Online Sub-Sampling for Reinforcement Learning with General Function Approximation

Dingwen Kong [*1] **Ruslan Salakhutdinov** [*2] **Ruosong Wang** [*2] **Lin F. Yang** [*3]

Abstract

Designing provably efficient algorithms with general function approximation is an important open problem in reinforcement learning. Recently, Wang et al. [2020c] establish a value-based algorithm with general function approximation that enjoys $\widetilde{O}(\mathrm{poly}(dH)\sqrt{K})$[1] regret bound, where d depends on the complexity of the function class, H is the planning horizon, and K is the total number of episodes. However, their algorithm requires $\Omega(K)$ computation time per round, rendering the algorithm inefficient for practical use. In this paper, by applying online sub-sampling techniques, we develop an algorithm that takes $\widetilde{O}(\mathrm{poly}(dH))$ computation time per round on average, and enjoys nearly the same regret bound. Furthermore, the algorithm achieves low switching cost, i.e., it changes the policy only $\widetilde{O}(\mathrm{poly}(dH))$ times during its execution, making it appealing to be implemented in real-life scenarios. Moreover, by using an upper-confidence based exploration-driven reward function, the algorithm provably explores the environment in the reward-free setting. In particular, after $\widetilde{O}(\mathrm{poly}(dH))/\epsilon^2$ rounds of exploration, the algorithm outputs an ϵ-optimal policy for any given reward function.

1. Introduction

Function approximation (FA) is one of the key techniques to scale up reinforcement learning (RL) in real-world applications. FA methods have achieved phenomenal empirical success (Mnih et al., 2013; 2015; Silver et al., 2017; Vinyals et al., 2019; Akkaya et al., 2019), where in these applications, RL agents learn to control complex systems by approximating value functions using deep neural networks. In contrast, the theoretical understanding of RL with general FA is still rudimentary, e.g., reasonable theoretical understandings have only been established in the tabular setting or the linear setting (Azar et al., 2017; Jin et al., 2018; Yang & Wang, 2020; Jin et al., 2020b). Designing RL algorithms with general FA with provable efficiency becomes increasingly important as it helps understanding the limits of existing algorithms while inspiring the design of better practical algorithms.

Recently, Wang et al. (2020c) establish a provably efficient algorithm with general function approximation that achieves a regret bound of $\widetilde{O}(\mathrm{poly}(dH)\sqrt{K})$, where d depends on the eluder dimension (Russo & Roy, 2013) and log-covering numbers of the function class, H is the planning horizon, and K is the total number of episodes. Here the regret measures the difference between the expected rewards collected by the optimal policy and that of the RL algorithm. Such a regret bound indicates that the algorithm learns a policy with suboptimality at most ϵ after interacting with the environment for $\widetilde{O}(\mathrm{poly}(dH))/\epsilon^2$ episodes. Their algorithm is based on least-squares value iteration (LSVI), and to balance exploration and exploitation, their algorithm employs the principle of "optimism in the face of uncertainty" and adds an exploration bonus to the learned value function. The main technical innovation in their paper is a stable exploration bonus, which is computed by first sub-sampling the replay buffer and then computing the uncertainty of functions in the confidence set, where the confidence set is defined by the sub-sampled dataset. Such sub-sampling process gives a dataset with limited complexity and thus ensures the stability of the exploration bonus.

A notable drawback of the work of Wang et al. (2020c) is the requirement of resampling and recomputing the bonus function in each round. Note that implementing this step requires making a full scan over the entire dataset, requiring $\Omega(KH)$ time per round. Furthermore, their algorithm computes the solution to a regression problem to obtain a new policy in each round, which requires $\Omega(KH^2)$ time. Hence, the overall running time of the algorithm is $\Omega(K^2H^2)$. Such running time becomes inefficient when K becomes large. The goal of the current paper is *to obtain an algorithm with*

[*]Alphabetical Order [1]Peking University [2]Carnegie Mellon University [3]University of California, Los Angeles. Correspondence to: Lin F. Yang <linyang@ee.ucla.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]Throughout the paper, we use $\widetilde{O}(\cdot)$ to suppress logarithm factors.

general FA that achieves the same regret bound but with better computational efficiency.

In this paper, we achieve the above goal by applying online sub-sampling techniques. The core idea is to maintain a small sub-sampled dataset, and each time a new datum arrives, instead of making a full scan over the whole dataset, the algorithm decides whether to keep the new data or not by scanning only the sub-sampled dataset, which is much smaller in size. The main difficulty here is how to define the sub-sampling probabilities using only the sub-sampled dataset so that (i) the size of the sub-sampled dataset is bounded and (ii) the sub-sampled dataset provides a good approximation to the confidence set. The sensitivity sampling framework employed in Wang et al. (2020c) can only deal with static datasets and therefore requires a full scan over the whole dataset in each round.

To achieve this goal, we borrow ideas from the online leverage score sampling technique introduced in the streaming algorithm literature (Cohen et al., 2016), which sub-samples a given $n \times d$ matrix ($n \gg d$) row by row and preserves the covariance matrix approximately. Each time a new row arrives, the algorithm computes a probability to keep that row according to the online leverage score, which depends only on the sub-sampled rows and the new row. However, the algorithm and analysis in (Cohen et al., 2016) works only in the linear setting, while the main focus of the current paper is to obtain an algorithm for RL with general FA. This renders the need for new techniques.

In this paper, we establish a novel notion of *online sensitivity score*, which measures the importance of a data point relative to the sub-sampled dataset over a given function class. Online sensitivity score generalizes the notion of online leverage score, which is defined specifically for linear functions (Cohen et al., 2016). We show that for a general function class, by sub-sampling according to online sensitivity scores, the size of the sub-sampled dataset is bounded by the complexity of the function class during the execution of the algorithm, while preserving the same accuracy as in Wang et al. (2020c). Therefore, defining the exploration bonus as the uncertainty of the function class on the sub-sampled dataset will be sufficient for achieving a similar regret guarantee. We also give efficient algorithms for computing online sensitivity scores as well as the bonus function. Using our techniques, our algorithm spends $\tilde{O}(\mathrm{poly}(dH))$ time per round for sub-sampling and computing the exploration bonus, where d depends on the complexity of the function class.

The online sub-sampling technique naturally implies a low switching property of the algorithm. Since the sub-sampled dataset only changes for $\tilde{O}(\mathrm{poly}(dH))$ times, the number of different policies we need to use is also $\tilde{O}(\mathrm{poly}(dH))$. Hence, we only need to solve $\tilde{O}(\mathrm{poly}(dH))$ different regression problems, and therefore our new algorithm spends at most $\tilde{O}(\mathrm{poly}(dH))$ time per round on average for solving regression problems. We note that the low switching property is desirable in many scenarios, like medical domains and recommendation systems (Almirall et al., 2014; Theocharous et al., 2015) (see also Bai et al. (2019) for a detailed survey). As a natural application, our algorithm can be applied in the concurrent RL setting (Silver et al., 2013; Guo & Brunskill, 2015; Dimakopoulou et al., 2018), where multiple agents act concurrently and apply the same policy to collect samples.

Finally, we show that the exploration bonus based on online sub-sampling can be used to explore the environment in an unsupervised fashion. More specifically, we show that even without the guidance of the reward function, after exploring the environment for $\widetilde{O}(\mathrm{poly}(dH))/\epsilon^2$ episodes, our algorithm computes an ϵ-optimal policy for any reward function from a prespecified function class. Reward-free exploration with general function approximation is new and previously was only shown for the tabular setting and the linear setting (Jin et al., 2020a; Wang et al., 2020b).

2. Preliminaries

Throughout the paper, we use $[N]$ to denote the set $\{1, 2, ..., N\}$. We use β as a global parameter, and the choice of β will be elaborated in the appendix.

2.1. Episodic Markov Decision Process

In this paper, we consider a finite-horizon Markov decision process (MDP) $M = (\mathcal{S}, \mathcal{A}, P, r, H, s_1)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P = \{P_h\}_{h=1}^H$ where $P_h : \mathcal{S} \times \mathcal{A} \to \triangle(\mathcal{S})$ are the transition operators, $r = \{r_h\}_{h=1}^H$ where $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ are the deterministic reward functions, H is the planning horizon. Without loss of generality, we assume that the initial state s_1 is fixed.[2]

The agent interacts with the environment episodically. Each episode consists of H time steps. A deterministic policy π chooses an action $a \in \mathcal{A}$ based on the current state $s \in \mathcal{S}$ at each time step $h \in [H]$. Formally, $\pi = \{\pi_h\}_{h=1}^H$ where for each $h \in [H]$, $\pi_h : \mathcal{S} \to \mathcal{A}$ maps a given state to an action. In each episode, the policy π induces a trajectory $s_1, a_1, r_1, s_2, a_2, r_2, ..., s_H, a_H, r_H, s_{H+1}$ where $a_1 = \pi_1(s_1)$, $r_1 = r_1(s_1, a_1)$, $s_2 \sim P_1(\cdot|s_1, a_1)$, etc.

We use Q-function and V-function to evaluate the long-term expected cumulative reward in terms of the current state (state-action pair) and the policy deployed. Concretely, the Q-function and V-function are defined as: $Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{h'=h}^H r_{h'} | s_h = s, a_h = a, \pi\right]$ and

[2] For a general initial distribution ρ, we can treat it as the first stage transition probability, P_1.

$V_h^\pi(s) = \mathbb{E}\left[\sum_{h'=h}^H r_{h'}|s_h = s, \pi\right]$. We also denote the optimal Q-function and V-function as $Q_h^*(s,a) = Q_h^{\pi^*}(s,a)$ and $V_h^*(s) = V_h^{\pi^*}(s)$, where π^* is the optimal policy. For a set of state-action pairs $\mathcal{Z} \subseteq \mathcal{S} \times \mathcal{A}$ and a function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define $\|f\|_{\mathcal{Z}} = \left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(s,a)^2\right)^{1/2}$.

In this paper we use *regret* to measure the effectiveness of the learning algorithm. The regret is defined as $\text{Regret}(K) = \sum_{k=1}^K \left(V_1^*(s_1) - V_1^{\pi^k}(s_1)\right)$, where K is the number of episodes, s_1 is the fixed initial state and π^k is the policy used in k-th episode. We further define $T := KH$ to be the total number of steps.

2.2. General Function Classes and Complexity Measures

In our setting, we assume that a function class $\mathcal{F} \subset \{f : \mathcal{S} \times \mathcal{A} \to [0, H+1]\}$ is given as a priori. In the algorithm we will use functions from $\mathcal{F}$ to approximate the optimal Q-function. We make the following assumption on the expressiveness of the function class $\mathcal{F}$.

Assumption 1. *For any $h \in [H]$ and $V : \mathcal{S} \to [0, H]$, there exists $f_V \in \mathcal{F}$ such that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $f_V(s,a) = \sum_{s'\in\mathcal{S}} P_h(s'|s,a)V(s') + r_h(s,a)$.*

This assumption guarantees that any value function lies in function class $\mathcal{F}$ after applying the Bellman operator. It poses some implicit constraint on the transition operators as well as the reward functions. As mentioned in Wang et al. (2020c), this assumption captures both the tabular setting and the linear MDP setting (Yang & Wang, 2019; Jin et al., 2020b). In practice, when $\mathcal{F}$ is a function class with sufficient expressive power (e.g. deep neural networks), Assumption 1 approximately holds. As we will show in Section I in the appendix, our algorithm is robust to the violation of the assumption, i.e., the algorithm still works well if the above assumption is satisfied approximately.

To measure the complexity of the RL problem, we need to assume the function class $\mathcal{F}$ has bounded eluder dimension $\dim_E(\mathcal{F}, \epsilon)$ ($\epsilon > 0$ is a parameter). We also need to assume bounded covering numbers for both the function class and the state-action space, denoted as $\mathcal{N}(\mathcal{F}, \epsilon)$ and $\mathcal{N}(\mathcal{S} \times \mathcal{A}, \epsilon)$ respectively. These two assumptions are standard in the literature (Russo & Roy, 2013; Wang et al., 2020c; Jin et al., 2021). The definitions will be given in the appendix.

2.3. Switching Cost

The concept of switching cost is first introduced in Bai et al. (2019). It is used to quantify the adaptability of reinforcement learning algorithms. In this paper we focus on the *global switching cost*, which counts the number of policy changes in the running of the algorithm in K episodes, namely: $N_{\text{switch}}^{\text{gl}} := \sum_{k=1}^{K-1} \mathbb{I}\{\pi_k \neq \pi_{k+1}\}$.

3. Computation-Efficient Algorithm via Online Sub-Sampling

In this section we introduce our online importance sub-sampling technique and several key components in the algorithm design. The full algorithms are deferred to the appendix. For every episode, our algorithm consists of three steps. In Step 1, we apply online sub-sampling to the data collected so far to reduce the size of the dataset. This sub-sampled dataset will be used to compute a stable exploration bonus in Step 2. The most interesting part is that we are able to decide whether or not we have collected substantial new information according to the sub-sampling procedure. In Step 2, if we have already collected enough new information, we do optimistic planning using those new information to calculate a new policy and then update the current policy to be the new policy. Otherwise, we keep using the old policy. In Step 3, we use the current policy to interact with the environment to collect new data. In this section, we define $d = \max(\log(\mathcal{N}(\mathcal{F}, \delta/T^2)), \dim_E(\mathcal{F}, 1/T), \log(\mathcal{N}(\mathcal{S} \times \mathcal{A}, \delta/T^2)))$ to be the complexity of the function class.

3.1. Online Importance Sub-Sampling

We use $\mathcal{Z}_h^k$ to denote the dataset of all the state-action pairs collected in step h up to episode k, i.e., $\mathcal{Z}_h^k = \{(s_h^\tau, a_h^\tau)\}_{\tau\in[k-1]}$. In Wang et al. (2020c), a key step is to obtain a good approximation of the dataset $\mathcal{Z}_h^k$, denoted as the *sub-sampled dataset* $\widehat{\mathcal{Z}}_h^k$, but with a much lower complexity (i.e., number of distinct points). Thus one can use $\widehat{\mathcal{Z}}_h^k$ to compute a more stable exploration bonus. Wang et al. (2020c) achieve this goal by resampling all existing data points each time a new data point is added. Therefore, the algorithm in Wang et al. (2020c) needs to scan the dataset for $H \cdot K$ times, and the dataset could have size as large as K. Such a method has two shortcomings. First, for the same h, the datasets $\{\mathcal{Z}_h^k\}_{k\in[K]}$ are similar to each other, and therefore sub-sampling each of them separately could be computationally inefficient. Moreover, because of the randomness of the sampling procedure, the exploration bonus changes in each episode, and therefore induces a high switching cost. We tackle these two problems by modifying the sub-sampling algorithm to an online version which immediately improves the computational efficiency. Furthermore, we can achieve low switching cost by switching the policy properly according to the sampling procedure.

Now we describe the procedure for constructing the sub-sampled dataset $\{\widehat{\mathcal{Z}}_h^k\}_{h=1}^H$, which is initialized to be an empty set for all $h \in [H]$. At the beginning of episode k, the algorithm receives $\{\widehat{\mathcal{Z}}_h^{k-1}\}_{h=1}^H$ and $\{(s_h^{k-1}, a_h^{k-1})\}_{h=1}^H$, i.e., the current sub-sampled datasets and the trajectory obtained in the previous episode. For each $h \in [H]$, we first compute the *online sensitivity score* (defined in (1)) of (s_h^{k-1}, a_h^{k-1}) with respect to $\widehat{\mathcal{Z}}_h^{k-1}$ by setting $\mathcal{Z} = \widehat{\mathcal{Z}}_h^{k-1}$

and $z = (s_h^{k-1}, a_h^{k-1})$ in (1), to measure the importance of (s_h^{k-1}, a_h^{k-1}) relative to $\widehat{\mathcal{Z}}_h^{k-1}$.

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

For each $h \in [H]$, starting with $\widehat{\mathcal{Z}}_h^k \leftarrow \widehat{\mathcal{Z}}_h^{k-1}$, our algorithm then adds (s_h^{k-1}, a_h^{k-1}) into $\widehat{\mathcal{Z}}_h^k$ with probability proportional to its online sensitivity score. We also set the weight (or equivalently, the number of copies added to the sub-sampled dataset) of (s_h^{k-1}, a_h^{k-1}) to be the reciprocal of the sampling probability, if added.

Sensitivity scores measure how much new information (s_h^{k-1}, a_h^{k-1}) contains relative to the sub-sampled dataset $\widehat{\mathcal{Z}}_h^{k-1}$. We recompute the policy if a data point is added to the sub-sampled dataset (explained in the next section). As will be demonstrated, the total number of added data points is bounded by $\widetilde{O}(d^2)$, and thus the algorithm achieves low switching cost and low running time.

3.2. Optimistic Planning

We now describe how to compute an optimistic policy once a new data point is added to the sub-sampled dataset. Following Wang et al. (2020c), we calculate the new Q-functions using least-squares value iteration. Formally, we calculate $Q_H^k, Q_{H-1}^k, ..., Q_1^k$ as optimistic approximation of the optimal Q-functions iteratively. For each $h = H, H-1, ..., 1$, we solve the following optimization problem:

$$f_h^k \leftarrow \arg\min_{f\in\mathcal{F}} \sum_{\tau=1}^{k-1}\left(f(s_h^\tau, a_h^\tau) - \left(r_h^\tau + \max_{a\in\mathcal{A}} Q_{h+1}^k(s_{h+1}^\tau, a)\right)\right)^2$$

and set the estimated Q-function to be $Q_h^k(\cdot,\cdot) \leftarrow \min\left\{f_h^k(\cdot,\cdot) + b_h^k(\cdot,\cdot), H\right\}$, where $b_h^k(\cdot,\cdot)$ is an exploration bonus defined by the sub-sampled dataset $\widehat{\mathcal{Z}}_h^k$. We will discuss how to compute the $b_h^k(\cdot,\cdot)$ shortly. The policy is defined as the greedy policy with respect to Q_h^k.

3.3. Exploration Bonus

In this section we introduce our design of the exploration bonus. The goal of the exploration bonus is such that our estimate of the Q-function is an upper bound of the optimal Q-function. Let $\bar{f}_h^k(\cdot,\cdot) = r_h(\cdot,\cdot) + \sum_{s'\in\mathcal{S}} P_h(s'|\cdot,\cdot)V_{h+1}^k(s')$. The bonus function should measure the difference between $\bar{f}_h^k(\cdot,\cdot)$ and $f_h^k(\cdot,\cdot)$. Here a natural choice would be

$$\sup_{f_1,f_2\in\mathcal{F}, \|f_1-f_2\|_{\mathcal{Z}_h^k}^2\le\beta} |(f_1(\cdot,\cdot) - f_2(\cdot,\cdot)|,$$

where β is a parameter determined by the error bound of the regression problem. As discussed in Wang et al. (2020c), the above bonus function is not only computationally expensive but also introduces technical difficulties in proving the regret guarantee of the algorithm. Fortunately, as we have already obtained a sub-sampled dataset, $\widehat{\mathcal{Z}}_h^k$, which is a simplified version of the true dataset $\mathcal{Z}_h^k$, the bonus function can be simply defined to be

$$b_h^k(\cdot,\cdot) \leftarrow \sup_{f_1,f_2\in\mathcal{F}, \|f_1-f_2\|_{\widehat{\mathcal{Z}}_h^k}^2\le\beta} |(f_1(\cdot,\cdot) - f_2(\cdot,\cdot)|.$$

3.4. Computational Efficiency

Our algorithm can be implemented efficiently by only assuming access to a *Regression Oracle*, which is a mild assumption and is common in the literature (Foster et al., 2018; Foster & Rakhlin, 2020; Foster et al., 2020). Benefiting from both the online sub-sampling and the low-switching property, our algorithm only takes $\widetilde{O}(\text{poly}(dH)\cdot|\mathcal{A}|)$ time per round on average with access to a regression oracle. The details are deferred to the appendix.

