On the Model-Misspecification in Reinforcement Learning

Yunfan Li
University of California, Los Angeles

Lin Yang
University of California, Los Angeles

Abstract

The success of reinforcement learning (RL) crucially depends on effective function approximation when dealing with complex ground-truth models. Existing sample-efficient RL algorithms primarily employ three approaches to function approximation: policy-based, value-based, and model-based methods. However, in the face of model misspecification—a disparity between the ground-truth and optimal function approximators— it is shown that policy-based approaches can be robust even when the policy function approximation is under a large *locally-bounded* misspecification error, with which the function class may exhibit a $\Omega(1)$ approximation error in specific states and actions, but remains small on average within a policy-induced state distribution. Yet it remains an open question whether similar robustness can be achieved with value-based and model-based approaches, especially with general function approximation.

To bridge this gap, in this paper we present a unified theoretical framework for addressing model misspecification in RL. We demonstrate that, through meticulous algorithm design and sophisticated analysis, value-based and model-based methods employing general function approximation can achieve robustness under local misspecification error bounds. In particular, they can attain a regret bound of $\widetilde{O}\left(\text{poly}(dH)\cdot(\sqrt{K}+K\cdot\zeta)\right)$, where d represents the complexity of the function class, H is the episode length, K is the total number of episodes, and ζ denotes the local bound for misspecification error. Furthermore, we propose an algorithmic framework that can achieve the same order of regret bound without prior knowledge of ζ, thereby enhancing its practical applicability.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 INTRODUCTION

Reinforcement Learning (RL) is a paradigm where agents learn to interact with an environment through state and reward feedback. In recent years, RL has seen significant success across various applications, such as control (Mnih et al., 2015; Gu et al., 2017), board games (Silver et al., 2016), video games (Mnih et al., 2013), and even the training of large language models like ChatGPT (Ouyang et al., 2022). In these applications, RL systems use deep neural networks (DNN) to approximate the policy, value, or models, thereby addressing the notorious "curse-of-dimensionality" issues associated with RL systems that have large state-action spaces (Bellman, 2010).

Despite these successes, the theoretical understanding of how RL operates in practice, particularly when deep neural networks are involved, remains incomplete. A key question that arises is how approximation error, or misspecification, in function approximators can affect the performance of a deep RL system. Although deep networks are known to be universal approximators, their performance can be influenced by a range of factors, such as the training algorithm, dropout, normalization, and other engineering techniques. Therefore, the robustness of RL systems under misspecification is an important concern, particularly in risk-sensitive domains.

Theoretical advancements have begun to address the misspecification issue (Du et al., 2019a; Jin et al., 2020; Agarwal et al., 2020a; Zanette et al., 2021). For example, Du et al. (2019a) demonstrated that a small misspecification error in a value function approximator can lead to exponential increases in learning complexity, even when the function approximator for the optimal Q-value is a linear function class. In contrast, several studies have presented positive results for a relaxed model-class, where the transition probability matrix or the Bellman operator are close to a function class, making the learning system more robust to misspecification

errors (Jin et al., 2020; Wang et al., 2020b).

However, these studies address misspecification errors of distinct natural. For instance, studies like (Jin et al., 2020; Wang et al., 2020b; Ayoub et al., 2020) require the misspecification error to be *globally bounded*, i.e., the function approximators should approximate the transition probability with a small error for all state-action pairs. Conversely, studies like Agarwal et al. (2020a); Zanette et al. (2021) only require the misspecification error to be *locally bounded*, where the errors need only be bounded at *relevant* state-action pairs, which can be reached by a policy with a high enough probability. Notably, under a locally bounded misspecification error, the approximation error for certain state-action pairs (and therefore the global misspecification bound) can be arbitrarily large, and the analysis in (Jin et al., 2020; Wang et al., 2020b; Ayoub et al., 2020) would fail to provide a meaningful learning guarantee.

Existing sample-efficient Reinforcement Learning (RL) algorithms can be categorized into three main types based on the targets they approximate: policy-based, value-based and model-based. Policy-based approaches distinguish themselves from the other two by utilizing Monte-Carlo (MC) sampling to estimate policy values during training. This approach is considered more robust compared to value bootstrapping in value-based methods. However, the reliance on MC sampling makes it challenging to reuse samples since each new policy necessitates fresh runs and data collection. Consequently, the state-of-the-art policy-based approach (Zanette et al., 2021) exhibits a statistical error bound scaling as $\propto 1/\epsilon^{-3}$ for an error level of ϵ. In contrast, value-based and model-based approaches offer sample bounds of $\propto 1/\epsilon^{-2}$ (Yang and Wang, 2019, 2020; Jin et al., 2020; Ayoub et al., 2020). This disparity underscores the necessity for designing RL algorithms that are both statistically efficient and robust against misspecifications. Recent developments (Vial et al., 2022; Agarwal et al., 2023) have enhanced the robustness and practicality of classical value-based method (Jin et al., 2020). However, these algorithms depend on well-designed linear feature extractors, significantly limiting their applicability. In practice, algorithms often specify a function class (e.g., deep neural networks with a specific architecture) rather than a linear feature mapping. To date, a fundamental question concerning RL with general function approximation remains largely unanswered: *"Is the policy-based approach inherently more robust than the value-based and model-based approaches when dealing with misspecifications?"* Or, equivalently, *"Is it possible to design an RL approach with general function approximation that is both statistically efficient and robust to misspecifications?"*

In this paper, we delve into these fundamental questions and offer a comprehensive response. We present a unified and robust algorithm framework, **LBM-UCB** (***L**ocally **B**ounded **M**isspecification-**U**pper **C**onfidence **B**ound*), catering to value-based and model-based methods with general function approximation, specifically tailored to handle model misspecifications, particularly in the context of locally-bounded misspecifications. Unlike the realizable setting, where the ground-truth model is assumed to be within the function class, our robust algorithm framework meticulously designs a high-probability confidence set to encompass the best approximator within the function class. We demonstrate that under our framework, classical value-based algorithm (Wang et al., 2020b) and model-based algorithm (Ayoub et al., 2020) can achieve a level of robustness similar to policy-based methods in the presence of locally bounded misspecifications. Importantly, they maintain a statistical rate scaling as $\propto 1/\epsilon^{-2}$. To be specific, for episodic RL with a total of K episodes, a horizon length of H, a function class complexity of d, and a locally bounded misspecification error bound of ζ, our regret bound is $\widetilde{O}\left(\text{poly}(dH)\cdot(\sqrt{K}+K\cdot\zeta)\right)$. This bound is almost optimal in terms of ζ and K, and provides a regret bound of $O(\zeta K)$ even when the misspecification error for certain states is on the order of $O(1)$. Furthermore, we devise a meta-algorithm within our framework that does not require prior knowledge of the misspecification parameter ζ. This enhancement increases its potential practical applicability, making it more accessible and versatile in real-world applications.

A core novelty of our analysis is that instead of making rough assumptions about all state-action pairs having a uniform upper bound with respect to misspecification, we carefully study which state-action pairs genuinely impact the algorithm's robustness. Interestingly, we find that, within the same episode, those state-action pairs that truly influence the algorithm's performance are drawn from the same policy-induced distribution. This allows us to obtain better bounds in average sense under the distribution of policies. Furthermore, since there is no global upper bound on misspecification errors, the global optimism (or near-optimism) property described in (Jin et al., 2020; Wang et al., 2020b; Ayoub et al., 2020) no longer holds. To address this, we introduce a new method where we attach a virtual random process to utilize the optimal policy for data collection. This approach enables us to achieve near-optimism in the average sense, considering the distribution induced by the optimal policy.

2 RELATED WORK

In this section, we present the recent works that are relevant to our paper.

Misspecified Bandit In the context of misspecified linear bandit problems, where the reward function can be approximated by a linear function with some worst-case error, a number of papers, including those (Ghosh et al., 2017; Foster and Rakhlin, 2020; Lattimore et al., 2020; Takemura et al., 2021; Zhang et al., 2023b), have explored the notion of a uniform upper bound on misspecification errors across all actions. Additionally, Foster et al. (2020a) have introduced a more lenient average-case concept of misspecification, which assesses the error associated with the specific sequence being considered. It is noteworthy that several other papers, such as those by (Lykouris et al., 2018; Gupta et al., 2019; Zhao et al., 2021; Wei et al., 2022; Ding et al., 2022; Ye et al., 2023), delve into the analysis of cumulative misspecification errors, often referred to as "corruption," within the context of bandit problems.

RL with Function Approximations. Sample-efficient reinforcement learning (RL) algorithms employing function approximations can be classified into three primary categories, each based on the specific target they aim to approximate: value-based, model-based, and policy-based.

With regards to recent value-based methods, there are rich literature designing and analyzing algorithms for RL with linear function approximation (Du et al., 2019b; Wang et al., 2019; Zanette et al., 2019; Yang and Wang, 2020; Modi et al., 2020; Wang et al., 2020a; Agarwal et al., 2020b; He et al., 2022; Zhang et al., 2023a). However, these papers heavily rely on the assumption that the value function or the model can be approximated by a linear function or a generalized linear function of the feature vectors and do not discuss when model misspecification happens. On the other hand, these papers (Yang and Wang, 2019; Jin et al., 2020; Jia et al., 2019; Vial et al., 2022) consider the model misspecification using the globally-bounded misspecification error, making their algorithms robust to a small range of misspecified linear models. Furthermore, Vial et al. (2022) has introduced an algorithm with the notable attribute of being parameter-free in relation to the global bound of the misspecification parameter. In a similar vein, Agarwal et al. (2023) has demonstrated that the classical algorithm LSVI-UCB (Jin et al., 2020) remains effective even in the presence of locally-bounded misspecification error. For recent general function approximations, complexity measures are essential for non-linear function class, and Russo and Van Roy (2013) proposed the concept of eluder dimension. Recent papers have extended it to more general framework (Jiang et al., 2017; Du et al., 2021; Jin et al., 2021; Foster et al., 2020b; Chen et al., 2022; Zhong et al., 2022; Liu et al., 2023). However, the use of eluder dimension allows computational tractable optimization methods. Based on the eluder dimension, Wang et al. (2020b) describes a UCB-VI style algorithm that can explore the environment driven by a well-designed width function and Kong et al. (2021) devises an online sub-sampling method which largely reduces the average computation time of Wang et al. (2020b). However, when considering the misspecified case, their work can only tolerate the approximation error between the assumed general function class and the truth model to have a uniform upper bound for all state-action pairs. In this paper, we analyze the regret bound of value-based methods under locally bounded misspecified MDP.

In the realm of model-based methods, several notable papers (Jia et al., 2020; Ayoub et al., 2020; Modi et al., 2020) have provided valuable statistical guarantees, primarily focusing on linear function approximation. These works concentrate on scenarios where the underlying transition probability kernel of the MDP is represented as a linear mixture model. Notably, (Zhou et al., 2021) has introduced an algorithm that achieves nearly minimax optimality for linear mixture MDPs, incorporating Bernstein-type concentration techniques. Furthermore, (Zhou and Gu, 2022) has designed computationally efficient horizon-free RL algorithms within the same linear mixture MDP framework. In the context of reward-free settings, (Zhang et al., 2021) has presented an algorithm based on the linear mixture MDP assumption. On a broader scale, (Ayoub et al., 2020) has delved into general function approximation for model-based methods, using the concept of value-targeted regression. However, among the aforementioned papers, only (Jia et al., 2020) and (Ayoub et al., 2020) have ventured into the realm of model misspecification, although their assumptions remain confined to scenarios featuring globally-bounded misspecification error. In this paper, we analyze the regret bound of model-based methods under locally bounded misspecified MDP.

For recent policy-based methods with function approximation, a series of papers provide statistical guarantees (Cai et al., 2020; Duan et al., 2020; Agarwal et al., 2021; Feng et al., 2021; Zanette et al., 2021). Among them, Agarwal et al. (2020a, 2021); Feng et al. (2021); Zanette et al. (2021) consider model misspecification using locally-bounded misspecification error but suffer from poor sample complexity due to policy evaluations. Specifically, Agarwal et al. (2020a) uses a notion called transfer error to measure the model misspecification in the linear setting, where they assume a good approximator under some policy cover has a bounded error in average sense when transferred to an arbitrary policy-induced distribution. Moreover, Feng et al. (2021) proposes a model-free algorithm applying

the indicator of width function (Wang et al., 2020b) under the bounded transfer error assumption which allows the use of general function approximation. To improve the poor sample complexity of Agarwal et al. (2020a), Zanette et al. (2021) uses the doubling trick for determinant of empirical cumulative covariance and importance sampling technique.

3 PRELIMINARIES

In this paper, we focus on the episodic RL with setting modeled by a finite-horizon Markov Decision Process. Below we present a brief introduction of problem settings.

3.1 Episodic RL with Finite-Horizon Markov Decision Process

We consider a finite-horizon Markov Decision Process (MDP) $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, \mu)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space which has a finite size[1], $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition operator, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the deterministic reward function, H is the planning horizon, i.e. episode length, and μ is the initial distribution.

An agent interacts with the environment episodically as follows. For each H-length episode, the agent adopts a policy π. To be specific, a policy $\pi = \{\pi_h\}_{h=1}^H$, where for each $h \in [H]$, $\pi_h : \mathcal{S} \to \mathcal{A}$ chooses an action a from the action space based on the current state s. The policy π induces a trajectory $s_1, a_1, r_1, s_2, a_2, r_2, \cdots s_H, a_H, r_H$, where $s_1 \sim \mu$, $a_1 = \pi_1(s_1)$, $r_1 = r(s_1, a_1)$, $s_2 \sim P(\cdot|s_1, a_1)$, $a_2 = \pi_2(s_2)$, etc.

We use V-function and Q-function to evaluate the long-term expected cumulative reward under the policy π with respect to the current state (state-action) pair. They are defined as:

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

and

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

For MDP, there always exists an optimal deterministic policy π^*, such that $V_h^{\pi^*}(s) = \sup_\pi V_h^\pi(s)$ for all $s \in \mathcal{S}$ and all $h \in [H]$ (Puterman, 2014). To simplify our notation, we denote the optimal Q-function and V-function as $Q_h^*(s, a) = Q_h^{\pi^*}(s, a)$ and $V_h^*(s) = V_h^{\pi^*}(s)$. We also denote $[\mathbb{P}V_{h+1}](s, a) := \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} V_{h+1}(s')$, and the Bellman equation can be written as :

$$Q_h^\pi(s, a) = r(s, a) + \mathbb{P}V_{h+1}^\pi(s, a)$$

and

$$V_h^\pi(s) = Q_h^\pi(s, \pi_h(s))$$

Besides values, we also consider the state-action distribution generated by a policy. Without loss of generality, we assume that the agent always starts from a fixed point s_1 for each episode k. Concretely, for each time step $h \in [H]$, we define the state-action distribution induced by a policy π as

$$d_h^\pi(s, a) = \mathbb{P}^\pi(s_h = s, a_h = a|s_1)$$

where $\mathbb{P}^\pi(s_h = s, a_h = a|s_1)$ is the probability of reaching (s, a) at the h-th step starting from s_1 under policy π. We also define the average distribution $d^\pi = \frac{1}{H}\sum_{h=1}^{H} d_h^\pi$.

The goal of the agent is to improve its performance with the environment. One way to measure the effectiveness of a learning algorithm is using the notion of regret. For $k \in [K]$, suppose the agent starts from state s_1^k and chooses the policy π^k to collect a trajectory. Then the regret is defined as

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

3.2 Function Approximation

In addressing optimization challenges within MDPs featuring large state-action spaces, we introduce function approximation methods. These methods can be categorized into two key groups: value-based (approximating the optimal value function) and model-based (approximating the MDP's transition kernel) function approximation. We will now provide detailed explanations of both approaches.

Value-based Function Approximation For the value-based setting, the function class $\mathcal{F}$ contains the approximators of optimal state-action value function of the MDP, which means $\mathcal{F} \subset \{f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$.

- We denote corresponding state-action value function $Q_f = f$.
- We denote corresponding value function $V_f(\cdot) = \max_{a \in \mathcal{A}} Q_f(\cdot, a)$. Moreover, we denote the corresponding optimal policy π_f with $\pi_f(\cdot) = \text{argmax}_{a \in \mathcal{A}} Q_f(\cdot, a)$.

[1] Our approach can be extended to infinite-sized or continuous action space with an efficient optimization oracle for computing the arg max operation.

- We denote f^* as the optimal state-action value function based on the ground-truth model, and it is possible that f^* does not belong to the set $\mathcal{F}$.

Model-based Function Approximation For the model-based setting, the function class $\mathcal{F}$ contains the approximators of transition kernels, for which we denote $f = \mathbb{P}_f \in \mathcal{F}$.

- We denote V_f^π as the value function induced by model $\mathbb{P}_f$ and policy π.
- We denote V_f as the optimal value function under model $\mathbb{P}_f$, i.e., $V_f = \sup_{\pi\in\Pi} V_f^\pi$. Moreover, we denote π_f as the corresponding optimal policy, i.e. $\pi_f = \text{argmax}_{\pi\in\Pi} V_f^\pi$.
- We denote the ground-truth model as f^*, and f^* may not belong to the function class $\mathcal{F}$.

Notation We use $[n]$ to represent index set $\{1, \cdots n\}$. For $x \in \mathbb{R}$, $\lfloor x \rfloor$ represents the largest integer not exceeding x and $\lceil x \rceil$ represents the smallest integer exceeding x. Given $a, b \in \mathbb{R}^d$, we denote by $a^\top b$ the inner product between a and b and $||a||_2$ the Euclidean norm of a. Given a matrix A, we use $||A||_2$ for the spectral norm of A, and for a positive definite matrix Σ and a vector x, we define $||x||_\Sigma = \sqrt{x^\top \Sigma x}$. We use O to represent leading orders in asymptotic upper bounds and $\widetilde{O}$ to hide the polylog factors. For a finite set $\mathcal{A}$, we denote the cardinality of $\mathcal{A}$ by $|\mathcal{A}|$, all distributions over $\mathcal{A}$ by $\Delta(\mathcal{A})$, and especially the uniform distribution over $\mathcal{A}$ by $\text{Unif}(\mathcal{A})$. For a function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define $\|f\|_\infty = \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |f(s, a)|$. Similarly, for a function $v : \mathcal{S} \to \mathbb{R}$, we define $\|v\|_\infty = \max_{s\in\mathcal{S}} |v(s)|$. For a set of state-action pairs $\mathcal{Z} \subseteq \mathcal{S} \times \mathcal{A}$, for a function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define the $\mathcal{Z}$-norm of f as $\|f\|_\mathcal{Z} = \left(\sum_{(s,a)\in\mathcal{Z}} (f(s, a))^2\right)^{1/2}$. Given a dataset $\mathcal{D} = \{(s_i, a_i, q_i)\}_{i=1}^{|\mathcal{D}|} \subset \mathcal{S} \times \mathcal{A} \times \mathbb{R}$, for a function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, define $||f||_\mathcal{D} = \left(\sum_{t=1}^{|\mathcal{D}|} (f(s_t, a_t) - q_t)^2\right)^{1/2}$.

4 ROBUST RL ALGORITHMS WITH GENERAL FUNCTION APPROXIMATION

4.1 Generic Framework: LBM-UCB

In this section, we provide a generic robust RL framework, **LBM-UCB** (***L**ocally **B**ounded **M**isspecification-**U**pper **C**onfidence **B**ound*), with general function approximation when locally-bounded misspecifications appear.

At the outset of each episode, denoted as $k = 1, 2, \cdots, K$, our algorithm identifies the optimal empirical approximator, denoted as f^k, from the hypothesis class $\mathcal{F}$. This selection is made by minimizing the loss function L_{k-1} with the current dataset $\mathcal{D}^{k-1}$. The form of the loss function L_{k-1} varies depending on the type of function approximation, typically employing the two-norm distance to measure the fitting error of the function within the hypothesis class.

