Reinforcement Learning with Logarithmic Regret and Policy Switches

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we study the problem of regret minimization for episodic Reinforcement Learning (RL) both in the model-free and the model-based setting. We focus on learning with general function classes and general model classes, and we derive results that scale with the *eluder dimension* of these classes. In contrast to the existing body of work that mainly establishes instance-independent regret guarantees, we focus on the instance-dependent setting and show that the regret scales logarithmically with the horizon T, provided that there is a gap between the best and the second best action in every state. In addition, we show that such a logarithmic regret bound is realizable by algorithms with $O(\log T)$ switching cost (also known as *adaptivity complexity*). In other words, these algorithms rarely switch their policy during the course of their execution. Finally, we complement our results with lower bounds which show that even in the tabular setting, we cannot hope for regret guarantees lower than $O(\log T)$.

1 Introduction

The main goal of Reinforcement Learning (RL) is the design and analysis of algorithms for automated decision making in complex and unknown environments. The environment is modeled as a *state space* and the available decisions are modeled as an *action space*. In recent years, RL has seen tremendous success in practical applications including, but not limited to, games and robotics [MKS+15, SHM+16, DCH+16, SSS+17, VBC+19]. Despite this success, a theoretical understanding of the algorithms that are deployed in these settings remains elusive. Traditionally, theoretical RL approaches have focused on the *tabular* setting where the complexity of the algorithms depends on the cardinality of the aforementioned spaces [SB18]. Thus, they are not suitable for applications where the state-action space is very large.

A different approach that has gained a lot of attention recently is the *function approximation* regime where the cumulative reward of the algorithm is modeled through a function, such as linear functions over some feature space. The advantage of this approach is that the algorithm can perform its search over a lower-dimensional space. There is a long line of work that provides regret guarantees for RL in the function approximation setting [OVR14, OVRW16, YW20, JYWJ20, AJS+20, CYJW20, KKL+20, ZBB+20, HZG21, KSWY21, ZHG21, ICN+21].

Most of these works have focused on establishing worst-case $\sqrt{T}$-regret guarantees, where T is the number of interactions with the environment. The caveat with these guarantees is that they are pessimistic since they neglect benign settings where even an *exponential* improvement over these bounds is achievable. To address this issue, there are some works that obtain instance-dependent regret bounds for RL in the tabular setting and in the linear function approximation setting [SJ19, YW20, HZG21]. However, getting logarithmic regret bounds in the context of *general* function approximation remains open. Hence, a natural question is the following:

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Can we establish instance-dependent logarithmic regret bounds with general function approximation?

In this paper, we provide an affirmative answer to this question. Following the assumptions in the existing literature, we model the RL problem as an MDP that enjoys the property that the optimal *policies* are at least $\text{gap}_{\min}$ better than the rest, where $\text{gap}_{\min} > 0$ is a parameter that captures the hardness of the underlying problem. We focus both on the *model-free* and the *model-based* settings with general function approximation, where we represent the value function or the transition model by a given function class, respectively.

In the model-free setting, we study the algorithm proposed in [KSWY21], which is a variant of the least-squares value iteration (LSVI) with upper confidence bound (UCB) bonuses that guide exploration. Here, the bonus functions are given by the width of a data-dependent confidence region for LSVI. For the model-based setting, we develop a similar algorithm which combines value-targeted regression with UCB bonuses. On top of the logarithmic regret guarantees they enjoy, our algorithms feature lazy policy updates, in the sense that the policy is updated rarely and only when certain conditions are met. For both settings, we establish $O(\text{poly}(\log T) \cdot \text{poly}(H) \cdot \text{poly}(d_{\mathcal{F}}) \cdot 1/\text{gap}_{\min})$ regret guarantees, where T is the number of interactions with the environment, H is the *planning horizon*, $d_{\mathcal{F}}$ is a term that captures the complexity of the function class $\mathcal{F}$ which is used to approximate either the value function or the transition model. In particular, $d_{\mathcal{F}}$ involves both the *eluder dimension* [RVR13] and the log-covering numbers of the function classes. That is, for benign MDPs these RL algorithms achieve logarithmic regret, which is exponentially better than the worst-case $O(\sqrt{T})$-regret. Moreover, we extend the result of [KSWY21] regarding the *adaptivity complexity* of the algorithm, i.e. the number of different policies it uses, to the model-based and show that it is logarithmic in T. We remark that in many real-world applications, like medical trials and personalized recommendation, there is a high cost and overhead associated with changing the underlying policy. Thus, coming up with algorithms that have low adaptivity complexity is an important task. To the best of our knowledge, this is the first work that establishes a logarithmic instance-dependent regret guarantee for RL with general function approximation.

1.1 Related Work

Logarithmic regret bounds for bandits. There is a long line of work that establishes logarithmic regret guarantees in bandit problems. Essentially, bandits are a special case of RL where the transition to the next state does not depend on the action that was taken by the agent. An extensive list of such algorithms can be found in [BCB12, Sli19, LS20].

Logarithmic regret bounds for RL. A series of works are devoted to proving instance-dependent logarithmic regret bounds in tabular RL. [OPT18, SJ19] prove lower bounds that show that a logarithmic dependence on T is unavoidable. Considering the upper bounds, [AO07, TB07] establish logarithmic regret guarantees in the average reward setting. Both of these guarantees are asymptotic as they require the number of interactions T with the MDP to be large enough. Regarding non-asymptotic bounds, [JOA10] provide such an algorithm that achieves $O(D^2|\mathcal{S}|^2|\mathcal{A}|\log(T)/\text{gap}_{\min})$ regret for the average-reward MDP, where D is the diameter of the MDP. For episodic MDPs, logarithmic regret upper bounds are established in [SJ19, YYD21]. The work that is probably the most closely related to ours is [HZG21]. It provides instance-dependent logarithmic regret guarantees both in the model-free and model-based setting with *linear* function approximation. Moreover, the proposed algorithms update the policy in every episode. Our work generalizes these results since the linear regime is a special case of the setting we are studying. Furthermore, we achieve an exponential improvement on the adaptivity complexity over their algorithms.

Bandits with limited adaptivity complexity. There is a lot of interest in obtaining bandit algorithms that update their policies rarely [AYPS11, PRCS16, AAAK17, GHRZ19, DLZZ20, RYZ21]. Notably, [DLZZ20] study rare policy switching constraints for a broader class of online learning and decision making problems such as logit bandits.

RL with limited adaptivity complexity. Recently, there has been a lot of interest in developing RL algorithms that achieve sub-linear regret and have low adaptivity complexity [BXJW19, ZZJ20, CK20, WZG21, KSWY21, GXDY21]. We develop an algorithm with low adaptivity complexity that works in the model-based, general function approximation setting.

RL with general function approximation. As we have alluded to already, it is important to develop and analyze algorithms in the function approximation regime. So far, the most commonly studied setting is RL with linear function approximation [YW20, JYWJ20, DLMW20, WSY20, ZLKB20, AKKS20]. Recently, there are also important results in RL with general function approximation. To

be specific, [JKA+17] design an efficient algorithm whose sample complexity is bounded in terms of the Bellman rank of the function class. [AJS+20] develop an algorithm for model-based RL, whose regret bound depends on the eluder dimension of the underlying class of models. [JLM21] propose an algorithm for problems where the underlying class has bounded Bellman eluder dimension. Recent works such as [WSY20, KSWY21, FRSLX20] develop algorithms in the model-free setting whose regret scale with the eluder dimension of the functions.

2 Preliminaries

2.1 Notation

We use the notation $[N] = \{1, 2, \ldots, N\}$. We also define the infinity norm of some function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is some domain, to be $||f||_\infty = \sup_{x \in \mathcal{X}} |f(x)|$. For a dataset $\mathcal{D} = \{(x_i, q_i)\}_{i=1}^n \subseteq \mathcal{X} \times \mathbb{R}$ and a function $f : \mathcal{X} \to \mathbb{R}$, we define the following norm $||f||_{\mathcal{D}} = \left(\sum_{i=1}^n (f(x_i) - q_i)^2\right)^{1/2}$. Given a set $\mathcal{Z} = \{x_i\}_{i=1}^n \subseteq \mathcal{X}$ we let $||f||_{\mathcal{Z}} = \left(\sum_{i=1}^n f(x_i)^2\right)^{1/2}$ be the data-dependent norm. Given a measurable set $\mathcal{X}$, we denote with $\Delta(\mathcal{X})$ the probability simplex over $\mathcal{X}$. We also denote by $\mathbb{1}[\mathcal{E}]$ the indicator function of the event $\mathcal{E}$ and by $\mathrm{poly}(x)$ a function that is a polynomial in x.

2.2 Episodic Markov Decision Processes

The learning agent interacts with the environment over a sequence of K rounds which we call *episodes*. We model the interaction of the agent with the environment in every episode as a *Markov Decision Process* (MDP). We denote an MDP by $M = (\mathcal{S}, \mathcal{A}, P, r, H, s_1)$, where $\mathcal{S}$ is the *state space*, $\mathcal{A}$ is the *action space*, $P = \{P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\}_{h=1}^H$ are the *transition kernels*, $r = \{r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]\}_{h=1}^H$ are the *reward functions* which we assume to be deterministic, H is the *planning horizon*, which is the length of every episode, and s_1 is the initial state of every episode. We let $T = K \cdot H$ be the total number of interactions with the MDP. During every episode, the agent uses a *policy* $\pi = \{\pi_h : \mathcal{S} \to \mathcal{A}\}_{h=1}^H$, to take an action at a given state. We use the Q-function and V-function to evaluate the expected total reward generated by a policy π. More specifically, we define

$$Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) \big| s_h = s, a_h = a, \pi\right], V_h^\pi(s) \quad = \mathbb{E}\left[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) \big| s_h = s, \pi\right],$$

where the actions are picked according to π and $s_{h'+1} \sim P_{h'}(\cdot | s_{h'}, a_{h'})$. For simplicity, we denote $\langle P_h(\cdot|s, a), V\rangle = \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V(s')]$. We denote the optimal policy for a given MDP with π^*. For the optimal Q-function, V-function we use $Q_h^*(s, a) = Q_h^{\pi^*}(s, a), V_h^*(s) = V_h^{\pi^*}(s)$, respectively.

The goal of the learner is to improve its performance as it interacts with the unknown environment. In the episodic setting, the agent commits to a policy at the beginning of every episode. We let π^k denote the policy that the agent uses in the k-th episode. At each step $h \in [H]$, the agent observes the state s_h^k, chooses an action according to π^k, and then observes the reward $r_h(s_h^k, a_h^k)$ and the next state $s_{h+1}^k \sim P_h(\cdot|s_h^k, a_h^k)$. In this work, to measure the performance of the agent we use the notion of *regret*, defined as

$$\mathrm{Regret}(K) = \sum_{k=1}^K \left(V_1^*(s_1) - V_1^{\pi^k}(s_1)\right).$$

The regret measures the difference between the total reward that the agent would have accumulated if she was following the optimal policy and the reward she actually accumulates. We strive for algorithms that guarantee sub-linear regret, because, as $K \to \infty$, the reward of the algorithm approaches that of the the optimal policy.

An assumption we make in order to achieve logarithmic regret guarantees is that the minimum *sub-optimality* gap is positive.

Definition 2.1. We define the sub-optimality gap of a state-action pair (s, a) at step h to be

$$\mathrm{gap}_h(s, a) = V_h^*(s) - Q_h^*(s, a).$$

The minimum sub-optimality gap is defined to be

$$\text{gap}_{\min} = \inf_{h,s,a} \{\text{gap}_h(s,a) : \text{gap}_h(s,a) \neq 0\}.$$

It is well-known that if we do not make any assumptions the best regret guarantee we can hope for is $O(\sqrt{T})$ [JOA10]. In this work, we derive instance-dependent regret guarantees that achieve an exponential improvement on T when $\text{gap}_{\min} > 0$. We remark that this assumption is *not* restrictive and it rules out some pathological cases. In particular, it allows us to handle MDPs with multiple optimal policies. For a detailed discussion, we kindly refer the reader to Appendix A.1.

2.3 Model-Free Assumption

In this paper, we deal with general function classes. In the model-free setting we assume that we have access to a function class $\mathcal{F} \subseteq \{f : \mathcal{S} \times \mathcal{A} \to [0, H+1]\}$. Our goal is to use the functions in $\mathcal{F}$ to approximate the optimal Q-function. In order to derive meaningful results we assume that this class has some structure. We follow the same assumption as in [WSY20, KSWY21].

Assumption 2.2 (Bellman Operator Assumption)**.** For any $h \in [H]$ and $V : \mathcal{S} \to [0, H]$ there exists some $f_V \in \mathcal{F}$ such that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ we have

$$f_V(s,a) = r_h(s,a) + \sum_{s' \in \mathcal{S}} P_h(s'|s,a)V(s').$$

The intuition behind this assumption is that if we apply the one-step Bellman backup operator to some value function V, i.e. $r_h(s,a) + \sum_{s' \in \mathcal{S}} P_h(s'|s,a)V(s')$, the result will remain in the function class. Thus, it implicitly poses some constraints both on the transition probabilities and the reward function. It is known that both the tabular setting and the linear MDP setting satisfy this assumption [YW19, JYWJ20]. Another assumption we make is that the function class and the state-action space have bounded covering numbers. We will show that the dependence of the regret on the covering number is poly-logarithmic. This assumption has also appeared in other works [RVR13, WSY20, JLM21, KSWY21].

Assumption 2.3 (Bounded Covering Number)**.** We say that $\mathcal{N}(\mathcal{F}, \varepsilon)$ is a bound on the ε-covering number of $\mathcal{F}$, if for any $\varepsilon > 0$ there is an ε-cover $\mathcal{C}(\mathcal{F}, \varepsilon) \subseteq \mathcal{F}$ with size $|\mathcal{C}(\mathcal{F}, \varepsilon)| \leq \mathcal{N}(\mathcal{F}, \varepsilon)$, so that for all $f \in \mathcal{F}$ there is some $f' \in \mathcal{C}(\mathcal{F}, \varepsilon)$ such that $||f - f'||_\infty \leq \varepsilon$. Similarly, we say that $\mathcal{N}(\mathcal{S} \times \mathcal{A}, \varepsilon)$ is a bound on the ε-covering number of $\mathcal{S} \times \mathcal{A}$ with respect to $\mathcal{F}$, if for any $\varepsilon > 0$ there is an ε-cover $\mathcal{C}(\mathcal{S} \times \mathcal{A}, \varepsilon) \subseteq \mathcal{S} \times \mathcal{A}$ with size $|\mathcal{C}(\mathcal{S} \times \mathcal{A}, \varepsilon)| \leq \mathcal{N}(\mathcal{S} \times \mathcal{A}, \varepsilon)$, so that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ there is some $(s', a') \in \mathcal{C}(\mathcal{S} \times \mathcal{A}, \varepsilon)$ such that $\sup_{f \in \mathcal{F}} |f(s,a) - f(s',a')| \leq \varepsilon$.

The intuition behind this assumption is straightforward: even if the function class or the state-action space are infinite, we can approximate them using a small number of points.

2.4 Model-Based Assumption

The assumptions in Section 2.3 are *model-free* since they impose some structure on the function class that approximates the Q-function instead of the transition kernel. In order to derive our results, we can also follow a different route and impose some structure directly on the transition kernel [AJS+20].

Assumption 2.4 (Known Transition Model Family)**.** For all $h \in [H]$, the transition model P_h belongs to a family of models $\mathcal{P}_h$ which is known to the learner. The elements of $\mathcal{P}_h$ are transition kernels that map state-action pairs to signed distributions over the state space $\mathcal{S}$.

We allow signed distributions in our model class to increase its generality. For example, this is useful when we are given access to a model class that can be compactly represented only when it includes non-probability kernels. For an extensive discussion about this, the reader is referred to [PS16].

Transition kernels have been used to model complex stochastic controlled systems. For instance, transitions in robotics systems are often modelled using parameters of the environment, such as friction. An important class that satisfies this assumption are the *linear mixture models*.

Definition 2.5. The class of models $\mathcal{P}$ with feature mapping $\phi(s'|s,a) : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and some $\theta^* \in \mathbb{R}^d$ whose euclidean norm is bounded, is called linear mixture model if:

- $P(s'|s,a) = \langle \phi(\cdot|s,a), \theta^* \rangle$.
- For any bounded function $V : \mathcal{S} \to [0,H]$ and any pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have $||\phi_V(s,a)||_2 \leq \sqrt{H}$, where $\phi_V(s,a) = \langle \phi(\cdot|s,a), V \rangle$.

One way to interpret the linear mixture model is as an aggregation of some basis models which are known to the designer [MJTS20]. Another interesting way to think about it comes from large-scale queuing networks where both the arrival rate of jobs and the processing speed for the queues are unknown. If we approximate this system in discrete time, then the transition matrix from timestep t to timestep $t + \Delta t$ approaches that of a linear function with respect to the arrival rate and the processing time [GK89]. Another interesting setting that satisfies this assumption is the linear-factored MDP [YW20].