3.5. Theoretical Guarantee

Theorem 1. *Under Assumption 1, for sufficiently large T, with probability $1-\delta$, the algorithm achieves a regret bound Regret$(K) = O(\sqrt{\iota_1 \cdot H^3 \cdot T})$ where*

$$\iota_1 = \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta)\cdot \dim_E^2(\mathcal{F}, 1/T)\cdot\log^2 T \cdot \log\left(\mathcal{N}(\mathcal{S}\times\mathcal{A}, \delta/T^2)\cdot T/\delta\right)$$

and the global switching cost is bounded by $N_{switch}^{gl} = O(\iota_2 \cdot H)$ where

$$\iota_2 = \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta)\cdot \dim_E(\mathcal{F}, 1/T)\cdot\log^2 T.$$

Furthermore, with probability $1-\delta$ the algorithm takes $\widetilde{O}(\text{poly}(dH)\cdot|\mathcal{A}|)$ time per round on average with access to a regression oracle.

Note that our regret bound is the same with that in Wang et al. (2020c) when applied to the same setting, whereas our running time and switching-cost are much lower.

4. Conclusion

We establish a novel RL algorithm with general function approximation using the online sub-sampling technique. Our algorithm greatly improves the computational efficiency compared to that in Wang et al. (2020c) and enjoys nearly the same regret bound. Furthermore, our algorithm achieves low switching cost, making it appealing to be implemented in real-life scenarios. Moreover, the algorithm can be easily modified to provably explore the environment without the guidance of a reward function.

References

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *arXiv preprint arXiv:2006.10814*, 2020a.

Agarwal, A., Kakade, S., and Yang, L. F. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83. PMLR, 2020b.

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Almirall, D., Nahum-Shani, I., Sherwood, N. E., and Murphy, S. A. Introduction to smart designs for the development of adaptive interventions: with application to weight loss research. *Translational behavioral medicine*, 4(3): 260–274, 2014.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Azar, M. G., Munos, R., and Kappen, H. J. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3): 325–349, 2013.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272, 2017.

Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. Provably efficient q-learning with low switching cost. In *Advances in Neural Information Processing Systems*, pp. 8004–8013, 2019.

Cohen, M. B., Musco, C., and Pachocki, J. Online row sampling. *arXiv preprint arXiv:1604.05448*, 2016.

Cui, Q. and Yang, L. F. Minimax sample complexity for turn-based stochastic game. *arXiv preprint arXiv:2011.14267*, 2020a.

Cui, Q. and Yang, L. F. Is plug-in solver sample-efficient for feature-based reinforcement learning? *arXiv preprint arXiv:2010.05673*, 2020b.

Dimakopoulou, M., Osband, I., and Van Roy, B. Scalable coordinated exploration in concurrent reinforcement learning. *arXiv preprint arXiv:1805.08948*, 2018.

Du, S. S., Kakade, S. M., Wang, R., and Yang, L. F. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2019.

Du, S. S., Lee, J. D., Mahajan, G., and Wang, R. Agnostic q-learning with function approximation in deterministic systems: Near-optimal bounds on approximation error and sample complexity. *Advances in Neural Information Processing Systems*, 33, 2020.

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

Feldman, D. and Langberg, M. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 569–578, 2011.

Feldman, D., Schmidt, M., and Sohler, C. Turning big data into tiny data: constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1434–1453, 2013.

Foster, D. and Rakhlin, A. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 3199–3210. PMLR, 2020.

Foster, D. J., Agarwal, A., Dudik, M., Haipeng, L., and Schapire, R. E. Practical contextual bandits with regression oracles. In *35th International Conference on Machine Learning, ICML 2018*, pp. 2482–2517. International Machine Learning Society (IMLS), 2018.

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

Freedman, D. A. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Gao, M., Xie, T., Du, S. S., and Yang, L. F. A provably efficient algorithm for linear markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494*, 2021.

Guo, Z. and Brunskill, E. Concurrent pac rl. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

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

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *Advances in neural information processing systems*, pp. 4863–4873, 2018.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. *arXiv preprint arXiv:2002.02794*, 2020a.

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

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

Kakade, S. M. *On the sample complexity of reinforcement learning*. PhD thesis, UCL (University College London), 2003.

Kaufmann, E., Ménard, P., Domingues, O. D., Jonsson, A., Leurent, E., and Valko, M. Adaptive reward-free exploration. *arXiv preprint arXiv:2006.06294*, 2020.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Krishnamurthy, A., Agarwal, A., Huang, T.-K., Daumé III, H., and Langford, J. Active learning for cost-sensitive classification. In *International Conference on Machine Learning*, pp. 1915–1924. PMLR, 2017.

Langberg, M. and Schulman, L. J. Universal ϵ-approximators for integrals. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pp. 598–607. SIAM, 2010.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 33, 2020.

Li, G., Kamath, P., Foster, D. J., and Srebro, N. Eluder dimension and generalized rank. *arXiv preprint arXiv:2104.06970*, 2021.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

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

Osband, I. and Van Roy, B. Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27:1466–1474, 2014.

Osband, I. and Van Roy, B. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Russo, D. and Roy, B. V. Eluder dimension and the sample complexity of optimistic exploration. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pp. 2256–2264, 2013.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pp. 5186–5196, 2018.

Sidford, A., Wang, M., Yang, L., and Ye, Y. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pp. 2992–3002. PMLR, 2020.

Silver, D., Newnham, L., Barker, D., Weller, S., and McFall, J. Concurrent reinforcement learning from customer interactions. In *International conference on machine learning*, pp. 924–932. PMLR, 2013.

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

Szita, I. and Szepesvari, C. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 1031–1038, 2010.

Theocharous, G., Thomas, P. S., and Ghavamzadeh, M. Ad recommendation systems for life-time value optimization. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 1305–1310, 2015.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Wang, R., Du, S. S., Yang, L., and Kakade, S. Is long horizon rl more difficult than short horizon rl? *Advances in Neural Information Processing Systems*, 33, 2020a.

Wang, R., Du, S. S., Yang, L. F., and Salakhutdinov, R. On reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2006.11274*, 2020b.

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

Wang, T., Zhou, D., and Gu, Q. Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *arXiv preprint arXiv:2101.02195*, 2021.

Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Yang, L. and Wang, M. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004, 2019.

Yang, L. and Wang, M. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pp. 10746–10756. PMLR, 2020.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR, 2019.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020a.

Zanette, A., Lazaric, A., Kochenderfer, M. J., and Brunskill, E. Provably efficient reward-agnostic navigation with linear value iteration. *arXiv preprint arXiv:2008.07737*, 2020b.

Zhang, Z., Ji, X., and Du, S. S. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*, 2020a.

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33, 2020b.

Organization. The appendix is organized as follows. In Section A, we present missing notations, definitions and algorithms. In Section B, we discuss related works and its comparisons with our work. In Section C, we extend our results to the reward-free RL setting and state the theoretical guarantee in this setting. In Section D, we show how to implement our algorithms efficiently and present computation time guarantee. In Section E, we provide a proof sketch for our main theorem. In Section F, we formalize and prove the properties of the online sub-sampling procedure. In Section G and Section H we present formal proofs of Theorem 1 and Theorem 2. In Section I, we present our results in the misspecified setting, i.e., the case when Assumption 1 and Assumption 3 only hold approximately.

A. Missing Notations, Definitions, and Algorithms

A.1. Missing Notations

We define the infinity-norm of function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $v : \mathcal{S} \to \mathbb{R}$ as: $\|f\|_\infty = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |f(s,a)|$ and $\|v\|_\infty = \sup_{s\in\mathcal{S}} |v(s)|$. Given a dataset $\mathcal{D} = \{(s_i, a_i, q_i)\}_{i=1}^n \subseteq \mathcal{S} \times \mathcal{A} \times \mathbb{R}$, for a function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define $\|f\|_\mathcal{D} = \left(\sum_{i=1}^n (f(s_i, a_i) - q_i)^2\right)^{1/2}$. For n events $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$, we write

$$\Pr(\mathcal{E}_1\mathcal{E}_2 \ldots \mathcal{E}_n) = \Pr(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \ldots \cap \mathcal{E}_n).$$

For a event $\mathcal{E}$, we use $\mathbb{I}\{\mathcal{E}\}$ to denote the indicator function, i.e.,

$$\mathbb{I}\{\mathcal{E}\} = \begin{cases} 1 & \mathcal{E} \text{ holds} \\ 0 & \text{otherwise} \end{cases}.$$

For a event $\mathcal{E}$, we use $\mathcal{E}^c$ to denote its complement. For a multiset $\mathcal{Z}$, we use $|\mathcal{Z}|$ to denote the cardinality of $\mathcal{Z}$, and $n_d(\mathcal{Z})$ the number of distinct elements in $\mathcal{Z}$.

A.2. Missing Definitions

The eluder dimension (Russo & Roy, 2013) of the function class $\mathcal{F}$ is defined as follows. We remark that a wide range of function classes, including linear functions, generalized linear functions and bounded degree polynomials, have bounded eluder dimension. For more examples and more discussion, see Russo & Roy (2013); Osband & Van Roy (2014); Li et al. (2021).

Definition 1 (Eluder Dimension). *Let $\epsilon \geq 0$ and $\mathcal{Z} = \{(s_i, a_i)\}_{i=1}^n \subseteq \mathcal{S} \times \mathcal{A}$ be a sequence of state-action pairs.*
(1) A state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ is ϵ-dependent *on $\mathcal{Z}$ with respect to $\mathcal{F}$ if any $f, f' \in \mathcal{F}$ satisfying $\|f - f'\|_\mathcal{Z} \leq \epsilon$ also satisfies $|f(s, a) - f'(s, a)| \leq \epsilon$.*
(2) An (s, a) is ϵ-independent *of $\mathcal{Z}$ with respect to $\mathcal{F}$ if (s, a) is not ϵ-dependent on $\mathcal{Z}$.*
(3) The ϵ-eluder dimension $\dim_E(\mathcal{F}, \epsilon)$ *of a function class $\mathcal{F}$ is the length of the longest sequence of elements in $\mathcal{S} \times \mathcal{A}$ such that, for some $\epsilon' \geq \epsilon$, every element is ϵ'-independent of its predecessors.*

The covering numbers are defined as follows. Since our final regret bound depends logarithmically on the covering numbers, it is acceptable for the covers to have exponential size.

Assumption 2 (Covering Numbers). *The function class $\mathcal{F}$, and state-action space $\mathcal{S} \times \mathcal{A}$ both have bounded covering numbers. Concretely, for any $\epsilon > 0$, there exists an ϵ-cover $\mathcal{C}(\mathcal{F}, \epsilon) \subseteq \mathcal{F}$ with size $|\mathcal{C}(\mathcal{F}, \epsilon)| \leq \mathcal{N}(\mathcal{F}, \epsilon)$, such that for any $f \in \mathcal{F}$, there exists $f' \in \mathcal{C}(\mathcal{F}, \epsilon)$ with $\|f - f'\|_\infty \leq \epsilon$. Also, there exists an ϵ-cover $\mathcal{C}(\mathcal{S} \times \mathcal{A}, \epsilon)$ with size $|\mathcal{C}(\mathcal{S} \times \mathcal{A}, \epsilon)| \leq \mathcal{N}(\mathcal{S} \times \mathcal{A}, \epsilon)$, such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, there exists $(s', a') \in \mathcal{C}(\mathcal{S} \times \mathcal{A}, \epsilon)$ with $\sup_{f\in\mathcal{F}} |f(s, a) - f(s', a')| \leq \epsilon$.*

A.3. Missing Algorithms

The main algorithm is presented in Algorithm 1. The online sub-sampling algorithm used in the main algorithm is presented in algorithm 2.

B. Related Works and Comparisons

Tabular RL. There is a long line of theoretical work on the sample complexity and regret bound for RL in the tabular setting. See, e.g., (Kearns & Singh, 2002; Kakade, 2003; Szita & Szepesvari, 2010; Jaksch et al., 2010; Azar et al., 2013;

Algorithm 1 Low Switching Cost Value Iteration

Input: Failure probability $\delta \in (0,1)$, number of episodes K, and function class $\mathcal{F}$.
$\tilde{k} \leftarrow 1$
$\widehat{\mathcal{Z}}_h^1 \leftarrow \{\} \quad \forall h \in [H]$
for episode $k = 1, 2, ..., K$ **do**
  for $h = H, H-1, ..., 1$ **do**
    $\widehat{\mathcal{Z}}_h^k \leftarrow$ **Online-Sample**$(\mathcal{F}, \widehat{\mathcal{Z}}_h^{k-1}, (s_h^{k-1}, a_h^{k-1}), \delta)$ (if $k \geq 2$)
  end for
  if $k = 1$ **or** $\exists h \in [H]$ $\widehat{\mathcal{Z}}_h^k \neq \widehat{\mathcal{Z}}_h^{\tilde{k}}$ **then**
    $\tilde{k} \leftarrow k$
    $Q_{H+1}^k(\cdot,\cdot) \leftarrow 0, V_{H+1}^k(\cdot) \leftarrow 0$
    for $h = H, H-1, ..., 1$ **do**
      $\mathcal{D}_h^k \leftarrow \{(s_h^\tau, a_h^\tau, r_h^\tau + V_{h+1}^k(s_{h+1}^\tau))\}_{\tau \in [k-1]}$
      $f_h^k \leftarrow \text{argmin}_{f \in \mathcal{F}} \|f\|_{\mathcal{D}_h^k}^2$
      $b_h^k(\cdot,\cdot) \leftarrow \sup_{f_1, f_2 \in \mathcal{F}, \|f_1 - f_2\|_{\widehat{\mathcal{Z}}_h^k}^2 \leq \beta} |(f_1(\cdot,\cdot) - f_2(\cdot,\cdot)|$
      $Q_h^k(\cdot,\cdot) \leftarrow \min\{f_h^k(\cdot,\cdot) + b_h^k(\cdot,\cdot), H\}$ and $V_h^k(\cdot) = \max_{a \in \mathcal{A}} Q_h^k(\cdot, a)$
      $\pi_h^k(\cdot) \leftarrow \text{argmax}_{a \in \mathcal{A}} Q_h^k(\cdot, a)$
    end for
  end if
  Receive initial state s_1
  for $h = 1, 2, ..., H$ **do**
    Take action $a_h^k \leftarrow \pi_h^{\tilde{k}}(s_h^k)$ and observe s_{h+1}^k and r_h^k
  end for
end for

Algorithm 2 Online-Sample$(\mathcal{F}, \widehat{\mathcal{Z}}, z, \delta)$

Input: Function class $\mathcal{F}$, current sub-sampled dataset $\widehat{\mathcal{Z}} \subseteq \mathcal{S} \times \mathcal{A}$, new state-action pair z, failure probability $\delta \in (0,1)$
Let p_z to be the smallest real number such that $1/p_z$ is an integer and

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

Let $\widehat{z} \in \mathcal{C}(\mathcal{S} \times \mathcal{A}, 1/16\sqrt{64T^3/\delta}))$ such that $\sup_{f \in \mathcal{F}} |f(z) - f(\widehat{z})| \leq 1/16\sqrt{64T^3/\delta}$
Add $1/p_z$ copies of $\widehat{z}$ into $\widehat{\mathcal{Z}}$ (or equivalently, set the weight of $\widehat{z}$ to be $1/p_z$) with probability p_z
return $\widehat{\mathcal{Z}}$

Osband & Van Roy, 2016; Azar et al., 2017; Jin et al., 2018; Sidford et al., 2018; Zanette & Brunskill, 2019; Agarwal et al., 2020b; Wang et al., 2020a; Zhang et al., 2020a; Sidford et al., 2020; Cui & Yang, 2020a; Li et al., 2020) and references therein. However, as these results all depend polynomially on the size of the state space, they can not be directly applied to real-world problems with large state spaces.

RL with Function Approximation. As mentioned above, because of the large state spaces in real-world problems, it is more desirable to design and analyze algorithms with function approximation. The most basic and frequently studied setting is RL with linear function approximation. See, e.g., (Yang & Wang, 2019; 2020; Jin et al., 2020b; Du et al., 2019; Wang et al., 2019; Zanette et al., 2020a;b; Du et al., 2020; Cui & Yang, 2020b; Agarwal et al., 2020a) for recent theoretical advances.

Recently, there are a number of results analyzing RL with general function approximation. Jiang et al. (2017) design a provably efficient algorithm whose sample complexity can be upper bounded in terms of the Bellman rank. Ayoub et al. (2020) propose an algorithm for model-based RL based on value-targeted regression. Jin et al. (2021) develop an algorithm for problems with bounded bellman eluder dimension. Du et al. (2021) propose an algorithm for Bilinear Classes. Unfortunately, the above methods in general require to solve computationally intractable optimization problems. Wang et al.

(2020c) design a LSVI-based model-free algorithm whose regret bound depends on the eluder dimension of the function class. However, as mentioned before, their algorithm requires at least $\Omega(K^2H^2)$ computation time and has $\Omega(K)$ switching cost. Foster et al. (2020) also propose a LSVI-based algorithm whose regret bound depends on the notion of disagreement coefficient which is upper bounded by the eluder dimension of the function class. However, the algorithm of Foster et al. (2020) also requires at least $\Omega(K^2H^2)$ computation time and has $\Omega(K)$ switching cost. Moreover, the algorithm of Foster et al. (2020) additionally requires the block MDP assumption.

Reward-Free and Low Switching Cost. The reward-free setting is proposed in (Jin et al., 2020a). Kaufmann et al. (2020) refine the algorithm proposed in (Jin et al., 2020a) with improved sample complexity. Wang et al. (2020b); Zanette et al. (2020b) further design provably efficient algorithms with linear function approximation in the reward-free setting. Bai et al. (2019) is the first work that studies switching cost in RL. This problem was later studied in (Zhang et al., 2020b; Gao et al., 2021; Wang et al., 2021). Our work focus on the *global* switching cost studied in (Gao et al., 2021).

Online Sub-Sampling. Our core technique, sub-sampling by online sensitivity score, is inspired by the sensitivity sampling technique introduced in (Langberg & Schulman, 2010; Feldman & Langberg, 2011; Feldman et al., 2013) and the online leverage score sampling technique introduced in (Cohen et al., 2016). However, as mentioned earlier, the algorithm and analysis in (Cohen et al., 2016) works only for linear functions, and the framework in (Langberg & Schulman, 2010; Feldman & Langberg, 2011; Feldman et al., 2013) can only deal with static datasets. On the other hand, our techniques can deal with general function classes while operate in an online manner.

Comparison with Wang et al. (2020c) on Sampling Procedure. Our sub-sampling procedure is different from that in Wang et al. (2020c) in the following two aspects. First, the algorithm in Wang et al. (2020c) resamples the whole dataset once a new data point is obtained, while in our algorithm, either the sub-sampled dataset keeps unchanged, or (multiple copies of) the new data point is added to the subsampled dataset. Moreover, the definition of the sensitivity score in Wang et al. (2020c) depends on the whole dataset, while in our algorithm, the definition depends only on the current sub-sampled dataset $\widehat{\mathcal{Z}}_h^{k-1}$. These two differences are crucial for the low switching cost and low running time of our algorithm. Furthermore, as we will show later, even under this new definition of sensitivity score, the size of the sub-sampled dataset is bounded and the sub-sampled dataset provides a good approximation to the confidence set.

On the Linear Setting. When $\mathcal{F}$ is the class of d-dimensional linear functions, the global switching cost bound given in Theorem 1 is $\widetilde{O}(d^2H)$, which is worse than the $\widetilde{O}(dH)$ bound given in Gao et al. (2021). However, for linear functions, our sampling procedure is equivalent to the online leverage score sampling (Cohen et al., 2016), and therefore, by using the analysis in (Cohen et al., 2016) which is specific to the linear setting, the switching cost bound can be improved to $\widetilde{O}(dH)$, matching the bound given in Gao et al. (2021). Using the same technique, our regret bound can be improved to $\widetilde{O}(\sqrt{d^3H^3T})$ in the linear setting, matching the bound given in Jin et al. (2020b); Gao et al. (2021).

C. Online Sub-Sampling in the Reward-Free Setting

In this section we show that our results can be extended to the reward-free exploration setting, in which the agent explores the environment without the guidance of a reward, while achieving both low switching cost and computation efficiency. We begin with some basics and notations of reward-free RL.

C.1. Reward-free RL

The reward-free RL contains two phases, the *exploration phase* and the *planning phase*. In the exploration phase, the agent interacts with the MDP in episodes as usual, but receives no reward signal. After the exploration phase, the agent is given a reward function in the planning phase. The goal of the reward-free RL is to output a near-optimal policy with respect to the given reward function with no additional access to the environment. As mentioned in Jin et al. (2020a) and Wang et al. (2020b), this paradigm is particular suitable for the batch RL setting and the setting where there are multiple reward functions of interest.