In the conventional approach, a high-probability confidence set is constructed to encompass the ground-truth. However, in our misspecified setting, we cannot assume realizability, meaning that f^* may not belong to the hypothesis class $\mathcal{F}$. Consequently, we construct a confidence set designed to encompass the best ground-truth approximator with high probability. This confidence set, denoted as $\mathcal{B}^k$, is centered around the best empirical approximator f^k. Its radius consists of two components: $\mathcal{E}^k_{\text{stat}}$ and $\mathcal{E}^k_{\text{bias}}$. Here, $\mathcal{E}^k_{\text{stat}}$ accounts for the statistical error arising from dataset randomness, while $\mathcal{E}^k_{\text{bias}}$ represents the error stemming from the mismatch between the ground-truth and the best ground-truth approximator.

Subsequently, the algorithm selects the optimistic approximator f^k_{op} from the confidence set $\mathcal{B}^k$. Unlike realizable or globally-bounded misspecified settings, which achieve optimism, our locally-bounded misspecified setting only allows us to establish average optimism. The algorithm then determines the optimal policy π^k based on the optimistic approximator f^k_{op} and collects new data denoted as $\mathcal{Z}^k$ by executing that policy. Finally, the newly collected data is merged into the dataset, and the empirical loss function is updated for the subsequent training episode.

4.2 LBM-UCB for Value-based Algorithm

In the context of value-based function approximation, our algorithm's objective is to learn the optimal state-value function. In this case, our **LBM-UCB** becomes Algorithm 3 (**Robust-LSVI**). Consequently, the loss function takes the form: $L_{k-1}(\mathcal{D}_{k-1}, f) = ||f||^2_{\mathcal{D}^k_h}$ where $\mathcal{D}^k_h = \{(s^{k'}_{h'}, a^{k'}_{h'}, r^{k'}_{h'} + V^k_{h+1}(s^{k'}_{h'+1}))\}_{(k',h')\in[k-1]\times[H]}$.

Before we proceed with constructing the confidence set, we adopt the sensitivity sampling technique as outlined in (Wang et al., 2020b). This technique enables us to substantially reduce the size of the dataset while approximately preserving the confidence region.

In our efforts to encompass the best ground-truth approximator within the confidence set, we meticulously design the radius of the confidence set, denoted as $\beta(\mathcal{F}, \delta)$. This radius is expressed as:

Algorithm 1 LBM-UCB

1: **Input:** The hypothesis class $\mathcal{F}$.
2: **for** episode $k = 1, \cdots, K$ **do**
3: Find the best empirical approximator f^k in the hypothesis class $\mathcal{F}$ by solving the problem (1).

$$f^k = \underset{f \in \mathcal{F}}{\operatorname{argmin}} L_{k-1}(\mathcal{D}^{k-1}, f) \tag{1}$$

4: Construct the high probability confidence set to contain the best ground-truth approximator.

$$\mathcal{B}^k = \{f \in \mathcal{F} |\ d(f, f^k) \leq \mathcal{E}^k_{\text{stat}} + \mathcal{E}^k_{\text{bias}}\} \tag{2}$$

5: Get the optimistic approximator f^k_{op} in the confidence set $\mathcal{B}^k$ and the corresponding optimal policy $\pi^k = \pi_{f^k_{\text{op}}}$.
6: Execute the policy π^k to collect new data $\mathcal{Z}^k = \{\mathcal{Z}^k_h\}_{h \in [H]}$, where $\mathcal{Z}^k_h = \{(s^k_h, a^k_h, r^k_h, s^k_{h+1})\}$.
7: Update the dataset $\mathcal{D}^k \leftarrow \mathcal{D}^{k-1} \cup \mathcal{Z}^k$ and the empirical loss funtion L_k.

$$\beta(\mathcal{F}, \delta) = L(d, K, H, \delta) \cdot \left(\underbrace{\sqrt{kH}\zeta}_{\mathcal{E}^k_{\text{bias}}} + \underbrace{dH^2}_{\mathcal{E}^k_{\text{stat}}} \right) \tag{3}$$

Here, ζ denotes locally bounded misspecification error, and its specific definition is detailed in the section 5, $L(d, K, H, \delta)$ is a function that exhibits logarithmic scaling with respect to all the involved variables.

4.3 LBM-UCB for Model-based Algorithm

In the model-based setting, our algorithm's primary objective is to learn the transition kernel of the underlying MDP. In this context, our **LBM-UCB** becomes Algorithm 4 (**Robust-UCRL-VTR**). To design the loss function, we incorporate the concept of value-targeted regression from (Ayoub et al., 2020). Specifically,

$$\widehat{P}^{(k)} = \underset{P \in \mathcal{P}}{\operatorname{argmin}} L_{k-1}(\mathcal{D}_{k-1}, P) \tag{4}$$

where

$$L_{k-1}(\mathcal{D}_{k-1}, P) = \sum_{k'=1}^{k-1} \sum_{h=1}^{H} \left(P V^{k'}_{h+1}(s^{k'}_h, a^{k'}_h) - V^{k'}_{h+1}(s^{k'}_{h+1}) \right)^2 \tag{5}$$

Subsequently, we define the model distance in relation to the estimated value functions as

$$d_k(P, \widehat{P}^{(k)}) = \sum_{k'=1}^{k-1} \sum_{h=1}^{H} \left(P V^{k'}_{h+1}(s^{k'}_h, a^{k'}_h) - \widehat{P}^{(k)} V^{k'}_{h+1}(s^{k'}_h, a^{k'}_h) \right)^2 \tag{6}$$

We then define the confidence set as

$$B_k = \{P' \in \mathcal{P} |\ d_k(P', \widehat{P}^{(k)}) \leq \beta_k\} \tag{7}$$

The selection of the radius of the confidence set is similar to (3), where $\beta_k = L'(d, K, H, \delta) \cdot \left(\underbrace{\sqrt{kH}\zeta}_{\mathcal{E}^k_{\text{bias}}} + \underbrace{dH^2}_{\mathcal{E}^k_{\text{stat}}} \right)$. Here, ζ denotes locally bounded misspecification error, $L'(d, K, H, \delta)$ is a function that exhibits logarithmic scaling with respect to all the involved variables.

To obtain the optimistic approximator, the algorithm identifies the model that maximizes the optimal value. In other words,

$$P^{(k)} = \underset{P' \in B_k}{\operatorname{argmax}} V^*_{P',1}(s^k_1) \tag{8}$$

where s^k_1 is the initial state at the beginning of episode k, and $V^*_{P',1}$ represents the optimal value function at stage one under transition kernel P'. After that, the algorithm will calculate the corresponding optimal policy for $P^{(k)}$ using dynamic programming. In particular, for each $h \in [1, H+1]$, and all $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned} Q^k_{H+1}(s, a) &= 0 \\ V^k_h(s) &= \max_{a \in \mathcal{A}} Q^k_h(s, a) \\ Q^k_h(s, a) &= r_h(s, a) + P^{(k)} V^k_{h+1}(s, a) \end{aligned} \tag{9}$$

5 THEORETICAL ANALYSIS OF ROBUST RL ALGORITHMS WITH GENERAL FUNCTION APPROXIMATION

In this section, we will provide our theoretical analysis of Robust RL Algorithms with general function approximation in Section 4 under the locally-bounded misspecification assumptions.

First of all, the sample complexity of algorithms with function approximation depends on the complexity of the function class. To measure this complexity, we adopt the notion of eluder dimension which is first mentioned in Russo and Van Roy (2013).

Definition 5.1 (Eluder dimension)**.** *Let $\varepsilon \geq 0$ and $\mathcal{Z} = \{(s_i, a_i)\}_{i=1}^n \subseteq \mathcal{S} \times \mathcal{A}$ be a sequence of state-action pairs.*

- *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 $\|f - f'\|_{\mathcal{Z}} \leq \varepsilon$ also satisfies $|f(s, a) - f'(s, a)| \leq \varepsilon$.*

- *An (s,a) is ε-independent of $\mathcal{Z}$ with respect to $\mathcal{F}$ if (s,a) is not ε-dependent on $\mathcal{Z}$.*
- *The eluder dimension $\dim_E(\mathcal{F},\varepsilon)$ 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 $\varepsilon'\geq\varepsilon$, every element is ε'-independent of its predecessors.*

Next, we discuss the regret bound of these two algorithms respectively.

5.1 Regret Bound of Robust-LSVI

First, we give a theorectical analysis for **Robust-LSVI** with general function approximation. We assume that the function class $\mathcal{F}$ and the state-actions $\mathcal{S}\times\mathcal{A}$ have bounded covering numbers.

Assumption 1 (ε-cover)**.** *For any $\varepsilon>0$, the following holds:*

1. there exists an ε-cover $\mathcal{C}(\mathcal{F},\varepsilon)\subseteq\mathcal{F}$ with size $|\mathcal{C}(\mathcal{F},\varepsilon)|\leq\mathcal{N}(\mathcal{F},\varepsilon)$, such that for any $f\in\mathcal{F}$, there exists $f'\in\mathcal{C}(\mathcal{F},\varepsilon)$ with $\|f-f'\|_\infty\leq\varepsilon$;

2. there exists an ε-cover $\mathcal{C}(\mathcal{S}\times\mathcal{A},\varepsilon)$ with size $|\mathcal{C}(\mathcal{S}\times\mathcal{A},\varepsilon)|\leq\mathcal{N}(\mathcal{S}\times\mathcal{A},\varepsilon)$, such that for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, there exists $(s',a')\in\mathcal{C}(\mathcal{S}\times\mathcal{A},\varepsilon)$ with $\max_{f\in\mathcal{F}}|f(s,a)-f(s',a')|\leq\varepsilon$.

Remark 5.2. *Assumption 1 is rather standard. Since our algorithm complexity depends only logarithmically on $\mathcal{N}(\mathcal{F},\cdot)$ and $\mathcal{N}(\mathcal{S}\times\mathcal{A},\cdot)$, it is even acceptable to have exponential size of these covering numbers.*

Assumption 2 (General value function approximation with LBM)**.** *Given the ground-truth MDP M with the transition model $\mathbb{P}$ and the reward function r, we assume that there exists a function class $\mathcal{F}\subset\{f:\mathcal{S}\times\mathcal{A}\to[0,H+1]\}$ and a real number $\zeta\in[0,1]$, such that for any $V:\mathcal{S}\to[0,H]$, there exists a non-empty function class $\bar{\mathcal{F}}_V\subset\mathcal{F}$, which satisfies : for all $f\in\bar{\mathcal{F}}_V$, and all $\beta\in[4]$,*

$$\sup_\pi \mathbb{E}_{(s,a)\sim d^\pi}\left|f(s,a)-(r(s,a)+\mathbb{P}V(s,a))\right|^\beta\leq\zeta^\beta$$

Theorem 5.3 (Regret bound with known ζ)**.** *Under our Assumption 1 and 2, for any fixed $\delta\in(0,1)$, with probability at least $1-\delta$, the total regret of Algorithm 3 is at most $\widetilde{O}\left(\sqrt{d_E H^3}K\zeta\log(1/\delta)+\sqrt{d_E^2KH^3}\log(1/\delta)\right)$, where d_E represents the eluder dimension of the function class.*

The comprehensive proof is presented in Appendix C.

Remark 5.4. *Our assumption is strictly weaker than the globally-bounded misspecification error in (Wang et al., 2020b), where they assumed that, for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, and $V:\mathcal{S}\to[0,H]$, $|f(s,a)-(r(s,a)+\mathbb{P}V(s,a))|\leq\zeta$.*

In other words, our Assumption 2 only needs the misspecification error to be locally *bounded at relevant state-action pairs, which can be reached by some policies with sufficiently high probability, whereas Wang et al. (2020b) requires the misspecification error to bounded* globally *in all state-action pairs, including those not even relevant to the learning.*

Remark 5.5. *A classical special case of the general function approximation setting is the linear MDP (Yang and Wang, 2019; Jin et al., 2020). In this case, (Agarwal et al., 2023) initially demonstrated the effectiveness of LSVI-UCB (Jin et al., 2020) even under conditions of locally-bounded misspecification error. Our assumptions and findings serve as a more general version of (Agarwal et al., 2023), extending the utility of function approximation from a linear setting to a more general context. For the reader's convenience, we present the result below. Under Assumption 5 and 6 , for any fixed $\delta\in(0,1)$, with probability at least $1-\delta$, the total regret of the algorithm Robust-LSVI (Algorithm 6) is at most $\widetilde{O}\left(dKH^2\zeta\log(1/\delta)+\sqrt{d^3KH^4}\log(1/\delta)\right)$. Here, d represents the dimensionality of the linear features. The comprehensive proof is elaborated upon in Appendix B.*

5.2 Regret Bound of Robust-UCRL-VTR

Next, we provide our assumption and the main theoretical result for **Robust-UCRL-VTR**. Let $\mathcal{V}$ be the set of optimal value functions under some model in the hypothesis class $\mathcal{P}$: $\mathcal{V}=\{V^*_{P'}:P'\in\mathcal{P}\}$. We define $\mathcal{X}=\mathcal{S}\times\mathcal{A}\times\mathcal{V}$, and choose

$$\begin{aligned}\mathcal{F}=\big\{f:\mathcal{X}\to\mathbb{R}:\exists\widetilde{P}\in\mathcal{P}\ \text{ s.t.}\\ f(s,a,V)=\widetilde{\mathbb{P}}V(s,a),\ \ \forall(s,a,V)\in\mathcal{X}\big\}\end{aligned}\tag{10}$$

Similar to Assumption 1, we assume the function class $\mathcal{F}$ defined in (10) has bounded covering numbers.

Assumption 3 (ε-cover)**.** *For any $\varepsilon>0$, the following holds:*

There exists an ε-cover $\mathcal{C}(\mathcal{F},\varepsilon)\subseteq\mathcal{F}$ with size $|\mathcal{C}(\mathcal{F},\varepsilon)|\leq\mathcal{N}(\mathcal{F},\varepsilon)$, such that for any $f\in\mathcal{F}$, there exists $f'\in\mathcal{C}(\mathcal{F},\varepsilon)$ with $\|f-f'\|_\infty\leq\varepsilon$

Assumption 4 (General model function approximation with LBM)**.** *Given the ground-truth MDP M with the transition model $\mathbb{P}$, we assume that there exists a real number $\zeta\in[0,1]$, and $\bar{f}\in\mathcal{F}$ (defined in 10), such that for any $V\in\mathcal{V}$, and any $\beta\in[4]$,*

$$\sup_\pi \mathbb{E}_{(s,a)\sim d^\pi}\left|\bar{f}(s,a,V)-\mathbb{P}V(s,a)\right|^\beta\leq\zeta^\beta$$

Algorithm 2 Meta-algorithm for unknown misspecified parameter ζ

Input: The base algorithm *Alg.*, the total number of episodes K, the length of one episode H, failure probability $\delta > 0$.
for epoch $i = 0, 1, 2, \cdots, \lfloor \log_2 \left(\sqrt{3K+1}\right) \rfloor$ **do**
  $\zeta^{(i)} \leftarrow \frac{1}{2^i}$, $K^{(i)} \leftarrow \frac{1}{(\zeta^{(i)})^2}$,
  $(\overline{V_1}^{(i)}, \pi^{(i)}) \leftarrow$ Algorithm 5 $(Alg., K^{(i)}, H, \delta, \zeta^{(i)})$
  if $i \geq 1$ **then**
    if $|\overline{V_1}^{(i)} - \overline{V_1}^{(i-1)}| > C(d, H, \delta) \cdot \zeta^{(i)}$ **then**
      $j \leftarrow i - 1$,
      break;
for the rest episodes $t = 1, 2, \cdots K - \sum_{i=0}^{j+1} K^{(i)}$ **do**
  for step $h = 1, \cdots, H$ **do**
    Take action $a_h^t \leftarrow \pi^{(j)}(s_h^t)$, and observe s_{h+1}^t.

Now we present the main theorem of our algorithm and the in-depth proof is showcased in Appendix D.

Theorem 5.6 (Regret bound with known ζ)**.** *Under our Assumption 3 and 4, for any fixed $\delta \in (0,1)$, with probability at least $1-\delta$, the total regret of Algorithm 4 is at most $\widetilde{O}\left(\sqrt{d_E}KH\zeta \log(1/\delta) + \sqrt{d_E^2 KH^3}\log(1/\delta)\right)$, where d_E represents the eluder dimension of the function class.*

Remark 5.7. *Our assumption is strictly weaker than that in (Ayoub et al., 2020), where they assume that given the truth model $\mathbb{P}$, there exists an approximator $\bar{P} \in \mathcal{P}$ such that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $||\bar{P}(\cdot|s,a) - \mathbb{P}(\cdot|s,a)||_{TV} \leq \zeta$. To clarify, our Assumption 4 merely necessitates the misspecification error to be locally bounded at the relevant state-action pairs. These pairs are those that can be reached by certain policies with a high probability.*

Remark 5.8. *A direct corollary of our result is for the Linearly-Parametrized Transition Model. Specifically, we assume there are d transition models, $P_1, P_2, \cdots, P_d$, $\Theta \subset \mathbb{R}^d$ is a bounded and nonempty set, and let $\mathcal{P} = \left\{\sum_j \theta_j P_j : \theta \in \Theta\right\}$. Given the ground-truth model $\mathbb{P}$, if there exists a d-dimension vector $\alpha \in \Theta$, such that $\mathbb{E}_{(s,a)\sim d^\pi} ||\mathbb{P}(\cdot|s,a) - \sum_{j=1}^{d} \alpha_j P_j(\cdot|s,a)||_1^\alpha \leq \zeta^\alpha$, $\forall \alpha \in [4]$. Then for any fixed $\delta \in (0,1)$, with probability at least $1-\delta$, the total regret of Algorithm 4 is at most $\widetilde{O}\left(\sqrt{d}KH\zeta\log(1/\delta) + \sqrt{d^2KH^3}\log(1/\delta)\right)$.*

6 META ALGORITHM WITHOUT KNOWING THE MISSPECIFIED PARAMETER

In real-world environments, we cannot assume that the misspecified parameter ζ is provided. This issue serves as motivation for our meta algorithm (Algorithm 2), which makes the base algorithm (e.g., Algorithm 3,4,6) have the parameter-free property by employing exponentially decreasing misspecified parameters and increasing training episodes. Without loss of generality, we assume the initial state s_1 remains fixed across all episodes. The entire training process is divided into multiple epochs. In each epoch, the meta algorithm interacts with the environment (using a small variation of the base algorithm, Algorithm 5 in Appendix A, which is almost the same as the base algorithm except that it outputs the policy and value of each round) a total of $K^{(i)} = 1/(\zeta^{(i)})^2$ times, where $\zeta^{(i)} = 1/2^i$ is the exponentially decreasing misspecified parameter. After each epoch, the real-time reward data is utilized to estimate the value function of the average policy for that round. Notably, when training with misspecified parameter that is roughly the true value (i.e., $\zeta^{(i)} \gtrsim \zeta$), the value estimates from adjacent epochs exhibit minimal variation. However, as the misspecified parameter $\zeta^{(i)}$ decreases below the ground-truth parameter ζ, the obtained policy may deteriorate since the base algorithms do not guarantee optimism in this scenario. Hence, when our stability condition is violated, defined as $|\overline{V_1}^{(i)} - \overline{V_1}^{(i-1)}| > C(d, H, \delta) \cdot \zeta^{(i)}$ (where $C(d, H, \delta)$ is a constant dependent on d, H, δ, see definition in Appendix E), we break out of the loop and execute the last average policy for the remaining episodes. According to our selection of the misspecified parameters, there must exist an accurate parameter $\zeta^{(s)}$ close to the true ζ ($\zeta \leq \zeta^{(s)} < 2\zeta$). As the last executed policy still satisfies the stability condition, it can serve as an approximate good policy for the previous policy with the parameter $\zeta^{(s)}$.

Based on the above analysis, our formal guarantee of Algorithm 2 for the unknown ζ case is presented as follows. The detailed proof is displayed in Appendix E.