2.5 Complexity Measure: Eluder Dimension

Our results depend on the complexity of the function classes and the model classes that we consider. To measure this complexity, we use the *eluder dimension* of these classes [RVR13].

Definition 2.6. Fix some $\varepsilon \geq 0$ and a sequence of n points $\mathcal{Z} = \{(x_i)\}_{i \in [n]} \subseteq \mathcal{X}$. Then:

1. A point $x \in \mathcal{X}$ is ε-dependent on $\mathcal{Z}$ with respect to $\mathcal{F}$ if for all $f, f' \in \mathcal{F}$ such that $||f - f'||_{\mathcal{Z}} \leq \varepsilon$ it holds that $|f(x) - f'(x)| \leq \varepsilon$.
2. A point x is ε-independent of $\mathcal{Z}$ with respect to $\mathcal{F}$ if x is not ε-dependent on $\mathcal{Z}$.
3. The ε-eluder dimension of $\mathcal{F}$, which is denoted by $\dim_E(\mathcal{F}, \varepsilon)$, is the length of the longest sequence of elements in $\mathcal{X}$ such that every element in this sequence is ε'-independent of its predecessors, for some $\varepsilon' \geq \varepsilon$.

Intuitively, the eluder dimension of $\mathcal{F}$ quantifies the largest set of elements $\mathcal{Z} \subseteq \mathcal{X}$ so that if all $f \in \mathcal{F}$ are close with respect to $\mathcal{Z}$, then they are close on all elements of $\mathcal{X}$.

It is known that when $\mathcal{X} = \mathcal{S} \times \mathcal{A}, f : \mathcal{S} \times \mathcal{A} \to [0,H]$, and $\mathcal{S}, \mathcal{A}$ are finite, we have that $\dim_E(\mathcal{F}, \varepsilon) \leq |\mathcal{S}| \cdot |\mathcal{A}|$, for all $\varepsilon > 0$ [RVR13, WSY20]. Moreover, when $\mathcal{F}$ is the class of linear functions, i.e., $f_\theta(s,a) = \theta^T \phi(s,a)$, for a given feauture vector $\phi(s,a)$, the eluder dimension of $\mathcal{F}$ is bounded by $\dim_E(\mathcal{F}, \varepsilon) = O(d \log(1/\varepsilon))$, for all $\varepsilon > 0$. We remark that classes which go beyond linear functions have bounded eluder dimension. Some examples are quadratic functions, generalized linear functions (e.g. sigmoids) which have the form $g(\langle \phi(s), x \rangle)$, where g is increasing and Lipshitz-continuous with non-zero derivative, all finite function classes, sparse linear combinations of features, linear thresholds with Gaussian i.i.d. features, etc. [RVR13, OVR14, LKFS21]. It is an ongoing line of work to understand what is the most general class of functions that has bounded eluder dimension. We refer to the setting we are working on as *general function approximation* to align with the previous works that study function classes with bounded eluder dimension.

2.6 Switching Cost

Essentially, the *switching cost* or the *adaptivity complexity* measures the number of episodes the algorithm has to update its policy in order to achieve the guaranteed regret bound [BXJW19, KSWY21].

Definition 2.7. We define the switching cost of an algorithm A over K episodes to be

$$N_{\text{switch}} = \sum_{k=1}^{K-1} \mathbb{1}[\pi_k \neq \pi_{k+1}].$$

3 Overview of the Algorithms and Main Results

In this section, we present our main results and give a high-level description of the techniques we use. We treat both the model-free and the model-based setting in a unified way. The algorithm for the model-free setting comes directly from [KSWY21]. We extend it appropriately in order to handle the model-based setting. The main algorithm is presented in Algorithm 1. The only differences between the two settings are the different sampling routine and Q-function estimator used by Algorithm 1. In a nutshell, the low-switching cost algorithm works as follows:

- After each round of the interaction with the MDP, Algorithm 1 adds elements to the current dataset with some probability that depends on their significance and updates the policy only if the dataset has changed. This guarantees that the adaptivity of the algorithm depends logarithmically on T, without hurting the regret guarantee. The sampling routines are postponed to Appendix A.2.
- There is a least-squares estimate of the Q-function (transition kernel) in the model-free (model-based) setting.
- A bonus is added to this estimate which encourages exploration and, with high probability, guarantees that the current estimate of the Q-function serves as an element-wise upper bound of Q^*. This bonus is based on a sub-sampled dataset that has been accumulated from previous interactions with the MDP.

Before we delve deeper into the two settings separately, we describe a parameter that is crucial for both of the algorithms we are using. Following [KSWY21], the *sensitivity* of an element z with respect to a dataset $\mathcal{Z}$ and a function class $\mathcal{F}$ is defined to be

$$\text{sensitivity}_{\mathcal{Z},\mathcal{F}}(z) = \min\left\{\sup_{f_1,f_2\in\mathcal{F}} \frac{(f_1(z)-f_2(z))^2}{\min\{||f_1-f_2||^2_{\mathcal{Z}}, T(H+1)^2\}+\beta}, 1\right\}.$$

Intuitively, this parameter captures the importance of the current element z relative to the dataset we are working with. We will elaborate on the choice of the parameter β for each of the two settings separately. To establish the regret guarantee, we propose a novel regret decomposition for this algorithm where we utilize the "peeling technique" that has been applied in prior works in local Rademacher complexities [BBM05] and in RL [HZG21, YYD21]. By doing that, we show how the regret of the algorithm relates to the suboptimality gap.

To establish the lower bound on the regret of any algorithm in the settings we are interested in, we utilize a result that was proved in [OPT18, SJ19]. It states that for all algorithms that achieve sublinear regret, there exists a tabular MDP where its regret is at least $\Omega(\text{poly}(\log(T)) \cdot \text{poly}(H) \cdot 1/\text{gap}_{\min})$. We remark that tabular MDPs satisfy both the model-free assumption [WSY20] and the model-based assumption [AJS+20], thus the lower bounds follow immediately from this result.

3.1 Model-Free Setting

We first present the approach we use in the model-free setting, i.e., where we have access to some function class $\mathcal{F}$ and state-action space $\mathcal{S} \times \mathcal{A}$ that satisfy Assumption 2.2 and Assumption 2.3, respectively. The algorithm we use comes from [KSWY21]. The Q-function estimator and the sampling routine that we use for this setting are presented in Algorithm 2 and Algorithm 4, respectively. The dataset includes pairs of the form $z_h^k = (s_h^k, a_h^k)$. The Q-function routine is a least-squares estimator that is based on all the previous interactions with the MDP. Notice that the bonus function is based only on the sub-sampled dataset and depends on a hardcoded parameter β. This parameter is chosen in a way that ensures the Q-function is an optimistic estimate of the actual one and the bonus we add is not too large. The choice of β is given in [KSWY21].

A crucial part of the algorithm is the online sub-sampling routine. The reason we are using this routine is twofold. Firstly, if we use the entire dataset there will be a huge number of distinct elements in it, which can make the exploration bonus unstable since it changes constantly and can take infinitely many different values. In order to establish the optimism of the Q-function estimation, namely,

$$Q_h^*(s,a) \le Q_h^k(s,a) \le \langle P_h(\cdot|s,a), V_h^k\rangle + 2b_h^k(s,a),$$

[KSWY21] show that it is crucial to bound the complexity of the exploration bonus. Secondly, if we sub-sample the dataset based on the importance of the elements, we can achieve the regret guarantees that we are aiming for by switching the policy only when an important element has been added. Notice that whenever an element is added to the dataset, multiple copies are included. This is to make the sub-sampled dataset behave like an unbiased estimator of the orignal one. Then, using concentration bounds one can show that it approximates the original one with high probability [KSWY21]. The full description of this procedure is presented in Algorithm 4. For a more detailed discussion about the importance of sub-sampling the dataset, the interested reader is referred to [WSY20, KSWY21].

Algorithm 1 Low Switching Cost Value Iteration (with parameters δ, K)

Require: Failure probability $\delta \in (0,1)$, number of episodes K, and setting of operation

$\widetilde{k} \leftarrow 1$
$\widehat{Z}_h^1 \leftarrow \emptyset, \forall h \in [H]$
for $k \in [K]$ **do**
  for $h = H, H-1, \ldots, 1$ **do**
    if $k \geq 2$ **then**
      $\widehat{\mathcal{Z}}_h^k \leftarrow$ **Sample**$(\mathcal{F}, \widehat{\mathcal{Z}}_h^{k-1}, z_h^{k-1}, \delta)$
    end if
  end for
  if $k = 1$ or $\exists h \in [H] : \widehat{Z}_h^k \neq \widehat{Z}_h^{\widetilde{k}}$ **then**
    $\widetilde{k} \leftarrow k$
    $Q_{H+1}^k(\cdot,\cdot) \leftarrow 0, V_{H+1}^k(\cdot) \leftarrow 0$
    for $h = H, H-1, \ldots, 1$ **do**
      $\mathcal{T}_h^k \leftarrow$ history of execution
      $Q_h^k(\cdot,\cdot) \leftarrow$ **Q-Estimator**$(\mathcal{T}_h^k, \mathcal{Z}_h^k)$
      $V_h^k(\cdot) = \max_{a \in \mathcal{A}} Q_h^k(\cdot, a)$
      $\pi_h^k(\cdot) \leftarrow \arg\max_{a \in \mathcal{A}} Q_h^k(\cdot, a)$
    end for
  end if
  Receive initial state s_1 of episode k
  for $h \in [H]$ **do**
    Take action $a_h^k \leftarrow \pi_h^{\widetilde{k}}(s_h^k)$
  end for
end for

Algorithm 2 Q-function Model-Free Estimator

Require: Current sub-sampled dataset $\widehat{\mathcal{Z}}$, history of execution $\mathcal{T}$

$\mathcal{D}_h^k \leftarrow \{(s_h^\tau, a_h^\tau, r_h^\tau + V_{h+1}^k(s_{h+1}^\tau))\}_{\tau \in [k-1]}$
$\widehat{f} \leftarrow \arg\min_{f \in \mathcal{F}} ||f||_{\mathcal{D}}^2$
$\widehat{F}_h^k \leftarrow \{f_1, f_2 \in \mathcal{F} : \min\{\|f_1 - f_2\|_{\widehat{\mathcal{Z}}_h^k}^2, T(H+1)^2 \leq \beta\}$
$b_h^k(\cdot,\cdot) \leftarrow \sup_{f_1, f_2 \in \widehat{\mathcal{F}}_h^k} |f_1(\cdot,\cdot) - f_2(\cdot,\cdot)|$
Return $\min\{f_h^k(\cdot,\cdot) + b_h^k(\cdot,\cdot), H\}$

Algorithm 3 Q-function Model-Based Estimator

Require: Function class $\mathcal{F}$, current sub-sampled dataset $\widehat{\mathcal{Z}}$, current regression dataset $\mathcal{D}$

$\widehat{P}_h^k \leftarrow \arg\min_{P \in \mathcal{P}_h} \sum_{k'=1}^{k-1} \left(\langle P(\cdot|s_h^{k'}, a_h^{k'}), V_{h+1}^{k'}\rangle - V_{h+1}^{k'}(s_{h+1}^{k'})\right)^2$
$\mathcal{F}_h^k = \{f_1, f_2 : \min\{||f_1 - f_2||_{\widehat{\mathcal{Z}}_h^k}, T(H+1)^2\} \leq \beta\}$
$b_h^k(\cdot,\cdot) \leftarrow \sup_{f_1, f_2 \in \mathcal{F}_h^k} |f_1(\cdot,\cdot,V_{h+1}) - f_2(\cdot,\cdot,V_{h+1})|$
Return $\min\{r_h(\cdot,\cdot) + \langle \widehat{P}_h^k(\cdot|\cdot,\cdot), V_{h+1}^k\rangle + b_h^k(\cdot,\cdot), H\}$

We are now ready to state our main result in this setting.

Theorem 3.1. *There exists an absolute constant $C > 0$, and a proper parameter β for Algorithm 1 such that with probability of at least $1 - \lceil \log T \rceil e^{-\tau} - \delta$ the regret of the algorithm is bounded by*

$$\text{Regret}(K) \leq \frac{C d_{\mathcal{F}} H^5 \log^4 T}{\text{gap}_{\min}} + \frac{16 H^2 \tau}{3} + 2,$$

for any $\delta, \tau > 0$, where $d_{\mathcal{F}} = \dim_E^2(\mathcal{F}, 1/T) \cdot \log(\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \cdot \log(\mathcal{N}(\mathcal{S} \times \mathcal{A}, \delta/T^2)/\delta)$ is a parameter that captures the complexity of the function class. The value of the parameter β is

$$\beta = C d'_{\mathcal{F}} H^2 \log^4 T,$$

where $d'_{\mathcal{F}} = \log(\mathcal{N}(\mathcal{F}, \delta/T^3)/\delta) \dim_E(\mathcal{F}, 1/T) \log(\mathcal{N}(\mathcal{S} \times \mathcal{A}, \delta/T^3)/\delta))$. Moreover, the number of switching policies is bounded by

$$O\left(H \log(T\mathcal{N}(\mathcal{F}, \sqrt{\delta}/T^2)/\delta) \dim_E(\mathcal{F}, 1/T) \log^2 T\right).$$

3.2 Model-Based Setting

In this regime, we assume that the MDP satisfies Assumption 2.4. We also assume that the reward function is known to the learner similar to [AJS+20]. If the reward is unknown, we just estimate it and construct a confidence region.

Before we discuss the details of our approach, we need to describe an important set of functions that show up in our algorithm and in the regret guarantee. Let $\mathcal{V}$ be the set of all measurable functions that are bounded by H. Let $\mathcal{P}_h$ be the set of potential models in step H. We also let $f : \mathcal{S} \times \mathcal{A} \times \mathcal{V} \to \mathbb{R}$ and define the following set:

$$\mathcal{F}_h = \left\{ f : \exists \widetilde{P}_h \in \mathcal{P}_h \text{ so that } f(s,a,V) = \int_{\mathcal{S}} \widetilde{P}_h(s'|s,a) V(s') ds', \forall (s,a,V) \in \mathcal{S} \times \mathcal{A} \times \mathcal{V} \right\}.$$

The bounds we state scale with the complexity of $\mathcal{F}_h$. The Q-function estimator and the sampling routine we use that are specific to this setting are presented in Algorithm 3 and Algorithm 5, respectively. The dataset includes elements of the form $z_h^k = (s_h^k, a_h^k, V_{h+1}^k(\cdot))$. The Q-function routine works in the following way. We first estimate a model $\widehat{P}_h \in \mathcal{P}_h$ using a least-squares estimator that is based on all the previous interactions with the MDP. Using a concentration argument for this estimator of the model, similar to [RVR13, AJS+20], we can show that for an appropriate choice of β, the estimated model lies in a data-dependent ball centered at $\widehat{P}_h$, with high probability (see Lemma C.11 in the Appendix). Thus, we can set the bonus function to be the diameter of this ball in order to ensure that $\widehat{Q}_h$ is an optimistic estimate of Q_h^*. In addition, the choice of β ensures that the bonus we add is not very large. Notice also that since the concentration argument in this setting differs with that in the model-free setting, we do not need to round the elements that we are adding to the sub-sampled dataset. In this setting, the main reason we sub-sample the dataset is to achieve logarithmic adaptivity. To bound the adaptivity complexity, we use a similar approach as in [KSWY21].

We are now ready to state our main result in this setting.

Theorem 3.2. *There exists an absolute constant C and a proper value of the parameter β for Algorithm 1 such that with probability at least $1 - \lceil \log T \rceil e^{-\tau} - \delta$ the regret of the algorithm is bounded by*

$$\text{Regret}(K) \leq \frac{C d_{\mathcal{F}} H^5 \log T}{\text{gap}_{\min}} + \frac{16 H^2 \tau}{3} + 2,$$

where $d_{\mathcal{F}} = \log(\mathcal{N}(\mathcal{F}, 1/T)/\delta) \dim_E^2(\mathcal{F}, 1/T)$. The value of the parameter β is

$$\beta = 4H^2 \log(2\mathcal{N}(\mathcal{F}, 1/T)/\delta) + 4/H \left(C + \sqrt{H^2/4 \log(T/\delta)} \right),$$

where $\mathcal{N}(\mathcal{F}, 1/T) = \max_{h \in [H]} \mathcal{N}(\mathcal{F}_h, 1/T), \dim_E(\mathcal{F}, 1/T) = \max_{h \in [H]} \dim_E(\mathcal{F}_h, 1/T)$.