Notations. Slightly changing the notation, we define the value (action-value) functions with respect to a given reward function $r = \{r_h\}_{h=1}^H$ as

$$V_h^\pi(s, r) = \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \mid s_h = s, \pi\right]$$

and

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

The optimal value (action-value) functions $V_h^*(s, r)$ and $Q_h^*(s, a, r)$ are defined similarly. We say a policy π is a ϵ-optimal policy with respect to r if $V_1^*(s_1, r) - V_1^\pi(s_1, r) \le \epsilon$. The global switching cost in the reward-free RL setting is defined as the number of policy changes in the exploration phase: $N_{\text{switch}}^{\text{gl}} := \sum_{k=1}^{K-1} \mathbb{I}\{\pi_k \neq \pi_{k+1}\}$, where π_k is the policy used in the k-th episode of the exploration phase.

C.2. Algorithm

The algorithm consists of two phases: an exploration phase and a planning phase. Below, we introduce our algorithm.

Algorithm 3 Exploration Phase

Input: Failure probability $\delta \in (0, 1)$, number of episodes K, and function class $\mathcal{F}$.
$\tilde{k} \leftarrow 1$
$\widehat{\mathcal{Z}}_h^1 \leftarrow \{\}$ $\quad \forall h \in [H]$
for episode $k = 1, 2, ..., K$ **do**
  for $h = H, H-1, ..., 1$ **do**
    $\widehat{\mathcal{Z}}_h^k \leftarrow$ **Online-Sample**$(\mathcal{F}, \widehat{\mathcal{Z}}_h^{k-1}, (s_h^{k-1}, a_h^{k-1}), \delta)$ (if $k \ge 2$)
  end for
  if $k = 1$ **or** $\exists h \in [H]$ $\widehat{\mathcal{Z}}_h^k \neq \widehat{\mathcal{Z}}_h^{\tilde{k}}$ **then**
    $\tilde{k} \leftarrow k$
    $Q_{H+1}^k(\cdot, \cdot) \leftarrow 0, V_{H+1}^k(\cdot) \leftarrow 0$
    for $h = H, H-1, ..., 1$ **do**
      $\mathcal{D}_h^k \leftarrow \{(s_h^\tau, a_h^\tau, V_{h+1}^k(s_{h+1}^\tau))\}_{\tau \in [k-1]}$
      $f_h^k \leftarrow \text{argmin}_{f \in \mathcal{F}} \|f\|_{\mathcal{D}_h^k}^2$
      $b_h^k(\cdot, \cdot) \leftarrow \sup_{\|f_1 - f_2\|_{\widehat{\mathcal{Z}}_h^k}^2 \le \beta} |(f_1(\cdot, \cdot) - f_2(\cdot, \cdot)|$
      $r_h^k(\cdot, \cdot) \leftarrow \min\{b_h^k(\cdot, \cdot)/H, 1\}$
      $Q_h^k(\cdot, \cdot) \leftarrow \min\{f_h^k(\cdot, \cdot) + b_h^k(\cdot, \cdot) + r_h^k(\cdot, \cdot), H\}$
      $V_h^k(\cdot) \leftarrow \max_{a \in \mathcal{A}} Q_h^k(\cdot, a)$
      $\pi_h^k(\cdot) \leftarrow \text{argmax}_{a \in \mathcal{A}} Q_h^k(\cdot, a)$
    end for
  end if
  Receive fixed initial state s_1
  for $h = 1, 2, ..., H$ **do**
    Take action $a_h^k \leftarrow \pi_h^{\tilde{k}}(s_h^k)$ and observe s_{h+1}^k
  end for
end for

Exploration Phase. Our algorithm for the exploration phase is quite similar to Algorithm 1[3]. The main difference is that without the guidance from the reward signal, we use the following *exploration-driven* reward function to encourage exploration:

$$r_h^k(\cdot, \cdot) \leftarrow \min\{b_h^k(\cdot, \cdot)/H, 1\}.$$

The full algorithm used in the exploration phase is presented in Algorithm 3.

[3]We need to slightly change the choice of β in the reward-free setting. See Section F for details.

Algorithm 4 Planning Phase

Input: Dataset $\mathcal{Z}_h^K, h \in [H]$, subsampled dataset $\widehat{\mathcal{Z}}_h^K, h \in [H]$, reward function $r = \{r_h\}_{h=1}^H$, and function class $\mathcal{F}$.
$Q_{H+1}(\cdot,\cdot) \leftarrow 0, V_{H+1}(\cdot) \leftarrow 0$
$\mathcal{Z}_h \leftarrow \mathcal{Z}_h^K, \widehat{\mathcal{Z}}_h \leftarrow \widehat{\mathcal{Z}}_h^K$
for $h = H, H-1, ..., 1$ **do**
  $\mathcal{D}_h \leftarrow \{(s_h^\tau, a_h^\tau, V_{h+1}(s_{h+1}^\tau))\}_{\tau \in [K-1]}$
  $f_h \leftarrow \text{argmin}_{f \in \mathcal{F}} \|f\|_{\mathcal{D}_h}^2$
  $b_h(\cdot,\cdot) \leftarrow \sup_{\|f_1 - f_2\|_{\widehat{\mathcal{Z}}_h}^2 \le \beta} |(f_1(\cdot,\cdot) - f_2(\cdot,\cdot)|$
  $Q_h(\cdot,\cdot) \leftarrow \min\{f_h(\cdot,\cdot) + b_h(\cdot,\cdot) + r_h(\cdot,\cdot), H\}$
  $V_h(\cdot) \leftarrow \max_{a \in \mathcal{A}} Q_h(\cdot, a)$
  $\pi_h(\cdot) \leftarrow \text{argmax}_{a \in \mathcal{A}} Q_h(\cdot, a)$
end for
return $\pi = \{\pi_h\}_{h=1}^H$

Planning Phase. In the planning phase we do optimistic planning similar to Algorithm 1, but with real reward instead of exploration-driven reward. We still add the bonus function to guarantee optimism.

C.3. Assumptions

Before presenting our theoretical guarantee, we need to make a few assumptions on the function classes used in the algorithm. The first assumption is on the expressiveness of the function class $\mathcal{F}$.

Assumption 3. *For any $h \in [H]$ and any $V : \mathcal{S} \to [0, H]$, there exists $f_V \in \mathcal{F}$ which satisfies*

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

for all $(s,a) \in S \times A$

Compared to Assumption 1, with no reward function, Assumption 3 can be regarded as a constrain on the transition core. Intuitively, this assumption guarantees that we can use function class $\mathcal{F}$ to effectively explore the transition operator.

In Linear MDPs, it is assumed that the reward function is linear in the feature extractor. Instead of making explicit assumption on the structure of the reward function, we assume the reward function given in the planning phase belongs to a function class with bounded covering number.

Assumption 4. *The reward function $r = \{r_h\}_{h=1}^H$ belongs to a function class $\mathcal{R} \subseteq \{\mathcal{S} \times \mathcal{A} \to [0,1]\}$, i.e., $r_h \in \mathcal{R}$ for all $h \in [H]$. And for any $\epsilon > 0$, there exists an ϵ-cover $\mathcal{C}(\mathcal{R}, \epsilon)$ with size $|\mathcal{C}(\mathcal{R}, \epsilon)| \le \mathcal{N}(\mathcal{R}, \epsilon)$.*

C.4. Theoretical Guarantee

Now we state our theoretical guarantee in the reward-free RL setting.

Theorem 2. *Suppose Assumption 3 holds and T is sufficiently large. For any given $\delta \in (0,1)$, after collecting K trajectories during the exploration phase (by Algorithm 3), with probability at least $1-\delta$, for any reward function $r = \{r_h\}_{h=1}^H$ satisfying Assumption 4, Algorithm 4 outputs an $O(H^3 \cdot \sqrt{\iota_1 / K})$-optimal policy for the MDP $(\mathcal{S}, \mathcal{A}, P, r, H, s_1)$. Here,*

$$\iota_1 = \log(\mathcal{N}(\mathcal{R}, 1/T)) \cdot \dim_E(\mathcal{F}, 1/T)$$
$$+ \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \cdot \log^2 T \cdot \dim_E^2(\mathcal{F}, 1/T) \cdot \log\left(\mathcal{N}(\mathcal{S} \times \mathcal{A}, \delta/T^2) \cdot T/\delta\right).$$

Moreover, the global switching cost of Algorithm 3 is upper bounded by

$$N_{switch}^{gl} = O(H \cdot \iota_2)$$

where

$$\iota_2 = \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \cdot \dim_E(\mathcal{F}, 1/T) \cdot \log^2 T.$$

Furthermore, with probability at least $1-\delta$, Algorithm 3 takes $\tilde{O}(\text{poly}(dH) \cdot |\mathcal{A}|)$ time per round with access to a regression oracle.

Using ideas in the proof of Theorem 1, the proof of Theorem 2 follows rather straightforwardly from Wang et al. (2020b). The high-level idea is to show that, after the exploration phase, for any reward function, the error of the planning policy is upper bounded by the expectation of the bonus functions, which is shown to be small enough using results proved in Theorem 1. The formal proof of Theorem 2 is presented in Section H.

Remark 1. *Let* $d(T,\delta) := \max(\log(\mathcal{N}(\mathcal{R},1/T)), \log(\mathcal{N}(\mathcal{F},\delta/T^2)), \dim_E(\mathcal{F},1/T), \log(\mathcal{N}(\mathcal{S}\times\mathcal{A},\delta/T^2)))$. *Then the output policy is guaranteed to be* $\widetilde{O}(H^3\cdot d(T,\delta)^2/\sqrt{K})$*-optimal with high probability. In the tabular case we have* $d(T,\delta)=O(|\mathcal{S}||\mathcal{A}|\cdot\operatorname{poly}\log(|\mathcal{S}||\mathcal{A}|T\delta^{-1}))=\widetilde{O}(|\mathcal{S}||\mathcal{A}|)$. *When* $\mathcal{F}$ *and* $\mathcal{R}$ *are both the class of d-dimensional linear functions we have* $d(T,\delta)=O(d\cdot\operatorname{poly}\log(dT\delta^{-1}))=\widetilde{O}(d)$. *However, it is hard to rigorously show this kind of property when* $\mathcal{F}$ *and* $\mathcal{R}$ *are both general function classes. Generally speaking, if* $d(T,\delta)=O(d^*\cdot\operatorname{poly}\log(d^*T\delta^{-1}))=\widetilde{O}(d^*)$ *where* d^* *depends only on the complexity of* $\mathcal{F}$ *and* $\mathcal{R}$*, then the output policy is guaranteed to be* $\widetilde{O}(H^3\cdot(d^*)^2/\sqrt{K})$*-optimal with high probability. Thus for any* $\epsilon>0$*, by taking* $K=C\cdot(d^*)^4H^6\cdot\epsilon^{-2}\cdot\operatorname{polylog}(d^*T\delta^{-1}\epsilon^{-1})$ *where* $C>0$ *is a sufficiently large constant, our algorithm guarantees to output an* ϵ*-optimal policy after exploring the environment for* $\widetilde{O}((d^*)^4H^6\epsilon^{-2})$ *episodes. In this case, our sample complexity bound and switching cost bound become* $\widetilde{O}((d^*)^4H^6\epsilon^{-2})$ *and* $\widetilde{O}((d^*)^2H)$. *In particular, when* $\mathcal{F}$ *and* $\mathcal{R}$ *are the class of d-dimensional linear functions, our sample complexity bound can be improved to* $\widetilde{O}(d^3H^6\epsilon^{-2})$ *with refined analysis using the technique mentioned in Section B, matching the bound given in Wang et al. (2020b).*

D. Computational Efficiency

In this section we show how to implement our algorithms efficiently by assuming access to the following Regression Oracle. We remark that this is a mild assumption since the regression problem is common in machine learning practice and can usually be solved efficiently. This assumption also commonly appears in the literature (Foster et al., 2018; Foster & Rakhlin, 2020; Foster et al., 2020).

Regression Oracle. We assume access to a weighted least-squares regression oracle over the function class $\mathcal{G}$, which takes a set U of weighted examples $(w,z,y)\in\mathbb{R}_+\times(\mathcal{S}\times\mathcal{A})\times\mathbb{R}$ as input, and outputs the function with the smallest weighted squared loss:

$$\text{ORACLE}(H,\mathcal{G})=\operatorname{argmin}_{g\in\mathcal{G}}\sum_{(w,z,y)\in U}w(g(z)-y)^2.$$

Furthermore, we assume that the time cost of an oracle call grows linear in $|U|$. Given this oracle, one can solve the following optimization problem efficiently using time proportional to the number of distinct elements in $\mathcal{Z}$ by directly calling the oracle with $\mathcal{G}=\mathcal{F}-\mathcal{F}:=\{f_1-f_2|f_1,f_2\in\mathcal{F}\}$:

$$\min\|f_1-f_2\|_{\mathcal{Z}}^2+\frac{w}{2}(f_1(s,a)-f_2(s,a)-2H)^2 \tag{2}$$

where $w>0$ is a parameter. Such a procedure can be then used to solve the following constrained optimization problem approximately:

$$\max f_1(s,a)-f_2(s,a)\text{ s.t. }\|f_1-f_2\|_{\mathcal{Z}}^2\le\epsilon,\ f_1,f_2\in\mathcal{F}. \tag{3}$$

We adopt ideas from Foster et al. (2018): we do binary search over w to find the proper value of w and then the solution of (2) gives an approximate solution to (3). The full algorithm is presented in Algorithm 5. The following theorem shows that when $\mathcal{F}$ is convex, algorithm 5 solves the constrained optimization problem up to a precision of α in $O(\log(1/\alpha))$ iterations, i.e., $O(\log(1/\alpha))$ oracle invocations. When $\mathcal{F}$ is not convex, the constrained optimization problem can be solved with $O(1/\alpha)$ oracle invocations using the techniques in (Krishnamurthy et al., 2017).

Theorem 3. *Assume that the optimal solution to the following constrained optimization problem is* $g^*=f_1^*-f_2^*$.

$$\textit{maximize } f_1(s,a)-f_2(s,a)$$

$$\textit{subject to } \|f_1-f_2\|_{\mathcal{Z}}^2\le\beta,\quad f_1,f_2\in\mathcal{F}$$

We run algorithm 5 to solve the above problem. If the function class $\mathcal{F}$ *is convex and closed under pointwise convergence, then algorithm 5 terminate after* $O(\log(1/\alpha))$ *oracle invocations and the returned values satisfy*

$$|z_H-g^*(s,a)|\le\alpha.$$

Algorithm 5 Binary Search

Input: Dataset $\mathcal{Z}$, objective (s, a), tolerance β, precision α
$\mathcal{G} \leftarrow \mathcal{F} - \mathcal{F}$
$R(g, w) := \|g\|_{\mathcal{Z}}^2 + \frac{w}{2}(g(s, a) - 2(H + 1))^2, \forall g \in \mathcal{G}$
$w_L \leftarrow 0, w_H \leftarrow \beta/(\alpha(H + 1))$
$g_L \leftarrow 0, z_L \leftarrow 0$
$g_H \leftarrow \text{argmin}_{g\in\mathcal{G}} R(g, w_H), z_H \leftarrow g_H(s, a)$
$\Delta \leftarrow \alpha\beta/(8(H + 1)^3)$
while $|z_H - z_L| > \alpha$ and $|w_H - w_L| > \Delta$ **do**
  $\widetilde{w} \leftarrow (w_H + w_L)/2$
  $\widetilde{g} \leftarrow \text{argmin}_{g\in\mathcal{G}} R(g, \widetilde{w}), \widetilde{z} \leftarrow \widetilde{g}(s, a)$
  if $\|\widetilde{g}\|_{\mathcal{Z}}^2 > \beta$ **then**
    $w_H \leftarrow \widetilde{w}, z_H \leftarrow \widetilde{z}$
  else
    $w_L \leftarrow \widetilde{w}, z_L \leftarrow \widetilde{z}$
  end if
end while
Output: z_H

Note that if $\mathcal{F}$ is convex, then $\mathcal{F} - \mathcal{F}$ is also convex due to the following equation:

$$\lambda(f_1 - f_2) + (1 - \lambda)(f_3 - f_4) = (\lambda f_1 + (1 - \lambda)f_3) - (\lambda f_2 + (1 - \lambda)f_4).$$

The rest of the proof is identical to Theorem 1 of (Foster et al., 2018). We omit it here for brevity.

Computing Exploration Bonus and Sensitivity. We now show how to reduce the computation of the exploration bonus and online sensitivity scores to the constrained optimization problem in (3). Given a dataset $\mathcal{Z}$ and a state-action pair (s, a), the exploration bonus is essentially the solution to the following constrained optimization problem:

$$\max f_1(s, a) - f_2(s, a) \text{ s.t. } \|f_1 - f_2\|_{\mathcal{Z}}^2 \le \beta,\ f_1, f_2 \in \mathcal{F},$$

which can be easily reduced to the constrained optimization problem in (3). On the other hand, the estimation of the online sensitivity score can be reduced to the following optimization problems:

$$\max f_1(s, a) - f_2(s, a) \text{ s.t. } \|f_1 - f_2\|_{\mathcal{Z}}^2 \le 2^\alpha,\ f_1, f_2 \in \mathcal{F}$$

for $\alpha \in \{0, 1, ..., \log(T(H + 1)^2), +\infty\}$. Indeed, assuming the solution of the above problem is f_1^α, f_2^α for some α, and let

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

We then have $1 \le \mathsf{sensitivity}_{\mathcal{Z},\mathcal{F}}(z)/\mathsf{sensitivity}^{\mathsf{est}}_{\mathcal{Z},\mathcal{F}}(z) \le 2$. Note that a 2-approximation of the sensitivity score is sufficient for our analysis. Hence, both the exploration bonus and the sensitivity scores can be computed in time proportional to the size of $\widehat{\mathcal{Z}}_h^k$, i.e., $\widetilde{O}(d^2)$.

Time Complexity of the Algorithm. For the online sub-sampling procedure (Algorithm 2), as the online sensitivity score is computed using the sub-sampled dataset and the size of the sub-sampled dataset is bounded by $\widetilde{O}(d^2)$, this step takes $\widetilde{O}(\text{poly}(dH))$ time per episode. For all $h \in [H]$, the computation of f_h^k can be done efficiently by directly calling the regression oracle with $\mathcal{G} = \mathcal{F}$. Since we only compute f_h^k when the sub-sampled dataset $\widehat{\mathcal{Z}}_h^k$ is changed, and each $\widehat{\mathcal{Z}}_h^k$ will be changed for at most $\widetilde{O}(d^2)$ times, this step takes time $\widetilde{O}(\text{poly}(dH) \cdot |\mathcal{A}|)$ on average for each episode. Similarly, the computation of the exploration bonus also takes time $\widetilde{O}(\text{poly}(dH) \cdot |\mathcal{A}|)$ on average.

E. Proof Sketch

Now we present the major steps for proving Theorem 1. The detailed Proof is given in the Section F and Section G.

The Online Sub-Sampling Algorithm. In order to establish the correctness and effectiveness of our algorithm, we need to show that (i) the sub-sampled datasets $\widehat{\mathcal{Z}}_h^k$ always have bounded size and (ii) each $\widehat{\mathcal{Z}}_h^k$ provides a good approximation to $\mathcal{Z}_h^k$. In our proof (Proposition 2), we first show that the summation of the online sensitivity scores is upper bounded by $\widetilde{O}(d^2)$ if we use $\mathcal{Z}_h^k$ (the original dataset) instead of $\widehat{\mathcal{Z}}_h^k$ (the sub-sampled dataset) to calculate the online sensitivity scores. This is established by a combinatorial argument (cf. Lemma 5) which draws a connection between the eluder dimension and the summation of the online sensitivity scores. However, in our algorithm, for efficiency considerations, we use $\widehat{\mathcal{Z}}_h^k$ instead of $\mathcal{Z}_h^k$ to calculate the sensitivity scores. Fortunately, as we will show, $\widehat{\mathcal{Z}}_h^k$ provides an accurate estimation to $\mathcal{Z}_h^k$. Moreover, thanks to the design of the online sensitivity scores, their summation is robust to perturbations on the datasets. Hence, the summation of the sensitivity scores can be bounded even if $\widehat{\mathcal{Z}}_h^k$ is used in replace of $\mathcal{Z}_h^k$. Note that the summation of the sensitivity scores provides an upper bound on the expected size of the sub-sampled dataset, and a high probability bound can be easily obtained by using martingale concentration bounds.