Theorem 6.1 (Regret bound with unknown ζ)**.** *Suppose the input base algorithm Alg. which needs to know the locally-bounded misspecified parameter ζ has a regret bound of $\widetilde{O}\left(d^\alpha H^\beta(\sqrt{K} + K\cdot\zeta)\right)$, then our meta-algorithm (Algorithm 2) can achieve the same order of regret bound $\widetilde{O}\left(d^\alpha H^\beta(\sqrt{K} + K\cdot\zeta)\right)$ without knowing the misspecified parameter ζ.*

7 CONCLUSION

In this paper, we have proposed a robust RL algorithm framework for value-based and model-based methods under locally-bounded misspecification error. Through a careful design of the high-probability confidence set and a refined analysis, we have significantly improved the regret bound of (Wang et al., 2020b; Ayoub et al., 2020) when the misspecification error is not *globally bounded*. Furthermore, we have developed a provably efficient meta algorithm to address scenarios where the misspecified parameter is unknown.

Acknowledgements

YL is supported in part by NSF grant 2221871. LY is supported in part by NSF grant 2221871, and an Amazon Research Grant.

References

A. Agarwal, M. Henaff, S. Kakade, and W. Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *Advances in neural information processing systems*, 33:13399–13412, 2020a.

A. Agarwal, S. Kakade, A. Krishnamurthy, and W. Sun. Flambe: Structural complexity and representation learning of low rank MDPs. *Advances in neural information processing systems*, 33:20095–20107, 2020b.

A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(98):1–76, 2021.

A. Agarwal, Y. Song, W. Sun, K. Wang, M. Wang, and X. Zhang. Provable benefits of representational transfer in reinforcement learning. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2114–2187. PMLR, 2023.

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

R. E. Bellman. *Dynamic programming*. Princeton university press, 2010.

Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

Z. Chen, C. J. Li, A. Yuan, Q. Gu, and M. I. Jordan. A general framework for sample-efficient function approximation in reinforcement learning. *arXiv preprint arXiv:2209.15634*, 2022.

Q. Ding, C.-J. Hsieh, and J. Sharpnack. Robust stochastic linear contextual bandits under adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 7111–7123. PMLR, 2022.

S. Du, S. Kakade, J. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

S. S. Du, S. M. Kakade, R. Wang, and L. F. Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019a.

S. S. Du, Y. Luo, R. Wang, and H. Zhang. Provably efficient q-learning with function approximation via distribution shift error checking oracle. *Advances in Neural Information Processing Systems*, 32, 2019b.

Y. Duan, Z. Jia, and M. Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

F. Feng, W. Yin, A. Agarwal, and L. Yang. Provably correct optimization and exploration with non-linear policies. In *International Conference on Machine Learning*, pages 3263–3273. PMLR, 2021.

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

D. J. Foster, C. Gentile, M. Mohri, and J. Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020a.

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

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

A. Ghosh, S. R. Chowdhury, and A. Gopalan. Misspecified linear bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

S. Gu, E. Holly, T. Lillicrap, and S. Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3389–3396. IEEE, 2017.

A. Gupta, T. Koren, and K. Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Proceedings of the Annual Conference on Learning Theory*, 2019.

J. He, H. Zhao, D. Zhou, and Q. Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. *arXiv preprint arXiv:2212.06132*, 2022.

Z. Jia, L. F. Yang, and M. Wang. Feature-based q-learning for two-player stochastic games. *arXiv preprint arXiv:1906.00423*, 2019.

Z. Jia, L. Yang, C. Szepesvari, and M. Wang. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pages 666–686. PMLR, 2020.

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

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

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

D. Kong, R. Salakhutdinov, R. Wang, and L. F. Yang. Online sub-sampling for reinforcement learning with general function approximation. *arXiv preprint arXiv:2106.07203*, 2021.

T. Lattimore, C. Szepesvari, and G. Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.

Y. Li, Y. Wang, Y. Cheng, and L. Yang. Low-switching policy gradient with exploration via online sensitivity sampling. *arXiv preprint arXiv:2306.09554*, 2023.

Z. Liu, M. Lu, W. Xiong, H. Zhong, H. Hu, S. Zhang, S. Zheng, Z. Yang, and Z. Wang. One objective to rule them all: A maximization objective fusing estimation and planning for exploration. *arXiv preprint arXiv:2305.18258*, 2023.

T. Lykouris, V. Mirrokni, and R. Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the Annual ACM SIGACT Symposium on Theory of Computing*, 2018.

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

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

A. Modi, N. Jiang, A. Tewari, and S. Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

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

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

K. Takemura, S. Ito, D. Hatano, H. Sumita, T. Fukunaga, N. Kakimura, and K.-i. Kawarabayashi. A parameter-free algorithm for misspecified linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3367–3375. PMLR, 2021.

D. Vial, A. Parulekar, S. Shakkottai, and R. Srikant. Improved algorithms for misspecified linear markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pages 4723–4746. PMLR, 2022.

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

R. Wang, R. R. Salakhutdinov, and L. Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020b.

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

C.-Y. Wei, C. Dann, and J. Zimmert. A model selection approach for corruption robust reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

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

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

C. Ye, W. Xiong, Q. Gu, and T. Zhang. Corruption-robust algorithms with uncertainty weighting for nonlinear contextual bandits and markov decision processes. In *International Conference on Machine Learning*, pages 39834–39863. PMLR, 2023.

A. Zanette, A. Lazaric, M. J. Kochenderfer, and E. Brunskill. Limiting extrapolation in linear approximate value iteration. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Zanette, C.-A. Cheng, and A. Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory*, pages 4473–4525. PMLR, 2021.

J. Zhang, W. Zhang, and Q. Gu. Optimal horizon-free reward-free exploration for linear mixture mdps. *arXiv preprint arXiv:2303.10165*, 2023a.

W. Zhang, D. Zhou, and Q. Gu. Reward-free model-based reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 34:1582–1593, 2021.

W. Zhang, J. He, Z. Fan, and Q. Gu. On the interplay between misspecification and sub-optimality gap in linear contextual bandits. *arXiv preprint arXiv:2303.09390*, 2023b.

H. Zhao, D. Zhou, and Q. Gu. Linear contextual bandits with adversarial corruptions. *arXiv preprint arXiv:2110.12615*, 2021.

H. Zhong, W. Xiong, S. Zheng, L. Wang, Z. Wang, Z. Yang, and T. Zhang. A posterior sampling framework for interactive decision making. *arXiv preprint arXiv:2211.01962*, 2022.

D. Zhou and Q. Gu. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *Advances in neural information processing systems*, 35:36337–36349, 2022.

D. Zhou, Q. Gu, and C. Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Remaining Algorithm Pseudocodes

We provide the remaining algorithms in this section.

Algorithm 3 Robust-LSVI with general function approximation (ζ is known)

Input: The function class $\mathcal{F}$, the number of episodes K, the length of one episode H, failure probability $\delta > 0$, misspecified parameter ζ.

for episode $k = 1, \cdots, K$ **do**

  Receive the initial state s_1^k.

  Initialize $Q_{H+1}^k(\cdot,\cdot) \leftarrow 0$, $V_{H+1}^k(\cdot) \leftarrow 0$.

  $\mathcal{Z}^k \leftarrow \{(s_{h'}^{k'}, a_{h'}^{k'})\}_{(k',h')\in[k-1]\times[H]}$

  for step $h = H, H-1, \cdots, 1$ **do**

    Find the best empirical approximator:

$$f_h^k \leftarrow \operatorname*{argmin}_{f\in\mathcal{F}} ||f||^2_{\mathcal{D}_h^k}$$

    where

$$\mathcal{D}_h^k = \{(s_{h'}^{k'}, a_{h'}^{k'}, r_{h'}^{k'} + V_{h+1}^k(s_{h'+1}^{k'}))\}_{(k',h')\in[k-1]\times[H]}$$

    $(\widehat{f}_h^k, \widehat{\mathcal{Z}}^k) \leftarrow$ Sensitivity-Sampling$(\mathcal{F}, f_h^k, \mathcal{Z}^k, \delta)$

    Construct the confidence set

$$\widehat{\mathcal{F}} = \left\{ f \in \mathcal{F} : ||f - \widehat{f}_h^k||_{\widehat{\mathcal{Z}}^k} \le \beta(\mathcal{F}, \delta) \right\}$$

    where $\beta(\mathcal{F},\delta) = L(d,K,H,\delta) \cdot \left(\underbrace{\sqrt{kH}\zeta}_{\mathcal{E}_{\text{bias}}^k} + \underbrace{dH^2}_{\mathcal{E}_{\text{stat}}^k} \right)$

    Get the optimistic approximator

$$Q_h^k(\cdot,\cdot) \leftarrow \min\{f_h^k(\cdot,\cdot) + b_h^k(\cdot,\cdot), H\}, \quad V_h^k(\cdot) = \max_{a\in\mathcal{A}} Q_h^k(\cdot,a)$$

    , where

$$b_h^k(\cdot,\cdot) = \sup_{f_1,f_2\in\widehat{\mathcal{F}}} |f_1(\cdot,\cdot) - f_2(\cdot,\cdot)|$$

    Get the corresponding optimal policy

$$\pi_h^k(\cdot) \leftarrow \text{argmax}_{a\in\mathcal{A}} Q_h^k(\cdot,a)$$

  for step $h = 1, \cdots, H$ **do**

    Take action $a_h^k \leftarrow \pi_h^k(s_h^k)$, and observe s_{h+1}^k and $r_h^k = r(s_h^k, a_h^k)$.

Algorithm 4 Robust-UCRL-VTR with general function approximation (ζ is known)

Input: Family of MDP models $\mathcal{P}$, The number of episodes K, the length of one episode H, failure probability $\delta > 0$, misspecified parameter ζ.

$B_1 = \mathcal{P}$

for episode $k = 1, \cdots, K$ **do**

Receive the initial state s_1^k. Find the best empirical approximator:

$$\widehat{P}^{(k)} \leftarrow \operatorname*{argmin}_{P \in \mathcal{P}} L_{k-1}(\mathcal{D}_{k-1}, P) \quad (5)$$

Construct the confidence set:

$$B_k = \{P' \in \mathcal{P} |\ d_k(P', \widehat{P}^{(k)}) \leq \beta_k\} \quad (6)$$

where

$$\beta_k = L'(d, K, H, \delta) \cdot (\underbrace{\sqrt{kH}\zeta}_{\mathcal{E}_{\text{bias}}^k} + \underbrace{H}_{\mathcal{E}_{\text{stat}}^k})$$

Get the optimistic approximator

$$P^{(k)} = \operatorname{argmax}_{P' \in B_k} V_{P',1}^*(s_1^k)$$

Compute $Q_1^k, Q_2^k, \cdots, Q_H^k$ for $P^{(k)}$ using dynamic programming (9).

Get the corresponding optimal policy

$$\pi_h^k(\cdot) \leftarrow \operatorname*{argmax}_{a \in \mathcal{A}} Q_h^k(\cdot, a) \quad , h = 1, 2, \cdots, H$$

for step $h = 1, 2, \cdots, H$ **do**

Take action $a_h^k \leftarrow \pi_h^k(s_h^k)$, and observe s_{h+1}^k and $r_h^k = r(s_h^k, a_h^k)$.

Algorithm 5 Single-epoch-Algorithm

Input: The base algorithm *Alg.*, The number of episodes K in a single epoch, the length of one episode H, failure probability $\delta > 0$, misspecified parameter ζ.
for $k = 1, \cdots, K$ **do**
  Update the policy $\{\pi_h^k\}_{h\in[H]}$ by using *Alg.*
  for step $h = 1, \cdots, H$ **do**
    Take action $a_h^k \leftarrow \pi_h^k(s_h^k)$, and observe s_{h+1}^k.
    Calculate $R^k \leftarrow R^k + r_h(s_h^k, a_h^k)$.
Output: **value** : $\overline{V_1} \leftarrow \frac{1}{K}\sum_{k=1}^{K} R^k$
  policy : $\mathrm{Unif}(\pi^1, \pi^2, \cdots, \pi^K)$

B Analysis of Robust Value-based Algorithm with Linear Function Approximation for Locally-bounded Misspecified MDP

Algorithm 6 Robust-LSVI with linear function approximation (ζ is known)

Input: The number of episodes K, the length of one episode H, failure probability $\delta > 0$, misspecified parameter ζ.
for episode $k = 1, \cdots, K$ **do**
  Receive the initial state s_1^k.
  Initialize $Q_{H+1}^k(\cdot,\cdot) \leftarrow 0$, $V_{H+1}^k(\cdot) \leftarrow 0$.
  Update the bonus parameter $\beta_k \leftarrow c_\beta \left(4\sqrt{kd}\zeta + \sqrt{(\lambda+1)d^2 \log\left(\frac{4dKH}{\delta}\right)}\right) H$.
  for step $h = H, H-1, \cdots, 1$ **do**
    $\Lambda_h^k \leftarrow \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda \cdot \mathbf{I}$.
    $\mathbf{w}_h^k \leftarrow (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)[r_h(s_h^\tau, a_h^\tau) + V_{h+1}^k(s_{h+1}^\tau)]$.
    $Q_h^k(\cdot,\cdot) \leftarrow \min\{(\mathbf{w}_h^k)^\top \phi(\cdot,\cdot) + \beta_k[\phi(\cdot,\cdot)^\top (\Lambda_h^k)^{-1}\phi(\cdot,\cdot)]^{1/2}, H\}$.
    $V_h^k(\cdot) \leftarrow \max_{a\in\mathcal{A}} Q_h^k(\cdot, a)$.
    $\pi_h^k(\cdot) \leftarrow \operatorname{argmax}_{a\in\mathcal{A}} Q_h^k(\cdot, a)$.
  for step $h = 1, \cdots, H$ **do**
    Take action $a_h^k \leftarrow \pi_h^k(s_h^k)$, and observe s_{h+1}^k.

In order to let readers better understand the core idea of our paper, we first give the proof for the linear case before giving the proof for the general case. We study the linear function approximation setting for MDPs introduced in Yang and Wang (2019); Jin et al. (2020), where the probability transition matrix can be approximated by a linear function class. To enable a much stronger *locally bounded* misspecification error, we consider the following notion of ζ-Average-Approximate Linear MDP. It is worth mentioning that Agarwal et al. (2023) also gives a positive result of LSVI-UCB (Jin et al., 2020) under average-misspecification, and here we present another version of the proof.

Assumption 5. *(ζ-Average-Approximate Linear MDP). For any $\zeta \le 1$, we say that the MDP $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ is a ζ-Average-Approximate Linear MDP with a feature map $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$, if for any $h \in [H]$, there exists a d-dimension measures $\boldsymbol{\mu}_h = (\mu_h^{(1)}, \cdots, \mu_h^{(d)})$ over $\mathcal{S}$, and an vector $\boldsymbol{\theta}_h \in \mathbb{R}^d$, such that for any policy π, and any $\alpha \in [4]$, we have*

$$\mathbb{E}_{(s,a)\sim d_h^\pi} ||\mathbb{P}_h(\cdot|s,a) - \langle \phi(s,a), \boldsymbol{\mu}_h(\cdot)\rangle||_{TV}^\alpha \le \zeta^\alpha \quad \text{and} \quad \mathbb{E}_{(s,a)\sim d_h^\pi} |r_h(s,a) - \langle \phi(s,a), \boldsymbol{\theta}_h\rangle|^\alpha \le \zeta^\alpha$$

Remark B.1. *We note that the assumption on the boundedness of the 4-th moments is minor: $\mathbb{E}_{(s,a)\sim d_h^\pi}|f(s,a)|^\alpha$ is bounded for any $\alpha > 1$ as long as f is bounded and $\mathbb{E}_{(s,a)\sim d_h^\pi}|f(s,a)|$ is bounded. We choose a 4-th moment bound for the ease of presentation and fair comparison with existing results.*

Remark B.2. *Our assumption is strictly weaker than the ζ-Approximate Linear MDP in Jin et al. (2020), where they assumed that, for all $(s,a) \in \mathcal{S}\times\mathcal{A}$: $||\mathbb{P}_h(\cdot|s,a) - \langle \phi(s,a), \boldsymbol{\mu}_h(\cdot)\rangle||_{TV} \le \zeta$, $|r_h(s,a) - \langle \phi(s,a), \boldsymbol{\theta}_h\rangle| \le \zeta$.*

In other words, ζ-Average-Approximate Linear MDP only needs the misspecification error to be locally *bounded at relevant state-action pairs, which can be reached by some policies with sufficiently high probability, whereas ζ-Approximate Linear MDP requires the misspecification error to bounded* globally *in all state-action pairs, including those not even relevant to the learning.*

We will also need the following standard assumptions for the regularity of the feature map:

Assumption 6. *(Boundness) Without loss of generality, we assume that $||\phi(s,a)|| \le 1$, and $\max\{||\boldsymbol{\mu}_h(\mathcal{S})||, ||\boldsymbol{\theta}_h||\} \le \sqrt{d}$, $\forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$.*

The following analysis in Section B are based on Assumption 5 and 6.

To simplify our notation, for each $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, we denote

$$\xi_h(s,a) = ||\mathbb{P}_h(\cdot|s,a) - \widetilde{\mathbb{P}_h}(\cdot|s,a)||_{\mathrm{TV}}$$

$$\eta_h(s,a) = |r_h(s,a) - \widetilde{r_h}(s,a)|$$

where $\widetilde{\mathbb{P}_h}(\cdot|s,a) := \langle \phi(s,a), \boldsymbol{\mu}_h(\cdot)\rangle$, and $\widetilde{r_h}(s,a) := \langle \phi(s,a), \boldsymbol{\theta}_h\rangle$.

With the above notation, $\xi_h(s,a)$, $\eta_h(s,a)$ denotes the model misspecification error for transition kernel $\mathbb{P}_h$ and reward function r_h of a fixed state-action pair (s,a).

Moreover, for all $(k,h) \in [K] \times [H]$, we denote $\phi_h^k := \phi(s_h^k, a_h^k)$.

In the following analysis, we consider an auxiliary stochastic process, which, although unattainable in reality, proves valuable for our analysis.

Auxiliary Stochastic Process For each episode, denoted by $k \in [K]$, we collect the dataset $D_k = \{(s_h^k, a_h^k)\}_{h=1}^H$ using policies $\{\pi_h^k\}_{h=1}^H$ trained by the algorithm in the last k episodes. Additionally, we allow the agent to gather data $D_k^* = \{(s_h^{k*}, a_h^{k*})\}_{h=1}^H$ using optimal policies $\{\pi_h^*\}_{h=1}^H$ within the MDP.

It is worth observing that this auxiliary stochastic process closely resembles the original training process, with the sole addition being a dataset sampled under optimal policies. However, it is crucial to emphasize that this additional dataset does not influence our training course. Consequently, the results obtained from the original stochastic process remain valid in this auxiliary stochastic process. To formalize this concept, we define the following filtration: $\mathcal{F}_0 = \{\emptyset, \Omega\}$, $\mathcal{F}_1 = \sigma\left(\{D_1, D_1^*\}\right), \cdots$, $\mathcal{F}_k = \sigma\left(\{D_1, D_1^*, \cdots, D_k, D_k^*\}\right), \cdots$, $\mathcal{F}_K = \sigma\left(\{D_1, D_1^*, \cdots, D_K, D_K^*\}\right)$. Here, $\sigma(D)$ represents the filtration induced by the dataset D.

B.1 Proof of Theorem 5.3

In this section, we present the comprehensive proof of Theorem 5.3. Prior to providing the proof for the main theorem (Theorem B.11), it is necessary to establish the foundation through the following lemmas.

