\right\|_\infty$. For any $f \in \mathcal{F}$, we select $\widehat{f} \in \mathcal{Z}$ with $\left\|f - \widehat{f}\right\|_\infty \le \zeta_{\mathrm{elim}}/8$. When condition on Event $\mathcal{E}_4$ in Equation (15), we have

$$\begin{aligned}\left|\widehat{\mathcal{E}}(f, \pi^t, h_t) - \mathcal{E}(f, \pi^t, h_t)\right| &\le \left|\widehat{\mathcal{E}}(\widehat{f}, \pi^t, h_t) - \mathcal{E}(\widehat{f}, \pi^t, h_t)\right| + \left|\mathcal{E}(\widehat{f}, \pi^t, h_t) - \mathcal{E}(f, \pi^t, h_t)\right| \\ &\quad + \left|\widehat{\mathcal{E}}(\widehat{f}, \pi^t, h_t) - \widehat{\mathcal{E}}(f, \pi^t, h_t)\right| \\ &\le \epsilon/(8H\sqrt{d}) + \zeta_{\mathrm{elim}}/2 \le \epsilon/(4H\sqrt{d})\end{aligned}$$

holds for all $(t, f) \in [dH+1] \times \mathcal{F}$ with probability at least $1-\delta/8$. Thus, we complete the proof of Lemma D.2. □