In order to show that the sub-sampled dataset provides a good approximation to the original dataset, in our proof (Proposition 1), we show that the confidence set induced by the sub-sampled dataset is close to that induced by the original dataset. To prove this, for each pair of $f_1, f_2 \in \mathcal{F}$, we show that $\|f_1 - f_2\|_{\widehat{\mathcal{Z}}_h^k}$ to close to $\|f_1 - f_2\|_{\mathcal{Z}_h^k}$, and therefore the confidence set is approximately preserved. In order to show that $\|f_1 - f_2\|_{\widehat{\mathcal{Z}}_h^k}$ is close to $\|f_1 - f_2\|_{\mathcal{Z}_h^k}$, we note that sub-sampling proportional to online sensitivity scores implies that the estimator is unbiased and has low variance, and thus the desired result follows by Bernstein-type martingale concentration bounds.

Regret Decomposition. The proof of Theorem 1 consists of a standard regret decomposition step that decomposes the regret as the summation of the exploration bonus which is common in optimistic algorithms. Here, one core step is to show optimism, which requires that our exploration bonus upper bounds the estimation error. To show this, one needs to show that the least-squares estimator f_h^k is close to $\mathcal{T}Q_{h+1}^k$, where $\mathcal{T}$ is the Bellman backup operator. To tackle the dependence between the collected samples and Q_{h+1}^k, we apply a uniform convergence argument which also appears in Jin et al. (2020b); Wang et al. (2020c). Here, we need to build a cover over all possible estimated Q-functions Q_h^k. This is possible since the sub-sampled dataset that defines the exploration bonus b_h^k has bounded complexity. Once optimism is established, we can then use the mechanism developed in Russo & Roy (2013) to bound the summation of the exploration bonus in terms of the eluder dimension.

The Switching Cost. In order to achieve low switching cost, we change the policy only when the sub-sampled dataset is changed. This is best understood from a regret decomposition perspective. Note that our exploration bonus depends only on the current state-action pair and the sub-sampled dataset $\widehat{\mathcal{Z}}_h^k$, and therefore, if the sub-sampled dataset $\widehat{\mathcal{Z}}_h^k$ does not change, the exploration bonus will also be unchanged. Hence, in terms of the summation of the exploration bonus, our algorithm (which switches policy only when the sub-sampled dataset is unchanged) is as good as the ideal algorithm which changes its policy in each round. For the ideal algorithm, the summation of the exploration bonus can be upper bounded in terms of the eluder dimension using ideas in Russo & Roy (2013). Thus, the regret of our algorithm can also be upper bounded in terms of the eluder dimension.

F. Properties of Online Sub-Sampling

In this section we first formalize the properties of the online sub-sampling procedure. We then give complete proofs of these properties.

F.1. Choice of the parameter β

We elaborate the choice of β. In our algorithms, β is used in the computation of the bonus function and the online sensitivity score (1). In the standard RL setting (Algorithm 1), we set β to be

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

In the reward-free RL setting (Algorithm 3 and Algorithm 4), we modify the value of β to be

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

F.2. Analysis and Propositions

As mentioned in Section 3.1, we need to show that the sub-sampled dataset $\widehat{\mathcal{Z}}_h^k$

- provides a good approximation to $\mathcal{Z}_h^k$; and
- has a much lower complexity than $\mathcal{Z}_h^k$ (in terms of number of distinct elements).

We define the following enlarged and shrunk confidence sets. For all $(k,h) \in [K] \times [H]$ and $\alpha \in [\beta, +\infty)$, define

$$\underline{\mathcal{B}}_h^k(\alpha) := \{(f_1, f_2) \in \mathcal{F} \times \mathcal{F} | \|f_1 - f_2\|_{\mathcal{Z}_h^k}^2 \le \alpha/100\},$$

$$\mathcal{B}_h^k(\alpha) := \{(f_1, f_2) \in \mathcal{F} \times \mathcal{F} | \min\{\|f_1 - f_2\|_{\widehat{\mathcal{Z}}_h^k}^2, T(H+1)^2\} \le \alpha\},$$

$$\overline{\mathcal{B}}_h^k(\alpha) := \{(f_1, f_2) \in \mathcal{F} \times \mathcal{F} | \|f_1 - f_2\|_{\mathcal{Z}_h^k}^2 \le 100\alpha\}.$$

For each $(k,h) \in [K] \times [H]$, we use $\mathcal{E}_h^k(\alpha)$ to denote the event that

$$\underline{\mathcal{B}}_h^k(\alpha) \subseteq \mathcal{B}_h^k(\alpha) \subseteq \overline{\mathcal{B}}_h^k(\alpha).$$

Furthermore, we denote that

$$\mathcal{E}_h^k := \bigcap_{n=0}^{\infty} \mathcal{E}_h^k(100^n \beta).$$

Event $\mathcal{E}_h^k$ characterizes the meaning of "good approximation". In fact, if $\mathcal{E}_h^k$ happens, we can show that $\|f_1 - f_2\|_{\widehat{\mathcal{Z}}_h^k}$ is close to $\|f_1 - f_2\|_{\mathcal{Z}_h^k}$ up to a constant factor, thus the confidence set induced by $\widehat{\mathcal{Z}}_h^k$ is accurate. The following proposition verifies that $\mathcal{E}_h^k$ happens with high probability.

Proposition 1.

$$\Pr\left(\bigcap_{h=1}^{H}\bigcap_{k=1}^{K} \mathcal{E}_h^k\right) \ge 1 - \delta/32.$$

Moreover, we define the following bonus functions calculated by $\mathcal{Z}_h^k$ instead of $\widehat{\mathcal{Z}}_h^k$:

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

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

If $\mathcal{E}_h^k$ happens, by taking $\alpha = \beta$, we have that

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

which verifies the correctness of our bonus function b_h^k used in the algorithm.

Proposition 2 bounds the size of $\widehat{\mathcal{Z}}_h^k$.

Proposition 2. *With probability at least $1 - \delta/8$, the following statements hold:*

1. For any fixed $h \in [H]$, the subsampled dataset $\widehat{\mathcal{Z}}_h^k$ ($k = 1, 2, ..., K$) changes for at most

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

times for some absolute constant $C > 0$.

As a result, for any pair $(k,h) \in [K] \times [H]$, $n_d(\widehat{\mathcal{Z}}_h^k) \le S_{\max}$.

2. For any $(h,k) \in [H] \times [K]$,

$$|\widehat{\mathcal{Z}}_h^k| \le 64T^3/\delta.$$

In the following two sections we prove Proposition 1 and Proposition 2. Throughout the proof, We use $\mathscr{F}_k$ to denote the filtration induced by the history up to episode k (include episode k) and use $\mathbb{E}_k$ to denote the expectation conditioned on $\mathscr{F}_k$.

F.3. Proof of Proposition 1

For completeness, we state a Bernstein-type martingale concentration inequality which will be frequently used in our proofs.

Lemma 1 ((Freedman, 1975)). *Consider a real-valued martingale $\{Y_k : k = 0, 1, 2...\}$ with difference sequence $\{X_k : k = 1, 2, ...\}$. Assume that the difference sequence is uniformly bounded:*

$$|X_k| \leq R \text{ almost surely for } k = 1, 2, 3, ...$$

For a fixed $n \in \mathbb{N}$, assume that

$$\sum_{k=1}^{n} \mathbb{E}_{k-1}(X_k^2) \leq \sigma^2$$

almost surely. Then for all $t \geq 0$,

$$P\{|Y_n - Y_0| \geq t\} \leq 2 \exp\left\{-\frac{t^2/2}{\sigma^2 + Rt/3}\right\}.$$

The next lemma upper bounds the size of $\widehat{\mathcal{Z}}_h^k$.

Lemma 2. *With probability at least $1 - \delta/64T$,*

$$|\widehat{\mathcal{Z}}_h^k| \leq 64T^3/\delta \quad \forall (k, h) \in [K] \times [H].$$

Proof. Consider a fixed pair $(k, h) \in [K] \times [H]$. By Markov's inequality we have that

$$|\widehat{\mathcal{Z}}_h^k| \leq 64T^2 |\mathcal{Z}_h^k|/\delta$$

holds with probability at least $1 - \delta/(64T^2)$. With a union bound for all $(k, h) \in [K] \times [H]$ we complete the proof. □

Now we start to analyze the events defined in Section F.2. Recall that in Section F.2 we mentioned that event $\mathcal{E}_h^k$ characterizes the meaning of "good approximation". Our next lemma formalizes this intuition.

Lemma 3. *If $\mathcal{E}_h^k$ happens, then*

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

and

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

Proof. If $\|f_1 - f_2\|_{\mathcal{Z}_h^k}^2 \leq 100\beta$, we have $(f_1, f_2) \in \underline{\mathcal{B}}_h^k(10000\beta)$. From $\mathcal{E}_h^k$ we know $(f_1, f_2) \in \mathcal{B}_h^k(10000\beta)$, which implies the desired result.

If $\|f_1 - f_2\|_{\mathcal{Z}_h^k}^2 > 100\beta$, assume that $100^n\beta < \|f_1 - f_2\|_{\mathcal{Z}_h^k}^2 \leq 100^{n+1}\beta$, $n \in \mathbb{N}^*$. Then we have $(f_1, f_2) \notin \overline{\mathcal{B}}_h^k(100^{n-1}\beta)$ and also $(f_1, f_2) \notin \mathcal{B}_h^k(100^{n-1}\beta)$. This implies that $\min\{\|f_1 - f_2\|_{\widehat{\mathcal{Z}}_h^k}^2, T(H+1)^2\} \geq 100^{n-1}\beta \geq \frac{1}{10000}\|f_1 - f_2\|_{\mathcal{Z}_h^k}^2$. Similarly, we have $(f_1, f_2) \in \underline{\mathcal{B}}_h^k(100^{n+2}\beta)$, then also $(f_1, f_2) \in \mathcal{B}_h^k(100^{n+2}\beta)$. Thus we have $\min\{\|f_1 - f_2\|_{\widehat{\mathcal{Z}}_h^k}^2, T(H+1)^2\} \leq 100^{n+2}\beta \leq 10000\|f_1 - f_2\|_{\mathcal{Z}_h^k}^2$. □

Recall that Proposition 1 states that $\bigcap_{h=1}^{H}\bigcap_{k=1}^{K}\mathcal{E}_h^k$ happens with high probability. As will be shown in the proof of Proposition 1 later, to bound the probability of $\bigcap_{h=1}^{H}\bigcap_{k=1}^{K}\mathcal{E}_h^k$ we only need to bound $\Pr\left(\mathcal{E}_h^1\mathcal{E}_h^2...\mathcal{E}_h^{k-1}(\mathcal{E}_h^k(100^n\beta))^c\right)$. Note that $\mathcal{E}_h^k(100^n\beta)$ always holds if $100^n\beta \geq T(H+1)^2$. We establish the following lemma.

Lemma 4. *For $\alpha \in [\beta, T(H+1)^2]$,*

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

Proof. We use $\overline{\mathcal{Z}}_h^k$ to denote the dataset without rounding, i.e., we replace every element $\hat{z}$ with z in $\widehat{\mathcal{Z}}_h^k$. Denote $C_1 := C \cdot \log(T\mathcal{N}(\mathcal{F}, \sqrt{\delta/64T^3})/\delta)$ to be the parameter used in Algorithm 2.

Consider a fixed pair $(f_1, f_2) \in \mathcal{C}(\mathcal{F}, \sqrt{\delta/(64T^3)}) \times \mathcal{C}(\mathcal{F}, \sqrt{\delta/(64T^3)})$.

For each $i \geq 2$, define

$$Z_i = \max\{\|f_1 - f_2\|^2_{\mathcal{Z}_h^i}, \min\{\|f_1 - f_2\|^2_{\widehat{\mathcal{Z}}_h^{i-1}}, T(H+1)^2\}\}$$

and

$$Y_i = \begin{cases} \frac{1}{p_{z_h^{i-1}}}(f_1(z_h^{i-1}) - f_2(z_h^{i-1}))^2 & z_h^{i-1} \text{ is added into } \overline{\mathcal{Z}}_h^i \text{ and } Z_i \leq 2000000\alpha \\ 0 & z_h^{i-1} \text{ is not added into } \overline{\mathcal{Z}}_h^i \text{ and } Z_i \leq 2000000\alpha \\ (f_1(z_h^{i-1}) - f_2(z_h^{i-1}))^2 & Z_i > 2000000\alpha \end{cases}.$$

Y_i's are used to characterize the sampling procedure. Note that Y_i is adapted to the filtration $\mathscr{F}_i$, and $\mathbb{E}_{i-1}[Y_i] = (f_1(z_h^{i-1}) - f_2(z_h^{i-1}))^2$. In order to use Freedman's inequality, we need to bound Y_i and its variance.

If $p_{z_h^{i-1}} = 1$ or $\min\{\|f_1 - f_2\|^2_{\widehat{\mathcal{Z}}_h^{i-1}}, T(H+1)^2\} > 2000000\alpha$, then $Y_i - \mathbb{E}_{i-1}[Y_i] = \text{Var}_{i-1}[Y_i - \mathbb{E}_{i-1}[Y_i]] = 0$. Otherwise from the definition of p_z in Algorithm 2 we have that:

$$\begin{aligned} |Y_i - \mathbb{E}_{i-1}[Y_i]| &\leq (\min\{\|f_1 - f_2\|^2_{\widehat{\mathcal{Z}}_h^{i-1}}, T(H+1)^2\} + \beta) \cdot 1/C_1 \\ &\leq 3000000\alpha/C_1 \end{aligned}$$

and

$$\begin{aligned} \text{Var}_{i-1}[Y_i - \mathbb{E}_{i-1}[Y_i]] &\leq \frac{1}{p_{z_h^{i-1}}}(f_1(z_h^{i-1}) - f_2(z_h^{i-1}))^4 \\ &\leq (f_1(z_h^{i-1}) - f_2(z_h^{i-1}))^2 \cdot 3000000\alpha/C_1. \end{aligned}$$

It is easy to verify that

$$\sum_{i=2}^{k} \text{Var}_{i-1}[Y_i - \mathbb{E}_{i-1}[Y_i]] \leq (3000000\alpha)^2/C_1.$$

By Freedman's inequality (Lemma 1), we have that

$$\begin{aligned} &\Pr\left\{\left|\sum_{i=2}^{k}(Y_i - \mathbb{E}_{i-1}[Y_i])\right| \geq \alpha/100\right\} \\ &\leq 2\exp\left\{-\frac{(\alpha/100)^2/2}{(3000000\alpha)^2/C_1 + \alpha \cdot 3000000\alpha/3C_1}\right\} \\ &\leq (\delta/64T^2)/(\mathcal{N}(\mathcal{F}, \sqrt{\delta/(64T^3)}))^2. \end{aligned}$$

With a union bound, the above inequality implies that with probability at least $1 - \delta/(64T^2)$, for any pair $(f_1, f_2) \in \mathcal{C}(\mathcal{F}, \sqrt{\delta/(64T^3)}) \times \mathcal{C}(\mathcal{F}, \sqrt{\delta/(64T^3)})$, the corresponding Y_i's satisfy

$$\left|\sum_{i=2}^{k}(Y_i - \mathbb{E}_{i-1}[Y_i])\right| \leq \alpha/100.$$

Now we condition on the above event and the event defined in Lemma 2 for the rest of the proof.

Part 1: ($\underline{\mathcal{B}}_h^k(\alpha) \subseteq \mathcal{B}_h^k(\alpha)$) Consider any pair $f_1, f_2 \in \mathcal{F}$ with $\|f_1 - f_2\|^2_{\mathcal{Z}_h^k} \leq \alpha/100$. From the definition we know that there exist $(\hat{f}_1, \hat{f}_2) \in \mathcal{C}(\mathcal{F}, \sqrt{\delta/(64T^3)}) \times \mathcal{C}(\mathcal{F}, \sqrt{\delta/(64T^3)})$ such that $\|\hat{f}_1 - f_1\|_\infty, \|\hat{f}_2 - f_2\|_\infty \leq \sqrt{\delta/(64T^3)}$. Then

we have that

$$\begin{aligned}\|\hat{f}_1-\hat{f}_2\|^2_{\mathcal{Z}_h^k} &\le (\|f_1-f_2\|_{\mathcal{Z}_h^k}+\|f_1-\hat{f}_1\|_{\mathcal{Z}_h^k}+\|\hat{f}_2-f_2\|_{\mathcal{Z}_h^k})^2\\ &\le (\|f_1-f_2\|_{\mathcal{Z}_h^k}+2\cdot\sqrt{|\mathcal{Z}_h^k|}\cdot\sqrt{\delta/(64T^3)})^2\\ &\le \alpha/50.\end{aligned}$$

We consider the Y_i's which corrspond to $\hat{f}_1$ and $\hat{f}_2$. Because $\|\hat{f}_1-\hat{f}_2\|^2_{\mathcal{Z}_h^k}\le\alpha/50$, we also have that $\|\hat{f}_1-\hat{f}_2\|^2_{\mathcal{Z}_h^{k-1}}\le\alpha/50$. From $\mathcal{E}_h^{k-1}$ we know that $\min\{\|\hat{f}_1-\hat{f}_2\|^2_{\widehat{\mathcal{Z}}_h^{k-1}},T(H+1)^2\}\le 10000\alpha$. Then from the definition of Y_i we have

$$\|\hat{f}_1-\hat{f}_2\|^2_{\overline{\mathcal{Z}}_h^k}=\sum_{i=2}^k Y_i.$$

Then $\|\hat{f}_1-\hat{f}_2\|^2_{\overline{\mathcal{Z}}_h^k}$ can be bounded in the following manner:

$$\begin{aligned}\|\hat{f}_1-\hat{f}_2\|^2_{\overline{\mathcal{Z}}_h^k} &= \sum_{i=2}^k Y_i\\ &\le \sum_{i=2}^k \mathbb{E}_{i-1}[Y_i]+\alpha/100\\ &= \|\hat{f}_1-\hat{f}_2\|^2_{\mathcal{Z}_h^k}+\alpha/100\\ &\le 3\alpha/100.\end{aligned}$$

As a result, $\|f_1-f_2\|^2_{\overline{\mathcal{Z}}_h^k}$ can also be bounded:

$$\begin{aligned}\|f_1-f_2\|^2_{\overline{\mathcal{Z}}_h^k} &\le (\|\hat{f}_1-\hat{f}_2\|_{\overline{\mathcal{Z}}_h^k}+\|f_1-\hat{f}_1\|_{\overline{\mathcal{Z}}_h^k}+\|\hat{f}_2-f_2\|_{\overline{\mathcal{Z}}_h^k})^2\\ &\le (\|\hat{f}_1-\hat{f}_2\|_{\overline{\mathcal{Z}}_h^k}+2\cdot\sqrt{|\overline{\mathcal{Z}}_h^k|}\cdot\sqrt{\delta/(64T^3)})^2\\ &\le \alpha/25.\end{aligned}$$

Finally we could bound $\|f_1-f_2\|^2_{\widehat{\mathcal{Z}}_h^k}$:

$$\begin{aligned}\|f_1-f_2\|^2_{\widehat{\mathcal{Z}}_h^k} &\le (\|f_1-f_2\|_{\overline{\mathcal{Z}}_h^k}+\sqrt{64T^3/\delta}/(8\sqrt{64T^3/\delta}))^2\\ &\le \alpha.\end{aligned}$$

We conclude that for any pair $f_1,f_2\in\mathcal{F}$ with $\|f_1-f_2\|^2_{\mathcal{Z}_h^k}\le\alpha/100$, it holds that $\|f_1-f_2\|^2_{\widehat{\mathcal{Z}}_h^k}\le\alpha$. Thus we must have $\underline{\mathcal{B}}_h^k(\alpha)\subseteq\mathcal{B}_h^k(\alpha)$.