Lemma B.3. *(Misspecification Error for Q-function). For a ζ-Average-Approximate Linear MDP (Assumption 5), for any fixed policy π, any $h \in [H]$, there exists weights $\{\mathbf{w}_h^\pi\}_{h\in[H]}$, where $\mathbf{w}_h^\pi = \boldsymbol{\theta}_h + \int V_{h+1}^\pi(s')d\boldsymbol{\mu}_h(s')$, such that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$|Q_h^\pi(s,a) - \langle \phi(s,a), \mathbf{w}_h^\pi\rangle| \le \eta_h(s,a) + H \cdot \xi_h(s,a)$$

Proof. This proof is straightforward by using the property of Q-function.

$$\begin{aligned} |Q_h^\pi(s,a) - \langle \phi(s,a), \mathbf{w}_h^\pi\rangle| &= \left| r_h(s,a) + \mathbb{P}_h V_{h+1}^\pi(s,a) - \left\langle \phi(s,a), \boldsymbol{\theta}_h + \int V_{h+1}^\pi(s')d\boldsymbol{\mu}_h(s')\right\rangle \right| \\ &\le |r_h(s,a) - \langle \phi(s,a), \boldsymbol{\theta}_h\rangle| + \left| \mathbb{P}_h V_{h+1}^\pi(s,a) - \left\langle \phi(s,a), \int V_{h+1}^\pi(s')d\boldsymbol{\mu}_h(s')\right\rangle \right| \\ &\le \eta_h(s,a) + H \cdot \xi_h(s,a) \end{aligned} \tag{11}$$

□

Lemma B.4. *For any $h \in [H]$,*

$$||\mathbf{w}_h^\pi|| \le 2H\sqrt{d}$$

Proof. For any policy π, $\mathbf{w}_h^\pi = \boldsymbol{\theta_h} + \int V_{h+1}^\pi(s')d\boldsymbol{\mu}_h(s')$, therefore, under Assumption 6, we have:

$$||\mathbf{w}_h^\pi|| \le ||\boldsymbol{\theta}_h|| + ||\int V_{h+1}^\pi(s')d\boldsymbol{\mu}_h(s')|| \le \sqrt{d} + H\sqrt{d} \le 2H\sqrt{d}$$

□

Lemma B.5. *Let c_β be a constant in the definition of β_k, where $\beta_k = c_\beta \left(4\sqrt{kd}\zeta + \sqrt{(\lambda+1)d^2 \log\left(\frac{4dKH}{\delta}\right)}\right) H$. Then under Assumption 5, 6, there exists an absolute constant C that is independent of c_β such that for any fixed $\delta \in [0,1]$, we have for all $(k,h) \in [K] \times [H]$, with probability at least $1-\delta$,*

$$\left\| \sum_{\tau=1}^{k-1} \phi_h^\tau [V_{h+1}^k(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^k(s_h^\tau, a_h^\tau)] \right\|_{(\Lambda_h^k)^{-1}} \leq C \cdot dH \sqrt{\log\left[2(c_\beta+1)dKH/\delta\right]}$$

Proof. Lemmas G.3 and Lemma G.5 together imply that for all $(k,h) \in [K] \times [H]$, with probability at least $1-\delta$,

$$\begin{aligned} &\left\| \sum_{\tau=1}^{k-1} \phi_h^\tau [V_{h+1}^k(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^k(s_h^\tau, a_h^\tau)] \right\|_{(\Lambda_h^k)^{-1}}^2 \\ &\leq 4H^2 \left(\frac{d}{2} \log(\frac{k+\lambda}{\lambda}) + d \log\left(1 + \frac{8H\sqrt{dk}}{\epsilon\sqrt{\lambda}}\right) + d^2 \log\left[1 + 8d^{1/2}B^2/(\lambda\epsilon^2)\right] + \log(\frac{1}{\delta}) \right) + \frac{8k^2\epsilon^2}{\lambda} \end{aligned} \tag{12}$$

We let $\epsilon = dH/k$, and $B = \max_k \beta_k = c_\beta \left(4\sqrt{Kd}\zeta + \sqrt{(\lambda+1)d^2 \log\left(\frac{4dKH}{\delta}\right)}\right) H$, then from Eq.(12), there exists a constant C which is independent of c_β such that

$$\left\| \sum_{\tau=1}^{k-1} \phi_h^\tau [V_{h+1}^k(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^k(s_h^\tau, a_h^\tau)] \right\|_{(\Lambda_h^k)^{-1}} \leq C \cdot dH \sqrt{\log\left[2(c_\beta+1)dKH/\delta\right]}$$

□

Lemma B.6. *During the course of training, we define the mixed misspecification error $\epsilon_h^\tau = (r_h - \widetilde{r_h})(s_h^\tau, a_h^\tau) + \left(\mathbb{P}_h - \widetilde{\mathbb{P}_h}\right) V_{h+1}^k(s_h^\tau, a_h^\tau)$, $\forall (\tau, h) \in [K] \times [H]$, then for any fixed policy π, conditioned on the event in Lemma B.5, we have for all $(s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, that*

$$\begin{aligned} &|\left\langle \phi(s,a), \mathbf{w}_h^k \right\rangle - Q_h^\pi(s,a) - \mathbb{P}_h(V_{h+1}^k - V_{h+1}^\pi)(s,a)| \\ &\leq \lambda_h^k \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1} \phi(s,a)} + 3H \cdot \xi_h(s,a) + \eta_h(s,a) \end{aligned} \tag{13}$$

where

$$\lambda_h^k = 4H\sqrt{\lambda d} + C \cdot dH \sqrt{\log\left[2(c_\beta+1)dKH/\delta\right]} + \sqrt{d} \sqrt{\sum_{\tau=1}^{k-1} (\epsilon_h^\tau)^2}$$

Proof. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have $\langle \phi(s,a), \mathbf{w}_h^\pi \rangle = \langle \phi(s,a), \boldsymbol{\theta}_h \rangle + \widetilde{\mathbb{P}_h} V_{h+1}^\pi(s,a)$. Therefore, we have

$$\begin{aligned} \mathbf{w}_h^k - \mathbf{w}_h^\pi &= (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau [r_h^\tau + V_{h+1}^k(s_{h+1}^\tau)] - \mathbf{w}_h^\pi \\ &= (\Lambda_h^k)^{-1} \{ -\lambda \mathbf{w}_h^\pi + \sum_{\tau=1}^{k-1} \phi_h^\tau [r_h^\tau + V_{h+1}^k(s_{h+1}^\tau) - (\phi_h^\tau)^\top \theta_h - \widetilde{\mathbb{P}_h} V_{h+1}^\pi(s_h^\tau, a_h^\tau)] \} \\ &= \underbrace{-\lambda (\Lambda_h^k)^{-1} \mathbf{w}_h^\pi}_{p_1} + \underbrace{(\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau [V_{h+1}^k(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^k(s_h^\tau, a_h^\tau)]}_{p_2} \\ &\quad + \underbrace{(\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau \widetilde{\mathbb{P}}_h (V_{h+1}^k - V_{h+1}^\pi)(s_h^\tau, a_h^\tau)}_{p_3} + \underbrace{(\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau [r_h^\tau - (\phi_h^\tau)^\top \boldsymbol{\theta}_h + \left(\mathbb{P}_h - \widetilde{\mathbb{P}_h}\right) V_{h+1}^k(s_h^\tau, a_h^\tau)]}_{p_4} \end{aligned} \tag{14}$$

For the last term, according to our definition in Lemma B.6, $r_h^\tau - (\phi_h^\tau)^\top \theta_h + \left(\mathbb{P}_h - \widetilde{\mathbb{P}_h}\right) V_{h+1}^k(s_h^\tau, a_h^\tau) = \epsilon_h^\tau$, and notice that $|\epsilon_h^\tau| \le \eta_h(s_h^\tau, a_h^\tau) + H \cdot \xi_h(s_h^\tau, a_h^\tau)$.

For the first term p_1, by Lemma B.4, we have

$$|\langle \phi(s,a), p_1 \rangle| = |\lambda \langle \phi(s,a), (\Lambda_h^k)^{-1} \mathbf{w}_h^\pi \rangle| \le \sqrt{\lambda} ||\mathbf{w}_h^\pi|| \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1} \phi(s,a)} \le 2H\sqrt{\lambda d} \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1} \phi(s,a)}$$

Conditioned on Lemma B.5, we have

$$|\langle \phi(s,a), p_2 \rangle| \le C \cdot dH \sqrt{\log\left[2(c_\beta+1)dKH/\delta\right]} \cdot \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1} \phi(s,a)}$$

For the third term,

$$\begin{aligned}
\langle \phi(s,a), p_3 \rangle &= \left\langle \phi(s,a), (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau \widetilde{\mathbb{P}}_h (V_{h+1}^k - V_{h+1}^\pi)(s_h^\tau, a_h^\tau) \right\rangle \\
&= \left\langle \phi(s,a), (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau (\phi_h^\tau)^\top \int (V_{h+1}^k - V_{h+1}^\pi)(s') d\boldsymbol{\mu}_h(s') \right\rangle \\
&= \underbrace{\left\langle \phi(s,a), \int (V_{h+1}^k - V_{h+1}^\pi)(s') d\boldsymbol{\mu}_h(s') \right\rangle}_{t_1} \underbrace{- \lambda \left\langle \phi(s,a), (\Lambda_h^k)^{-1} \int (V_{h+1}^k - V_{h+1}^\pi)(s') d\boldsymbol{\mu}_h(s') \right\rangle}_{t_2}
\end{aligned} \tag{15}$$

Notice that $t_1 = \widetilde{\mathbb{P}_h}(V_{h+1}^k - V_{h+1}^\pi)(s,a)$, $|t_2| \le 2H\sqrt{d\lambda}\sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1}\phi(s,a)}$, and

$$\left| t_1 - \mathbb{P}_h(V_{h+1}^k - V_{h+1}^\pi)(s,a) \right| = \left| (\widetilde{\mathbb{P}}_h - \mathbb{P}_h)(V_{h+1}^k - V_{h+1}^\pi)(s,a) \right| \le 2H \cdot \xi_h(s,a)$$

For the last term p_4,

$$\begin{aligned}
|\langle \phi(s,a), p_4 \rangle| &= \left| \left\langle \phi(s,a), (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau \epsilon_h^\tau \right\rangle \right| \\
&\le \sum_{\tau=1}^{k-1} |(\epsilon_h^\tau \phi)^\top (\Lambda_h^k)^{-1} \phi_h^\tau| \\
&\le \sqrt{\left(\sum_{\tau=1}^{k-1} (\epsilon_h^\tau \phi)^\top (\Lambda_h^k)^{-1} (\epsilon_h^\tau \phi) \right) \left(\sum_{\tau=1}^{k-1} (\phi_h^\tau)^\top (\Lambda_h^k)^{-1} \phi_h^\tau \right)} \\
&= \sqrt{\sum_{\tau=1}^{k-1} (\epsilon_h^\tau)^2} \cdot \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1} \phi(s,a)} \cdot \sqrt{\sum_{\tau=1}^{k-1} (\phi_h^\tau)^\top (\Lambda_h^k)^{-1} \phi_h^\tau} \\
&\le \sqrt{\sum_{\tau=1}^{k-1} (\epsilon_h^\tau)^2} \cdot \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1} \phi(s,a)} \cdot \sqrt{d} \quad \text{(By Lemma G.1)}
\end{aligned} \tag{16}$$

Finally, Combined with the result in Lemma B.3, we have

$$\begin{aligned}
&|\langle \phi(s,a), \mathbf{w}_h^k \rangle - Q_h^\pi(s,a) - \mathbb{P}_h(V_{h+1}^k - V_{h+1}^\pi)(s,a)| \\
&\le |\langle \phi(s,a), \mathbf{w}_h^k - \mathbf{w}_h^\pi \rangle - \mathbb{P}_h(V_{h+1}^k - V_{h+1}^\pi)(s,a)| + |Q_h^\pi(s,a) - \langle \phi(s,a), \mathbf{w}_h^\pi \rangle| \\
&\le \underbrace{\left\{ 4H\sqrt{\lambda d} + C \cdot dH \sqrt{\log\left[2(c_\beta+1)dKH/\delta\right]} + \sqrt{d} \sqrt{\sum_{\tau=1}^{k-1} (\epsilon_h^\tau)^2} \right\}}_{\lambda_h^k} \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1} \phi(s,a)} + 3H \cdot \xi_h(s,a) + \eta_h(s,a)
\end{aligned} \tag{17}$$

□

Lemma B.7. *(Recursive formula). We define* $\delta_h^k = V_h^k(s_h^k) - V_h^{\pi_k}(s_h^k)$, *and* $\zeta_{h+1}^k = \mathbb{E}[\delta_{h+1}^k | s_h^k, a_h^k] - \delta_{h+1}^k$. *Then, conditioned on the event in Lemma B.5, we have for any* $(k,h) \in [K] \times [H]$*:*

$$\delta_h^k \leq \delta_{h+1}^k + \zeta_{h+1}^k + (\lambda_h^k + \beta_k)\sqrt{(\phi_h^k)^\top (\Lambda_h^k)^{-1}\phi_h^k} + 3H \cdot \xi_h(s_h^k, a_h^k) + \eta_h(s_h^k, a_h^k) \tag{18}$$

Proof. By Lemma B.6, we have for any $(s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$:

$$\begin{aligned} Q_h^k(s,a) - Q_h^{\pi_k}(s,a) &\leq \left\langle \phi(s,a), \mathbf{w}_h^k \right\rangle + \beta_k \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1}\phi(s,a)} - Q_h^{\pi_k}(s,a) \\ &\leq \mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi_k})(s,a) + (\lambda_h^k + \beta_k)\sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1}\phi(s,a)} + 3H \cdot \xi_h(s,a) + \eta_h(s,a) \end{aligned} \tag{19}$$

Notice that $\delta_h^k = Q_h^k(s_h^k, a_h^k) - Q_h^{\pi_k}(s_h^k, a_h^k)$, and $\mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi_k})(s_h^k, a_h^k) = \mathbb{E}[\delta_{h+1}^k | s_h^k, a_h^k]$,

which finishes the proof. □

Lemma B.8. *(Bound of cumulative misspecification error). With probability at least* $1-\delta$*,*

$$\sum_{k=1}^K \sum_{h=1}^H \xi_h(s_h^k, a_h^k) \leq \sqrt{8KH^2 \log(\frac{4}{\delta})} + KH\zeta \tag{20}$$

and

$$\sum_{k=1}^K \sum_{h=1}^H \eta_h(s_h^k, a_h^k) \leq \sqrt{32dKH^2 \log(\frac{4}{\delta})} + KH\zeta \tag{21}$$

Proof. We define $X_k = \sum_{h=1}^H \xi_h(s_h^k, a_h^k)$, $Z_0 = 0$, $Z_k = \sum_{i=1}^k X_i - \sum_{i=1}^k \mathbb{E}[X_i | \mathcal{F}_{i-1}]$, $k = 1, 2, \cdots, K$. Notice that $\{Z_k\}_{k=1}^K$ is a martingale, and $|Z_k - Z_{k-1}| = |X_k - \mathbb{E}[X_k|\mathcal{F}_{k-1}]| \leq 2H$, $\forall k \in [K]$.

Then by Azuma-Hoeffding's inequality: For any $\epsilon > 0$,

$$\mathbb{P}\left(|Z_K - Z_0| \geq \epsilon\right) \leq 2 \exp\left\{\frac{-\epsilon^2}{2K \cdot 4H^2}\right\}$$

which means that, with probability at least $1-\delta$,

$$|\sum_{k=1}^K X_k - \sum_{k=1}^K \mathbb{E}[X_k|\mathcal{F}_{k-1}]| \leq \sqrt{8KH^2 \log(\frac{2}{\delta})}$$

For the term $\sum_{k=1}^K \mathbb{E}[X_k|\mathcal{F}_{k-1}]$, notice that

$$\mathbb{E}[X_k|\mathcal{F}_{k-1}] = \sum_{h=1}^H \mathbb{E}[\xi_h(s_h^k, a_h^k)|\mathcal{F}_{k-1}] = \sum_{h=1}^H \mathbb{E}_{(s,a)\sim d_h^{\pi_k}}[\xi_h(s,a)] \leq H\zeta \quad \text{(By Assumption 5)}$$

Therefore,

$$\sum_{k=1}^K \mathbb{E}[X_k|\mathcal{F}_{k-1}] \leq KH\zeta$$

Finally, with probablity at least $1-\delta$,

$$\sum_{k=1}^K \sum_{h=1}^H \xi_h(s_h^k, a_h^k) \leq \sqrt{8KH^2 \log(\frac{2}{\delta})} + KH\zeta$$

and similarly, with probability at least $1-\delta$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\eta_h(s_h^k,a_h^k) \le \sqrt{32dKH^2\log(\frac{2}{\delta})} + KH\zeta$$

By taking the union bound, we achieve the result. □

Lemma B.9. *(Bound of bonus parameter). With probability at least $1-\delta$, for all $(k,h)\in[K]\times[H]$, it holds that*

$$\lambda_h^k \le \beta_k$$

where

$$\beta_k = c_\beta\left(4\sqrt{kd}\zeta + \sqrt{(\lambda+1)d^2\log\left(\frac{4dKH}{\delta}\right)}\right)H$$

Proof. For any fixed $(k,h)\in[K]\times[H]$, we have

$$\begin{aligned}(\lambda_h^k)^2 &= \left(4H\sqrt{\lambda d} + C\cdot dH\sqrt{\log\left[2(c_\beta+1)dKH/\delta\right]} + \sqrt{d}\sqrt{\sum_{\tau=1}^{k-1}(\epsilon_h^\tau)^2}\right)^2 \\ &\le 2\left(\left(4H\sqrt{\lambda d} + C\cdot dH\sqrt{\log\left[2(c_\beta+1)dKH/\delta\right]}\right)^2 + d\sum_{\tau=1}^{k-1}(\epsilon_h^\tau)^2\right)\end{aligned} \tag{22}$$

Notice that

$$\begin{aligned}\sum_{\tau=1}^{k-1}(\epsilon_h^\tau)^2 &\le \sum_{\tau=1}^{k-1}\left(\eta_h(s_h^\tau,a_h^\tau) + H\cdot\xi_h(s_h^\tau,a_h^\tau)\right)^2 \\ &\le 2\sum_{\tau=1}^{k-1}\left(\eta_h^2(s_h^\tau,a_h^\tau) + H^2\xi_h^2(s_h^\tau,a_h^\tau)\right)\end{aligned} \tag{23}$$

Case 1 $k \ge \frac{64d^2\log(\frac{4}{\delta})}{\zeta^4}$ By applying Azuma-Hoeffding's inequality, with probability at least $1-\delta/2$, we have

$$\begin{aligned}\sum_{\tau=1}^{k-1}\xi_h^2(s_h^\tau,a_h^\tau) &\le \sum_{\tau=1}^{k-1}\mathbb{E}\left[\xi_h^2(s_h^\tau,a_h^\tau)|\mathcal{F}_{\tau-1}\right] + \sqrt{8k\log(\frac{4}{\delta})} \\ &\le \sum_{\tau=1}^{k}\mathbb{E}_{(s_h^\tau,a_h^\tau)\sim d_h^{\pi_\tau}}\left[\xi_h^2(s_h^\tau,a_h^\tau)\right] + \sqrt{8k\log(\frac{4}{\delta})} \\ &\le k\zeta^2 + \sqrt{8k\log(\frac{4}{\delta})} \quad \text{(By Assumption 5)}\end{aligned} \tag{24}$$

In the same way, with probability at least $1-\delta/2$, we have

$$\sum_{\tau=1}^{k-1}\eta_h^2(s_h^\tau,a_h^\tau) \le k\zeta^2 + \sqrt{128kd^2\log(\frac{4}{\delta})} \tag{25}$$

Therefore, with probability at least $1-\delta$,

$$\begin{aligned}\sum_{\tau=1}^{k-1}(\epsilon_h^\tau)^2 &\le 2\left(k\zeta^2 + \sqrt{128kd^2\log(\frac{4}{\delta})}\right) + 2H^2\left(k\zeta^2 + \sqrt{8k\log(\frac{4}{\delta})}\right) \\ &\le 4kH^2\zeta^2 + 32dH^2\sqrt{k\log(\frac{4}{\delta})} \\ &\le 8kH^2\zeta^2\end{aligned} \tag{26}$$

Case 2 $k < \frac{64d^2\log(\frac{4}{\delta})}{\zeta^4}$ We denote $X_\tau = \xi_h^2(s_h^\tau, a_h^\tau) - \mathbb{E}[\xi_h^2(s_h^\tau, a_h^\tau)|\mathcal{F}_{\tau-1}]$, and $Y_0 = 0$, $Y_\tau = \sum_{i=1}^{\tau} X_i$, $\tau = 1, 2, \cdots, k-1$.