The number of switching policies is bounded by

$$O\left(H \log(T\mathcal{N}(F, \sqrt{\delta/(64T^3)})/\delta) \dim_E(\mathcal{F}, 1/T) \log^2 T \right).$$

We remark that the eluder dimension and the log-covering numbers, which show up in both of our bounds, depend on T. To the best of our knowledge, in all the known cases that the eluder dimension is bounded (e.g. linear functions, quadratic functions, generalized linear functions), the scaling is $\text{poly}(\log T)$ (see e.g. [RVR13]).

3.3 Technical Contributions

We briefly highlight the main technical contributions of this work. Regarding algorithmic contributions, we extend the algorithm from [KSWY21], which works in the model-free setting, to the model-based setting. Even though these two algorithms look similar, the approach to prove concentration, optimism, and correctness in the model-based setting is different (see, e.g., Lemma C.12). This perspective provides a unified treatment of both settings that can be beneficial and leads to new results. For instance, if we wish to get instance-independent regret guarantees using the standard regret decomposition, we can get $\sqrt{T}$-regret with $\text{poly}(\log T)$ adaptivity complexity in the model-based setting. This is an *exponential* improvement in the adaptivity complexity compared to [AJS+20]. Furthermore, the analysis to get the instance-dependent guarantees of both algorithms diverges significantly from [KSWY21] since we use a different regret decomposition. Logarithmic regret guarantees in the linear function approximation setting have been established in [HZG21]. However, the technical results that are needed to derive the guarantees in this setting are tailor-made to linear functions. Importantly, [HZG21] bound several quantities in terms of $\phi(s,a)^\top (\Lambda_h^k)^{-1} \phi(s,a)$ where $\phi(s,a)$ are the features of the state-action pair (s,a) and $\Lambda_h^k = \sum_{i=1}^{k-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$. Since we are working with general functions, we need to derive bounds that scale with the *eluder dimension* of these classes (see, e.g. Lemma C.13).

4 Proof Sketch of Main Results

In this section, we sketch the proof of our main results. Due to space limitation, we only discuss the model-free setting. The full proofs in the model-free, model-based setting can be found in Appendix B, Appendix C, respectively.

The first step in our analysis is the regret decomposition of the algorithm. Lemma B.1 shows that $\mathbb{E}\left[\text{Regret}(K)\right] = \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\text{gap}_h(s_h^k, a_h^k)\right]$. Thus, we see that to bound the regret it is enough to bound $\sum_{k=1}^{K}\text{gap}_h(s_h^k, a_h^k)$ for every $h \in [H]$. Towards this end, notice that $\text{gap}_h(s_h^k, a_h^k) = 0$ or $\text{gap}_h(s_h^k, a_h^k) \in [\text{gap}_{\min}, H]$. We apply the "peeling technique" that has also been used in local Rademacher complexities [BBM05] and in [HZG21, YYD21]. The idea is to split the interval $[0, H]$ into $\log(H/\text{gap}_{\min})$ intervals, where the i-th interval is $[2^{i-1}\text{gap}_{\min}, 2^i\text{gap}_{\min}]$. Hence, for every $\text{gap}_h(s_h^k, a_h^k)$ that falls in the i-th interval its contribution to the regret is at most $2^i\text{gap}_{\min}$. Thus, to bound the regret it suffices to bound the number of suboptimalities that fall into every interval. Notice that for some $\text{gap}_h(s_h^k, a_h^k)$ in this interval we have that

$$V_h^*(s_h^k) - Q_h^{\pi_k}(s_h^k, a_h^k) \geq \text{gap}_h(s_h^k, a_h^k) \geq 2^{i-1}\text{gap}_{\min},$$

so it suffices to bound the number of sub-optimalities $V_h^*(s_h^k) - Q_h^{\pi_k}(s_h^k, a_h^k)$ that fall into the i-th interval. Both for the model-free and the model-based setting, we can derive such a bound. Finally, notice that once we have bounded the number of suboptimilaties in every interval, it is not difficult to bound the total regret. Let $C_i = [2^{i-1}\text{gap}_{\min}, 2^i\text{gap}_{\min})$ and $N = \log(H/\text{gap}_{\min})$. Then, we know that $\text{Regret}(K)$ can be upper bounded by

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\text{gap}_h(s_h^k, a_h^k) = \sum_{i=1}^{N}\sum_{\text{gap}_h(s_h^k,a_h^k)\in C_i}\text{gap}_h(s_h^k, a_h^k) \leq \sum_{i=1}^{N}\sum_{k=1}^{K}2^i\,\mathbb{1}\left[\text{gap}_h(s_h^k, a_h^k) \in C_i\right]$$
$$\leq \sum_{i=1}^{N}2^i\sum_{k=1}^{K}\mathbb{1}\left[V_h^*(s_h^k) - Q_h^{\pi_k}(s_h^k, a_h^k) \geq 2^{i-1}\text{gap}_{\min}\right].$$

Hence, deriving the regret guarantee boils down to bounding

$$\sum_{k=1}^{K}\mathbb{1}\left[V_h^*(s_h^k) - Q_h^{\pi_k}(s_h^k, a_h^k) \geq 2^i\text{gap}_{\min}\right].$$

We provide such a bound in Lemma B.9, which depends polynomially on $\log T, 1/\text{gap}_{\min}$, and the complexity parameters of the function class. The outline of the proof is the following. We fix some episode $h \in [H]$ and let K' be the set of rounds where $V_h^*(s_h^k) - Q_h^{\pi_k}(s_h^k, a_h^k) \geq 2^i\text{gap}_{\min}$. To get a bound on $|K'|$, our approach is to lower bound and upper bound the quantity $\sum_{i=1}^{|K'|}Q_h^{k_i}(s_h^{k_i}, a_h^{k_i}) - Q_h^{\pi_{k_i}}(s_h^{\pi_{k_i}}, a_h^{\pi_{k_i}})$ by $f_1(|K'|), f_2(|K'|)$, respectively. Then, we use the fact that $f_1(|K'|) \leq f_2(|K'|)$ to establish our bound. For the lower bound, using the definition of K' we get $f_1(|K'|) = 2^i\text{gap}_{\min}|K'|$. For the upper bound, we leverage the fact that $Q_h^{k_i}(s_h^{k_i}, a_h^{k_i}) \leq \langle P_h(\cdot|s_h^{k_i}, a_h^{k_i}), V_{h+1}^{k_i}\rangle + 2b_h^{k_i}(s_h^{k_i}, a_h^{k_i})$ (cf. Lemma B.8) and obtain $f_2(|K'|) \leq \sum_{i=1}^{|K'|}\sum_{h'=h}^{H}\varepsilon_{h'}^{k_i} + \sum_{i=1}^{|K'|}\sum_{h'=h}^{H}b_{h'}^{k_i}(s_{h'}^{k_i}, a_{h'}^{k_i})$, where $\varepsilon_h^{k_i}$ forms a bounded martingale difference sequence. We bound each term on the RHS separately. For the first one, we use the Azuma-Hoeffding inequality which can be found in Lemma B.7. For the second term, we generalize the bound on the summation of the bonus functions over all the episodes from [KSWY21], and show that a similar bound holds for the summation of the bonus over *any* set of episodes K' (see Lemma B.6). Putting everything together, we get that $|K'| = O\left(1/(4^i\text{gap}_{\min}) \cdot H^4 \cdot \log^4 T \cdot \text{poly}(d_{\mathcal{F}})\right)$, where $d_{\mathcal{F}}$ is the complexity parameter of the class.

5 Conclusion and Societal Impact

In this paper, we consider episodic RL with general function approximation. We prove that there are algorithms with logarithmic adaptivity complexity both in the model-free and model-based settings that achieve logarithmic instance-dependent regret guarantees. This is theoretical work and does not have any negative societal implications.

References

[AAAK17] Arpit Agarwal, Shivani Agarwal, Sepehr Assadi, and Sanjeev Khanna. Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *Conference on Learning Theory*, pages 39–75. PMLR, 2017.

[AJS+20] Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

[AKKS20] Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *arXiv preprint arXiv:2006.10814*, 2020.

[AO07] Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in neural information processing systems*, pages 49–56, 2007.

[AYPS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

[BBM05] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[BCB12] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

[BXJW19] Yu Bai, Tengyang Xie, Nan Jiang, and Yu-Xiang Wang. Provably efficient q-learning with low switching cost. *arXiv preprint arXiv:1905.12849*, 2019.

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[CK20] Tongyi Cao and Akshay Krishnamurthy. Provably adaptive reinforcement learning in metric spaces. *Advances in Neural Information Processing Systems*, 33:9736–9744, 2020.

[CYJW20] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

[DCH+16] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, pages 1329–1338. PMLR, 2016.

[DLMW20] Simon S Du, Jason D Lee, Gaurav Mahajan, and Ruosong Wang. Agnostic q-learning with function approximation in deterministic systems: Near-optimal bounds on approximation error and sample complexity. *Advances in Neural Information Processing Systems*, 2020, 2020.

[DLZZ20] Kefan Dong, Yingkai Li, Qin Zhang, and Yuan Zhou. Multinomial logit bandit with low switching cost. In *International Conference on Machine Learning*, pages 2607–2615. PMLR, 2020.

[Fre75] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[FRSLX20] Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

[GHRZ19] Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *arXiv preprint arXiv:1904.01763*, 2019.

[GK89] Boris Vladimirovich Gnedenko and Igor Nikolaevich Kovalenko. *Introduction to queueing theory*. Birkhauser Boston Inc., 1989.

[GXDY21] Minbo Gao, Tianle Xie, Simon S Du, and Lin F Yang. A provably efficient algorithm for linear markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494*, 2021.

[HZG21] Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR, 2021.

[ICN+21] Haque Ishfaq, Qiwen Cui, Viet Nguyen, Alex Ayoub, Zhuoran Yang, Zhaoran Wang, Doina Precup, and Lin Yang. Randomized exploration in reinforcement learning with general value function approximation. In *International Conference on Machine Learning*, pages 4607–4616. PMLR, 2021.

[JKA+17] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

[JLM21] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815*, 2021.

[JOA10] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

[JYWJ20] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[KKL+20] Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *arXiv preprint arXiv:2006.12466*, 2020.

[KSWY21] Dingwen Kong, Ruslan Salakhutdinov, Ruosong Wang, and Lin F Yang. Online sub-sampling for reinforcement learning with general function approximation. *arXiv preprint arXiv:2106.07203*, 2021.

[LKFS21] Gene Li, Pritish Kamath, Dylan J Foster, and Nathan Srebro. Eluder dimension and generalized rank. *arXiv preprint arXiv:2104.06970*, 2021.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[MJTS20] Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

[MKS+15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[OPT18] Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. *arXiv preprint arXiv:1806.00775*, 2018.

[OVR14] Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. *arXiv preprint arXiv:1406.1853*, 2014.

[OVRW16] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386. PMLR, 2016.

[PRCS16] Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *The Annals of Statistics*, 44(2):660–681, 2016.

[PS16] Bernardo Ávila Pires and Csaba Szepesvári. Policy error bounds for model-based reinforcement learning with factored linear models. In *Conference on Learning Theory*, pages 121–151. PMLR, 2016.

[RVR13] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *NIPS*, pages 2256–2264. Citeseer, 2013.

[RYZ21] Yufei Ruan, Jiaqi Yang, and Yuan Zhou. Linear bandits with limited adaptivity and learning distributional optimal design. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 74–87, 2021.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[SHM+16] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[SJ19] Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32:1153–1162, 2019.

[Sli19] Aleksandrs Slivkins. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*, 2019.

[SSS+17] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[TB07] Ambuj Tewari and Peter L Bartlett. Optimistic linear programming gives logarithmic regret for irreducible mdps. In *NIPS*, pages 1505–1512. Citeseer, 2007.

[VBC+19] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[WSY20] Ruosong Wang, Ruslan Salakhutdinov, and Lin F Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *arXiv preprint arXiv:2005.10804*, 2020.

[WZG21] Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *arXiv preprint arXiv:2101.02195*, 2021.

[YW19] Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

[YW20] Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

[YYD21] Kunhe Yang, Lin Yang, and Simon Du. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pages 1576–1584. PMLR, 2021.

[ZBB+20] Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020.

[ZHG21] Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021.

[ZLKB20] Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

[ZZJ20] Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *arXiv preprint arXiv:2004.10019*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Yes, for example we have described which function classes have bounded eluder dimension.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] In Section 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We have described the assumptions under which our results are correct(Assumption 2.3, 2.2 , 2.4, positive minimum sub-optimality gap).
   (b) Did you include complete proofs of all theoretical results? [Yes] Due to space limitations, the full proofs are postponed to the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Omitted Details from Main Body

A.1 Minimum Suboptimality Gap Assumption

It is known that without any instance-specific assumptions we cannot get $o(\sqrt{T})$ regret. The sub-optimality gap assumption is standard in the bandit literature and in the RL-literature that provides $o(\sqrt{T})$ regret guarantees. We remark that $\mathrm{gap}_{\inf} > 0$ for all non-degenerate finite action-state spaces. Notice that $\mathrm{gap}_h(s,a) = V_h^*(s) - Q_h^*(s,a), \mathrm{gap}_{\min} = \inf h, s, a\{\mathrm{gap}_h(s,a) : \mathrm{gap}_h(s,a) \neq 0\}$ (if this set is empty the MDP is degenerate). Notice that if there are multiple optimal actions a_i at some state s, then $\mathrm{gap}_h(s,a_i) = 0$, so they will not be considered in the minimization. Thus, the multiplicity of the optimal policies does not break the assumption. Note that $\mathrm{gap}_{\min} = 0$ only in an infinite state-action space such that for all $\varepsilon > 0$, there are some $h_\varepsilon, s_\varepsilon, a_\varepsilon$ such that $0 < V_{h_\varepsilon}^*(s_\varepsilon) - Q_{h_\varepsilon}^*(s_\varepsilon, a_\varepsilon) \leq \varepsilon$ (i.e. we get arbitrarily close to an optimal action). Hence, it quantifies the hardness of the underlying problem because it provides a gap between the reward of an optimal policy and the reward of the non-optimal policies, so it shows how "easy" it is to distinguish an optimal one.

A.2 Omitted Algorithms

Algorithm 4 Model-Free Sampling Routine

Require: Function class $\mathcal{F}$, current sub-sampled dataset $\widehat{\mathcal{Z}}$, new element $z = (s,a)$, failure probability $\delta \in (0,1)$

1: Let p_z be the smallest number such that $1/p_z$ is an integer and p_z is greater than

$$\min\{1, C\text{sensitivity}_{\widehat{\mathcal{Z}},\mathcal{F}} \log(T\mathcal{N}(\mathcal{F}, \sqrt{\delta/(64T^3)})/\delta)\}$$

2: Let $\widehat{z} \in \mathcal{C}(\mathcal{S} \times \mathcal{A}, 1/16\sqrt{64T^3/\delta})$ such that

$$\sup_{f \in \mathcal{F}} |f(z) - f(\widehat{z})| \leq 1/16\sqrt{64T^3/\delta}$$

3: Add $1/p_z$ copies of $\widehat{z}$ into $\widehat{\mathcal{Z}}$ with probability p_z

4: **Return** $\widehat{\mathcal{Z}}$

Algorithm 5 Model-Based Sampling Routine

Require: Function class $\mathcal{F}$, current sub-sampled dataset $\widehat{\mathcal{Z}}$, new element $z = (s,a,V)$, failure probability $\delta \in (0,1)$

1: Let p_z be the smallest number such that $1/p_z$ is an integer and p_z is greater than

$$\min\{1, C \cdot \text{sensitivity}_{\widehat{\mathcal{Z}},\mathcal{F}} \cdot \log(T\mathcal{N}(\mathcal{F}, \sqrt{\delta/(64T^3)})/\delta)\}$$

2: Add $1/p_z$ copies of z into $\widehat{\mathcal{Z}}$ with probability p_z

3: **Return** $\widehat{\mathcal{Z}}$

B Proof of Theorem 3.1

In this section, our main goal is to prove Theorem 3.1. Recall that we assume we have access to a set $\mathcal{F} \subseteq \{f : \mathcal{S} \times \mathcal{A} \to [0, H+1]\}$, which we use to approximate the Q-function. For this set, we work with Assumption 2.2 and Assumption 2.3.

The proofs of the supporting lemmas are postponed to Appendix B.1. Before we are ready to prove our result, we need to discuss some results of prior works that are crucial to our proof.

The regret decomposition in [HZG21], gives us that

Lemma B.1. *[HZG21] For any MDP M we have that*

$$\mathbb{E}\left[\text{Regret}(K)\right] = \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\text{gap}_h(s_h^k, a_h^k)\right].$$

Moreover, for any $\tau > 0$ it holds with probability at least $1 - \lceil \log T \rceil e^{-\tau}$ that

$$\text{Regret}(K) \leq 2\sum_{k=1}^{K}\sum_{h=1}^{H}\text{gap}_h(s_h^k, a_h^k) + \frac{16H^2\tau}{3} + 2.$$

The following lemma resembles Lemma 6.3 [HZG21]. Its proof is postponed to Appendix B.1.