Part 2: ($\mathcal{B}_h^k(\alpha)\subseteq\overline{\mathcal{B}}_h^k(\alpha)$) Consider any pair $f_1,f_2\in\mathcal{F}$ with $\|f_1-f_2\|^2_{\mathcal{Z}_h^k}>100\alpha$. From the definition we know that there exist $(\hat{f}_1,\hat{f}_2)\in\mathcal{C}(\mathcal{F},\sqrt{\delta/(64T^3)})\times\mathcal{C}(\mathcal{F},\sqrt{\delta/(64T^3)})$ such that $\|\hat{f}_1-f_1\|_\infty,\|\hat{f}_2-f_2\|_\infty\le\sqrt{\delta/(64T^3)}$. Then we have that

$$\begin{aligned}\|\hat{f}_1-\hat{f}_2\|^2_{\mathcal{Z}_h^k} &\ge (\|f_1-f_2\|_{\mathcal{Z}_h^k}-\|f_1-\hat{f}_1\|_{\mathcal{Z}_h^k}-\|\hat{f}_2-f_2\|_{\mathcal{Z}_h^k})^2\\ &\ge (\|f_1-f_2\|_{\mathcal{Z}_h^k}-2\cdot\sqrt{|\mathcal{Z}_h^k|}\cdot\sqrt{\delta/(64T^3)})^2\\ &> 50\alpha.\end{aligned}$$

Thus we have $\|\hat{f}_1-\hat{f}_2\|^2_{\mathcal{Z}_h^k}>50\alpha$. We consider the Y_i's which corrspond to $\hat{f}_1$ and $\hat{f}_2$. Here we want to prove that $\|\hat{f}_1-\hat{f}_2\|^2_{\widehat{\mathcal{Z}}_h^k}>40\alpha$. For the sake of contradicition we assume that $\|\hat{f}_1-\hat{f}_2\|^2_{\widehat{\mathcal{Z}}_h^k}\le 40\alpha$.

Case 1: $\|\hat{f}_1 - \hat{f}_2\|^2_{\mathcal{Z}_h^k} \leq 2000000\alpha$**.** From the definition of Y_i we have that

$$\|\hat{f}_1 - \hat{f}_2\|^2_{\overline{\mathcal{Z}}_h^k} = \sum_{i=2}^{k} Y_i.$$

Combined with the former result, we conclude that

$$\|\hat{f}_1 - \hat{f}_2\|^2_{\overline{\mathcal{Z}}_h^k} = \sum_{i=2}^{k} Y_i \geq \sum_{i=2}^{k} \mathbb{E}_{i-1}[Y_i] - \alpha/100 = \|\hat{f}_1 - \hat{f}_2\|^2_{\mathcal{Z}_h^k} - \alpha/100 > 50\alpha - \alpha/100 = 49\alpha.$$

Then we have

$$\begin{aligned}\|\hat{f}_1 - \hat{f}_2\|^2_{\widehat{\mathcal{Z}}_h^k} &\geq (\|\hat{f}_1 - \hat{f}_2\|_{\overline{\mathcal{Z}}_h^k} - \sqrt{64T^3/\delta}/(8\sqrt{64T^3/\delta}))^2 \\ &> 40\alpha.\end{aligned}$$

This leads to a contradiction.

Case 2: $\|\hat{f}_1 - \hat{f}_2\|^2_{\mathcal{Z}_h^{k-1}} > 1000000\alpha$**.** From $\mathcal{E}_h^{k-1}$ we deduce that $\|\hat{f}_1 - \hat{f}_2\|^2_{\widehat{\mathcal{Z}}_h^{k-1}} > 100\alpha$ which directly leads to a contradiction.

Case 3: $\|\hat{f}_1 - \hat{f}_2\|^2_{\mathcal{Z}_h^k} > 2000000\alpha$ **and** $\|\hat{f}_1 - \hat{f}_2\|^2_{\mathcal{Z}_h^{k-1}} \leq 1000000\alpha$**.** It is clear that $(\hat{f}_1(z_h^{k-1}) - \hat{f}_2(z_h^{k-1}))^2 > 1000000\alpha$. From the definition of sensitivity we know that z_h^{k-1} will be added into $\overline{\mathcal{Z}}_h^k$ almost surely. This clearly leads to a contradiction. We conclude that $\|\hat{f}_1 - \hat{f}_2\|^2_{\widehat{\mathcal{Z}}_h^k} > 40\alpha$.

Finally we could bound $\|f_1 - f_2\|^2_{\widehat{\mathcal{Z}}_h^k}$:

$$\begin{aligned}\|f_1 - f_2\|^2_{\widehat{\mathcal{Z}}_h^k} &\geq (\|\hat{f}_1 - \hat{f}_2\|_{\widehat{\mathcal{Z}}_h^k} - \|f_1 - \hat{f}_1\|_{\widehat{\mathcal{Z}}_h^k} - \|\hat{f}_2 - f_2\|_{\widehat{\mathcal{Z}}_h^k})^2 \\ &\geq (\|\hat{f}_1 - \hat{f}_2\|_{\widehat{\mathcal{Z}}_h^k} - 2 \cdot \sqrt{|\widehat{\mathcal{Z}}_h^k|} \cdot \sqrt{\delta/(64T^3)})^2 \\ &> \alpha.\end{aligned}$$

We conclude that for any pair $f_1, f_2 \in \mathcal{F}$ with $\|f_1 - f_2\|^2_{\mathcal{Z}_h^k} > 10000\beta$, it holds that $\|f_1 - f_2\|^2_{\widehat{\mathcal{Z}}_h^k} > 100\beta$. This implies that $\mathcal{B}_h^k(\alpha) \subseteq \overline{\mathcal{B}}_h^k(\alpha)$. □

Proof of Proposition 1. Note that for all $(k, h) \in [K] \times [H]$, we have

$$\begin{aligned}&\Pr(\mathcal{E}_h^1\mathcal{E}_h^2...\mathcal{E}_h^{k-1}) - \Pr(\mathcal{E}_h^1\mathcal{E}_h^2...\mathcal{E}_h^k) \\ =&\Pr(\mathcal{E}_h^1\mathcal{E}_h^2...\mathcal{E}_h^{k-1}(\mathcal{E}_h^k)^c) \\ =&\Pr\left(\mathcal{E}_h^1\mathcal{E}_h^2...\mathcal{E}_h^{k-1}\left(\bigcap_{n=0}^{\infty}\mathcal{E}_h^k(100^n\beta)\right)^c\right) \\ =&\Pr\left(\mathcal{E}_h^1\mathcal{E}_h^2...\mathcal{E}_h^{k-1}\bigcup_{n=0}^{\infty}(\mathcal{E}_h^k(100^n\beta))^c\right) \\ \leq&\sum_{n=0}^{\infty}\Pr\left(\mathcal{E}_h^1\mathcal{E}_h^2...\mathcal{E}_h^{k-1}(\mathcal{E}_h^k(100^n\beta))^c\right) \\ =&\sum_{n\geq 0, 100^n\beta\leq T(H+1)^2}\Pr\left(\mathcal{E}_h^1\mathcal{E}_h^2...\mathcal{E}_h^{k-1}(\mathcal{E}_h^k(100^n\beta))^c\right). \end{aligned} \tag{4}$$

Combining Equation (4) and Lemma 4, we have that for all $(k,h) \in [K] \times [H]$,

$$\Pr(\mathcal{E}_h^1 \mathcal{E}_h^2 ... \mathcal{E}_h^{k-1}) - \Pr(\mathcal{E}_h^1 \mathcal{E}_h^2 ... \mathcal{E}_h^k) \leq \delta/(32T^2) \cdot (\log(T(H+1)^2/\beta) + 2) \leq \delta/32T.$$

Thus for all $h \in [H]$ we have

$$\begin{aligned}
&\Pr\left(\bigcap_{k=1}^{K} \mathcal{E}_h^k\right) \\
&= 1 - \sum_{k=1}^{K} (\Pr(\mathcal{E}_h^1 \mathcal{E}_h^2 ... \mathcal{E}_h^{k-1}) - \Pr(\mathcal{E}_h^1 \mathcal{E}_h^2 ... \mathcal{E}_h^k)) \\
&\geq 1 - K \cdot (\delta/32T) \\
&= 1 - \delta/32H.
\end{aligned}$$

By applying a union bound for all $h \in [H]$ we complete the proof. □

F.4. Proof of Proposition 2

We start our proof by showing that the summation of online sensitivity scores can be upper bounded if $\mathcal{Z}_h^k$, i.e, the dataset without sub-sampling, is used.

Lemma 5. *For all $h \in [H]$, we have*

$$\sum_{k=1}^{K-1} \text{sensitivity}_{\mathcal{Z}_h^k, \mathcal{F}}(z_h^k) \leq C \cdot \dim_E(\mathcal{F}, 1/T) \log((H+1)^2 T) \log T$$

for some absolute constant $C > 0$.

Proof. Note that $|\mathcal{Z}_h^k| \leq T$, thus we have that

$$\begin{aligned}
\mathsf{sensitivity}_{\mathcal{Z}_h^k, \mathcal{F}}(z_h^k) &= \min\left\{ \sup_{f_1, f_2 \in \mathcal{F}} \frac{(f_1(z_h^k) - f_2(z_h^k))^2}{\min\{\|f_1 - f_2\|_{\mathcal{Z}_h^k}^2, T(H+1)^2\} + \beta}, 1 \right\} \\
&\leq \min\left\{ \sup_{f_1, f_2 \in \mathcal{F}} \frac{(f_1(z_h^k) - f_2(z_h^k))^2}{\|f_1 - f_2\|_{\mathcal{Z}_h^k}^2 + 1}, 1 \right\}.
\end{aligned}$$

For each $k \in [K-1]$, let $f_1, f_2 \in \mathcal{F}$ be an arbitrary pair of functions such that

$$\frac{(f_1(z_h^k) - f_2(z_h^k))^2}{\|f_1 - f_2\|_{\mathcal{Z}_h^k}^2 + 1}$$

is maximized, and we define $L(z_h^k) = (f_1(z_h^k) - f_2(z_h^k))^2$ for such f_1 and f_2. Note that $0 \leq L(z_h^k) \leq (H+1)^2$. Let $\mathcal{Z}_h^K = \bigcup_{\alpha=0}^{\log((H+1)^2 T)-1} \mathcal{Z}^\alpha \cup \mathcal{Z}^\infty$ be a dyadic decomposition with respect to $L(\cdot)$ (we assume $\log((H+1)^2 T)$ is an integer for simplicity), where for each $0 \leq \alpha < \log((H+1)^2 T)$, define

$$\mathcal{Z}^\alpha = \{z_h^k \in \mathcal{Z}_h^K \mid L(z_h^k) \in ((H+1)^2 \cdot 2^{-\alpha-1}, (H+1)^2 \cdot 2^{-\alpha}]\}$$

and

$$\mathcal{Z}^\infty = \{z_h^k \in \mathcal{Z}_h^K \mid L(z_h^k) \leq 1/T\}.$$

Clearly, for any $z_h^k \in \mathcal{Z}^\infty$, $\mathsf{sensitivity}_{\mathcal{Z}_h^k, \mathcal{F}}(z_h^k) \leq 1/T$ and thus

$$\sum_{z_h^k \in \mathcal{Z}^\infty} \mathsf{sensitivity}_{\mathcal{Z}_h^k, \mathcal{F}}(z_h^k) \leq 1.$$

Now we bound $\sum_{z_h^k \in \mathcal{Z}^\alpha} \mathsf{sensitivity}_{\mathcal{Z}_h^k,\mathcal{F}}(z_h^k)$ for each $0 \le \alpha < \log((H+1)^2 T)$ separately. For each α, let

$$N_\alpha = |\mathcal{Z}^\alpha| / \dim_E(\mathcal{F}, (H+1)^2 \cdot 2^{-\alpha-1})$$

and we decompose $\mathcal{Z}^\alpha$ into $N_\alpha + 1$ disjoint subsets, i.e., $\mathcal{Z}^\alpha = \bigcup_{j=1}^{N_\alpha+1} \mathcal{Z}_j^\alpha$, by using the following procedure. We initialize $\mathcal{Z}_j^\alpha = \{\}$ for all j and consider each $z_h^k \in \mathcal{Z}^\alpha$ sequentially. For each $z_h^k \in \mathcal{Z}^\alpha$, we find the smallest $1 \le j \le N_\alpha$ such that z_h^k is $(H+1)^2 \cdot 2^{-\alpha-1}$-independent of $\mathcal{Z}_j^\alpha$ with respect to $\mathcal{F}$. We set $j = N_\alpha + 1$ if such j does not exist, and use $j(z_h^k) \in [N_\alpha + 1]$ to denote the choice of j for z_h^k. We then add z_h^k into $\mathcal{Z}_j^\alpha$. For each $z_h^k \in \mathcal{Z}^\alpha$, it is clear that z_h^k is dependent on each of $\mathcal{Z}_1^\alpha, \mathcal{Z}_2^\alpha, \ldots, \mathcal{Z}_{j(z_h^k)-1}^\alpha$.

Now we show that for each $z_h^k \in \mathcal{Z}^\alpha$,

$$\mathsf{sensitivity}_{\mathcal{Z}_h^k,\mathcal{F}}(z_h^k) \le \frac{4}{j(z_h^k)}.$$

For any $z_h^k \in \mathcal{Z}^\alpha$, we use $f_1, f_2 \in \mathcal{F}$ to denote the pair of functions in $\mathcal{F}$ such that

$$\frac{(f_1(z_h^k) - f_2(z_h^k))^2}{\|f_1 - f_2\|_{\mathcal{Z}_h^k}^2 + 1}$$

is maximized. Since $z_h^k \in \mathcal{Z}^\alpha$, we must have $(f_1(z_h^k) - f_2(z_h^k))^2 > (H+1)^2 \cdot 2^{-\alpha-1}$. Since z_h^k is dependent on each of $\mathcal{Z}_1^\alpha, \mathcal{Z}_2^\alpha, \ldots, \mathcal{Z}_{j(z_h^k)-1}^\alpha$, and for each $1 \le t < j(z_h^k)$, we have

$$\|f_1 - f_2\|_{\mathcal{Z}_t^\alpha} \ge (H+1)^2 \cdot 2^{-\alpha-1}.$$

It is important to note that $\mathcal{Z}_1^\alpha, \mathcal{Z}_2^\alpha, \ldots, \mathcal{Z}_{j(z_h^k)-1}^\alpha \subseteq \mathcal{Z}_h^k$ due to the design of the partition procedure. Thus the online sensitivity score can be bounded by:

$$\begin{aligned}
\mathsf{sensitivity}_{\mathcal{Z}_h^k,\mathcal{F}}(z_h^k) \le & \frac{(f_1(z_h^k) - f_2(z_h^k))^2}{\|f_1 - f_2\|_{\mathcal{Z}_h^k}^2 + 1} \le \frac{(H+1)^2 \cdot 2^{-\alpha}}{\|f_1 - f_2\|_{\mathcal{Z}_h^k}^2} \\
\le & \frac{(H+1)^2 \cdot 2^{-\alpha}}{\sum_{t=1}^{j(z_h^k)-1} \|f_1 - f_2\|_{\mathcal{Z}_t^\alpha}} \le 2/(j(z_h^k) - 1).
\end{aligned}$$

By the definition of online sensitivity score we have $\mathsf{sensitivity}_{\mathcal{Z}_h^k,\mathcal{F}}(z_h^k) \le 1$, thus we conclude that:

$$\mathsf{sensitivity}_{\mathcal{Z}_h^k,\mathcal{F}}(z_h^k) \le \min\{\frac{2}{j(z_h^k) - 1}, 1\} \le \frac{4}{j(z_h^k)}.$$

Moreover, by the definition of $(H+1)^2 \cdot 2^{-\alpha-1}$-independent, we have $|\mathcal{Z}_j^\alpha| \le \dim_E(\mathcal{F}, (H+1)^2 \cdot 2^{-\alpha-1})$ for all $1 \le j \le N_\alpha$. Therefore,

$$\begin{aligned}
& \sum_{z_h^k \in \mathcal{Z}^\alpha} \mathsf{sensitivity}_{\mathcal{Z}_h^k,\mathcal{F}}(z_h^k) \le \sum_{1 \le j \le N_\alpha} |\mathcal{Z}_j^\alpha| \cdot 4/j + \sum_{z \in \mathcal{Z}_{N_\alpha+1}^\alpha} 4/N_\alpha \\
\le & 4\dim_E(\mathcal{F}, (H+1)^2 \cdot 2^{-\alpha-1}) \ln(N_\alpha) + |\mathcal{Z}^\alpha| \cdot \frac{4\dim_E(\mathcal{F}, (H+1)^2 \cdot 2^{-\alpha-1})}{|\mathcal{Z}^\alpha|} \\
\le & 8\dim_E(\mathcal{F}, (H+1)^2 \cdot 2^{-\alpha-1}) \log T.
\end{aligned}$$

By the monotonicity of eluder dimension, it follows that

$$\begin{aligned}
& \sum_{k=1}^{K-1} \mathrm{sensitivity}_{\mathcal{Z}_h^k,\mathcal{F}}(z_h^k) \\
\le & \sum_{\alpha=0}^{\log((H+1)^2 T)-1} \sum_{z_h^k \in \mathcal{Z}^\alpha} \mathsf{sensitivity}_{\mathcal{Z}_h^k,\mathcal{F}}(z_h^k) + \sum_{z_h^k \in \mathcal{Z}^\infty} \mathsf{sensitivity}_{\mathcal{Z}_h^k,\mathcal{F}}(z_h^k) \\
\le & 8 \log((H+1)^2 T) \dim_E(\mathcal{F}, 1/T) \log T + 1 \\
\le & 9 \log((H+1)^2 T) \dim_E(\mathcal{F}, 1/T) \log T
\end{aligned}$$

as desired. □

With Lemma 5, now we are ready to prove Proposition 2.

Proof of Proposition 2. Firstly, note that conditioned on $\mathcal{E}_h^k$, which means $\widehat{\mathcal{Z}}_h^k$ is a good approximation to $\mathcal{Z}_h^k$, the online sensitivity score computed with $\widehat{\mathcal{Z}}_h^k$ will be relatively accurate. Formally, this argument can be stated as

$$\mathbb{I}\{\mathcal{E}_h^k\} \cdot \text{sensitivity}_{\widehat{\mathcal{Z}}_h^k,\mathcal{F}}(z_h^k) \leq C \cdot \text{sensitivity}_{\mathcal{Z}_h^k,\mathcal{F}}(z_h^k)$$

for some absolute constant $C > 0$. This property can be easily derived from Lemma 3.

Note that in Algorithm 2,[4]

$$p_z \lesssim \text{sensitivity}_{\mathcal{Z},\mathcal{F}}(z) \cdot \log(T\mathcal{N}(\mathcal{F}, \sqrt{\delta/64T^3})/\delta).$$

Thus from Lemma 5 we know that

$$\begin{aligned}
\sum_{k=1}^{K-1} \mathbb{I}\{\mathcal{E}_h^k\} \cdot p_{z_h^k} &\lesssim \sum_{k=1}^{K-1} \mathbb{I}\{\mathcal{E}_h^k\} \cdot \text{sensitivity}_{\widehat{\mathcal{Z}}_h^k,\mathcal{F}}(z_h^k) \cdot \log(T\mathcal{N}(\mathcal{F}, \sqrt{\delta/64T^3})/\delta) \\
&\lesssim \sum_{k=1}^{K-1} \text{sensitivity}_{\mathcal{Z}_h^k,\mathcal{F}}(z_h^k) \cdot \log(T\mathcal{N}(\mathcal{F}, \sqrt{\delta/64T^3})/\delta) \\
&\lesssim \log(T\mathcal{N}(\mathcal{F}, \sqrt{\delta/64T^3})/\delta) \dim_E(\mathcal{F}, 1/T) \log^2 T.
\end{aligned}$$

As we can adjust the constant C in Proposition 2, we assume that

$$\sum_{k=1}^{K-1} \mathbb{I}\{\mathcal{E}_h^k\} \cdot p_{z_h^k} \leq S_{\max}/3.$$

For $2 \leq k \leq K$, let X_h^k be a random variable defined as:

$$X_h^k = \begin{cases} \mathbb{I}\{\mathcal{E}_h^{k-1}\} & \hat{z}_h^{k-1} \text{ is added into } \widehat{\mathcal{Z}}_h^k \\ 0 & \text{otherwise} \end{cases}.$$

Then X_h^k is adapted to the filtration $\mathscr{F}_k$. We have that $\mathbb{E}_{k-1}[X_h^k] = p_{z_h^{k-1}} \cdot \mathbb{I}\{\mathcal{E}_h^{k-1}\}$ and $\mathbb{E}_{k-1}[(X_h^k - \mathbb{E}_{i-1}[X_h^k])^2] = \mathbb{I}\{\mathcal{E}_h^{k-1}\} \cdot p_{z_h^{k-1}}(1 - p_{z_h^{k-1}})$. Note that $X_h^k - \mathbb{E}_{k-1}[X_h^k]$ is a martingale difference sequence with respect to $\mathscr{F}_k$ and

$$\sum_{k=2}^{K} \mathbb{E}_{k-1}[(X_h^k - \mathbb{E}_{k-1}[X_h^k])^2] = \sum_{k=2}^{K} \mathbb{I}\{\mathcal{E}_h^{k-1}\} p_{z_h^{k-1}}(1 - p_{z_h^{k-1}}) \leq \sum_{k=2}^{K} \mathbb{I}\{\mathcal{E}_h^{k-1}\} \cdot p_{z_h^{k-1}} \leq S_{\max}/3,$$

$$\sum_{k=2}^{K} \mathbb{E}_{k-1}[X_h^k] = \sum_{k=2}^{K} p_{z_h^{k-1}} \mathbb{I}\{\mathcal{E}_h^{k-1}\} \leq S_{\max}/3.$$

By Freedman's inequality (Lemma 1), we have that

$$\begin{aligned}
&\Pr\left\{\sum_{k=2}^{K} X_h^k \geq S_{\max}\right\} \\
\leq &\Pr\left\{\left|\sum_{k=2}^{K} (X_h^k - \mathbb{E}_{k-1}[X_h^k])\right| \geq 2S_{\max}/3\right\} \\
\leq &2\exp\left\{-\frac{(2S_{\max}/3)^2/2}{S_{\max}/3 + 2S_{\max}/9}\right\} \\
\leq &\delta/(32T).
\end{aligned}$$

[4] We use $f \lesssim g$ to donote that $f \leq Cg$ for some absolute constant $C > 0$.