Notice that $\{Y_\tau\}_{\tau=1}^{k-1}$ is a martingale.

$$\begin{aligned}
\sum_{\tau=1}^{k-1}\mathbb{E}[X_\tau^2|\mathcal{F}_{\tau-1}] &= \sum_{\tau=1}^{k-1}\mathbb{E}\left[\left(\xi_h^2(s_h^\tau, a_h^\tau) - \mathbb{E}[\xi_h^2(s_h^\tau, a_h^\tau)|\mathcal{F}_{\tau-1}]\right)^2|\mathcal{F}_{\tau-1}\right] \\
&\le \sum_{\tau=1}^{k-1}\mathbb{E}\left[\xi_h^4(s_h^\tau, a_h^\tau) + \left(\mathbb{E}[\xi_h^2(s_h^\tau, a_h^\tau)|\mathcal{F}_{\tau-1}]\right)^2|\mathcal{F}_{\tau-1}\right] \\
&= \sum_{\tau=1}^{k-1}\mathbb{E}\left[\xi_h^4(s_h^\tau, a_h^\tau)|\mathcal{F}_{\tau-1}\right] + \sum_{\tau=1}^{k-1}\left(\mathbb{E}[\xi_h^2(s_h^\tau, a_h^\tau)|\mathcal{F}_{\tau-1}]\right)^2 \\
&\le 2\sum_{\tau=1}^{k-1}\mathbb{E}\left[\xi_h^4(s_h^\tau, a_h^\tau)|\mathcal{F}_{\tau-1}\right] \quad \text{(By Jensen's inequality)} \\
&\le 2k\zeta^4 \quad \text{(By Assumption 5)}
\end{aligned} \tag{27}$$

By applying Freedman's inequality (Lemma G.6), we have for any $t \ge 0$, that

$$\mathbb{P}\left(|Y_k - Y_0| \ge t\right) \le 2\exp\left\{-\frac{t^2/2}{2k\zeta^4 + 2t/3}\right\} \le 2\exp\left\{-\frac{t^2/2}{128d^2\log\left(\frac{4}{\delta}\right) + 2t/3}\right\}$$

We let the rightmost term in the above formula to be $\delta/2$, and by solving the quadratic equation with respect to t, we get $t = C_1 \cdot d\log(\frac{4}{\delta})$, where C_1 is some constant. Therefore, we have with probability at least $1 - \delta/2$,

$$\sum_{\tau=1}^{k-1}\xi_h^2(s_h^\tau, a_h^\tau) \le k\zeta^2 + C_1 \cdot d\log(\frac{4}{\delta})$$

Similarly, with probability at least $1 - \delta/2$,

$$\sum_{\tau=1}^{k-1}\eta_h^2(s_h^\tau, a_h^\tau) \le k\zeta^2 + C_2 \cdot d\log(\frac{4}{\delta})$$

where C_2 is some constant. Therefore, in this case, with probability at least $1 - \delta$,

$$\sum_{\tau=1}^{k-1}(\epsilon_h^\tau)^2 \le 4kH^2\zeta^2 + C'dH^2\log(\frac{4}{\delta})$$

where $C' = 2(C_1 + C_2)$ is also some constant.

Combining two cases, we have for any fixed $(k, h) \in [K] \times [H]$, with probability at least $1 - \delta$,

$$\sum_{\tau=1}^{k-1}(\epsilon_h^\tau)^2 \le 8kH^2\zeta^2 + C'dH^2\log(\frac{4}{\delta}) \tag{28}$$

Finally, by taking the union bound of all $(k, h) \in [K] \times [H]$, we have with probability at least $1 - \delta$,

$$\begin{aligned}
(\lambda_h^k)^2 &\le 32dH^2\lambda + 4C^2d^2H^2\log\left(\frac{2(c_\beta+1)dKH}{\delta}\right) + 16dkH^2\zeta^2 + 2C'd^2H^2\log(\frac{4KH}{\delta}) \\
&\le 16dkH^2\zeta^2 + 32(\lambda+1)(C+C')^2d^2H^2\log\left(\frac{4(c_\beta+1)dKH}{\delta}\right)
\end{aligned} \tag{29}$$

This means that

$$\lambda_h^k \le 4\sqrt{kd}H\zeta + \sqrt{32(\lambda+1)(C+C')^2}dH\sqrt{\log\left(\frac{4(c_\beta+1)dKH}{\delta}\right)}$$

By choosing an appropriate c_β, we have

$$\lambda_h^k \le c_\beta \left(4\sqrt{kd}\zeta + \sqrt{(\lambda+1)d^2 \log\left(\frac{4dKH}{\delta}\right)}\right) H = \beta_k$$

□

Lemma B.10. *(Near-Optimism) Given K initial points $\{s_1^k\}_{k=1}^K$, we use $\{(s_h^{k*}, a_h^{k*})\}_{(k,h)\in[K]\times[H]}$ (where $s_1^{k*} = s_1^k$, $\forall k \in [K]$) to represent the dataset sampled by the optimal policy π^* in the true environment (This dataset is impossible to obtain in reality, but it can be used for our analysis), and we denote $\delta_h^{k*} = V_h^*(s_h^{k*}) - V_h^k(s_h^{k*})$, and $\zeta_h^{k*} = \mathbb{E}\left[\delta_{h+1}^{k*} | s_h^{k*}, a_h^{k*}\right] - \delta_{h+1}^{k*}$. Then conditioned on the event in Lemma B.9, with probability at least $1-\delta$, we have*

$$\sum_{k=1}^K [V_1^*(s_1^k) - V_1^k(s_1^k)] \le 4KH^2\zeta + 12H^2\sqrt{dK\log(\frac{8}{\delta})} \tag{30}$$

Proof. First of all, by the definition of V_1^k, we have

$$V_1^k(s_1^{k*}) = \max_{a\in\mathcal{A}} Q_1^k(s_1^{k*}, a) \ge Q_1^k(s_1^{k*}, a_1^{k*})$$

Therefore,

$$\sum_{k=1}^K [V_1^*(s_1^{k*}) - V_1^k(s_1^{k*})] \le \sum_{k=1}^K \left(Q_1^*(s_1^{k*}, a_1^{k*}) - Q_1^k(s_1^{k*}, a_1^{k*})\right) \tag{31}$$

Notice that if we have $Q_1^k(s_1^{k*}, a_1^{k*}) = H$ for some k, then $Q_1^*(s_1^{k*}, a_1^{k*}) \le Q_1^k(s_1^{k*}, a_1^{k*})$, which means we achieve an optimistic estimate, and this term will less than or equal to 0 in Eq.(31). Therefore, we only need to consider the situation when $Q_1^k(s_1^{k*}, a_1^{k*}) < H$, $\forall k \in [K]$.

In this case,

$$Q_1^k(s_1^{k*}, a_1^{k*}) = \left\langle \phi(s_1^{k*}, a_1^{k*}), \mathbf{w}_1^k \right\rangle + \beta_k \sqrt{\phi(s_1^{k*}, a_1^{k*})^\top (\Lambda_1^k)^{-1} \phi(s_1^{k*}, a_1^{k*})}$$

Then we have

$$\begin{aligned}
&\sum_{k=1}^K \left(Q_1^*(s_1^{k*}, a_1^{k*}) - Q_1^k(s_1^{k*}, a_1^{k*})\right) \\
=& \sum_{k=1}^K \left(Q_1^*(s_1^{k*}, a_1^{k*}) - \left\langle \phi(s_1^{k*}, a_1^{k*}), \mathbf{w}_1^k \right\rangle - \beta_k \sqrt{\phi(s_1^{k*}, a_1^{k*})^\top (\Lambda_1^k)^{-1} \phi(s_1^{k*}, a_1^{k*})}\right) \\
\le& \sum_{k=1}^K \left((\lambda_h^k - \beta_k)\sqrt{\phi(s_1^{k*}, a_1^{k*})^\top (\Lambda_1^k)^{-1} \phi(s_1^{k*}, a_1^{k*})} + 3H\xi_1(s_1^{k*}, a_1^{k*}) + \eta_1(s_1^{k*}, a_1^{k*}) - \mathbb{P}_1(V_2^k - V_2^*)(s_1^{k*}, a_1^{k*})\right)
\end{aligned} \tag{32}$$

where the last inequality is derived by Lemma B.6.

By conditioning on the event in Lemma B.9, we know that $\lambda_h^k \le \beta_k$, $\forall (k,h) \in [K]\times[H]$.

In addition,

$$\mathbb{P}_1(V_2^* - V_2^k)(s_1^{k*}, a_1^{k*}) = \mathbb{E}_{s_2^{k*}\sim\mathbb{P}_1(\cdot|s_1^{k*}, a_1^{k*})}\left[V_2^*(s_2^{k*}) - V_2^k(s_2^{k*})\right] \tag{33}$$

and

$$V_2^*(s_2^{k*}) - V_2^k(s_2^{k*}) \le Q_2^*(s_2^{k*}, a_2^{k*}) - Q_2^k(s_2^{k*}, a_2^{k*})$$

Similar to Eq.(31), we only need to consider the case when $Q_2^k(s_2^{k*}, a_2^{k*}) < H$, $\forall k \in [K]$. In this way, we can recursively use Lemma B.6.

Therefore,

$$\begin{aligned}
&\sum_{k=1}^{K}\left(Q_1^*(s_1^{k*},a_1^{k*})-Q_1^k(s_1^{k*},a_1^{k*})\right)\\
&\le\sum_{k=1}^{K}\left(3H\cdot\xi_1(s_1^{k*},a_1^{k*})+\eta_1(s_1^{k*},a_1^{k*})+\delta_2^{k*}+\zeta_1^{k*}\right)\\
&\le\sum_{k=1}^{K}\left(3H\cdot\xi_1(s_1^{k*},a_1^{k*})+\eta_1(s_1^{k*},a_1^{k*})+3H\cdot\xi_2(s_2^{k*},a_2^{k*})+\eta_2(s_2^{k*},a_2^{k*})+\delta_3^{k*}+\zeta_2^{k*}+\zeta_1^{k*}\right)\\
&\le\cdots\le 3H\sum_{k=1}^{K}\sum_{h=1}^{H}\xi_h(s_h^{k*},a_h^{k*})+\sum_{k=1}^{K}\sum_{h=1}^{H}\eta_h(s_h^{k*},a_h^{k*})+\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_h^{k*}
\end{aligned}\tag{34}$$

Similar to Lemma B.8, We define $X_k^*=\sum_{h=1}^{H}\xi_h(s_h^{k*},a_h^{k*})$, $Z_0^*=0$, $Z_k^*=\sum_{i=1}^{k}X_i^*-\sum_{i=1}^{k}\mathbb{E}[X_i^*|\mathcal{F}_{i-1}]$, $k=1,2,\cdots,K$. Notice that $\{Z_k^*\}_{k=1}^K$ is a martingale, and $|Z_k^*-Z_{k-1}^*|=|X_k^*-\mathbb{E}[X_k^*|\mathcal{F}_{k-1}]|\le 2H$, $\forall k\in[K]$.

Then by Azuma-Hoeffding's inequality: For any $\epsilon>0$,

$$\mathbb{P}\left(|Z_K^*-Z_0^*|\ge\epsilon\right)\le 2\exp\left\{\frac{-\epsilon^2}{2K\cdot 4H^2}\right\}$$

which means that, with probability at least $1-\delta$,

$$|\sum_{k=1}^{K}X_k^*-\sum_{k=1}^{K}\mathbb{E}[X_k^*|\mathcal{F}_{k-1}]|\le\sqrt{8KH^2\log(\frac{2}{\delta})}$$

For the term $\sum_{k=1}^{K}\mathbb{E}[X_k^*|\mathcal{F}_{k-1}]$, notice that

$$\mathbb{E}[X_k^*|\mathcal{F}_{k-1}]=\sum_{h=1}^{H}\mathbb{E}[\xi_h(s_h^{k*},a_h^{k*})|\mathcal{F}_{k-1}]=\sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim d_h^{\pi^*}}[\xi_h(s,a)]\le H\zeta\quad\text{(By Assumption 5)}$$

Therefore,

$$\sum_{k=1}^{K}\mathbb{E}[X_k^*|\mathcal{F}_{k-1}]\le KH\zeta$$

Then, with probability at least $1-\delta$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\xi_h(s_h^{k*},a_h^{k*})\le\sqrt{8KH^2\log(\frac{2}{\delta})}+KH\zeta$$

and similarly, with probability at least $1-\delta$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\eta_h(s_h^{k*},a_h^{k*})\le\sqrt{32dKH^2\log(\frac{2}{\delta})}+KH\zeta$$

For the last term $\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_h^{k*}$, notice that $\{\zeta_h^{k*}\}$ is a martingale difference sequence, and each term is upper bounded by $2H$. By using Azuma-Hoeffding's inequality, with probability at least $1-\delta/4$, the following inequality holds:

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_h^{k*}\le\sqrt{8KH^3\cdot\log(8/\delta)}\tag{35}$$

Finally, with probability at least $1-\delta$,

$$\begin{aligned}\sum_{k=1}^{K}[V_1^*(s_1^k)-V_1^k(s_1^k)] &\le 3H\sum_{k=1}^{K}\sum_{h=1}^{H}\xi_h(s_h^{k*},a_h^{k*})+\sum_{k=1}^{K}\sum_{h=1}^{H}\eta_h(s_h^{k*},a_h^{k*})+\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_h^{k*}\\ &\le 4KH^2\zeta+12H^2\sqrt{dK\log(\frac{8}{\delta})}\end{aligned} \tag{36}$$

□

Theorem B.11. ***(Regret Bound under ζ-Average-Approximate Linear MDP).***

Under our Assumption 5, 6 , for any fixed $\delta\in(0,1)$, with probability at least $1-\delta$, the total regret of the algorithm Robust-LSVI (Algorithm 6) is at most $\widetilde{O}\left(dKH^2\zeta+\sqrt{d^3KH^4}\right)$.

Proof. First of all, we do the following decomposition.

$$\text{Regret}(K)=\sum_{k=1}^{K}[V_1^*(s_1^k)-V_1^{\pi_k}(s_1^k)]=\underbrace{\sum_{k=1}^{K}[V_1^*(s_1^k)-V_1^k(s_1^k)]}_{A}+\underbrace{\sum_{k=1}^{K}[V_1^k(s_1^k)-V_1^{\pi_k}(s_1^k)]}_{B} \tag{37}$$

For the term A, from Lemma B.10 (Near-Optimism), we have with probability at least $1-\frac{\delta}{2}$,

$$\sum_{k=1}^{K}[V_1^*(s_1^k)-V_1^k(s_1^k)]\le 4KH^2\zeta+12H^2\sqrt{dK\log(\frac{16}{\delta})} \tag{38}$$

For the term B, from Lemma B.7 (Recursive formula), we have

$$\begin{aligned}&\sum_{k=1}^{K}[V_1^k(s_1^k)-V_1^{\pi_k}(s_1^k)]=\sum_{k=1}^{K}\delta_1^k\\ &\le\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_h^k+\sum_{k=1}^{K}\beta_k\sum_{h=1}^{H}\sqrt{(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k}+\sum_{k=1}^{K}\sum_{h=1}^{H}\lambda_h^k\sqrt{(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k}\\ &+3H\sum_{k=1}^{K}\sum_{h=1}^{H}\xi_h(s_h^k,a_h^k)+\sum_{k=1}^{K}\sum_{h=1}^{H}\eta_h(s_h^k,a_h^k)\end{aligned} \tag{39}$$

Notice that $\{\zeta_h^k\}$ is a martingale difference sequence, and each term is upper bounded by $2H$. By using Azuma-Hoeffding's inequality, with probability at least $1-\delta/4$, the following inequality holds:

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_h^k\le\sqrt{8KH^3\cdot\log(8/\delta)} \tag{40}$$

By Lemma B.9, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\lambda_h^k\sqrt{(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k}\le\sum_{k=1}^{K}\beta_k\sum_{h=1}^{H}\sqrt{(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k} \tag{41}$$

and by using Cauchy-Schwarz inequality, we have

$$\sum_{k=1}^{K}\beta_k\sqrt{(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k}\le\left[\sum_{k=1}^{K}\beta_k^2\right]^{\frac{1}{2}}\cdot\left[\sum_{k=1}^{K}(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k\right]^{\frac{1}{2}} \tag{42}$$

and

$$\begin{aligned}\sum_{k=1}^{K}\beta_k^2 &= \sum_{k=1}^{K}\left(c_\beta^2\left(4\sqrt{kd}\zeta+\sqrt{(\lambda+1)d^2\log\left(\frac{4dKH}{\delta}\right)}\right)^2 H^2\right)\\ &\le 2\sum_{k=1}^{K}c_\beta^2H^2\left(16kd\zeta^2+(\lambda+1)d^2\log\left(\frac{4dKH}{\delta}\right)\right)\\ &\le 32c_\beta^2K^2H^2d\zeta^2+2c_\beta^2(\lambda+1)KH^2d^2\log\left(\frac{4dKH}{\delta}\right)\end{aligned} \tag{43}$$

Therefore,

$$\left[\sum_{k=1}^{K}\beta_k^2\right]^{\frac{1}{2}} \le 8c_\beta\sqrt{d}KH\zeta+2c_\beta(\lambda+1)Hd\sqrt{K\log\left(\frac{4dKH}{\delta}\right)} \tag{44}$$

By Lemma G.2, we have for any $h\in[H]$,

$$\sum_{k=1}^{K}(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k \le 2\log\left[\frac{\det(\Lambda_h^{k+1})}{\det(\Lambda_h^1)}\right] \le 2d\log\left[\frac{\lambda+k}{\lambda}\right] \le 2d\log\left(\frac{2dKH}{\delta}\right)$$

thus

$$\left[\sum_{k=1}^{K}(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k\right]^{\frac{1}{2}} \le \sqrt{2d\log\left(\frac{2dKH}{\delta}\right)} \tag{45}$$

By combining (42), (44), and (45), we have

$$\sum_{k=1}^{K}\beta_k\sum_{h=1}^{H}\sqrt{(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k} \le 16c_\beta dKH^2\zeta\sqrt{\log\left(\frac{2dKH}{\delta}\right)}+4c_\beta(\lambda+1)\sqrt{Kd^3H^4}\cdot\log\left(\frac{4dKH}{\delta}\right) \tag{46}$$

Using Lemma B.8 (Bound of cumulative misspecification error), (40), (41), and (46), we can give a bound of (39), that with probability at least $1-\delta/2$,

$$\sum_{k=1}^{K}[V_1^k(s_1^k)-V_1^{\pi_k}(s_1^k)] \le 36c_\beta dKH^2\zeta\sqrt{\log\left(\frac{2dKH}{\delta}\right)}+20c_\beta(\lambda+1)\sqrt{Kd^3H^4}\cdot\log\left(\frac{4dKH}{\delta}\right) \tag{47}$$

Finally, by combining (38) and (47), we get the regret bound:

$$\mathrm{Regret}(K) \le 40c_\beta dKH^2\zeta\sqrt{\log\left(\frac{2dKH}{\delta}\right)}+32c_\beta(\lambda+1)\sqrt{Kd^3H^4}\cdot\log\left(\frac{4dKH}{\delta}\right) \tag{48}$$

This indicates that our regret bound is $\widetilde{O}(dKH^2\zeta+\sqrt{d^3KH^4})$, which finishes our proof.