Lemma B.2. *If we pick*

$$\beta = CH^2 \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \dim_E(\mathcal{F}, 1/T) \log^2 T \log\left(\mathcal{C}(\mathcal{S}\times\mathcal{A}, \delta/T^2)T/\delta)\right),$$

for come constant C and for $h \in [H]$, then we have that with probability at least $1 - 2K\delta$

$$\sum_{k=1}^{K}\left(V_h^*(s_h^k) - Q_h^*(s_h^k, a_h^k)\right) \leq \frac{CH^4 \dim_E^2(\mathcal{F}, 1/T)\log^2 T \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta)\log(\mathcal{C}(\mathcal{S}\times\mathcal{A}, \delta/T^2)T/\delta)}{\text{gap}_{\min}}.$$

We are now ready to state the regret bound of our algorithm. In particular the regret guarantee follows from the regret decomposition and the bound we established before.

Lemma B.3. *There exists a constant C and proper values of the parameter β of Algorithm 1 such that with probability at least $1 - \lceil \log T \rceil e^{-\tau} - 2K\delta$ the regret of the algorithm is bounded by*

$$\text{Regret}(K) \leq \frac{CH^5 \dim_E^2(\mathcal{F}, 1/T)\log^2 T \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta)\log(\mathcal{C}(\mathcal{S}\times\mathcal{A}, \delta/T^2)T/\delta)}{\text{gap}_{\min}} + \frac{16H^2\tau}{3} + 2.$$

The choice of the parameter is

$$\beta = CH^2 \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \dim_E(\mathcal{F}, 1/T) \log^2 T \log\left(\mathcal{C}(\mathcal{S}\times\mathcal{A}, \delta/T^2)T/\delta)\right).$$

Proof. Throughout the proof, we condition on the events described in Lemma B.1, B.2 which happen with probability at least $1 - \lceil \log T \rceil e^{-\tau} - 2K\delta$.

From Lemma B.1 we have that

$$\text{Regret}(K) \leq 2\sum_{k=1}^{K}\sum_{h=1}^{H}\text{gap}_h(s_h^k, a_h^k) + \frac{16H^2\tau}{3} + 2.$$

We can bound the first term on the RHS using Lemma B.2 as follows

$$\begin{aligned}
2\sum_{k=1}^{K}\sum_{h=1}^{H}\text{gap}_h(s_h^k, a_h^k) &= 2\sum_{k=1}^{K}\sum_{h=1}^{H}\left(V_h^*(s_h^k) - Q_h^*(s_h^k, a_h^k)\right) \\
&\leq \frac{CH^5 \dim_E^2(\mathcal{F}, 1/T)\log^2 T \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta)\log(\mathcal{C}(\mathcal{S}\times\mathcal{A}, \delta/T^2)T/\delta)}{\text{gap}_{\min}}.
\end{aligned}$$

Hence, for the total regret we have that

$$\text{Regret}(K) \leq \frac{CH^5 \dim_E^2(\mathcal{F}, 1/T)\log^2 T \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta)\log(\mathcal{C}(\mathcal{S}\times\mathcal{A}, \delta/T^2)T/\delta)}{\text{gap}_{\min}} + \frac{16H^2\tau}{3} + 2.$$

□

We remark that the value of β we use in the main body is simply an upper bound on this value.

Finally, we state the bound on the adaptivity of our algorithm. In particular, since our algorithm is the same as in [KSWY21], the logarithmic switching cost follows directly from their result.

Lemma B.4. *[KSWY21] For any fixed $h \in [H]$, With probability $1-\delta$, the sub-sampled dataset $\widehat{\mathcal{Z}}_h^k$ changes at most*

$$O\left(\log(T\mathcal{N}(\mathcal{F}, \sqrt{\delta/T^3})/\delta)\dim_E(\mathcal{F}, 1/T)\log^2 T\right)$$

times.

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1:

The proof of this theorem follows by using Lemma B.3 and taking a union on the result of Lemma B.4 and setting the error probability accordingly. □

B.1 Supporting Lemmas: Theorem 3.1

In this section, we present the proof of Lemma B.2.

First, we need to show that the sub-sampled dataset is a good approximation of the original one. To this end, we use Proposition 1 from [KSWY21].

Proposition B.5. *[KSWY21] For any $h, k \in [H] \times [K]$ we let*

$$\underline{b}_h^k(\cdot,\cdot) = \sup_{\|f_1-f_2\|^2_{\mathcal{Z}_h^k} \le \beta/100} |f_1(\cdot,\cdot) - f_2(\cdot,\cdot)|,$$

$$\overline{b}_h^k(\cdot,\cdot) = \sup_{\|f_1-f_2\|^2_{\mathcal{Z}_h^k} \le 100\beta} |f_1(\cdot,\cdot) - f_2(\cdot,\cdot)|.$$

Then, with probability at least $1-\delta/32$ we have that

$$\underline{b}_h^k(\cdot,\cdot) \le b_h^k(\cdot,\cdot) \le \overline{b}_h^k(\cdot,\cdot).$$

We now present a generalized version of Lemma 11 [KSWY21] that will be used to bound the regret of our algorithm. Essentially, this gives a bound on the summation of the bonus terms over a set of episodes $K' \subseteq [K]$ in terms of the eluder dimension of the function class and the number of episodes.

Lemma B.6. *For any set $K' \subseteq [K]$, with probability at least $1-\delta/32$ we have that*

$$\sum_{i=1}^{|K'|}\sum_{h=1}^{H} b_h^{k_i}(s_h^{k_i}, a_h^{k_i}) \le H + H(H+1)\cdot \dim_E(\mathcal{F}, 1/T) + CH\sqrt{\dim_E(\mathcal{F}, 1/T)|K'|\cdot\beta},$$

where $C > 0$ is some constant.

Proof. Throughout the proof, we condition on the event defined in Proposition B.5. This gives us that for any $k \in K', h \in H$

$$b_h^k(s_h^k, a_h^k) \le \overline{b}_h^k(s_h^k, a_h^k) = \sup_{||f_1-f_2||^2_{\mathcal{Z}_h^k} \le 100\beta} |f_1(s_h^k, a_h^k) - f_2(s_h^k, a_h^k)|.$$

We bound $\sum_{i=1}^{K'} \overline{b}_h^{k_i}(s_h^{k_i}, a_h^{k_i})$ for each $h \in [H]$ separately.

Given some $\epsilon > 0$, we define $\mathcal{L}_h = \{(s_h^{k_i}, a_h^{k_i}) : k_i \in K', \overline{b}_h^{k_i}(s_h^{k_i}, a_h^{k_i}) > \epsilon\}$, i.e. the set of state-action pairs at step h and some episode in K' where the bonus function has value greater than ϵ. Consider some $k \in K'$. We denote $L_h = |\mathcal{L}_h|$, $\widetilde{\mathcal{Z}}_h^k = \{(s_h^k, a_h^k) \in \mathcal{Z}_h^k, k \in K'\}$. Our goal is to show that there is some $z_h^k = (s_h^k, a_h^k) \in \mathcal{L}_h$ that is ϵ-dependent on at least $L_h/\dim_E(\mathcal{F}, \epsilon) - 1$ disjoint subsequences in $\mathcal{Z}_h^k \cap \mathcal{L}_h$. We also denote $N = L_h/\dim_E(\mathcal{F}, \epsilon) - 1$.

To do that, we decompose $\mathcal{L}_h$ into $N+1$ disjoint subsets and we denote the j-th subset by $\mathcal{L}_h^j$. We use the following procedure. Initially we set $\mathcal{L}_h^j = \emptyset$ for all $j \in [N+1]$ and consider every $z_h^k \in \mathcal{L}_h$ in a sequential manner. For each such z_h^k we find the smallest index $j, 1 \leq j \leq N$, such that z_h^k is ϵ-independent of the elements in $\mathcal{L}_h^j$ with respect to $\mathcal{F}$. If there is no such j, we set $j = N+1$. Then, we update $\mathcal{L}_h^j \leftarrow \mathcal{L}_h^j \cup z_h^k$. Notice that after we go through all the elements of $\mathcal{L}_h$, we must have that $\mathcal{L}_h^{N+1} \neq \emptyset$. This is because every set $\mathcal{L}_h^j, 1 \leq j \leq N$, contains at most $\dim_E(\mathcal{F}, \epsilon)$ elements. Moreover, by definition, every element $z_h^k \in \mathcal{L}_h^{N+1}$ is ϵ-dependent on $N = L_h/\dim_E(\mathcal{F}, \epsilon) - 1$ disjoint subsequences in $\mathcal{Z}_h^k \cap \mathcal{L}^h$.

Furthermore, since $\bar{b}_h^k(s_h^k, a_h^k) > \epsilon$ for all $z_h^k \in \mathcal{L}_h$ there must exist $f_1, f_2 \in \mathcal{F}$ such that $|f_1(s_h^k, a_h^k) - f_2(s_h^k, a_h^k)| > \epsilon$ and $||f_1 - f_2||^2_{\mathcal{Z}_h^k} \leq 100\beta$. Hence, since $z_h^k \in \mathcal{L}_h^{N+1}$ is ϵ-dependent on N disjoint subsequences $\mathcal{L}_h^j$ and for each such subsequence, by the definition of ϵ-dependence, it holds that $||f_1 - f_2||_{\mathcal{L}_h^j} > \epsilon^2$ we have that

$$N\epsilon^2 \leq ||f_1 - f_2||^2_{\mathcal{Z}_h^k} \leq 100\beta \implies (L_h/\dim_E(\mathcal{F}, \epsilon) - 1)\epsilon^2 \leq 100\beta \implies L_h \leq \left(\frac{100\beta}{\epsilon^2} + 1\right) \dim_E(\mathcal{F}, \epsilon).$$

We now pick a permutation $\bar{b}_1 \geq \bar{b}_2 \geq \ldots \geq \bar{b}_{|K'|}$ of the bonus functions $\{\bar{b}_h^k(s_h^k, a_h^k)\}_{k \in K'}$. For all $\bar{b}_k \geq 1/|K'|$ it holds that

$$k \leq \left(\frac{100\beta}{\bar{b}_k^2} + 1\right) \dim_E(\mathcal{F}, \bar{b}_k) \leq \left(\frac{100\beta}{\bar{b}_k^2} + 1\right) \dim_E(\mathcal{F}, 1/K') \implies$$
$$\bar{b}_k \leq \left(\frac{k}{\dim_E(\mathcal{F}, 1/K')} - 1\right)^{-1/2} \sqrt{100\beta}.$$

Moreover, notice that we get by definition that $\bar{b}_k \leq H + 1$. Hence, we have that

$$\begin{aligned}
\sum_{i=1}^{|K'|} \bar{b}_h^{k_i}(s_h^{k_i}, a_h^{k_i}) &= \sum_{i: \bar{b}_{k_i} < 1/|K'|} \bar{b}_h^{k_i}(s_h^{k_i}, a_h^{k_i}) + \sum_{i: \bar{b}_{k_i} \geq 1/|K'|} \bar{b}_h^{k_i}(s_h^{k_i}, a_h^{k_i}) \\
&\leq |K'| \cdot 1/|K'| + \sum_{i: \bar{b}_{k_i} \geq 1/|K'|, i \leq \dim_E(\mathcal{F}, 1/|K'|)} \bar{b}_h^{k_i}(s_h^{k_i}, a_h^{k_i}) \\
&+ \sum_{i: \bar{b}_{k_i} \geq 1/|K'|, |K'| \geq i > \dim_E(\mathcal{F}, 1/|K'|)} \bar{b}_h^{k_i}(s_h^{k_i}, a_h^{k_i}) \\
&\leq 1 + (H+1) \cdot \dim_E(\mathcal{F}, 1/|K'|) + \sum_{|K'| \geq i > \dim_E(\mathcal{F}, 1/|K'|)} \left(\frac{k}{\dim_E(\mathcal{F}, 1/|K'|)} - 1\right)^{-1/2} \sqrt{100\beta} \\
&\leq 1 + (H+1) \cdot \dim_E(\mathcal{F}, 1/|K'|) + C\sqrt{\dim_E(\mathcal{F}, 1/|K'|)|K'|\beta} \\
&\leq 1 + (H+1) \cdot \dim_E(\mathcal{F}, 1/T) + C\sqrt{\dim_E(\mathcal{F}, 1/T)|K'|\beta}
\end{aligned}$$

for some constant $C > 0$, where the second to last inequality can be obtained by bounding the summation by the integral and the last one by the definition of the eluder dimension. Summing up all the inequalities for $h \in [H]$, we get the result. □

We need the Azuma-Hoeffding inequality to bound a martingale difference sequence. For completeness, we present it here as well.

Lemma B.7. *[CBL06] Let $\{x_i\}_{i=1}^n$ be a martingale difference sequence with respect to some filtration $\{\mathcal{F}_i\}$ for which $|x_i| \leq M$ for some constant M, x_i is $\mathcal{F}_{i+1}$ measurable and $\mathbb{E}[x_i|\mathcal{F}_i] = 0$.*

Then, for any $0 < \delta < 1$, we have that with probability at least $1 - \delta$ it holds that

$$\sum_{i=1}^{n} x_i \leq M\sqrt{2n\log(1/\delta)}.$$

The following lemma that appears in [KSWY21] shows that the estimate of the Q-function upper bounds the optimal one.

Lemma B.8. *[KSWY21] With probability at least $1 - \delta/2$ we have that for all $(k, h) \in [K] \times [H]$ and all $(s, a) \in \mathcal{S} \times \mathcal{A}$*

$$Q_h^*(s,a) \leq Q_h^k(s,a) \leq \bar{f}_h^k(s,a) + 2b_h^k(s,a)$$

where $\bar{f}_h^k(\cdot,\cdot) = \sum_{s' \in \mathcal{S}} P_h(s'|\cdot,\cdot)V_{h+1}^k(s') + r_h(\cdot,\cdot)$.

We now present a lemma that bounds the number of rounds that the suboptimilaty gap falls in some interval. It is inspired by Lemma 6.2 [HZG21].

Lemma B.9. *If we pick*

$$\beta = CH^2 \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \dim_E(\mathcal{F}, 1/T) \log^2 T \log\left(\mathcal{C}(\mathcal{S} \times \mathcal{A}, \delta/T^2)T/\delta)\right),$$

then there exists a constant $\widetilde{C}$ such that for all $h \in [H], n \in [N]$ with probability at least $1 - 2K\delta$, we have that

$$\sum_{k=1}^{K} \mathbb{1}[V_h^*(s_h^k) - Q_h^{\pi_k}(s_h^k, a_h^k) \geq 2^n \text{gap}_{\min}] \leq \frac{\widetilde{C}H^4 \dim_E^2(\mathcal{F}, 1/T) \log^2 T \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \log(\mathcal{C}(\mathcal{S} \times \mathcal{A}, \delta/T^2)T/\delta)}{4^n \text{gap}_{\min}^2}.$$

Proof. We keep h fixed.

We denote by K' the set of episodes where the gap at step h is at least 2^n, i.e.

$$K' = \left\{k \in [K] : V_h^*(s_h^k) - Q_h^{\pi_k}(s_h^k, a_h^k) \geq 2^n \text{gap}_{\min}\right\}.$$

The goal is to bound the quantity $\sum_{i=1}^{|K'|}\left(Q_h^{k_i}(s_h^{k_i}, a_h^{k_i}) - Q_h^{\pi_{k_i}}(s_h^{k_i}, a_h^{k_i})\right)$ from below and above with functions $f_1(|K'|), f_2(|K'|)$ and then use the fact that $f_1(|K'|) \leq f_2(|K'|)$ to derive an upper bound on $|K'|$.

For the lower bound, we have that

$$\begin{aligned}
\sum_{i=1}^{|K'|}\left(Q_h^{k_i}(s_h^{k_i}, a_h^{k_i}) - Q_h^{\pi_{k_i}}(s_h^{k_i}, a_h^{k_i})\right) &\geq \sum_{i=1}^{|K'|}\left(Q_h^{k_i}(s_h^{k_i}, \pi_h^*(s_h^{k_i})) - Q_h^{\pi_{k_i}}(s_h^{k_i}, a_h^{k_i})\right) \\
&\geq \sum_{i=1}^{|K'|}\left(Q_h^{*}(s_h^{k_i}, \pi_h^*(s_h^{k_i})) - Q_h^{\pi_{k_i}}(s_h^{k_i}, a_h^{k_i})\right) \\
&= \sum_{i=1}^{|K'|}\left(V_h^{*}(s_h^{k_i}) - Q_h^{\pi_{k_i}}(s_h^{k_i}, a_h^{k_i})\right) \geq 2^n \text{gap}_{\min}|K'|,
\end{aligned}$$

where the first inequality holds by the definition of the policy π_{k_i}, the second one follows because $Q_h^{k_i}(\cdot,\cdot)$ is an optimistic estimate of $Q_h^*(\cdot,\cdot)$ which happens with probability at least $1 - \delta/2$ (see Lemma B.8) and the third one by the definition of k_i.