With a union bound we know that with probability at least $1-\delta/32$,

$$\sum_{k=2}^{K} X_h^k < S_{\max} \quad \forall h \in [H].$$

We condition on the above event and $\bigcap_{h=1}^{H}\bigcap_{k=1}^{K}\mathcal{E}_h^k$. In this case, it is clear that for all $h \in [H]$, we add elements into $\widehat{\mathcal{Z}}_h^k$ for at most $S_{\max}$ times. Combining the above result with Lemma 2 completes the proof. □

G. Proof of Theorem 1

G.1. Analysis of the Optimistic Planning

Our next lemma bounds the complexity of the bonus function. This step is essential for showing optimism, as we need to establish a uniform convergence argument to deal with the dependency in the data sequence.

Lemma 6. *With probability at least $1-\delta/8$, for all $(h,k) \in [H]\times[K]$, $b_h^k(\cdot,\cdot) \in \mathcal{M}$.*

Here $\mathcal{M}$ is a prespecified function class with bounded size:

$$\begin{aligned}
&\log|\mathcal{M}| \\
\leq& C' \cdot \log(T\mathcal{N}(\mathcal{F}, \sqrt{\delta/64T^3})/\delta) \cdot \dim_E(\mathcal{F}, 1/T) \cdot \log^2 T \cdot \log\left(\mathcal{C}(\mathcal{S}\times\mathcal{A}, 1/(16\sqrt{64T^3/\delta})) \cdot 64T^3/\delta\right) \\
\leq& C \cdot \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \cdot \dim_E(\mathcal{F}, 1/T) \cdot \log^2 T \cdot \log\left(\mathcal{C}(\mathcal{S}\times\mathcal{A}, \delta/T^2) \cdot T/\delta\right)
\end{aligned}$$

for some absolute constant $C', C > 0$ if T is sufficiently large.

Proof. Define $\mathcal{M}$ to be the set of all functions with the form:

$$\left\{|f_1(\cdot,\cdot) - f_2(\cdot,\cdot)| \mid \|f_1 - f_2\|_{\mathcal{Z}}^2 \leq \beta\right\}, \quad \mathcal{Z} \in \Omega$$

where Ω contains all set with bounded size:

$$\Omega := \{\mathcal{Z} \subseteq \mathcal{C}(\mathcal{S}\times\mathcal{A}, 1/(16\sqrt{64T^3/\delta})) \mid |\mathcal{Z}| \leq 64T^3/\delta, n_d(\mathcal{Z}) \leq S_{\max}\}$$

where $S_{\max}$ is defined in Proposition 2.

Conditioned on the event defined in Proposition 2, $\widehat{\mathcal{Z}}_h^k \in \Omega$, and $b_h^k(\cdot,\cdot) \in \mathcal{M}$ for all $(h,k) \in [H]\times[K]$. □

The next lemma estimates the error of the one-step bellman backup.

Lemma 7. *Consider a fixed pair $(k,h) \in [K]\times[H]$. For any $V : \mathcal{S} \to [0,H]$, define*

$$\mathcal{D}_h^k(V) := \{(s_h^\tau, a_h^\tau, r_h^\tau + V(s_{h+1}^\tau))\}_{\tau\in[k-1]}$$

and also

$$\widehat{f_V} := \operatorname{argmin}_{f\in\mathcal{F}} \|f\|_{\mathcal{D}_h^k(V)}^2.$$

For any $V : \mathcal{S} \to [0,H]$ and $\delta \in (0,1)$, there is an event $\mathcal{E}_{V,\delta}$ which holds with probability at least $1-\delta$, such that for any $V' : \mathcal{S} \to [0,H]$ with $\|V'-V\|_\infty \leq 1/T$, we have

$$\left\|\widehat{f_{V'}}(\cdot,\cdot) - r_h(\cdot,\cdot) - \sum_{s'\in\mathcal{S}} P_h(s'|\cdot,\cdot)V'(s')\right\|_{\mathcal{Z}_h^k} \leq C\cdot(H\sqrt{\log(1/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)}).$$

for some absolute constant $C > 0$.

Proof. The proof is almost identical to that of Lemma 5 in (Wang et al., 2020c). We provide a proof here for completeness.

In our proof, we consider a fixed $V : \mathcal{S} \to [0, H]$, and define

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

By Assumption 1, we have that $f_V(\cdot, \cdot) \in \mathcal{F}$.

For any $f \in \mathcal{F}$, we consider $\sum_{\tau=1}^{k-1} \xi_h^\tau(f)$ where

$$\xi_h^\tau(f) := 2(f(s_h^\tau, a_h^\tau) - f_V(s_h^\tau, a_h^\tau)) \cdot (f_V(s_h^\tau, a_h^\tau) - r_h^\tau - V(s_{h+1}^\tau)).$$

Then $\xi_h^\tau(f)$ is adapted to the filtration $\mathscr{F}_\tau$ and $\mathbb{E}_{\tau-1}[\xi_h^\tau(f)] = 0$. Moreover,

$$|\xi_h^\tau(f)| \le 2(H+1) \left| f(s_h^\tau, a_h^\tau) - f_V(s_h^\tau, a_h^\tau) \right|.$$

By Azuma-Hoeffding inequality, we have

$$\Pr\left[\left| \sum_{\tau=1}^{k-1} \xi_h^\tau(f) \right| \ge \epsilon \right] \le 2 \exp\left(- \frac{\epsilon^2}{8(H+1)^2 \|f - f_V\|_{\mathcal{Z}_h^k}^2} \right).$$

Let

$$\begin{aligned} \epsilon &= \left(8(H+1)^2 \log\left(\frac{2\mathcal{N}(\mathcal{F}, 1/T)}{\delta} \right) \cdot \|f - f_V\|_{\mathcal{Z}_h^k}^2 \right)^{1/2} \\ &\le 4(H+1) \|f - f_V\|_{\mathcal{Z}_h^k} \cdot \sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)}. \end{aligned}$$

We have, with probability at least $1 - \delta$, for all $f \in \mathcal{C}(\mathcal{F}, 1/T)$,

$$\left| \sum_{\tau=1}^{k-1} \xi_h^\tau(f) \right| \le 4(H+1) \|f - f_V\|_{\mathcal{Z}_h^k} \cdot \sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)}.$$

We define the above event to be $\mathcal{E}_{V,\delta}$, and we condition on this event for the rest of the proof.

For all $f \in \mathcal{F}$, there exists $g \in \mathcal{C}(\mathcal{F}, 1/T)$, such that $\|f - g\|_\infty \le 1/T$, and we have

$$\begin{aligned} \left| \sum_{\tau=1}^{k-1} \xi_h^\tau(f) \right| &\le \left| \sum_{\tau=1}^{k-1} \xi_h^\tau(g) \right| + 2(H+1) \\ &\le 4(H+1) \|g - f_V\|_{\mathcal{Z}_h^k} \cdot \sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} + 2(H+1) \\ &\le 4(H+1)(\|f - f_V\|_{\mathcal{Z}_h^k} + 1) \cdot \sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} + 2(H+1). \end{aligned}$$

Consider $V' : \mathcal{S} \to [0, H]$ with $\|V' - V\|_\infty \le 1/T$. We have

$$\|f_{V'} - f_V\|_\infty \le \|V' - V\|_\infty \le 1/T.$$

For any $f \in \mathcal{F}$,

$$\|f\|_{\mathcal{D}_h^k(V')}^2 - \|f_{V'}\|_{\mathcal{D}_h^k(V')}^2 = \|f - f_{V'}\|_{\mathcal{Z}_h^k}^2 + 2 \sum_{\tau=1}^{k-1} (f(s_h^\tau, a_h^\tau) - f_{V'}(s_h^\tau, a_h^\tau)) \cdot (f_{V'}(s_h^\tau, a_h^\tau) - r_h^\tau - V'(s_{h+1}^\tau)).$$

For the second term, we have,

$$
\begin{aligned}
&2\sum_{\tau=1}^{k-1}(f(s_h^\tau,a_h^\tau)-f_{V'}(s_h^\tau,a_h^\tau))\cdot(f_{V'}(s_h^\tau,a_h^\tau)-r_h^\tau-V'(s_{h+1}^\tau))\\
\geq&2\sum_{\tau=1}^{k-1}(f(s_h^\tau,a_h^\tau)-f_V(s_h^\tau,a_h^\tau))\cdot(f_V(s_h^\tau,a_h^\tau)-r_h^\tau-V(s_{h+1}^\tau))-4(H+1)\cdot\|V'-V\|_\infty\cdot k\\
=&\sum_{\tau=1}^{k-1}\xi_h^\tau(f)-4(H+1)\cdot\|V'-V\|_\infty\cdot k\\
\geq&-4(H+1)(\|f-f_V\|_{\mathcal{Z}_h^k}+1)\cdot\sqrt{\log(2/\delta)+\log\mathcal{N}(\mathcal{F},1/T)}-2(H+1)-4(H+1)\cdot\|V'-V\|_\infty\cdot k\\
\geq&-4(H+1)(\|f-f_{V'}\|_{\mathcal{Z}_h^k}+2)\cdot\sqrt{\log(2/\delta)+\log\mathcal{N}(\mathcal{F},1/T)}-6(H+1).
\end{aligned}
$$

Recall that $\widehat{f}_{V'}=\arg\min_{f\in\mathcal{F}}\|f\|^2_{\mathcal{D}_h^k(V')}$. We have $\|\widehat{f}_{V'}\|^2_{\mathcal{D}_h^k(V')}-\|f_{V'}\|^2_{\mathcal{D}_h^k(V')}\leq 0$, which implies,

$$
\begin{aligned}
0&\geq\|\widehat{f}_{V'}\|^2_{\mathcal{D}_h^k(V')}-\|f_{V'}\|^2_{\mathcal{D}_h^k(V')}\\
&=\|\widehat{f}_{V'}-f_{V'}\|^2_{\mathcal{Z}_h^k}+2\sum_{\tau=1}^{k-1}(\widehat{f}_{V'}(s_h^\tau,a_h^\tau)-f_{V'}(s_h^\tau,a_h^\tau))\cdot(f_{V'}(s_h^\tau,a_h^\tau)-r_h^\tau-V'(s_{h+1}^\tau))\\
&\geq\|\widehat{f}_{V'}-f_{V'}\|^2_{\mathcal{Z}_h^k}-4(H+1)(\|\widehat{f}_{V'}-f_{V'}\|_{\mathcal{Z}_h^k}+2)\cdot\sqrt{\log(2/\delta)+\log\mathcal{N}(\mathcal{F},1/T)}-6(H+1).
\end{aligned}
$$

Solving the above inequality, we have,

$$
\|\widehat{f}_{V'}-f_{V'}\|_{\mathcal{Z}_h^k}\leq C\cdot\big(H\cdot\sqrt{\log(1/\delta)+\log\mathcal{N}(\mathcal{F},1/T)}\big)
$$

for some absolute constant $C>0$. □

Our next lemma shows that f_h^k belongs to the desired confidence region.

Lemma 8. *Let $\mathcal{E}_2$ denote the event that for all $(k,h)\in[K]\times[H]$,*

$$
\big\|f_h^k-\bar{f}_h^k\big\|_{\mathcal{Z}_h^k}\leq\beta/100
$$

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

Then $\Pr[\mathcal{E}_2]\geq 1-\delta/4$ provided

$$
\beta\geq C\cdot H^2\cdot(\log(T/\delta)+\log\mathcal{N}(\mathcal{F},1/T)+\log|\mathcal{M}|)
$$

for some absolute constant $C>0$.

Proof. Note that for all $(k,h)\in[K]\times[H]$,

$$
\begin{aligned}
Q_h^k(\cdot,\cdot)&=\min\{f_h^k(\cdot,\cdot)+b_h^k(\cdot,\cdot),H\},\text{and}\\
V_h^k(\cdot)&=\max_{a\in\mathcal{A}}Q_h^k(\cdot,a).
\end{aligned}
$$

We define

$$
\begin{aligned}
\mathcal{Q}&:=\{\min\{f(\cdot,\cdot)+m(\cdot,\cdot),H\}|f\in\mathcal{C}(\mathcal{F},1/T),m\in\mathcal{M}\}\cup\{0\},\text{and}\\
\mathcal{V}&:=\{\max_{a\in\mathcal{A}}q(\cdot,a)|q\in\mathcal{Q}\}.
\end{aligned}
$$

Then $\log|\mathcal{V}|\leq\log|\mathcal{M}|+\log\mathcal{N}(\mathcal{F},1/T)+1$.

Conditioned on the event defined in Lemma 6, we have that $b_h^k(\cdot,\cdot)\in\mathcal{M}$ for all $(h,k)\in[H]\times[K]$. Thus for all $(k,h)\in[K]\times[H]$, $\mathcal{V}$ is a $(1/T)$-cover of $V_h^k(\cdot)$.

For each $V \in \mathcal{V}$, let $\mathcal{E}_{V,\delta/(8|\mathcal{V}|T)}$ be the event defined in Lemma 7. Note that $\mathcal{E}_{V,\delta/(8|\mathcal{V}|T)}$ also relates to a fixed pair (k,h). By applying Lemma 7 and a union bound, we have $\Pr\left[\bigcap_{(k,h)\in[K]\times[H]}\bigcap_{V\in\mathcal{V}}\mathcal{E}_{V,\delta/(8|\mathcal{V}|T)}\right] \geq 1-\delta/8$. We also condition on this event for the rest of the proof.

For $(k,h)\in[K]\times[H]$, recall that f_h^k is the solution to the regression problem in Algorithm 1, i.e., $f_h^k = \arg\min_{f\in\mathcal{F}}\|f\|^2_{\mathcal{D}_h^k}$. Let $V\in\mathcal{V}$ such that $\|V - V_{h+1}^k\|_\infty \leq 1/T$. By the definition of $\mathcal{E}_{V,\delta/(8|\mathcal{V}|T)}$ (the one relats to this (k,h) pair), we have that

$$\begin{aligned}
&\left\| f_h^k(\cdot,\cdot) - r_h(\cdot,\cdot) - \sum_{s'\in\mathcal{S}} P_h(s'|\cdot,\cdot)V_{h+1}^k(s') \right\|_{\mathcal{Z}_h^k} \\
\lesssim & H\sqrt{\log(8|\mathcal{V}|T/\delta) + \log\mathcal{N}(\mathcal{F},1/T)} \\
\lesssim & H\sqrt{\log(T/\delta) + \log\mathcal{N}(\mathcal{F},1/T) + \log|\mathcal{M}|}.
\end{aligned}$$

as desired. □

Finally, we use the above result to show optimism.

Lemma 9. *Let $\mathcal{E}_3$ denote the event 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 $\bar{f}_h^k(\cdot,\cdot) = \sum_{s\in\mathcal{S}'} P_h(s'|\cdot,\cdot)V_{h+1}^k(s') + r_h(\cdot,\cdot)$.
Then $\Pr[\mathcal{E}_3]\geq 1-\delta/2$.

Proof. We condition on the event defined in Proposition 1 and $\mathcal{E}_2$ defined in Lemma 8. Because $\left\|f_h^k - \bar{f}_h^k\right\|_{\mathcal{Z}_h^k} \leq \beta/100$, from the definition of $\underline{b}_h^k$ we have that $\left|\bar{f}_h^k(\cdot,\cdot) - f_h^k(\cdot,\cdot)\right| \leq \underline{b}_h^k(\cdot,\cdot)$. Moreover, by Proposition 1 we have $\underline{b}_h^k(\cdot,\cdot)\leq b_h^k(\cdot,\cdot)$. Thus for all $(k,h)\in[K]\times[H]$, $(s,a)\in\mathcal{S}\times\mathcal{A}$, we have

$$\begin{aligned}
&Q_h^k(s,a) \\
=&\min\{f_h^k(s,a) + b_h^k(s,a), H\} \\
\leq&\min\left\{\bar{f}_h^k(s,a) + b_h^k(s,a) + \left|\bar{f}_h^k(s,a) - f_h^k(s,a)\right|, H\right\} \\
\leq&\min\left\{\bar{f}_h^k(s,a) + b_h^k(s,a) + \underline{b}_h^k(s,a), H\right\} \\
\leq&\min\left\{\bar{f}_h^k(s,a) + 2b_h^k(s,a), H\right\} \\
\leq&\bar{f}_h^k(s,a) + 2b_h^k(s,a).
\end{aligned}$$

Next we use induction on h to prove $Q_h^*(\cdot,\cdot)\leq Q_h^k(\cdot,\cdot)$. The inequality clearly holds when $h = H+1$. Now we assume $Q_{h+1}^*(\cdot,\cdot)\leq Q_{h+1}^k(\cdot,\cdot)$ for some $h\in[H]$. Then obviously we have $V_{h+1}^*(\cdot)\leq V_{h+1}^k(\cdot)$. Therefore for all $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
&Q_h^*(s,a) \\
=&r_h(s,a) + \sum_{s'\in\mathcal{S}} P_h(s'|s,a)V_{h+1}^*(s') \\
\leq&\min\left\{r_h(s,a) + \sum_{s'\in\mathcal{S}} P_h(s'|s,a)V_{h+1}^k(s'), H\right\} \\
=&\min\left\{\bar{f}_h^k(s,a), H\right\} \\
\leq&\min\left\{f_h^k(s,a) + b_h^k(s,a), H\right\} \\
=&Q_h^k(s,a).
\end{aligned}$$

□

G.2. Regret decomposition

Now we are ready to bound the regret. For any $k \in [K]$, we let $\tilde{k}$ represents the episode index we update the policy to the one used in the k-th episode.