□

C Analysis of Robust Value-based Algorithm with General Function Approximation for Locally-bounded Misspecified MDP

Assumption 7. *(General function approximation with locally-bounded misspecification error)*

Given the MDP M with the transition model P, we assume that there exists a function class $\mathcal{F}\subset\{f:\mathcal{S}\times\mathcal{A}\to[0,H]\}$ and a real number $\zeta\in[0,1]$, such that for any $V:\mathcal{S}\to[0,H]$, there exists $\bar{f}_V\in\mathcal{F}$, which satisfies : $\forall\beta\in[4]$,

$$\sup_\pi \mathbb{E}_{(s,a)\sim d^\pi}\left|\bar{f}_V(s,a)-(r(s,a)+\mathbb{P}V(s,a))\right|^\beta \le \zeta^\beta$$

To simplify the notation, for a fixed $k \in [K]$, we let $\mathcal{Z}^k = \{(s_h^{k'}, a_h^{k'})\}_{(k',h)\in[k-1]\times[H]}$.

For any $V : \mathcal{S} \to [0, H]$, define

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

and the accordingly minimizer

$$\widehat{f}_V := \arg\min_{f\in\mathcal{F}} ||f||^2_{\mathcal{D}_V^k}$$

Lemma C.1. *For any fixed $f \in \mathcal{F}$, and fixed $V : \mathcal{S} \to [0, H]$, with probability at least $1 - \delta$, we have for all $k \in [K]$ that*

$$||\bar{f}_V||^2_{\mathcal{D}_V^k} - ||f||^2_{\mathcal{D}_V^k} - \frac{2}{3}H^2\log(\frac{1}{\delta}) - ||f - \bar{f}_V||_{\mathcal{Z}^k} \cdot \sqrt{kH\zeta^2 + C'H \cdot \log(\frac{4KH}{\delta})} + \frac{1}{4}||f - \bar{f}_V||^2_{\mathcal{Z}^k} \le 0 \tag{49}$$

Proof. For each $(k, h) \in [K] \times [H]$, and a fixed $f \in \mathcal{F}$, we define

$$\mathcal{Z}_h^k = \left(\bar{f}_V(s_h^k, a_h^k) - r(s_h^k, a_h^k) - V(s_{h+1}^k)\right)^2 - \left(f(s_h^k, a_h^k) - r(s_h^k, a_h^k) - V(s_{h+1}^k)\right)^2$$

$$\epsilon_h^k = V(s_{h+1}^k) - \mathbb{P}V(s_h^k, a_h^k), \quad \xi_h^k = r(s_h^k, a_h^k) + \mathbb{P}V(s_h^k, a_h^k) - \bar{f}_V(s_h^k, a_h^k)$$

and set $\mathbb{F}_h^k$ be the σ-algebra generated by $\{(s_{h'}^{k'}, a_{h'}^{k'})\}_{(h',k')\in[H]\times[k-1]} \cup \{(s_1^k, a_1^k), (s_2^k, a_2^k), \cdots, (s_h^k, a_h^k)\}$.

Actually, $\sum_{k'=1}^{k-1}\sum_{h=1}^{H} \mathcal{Z}_h^{k'} = ||\bar{f}_V||^2_{\mathcal{D}_V^k} - ||f||^2_{\mathcal{D}_V^k}$, and after a simple calculation, we have

$$\mathcal{Z}_h^k = -\left(f(s_h^k, a_h^k) - \bar{f}_V(s_h^k, a_h^k)\right)^2 + 2\left(f(s_h^k, a_h^k) - \bar{f}_V(s_h^k, a_h^k)\right)\left(\underbrace{r(s_h^k, a_h^k) + V(s_{h+1}^k) - \bar{f}_V(s_h^k, a_h^k)}_{\epsilon_h^k + \xi_h^k}\right)$$

Notice that $\mathbb{E}[\epsilon_h^k|\mathbb{F}_h^k] = 0$, and since ϵ_h^k is bounded in $[-H, H]$, hence, ϵ_h^k is H-subguassian. That is to say, for any $\lambda \in \mathbb{R}$, we have $\mathbb{E}[\exp\{\lambda\epsilon_h^k\}|\mathbb{F}_h^k] \le \exp\{\frac{\lambda^2H^2}{8}\}$.

Moreover, under Assumption 2, using the same argument in Lemma B.9, with probability at least $1 - \delta/2$, for all $(k, h) \in [K] \times [H]$, we have

$$\sum_{k'=1}^{k}\sum_{h=1}^{H}(\xi_h^{k'})^2 \le kH\zeta^2 + C'H \cdot \log(\frac{8KH}{\delta}) \tag{50}$$

where C' is some constant.

Therefore, the conditional mean and the conditional cumulant generating function of the centered random variable can be calculated.

$$\mu_h^k = \mathbb{E}[\mathcal{Z}_h^k|\mathbb{F}_h^k] = -\left(f(s_h^k, a_h^k) - \bar{f}_V(s_h^k, a_h^k)\right)^2 + 2\left(f(s_h^k, a_h^k) - \bar{f}_V(s_h^k, a_h^k)\right)\xi_h^k$$

and

$$\begin{aligned}\phi_h^k(\lambda) &= \log\mathbb{E}\left[\exp\{\lambda(\mathcal{Z}_h^k - \mu_h^k)\}|\mathbb{F}_h^k\right] \\ &= \log\mathbb{E}\left[\exp\{2\lambda\left(f(s_h^k, a_h^k) - \bar{f}_V(s_h^k, a_h^k)\right)\epsilon_h^k\}|\mathbb{F}_h^k\right] \\ &\le \frac{\lambda^2\left(f(s_h^k, a_h^k) - \bar{f}_V(s_h^k, a_h^k)\right)^2H^2}{2}\end{aligned} \tag{51}$$

By using Lemma G.7, we have for any $x \geq 0$, $\lambda \geq 0$,

$$\begin{aligned}\mathbb{P}\Big(\lambda \sum_{k'=1}^{k-1}\sum_{h=1}^{H} \mathcal{Z}_h^{k'} \leq x - \lambda \sum_{k'=1}^{k-1}\sum_{h=1}^{H}\Big(f(s_h^{k'},a_h^{k'}) - \bar{f}_V(s_h^{k'},a_h^{k'})\Big)^2 \\ + 2\lambda \sum_{k'=1}^{k-1}\sum_{h=1}^{H}\Big(f(s_h^{k'},a_h^{k'}) - \bar{f}_V(s_h^{k'},a_h^{k'})\Big)\xi_h^{k'} \\ + \frac{\lambda^2 H^2}{2}\sum_{k'=1}^{k-1}\sum_{h=1}^{H}\Big(f(s_h^{k'},a_h^{k'}) - \bar{f}_V(s_h^{k'},a_h^{k'})\Big)^2, \forall (k,h) \in [K]\times[H]\Big) \geq 1 - e^{-x}\end{aligned} \tag{52}$$

After setting $x = \log(\frac{2}{\delta})$, $\lambda = \frac{3}{2H^2}$, and conditioned the event that Eq.(50) holds, that is to say,

$$\begin{aligned}&\sum_{k'=1}^{k-1}\sum_{h=1}^{H}\Big(f(s_h^{k'},a_h^{k'}) - \bar{f}_V(s_h^{k'},a_h^{k'})\Big)\xi_h^{k'} \\ &\leq \sqrt{\sum_{k'=1}^{k-1}\sum_{h=1}^{H}\big(f(s_h^{k'},a_h^{k'}) - \bar{f}_V(s_h^{k'},a_h^{k'})\big)^2}\cdot\sqrt{\sum_{k'=1}^{k-1}\sum_{h=1}^{H}(\xi_h^{k'})^2} \\ &\leq \sqrt{\sum_{k'=1}^{k-1}\sum_{h=1}^{H}\big(f(s_h^{k'},a_h^{k'}) - \bar{f}_V(s_h^{k'},a_h^{k'})\big)^2}\cdot\sqrt{kH\zeta^2 + C'H\cdot\log(\frac{8KH}{\delta})}\end{aligned} \tag{53}$$

Then we can derive the following result by using Eq.(52) :

$$\begin{aligned}&\mathbb{P}\Big(||\bar{f}_V||^2_{\mathcal{D}_V^k} - ||f||^2_{\mathcal{D}_V^k} - \frac{2}{3}H^2\log(\frac{2}{\delta}) - ||f-\bar{f}_V||_{\mathcal{Z}^k}\cdot\sqrt{kH\zeta^2 + C'H\cdot\log(\frac{8KH}{\delta})} \\ &+ \frac{1}{4}||f-\bar{f}_V||^2_{\mathcal{Z}^k} \leq 0, \forall k \in [K]\Big) \geq 1-\delta\end{aligned} \tag{54}$$

□

Lemma C.2. *(Discretization error) If $g \in \mathcal{C}(\mathcal{F}, 1/T)$ satisfies $||f-g||_\infty \leq 1/T$, then*

$$\begin{aligned}&\Big|\frac{1}{4}||g-\bar{f}_V||^2_{\mathcal{Z}^k} - \frac{1}{4}||f-\bar{f}_V||^2_{\mathcal{Z}^k} + ||f-\bar{f}_V||_{\mathcal{Z}^k}\cdot\sqrt{kH\zeta^2 + C'H\cdot\log(\frac{8KH\mathcal{N}(\mathcal{F},1/T)}{\delta})} \\ &- ||g-\bar{f}_V||_{\mathcal{Z}^k}\cdot\sqrt{kH\zeta^2 + C'H\cdot\log(\frac{8KH\mathcal{N}(\mathcal{F},1/T)}{\delta})} + ||f||^2_{\mathcal{D}_V^k} - ||g||^2_{\mathcal{D}_V^k}\Big| \\ &\leq 5(H+1) + 2\sqrt{kH^2\zeta^2 + C'H^2\cdot\log(\frac{8KH\mathcal{N}(\mathcal{F},1/T)}{\delta})}\end{aligned} \tag{55}$$

Proof. For any $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}&\Big|\big(g(s,a) - \bar{f}_V(s,a)\big)^2 - \big(f(s,a) - \bar{f}_V(s,a)\big)^2\Big| \\ &\leq \big|\big[g(s,a) + f(s,a) - 2\bar{f}_V(s,a)\big]\cdot\big[g(s,a) - f(s,a)\big]\big| \\ &\leq 4H\cdot\frac{1}{T}\end{aligned} \tag{56}$$

Therefore,

$$\big|||f-\bar{f}_V||^2_{\mathcal{Z}^k} - ||g-\bar{f}||^2_{\mathcal{Z}^k}\big| \leq 4H\cdot\frac{1}{T}\cdot|\mathcal{Z}^k| = 4H \tag{57}$$

$$
\begin{aligned}
&\left|||f||^2_{\mathcal{D}_V^k} - ||g||^2_{\mathcal{D}_V^k}\right| \\
&= \left|\sum_{k'=1}^{k-1}\sum_{h=1}^{H}\left(f(s_h^{k'},a_h^{k'}) - r(s_h^{k'},a_h^{k'}) - V(s_{h+1}^{k'})\right)^2 - \sum_{k'=1}^{k-1}\sum_{h=1}^{H}\left(g(s_h^{k'},a_h^{k'}) - r(s_h^{k'},a_h^{k'}) - V(s_{h+1}^{k'})\right)^2\right| \\
&\le \left|\sum_{k'=1}^{k-1}\sum_{h=1}^{H}\left(f(s_h^{k'},a_h^{k'}) + g(s_h^{k'},a_h^{k'}) - 2r(s_h^{k'},a_h^{k'}) - 2V(s_{h+1}^{k'})\right)\left(f(s_h^{k'},a_h^{k'}) - g(s_h^{k'},a_h^{k'})\right)\right| \\
&\le 4(H+1)\cdot|\mathcal{Z}^k|\cdot\frac{1}{T} = 4(H+1)
\end{aligned}
\tag{58}
$$

Moreover,

$$
\left|||f-\bar{f}_V||_{\mathcal{Z}^k} - ||g-\bar{f}_V||_{\mathcal{Z}^k}\right| \le \sqrt{\left|||f-\bar{f}_V||^2_{\mathcal{Z}^k} - ||g-\bar{f}||^2_{\mathcal{Z}^k}\right|} \le 2\sqrt{H} \tag{59}
$$

By combining (57), (58) and (59), we have (55).

□

Lemma C.3. *For a fixed $V : \mathcal{S} \to [0, H]$, with probability at least $1-\delta$, for all $k \in [K]$ and all $f \in \mathcal{F}$, that*

$$
\begin{aligned}
||f||^2_{\mathcal{D}_V^k} - ||\bar{f}_V||^2_{\mathcal{D}_V^k} \ge \frac{1}{4}||f-\bar{f}_V||^2_{\mathcal{Z}^k} - ||f-\bar{f}_V||_{\mathcal{Z}^k}\cdot\sqrt{kH\zeta^2 + C'H\cdot\log(\frac{8KH\mathcal{N}(\mathcal{F},1/T)}{\delta})} \\
- \frac{2}{3}H^2\log(\frac{2\mathcal{N}(\mathcal{F},1/T)}{\delta}) - 5(H+1) - 2\sqrt{kH^2\zeta^2 + C'H^2\cdot\log(\frac{8KH\mathcal{N}(\mathcal{F},1/T)}{\delta})}
\end{aligned}
\tag{60}
$$

We define the above event to be $\varepsilon_{V,\delta}$.

Proof. For any $f \in \mathcal{F}$, there exists a $g \in \mathcal{C}(\mathcal{F}, 1/T)$, such that $||f-g||_\infty \le \frac{1}{T}$. By taking a union bound on all $g \in \mathcal{C}(\mathcal{F}, 1/T)$ and using the result from Lemma C.1, we have, with probability at least $1-\delta$, for all $g \in \mathcal{C}(\mathcal{F}, 1/T)$, any $k \in [K]$, that

$$
\begin{aligned}
&||\bar{f}_V||^2_{\mathcal{D}_V^k} - ||g||^2_{\mathcal{D}_V^k} - \frac{2}{3}H^2\log(\frac{2\mathcal{N}(\mathcal{F},1/T)}{\delta}) - \\
&||g-\bar{f}_V||_{\mathcal{Z}^k}\cdot\sqrt{kH\zeta^2 + C'H\cdot\log(\frac{8KH\mathcal{N}(\mathcal{F},1/T)}{\delta})} + \frac{1}{4}||g-\bar{f}_V||^2_{\mathcal{Z}^k} \le 0
\end{aligned}
\tag{61}
$$

Therefore, with probability at least $1-\delta$, for all $k \in [K]$, and all $f \in \mathcal{F}$, we have:

$$\begin{aligned}
||f||^2_{\mathcal{D}_V^k} - ||\bar{f}_V||^2_{\mathcal{D}_V^k} \geq & \frac{1}{4}||f-\bar{f}_V||^2_{\mathcal{Z}^k} - \frac{2}{3}H^2\log(\frac{2\mathcal{N}(\mathcal{F},1/T)}{\delta}) \\
& - ||f-\bar{f}_V||_{\mathcal{Z}^k} \cdot \sqrt{kH\zeta^2 + C'H\cdot\log(\frac{8KH\mathcal{N}(\mathcal{F},1/T)}{\delta})} \\
& + \Big\{\frac{1}{4}||g-\bar{f}_V||^2_{\mathcal{Z}^k} - \frac{1}{4}||f-\bar{f}_V||^2_{\mathcal{Z}^k} \\
& + ||f-\bar{f}_V||_{\mathcal{Z}^k} \cdot \sqrt{kH\zeta^2 + C'H\cdot\log(\frac{8KH\mathcal{N}(\mathcal{F},1/T)}{\delta})} \\
& - ||g-\bar{f}_V||_{\mathcal{Z}^k} \cdot \sqrt{kH\zeta^2 + C'H\cdot\log(\frac{8KH\mathcal{N}(\mathcal{F},1/T)}{\delta})} \\
& + ||f||^2_{\mathcal{D}_V^k} - ||g||^2_{\mathcal{D}_V^k}\Big\} \\
\geq & \frac{1}{4}||f-\bar{f}_V||^2_{\mathcal{Z}^k} - \frac{2}{3}H^2\log(\frac{2\mathcal{N}(\mathcal{F},1/T)}{\delta}) \\
& - ||f-\bar{f}_V||_{\mathcal{Z}^k} \cdot \sqrt{kH\zeta^2 + C'H\cdot\log(\frac{8KH\mathcal{N}(\mathcal{F},1/T)}{\delta})} \\
& - 5(H+1) - 2\sqrt{kH^2\zeta^2 + C'H^2\cdot\log(\frac{8KH\mathcal{N}(\mathcal{F},1/T)}{\delta})} \quad \text{(By (55))}
\end{aligned} \tag{62}$$

□

Lemma C.4. *Conditioned on the event $\varepsilon_{V,\delta}$ (defined in Lemma C.3), then for any $V' : \mathcal{S} \to [0,H]$ with $||V'-V||_\infty \leq \frac{1}{T}$, we have for any $k \in [K]$,*

$$||\widehat{f}_{V'} - \bar{f}_V||_{\mathcal{Z}^k} \leq C' \cdot \sqrt{kH\zeta^2 + H^2\log(\frac{KH\mathcal{N}(\mathcal{F},1/T)}{\delta})} \tag{63}$$

Proof. For a fixed $V : \mathcal{S} \to [0,H]$, we consider any $V' : \mathcal{S} \to [0,H]$ with $||V'-V||_\infty \leq \frac{1}{T}$.
Then for any $f \in \mathcal{F}$,

$$\begin{aligned}
||f||^2_{\mathcal{D}_{V'}^k} - ||\bar{f}_V||^2_{\mathcal{D}_{V'}^k} &= ||f-\bar{f}_V||^2_{\mathcal{Z}^k} + 2\sum_{k'=1}^{k-1}\sum_{h=1}^{H}\Big(f(s_h^{k'},a_h^{k'}) - \bar{f}_V(s_h^{k'},a_h^{k'})\Big)\cdot\Big(\bar{f}_V(s_h^{k'},a_h^{k'}) - r_h^{k'} - V'(s_{h+1}^{k'})\Big) \\
&= ||f-\bar{f}_V||^2_{\mathcal{Z}^k} + \underbrace{2\sum_{k'=1}^{k-1}\sum_{h=1}^{H}\Big(f(s_h^{k'},a_h^{k'}) - \bar{f}_V(s_h^{k'},a_h^{k'})\Big)\cdot\Big(\bar{f}_V(s_h^{k'},a_h^{k'}) - r_h^{k'} - V(s_{h+1}^{k'})\Big)}_{\text{term A}} \\
&\quad \underbrace{- 2\sum_{k'=1}^{k-1}\sum_{h=1}^{H}\Big(f(s_h^{k'},a_h^{k'}) - \bar{f}_V(s_h^{k'},a_h^{k'})\Big)\Big(V'(s_{h+1}^{k'}) - V(s_{h+1}^{k'})\Big)}_{\text{term B}}
\end{aligned} \tag{64}$$

For the term A, by using (60), we have

$$\begin{aligned}
\text{term A} &= ||f||^2_{\mathcal{D}_V^k} - ||\bar{f}_V||^2_{\mathcal{D}_V^k} - ||f-\bar{f}_V||^2_{\mathcal{Z}^k} \\
&\geq \frac{1}{4}||f-\bar{f}_V||^2_{\mathcal{Z}^k} - ||f-\bar{f}_V||_{\mathcal{Z}^k}\cdot\sqrt{kH\zeta^2 + C'H\cdot\log(\frac{8KH\mathcal{N}(\mathcal{F},1/T)}{\delta})} \\
&\quad - \frac{2}{3}H^2\log(\frac{2\mathcal{N}(\mathcal{F},1/T)}{\delta}) - 5(H+1) - 2\sqrt{kH^2\zeta^2 + C'H^2\cdot\log(\frac{8KH\mathcal{N}(\mathcal{F},1/T)}{\delta})} - ||f-\bar{f}_V||^2_{\mathcal{Z}^k}
\end{aligned} \tag{65}$$

By using Cauchy-Schwarz's inequality, we can derive the upper bound for term B.

$$\text{term B} \le 2||f-\bar{f}_V||_{\mathcal{Z}^k} \cdot \sqrt{\sum_{k'=1}^{k-1}\sum_{h=1}^{H}\left(V'(s_{h+1}^{k'})-V(s_{h+1}^{k'})\right)^2} \le 2||f-\bar{f}_V||_{\mathcal{Z}^k} \tag{66}$$

Therefore, Eq.(64) can be bounded by

$$\begin{aligned}||f||^2_{\mathcal{D}^k_{V'}} - ||\bar{f}_V||^2_{\mathcal{D}^k_{V'}} \ge\ & ||f-\bar{f}_V||^2_{\mathcal{Z}^k} + \frac{1}{4}||f-\bar{f}_V||^2_{\mathcal{Z}^k} - ||f-\bar{f}_V||_{\mathcal{Z}^k}\cdot\sqrt{kH\zeta^2 + C'H\cdot\log(\frac{8KH\mathcal{N}(\mathcal{F},1/T)}{\delta})} \\ & - \frac{2}{3}H^2\log(\frac{2\mathcal{N}(\mathcal{F},1/T)}{\delta}) - 5(H+1) - 2\sqrt{kH^2\zeta^2 + C'H^2\cdot\log(\frac{8KH\mathcal{N}(\mathcal{F},1/T)}{\delta})} - ||f-\bar{f}_V||^2_{\mathcal{Z}^k} \\ & - 2||f-\bar{f}_V||_{\mathcal{Z}^k}\end{aligned} \tag{67}$$

Notice that

$$\widehat{f}_{V'} := \arg\min_{f\in\mathcal{F}} ||f||^2_{\mathcal{D}^k_{V'}}$$

, and by solving the quadratic equation for $||\widehat{f}_{V'}-\bar{f}_V||_{\mathcal{Z}^k}$, we have

$$||\widehat{f}_{V'}-\bar{f}_V||_{\mathcal{Z}^k} \le C'\cdot\sqrt{kH\zeta^2 + H^2\log(\frac{KH\mathcal{N}(\mathcal{F},1/T)}{\delta})} \tag{68}$$

where C' is an absolute constant.