We get the upper bound on this quantity in the following way. For any $h' \in [H]$ we have

$$\begin{aligned}
Q_{h'}^k(s_{h'}^k, a_{h'}^k) - Q_{h'}^{\pi_k}(s_{h'}^k, a_{h'}^k) &\leq \sum_{s' \in \mathcal{S}} P_{h'}(s'|s_{h'}^k, a_{h'}^k)V_{h'+1}^k(s') + r_{h'}(s_{h'}^k, a_{h'}^k) + 2b_{h'}^k(s_{h'}^k, a_{h'}^k) - Q_{h'}^{\pi_k}(s_{h'}^k, a_{h'}^k) \\
&= \left\langle P_{h'}(\cdot|s_{h'}^k, a_{h'}^k), V_{h'+1}^k - V_{h'+1}^{\pi_k}\right\rangle + 2b_{h'}^k(s_{h'}^k, a_{h'}^k) \\
&= V_{h'+1}^k(s_{h'+1}^k) - V_{h'+1}^{\pi_k}(s_{h'+1}^k) + \epsilon_{h'}^k + 2b_{h'}^k(s_{h'}^k, a_{h'}^k) \\
&= Q_{h'+1}^k(s_{h'+1}^k, a_{h'+1}^k) - Q_{h'+1}^{\pi_k}(s_{h'+1}^k, a_{h'+1}^k) + \epsilon_{h'}^k + 2b_{h'}^k(s_{h'}^k, a_{h'}^k)
\end{aligned}$$

where we define $\epsilon_{h'}^k = \left\langle P_{h'}(\cdot|s_{h'}^k, a_{h'}^k), V_{h'+1}^k - V_{h'+1}^{\pi_k} \right\rangle - (V_{h'+1}^k(s_{h'+1}^k) - V_{h'+1}^{\pi_k}(s_{h'+1}^k))$ and the inequality follows from Lemma B.8.

We now take the summation over all $k \in |K'|, h \leq h' \leq H$ and we get

$$\sum_{i=1}^{|K'|} \left(Q_h^{k_i}(s_h^{k_i}, a_h^{k_i}) - Q_h^{\pi_{k_i}}(s_h^{k_i}, a_h^{k_i}) \right) \leq \sum_{i=1}^{|K'|} \sum_{h'=h}^{H} \epsilon_{h'}^{k_i} + \sum_{i=1}^{|K'|} \sum_{h'=h}^{H} b_{h'}^{k_i}(s_{h'}^{k_i}, a_{h'}^{k_i}).$$

We will bound each of the two terms on the RHS separately.

For the first term, we notice that $x_j = \left\langle P_j(\cdot|s_j^{k_i}, a_j^{k_i}), V_{j+1}^{k_i} - V_{j+1}^{\pi_{k_i}} \right\rangle - (V_{j+1}^{k_i}(s_{j+1}^{k_i}) - V_{j+1}^{\pi_{k_i}}(s_{j+1}^{k_i}))$ forms a martingale difference sequence with zero mean and $|x_j| \leq 2H$. Hence, we can use Lemma B.7 and that for each $k \in K'$, with probability at least $1 - \delta$ we have that

$$\sum_{i=1}^{k} \sum_{j=1}^{H} \left(\left\langle P_j(\cdot|s_j^{k_i}, a_j^{k_i}), V_{j+1}^{k_i} - V_{j+1}^{\pi_{k_i}} \right\rangle - \left(V_{j+1}^{k_i}(s_{j+1}^{k_i}) - V_{j+1}^{\pi_{k_i}}(s_{j+1}^{k_i}) \right) \right) \leq \sqrt{8kH^3 \log(1/\delta)}.$$

If we take the union bound over all $k \in [K]$ we have that with probability at least $1 - |K'|\delta$

$$\sum_{i=1}^{|K'|} \sum_{h'=h}^{H} \epsilon_{h'}^{k_i} \leq \sqrt{8|K'|H^3 \log(1/\delta)}.$$

We now focus on the second term. Using Lemma B.6 we get that

$$\sum_{i=1}^{|K'|} \sum_{h'=h}^{H} b_{h'}^{k_i}(s_{h'}^{k_i}, a_{h'}^{k_i}) \leq H + H(H+1)\dim_E(\mathcal{F}, 1/T) + CH\sqrt{\dim_E(\mathcal{F}, 1/T)|K'|\beta}$$

and this happens with probability at least $1 - \delta/32$. Hence, combining the upper and lower bound of

$$\sum_{i=1}^{|K'|} \left(Q_h^{k_i}(s_h^{k_i}, a_h^{k_i}) - Q_h^{\pi_{k_i}}(s_h^{k_i}, a_h^{k_i}) \right)$$

we get that

$$2^n \text{gap}_{\min}|K'| \leq \sqrt{8|K'|H^3 \log(1/\delta)} + H + H(H+1)\dim_E(\mathcal{F}, 1/T) + CH\sqrt{\dim_E(\mathcal{F}, 1/T)|K'|\beta}.$$

Solving for $|K'|$ gives us that

$$|K'| \leq \frac{\widetilde{C}H^4 \dim_E^2(\mathcal{F}, 1/T) \log^2 T \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \log(\mathcal{C}(\mathcal{S} \times \mathcal{A}, \delta/T^2)T/\delta)}{4^n \text{gap}_{\min}^2}.$$

□

We are now ready to prove Lemma B.2.

Proof of Lemma B.2: Throughout this proof we condition on the event described in Lemma B.9 which happens with probability at least $1 - 2K\delta$. Since $\text{gap}_{\min} > 0$ whenever we do not take the optimal action, we have that either $V_h^*(s_k) - Q_h^*(s_h^k, a_h^k) = 0$ or $V_h^*(s_k) - Q_h^*(s_h^k, a_h^k) \geq \text{gap}_{\min}$. Our approach is to divide the interval $[0, H]$ into $N = \lceil \log(H/\text{gap}_{\min}) \rceil$ intervals and count the number

of $V_h^*(s_k) - Q_h^*(s_h^k, a_h^k)$ that fall into each interval. Notice that for every $V_h^*(s_k) - Q_h^*(s_h^k, a_h^k)$ that falls into interval i we can get an upper bound of $V_h^*(s_k) - Q_h^*(s_h^k, a_h^k) \leq 2^i \text{gap}_{\min}$ and this upper bound is essentially tight. Hence, we have that

$$\begin{aligned}
\sum_{k=1}^{K} \left(V_h^*(s_h^k) - Q_h^*(s_h^k, a_h^k)\right) &\leq \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{1}\left[2^i \text{gap}_{\min} \geq V_h^*(s_h^k) - Q_h^*(s_h^k, a_h^k) \geq 2^{i-1}\text{gap}_{\min}\right] \cdot 2^i \text{gap}_{\min} \\
&\leq \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{1}\left[V_h^*(s_h^k) - Q_h^*(s_h^k, a_h^k) \geq 2^{i-1}\text{gap}_{\min}\right] \cdot 2^i \text{gap}_{\min} \\
&\leq \sum_{i=1}^{N} \frac{\widetilde{C} H^4 \dim_E^2(\mathcal{F}, 1/T) \log^2 T \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \log(\mathcal{C}(\mathcal{S}\times\mathcal{A}, \delta/T^2)T/\delta)}{4^{i-1}\text{gap}_{\min}^2} \cdot 2^i \text{gap}_{\min} \\
&= \sum_{i=1}^{N} \frac{\widetilde{C} H^4 \dim_E^2(\mathcal{F}, 1/T) \log^2 T \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \log(\mathcal{C}(\mathcal{S}\times\mathcal{A}, \delta/T^2)T/\delta)}{2^{i}\text{gap}_{\min}} \\
&\leq \frac{C H^4 \dim_E^2(\mathcal{F}, 1/T) \log^2 T \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \log(\mathcal{C}(\mathcal{S}\times\mathcal{A}, \delta/T^2)T/\delta)}{\text{gap}_{\min}}
\end{aligned}$$

where the first inequality holds by the definition of the intervals, the second due to the properties of the indicator function, the third because of Lemma B.9 and in the last two steps we just manipulate the constants. □

C Proof of Theorem 3.2

In this section, our main goal is to prove Theorem 3.2. We work with Assumption 2.4. We follow the same regret decomposition as in Appendix B.

We first present a lemma that is crucial in bounding the regret of the algorithm.

Lemma C.1. *If we pick*

$$\beta = 4H^2 \log(2\mathcal{N}(\mathcal{F}, 1/T)/\delta) + 4/H\left(C + \sqrt{H^2/4\log(4(K(K+1)/\delta))}\right),$$

for come constant C and for $h \in [H]$, then we have that with probability at least $1 - (K+3)\delta$

$$\sum_{k=1}^{K} \left(V_h^*(s_h^k) - Q_h^*(s_h^k, a_h^k)\right) \leq \frac{C H^4 \log(T\mathcal{N}(\mathcal{F}, 1/T)/\delta) \dim_E^2(\mathcal{F}, 1/T)}{\text{gap}_{\min}}.$$

We are now ready to state the regret bound of our algorithm.

Lemma C.2. *There exists a constant C and proper values of the parameter β of Algorithm 1 such that with probability at least $1 - \lceil \log T \rceil e^{-\tau} - H(K+3)\delta$ the regret of the algorithm is bounded by*

$$\text{Regret}(K) \leq \frac{C H^5 \log(T\mathcal{N}(\mathcal{F}, 1/T)/\delta) \dim_E^2(\mathcal{F}, 1/T)}{\text{gap}_{\min}} + \frac{16H^2\tau}{3} + 2.$$

The value of the parameter is

$$\beta = 4H^2 \log(2\mathcal{N}(\mathcal{F}, 1/T)/\delta) + 4/H\left(C + \sqrt{H^2/4\log(4(K(K+1)/\delta))}\right).$$

where $\mathcal{N}(\mathcal{F}, 1/T) = \arg\max_{h\in[H]} \mathcal{N}(\mathcal{F}_h, 1/T)$.

In particular, the dependence of the regret in the time horizon T is logarithmic.

Proof. Throughout the proof, we condition on the events described in Lemma B.1, C.1 which happen with probability at least $1 - \lceil \log T \rceil e^{-\tau} - H(K+3)\delta$.

From Lemma B.1 we have that

$$\text{Regret}(K) \leq 2\sum_{k=1}^{K}\sum_{h=1}^{H} \text{gap}_h(s_h^k, a_h^k) + \frac{16H^2\tau}{3} + 2.$$

We can bound the first term on the RHS using Lemma C.1 as follows

$$2\sum_{k=1}^{K}\sum_{h=1}^{H} \text{gap}_h(s_h^k, a_h^k) = 2\sum_{k=1}^{K}\sum_{h=1}^{H} \left(V_h^*(s_h^k) - Q_h^*(s_h^k, a_h^k)\right) \leq \frac{2CH^5 \log(T\mathcal{N}(\mathcal{F}, 1/T)/\delta)\dim_E^2(\mathcal{F}, 1/T)}{\text{gap}_{\min}}.$$

This gives us the result.

□

We remark that the value of β we use in the main body is simply an upper bound on this value.

The only thing that we need to do now is to bound the number of rounds that we update our policy. Since we are using exactly the same sensitivity score and update probability as in [KSWY21], this follows from their result.

Lemma C.3. *[KSWY21] With probability at least $1 - \delta/32$ for any fixed $h \in [H]$ we have that the sub-sampled dataset $\widehat{Z}_h^k, k \in [K]$ changes at most*

$$S_{\max} = C \cdot \log(T\mathcal{N}(F_h, \sqrt{\delta/(64T^3)})/\delta)\dim_E(\mathcal{F}_h, 1/T)\log^2 T$$

times.

We are now ready to prove Theorem 3.2.

Proof of Theorem 3.2:

=The proof of this theorem follows by combining Lemma C.2 and taking a union on the result of Lemma C.3 and setting the error probability accordingly. □

C.1 Supporting Lemmas: Theorem 3.2

In this section our goal is to prove the supporting lemmas of Theorem 3.2.

Recall that our approach is to modify the algorithm in [KSWY21] to work in this setting and use a similar analysis as in Appendix B. Unlike Appendix B where we approximate the optimal Q-function, here we try to estimate the true transition kernel. Let $\mathcal{Z}_h^k = \{\left(s_h^\tau, a_h^\tau, V_{h+1}^\tau(\cdot)\right)\}_{\tau\in[k-1]}$ be the dataset up to episode k and $\widehat{\mathcal{Z}}_h^k$ the sub-sampled dataset. In each episode k, we update our policy whenever we add an element in the dataset for some $h \in [H]$. Recall that whenever we perform an update our policy becomes:

$$\begin{aligned}
&Q_{H+1}^k(s,a) = 0,\\
&V_{H+1}^k(s) = 0,\\
&Q_h^k(s,a) = \min\{r_h(s,a) + \langle \widehat{P}_h^k(\cdot|s,a), V_{h+1}^k\rangle + b_h^k(s,a), H\},\\
&V_h^k(s) = \max_{a\in\mathcal{A}} Q_h^k(s,a)
\end{aligned}$$

for some $\widehat{P}_h^k, b_h^k(\cdot,\cdot)$ that we will define shortly. We get the policy $\pi_h^k(s)$ by picking greedily the action that maximizes the estimate $Q_h^k(s,a)$.

The least-squares estimate of the model is

$$\widehat{P}_h^{k+1} = \arg\min_{P\in\mathcal{P}_h} \sum_{k'=1}^{k} \left(\langle P(\cdot|s_h^{k'}, a_h^{k'}), V_{h+1}^{k'}\rangle - y_h^{k'}\right)^2, y_h^{k'} = V_{h+1}^{k'}(s_{h+1}^{k'}).$$

Recall the definition of the function class that we use in the derivation of our results.

Definition C.4. Let $\mathcal{V}$ be the set all measurable functions that are bounded by H. We now let $f : \mathcal{S} \times \mathcal{A} \times \mathcal{V} \to \mathbb{R}$ and define the following set:

$$\mathcal{F}_h = \left\{ f : \exists P_h \in \mathcal{P}_h \text{ so that } f(s,a,V) = \int_{\mathcal{S}} P_h(s'|s,a)V(s')ds', \forall (s,a,V) \in \mathcal{S} \times \mathcal{A} \times \mathcal{V} \right\}. \tag{1}$$

Recall that in this setting the norm of a function with respect to a dataset $\mathcal{Z}$ is

$$||f||_{\mathcal{Z}} = \sqrt{\sum_{z=(s_z,a_z,V_z(\cdot))\in\mathcal{Z}} (f(s_z,a_z,V_z(\cdot))^2}.$$

Recall also that the bonus function is

$$b_h^k(s,a) = \sup_{f_1,f_2\in\mathcal{F}_h : \min\{||f_1-f_2||_{\widehat{\mathcal{Z}}_h^k}, T(H+1)^2\}\le\beta} |f_1(s,a,V_{h+1}^k(\cdot)) - f_2(s,a,V_{h+1}^k(\cdot))|.$$

The parameter β will be defined later in a way that will ensure optimism.

First, we need to show that at for every $k \in [K], h \in [H]$, the sub-sampled dataset approximates the original one. Our approach is inspired by [KSWY21].

We define the following quantities

$$\begin{aligned}
\underline{\mathcal{C}}_h^k(\alpha) &= \left\{ (f_1,f_2) \in \mathcal{F}_h \times \mathcal{F}_h : ||f_1 - f_2||_{\mathcal{Z}_h^k}^2 \le \alpha/100 \right\} \\
\widehat{\mathcal{C}}_h^k(\alpha) &= \left\{ (f_1,f_2) \in \mathcal{F}_h \times \mathcal{F}_h : \min\{||f_1 - f_2||_{\widehat{\mathcal{Z}}_h^k}^2, T(H+1)^2\} \le \alpha \right\} \\
\overline{\mathcal{C}}_h^k(\alpha) &= \left\{ (f_1,f_2) \in \mathcal{F}_h \times \mathcal{F}_h : ||f_1 - f_2||_{\mathcal{Z}_h^k}^2 \le 100\alpha \right\}.
\end{aligned}$$

We also let

$$\begin{aligned}
\underline{b}_h^k(s,a) &= \sup_{f_1,f_2\in\underline{\mathcal{C}}_h^k(\beta)} |f_1(s,a,V_{h+1}^k(\cdot)) - f_2(s,a,V_{h+1}^k(\cdot))| \\
\overline{b}_h^k(s,a) &= \sup_{f_1,f_2\in\overline{\mathcal{C}}_h^k(\beta)} |f_1(s,a,V_{h+1}^k(\cdot)) - f_2(s,a,V_{h+1}^k(\cdot))|.
\end{aligned}$$

Our goal is to show that $\underline{\mathcal{C}}_h^k(\alpha) \subseteq \widehat{\mathcal{C}}_h^k(\alpha) \subseteq \overline{\mathcal{C}}_h^k(\alpha)$ with high probability. Let $\mathcal{E}_h^k(\alpha)$ denote the event that this holds. We also denote by $\mathcal{E}_h^k = \cap_{n=0}^{\infty} \mathcal{E}_h^k(100^n\beta)$. This event will show us that $\widehat{\mathcal{Z}}_h^k$ is a good approximation to $\mathcal{Z}_h^k$.