Lemma 10. *With probability at least* $1 - 5\delta/8$*, we have*

$$\textit{Regret}(K) \le 4H\sqrt{KH \cdot \log(16/\delta)} + 2\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k).$$

Proof. $\forall (k,h) \in [K] \times [H-1]$, define

$$\xi_h^k = \sum_{s' \in \mathcal{S}} P_h(s'|s_h^k, a_h^k)(V_{h+1}^{\tilde{k}}(s') - V_{h+1}^{\pi_{\tilde{k}}}(s')) - (V_{h+1}^{\tilde{k}}(s_{h+1}^k) - V_{h+1}^{\pi_{\tilde{k}}}(s_{h+1}^k)).$$

Note that $\{\xi_h^k\}$ is a martingale difference sequence and $|\xi_h^k| \le 2H$. By Azuma-Hoeffding inequality, with probability at least $1 - \delta/8$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H-1} \xi_h^k \le 4H\sqrt{KH \cdot \log(16/\delta)}.$$

Conditioned on the above event and $\mathcal{E}_3$ defined in Lemma 9, we have

$$\begin{aligned}
&\text{Regret}(K)\\
=&\sum_{k=1}^{K}\left(V_1^*(s_1^k) - V_1^{\pi_{\tilde{k}}}(s_1^k)\right)\\
\le&\sum_{k=1}^{K}\left(V_1^{\tilde{k}}(s_1^k) - V_1^{\pi_{\tilde{k}}}(s_1^k)\right)\\
=&\sum_{k=1}^{K}\left(Q_1^{\tilde{k}}(s_1^k, a_1^k) - Q_1^{\pi_{\tilde{k}}}(s_1^k, a_1^k)\right) \quad \text{(note that we played } \pi_{\tilde{k}} \text{ in epsiode } k)\\
=&\sum_{k=1}^{K}\left(\sum_{s' \in \mathcal{S}} P_1(s'|s_1^k, a_1^k)(V_2^{\tilde{k}}(s') - V_2^{\pi_{\tilde{k}}}(s')) + 2b_1^{\tilde{k}}(s_1^k, a_1^k)\right)\\
=&\sum_{k=1}^{K}\left((V_2^{\tilde{k}}(s_2^k) - V_2^{\pi_{\tilde{k}}}(s_2^k)) + \xi_1^k + 2b_1^{\tilde{k}}(s_1^k, a_1^k)\right)\\
\le&...\\
\le&\sum_{k=1}^{K}\sum_{h=1}^{H-1}\xi_h^k + 2\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^{\tilde{k}}(s_h^k, a_h^k)\\
=&\sum_{k=1}^{K}\sum_{h=1}^{H-1}\xi_h^k + 2\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) \quad \text{(note that } b_h^{\tilde{k}}(\cdot,\cdot) = b_h^k(\cdot,\cdot))\\
\le&4H\sqrt{KH \cdot \log(16/\delta)} + 2\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k)
\end{aligned}$$

as desired. □

The next lemma bounds the summation of the exploration bonus in terms of the eluder dimension.

Lemma 11. *With probability at least* $1 - \delta/32$*,*

$$\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) \le H + H(H+1)\dim_E(\mathcal{F}, 1/T) + C \cdot \sqrt{\dim_E(\mathcal{F}, 1/T) \cdot TH \cdot \beta}$$

for some absolute constant $C > 0$.

Proof. We condition on the event defined in Proposition 1 in the proof. Then we have

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

In the rest of the proof, we bound $\sum_{k=1}^K b_h^k(s_h^k, a_h^k)$ for each $h \in [H]$ separately.

For any given $\epsilon > 0$ and $h \in [H]$, let $\mathcal{L}_h = \{(s_h^k, a_h^k) | k \in [K], b_h^k(s_h^k, a_h^k) > \epsilon\}$ with $|\mathcal{L}_h| = L_h$. We will show that there exists $z_h^k := (s_h^k, a_h^k) \in \mathcal{L}_h$ such that (s_h^k, a_h^k) is ϵ-dependent on at least $L_h / \dim_E(\mathcal{F}, \epsilon) - 1$ disjoint subsequences in $\mathcal{Z}_h^k \cap \mathcal{L}_h$. Denote $N = L_h / \dim_E(\mathcal{F}, \epsilon) - 1$.

We decompose $\mathcal{L}_h$ into $N+1$ disjoint subsets, $\mathcal{L}_h = \cup_{j=1}^{N+1} \mathcal{L}_h^j$ by the following procedure. We initialize $\mathcal{L}_h^j = \{\}$ for all j and consider each $z_h^k \in \mathcal{L}_h$ sequentially. For each $z_h^k \in \mathcal{L}_h$, we find the smallest $1 \le j \le N$ such that z_h^k is ϵ-independent on $\mathcal{L}_h^j$ with respect to $\mathcal{F}$. We set $j = N+1$ if such j does not exist. We add z_h^k into $\mathcal{L}_h^j$ afterwards. When the decomposition of $\mathcal{L}_h$ is finished, $\mathcal{L}_h^{N+1}$ must be nonempty as $\mathcal{L}_h^j$ contains at most $\dim_E(\mathcal{F}, \epsilon)$ elements for $j \in [N]$. For any $z_h^k \in \mathcal{L}_h^{N+1}$, z_h^k is ϵ-dependent on at least $L_h / \dim_E(\mathcal{F}, \epsilon) - 1$ disjoint subsequences in $\mathcal{Z}_h^k \cap \mathcal{L}_h$.

On the other hand, there 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 $\|f_1 - f_2\|^2_{\mathcal{Z}_h^k} \le 100\beta$. By the definition of ϵ-dependent we have

$$(L_h / \dim_E(\mathcal{F}, \epsilon) - 1)\epsilon^2 \le \|f_1 - f_2\|^2_{\mathcal{Z}_h^k} \le 100\beta$$

which implies

$$L_h \le \left(\frac{100\beta}{\epsilon^2} + 1\right) \dim_E(\mathcal{F}, \epsilon).$$

Let $b_1 \ge b_2 \ge ... \ge b_K$ be a permutation of $\{b_h^k(s_h^k, a_h^k)\}_{k \in [K]}$. For any $b_k \ge 1/K$, we have

$$k \le \left(\frac{100\beta}{b_k^2} + 1\right) \dim_E(\mathcal{F}, b_k) \le \left(\frac{100\beta}{b_k^2} + 1\right) \dim_E(\mathcal{F}, 1/K)$$

which implies

$$b_k \le \left(\frac{t}{\dim_E(\mathcal{F}, 1/K)} - 1\right)^{-1/2} \cdot \sqrt{100\beta}.$$

Moreover, we have $b_k \le H + 1$. Therefore,

$$\begin{aligned} \sum_{k=1}^K b_k \le& 1 + (H+1)\dim_E(\mathcal{F}, 1/K) + \sum_{\dim_E(\mathcal{F}, 1/K) < k \le K} \left(\frac{k}{\dim_E(\mathcal{F}, 1/K)} - 1\right)^{-1/2} \cdot \sqrt{100\beta} \\ \le& 1 + (H+1)\dim_E(\mathcal{F}, 1/K) + C \cdot \sqrt{\dim_E(\mathcal{F}, 1/K) \cdot K \cdot \beta}. \end{aligned}$$

Summing up for all $h \in [H]$, we conclude that with probability at least $1 - \delta/8$,

$$\begin{aligned} \sum_{k=1}^K \sum_{h=1}^H b_h^k(s_h^k, a_h^k) &\le H + H(H+1)\dim_E(\mathcal{F}, 1/K) + CH \cdot \sqrt{\dim_E(\mathcal{F}, 1/K) \cdot K \cdot \beta} \\ &= H + H(H+1)\dim_E(\mathcal{F}, 1/K) + C \cdot \sqrt{\dim_E(\mathcal{F}, 1/K) \cdot TH \cdot \beta} \\ &= H + H(H+1)\dim_E(\mathcal{F}, 1/T) + C \cdot \sqrt{\dim_E(\mathcal{F}, 1/T) \cdot TH \cdot \beta} \end{aligned}$$

as desired. □

Proof of Theorem 1. By Lemma 10 and Lemma 11, with probability at least $1-3\delta/4$, we have

$$\begin{aligned}
&\text{Regret}(K)\\
\leq& 4H\sqrt{KH\cdot\log(16/\delta)}+2\sum_{k=1}^{K}\sum_{h=1}^{H}b_h^k(s_h^k,a_h^k)\\
=& 4H\sqrt{KH\cdot\log(16/\delta)}+2\left(H+H(H+1)\dim_E(\mathcal{F},1/T)+C\cdot\sqrt{\dim_E(\mathcal{F},1/T)\cdot TH\cdot\beta}\right)\\
\lesssim&\sqrt{\dim_E(\mathcal{F},1/T)\cdot T\cdot H^3\cdot\log(T\mathcal{N}(\mathcal{F},\delta/T^2)/\delta)\cdot\dim_E(\mathcal{F},1/T)\cdot\log^2 T\cdot\log\left(\mathcal{C}(\mathcal{S}\times\mathcal{A},\delta/(T^2))\cdot T/\delta\right)}\\
\lesssim&\sqrt{T\cdot H^3\cdot\log(T\mathcal{N}(\mathcal{F},\delta/T^2)/\delta)\cdot\dim_E^2(\mathcal{F},1/T)\cdot\log^2 T\cdot\log\left(\mathcal{C}(\mathcal{S}\times\mathcal{A},\delta/(T^2))\cdot T/\delta\right)}.
\end{aligned}$$

We also condition on the event defined in Proposition 2, in which case the global switching cost is bounded by $O(H\cdot\iota_2)$. At the same time, under the event defined in Proposition 2, the size of the sub-sampled dataset $\widehat{\mathcal{Z}}_h^k$ is at most $\widetilde{O}(\text{poly}(dH))$. Thus by the analysis in Section 3.4, the algorithm takes $\widetilde{O}(\text{poly}(dH)\cdot|\mathcal{A}|)$ time per round on average with an access to a regression oracle. □

H. Proof of Theorem 2

Firstly, note that the online sub-sampling procedure used in Algorithm 3 is exactly the same with Algorithm 1. Thus the properties of online sub-sampling, i.e., Proposition 1 and Proposition 2 still hold. As a special case, Proposition 1 and Proposition 2 also imply the corresponding results for b_h and $\widehat{\mathcal{Z}}_h$ used in Algorithm 4.

Lemma 12. *With probability at least $1-\delta/8$, for all $(h,k)\in[H]\times[K]$, $b_h^k(\cdot,\cdot)\in\mathcal{M}$.*

Here $\mathcal{M}$ is a prespecified function class with bounded size:

$$\begin{aligned}
&\log|\mathcal{M}|\\
\leq& C'\cdot\log(T\mathcal{N}(\mathcal{F},\sqrt{\delta/64T^3})/\delta)\cdot\dim_E(\mathcal{F},1/T)\cdot\log^2 T\cdot\log\left(\mathcal{C}(\mathcal{S}\times\mathcal{A},1/(16\sqrt{64T^3/\delta}))\cdot 64T^3/\delta\right)\\
\leq& C\cdot\log(T\mathcal{N}(\mathcal{F},\delta/T^2)/\delta)\cdot\dim_E(\mathcal{F},1/T)\cdot\log^2 T\cdot\log\left(\mathcal{C}(\mathcal{S}\times\mathcal{A},\delta/T^2)\cdot T/\delta\right).
\end{aligned}$$

for some absolute constant $C',C>0$ if T is sufficiently large.

As a special case, $b_h(\cdot,\cdot)\in\mathcal{M}$ as well.

Proof. The proof is identical to Lemma 6. □

The next lemma estimate the error of the one-step bellman backup.

Lemma 13. *Consider a fixed pair $(k,h)\in[K]\times[H]$. For any $V:\mathcal{S}\to[0,H]$, define*

$$\mathcal{D}_h^k(V):=\{(s_h^\tau,a_h^\tau,V(s_{h+1}^\tau))\}_{\tau\in[k-1]}$$

and also

$$\widehat{f_V}:=\text{argmin}_{f\in\mathcal{F}}\|f\|^2_{\mathcal{D}_h^k(V)}.$$

For any $V:\mathcal{S}\to[0,H]$ and $\delta\in(0,1)$, there is an event $\mathcal{E}_{V,\delta}$ which holds with probability at least $1-\delta$, such that for any $V':\mathcal{S}\to[0,H]$ with $\|V'-V\|_\infty\leq 2/T$, we have

$$\left\|\widehat{f_{V'}}(\cdot,\cdot)-\sum_{s'\in\mathcal{S}}P_h(s'|\cdot,\cdot)V'(s')\right\|_{\mathcal{Z}_h^k}\leq C\cdot(H\sqrt{\log(1/\delta)+\log\mathcal{N}(\mathcal{F},1/T)}).$$

for some absolute constant $C>0$

Proof. A key observation is that $\sum_{s'\in\mathcal{S}}P_h(s'|\cdot,\cdot)V'(s')\in\mathcal{F}$ due to Assumption 3. The rest of the proof is identical to Lemma 7. □

The next lemma verifies the confidence region.

Lemma 14. *Let $\mathcal{E}_2$ denote the event that for all $(k,h) \in [K] \times [H]$,*

$$\left\|f_h^k - \bar{f}_h^k\right\|_{\mathcal{Z}_h^k} \leq \beta/100$$

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

For the planning phase, for all $h \in [H]$, all reward function r in the function class $\mathcal{R}$

$$\left\|f_h - \bar{f}_h\right\|_{\mathcal{Z}_h} \leq \beta/100$$

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

Then we have $\Pr[\mathcal{E}_2] \geq 1 - \delta/4$ provided

$$\beta \geq C \cdot H^2 \cdot (\log(T/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T) + \log \mathcal{N}(\mathcal{R}, 1/T) + \log|\mathcal{M}|).$$

for some absolute constant $C > 0$.

Proof. We condition on the event defined in Lemma 12 in the whole proof.

Note that for all $(k,h) \in [K] \times [H]$,

$$Q_h^k(\cdot,\cdot) = \min\{f_h^k(\cdot,\cdot) + b_h^k(\cdot,\cdot) + \Pi_{[0,1]}[b_h^k(\cdot,\cdot)/H], H\}, \text{and}$$

$$V_h^k(\cdot) = \max_{a \in \mathcal{A}} Q_h^k(\cdot, a).$$

We define

$$\mathcal{Q} := \{\min\{f(\cdot,\cdot) + m(\cdot,\cdot) + \Pi_{[0,1]}[m(\cdot,\cdot)/H], H\} | f \in \mathcal{C}(\mathcal{F}, 1/T), m \in \mathcal{M}\} \cup \{0\}, \text{and}$$

$$\mathcal{V} := \{\max_{a \in \mathcal{A}} q(\cdot, a) | q \in \mathcal{Q}\}.$$

Then $\log|\mathcal{V}| \leq \log|\mathcal{M}| + \log \mathcal{N}(\mathcal{F}, 1/T) + 1$.

Because $b_h^k(\cdot,\cdot) \in \mathcal{M}$, $\mathcal{V}$ is a $(1/T)$-cover of $V_h^k(\cdot)$ for all $(k,h) \in [K] \times [H+1]$. For each $V \in \mathcal{V}$, let $\mathcal{E}_{V,\delta/(16|\mathcal{V}|T)}$ be the event defined in Lemma 13. Note that $\mathcal{E}_{V,\delta/(16|\mathcal{V}|T)}$ relates to a fixed pair (k,h). By Lemma 13 and a union bound, we have $\Pr\left[\bigcap_{(k,h) \in [K] \times [H]} \bigcap_{V \in \mathcal{V}} \mathcal{E}_{V,\delta/(16|\mathcal{V}|T)}\right] \geq 1 - \delta/16$. We also condition on this event.

For $(k,h) \in [K] \times [H]$, recall that f_h^k is the solution to the regression problem in Algorithm 3, i.e., $f_h^k = \arg\min_{f \in \mathcal{F}} \|f\|_{\mathcal{D}_h^k}^2$. Let $V \in \mathcal{V}$ such that $\|V - V_{h+1}^k\|_\infty \leq 1/T$. By the definition of $\mathcal{E}_{V,\delta/(16|\mathcal{V}|T)}$ (the one relates to this (k,h) pair), we have that

$$\begin{aligned}
&\left\|f_h^k(\cdot,\cdot) - \sum_{s' \in \mathcal{S}} P_h(s'|\cdot,\cdot) V_{h+1}^k(s')\right\|_{\mathcal{Z}_h^k} \\
\lesssim & H\sqrt{\log(16|\mathcal{V}|T/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} \\
\lesssim & H\sqrt{\log(T/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T) + \log|\mathcal{M}|}.
\end{aligned}$$

For the planning phase, we define a new function class:

$$\mathcal{F}^* := \{f(\cdot,\cdot) + r(\cdot,\cdot) | f \in \mathcal{F}, r \in \mathcal{R}\}.$$

Then from the definition of covering number we have that:

$$\mathcal{C}(\mathcal{F}^*, 2/T) \leq \mathcal{C}(\mathcal{F}, 1/T)\mathcal{C}(\mathcal{R}, 1/T).$$

Thus for all $h \in [H]$,

$$\begin{aligned} Q_h(\cdot,\cdot) &= \min\{f_h(\cdot,\cdot) + b_h(\cdot,\cdot) + r_h(\cdot,\cdot), H\} \\ &= \min\{f^*(\cdot,\cdot) + b_h(\cdot,\cdot), H\} \quad (f^* \in \mathcal{F}^*), \text{and} \\ V_h(\cdot) &= \max_{a\in\mathcal{A}} Q_h(\cdot,a). \end{aligned}$$

We define

$$\begin{aligned} \mathcal{Q}^* &:= \{\min\{f^*(\cdot,\cdot) + m(\cdot,\cdot), H\} | f \in \mathcal{C}(\mathcal{F}^*, 2/T), m \in \mathcal{M}\} \cup \{0\}, \text{and} \\ \mathcal{V}^* &:= \{\max_{a\in\mathcal{A}} q(\cdot,a) | q \in \mathcal{Q}^*\}. \end{aligned}$$

Then we have

$$\begin{aligned} \log|\mathcal{V}^*| &\le \log|\mathcal{M}| + \log\mathcal{N}(\mathcal{F}^*, 2/T) + 1 \\ &\le \log|\mathcal{M}| + \log\mathcal{N}(\mathcal{F}, 1/T) + \log\mathcal{N}(\mathcal{R}, 1/T) + 1. \end{aligned}$$

Because $b_h(\cdot,\cdot) \in \mathcal{M}$, $\mathcal{V}^*$ is a $(2/T)$-cover of $V_h(\cdot)$ for all $h \in [H+1]$. Similarly, for each $V \in \mathcal{V}^*$, let $\mathcal{E}_{V,\delta/(16|\mathcal{V}^*|T)}$ be the event defined in Lemma 13. Note that $\mathcal{E}_{V,\delta/(16|\mathcal{V}^*|T)}$ relates to a fixed pair $(k,h) \in [K]\times[H]$. We only consider those with $k = K$. By Lemma 13 and a union bound, we have $\Pr\left[\bigcap_{h\in[H],k=K}\bigcap_{V\in\mathcal{V}} \mathcal{E}_{V,\delta/(16|\mathcal{V}^*|T)}\right] \ge 1-\delta/16K \ge 1-\delta/16$. We also condition on this event.

For $h \in [H]$, recall that f_h is the solution to the regression problem in Algorithm 4, i.e., $f_h = \arg\min_{f\in\mathcal{F}} \|f\|^2_{\mathcal{D}_h}$. Let $V \in \mathcal{V}^*$ such that $\|V - V_{h+1}\|_\infty \le 1/T$. By the definition of $\mathcal{E}_{V,\delta/(16|\mathcal{V}^*|T)}$ (the one relates to this h and $k = K$, note that $\mathcal{Z}_h^K = \mathcal{Z}_h$), we have that

$$\begin{aligned} &\left\| f_h(\cdot,\cdot) - \sum_{s'\in\mathcal{S}} P_h(s'|\cdot,\cdot)V_{h+1}(s') \right\|_{\mathcal{Z}_h} \\ \lesssim& H\sqrt{\log(16|\mathcal{V}^*|T/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} \\ \lesssim& H\sqrt{\log(T/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T) + \log\mathcal{N}(\mathcal{R}, 1/T) + \log|\mathcal{M}|}. \end{aligned}$$

Combining the above two parts we complete the proof. □

Finally, we use the above result to show optimism, in both the exploration and planning phase.

Lemma 15. *Let $\mathcal{E}_3$ denote the event that for all $(k,h) \in [K]\times[H]$, and all $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$Q_h^*(s,a,r^k) \le Q_h^k(s,a) \le \bar{f}_h^k(s,a) + r_h^k(\cdot,\cdot) + 2b_h^k(s,a)$$

where $\bar{f}_h^k(\cdot,\cdot) = \sum_{s\in\mathcal{S}'} P_h(s'|\cdot,\cdot)V_{h+1}^k(s')$.
And for the planning phase, for all $h \in [H]$, all $(s,a) \in \mathcal{S}\times\mathcal{A}$, and all reward function r in the function class $\mathcal{R}$,

$$Q_h^*(s,a,r) \le Q_h(s,a) \le \bar{f}_h(s,a) + r_h(\cdot,\cdot) + 2b_h(s,a)$$

where $\bar{f}_h(\cdot,\cdot) = \sum_{s\in\mathcal{S}'} P_h(s'|\cdot,\cdot)V_{h+1}(s')$.
Then $\Pr[\mathcal{E}_3] \ge 1 - 3\delta/8$.