□

Lemma C.5. *Let $\mathcal{F}_h^k$ be the confidence region defined as*

$$\mathcal{F}_h^k = \left\{f\in\mathcal{F} : ||f-f_h^k||_{\mathcal{Z}_h^k} \le \beta(\mathcal{F},\delta)\right\}$$

where

$$\beta(\mathcal{F},\delta) = C'\cdot\sqrt{kH\zeta^2 + H^2\left(\log(\frac{4T^2}{\delta}) + 2\log\mathcal{N}(\mathcal{F},1/T) + \log|\mathcal{W}| + 1\right)}$$

Then with probability at least $1-\frac{\delta}{4}$, we have for all $(k,h)\in[K]\times[H]$,

$$\bar{f}_{(V_{h+1}^k)^\dagger} \in \mathcal{F}_h^k \tag{69}$$

where $(V)^\dagger$ denotes the closest function to V in the set $\mathcal{V}$ (the $1/T$-net of $\{V_h^k\}$).

Proof. We denote

$$\mathcal{Q} := \{\min\{f(\cdot,\cdot)+\omega(\cdot,\cdot), H\} |\ \omega\in\mathcal{W}, f\in\mathcal{C}(\mathcal{F},1/T)\cup\{0\}\} \tag{70}$$

Notice that $\mathcal{Q}$ is a $(1/T)$-cover of $Q_{h+1}^k(\cdot,\cdot)$. This implies that

$$\mathcal{V} := \left\{\max_{a\in\mathcal{A}} q(\cdot,a) | q\in\mathcal{Q}\right\} \tag{71}$$

is also a $(1/T)$-cover of V_{h+1}^k, and we have $\log(|\mathcal{V}|) \le \log|\mathcal{W}| + \log\mathcal{N}(\mathcal{F},1/T) + 1$.

By taking the union bound for all $V\in\mathcal{V}$ in the event defined in Lemma C.3, we have $Pr(\bigcap_{V\in\mathcal{V}}\varepsilon_{V,\delta/(4|\mathcal{V}|T)}) \ge 1-\delta/(4T)$. We condition on $\bigcap_{V\in\mathcal{V}}\varepsilon_{V,\delta/(4|\mathcal{V}|T)})$ in the rest part of the proof.

Recall that f_h^k is the minimizer of the empirical loss, i.e., $f_h^k = \text{argmin}_{f\in\mathcal{F}}||f||^2_{\mathcal{D}_h^k}$. Let $(V_{h+1}^k)^\dagger\in\mathcal{V}$ such that $||V_{h+1}^k - (V_{h+1}^k)^\dagger||_\infty \le 1/T$. Then, by lemma C.4, we have

$$\begin{aligned}||f_h^k - \bar{f}_{(V_{h+1}^k)^\dagger}||_{\mathcal{Z}^k} &\le C'\cdot\sqrt{kH\zeta^2 + H^2\log(\frac{4\mathcal{N}(\mathcal{F},1/T)|\mathcal{V}|T^2}{\delta})} \\ &\le C'\cdot\sqrt{kH\zeta^2 + H^2\left(\log(\frac{4T^2}{\delta}) + 2\log\mathcal{N}(\mathcal{F},1/T) + \log|\mathcal{W}| + 1\right)}\end{aligned} \tag{72}$$

This completes the proof.

□

Lemma C.6. *(Proposition 2 in (Wang et al., 2020b)) With probability at least $1-\delta/8$, for all $(s,a)\in\mathcal{S}\times\mathcal{A}$,*

$$\omega(\mathcal{F}_h^k, s, a) \le b_h^k(s,a)$$

Lemma C.7. *(Lemma 10 in (Wang et al., 2020b)) With probability at least $1-\delta/8$,*

$$\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) \le 1 + 4H^2\, dim_E(\mathcal{F}, 1/T) + \sqrt{c\cdot dim_E(\mathcal{F},1/T)\cdot T}\cdot\beta(\mathcal{F},\delta) \tag{73}$$

Theorem C.8. *(Regret bound of robust value-based methods) Under our Assumption 1 and 2, for any fixed $\delta\in(0,1)$, with probability at least $1-\delta$, the total regret of Algorithm 3 is at most $\widetilde{O}\left(\sqrt{d_E H^3}K\zeta\log(1/\delta) + \sqrt{d_E^2 K H^3}\log(1/\delta)\right)$, where d_E represents the eluder dimension of the function class.*

Proof. First of all, we do the following decomposition.

$$\text{Regret}(K) = \sum_{k=1}^{K}[V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)] = \underbrace{\sum_{k=1}^{K}[V_1^*(s_1^k) - V_1^k(s_1^k)]}_{A} + \underbrace{\sum_{k=1}^{K}[V_1^k(s_1^k) - V_1^{\pi_k}(s_1^k)]}_{B} \tag{74}$$

We denote $\mathcal{E}$ to be the event that (69) holds, and $\mathcal{E}'$ to be the event that for all $(k,h)\in[K]\times[H]$, all $(s,a)\in\mathcal{S}\times\mathcal{A}$, $b_h^k(s,a)\ge\omega(\mathcal{F}_h^k,s,a)$. From Lemma C.5 and Lemma C.6, we have $Pr(\mathcal{E}\cap\mathcal{E}')\ge 1-\frac{\delta}{2}$. For the rest of the proof, we condition on the above event.

Note that

$$\max_{f\in\mathcal{F}_h^k}|f(s,a) - f_h^k(s,a)| \le \omega(\mathcal{F}_h^k,s,a) \le b_h^k(s,a)$$

Since $\bar{f}_{(V_{h+1}^k)^\dagger}\in\mathcal{F}_h^k$ for all $(k,h)\in[K]\times[H]$, we have for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, all $(k,h)\in[K]\times[H]$,

$$|\bar{f}_{(V_{h+1}^k)^\dagger}(s,a) - f_h^k(s,a)| \le \omega(\mathcal{F}_h^k,s,a) \le b_h^k(s,a) \tag{75}$$

To simplify our notation, for each $(k,h)\in[K]\times[H]$, we denote

$$\xi^\dagger_{V_{h+1}^k}(s,a) := \bar{f}_{(V_{h+1}^k)^\dagger}(s,a) - r(s,a) - \mathbb{P}V_{h+1}^k(s,a)$$

,

$$\zeta_{h+1}^k = \mathbb{P}(V_{h+1}^k - V_{h+1}^{\pi_k})(s_h^k,a_h^k) - \left(V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi_k}(s_{h+1}^k)\right)$$

and

$$\zeta_{h+1}^{k*} = \mathbb{P}(V_{h+1}^k - V_{h+1}^{\pi_k})(s_h^{k*},a_h^{k*}) - \left(V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi_k}(s_{h+1}^{k*})\right)$$

For the term $\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_{h+1}^k$ and $\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_{h+1}^{k*}$, notice that $\{\zeta_{h+1}^{k*}\}$ and $\{\zeta_{h+1}^k\}$ are martingale difference sequences, and each term is upper bounded by $2H$. By using Azuma-Hoeffding's inequality, with probability at least $1-\delta/4$, the following inequality holds:

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_{h+1}^k \le \sqrt{8KH^3\cdot\log(8/\delta)} \tag{76}$$

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_{h+1}^{k*} \le \sqrt{8KH^3\cdot\log(8/\delta)} \tag{77}$$

Notice that for all $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}|\xi^{\dagger}_{V^k_{h+1}}(s,a)| &= \left|\bar{f}_{(V^k_{h+1})^\dagger}(s,a) - r(s,a) - \mathbb{P}(V^k_{h+1})^\dagger(s,a) + \mathbb{P}(V^k_{h+1})^\dagger(s,a) - \mathbb{P}V^k_{h+1}(s,a)\right| \\ &\le \left|\xi_{(V^k_{h+1})^\dagger}(s,a)\right| + 1/T\end{aligned} \tag{78}$$

Therefore,

$$\left|\sum_{k=1}^{K}\sum_{h=1}^{H}\xi^{\dagger}_{V^k_{h+1}}(s^k_h,a^k_h)\right| \le \sum_{k=1}^{K}\sum_{h=1}^{H}\left|\xi_{(V^k_{h+1})^\dagger}(s^k_h,a^k_h)\right| + 1 \tag{79}$$

By using Assumption 2 and Azuma-Hoeffding's inequality, we have with probability at least $1-\delta/4$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left|\xi_{(V^k_{h+1})^\dagger}(s^k_h,a^k_h)\right| \le \sqrt{8KH^2\log(\frac{16}{\delta})} + KH\zeta \tag{80}$$

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left|\xi_{(V^k_{h+1})^\dagger}(s^{k*}_h,a^{k*}_h)\right| \le \sqrt{8KH^2\log(\frac{16}{\delta})} + KH\zeta \tag{81}$$

For the term A, we only need to consider when $V^k_h = f^k_h + b^k_h$, for all $(k,h)\in[K]\times[H]$.

$$\begin{aligned}A &= \sum_{k=1}^{K}\left(r(s^{k*}_1,a^{k*}_1) + \mathbb{P}V^*_2(s^{k*}_1,a^{k*}_1) - f^k_1(s^{k*}_1,a^{k*}_1) - b^k_1(s^{k*}_1,a^{k*}_1)\right) \\ &= \sum_{k=1}^{K}\Big(r(s^{k*}_1,a^{k*}_1) + \mathbb{P}V^k_2(s^{k*}_1,a^{k*}_1) - \bar{f}_{(V^k_2)^\dagger}(s^{k*}_1,a^{k*}_1) \\ &+ \bar{f}_{(V^k_2)^\dagger}(s^{k*}_1,a^{k*}_1) - f^k_1(s^{k*}_1,a^{k*}_1) - b^k_1(s^{k*}_1,a^{k*}_1) + \mathbb{P}(V^*_2 - V^k_2)(s^{k*}_1,a^{k*}_1)\Big) \\ &\le \sum_{k=1}^{K}\left(-\xi^{\dagger}_{V^k_2}(s^{k*}_1,a^{k*}_1) + \mathbb{P}(V^*_2 - V^k_2)(s^{k*}_1,a^{k*}_1)\right) \\ &\le \sum_{k=1}^{K}\left(-\xi^{\dagger}_{V^k_2}(s^{k*}_1,a^{k*}_1) + V^*_2(s^{k*}_2) - V^k_2(s^{k*}_2) + \zeta^{k*}_2\right) \le \cdots \\ &\le \sum_{k=1}^{K}\sum_{h=1}^{H}|\xi^{\dagger}_{V^k_{h+1}}(s^{k*}_h,a^{k*}_h)| + \sum_{k=1}^{K}\sum_{h=1}^{H}\zeta^{k*}_{h+1}\end{aligned} \tag{82}$$

For the term B,

$$\begin{aligned}
B &= \sum_{k=1}^{K} \left(Q_1^k(s_1^k, a_1^k) - Q_1^{\pi_k}(s_1^k, a_1^k)\right) \\
&\leq \sum_{k=1}^{K} \left(f_1^k(s_1^k, a_1^k) + b_1^k(s_1^k, a_1^k) - r(s_1^k, a_1^k) - \mathbb{P}V_2^{\pi_k}(s_1^k, a_1^k)\right) \\
&\leq \sum_{k=1}^{K} \left(\bar{f}_{(V_2^k)^\dagger}(s_1^k, a_1^k) + 2b_1^1(s_1^k, a_1^k) - r(s_1^k, a_1^k) - \mathbb{P}V_2^{\pi_k}(s_1^k, a_1^k)\right) \quad \text{(By Eq.(75))} \\
&= \sum_{k=1}^{K} \left(\bar{f}_{(V_2^k)^\dagger}(s_1^k, a_1^k) - r(s_1^k, a_1^k) - \mathbb{P}V_2^k(s_1^k, a_1^k) + \mathbb{P}V_2^k(s_1^k, a_1^k) - \mathbb{P}V_2^{\pi_k}(s_1^k, a_1^k) + 2b_1^1(s_1^k, a_1^k)\right) \\
&= \sum_{k=1}^{K} \left(\xi_{V_2^k}^\dagger(s_1^k, a_1^k) + 2b_1^1(s_1^k, a_1^k) + V_2^k(s_2^k) - V_2^{\pi_k}(s_2^k) + \zeta_2^k\right) \leq \cdots \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} |\xi_{V_{h+1}^k}^\dagger(s_h^k, a_h^k)| + 2\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) + \sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_{h+1}^k
\end{aligned} \tag{83}$$

By using (73), (76), (77), (79), (80), and (81), we complete our proof.

□

D Analysis of Robust Model-based Algorithm with General Function Approximation for Locally-bounded Misspecified MDP

In this section, we will provide the theoretical analysis for Algorithm 4 under locally-bounded misspecification error assumption (Assumption 4).

Confidence sets for non-linear regression with locally-bounded misspecification error

Let $\mathcal{V}$ be the set of optimal value functions under some model in $\mathcal{P}$: $\mathcal{V} = \{V_{P'}^* : P' \in \mathcal{P}\}$. We define $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{V}$, and choose

$$\mathcal{F} = \left\{f : \mathcal{X} \to \mathbb{R} : \exists \widetilde{P} \in \mathcal{P} \ \text{ s.t. } \ f(s, a, V) = \widetilde{\mathbb{P}}V(s, a), \ \ \forall (s, a, V) \in \mathcal{X}\right\} \tag{84}$$

Let $\phi : \mathcal{P} \to \mathcal{F}$ be the natural surjection to $\mathcal{F}$: $\phi(P) = f$, such that $f(s, a, V) = \mathbb{P}V(s, a), \ \ \forall (s, a, V) \in \mathcal{X}$. In fact, ϕ is a bijection, and for convenience to the reader, we denote $f_P = \phi(P)$.

For any $f \in \mathcal{F}$, we define the empirical loss as

$$L_{2,k}(f) = \sum_{k'=1}^{k}\sum_{h=1}^{H} \left(f(s_h^{k'}, a_h^{k'}, V_{h+1}^{k'}) - V_{h+1}^{k'}(s_{h+1}^{k'})\right)^2$$

and the minimizer $\widehat{f}_{k+1} = \operatorname{argmin}_{f \in \mathcal{F}} L_{2,k}(f)$. Since ϕ is a bijection, $\widehat{f}_{k+1} = \phi(\widehat{P}^{(k+1)}) = f_{\widehat{P}^{(k+1)}}$, where $\widehat{P}^{(k+1)}$ is defined in (4).

We also define the norm

$$||f||_{D_h^k} = \sqrt{\sum_{k'=1}^{k}\sum_{h'=1}^{h} \left(f(s_{h'}^{k'}, a_{h'}^{k'}, V_{h'+1}^{k'})\right)^2}$$

Now we are able to define the confidence set for each episode, which is also introduced in (6).

$$B_k = \{\widetilde{P} \in \mathcal{P} : L_k(\widetilde{P}, \widehat{P}^{(k)}) \leq \beta_k^2\} = \{\phi^{-1}(f) : f \in \mathcal{F} \ \text{ and } \ ||f - \widehat{f}_k||_{D_H^k} \leq \beta_k\}$$

We set $\mathbb{F}_h^k$ to be the σ-algebra generated by $\{(s_{h'}^{k'}, a_{h'}^{k'})\}_{(h',k')\in[H]\times[k-1]} \cup \{(s_1^k, a_1^k), (s_2^k, a_2^k), \cdots, (s_h^k, a_h^k)\}$. For each $(h,k) \in [H]\times[K]$, we define $\mathcal{Z}_h^k = \left(\bar{f}(s_h^k, a_h^k, V_{h+1}^k) - V_{h+1}^k(s_{h+1}^k)\right)^2 - \left(f(s_h^k, a_h^k, V_{h+1}^k) - V_{h+1}^k(s_{h+1}^k)\right)^2$

Lemma D.1. *Under Assumption 4, for any fixed $f \in \mathcal{F}$, with probability at least $1-\delta$, we have for all $k \in [K]$ that*

$$L_{2,k}(\bar{f}) - L_{2,k}(f) - H^2\log(\frac{1}{\delta}) - ||f-\bar{f}||_{D_H^k}\cdot\sqrt{kH\zeta^2 + C'H\cdot\log(\frac{4KH}{\delta})} + \frac{1}{2}||f-\bar{f}||_{D_H^k}^2 \le 0 \tag{85}$$

Proof. According to the definition of $\mathcal{Z}_h^k$, $\sum_{k'=1}^{k}\sum_{h=1}^{H}\mathcal{Z}_h^{k'} = L_{2,k}(\bar{f}) - L_{2,k}(f)$. After a simple calculation, we have

$$\mathcal{Z}_h^k = -\left(f(s_h^k, a_h^k, V_{h+1}^k) - \bar{f}(s_h^k, a_h^k, V_{h+1}^k)\right)^2 + 2\left(f(s_h^k, a_h^k, V_{h+1}^k) - \bar{f}(s_h^k, a_h^k, V_{h+1}^k)\right)\left(\underbrace{V_{h+1}^k(s_{h+1}^k) - \bar{f}(s_h^k, a_h^k, V_{h+1}^k)}_{\epsilon_h^k + \xi_h^k}\right)$$

where $\epsilon_h^k = V_{h+1}^k(s_{h+1}^k) - \mathbb{P}_h V_{h+1}^k(s_h^k, a_h^k)$, $\xi_h^k = \mathbb{P}_h V_{h+1}^k(s_h^k, a_h^k) - \bar{f}(s_h^k, a_h^k, V_{h+1}^k)$.

Notice that $\mathbb{E}[\epsilon_h^k|\mathbb{F}_h^k] = 0$, and since ϵ_h^k is bounded in $[0,H]$, hence, ϵ_h^k is $\frac{H}{2}$-subgaussian. That is to say, for any $\lambda\in\mathbb{R}$, we have $\mathbb{E}[\exp\{\lambda\epsilon_h^k\}|\mathbb{F}_h^k] \le \exp\{\frac{\lambda^2H^2}{8}\}$.