Notice that whenever this happens, it holds that $\underline{b}_h^k(s,a) \le b_h^k(s,a) \le \overline{b}_h^k(s,a)$.

The following lemma which is inspired by [KSWY21] establishes that fact.

Lemma C.5. *The probability that all the events $\mathcal{E}_h^k$ happen satisfies*

$$\Pr\left(\bigcap_{k=1}^{K} \bigcap_{h=1}^{H} \mathcal{E}_h^k \right) \ge 1 - \delta.$$

To prove Lemma C.5 we need the following concentration inequality proved in [Fre75].

Lemma C.6. *Let $\{Y_i\}_{i\in\mathbb{N}}$ be a real-valued martingale with difference sequence $\{X_i\}_{i\in\mathbb{N}}$. Let R be a uniform bound on X_i. Fix some $n \in \mathbb{N}$ and let σ^2 be a number such that*

$$\sum_{i=1}^{n} \mathbb{E}[X_i^2 | Y_0, \ldots, Y_{i-1}] \le \sigma^2.$$

Then, for all $t \ge 0$ we have that

$$\Pr(|Y_n - Y_0| \ge t) \le 2\exp\left\{ -\frac{t^2/2}{\sigma^2 + Rt/3} \right\}.$$

Moreover, we need a bound on the number of elements that are in the sub-sampled dataset. This is established in [KSWY21].

Lemma C.7. *[KSWY21] We have that with probability at least $1-\delta/(64T)$, we have $|\widehat{\mathcal{Z}}_h^k| \leq 64T^3/\delta$ for all $\delta > 0$.*

The subsequent lemma shows that, indeed, whenever $\mathcal{E}_h^k$ happens the sub-sampled dataset is a good approximation of the original one. It was proved in [KSWY21].

Lemma C.8. *[KSWY21] Whenever the event $\mathcal{E}_h^k$ happens, it holds that*

$$\frac{1}{10000}||f_1 - f_2||^2_{\mathcal{Z}_h^k} \leq \min\{||f_1 - f_2||^2_{\widehat{\mathcal{Z}}_h^k}, T(H+1)^2\} \leq 10000||f_1 - f_2||^2_{\mathcal{Z}_h^k}, \text{ if } ||f_1 - f_2||^2_{\mathcal{Z}_h^k} > 100\beta$$

and

$$\min\{||f_1 - f_2||^2_{\widehat{\mathcal{Z}}_h^k}, T(H+1)^2\} \leq 10000\beta, \text{ if } ||f_1 - f_2||^2_{\mathcal{Z}_h^k} \leq 100\beta.$$

To establish our result, we need the following lemma. The proof follows the approach of [KSWY21]. We present it here for completeness.

Lemma C.9. *For any $\alpha \in [\beta, T(H+1)^2]$, a fixed $h \in [H]$ and $k \in [K]$ we have the following bound for the probability that all the events $\{\mathcal{E}_h^i\}_{i \leq k-1}$ happen and the last one does not happen*

$$\Pr\left(\left(\bigcap_{i=1}^{k-1} \mathcal{E}_h^i\right) \mathcal{E}_h^k(\alpha)^c\right) \leq \delta/(32T^2).$$

Proof. Let C_1 be the quantity the sensitivity in the sampling probability. We fix some $h \in [H]$ throughout the proof.

We consider a fixed pair of functions f_h^1, f_h^2 in the discretized set $\mathcal{C}(\mathcal{F}_h, \sqrt{\delta/(64T^3)})$ and for $i \geq 2$ we let

$$Z_i = \max\left\{||f_h^1 - f_h^2||^2_{\mathcal{Z}_h^i}, \min\{||f_h^1 - f_h^2||^2_{\widehat{\mathcal{Z}}_h^{i-1}}, T(H+1)^2\}\right\}.$$

We also define

$$Y_i = \begin{cases} \frac{1}{p_{z_h^{i-1}}}(f_h^1(z_h^{i-1}) - f_h^2(z_h^{i-1}))^2 & z_h^{i-1} \text{ is added to } \widehat{\mathcal{Z}}_h^i \text{ and } Z_i \leq 2000000\alpha \\ 0, & z_h^{i-1} \text{ is not added to } \widehat{\mathcal{Z}}_h^i \text{ and } Z_i \leq 2000000\alpha \\ (f_h^1(z_h^{i-1}) - f_h^2(z_h^{i-1}))^2 & \text{otherwise} \end{cases}$$

Let $\mathbb{F}_i$ be the filtration that Y_i is adapted to. Our goal is to use Freedman's inequality (i.e. Lemma C.6) for Y_i. Notice that $\mathbb{E}[Y_i|\mathbb{F}_i] = (f_h^1(z_h^{i-1}) - f_h^2(z_h^{i-1}))^2$. Now we focus on the variance of Y_i. Notice that if $p_{z_h^{i-1}} = 1$ or $Z_i > 2000000\alpha$ then Y_i is deterministic so $Y_i - \mathbb{E}[Y_i|\mathbb{F}_{i-1}] = \text{Var}[Y_i - \mathbb{E}[Y_i|\mathbb{F}_{i-1}]] = 0$. For the other case, recall that

$$p_{z_h^i} = \min\{1, C' \cdot \text{sensitivity}_{\widehat{\mathcal{Z}}_h^{i-1}, \mathcal{F}_h}(z_h^i) \cdot \log(T\mathcal{N}(\mathcal{F}, \sqrt{\delta/64T^3})/\delta)\} = \min\{1, C_1 \cdot \text{sensitivity}_{\widehat{\mathcal{Z}}_h^{i-1}, \mathcal{F}_h}(z_h^i)\}$$

and

$$\text{sensitivity}_{\mathcal{Z},\mathcal{F}}(z) = \min\left\{\sup_{f_1, f_2 \in \mathcal{F}} \frac{(f_1(z) - f_2(z))^2}{\min\{||f_1 - f_2||_{\mathcal{Z}}, T(H+1)^2\} + \beta}, 1\right\}$$

Since $p_{z_h^{i-1}} < 1 \implies C_1 \cdot \text{sensitivity}_{\widehat{\mathcal{Z}}_h^{i-1}, \mathcal{F}_h}(z_h^i) < 1$. We consider two cases. If $Y_i \neq 0$ we can see that $Y_i \geq \mathbb{E}[Y_i|\mathbb{F}_i]$ so $|Y_i - \mathbb{E}[Y_i]| \leq Y_i$. Moreover,

$$
\begin{aligned}
Y_i &\leq \frac{(f_h^1(z_h^{i-1}) - f_h^2(z_h^{i-1}))^2}{p_{z_h^{i-1}}} \\
&\leq \frac{(f_h^1(z_h^{i-1}) - f_h^2(z_h^{i-1}))^2}{C_1 \sup_{f_1, f_2 \in \mathcal{F}_h} \frac{(f_1(z_h^{i-1}) - f_2(z_h^{i-1}))^2}{\min\{||f_1 - f_2||_{\widehat{\mathcal{Z}}_h^{i-1}}, T(H+1)^2\} + \beta}} \\
&\leq \frac{(f_h^1(z_h^{i-1}) - f_h^2(z_h^{i-1}))^2 \min\{||f_1 - f_2||_{\widehat{\mathcal{Z}}_h^{i-1}}, T(H+1)^2\} + \beta\}}{C_1 (f_h^1(z_h^{i-1}) - f_h^2(z_h^{i-1}))^2} \\
&= \left(\min\{||f_1 - f_2||_{\widehat{\mathcal{Z}}_h^{i-1}}, T(H+1)^2\} + \beta\right) \cdot 1/C_1 \\
&\leq 2000001\alpha/C_1 < 3000000\alpha/C_1
\end{aligned}
$$

Thus, we see that $|Y_i - \mathbb{E}[Y_i]| \leq 3000000\alpha/C_1$. On the other hand, we can see that if $Y_i = 0$ then $|\,\mathbb{E}_{i-1}[Y_i] - Y_i| = (f_h^1(z_h^{i-1}) - f_h^2(z_h^{i-1}))^2$ and the inequality we derived above still holds.

For the variance, we can see that

$$
\begin{aligned}
\text{Var}[Y_i - \mathbb{E}[Y_i|\mathbb{F}_i]|\mathbb{F}_i] &= p_{z_h}^{i-1} \left(\frac{1}{p_{z_h^{i-1}}} (f_h^1(z_h^{i-1}) - f_h^2(z_h^{i-1}))^2 \right)^2 + (1 - p_{z_h}^{i-1}) \cdot 0 \\
&\leq \frac{1}{p_{z_h^{i-1}}} (f_h^1(z_h^{i-1}) - f_h^2(z_h^{i-1}))^4 \\
&\leq 3000000\alpha \left(f_h^1(z_h^{i-1}) - f_h^2(z_h^{i-1})\right)^2 / C_1
\end{aligned}
$$

where the first equality follows from the definition, the first inequality is trivial and the third one from the inequality we derived above. Let k' be the maximum number $\leq k$ such that $Z_{k'} \leq 2000000\alpha$. Summing up the above inequalities for $i = 2, \ldots, k$ we get

$$
\begin{aligned}
\sum_{i=2}^{k} \text{Var}[Y_i - \mathbb{E}[Y_i|\mathbb{F}_i]|\mathbb{F}_i] &= \sum_{i=2}^{k'} \text{Var}[Y_i - \mathbb{E}[Y_i|\mathbb{F}_i]|\mathbb{F}_i] \\
&\leq \frac{3000000\alpha}{C_1} \sum_{i=2}^{k'} (f_h^1(z_h^{i-1}) - f_h^2(z_h^{i-1}))^2 \\
&\leq \frac{3000000\alpha \cdot 2000000\alpha}{C_1} \\
&\leq \frac{(3000000\alpha)^2}{C_1}
\end{aligned}
$$

where the the first equality follows from the fact that for $i > k'$ the random variable is deterministic, the first inequality follows by the summation of the previous one and the second one by the fact that $\sum_{i=2}^{k'} (f_h^1(z_h^{i-1}) - f_h^2(z_h^{i-1}))^2 \leq ||f_h^1 - f_h^2||_{\mathcal{Z}_h^{k'}} \leq Z_{k'}$.

We are now ready to use Freedman's inequality (Lemma C.6) with $R = \frac{3000000\alpha}{C_1}, \sigma^2 = \frac{(3000000\alpha)^2}{C_1}$. We get

$$\begin{aligned}
\Pr\left(\left|\sum_{i=1}^{k}(Y_i - \mathbb{E}[Y_i|\mathbb{F}_i])\right| \geq \alpha/100\right) &= \Pr\left(\left|\sum_{i=1}^{k'}(Y_i - \mathbb{E}[Y_i|\mathbb{F}_i])\right| \geq \alpha/100\right) \\
&\leq 2\exp\left\{-\frac{(\alpha/100)^2/2}{(3000000\alpha)^2/C_1 + \alpha^2 3000000/300C_1}\right\} \\
&= 2\exp\left\{-\frac{C_1}{20000(3000000+10000)}\right\} \\
&= 2\exp\left\{-\frac{C\log(T\mathcal{N}(\mathcal{F}_h, \sqrt{\delta/64T^3})/\delta)}{20000(3000000+10000)}\right\} \\
&= 2\exp\left\{-\frac{C(\log((T\mathcal{N}(\mathcal{F}_h, \sqrt{\delta/64T^3})/\delta)^2))}{40000(3000000+10000)}\right\} \\
&\leq (\delta/64T^2)/(\mathcal{N}(\mathcal{F}_h, \sqrt{\delta/64T^3}))^2
\end{aligned}$$

for some choice of C. Now we can take a union bound over all the functions in the discretized set and conclude that with probability at least $1 - \delta/(64T^2)$ we have that

$$\left|\sum_{i=1}^{k}(Y_i - \mathbb{E}[Y_i|\mathbb{F}_i])\right| \leq \alpha/100$$

for all pairs of functions in this set. We condition on this event and on the event in Lemma C.7. We first show that when this event happens, we have that $\underline{\mathcal{C}}_h^k(\alpha) \subseteq \widehat{\mathcal{C}}_h^k(\alpha)$. Consider $f_1, f_2 \in \underline{\mathcal{C}}_h^k(\alpha)$. We know that there exist $f_1', f_2' \in \mathcal{C}(\mathcal{F}, \sqrt{\delta/(64T^3)}) \times \mathcal{C}(\mathcal{F}, \sqrt{\delta/(64T^3)})$ with $||f_1 - f_1'||_\infty, ||f_2 - f_2'||_\infty \leq \sqrt{\delta/64T^3}$. Hence, we get that

$$\begin{aligned}
||f_1' - f_2'||^2_{\mathcal{Z}_h^k} &\leq \left(||f_1 - f_1'||_{\mathcal{Z}_h^k} + ||f_2 - f_2'||_{\mathcal{Z}_h^k} + ||f_1 - f_2||_{\mathcal{Z}_h^k}\right)^2 \\
&\leq \left(||f_1 - f_2||_{\mathcal{Z}_h^k} + 2\sqrt{\delta|\mathcal{Z}_h^k|/(64T^3)}\right)^2 \leq \alpha/50
\end{aligned}$$

We now consider the Y_i's that are generated by f_1', f_2'. It holds that $||f_1' - f_2'||^2_{\mathcal{Z}_h^k} \leq \alpha/50 \implies ||f_1' - f_2'||^2_{\mathcal{Z}_h^{k-1}} \leq \alpha/50$. Since the event $\mathcal{E}_h^{k-1}$ happens it follows that $\min\{||f_1' - f_2'||^2_{\widehat{\mathcal{Z}}_h^{k-1}}, T(H+1)^2\} \leq 100(\alpha/50) = 2\alpha < 2000000\alpha \implies Z_k \leq 2000000\alpha$. Thus, every Y_i is exactly $(f_1'(z_h^i) - f_2'(z_h^i))^2$ multiplied by the number of times z_h^i is in the sub-sampled dataset. Hence, we get

$$\begin{aligned}
||f_1' - f_2'||^2_{\widehat{\mathcal{Z}}_h^k} = \sum_{i=2}^{k} Y_i &\leq \sum_{i=2}^{k} \mathbb{E}[Y_i|\mathbb{F}_i] + \alpha/100 \\
&\leq ||f_1' - f_2'||^2_{\mathcal{Z}_h^k} + \alpha/100 \leq 3\alpha/100
\end{aligned}$$

where the first inequality follows from the concentration bound we have derived and the other two simply from the definitions of these quantities.

We now bound $||f_1 - f_2||^2_{\widehat{\mathcal{Z}}_h^k}$. We have that

$$\begin{aligned}
||f_1 - f_2||^2_{\widehat{\mathcal{Z}}_h^k} &\leq \left(||f_1' - f_2'||_{\widehat{\mathcal{Z}}_h^k} + ||f_1 - f_1'||_{\widehat{\mathcal{Z}}_h^k} + ||f_2 - f_2'||_{\widehat{\mathcal{Z}}_h^k}\right)^2 \\
&\leq (||f_1' - f_2'||_{\widehat{\mathcal{Z}}_h^k} + 2\sqrt{|\widehat{\mathcal{Z}}_h^k|} \cdot \sqrt{\delta/(64T^3)})^2 \\
&\leq (||f_1' - f_2'||_{\widehat{\mathcal{Z}}_h^k} + 2)^2 \leq (\sqrt{3\alpha/100} + 2)^2 \leq \alpha
\end{aligned}$$

Hence, we have shown that $\underline{\mathcal{C}}_h^k(\alpha) \subseteq \widehat{\mathcal{C}}_h^k(\alpha)$. So in this case, the one inequality that define $\mathcal{E}_h^k$ holds.