Proof. We condition on the event defined in Proposition 1 and $\mathcal{E}_2$ defined in Lemma 14. Because $\left\|f_h^k - \bar{f}_h^k\right\|_{\mathcal{Z}_h^k} \le \beta/100$, from the definition of $\underline{b}_h^k$ we have that $\left|\bar{f}_h^k(\cdot,\cdot) - f_h^k(\cdot,\cdot)\right| \le \underline{b}_h^k(\cdot,\cdot)$. Moreover, by Proposition 1 we have $\underline{b}_h^k(\cdot,\cdot) \le b_h^k(\cdot,\cdot)$. Thus for all $(k,h) \in [K]\times[H]$, $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$\begin{aligned} Q_h^k(s,a) &= \min\{f_h^k(s,a) + r_h^k(s,a) + b_h^k(s,a), H\} \\ &\le \min\left\{\bar{f}_h^k(s,a) + r_h^k(s,a) + b_h^k(s,a) + \left|\bar{f}_h^k(s,a) - f_h^k(s,a)\right|, H\right\} \\ &\le \min\left\{\bar{f}_h^k(s,a) + r_h^k(s,a) + b_h^k(s,a) + \underline{b}_h^k(s,a), H\right\} \\ &\le \min\left\{\bar{f}_h^k(s,a) + r_h^k(s,a) + 2b_h^k(s,a), H\right\} \\ &\le \bar{f}_h^k(s,a) + r_h^k(s,a) + 2b_h^k(s,a). \end{aligned}$$

Next we use induction on h to prove $Q_h^*(\cdot,\cdot,r^k) \le Q_h^k(\cdot,\cdot)$. The inequality clearly holds when $h = H+1$. Now we assume $Q_{h+1}^*(\cdot,\cdot,r^k) \le Q_{h+1}^k(\cdot,\cdot)$ for some $h \in [H]$. Then obviously we have $V_{h+1}^*(\cdot,r^k) \le V_{h+1}^k(\cdot)$. Therefore for all $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
Q_h^*(s,a,r^k) &= r_h^k(s,a) + \sum_{s'\in\mathcal{S}} P_h(s'|s,a)V_{h+1}^*(s',r^k) \\
&\le \min\left\{ r_h^k(s,a) + \sum_{s'\in\mathcal{S}} P_h(s'|s,a)V_{h+1}^k(s'), H \right\} \\
&= \min\left\{ r_h^k(s,a) + \bar{f}_h^k(s,a), H \right\} \\
&\le \min\left\{ r_h^k(s,a) + f_h^k(s,a) + b_h^k(s,a), H \right\} \\
&= Q_h^k(s,a).
\end{aligned}$$

The proof of the second inequality is identical. One only need to discard the superscript k in the above argument. □

Lemma 16. *With probability at least* $1-\delta/32$*,*

$$\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k,a_h^k) \le H + H(H+1)\dim_E(\mathcal{F},1/T) + C\cdot\sqrt{\dim_E(\mathcal{F},1/T)\cdot TH\cdot\beta}$$

for some absolute constant $C>0$*.*

Proof. The proof is identical to Lemma 11. □

The next lemma bounds the summation of the optimistic value functions in the exploration phase. The techniques are similar to the standard regret decomposition for optimistic algorithms.

Lemma 17. *With probability at least* $1-\delta/2$*,*

$$\sum_{k=1}^{K} V_1^{\tilde{k}}(s_1) = O(\sqrt{T\cdot H^3\cdot\iota_1})$$

where

$$\begin{aligned}
\iota_1 &= \log(\mathcal{N}(\mathcal{R},1/T))\cdot\dim_E(\mathcal{F},1/T) \\
&+ \log(T\mathcal{N}(\mathcal{F},\delta/T^2)/\delta)\cdot\dim_E^2(\mathcal{F},1/T)\cdot\log^2 T\cdot\log\left(\mathcal{N}(\mathcal{S}\times\mathcal{A},\delta/T^2)\cdot T/\delta\right).
\end{aligned}$$

Proof. For all $(k,h)\in[K]\times[H-1]$, denote

$$\xi_h^k = \sum_{s'\in\mathcal{S}} P_h(s'|s_h^k,a_h^k)V_{h+1}^{\tilde{k}}(s') - V_{h+1}^{\tilde{k}}(s_{h+1}^k).$$

If $\mathcal{E}_3$ defined in Lemma 15 happens, we have

$$\begin{aligned}
\sum_{k=1}^{K} V_1^{\tilde{k}}(s_1) &= \sum_{k=1}^{K} V_1^{\tilde{k}}(s_1^k) \\
&= \sum_{k=1}^{K} Q_1^{\tilde{k}}(s_1^k, a_1^k) \\
&\le \sum_{k=1}^{K} \left(r_1^{\tilde{k}}(s_1^k, a_1^k) + \sum_{s' \in \mathcal{S}} P_1(s'|s_1^k, a_1^k) V_2^{\tilde{k}}(s') + 2b_1^{\tilde{k}}(s_1^k, a_1^k) \right) \\
&\le \sum_{k=1}^{K} \left(\sum_{s' \in \mathcal{S}} P_1(s'|s_1^k, a_1^k) V_2^{\tilde{k}}(s') + (2 + 1/H) b_1^k(s_1^k, a_1^k) \right) \\
&= \sum_{k=1}^{K} \left(V_2^{\tilde{k}}(s_2^k) + \xi_1^k + (2 + 1/H) b_1^k(s_1^k, a_1^k) \right) \\
&\le \dots \\
&\le \sum_{k=1}^{K} \sum_{h=1}^{H-1} \xi_h^k + \sum_{k=1}^{K} \sum_{h=1}^{H} (2 + 1/H) b_h^k(s_h^k, a_h^k).
\end{aligned}$$

Note that $\{\xi_h^k\}_{(k,h) \in [K] \times [H]}$ (arranged in lexicographical order) is a martingale difference sequence with $|\xi_h^k| \le H$. By Azuma-Hoeffding inequality, we have

$$\Pr\left\{ \left| \sum_{k=1}^{K} \sum_{h=1}^{H-1} \xi_h^k \right| \le C' \cdot \sqrt{KH^3 \log(1/\delta)} \right\} \ge 1 - \delta/16.$$

for some absolute constant $C' > 0$. Conditioned on the above event, the event defined in Lemma 16, and $\mathcal{E}_3$ defined in Lemma 15 we conclude that with probability at least $1 - \delta/2$,

$$\begin{aligned}
&\sum_{k=1}^{K} V_1^{\tilde{k}}(s_1) \\
\le &\sum_{k=1}^{K} \sum_{h=1}^{H-1} \xi_h^k + \sum_{k=1}^{K} \sum_{h=1}^{H} (2 + 1/H) b_h^k(s_h^k, a_h^k) \\
= &C' \cdot \sqrt{KH^3 \log(1/\delta)} + (2 + 1/H) \cdot \left(H + H(H+1) \dim_E(\mathcal{F}, 1/T) + C \cdot \sqrt{\dim_E(\mathcal{F}, 1/T) \cdot TH \cdot \beta} \right) \\
\lesssim &\sqrt{T \cdot H^3 \cdot \iota_1}
\end{aligned}$$

as desired. □

The next lemma bounds the error of the planning policy in terms of the expectation of the bonus functions.

Lemma 18. *If $\mathcal{E}_3$ defined in Lemma 15 (optimism) happens, then for all reward function r in the function class $\mathcal{R}$,*

$$V_1^*(s_1, r) - V_1^{\pi}(s_1, r) \le 2H V_1^*(s_1, \Pi_{[0,1]}[b/H]).$$

Here b is the bonus function computed during the planning phase, and π is the output policy in the planning phase.

Proof. We generalize the lemma and use induction on h to prove that for any $h \in [H+1]$,

$$V_h(s) - V_h^{\pi}(s, r) - 2H V_h^{\pi}(s, \Pi_{[0,1]}[b/H]) \le 0 \quad \forall s \in \mathcal{S}.$$

The result is obvious for $h = H + 1$. Suppose for some $h \in [H]$, it holds that

$$V_{h+1}(s) - V_{h+1}^{\pi}(s, r) - 2H V_{h+1}^{\pi}(s, \Pi_{[0,1]}[b/H]) \le 0 \quad \forall s \in \mathcal{S}.$$

Then for all $s \in \mathcal{S}$, we have

$$\begin{aligned}
&V_h(s) - V_h^\pi(s,r) - 2HV_h^\pi(s, \Pi_{[0,1]}[b/H]) \\
=&Q_h(s,\pi(s)) - Q_h^\pi(s,\pi(s),r) - 2HQ_h^\pi(s,\pi(s),\Pi_{[0,1]}[b/H]) \\
\leq& \left(r_h(s,\pi(s)) + \sum_{s' \in \mathcal{S}} P_h(s'|s,\pi(s)) V_{h+1}(s') + 2\Pi_{[0,H]}[b_h(s,\pi(s))] \right) \\
&- \left(r_h(s,\pi(s)) + \sum_{s' \in \mathcal{S}} P_h(s'|s,\pi(s)) V_{h+1}^\pi(s',r) \right) \\
&-2H \left(\Pi_{[0,1]}[b_h(s,\pi(s))/H] + \sum_{s' \in \mathcal{S}} P_h(s'|s,\pi(s)) V_{h+1}(s', \Pi_{[0,1]}[b/H]) \right) \\
=& \sum_{s' \in \mathcal{S}} P_h(s'|s,\pi(s)) \left(V_{h+1}(s') - V_{h+1}^\pi(s',r) - 2HV_{h+1}^\pi(s', \Pi_{[0,1]}[b/H]) \right) \\
\leq& 0.
\end{aligned}$$

as desired.

By taking $h = 1$ and $s = s_1$ as a special case, we have that

$$V_1(s_1) - V_1^\pi(s_1,r) - 2HV_1^\pi(s_1, \Pi_{[0,1]}[b/H]) \leq 0.$$

Then we conclude

$$V_1^*(s_1,r) - V_1^\pi(s_1,r) \leq V_1(s_1,r) - V_1^\pi(s_1,r) \leq 2HV_1^\pi(s_1, \Pi_{[0,1]}[b/H]) \leq 2HV_1^*(s_1, \Pi_{[0,1]}[b/H]).$$

□

Lemma 19. *With probability at least* $1 - 7\delta/8$*, for all reward function* r *in the function class* $\mathcal{R}$*,*

$$V_1^*(s_1,r) - V_1^\pi(s_1,r) = O(H^3 \cdot \sqrt{\iota_1/K})$$

where

$$\begin{aligned}
\iota_1 = &\log(\mathcal{N}(\mathcal{R}, 1/T)) \cdot \dim_E(\mathcal{F}, 1/T) \\
&+ \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \cdot \dim_E^2(\mathcal{F}, 1/T) \cdot \log^2 T \cdot \log\left(\mathcal{N}(\mathcal{S} \times \mathcal{A}, \delta/T^2) \cdot T/\delta\right).
\end{aligned}$$

Proof. We condition on $\mathcal{E}_3$ defined in Lemma 15 and the event defined in Lemma 17. By Lemma 18, we have

$$V_1^*(s_1,r) - V_1^\pi(s_1,r) \leq 2HV_1^*(s_1, \Pi_{[0,1]}[b/H]).$$

By the monotonicity of the bonus function, we have that

$$\Pi_{[0,1]}[b_h(\cdot,\cdot)/H] \leq \Pi_{[0,1]}[b_h^k(\cdot,\cdot)/H] = r_h^k(\cdot,\cdot) \quad \forall (k,h) \in [K] \times [H].$$

Then the right hand side can be bounded in the following manner:

$$2HV_1^*(s_1, \Pi_{[0,1]}[b/H]) \leq 2\frac{H}{K} \sum_{k=1}^K V_1^*(s_1, r^{\tilde{k}}) \leq 2\frac{H}{K} \sum_{k=1}^K V_1^{\tilde{k}}(s_1).$$

Substituting the bound for $\sum_{k=1}^K V_1^{\tilde{k}}(s_1)$ completes the proof. □

Proof of Theorem 2. Simply combing Proposition 2 and Lemma 19 with a union bound completes the proof. □

I. Model Misspecification

In this section we study the case when there is a misspecification error in our model. We show that our algorithms are robust to the violation of the assumptions. Our result are similar to that in Wang et al. (2020c).The proofs are also essentially identical to that in Wang et al. (2020c).

In the standard RL setting, Assumption 1 with a misspecification error is stated as:

Assumption 5. *There exists a set of functions $\mathcal{F} \subseteq \{f : \mathcal{S} \times \mathcal{A} \to [0, H+1]\}$ and a real number $\zeta > 0$, such that for any $V : \mathcal{S} \to [0, H]$ and all $h \in [H]$, there exists $f_V \in \mathcal{F}$ which satisfies*

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| f_V(s,a) - r_h(s,a) - \sum_{s'\in\mathcal{S}} P_h(s' \mid s,a)V(s') \right| \le \zeta.$$

We call ζ the misspecification error.

In the reward-free RL setting, Assumption 3 with a misspecification error is stated as:

Assumption 6. *There exists a set of functions $\mathcal{F} \subseteq \{f : \mathcal{S} \times \mathcal{A} \to [0, H+1]\}$ and a real number $\zeta > 0$, such that for any $V : \mathcal{S} \to [0, H]$ and all $h \in [H]$, there exists $f_V \in \mathcal{F}$ which satisfies*

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| f_V(s,a) - \sum_{s'\in\mathcal{S}} P_h(s' \mid s,a)V(s') \right| \le \zeta.$$

We call ζ the misspecification error.

Our algorithms for the misspecification case are same with the original algorithms except for the change of the global parameter β.

In the standard RL setting (Algorithm 1), we set β to be:

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

In the reward-free RL setting (Algorithm 4 and Algorithm 3), we set β to be:

$$\begin{aligned}\beta = C \cdot (H^2 \cdot \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \cdot \dim_E(\mathcal{F}, 1/T) \cdot \log^2 T \cdot \log\left(\mathcal{N}(\mathcal{S}\times\mathcal{A}, \delta/T^2)\cdot T/\delta\right) + T\zeta) \\ + C \cdot H^2 \cdot \log(\mathcal{N}(\mathcal{R}, 1/T)) \cdot \dim_E(\mathcal{F}, 1/T)\end{aligned}$$

In Lemma 20 and Lemma 21, we bound the single-step optimization error in the misspecified case. The proofs are identical to that of Lemma 11 in (Wang et al., 2020c).

Lemma 20. *Suppose $\mathcal{F}$ satisfies Assumption 5. Consider a fixed pair $(k,h) \in [K]\times[H]$. For any $V : \mathcal{S} \to [0, H]$, define*

$$\mathcal{D}_h^k(V) := \{(s_h^\tau, a_h^\tau, r_h^\tau + V(s_{h+1}^\tau))\}_{\tau\in[k-1]}$$

and also

$$\widehat{f_V} := \operatorname{argmin}_{f\in\mathcal{F}} \|f\|^2_{\mathcal{D}_h^k(V)}.$$

For any $V : \mathcal{S} \to [0, H]$ and $\delta \in (0,1)$, there is an event $\mathcal{E}_{V,\delta}$ which holds with probability at least $1-\delta$, such that for any $V' : \mathcal{S} \to [0, H]$ with $\|V' - V\|_\infty \le 1/T$, we have

$$\left\| \widehat{f_{V'}}(\cdot,\cdot) - r_h(\cdot,\cdot) - \sum_{s'\in\mathcal{S}} P_h(s'|\cdot,\cdot)V'(s') \right\|_{\mathcal{Z}_h^k} \le C\cdot(\sqrt{H^2(\log(1/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)) + T\zeta}).$$

Lemma 21. *Suppose $\mathcal{F}$ satisfies Assumption 6. Consider a fixed pair $(k,h) \in [K]\times[H]$. For any $V : \mathcal{S} \to [0, H]$, define*

$$\mathcal{D}_h^k(V) := \{(s_h^\tau, a_h^\tau, V(s_{h+1}^\tau))\}_{\tau\in[k-1]}$$

and also

$$\widehat{f_V} := \operatorname{argmin}_{f\in\mathcal{F}} \|f\|^2_{\mathcal{D}_h^k(V)}.$$

For any $V : \mathcal{S} \to [0, H]$ and $\delta \in (0, 1)$, there is an event $\mathcal{E}_{V,\delta}$ which holds with probability at least $1 - \delta$, such that for any $V' : \mathcal{S} \to [0, H]$ with $\|V' - V\|_\infty \leq 2/T$, we have

$$\left\| \widehat{f_{V'}}(\cdot, \cdot) - \sum_{s' \in \mathcal{S}} P_h(s'|\cdot, \cdot)V'(s') \right\|_{\mathcal{Z}_h^k} \leq C \cdot (\sqrt{H^2(\log(1/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)) + T\zeta}).$$

Then similar to Lemma 8 and Lemma 14 we can verifies that the new value of β can derive the desired confidence region.

With the new value of β, the new regret bound in the regular RL setting can be easily derived. In the reward-free setting the bound for $V_1^*(s_1, r) - V_1^\pi(s_1, r)$ will have an additional $O(\sqrt{H^5 \cdot \dim_E(\mathcal{F}, 1/T) \cdot \zeta})$ term. Therefore there will be an irreducible error in the error bound.

Formally, our results in the misspecification case is stated in Theorem 4 and Theorem 5.

Theorem 4. *Assume Assumption 5 holds, and T is sufficiently large.With probability at least $1 - \delta$, the algorithm achieves a regret bound,*

$$\textit{Regret}(K) = O(\sqrt{\iota_1 \cdot H^3 \cdot T} + \sqrt{\dim_E(\mathcal{F}, 1/T) \cdot H \cdot \zeta} \cdot T)$$

where

$$\iota_1 = \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \cdot \dim_E^2(\mathcal{F}, 1/T) \cdot \log^2 T \cdot \log\left(\mathcal{N}(\mathcal{S} \times \mathcal{A}, \delta/T^2) \cdot T/\delta\right)$$

and the global switching cost is upper bounded by

$$N_{\textit{switch}}^{gl} = O(\iota_2 \cdot H)$$

where

$$\iota_2 = \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \cdot \dim_E(\mathcal{F}, 1/T) \cdot \log^2 T$$

Furthermore, with probability at least $1-\delta$ the algorithm takes $\widetilde{O}(\text{poly}(dH) \cdot |\mathcal{A}|)$ time per round with access to a regression oracle.

Theorem 5. *Suppose Assumption 6 holds and T is sufficiently large. For any given $\delta \in (0, 1)$, after collecting K trajectories during the exploration phase (by Algorithm 3), with probability at least $1 - \delta$, for any reward function $r = \{r_h\}_{h=1}^H$ satisfying Assumption 4, Algorithm 4 outputs an $O(H^3 \cdot \sqrt{\iota_1/K} + \epsilon_i)$-optimial policy for the MDP $(\mathcal{S}, \mathcal{A}, P, r, H, s_1)$. Here,*

$$\begin{aligned}\iota_1 &= \log(\mathcal{N}(\mathcal{R}, 1/T)) \cdot \dim_E(\mathcal{F}, 1/T) \\ &+ \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \cdot \log^2 T \cdot \dim_E^2(\mathcal{F}, 1/T) \cdot \log\left(\mathcal{N}(\mathcal{S} \times \mathcal{A}, \delta/T^2) \cdot T/\delta\right)\end{aligned}$$

and ϵ_i is the irreducible error:

$$\epsilon_i = \sqrt{H^5 \cdot \dim_E(\mathcal{F}, 1/T) \cdot \zeta}.$$

Moreover, the global switching cost of Algorithm 3 is upper bounded by

$$N_{\textit{switch}}^{gl} = O(H \cdot \iota_2)$$

where

$$\iota_2 = \log(T\mathcal{N}(\mathcal{F}, \delta/T^2)/\delta) \cdot \dim_E(\mathcal{F}, 1/T) \cdot \log^2 T.$$

Furthermore, with probability at least $1-\delta$, Algorithm 3 takes $\tilde{O}(\text{poly}(dH) \cdot |\mathcal{A}|)$ time per round with access to a regression oracle.