We shift our attention to the second inequality now. We will show the contrapositive of our claim, i.e. if $f_1, f_2 \notin \overline{C}_h^k(\alpha) \implies f_1, f_2 \notin \widehat{C}_h^k(\alpha)$. Let $f_1, f_2 \in \mathcal{F}_h \times \mathcal{F}_h$ such that $||f_1 - f_2||_{\mathcal{Z}_h^k} > 100\alpha$. We know that there exist $f_1', f_2' \in \mathcal{C}(\mathcal{F}, \sqrt{\delta/(64T^3)}) \times \mathcal{C}(\mathcal{F}, \sqrt{\delta/(64T^3)})$ with $||f_1 - f_1'||_\infty, ||f_2 - f_2'||_\infty \leq \sqrt{\delta/64T^3}$. Hence, using the triangle inequality we get that

$$\begin{aligned}
||f_1' - f_2'||^2_{\mathcal{Z}_h^k} &\geq (||f_1 - f_2||_{\mathcal{Z}_h^k} - ||f_1 - f_1'||_{\mathcal{Z}_h^k} - ||f_2 - f_2'||_{\mathcal{Z}_h^k})^2 \\
&\geq (||f_1 - f_2||_{\mathcal{Z}_h^k} - 2\sqrt{|\mathcal{Z}_h^k|}\sqrt{\delta/(64T^3)})^2 \\
&= (\sqrt{100\alpha} - 2\sqrt{\delta/(64T^2)})^2 > 50\alpha
\end{aligned}$$

Again, consider the Y_i's that are generated by f_1', f_2'. We want to show that $||f_1' - f_2'||^2_{\widehat{\mathcal{Z}}_h^k} > 40\alpha$. Assume towards contradiction that $||f_1' - f_2'||^2_{\widehat{\mathcal{Z}}_h^k} \leq 40\alpha$. We consider three different cases.

First Case: $||f_1' - f_2'||^2_{\mathcal{Z}_h^k} \leq 2000000\alpha$. Similarly as before, we have that

$$\begin{aligned}
||f_1' - f_2'||^2_{\widehat{\mathcal{Z}}_h^k} = \sum_{i=2}^{k} Y_i &\geq \mathbb{E}[Y_i|\mathbb{F}_i] - \alpha/100 \\
&> 50\alpha - \alpha/100 > 40\alpha
\end{aligned}$$

So we get a contradiction.

Second Case: $||f_1' - f_2'||^2_{\mathcal{Z}_h^{k-1}} > 10000\alpha$. The contradiction comes directly from the fact that $\mathcal{E}_h^{k-1}$ holds, so

$$||f_1' - f_2'||^2_{\widehat{\mathcal{Z}}_h^k} \geq ||f_1' - f_2'||^2_{\widehat{\mathcal{Z}}_h^{k-1}} > 100\alpha$$

.

Third Case: $||f_1' - f_2'||^2_{\mathcal{Z}_h^{k-1}} \leq 10000\alpha$ **and** $||f_1' - f_2'||^2_{\mathcal{Z}_h^k} > 2000000\alpha$. We can directly see that for this case $(f_1'(z_h^k) - f_2'(z_h^k))^2 \geq 1900000\alpha$. Since $||f_1' - f_2'||^2_{\mathcal{Z}_h^{k-1}} \leq 10000\alpha \implies ||f_1' - f_2'||^2_{\widehat{\mathcal{Z}}_h^{k-1}} \leq 1000000\alpha$. Thus, since $\alpha \geq \beta$ we can see that the sensitivity is 1 so the element will be added to the sub-sampled dataset. Hence, $||f_1' - f_2'||^2_{\mathcal{Z}_h^k} \geq (f_1'(z_h^k) - f_2'(z_h^k))^2 > 40\alpha$.

Thus, in any case we have that $||f_1' - f_2'||^2_{\widehat{\mathcal{Z}}_h^k} > 40\alpha > \alpha$, so we get the result. □

We are now ready to prove Lemma C.5.

Proof of Lemma C.5:

We know that for all $k \in [K], k \neq 1, h \in [H]$ it holds that

$$\begin{aligned}
\Pr(\mathcal{E}_h^1\mathcal{E}_h^2 \dots \mathcal{E}_h^{k-1}) - \Pr(\mathcal{E}_h^1\mathcal{E}_h^2 \dots \mathcal{E}_h^k) &= \Pr\left(\mathcal{E}_h^1\mathcal{E}_h^2 \dots \mathcal{E}_h^{k-1}(\mathcal{E}_h^k)^c\right) \\
&= \Pr\left(\mathcal{E}_h^1\mathcal{E}_h^2 \dots \mathcal{E}_h^{k-1}\left(\cap_{n=0}^{\infty}\mathcal{E}_h^k(100^n\beta)\right)^c\right) \\
&= \Pr\left(\mathcal{E}_h^1\mathcal{E}_h^2 \dots \mathcal{E}_h^{k-1} \cup_{n=0}^{\infty} \mathcal{E}_h^k(100^n\beta)^c\right) \\
&\leq \sum_{n=0}^{\infty} \Pr\left(\mathcal{E}_h^1\mathcal{E}_h^2 \dots \mathcal{E}_h^{k-1}(\mathcal{E}_h^k(100^n\beta))^c\right) \\
&= \sum_{n \geq 0, 100^n\beta \leq T(H+1)^2} \Pr\left(\mathcal{E}_h^1\mathcal{E}_h^2 \dots \mathcal{E}_h^{k-1}(\mathcal{E}_h^k(100^n\beta))^c\right).
\end{aligned}$$

Thus, using Lemma C.9 we see that $\Pr(\mathcal{E}_h^1\mathcal{E}_h^2 \dots \mathcal{E}_h^{k-1}) - \Pr(\mathcal{E}_h^1\mathcal{E}_h^2 \dots \mathcal{E}_h^k) \leq \delta/(32T^2)(\log(T(H+1)^2/\beta) + 2) \leq \delta/32T$.

Hence, for any fixed $h \in [H]$ we get

$$\Pr\left(\bigcap_{k=1}^{K} \mathcal{E}_h^k\right) = 1 - \sum_{k=1}^{K} \left(\Pr(\mathcal{E}_h^1 \mathcal{E}_h^2 \dots \mathcal{E}_h^{k-1}) - \Pr(\mathcal{E}_h^1 \mathcal{E}_h^2 \dots \mathcal{E}_h^k)\right)$$
$$\geq 1 - K(\delta/32T) = 1 - \delta/(32H)$$

and by taking a union bound over $h \in [H]$ we get the result. □

Now that we have shown that the sub-sampled dataset approximates well the original one, we shift our attention back to showing that our approach achieves optimism.

We first need a definition and a concetration lemma that is related to least-squares-estimators from prior work.

Definition C.10. A random variable X is conditionally σ-subgaussian with respect to some filtration $\mathbb{F}$ if for all $\lambda \in \mathbb{R}$ it holds that $\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2 \sigma^2/2)$.

Lemma C.11 ([RVR13], [AJS+20]). *Let $\mathbb{F} = \{\mathbb{F}_p\}_{p=0,1,\dots}$ be a filtration, $\{(X_p, Y_p)\}_p$ measurable random variables where $X_p \in \mathcal{X}, Y_p \in \mathbb{R}$. Let $\widetilde{\mathcal{F}}$ be a set of measurable functions from $\mathcal{X}$ to $\mathbb{R}$ and assume that $\mathbb{E}[Y_p|\mathbb{F}_{p-1}] = f^*(X_p)$ for some $f^* \in \widetilde{\mathcal{F}}$. Assume that $\{Y_p - f^*(X_p)\}_{p=1,\dots}$ is conditionally σ-subgaussian given $\mathbb{F}_{p-1}$. Let $\widehat{f}_t = \arg\min_{f \in \widetilde{\mathcal{F}}} \sum_{p=1}^{t} \left(f(X_p) - Y_p\right)^2$ and $\widetilde{\mathcal{F}}_t(\beta) = \left\{ f \in \widetilde{F} : \sum_{p=1}^{t} \left(f(X_p) - \widehat{f}(X_p))\right)^2 \leq \beta \right\}$. Then, for any $\alpha > 0$, with probability $1 - \delta$, for all $t \geq 1$ it holds that $f^* \in \widetilde{\mathcal{F}}_t(\beta_t(\delta, \alpha))$, where*

$$\beta_t(\delta, \alpha) = 8\sigma^2 \log(2\mathcal{N}(\widetilde{\mathcal{F}}, \alpha)/\delta) + 4t\alpha \left(C + \sqrt{\sigma^2 \log(4t(t+1)/\delta)}\right).$$

We are now ready to prove that our algorithm ensures optimism.

Lemma C.12. *With probability at least $1 - 2\delta$, we have that for all $h \in [H], k \in [K], s \in \mathcal{S}, a \in \mathcal{A}$*

$$Q_h^k(s,a) - Q_h^{\pi_k}(s,a) \leq \langle P_h(\cdot|s,a), V_{h+1}^k(\cdot) - V_{h+1}^{\pi_k}(\cdot)\rangle + 2b_h^k(s,a).$$

Moreover, it holds that $Q_h^k(s,a) \geq Q_h^*(s,a)$.

Proof. Fix some $h \in [H], k \in [K], s \in \mathcal{S}, a \in \mathcal{A}$. Throughout the proof, we condition on the events in Lemma C.5 and Lemma C.11. We assume that k is a round that we perform an update.

We define $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{V}, X_h^k = (s_h^k, a_h^k, V_{h+1}^k(\cdot)), Y_h^k = V_{h+1}^k(s_{h+1}^k)$. We also pick $\widetilde{\mathcal{F}} = \mathcal{F}_h$, where $\mathcal{F}_h$ is defined in Definition C.4. Then, we see that $\mathbb{E}[Y_h^k|\mathbb{F}_{k-1}] = f_h^*(X_h^k)$, where f_h^* is the function that corresponds to the true model P_h, and we know that $f_h^* \in \mathcal{F}_h$. Recall that the optimization problem we solve in Algorithm 1 for every round k we update our policy is

$$\widehat{P}_h^k = \arg\min_{P \in \mathcal{P}_h} \sum_{p=1}^{k} \left(\langle P(\cdot|s_h^p, a_h^p), V_{h+1}^p\rangle - V_{h+1}^p(s_{h+1}^p)\right)^2.$$

Based on the definition of $\mathcal{F}_h$, we can see that this is equivalent to

$$f_h^k = \arg\min_{f \in \mathcal{F}_h} \sum_{p=1}^{k} \left(f(s_h^p, a_h^p, V_{h+1}^p) - V_{h+1}^p(s_{h+1}^p)\right)^2.$$

Moreoever, $Y_h^k \in [0, H]$, so $Z_h^k = Y_h^k - f_h^*(X_h^k)$ is $H/2$-conditionally subgaussian. Thus, if we pick $\alpha = 1/T$ and

$$\beta_h^k = 4H^2 \log(2\mathcal{N}(\mathcal{F}_h, 1/T)/\delta) + 4k/T \left(C + \sqrt{H^2/4 \log(4(k(k+1)/\delta))}\right)$$

then Lemma C.11 gives us that $||f_h^* - \widehat{f}_h^k||_{\mathcal{Z}_h^k}^2 \leq \beta_h^k$. In particular, can pick

$$\widetilde{\beta} = \beta_h^K = 4H^2 \log(2\mathcal{N}(\mathcal{F}_h, 1/T)/\delta) + 4K/T \left(C + \sqrt{H^2/4 \log(4(K(K+1)/\delta))}\right)$$

and get a parameter that is independent of k. Moreover, Lemma C.5

$$||f_h^* - \widehat{f}_h^k||_{\widehat{\mathcal{Z}}_h^k} \leq 100\widetilde{\beta} = \beta.$$

This implies that for our bonus function we have that $|\widehat{f}_h^k(s, a, V_{h+1}^k(\cdot)) - f_h^*(s, a, V_{h+1}^k(\cdot))| \leq b_h^k(s, a)$.

Hence, we have that

$$\begin{aligned}\langle \widehat{P}_h^k(\cdot|s,a), V_{h+1}^k(\cdot)\rangle - \langle P_h(\cdot|s,a), V_{h+1}^k(\cdot)\rangle &= \widehat{f}_h^k(s, a, V_{h+1}^k) - f_h^*(s, a, V_{h+1}^k) \\ &\leq |\widehat{f}_h^k(s, a, V_{h+1}^k) - f_h^*(s, a, V_{h+1}^k)| \\ &\leq b_h^k(s,a).\end{aligned}$$

Now we use the definition of $Q_h^k(s,a), Q_h^{\pi_k}(s,a)$ to get that

$$\begin{aligned}Q_h^k(s,a) &\leq r_h^k(s,a) + \langle \widehat{P}_h^k(\cdot|s,a), V_{h+1}^k(\cdot)\rangle + b_h^k(s,a) \\ &\leq r_h^k(s,a) + \langle P_h(\cdot|s,a), V_{h+1}^k(\cdot)\rangle + 2b_h^k(s,a) \\ Q_h^{\pi_k}(s,a) &= r_h^k(s,a) + \langle P_h(\cdot|s,a), V_{h+1}^{\pi_k}(\cdot)\rangle.\end{aligned}$$

Combining these two, we get that

$$Q_h^k(s,a) - Q_h^{\pi_k}(s,a) \leq \langle P_h(\cdot|s,a), V_{h+1}^k(\cdot) - V_{h+1}^{\pi_k}(\cdot)\rangle + 2b_h^k(s,a)$$

which proves the first part of the result.

For the second part, notice that if $Q_h^k(s,a) = H$ then the statement holds trivially since $Q_h^*(s,a) \leq H$. So we can assume without loss of generality that $Q_h^k(s,a) = r_h^k(s,a) + \langle \widehat{P}_h^k(\cdot|s,a), V_{h+1}^k(\cdot)\rangle + b_h^k(s,a)$. The Bellman optimality condition gives us that $Q_h^*(s,a) = r_h^k(s,a) + \langle P_h^*(\cdot|s,a), V_{h+1}^*(\cdot)\rangle$. Hence, we have that

$$\begin{aligned}Q_h^k(s,a) - Q_h^*(s,a) &= \langle \widehat{P}_h^k(\cdot|s,a), V_{h+1}^k(\cdot)\rangle - \langle P_h(\cdot|s,a), V_{h+1}^*(\cdot)\rangle + b_h^k(s,a) \\ &= \langle \widehat{P}_h^k(\cdot|s,a), V_{h+1}^k(\cdot)\rangle - \langle P_h(\cdot|s,a), V_{h+1}^k(\cdot)\rangle + \langle P_h(\cdot|s,a), V_{h+1}^k(\cdot)\rangle \\ &\quad - \langle P_h(\cdot|s,a), V_{h+1}^*(\cdot)\rangle + b_h^k(s,a) \\ &= \langle \widehat{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a), V_{h+1}^k(\cdot)\rangle + \langle P_h(\cdot|s,a), V_{h+1}^k(\cdot) - V_{h+1}^*(\cdot)\rangle + b_h^k(s,a).\end{aligned}$$

Now from our previous discussion it follows that $b_h^k(s,a) + \langle \widehat{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a), V_{h+1}^k(\cdot)\rangle \geq 0$. Hence, it suffices to show that $\langle P_h(\cdot|s,a), V_{h+1}^k(\cdot) - V_{h+1}^*(\cdot)\rangle \geq 0$. To do that, we can just prove that $V_{h+1}^k(s') - V_{h+1}^*(s') \geq 0, \forall s' \in \mathcal{S}$. Since $V_{H+1}^k(s') = V_{H+1}^*(s') = 0, \forall s' \in \mathcal{S}$ we get that $Q_H^k(s,a) \geq Q_H^*(s,a)$. Thus, if we combine this with the update rule for V_h^k, V_h^* we get the claim by induction. □

Now that we have established the previous lemma, we need to bound the bonus that we are using in every round. The issue is that we do not update our policy in every round.

To do that, we follow a similar approach as in Appendix B.

Lemma C.13. *For every set $K' \subseteq [K]$ With probability at least $1-\delta$, we have that*

$$\sum_{i=1}^{|K'|}\sum_{h=1}^{H} b_h^{k_i}(s_h^{k_i}, a_h^{k_i}) \leq H + H(H+1)\dim_E(\mathcal{F}, 1/T) + CH\sqrt{\dim_E(\mathcal{F}, 1/T)|K'|\beta}$$

where $\dim_E(\mathcal{F}, 1/T) = \max_{h\in[H]} \dim_E(\mathcal{F}_h, 1/T)$.

Proof. We condition on the event described in Lemma C.5. From the definition of the bonus function, we have that for any $k \in [K]$

$$b_h^k(s_h^k, a_h^k) \leq \bar{b}_h^k(s_h^k, a_h^k) = \sup_{||f_1 - f_2||^2_{\mathcal{Z}_h^k} \leq 100\beta} |f_1(s_h^k, a_h^k) - f_2(s_h^k, a_h^k)|.$$

We bound $\sum_{i=1}^{K'} \bar{b}_h^{k_i}(s_h^{k_i}, a_h^{k_i})$ for each $h \in [H]$ separately.

Given some $\epsilon > 0$, we define $\widetilde{K} = \{k \in K' : \bar{b}_h^k(s_h^k, a_h^k) > \epsilon\}$, i.e. the set of episodes in K' where the bonus function at h has value greater than ϵ. Consider some $k \in K'$. We denote $\mathcal{L}_h = \{(s_h^k, a_h^k, V_{h+1}^k(\cdot)) : k \in \widetilde{K}\}, L_h = |\mathcal{L}_h|$, and $N = L_h/\dim_E(\mathcal{F}_h, \epsilon) - 1$. Our goal is to show that there is some $z_h^k = (s_h^k, a_h^k, V_{h+1}^k(\cdot)) \in \mathcal{L}_h$ that is ϵ-dependent on at least N disjoint subsequences in $\mathcal{Z}_h^k \cap \mathcal{L}_h$.

To do that, we decompose $\mathcal{L}_h$ into $N+1$ disjoint subsets and we denote the j-th subset by $\mathcal{L}_{h,j}$. We use the following procedure. Initially we set $\mathcal{L}_{h,j} = \emptyset$ for all $j \in [N+1]$ and consider every $z_h^k \in \mathcal{L}_h$ in a sequential manner. For each such z_h^k we find the smallest index $j, 1 \leq j \leq N$, such that z_h^k is ϵ-independent of the elements in $\mathcal{L}_{h,j}$ with respect to $\mathcal{F}_h$. If there is no such j, we set $j = N+1$. Then, we update $\mathcal{L}_{h,j} \leftarrow \mathcal{L}_{h,j} \cup z_h^k$. Notice that after we go through all the elements of $\mathcal{L}_h$, we must have that $\mathcal{L}_{h,N+1} \neq \emptyset$. This is because every set $\mathcal{L}_{h,j}, 1 \leq j \leq N$, contains at most $\dim_E(\mathcal{F}_h, \epsilon)$ elements. Moreover, by definition, every element $z_h^k \in \mathcal{L}_{h,N+1}$ is ϵ-dependent on at least N disjoint subsequences in $\mathcal{L}_h$.

Furthermore, since $\bar{b}_h^k(s_h^k, a_h^k) > \epsilon$ for all $z_h^k \in \mathcal{L}_h$ there must exist $f_1, f_2 \in \mathcal{F}_h$ such that $|f_1(s_h^k, a_h^k, V_{h+1}^k(\cdot)) - f_2(s_h^k, a_h^k, V_{h+1}^k(\cdot))| > \epsilon$ and $||f_1 - f_2||_{\mathcal{Z}_h^k}^2 \leq 100\beta$. Hence, since $z_h^k \in \mathcal{L}_{h,N+1}$ is ϵ-dependent on N disjoint subsequences $\mathcal{L}_h$ and for each such subsequence $\mathcal{L}$, by the definition of ϵ-dependence, it holds that $||f_1 - f_2||_{\mathcal{L}}^2 > \epsilon^2$ we have that

$$
\begin{aligned}
N\epsilon^2 &\leq ||f_1 - f_2||_{\mathcal{Z}_h^k}^2 \leq 100\beta \\
\implies (L_h/\dim_E(\mathcal{F}_h, \epsilon) - 1)\epsilon^2 &\leq 100\beta \\
\implies L_h &\leq \left(\frac{100\beta}{\epsilon^2} + 1\right) \dim_E(\mathcal{F}_h, \epsilon).
\end{aligned}
$$

We now pick a permutation $\bar{b}_1 \geq \bar{b}_2 \geq \ldots \geq \bar{b}_{|K'|}$ of the bonus functions $\{\bar{b}_h^k(s_h^k, a_h^k)\}_{k \in K'}$. For all $\bar{b}_k \geq 1/|K'|$ it holds that

$$
\begin{aligned}
k \leq \left(\frac{100\beta}{\bar{b}_k^2} + 1\right) \dim_E(\mathcal{F}_h, \bar{b}_k) \leq \left(\frac{100\beta}{\bar{b}_k^2} + 1\right) \dim_E(\mathcal{F}_h, 1/K') \implies \\
\bar{b}_k \leq \left(\frac{k}{\dim_E(\mathcal{F}_h, 1/K')} - 1\right)^{-1/2} \sqrt{100\beta}.
\end{aligned}
$$

Moreover, notice that we get by definition that $\bar{b}_k \leq H + 1$. Hence, we have that

$$
\begin{aligned}
\sum_{i=1}^{|K'|} \bar{b}_h^{k_i}(s_h^{k_i}, a_h^{k_i}) &= \sum_{i:\bar{b}_{k_i} < 1/|K'|} \bar{b}_h^{k_i}(s_h^{k_i}, a_h^{k_i}) + \sum_{i:\bar{b}_{k_i} \geq 1/|K'|} \bar{b}_h^{k_i}(s_h^{k_i}, a_h^{k_i}) \\
&\leq |K'| \cdot 1/|K'| + \sum_{i:\bar{b}_{k_i} \geq 1/|K'|, i \leq \dim_E(\mathcal{F}_h, 1/|K'|)} \bar{b}_h^{k_i}(s_h^{k_i}, a_h^{k_i}) \\
&+ \sum_{i:\bar{b}_{k_i} \geq 1/|K'|, |K'| \geq i > \dim_E(\mathcal{F}_h, 1/|K'|)} \bar{b}_h^{k_i}(s_h^{k_i}, a_h^{k_i}) \\
&\leq 1 + (H+1)\dim_E(\mathcal{F}_h, 1/|K'|) + \sum_{|K'| \geq i > \dim_E(\mathcal{F}_h, 1/|K'|)} \left(\frac{k}{\dim_E(\mathcal{F}_h, 1/|K'|)} - 1\right)^{-1/2} \sqrt{100\beta} \\
&\leq 1 + (H+1)\dim_E(\mathcal{F}_h, 1/|K'|) + C\sqrt{\dim_E(\mathcal{F}_h, 1/|K'|)|K'|\beta} \\
&\leq 1 + (H+1)\dim_E(\mathcal{F}_h, 1/T) + C\sqrt{\dim_E(\mathcal{F}_h, 1/T)|K'|\beta}
\end{aligned}
$$

for some constant $C > 0$, where the second to last inequality can be obtained by bounding the summation by the integral and the last one by the definition of the eluder dimension. We get the final result by summing up all the inequalities over H. $\square$

The next step in our proof, is to bound the number of episodes that our policy can be worse than the optimal one by $2^n \text{gap}_{\min}$, for all $n \in \mathbb{N}$. This is inspired by [HZG21].

Lemma C.14. *If we pick*

$$\beta = 4H^2 \log(2\mathcal{N}(\mathcal{F}, 1/T)/\delta) + 4/H \left(C + \sqrt{H^2/4 \log(4(K(K+1)/\delta))}\right),$$

then for every $h \in [H]$ and $n \in \mathbb{N}$, with probability at least $1 - (K+3)\delta$, we have that

$$\sum_{k=1}^{K} \mathbb{1}\left[V_h^*(s_h^k) - Q_h^{\pi_k}(s_h^k, a_h^k) \geq 2^n \text{gap}_{\min}\right] \leq \frac{\widetilde{C} H^4 \log(T\mathcal{N}(\mathcal{F}_h, 1/T)/\delta) \dim_E^2(\mathcal{F}, 1/T)}{4^n \text{gap}_{\min}^2}.$$

Proof. We keep h fixed.

We denote by K' the set of episodes where the gap at step h is at least 2^n, i.e.

$$K' = \left\{k \in [K] : V_h^*(s_h^k) - Q_h^{\pi_k}(s_h^k, a_h^k) \geq 2^n \text{gap}_{\min}\right\}.$$

The goal is to bound the quantity $\sum_{i=1}^{|K'|} \left(Q_h^{k_i}(s_h^{k_i}, a_h^{k_i}) - Q_h^{\pi_{k_i}}(s_h^{k_i}, a_h^{k_i})\right)$ from below and above with functions $f_1(|K'|), f_2(|K'|)$ and then use the fact that $f_1(|K'|) \leq f_2(|K'|)$ to derive an upper bound on $|K'|$.

For the lower bound, we have that

$$\begin{aligned}
\sum_{i=1}^{|K'|} \left(Q_h^{k_i}(s_h^{k_i}, a_h^{k_i}) - Q_h^{\pi_{k_i}}(s_h^{k_i}, a_h^{k_i})\right) &\geq \sum_{i=1}^{|K'|} \left(Q_h^{k_i}(s_h^{k_i}, \pi_h^*(s_h^{k_i})) - Q_h^{\pi_{k_i}}(s_h^{k_i}, a_h^{k_i})\right) \\
&\geq \sum_{i=1}^{|K'|} \left(Q_h^{*}(s_h^{k_i}, \pi_h^*(s_h^{k_i})) - Q_h^{\pi_{k_i}}(s_h^{k_i}, a_h^{k_i})\right) \\
&= \sum_{i=1}^{|K'|} \left(V_h^{*}(s_h^{k_i}) - Q_h^{\pi_{k_i}}(s_h^{k_i}, a_h^{k_i})\right) \\
&\geq 2^n \text{gap}_{\min} |K'|
\end{aligned}$$

where the first inequality holds by the definition of the policy π_{k_i}, the second one follows because $Q_h^{k_i}(\cdot,\cdot)$ is an optimistic estimate of $Q_h^*(\cdot,\cdot)$ which happens with probability at least $1 - 2\delta$ (see Lemma C.12) and the third one by the definition of k_i.

We get the upper bound on this quantity in the following way. For any $h' \in [H]$ we have

$$\begin{aligned}
Q_{h'}^k(s_{h'}^k, a_{h'}^k) - Q_{h'}^{\pi_k}(s_{h'}^k, a_{h'}^k) &\leq \sum_{s' \in \mathcal{S}} P_{h'}(s'|s_{h'}^k, a_{h'}^k) V_{h'+1}^k(s') + r_{h'}(s_{h'}^k, a_{h'}^k) + 2b_{h'}^k(s_{h'}^k, a_{h'}^k) - Q_{h'}^{\pi_k}(s_{h'}^k, a_{h'}^k) \\
&= \left\langle P_{h'}(\cdot|s_{h'}^k, a_{h'}^k), V_{h'+1}^k - V_{h'+1}^{\pi_k} \right\rangle + 2b_{h'}^k(s_{h'}^k, a_{h'}^k) \\
&= V_{h'+1}^k(s_{h'+1}^k) - V_{h'+1}^{\pi_k}(s_{h'+1}^k) + \epsilon_{h'}^k + 2b_{h'}^k(s_{h'}^k, a_{h'}^k) \\
&= Q_{h'+1}^k(s_{h'+1}^k, a_{h'+1}^k) - Q_{h'+1}^{\pi_k}(s_{h'+1}^k, a_{h'+1}^k) + \epsilon_{h'}^k + 2b_{h'}^k(s_{h'}^k, a_{h'}^k)
\end{aligned}$$

where we define $\epsilon_{h'}^k = \left\langle P_{h'}(\cdot|s_{h'}^k, a_{h'}^k), V_{h'+1}^k - V_{h'+1}^{\pi_k} \right\rangle - \left(V_{h'+1}^k(s_{h'+1}^k) - V_{h'+1}^{\pi_k}(s_{h'+1}^k)\right)$ and the inequality follows from Lemma C.12.

We now take the summation over all $k \in |K'|, h \leq h' \leq H$ and we get

$$\sum_{i=1}^{|K'|} \sum_{h'=h}^{H} \left(Q_h^{k_i}(s_h^{k_i}, a_h^{k_i}) - Q_h^{\pi_{k_i}}(s_h^{k_i}, a_h^{k_i})\right) \leq \sum_{i=1}^{|K'|} \sum_{h'=h}^{H} \epsilon_{h'}^{k_i} + \sum_{i=1}^{|K'|} \sum_{h'=h}^{H} b_{h'}^{k_i}(s_{h'}^{k_i}, a_{h'}^{k_i}).$$

We will bound each of the two terms on the RHS separately.

For the first term, we notice that $x_j = \left\langle P_j(\cdot|s_j^{k_i}, a_j^{k_i}), V_{j+1}^{k_i} - V_{j+1}^{\pi_{k_i}} \right\rangle - \left(V_{j+1}^{k_i}(s_{j+1}^{k_i}) - V_{j+1}^{\pi_{k_i}}(s_{j+1}^{k_i})\right)$ forms a martingale difference sequence with zero mean and

$|x_j| \leq 2H$. Hence, we can use Lemma B.7 and that for each $k \in K'$, with probability at least $1-\delta$ we have that

$$\sum_{i=1}^{k}\sum_{j=1}^{H}\left(\left\langle P_j(\cdot|s_j^{k_i}, a_j^{k_i}), V_{j+1}^{k_i} - V_{j+1}^{\pi_{k_i}}\right\rangle - \left(V_{j+1}^{k_i}(s_{j+1}^{k_i}) - V_{j+1}^{\pi_{k_i}}(s_{j+1}^{k_i})\right)\right) \leq \sqrt{8kH^3\log(1/\delta)}.$$

If we take the union bound over all $k \in [K]$ we have that with probability at least $1-|K'|\delta$

$$\sum_{i=1}^{|K'|}\sum_{h'=h}^{H}\epsilon_{h'}^{k_i} \leq \sqrt{8|K'|H^3\log(1/\delta)}.$$

We now focus on the second term. Using Lemma C.13 we get that

$$\sum_{i=1}^{|K'|}\sum_{h'=h}^{H} b_{h'}^{k_i}(s_{h'}^{k_i}, a_{h'}^{k_i}) \leq H + H(H+1)\dim_E(\mathcal{F}, 1/T) + CH\sqrt{\dim_E(\mathcal{F}, 1/T)|K'|\beta}$$

and this happens with probability at least $1-\delta$. Hence, combining the upper and lower bound of

$$\sum_{i=1}^{|K'|}\left(Q_h^{k_i}(s_h^{k_i}, a_h^{k_i}) - Q_h^{\pi_{k_i}}(s_h^{k_i}, a_h^{k_i})\right)$$

we get that

$$2^n\text{gap}_{\min}|K'| \leq \sqrt{8|K'|H^3\log(1/\delta)} + H + H(H+1)\dim_E(\mathcal{F}, 1/T) + CH\sqrt{\dim_E(\mathcal{F}, 1/T)|K'|\beta}$$

$$\implies |K'| \leq \frac{\widetilde{C}H^4\log(T\mathcal{N}(\mathcal{F}, 1/T)/\delta)\dim_E^2(\mathcal{F}, 1/T)}{4^n\text{gap}_{\min}^2}.$$

□

We are now ready to prove Lemma C.1.

Proof of Lemma C.1: Throughout this proof we condition on the event described in Lemma C.14 which happens with probability at least $1-(K+3)\delta$. Since $\text{gap}_{\min} > 0$ whenever we do not take the optimal action, we have that either $V_h^*(s_k) - Q_h^*(s_h^k, a_h^k) = 0$ or $V_h^*(s_k) - Q_h^*(s_h^k, a_h^k) \geq \text{gap}_{\min}$. Our approach is to divide the interval $[0, H]$ into $N = \lceil\log(H/\text{gap}_{\min})\rceil$ intervals and count the number of $V_h^*(s_k) - Q_h^*(s_h^k, a_h^k)$ that fall into each interval. Notice that for every $V_h^*(s_k) - Q_h^*(s_h^k, a_h^k)$ that falls into interval i we can get an upper bound of $V_h^*(s_k) - Q_h^*(s_h^k, a_h^k) \leq 2^i\text{gap}_{\min}$ and this upper bound is essentially tight. Hence, we have that

$$\begin{aligned}
\sum_{k=1}^{K}\left(V_h^*(s_h^k) - Q_h^*(s_h^k, a_h^k)\right) &\leq \sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{1}\left[2^i\text{gap}_{\min} \geq V_h^*(s_h^k) - Q_h^*(s_h^k, a_h^k) \geq 2^{i-1}\text{gap}_{\min}\right]\cdot 2^i\text{gap}_{\min} \\
&\leq \sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{1}\left[V_h^*(s_h^k) - Q_h^*(s_h^k, a_h^k) \geq 2^{i-1}\text{gap}_{\min}\right]\cdot 2^i\text{gap}_{\min} \\
&\leq \sum_{i=1}^{N}\frac{\widetilde{C}H^6\log(T\mathcal{N}(\mathcal{F}, 1/T)/\delta)\dim_E^2(\mathcal{F}, 1/T)}{4^{i-1}\text{gap}_{\min}^2}\cdot 2^i\text{gap}_{\min} \\
&= \sum_{i=1}^{N}\frac{C'H^4\log(T\mathcal{N}(\mathcal{F}, 1/T)/\delta)\dim_E^2(\mathcal{F}, 1/T)}{2^i\text{gap}_{\min}} \\
&\leq \frac{CH^4\log(T\mathcal{N}(\mathcal{F}, 1/T)/\delta)\dim_E^2(\mathcal{F}, 1/T)}{\text{gap}_{\min}}
\end{aligned}$$

where the first inequality holds by the definition of the intervals, the second due to the properties of the indicator function, the third because of Lemma C.14 and in the last two steps we just manipulate the constants. □