Moreover, under Assumption 4, using the same argument in Lemma B.9, with probability at least $1-\delta$, for all $(k,h)\in[K]\times[H]$, we have

$$\sum_{k'=1}^{k}\sum_{h=1}^{H}(\xi_h^{k'})^2 \le kH\zeta^2 + C'H\cdot\log(\frac{4KH}{\delta}) \tag{86}$$

where C' is some constant.

Therefore,

$$\mu_h^k = \mathbb{E}[\mathcal{Z}_h^k|\mathbb{F}_h^k] = -\left(f(s_h^k, a_h^k, V_{h+1}^k) - \bar{f}(s_h^k, a_h^k, V_{h+1}^k)\right)^2 + 2\left(f(s_h^k, a_h^k, V_{h+1}^k) - \bar{f}(s_h^k, a_h^k, V_{h+1}^k)\right)\xi_h^k$$

and

$$\begin{aligned}\phi_h^k(\lambda) &= \log\mathbb{E}\left[\exp\{\lambda(\mathcal{Z}_h^k - \mu_h^k)\}|\mathbb{F}_h^k\right]\\ &= \log\mathbb{E}\left[\exp\{2\lambda\left(f(s_h^k, a_h^k, V_{h+1}^k) - \bar{f}(s_h^k, a_h^k, V_{h+1}^k)\right)\epsilon_h^k\}|\mathbb{F}_h^k\right]\\ &\le \frac{\lambda^2\left(f(s_h^k, a_h^k, V_{h+1}^k) - \bar{f}(s_h^k, a_h^k, V_{h+1}^k)\right)^2H^2}{2}\end{aligned} \tag{87}$$

By using Lemma G.7, we have for any $x\ge 0$, $\lambda\ge 0$,

$$\begin{aligned}\mathbb{P}\Big(\lambda\sum_{k'=1}^{k}\sum_{h=1}^{H}\mathcal{Z}_h^{k'} \le x &- \lambda\sum_{k'=1}^{k}\sum_{h=1}^{H}\left(f(s_h^{k'}, a_h^{k'}, V_{h+1}^{k'}) - \bar{f}(s_h^{k'}, a_h^{k'}, V_{h+1}^{k'})\right)^2\\ &+ 2\lambda\sum_{k'=1}^{k}\sum_{h=1}^{H}\left(f(s_h^{k'}, a_h^{k'}, V_{h+1}^{k'}) - \bar{f}(s_h^{k'}, a_h^{k'}, V_{h+1}^{k'})\right)\xi_h^{k'}\\ &+ \frac{\lambda^2H^2}{2}\sum_{k'=1}^{k}\sum_{h=1}^{H}\left(f(s_h^{k'}, a_h^{k'}, V_{h+1}^{k'}) - \bar{f}(s_h^{k'}, a_h^{k'}, V_{h+1}^{k'})\right)^2, \forall(k,h)\in[K]\times[H]\Big) \ge 1 - e^{-x}\end{aligned} \tag{88}$$

After setting $x = \log(\frac{1}{\delta})$, $\lambda = \frac{1}{H^2}$, and conditioned the event that Eq.(86) holds, that is to say,

$$\begin{aligned}
&\sum_{k'=1}^{k}\sum_{h=1}^{H}\Big(f(s_h^{k'},a_h^{k'},V_{h+1}^{k'}) - \bar{f}(s_h^{k'},a_h^{k'},V_{h+1}^{k'})\Big)\,\xi_h^{k'} \\
&\leq \sqrt{\sum_{k'=1}^{k}\sum_{h=1}^{H}\big(f(s_h^{k'},a_h^{k'},V_{h+1}^{k'}) - \bar{f}(s_h^{k'},a_h^{k'},V_{h+1}^{k'})\big)^2}\cdot\sqrt{\sum_{k'=1}^{k}\sum_{h=1}^{H}(\xi_h^{k'})^2} \\
&\leq \sqrt{\sum_{k'=1}^{k}\sum_{h=1}^{H}\big(f(s_h^{k'},a_h^{k'},V_{h+1}^{k'}) - \bar{f}(s_h^{k'},a_h^{k'},V_{h+1}^{k'})\big)^2}\cdot\sqrt{kH\zeta^2 + C'H\cdot\log(\frac{4KH}{\delta})}
\end{aligned} \tag{89}$$

Then we can derive the following result by using Eq.(88) :

$$\begin{aligned}
&\mathbb{P}\Big(L_{2,k}(\bar{f}) - L_{2,k}(f) - H^2\log(\frac{1}{\delta}) - \sqrt{\sum_{k'=1}^{k}\sum_{h=1}^{H}\big(f(s_h^{k'},a_h^{k'},V_{h+1}^{k'}) - \bar{f}(s_h^{k'},a_h^{k'},V_{h+1}^{k'})\big)^2}\cdot\sqrt{kH\zeta^2 + C'H\cdot\log(\frac{4KH}{\delta})} \\
&+\frac{1}{2}\sum_{k'=1}^{k}\sum_{h=1}^{H}\Big(f(s_h^{k'},a_h^{k'},V_{h+1}^{k'}) - \bar{f}(s_h^{k'},a_h^{k'},V_{h+1}^{k'})\Big)^2 \leq 0, \forall k\in[K]\Big) \geq 1-\delta
\end{aligned} \tag{90}$$

which finishes the proof.

□

Lemma D.2. *(Discretization error)*
We denote $\mathcal{F}^\alpha$ as the α-cover of function class $\mathcal{F}$. If $f^\alpha \in \mathcal{F}^\alpha$ satisfies $||f - f^\alpha||_\infty \leq \alpha$, then

$$\begin{aligned}
&\Big|\frac{1}{2}||f^\alpha - \bar{f}||^2_{D_H^k} - \frac{1}{2}||f-\bar{f}||^2_{D_H^k} + ||f-\bar{f}||_{D_H^k}\cdot\sqrt{kH\zeta^2 + C'H\cdot\log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})} \\
&- ||f^\alpha - \bar{f}||_{D_H^k}\cdot\sqrt{kH\zeta^2 + C'H\cdot\log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})} + L_{2,k}(f) - L_{2,k}(f^\alpha)\Big| \\
&\leq 4\alpha kH^2 + \sqrt{4\alpha kH^2}\cdot\sqrt{kH\zeta^2 + C'H\cdot\log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})} \qquad ,\forall(k,h)\in[K]\times[H]
\end{aligned} \tag{91}$$

Proof. If $f^\alpha \in \mathcal{F}^\alpha$ satisfies $||f - f^\alpha||_\infty \leq \alpha$, then for any $(s,a,V)\in\mathcal{S}\times\mathcal{A}\times\mathcal{V}$, we have

$$\big|(f^\alpha)^2(s,a,V) - (f)^2(s,a,V)\big| \leq 2\alpha H \tag{92}$$

This implies that

$$\begin{aligned}
&\Big|\big(f^\alpha(s,a,V) - \bar{f}(s,a,V)\big)^2 - \big(f(s,a,V) - \bar{f}(s,a,V)\big)^2\Big| \\
&= \big|[(f^\alpha)(s,a,V)^2 - f(s,a,V)^2] + 2\bar{f}(s,a,V)\,(f(s,a,V) - f^\alpha(s,a,V))\big| \\
&\leq 2\alpha H + 2\alpha H = 4\alpha H
\end{aligned} \tag{93}$$

and for any $(k,h)\in[K]\times[H]$,

$$\begin{aligned}
&\big|(V_{h+1}^k(s_{h+1}^k) - f(s,a,V))^2 - (V_{h+1}^k(s_{h+1}^k) - f^\alpha(s,a,V))^2\big| \\
&= \big|2V_{h+1}^k(s_{h+1}^k)\,(f^\alpha(s,a,V) - f(s,a,V)) + f(s,a,V)^2 - f^\alpha(s,a,V)^2\big| \\
&\leq 2\alpha H + 2\alpha H = 4\alpha H
\end{aligned} \tag{94}$$

Moreover,

$$\left|||f-\bar{f}||_{D_H^k}-||f^\alpha-\bar{f}||_{D_H^k}\right| \le \sqrt{\left|||f-\bar{f}||^2_{D_H^k}-||f^\alpha-\bar{f}||^2_{D_H^k}\right|} \tag{95}$$

By taking the sum over k and H, we can find that the left hand side of (91) is bounded by

$$4\alpha kH^2+\sqrt{4\alpha kH^2}\cdot\sqrt{kH\zeta^2+C'H\cdot\log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})}$$

□

Lemma D.3. *With probability at least $1-\delta$, for all $k\in[K]$, we have*

$$||\widehat{f}_{k+1}-\bar{f}||_{D_H^k}\le\beta_k \tag{96}$$

where

$$\beta_k=3\sqrt{kH}\zeta+5\sqrt{C'H^2\cdot\log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})}+4\sqrt{\alpha kH^2} \tag{97}$$

Remark D.4. *Since the mapping $\phi:\mathcal{P}\to\mathcal{F}$ is a bijection, $\bar{P}=\phi^{-1}(\bar{f})$ satisfies*

$$\mathbb{P}\left(\bar{P}\in\bigcap_{k\in[K]}B_k\right)\ge 1-\delta \tag{98}$$

Proof. Let $\mathcal{F}^\alpha\subset\mathcal{F}$ be an α-cover of $\mathcal{F}$ in the sup-norm. In other words, for any $f\in\mathcal{F}$, there is an $f^\alpha\in\mathcal{F}^\alpha$, such that $||f^\alpha-f||_\infty\le\alpha$. By a union bound and from Lemma D.1, with probability at least $1-\delta$, we have for any $f^\alpha\in\mathcal{F}^\alpha$, any $k\in[K]$, that

$$L_{2,k}(f^\alpha)-L_{2,k}(\bar{f})\ge -H^2\log(\frac{|\mathcal{F}^\alpha|}{\delta})-||f^\alpha-\bar{f}||_{D_H^k}\cdot\sqrt{kH\zeta^2+C'H\cdot\log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})}+\frac{1}{2}||f^\alpha-\bar{f}||^2_{D_H^k} \tag{99}$$

Therefore, with probability at least $1-\delta$, for all $k\in[K]$, and all $f\in\mathcal{F}$, we have:

$$\begin{aligned}L_{2,k}(f)-L_{2,k}(\bar{f})\ge{}&\frac{1}{2}||f-\bar{f}||^2_{D_H^k}-H^2\log(\frac{|\mathcal{F}^\alpha|}{\delta})-||f-\bar{f}||_{D_H^k}\cdot\sqrt{kH\zeta^2+C'H\cdot\log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})}\\&+\Big\{\frac{1}{2}||f^\alpha-\bar{f}||^2_{D_H^k}-\frac{1}{2}||f-\bar{f}||^2_{D_H^k}\\&+||f-\bar{f}||_{D_H^k}\cdot\sqrt{kH\zeta^2+C'H\cdot\log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})}-||f^\alpha-\bar{f}||_{D_H^k}\cdot\sqrt{kH\zeta^2+C'H\cdot\log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})}\\&+L_{2,k}(f)-L_{2,k}(f^\alpha)\Big\}\end{aligned} \tag{100}$$

For the last term in (100), it can be bounded by Lemma D.2. Moreover, since $\widehat{f}_{k+1}=\operatorname{argmin}_{f\in\mathcal{F}}L_{2,k}(f)$, $L_{2,k}(\widehat{f}_{k+1})-L_{2,k}(\bar{f})\le 0$.

Therefore we have: with probability at least $1-\delta$, for all $k\in[K]$ and all $f\in\mathcal{F}$, that

$$\begin{aligned}&\frac{1}{2}||\widehat{f}_{k+1}-\bar{f}||^2_{D_H^k}-\sqrt{kH\zeta^2+C'H\cdot\log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})}\cdot||\widehat{f}_{k+1}-\bar{f}||_{D_H^k}\\&-H^2\log(\frac{|\mathcal{F}^\alpha|}{\delta})-4\alpha kH^2-\sqrt{4\alpha kH^2}\cdot\sqrt{kH\zeta^2+C'H\cdot\log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})}\le 0\end{aligned} \tag{101}$$

By solving the quadratic equation for $||\widehat{f}_{k+1} - \bar{f}||_{D_H^k}$, we can get

$$
\begin{aligned}
&||\widehat{f}_{k+1} - \bar{f}||_{D_H^k} \leq \sqrt{kH\zeta^2 + C'H \cdot \log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})} \\
&+ \sqrt{kH\zeta^2 + C'H \cdot \log(\frac{4KH|\mathcal{F}^\alpha|}{\delta}) + 2H^2 \log(\frac{|\mathcal{F}^\alpha|}{\delta}) + 8\alpha kH^2 + 4\sqrt{\alpha kH^2} \cdot \sqrt{kH\zeta^2 + C'H \cdot \log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})}} \\
&\leq \sqrt{kH}\zeta + \sqrt{C'H \cdot \log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})} + H\sqrt{2\log(\frac{|\mathcal{F}^\alpha|}{\delta})} + \sqrt{2\left(kH\zeta^2 + C'H \cdot \log(\frac{4KH|\mathcal{F}^\alpha|}{\delta}) + 8\alpha kH^2\right)} \\
&\leq \sqrt{kH}\zeta + \sqrt{C'H \cdot \log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})} + H\sqrt{2\log(\frac{|\mathcal{F}^\alpha|}{\delta})} + 2\sqrt{kH}\zeta + 2\sqrt{C'H \cdot \log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})} + 4\sqrt{\alpha kH^2} \\
&\leq 3\sqrt{kH}\zeta + 5\sqrt{C'H^2 \cdot \log(\frac{4KH|\mathcal{F}^\alpha|}{\delta})} + 4\sqrt{\alpha kH^2}
\end{aligned} \tag{102}
$$

□

Lemma D.5. *(Near-optimism) Given K initial points $\{s_1^k\}_{k=1}^K$, we use $\{(s_h^{k*}, a_h^{k*})\}_{(k,h)\in[K]\times[H]}$ (where $s_1^{k*} = s_1^k$, $\forall k \in [K]$) to represent the dataset sampled by the optimal policy π^* in the true model. For each $(k,h) \in [K]\times[H-1]$, we define $\zeta_{h+1}^{k*} = \mathbb{P}_h\left((V_{h+1}^* - V_{\bar{P},h+1}^*)(s_h^{k*}, a_h^{k*})\right) - \left((V_{h+1}^* - V_{\bar{P},h+1}^*)(s_{h+1}^{k*})\right)$, and $\xi_h^{k*} = \mathbb{P}_h V_{\bar{P},h+1}^*(s_h^{k*}, a_h^{k*}) - \bar{f}_h(s_h^{k*}, a_h^{k*}, V_{\bar{P},h+1}^*)$. Conditioned on the event that $\bar{P} \in \cap_{k=1}^K B_k$ (98) holds. Then we have*

$$
\sum_{k=1}^K \left(V_1^*(s_1^{k*}) - V_1^k(s_1^{k*})\right) \leq \sum_{k=1}^K \sum_{h=1}^{H-1} \zeta_{h+1}^{k*} + \sum_{k=1}^K \sum_{h=1}^H \xi_h^{k*}
$$

Proof. First, according to the definition of $P^{(k)}$ (defined in 8), and since we condition on Eq.(98), $\bar{P} \in \cap_{k=1}^K B_k$, we have $V_1^k(s_1^k) \geq V_{\bar{P},1}^*(s_1^k)$, $\forall k \in [K]$.

$$
\begin{aligned}
&\sum_{k=1}^K \left(V_1^*(s_1^{k*}) - V_1^k(s_1^{k*})\right) \\
&\leq \sum_{k=1}^K \left(V_1^*(s_1^{k*}) - V_{\bar{P},1}^*(s_1^{k*})\right) \\
&\leq \sum_{k=1}^K \left(Q_1^*(s_1^{k*}, a_1^{k*}) - Q_{\bar{P},1}^*(s_1^{k*}, a_1^{k*})\right) \\
&= \sum_{k=1}^K \left(r_1(s_1^{k*}, a_1^{k*}) + \mathbb{P}_1 V_2^*(s_1^{k*}, a_1^{k*}) - \left[r_1(s_1^{k*}, a_1^{k*}) + \bar{\mathbb{P}}_1 V_{\bar{P},2}^*(s_1^{k*}, a_1^{k*})\right]\right) \\
&= \sum_{k=1}^K \left(\mathbb{P}_1 V_2^*(s_1^{k*}, a_1^{k*}) - \mathbb{P}_1 V_{\bar{P},2}^*(s_1^{k*}, a_1^{k*}) + \mathbb{P}_1 V_{\bar{P},2}^*(s_1^{k*}, a_1^{k*}) - \bar{\mathbb{P}}_1 V_{\bar{P},2}^*(s_1^{k*}, a_1^{k*})\right) \\
&= \sum_{k=1}^K \left(V_2^*(s_2^{k*}) - V_{\bar{P},2}^*(s_2^{k*}) + \zeta_2^{k*} + \mathbb{P}_1 V_{\bar{P},2}^*(s_1^{k*}, a_1^{k*}) - \bar{f}_1(s_1^{k*}, a_1^{k*}, V_{\bar{P},2}^*)\right) \leq \cdots \\
&\leq \sum_{k=1}^K \sum_{h=1}^{H-1} \zeta_{h+1}^{k*} + \sum_{k=1}^K \sum_{h=1}^H \xi_h^{k*}
\end{aligned} \tag{103}
$$

□

Lemma D.6. *For each $(k,h) \in [K] \times [H-1]$, we define $\zeta_{h+1}^k = \mathbb{P}_h\left((V_{h+1}^k - V_{h+1}^{\pi_k})(s_h^k, a_h^k)\right) - \left((V_{h+1}^k - V_{h+1}^{\pi_k})(s_{h+1}^k)\right)$, and $W_k = \sup_{\widetilde{\mathbb{P}}^k \in B_k} \sum_{h=1}^{H-1} \left(\widetilde{\mathbb{P}}_h^k - \mathbb{P}_h\right) V_{h+1}^k(s_h^k, a_h^k)$. Then we have*

$$\sum_{k=1}^K \left(V_1^k(s_1^k) - V_1^{\pi_k}(s_1^k)\right) \le \sum_{k=1}^K \sum_{h=1}^{H-1} \zeta_{h+1}^k + \sum_{k=1}^K W_k$$

Proof. First, for any $k \in [K]$, we have the following decomposition:

$$\begin{aligned}
&V_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \\
&= r_1(s_1^k, a_1^k) + \mathbb{P}_1^{(k)} V_2^k(s_h^k, a_h^k) - \left(r_1(s_1^k, a_1^k) + \mathbb{P}_1 V_2^{\pi_k}(s_1^k, a_1^k)\right) \\
&= (\mathbb{P}_1^{(k)} - \mathbb{P}_1) V_2^k(s_1^k, a_1^k) + \mathbb{P}_1(V_2^k - V_2^{\pi_k})(s_1^k, a_1^k) \\
&= (\mathbb{P}_1^{(k)} - \mathbb{P}_1) V_2^k(s_1^k, a_1^k) + \zeta_2^k + V_2^k(s_2^k) - V_2^{\pi_k}(s_2^k) = \cdots \\
&= \sum_{h=1}^{H-1} \left(\mathbb{P}_h^{(k)} - \mathbb{P}_h\right) V_{h+1}^k(s_h^k, a_h^k) + \sum_{h=1}^{H-1} \zeta_{h+1}^k \\
&\le W_k + \sum_{h=1}^{H-1} \zeta_{h+1}^k
\end{aligned} \tag{104}$$

Therefore, we have

$$\sum_{k=1}^K \left(V_1^k(s_1^k) - V_1^{\pi_k}(s_1^k)\right) \le \sum_{k=1}^K \sum_{h=1}^{H-1} \zeta_{h+1}^k + \sum_{k=1}^K W_k$$

□

Lemma D.7. *Let $\alpha > 0$, and $d := \dim_E(\mathcal{F}, \alpha)$, where $\mathcal{F}$ is the function class the algorithm used to approximate the ground-truth model (84). Then for any non-decreasing sequences $(\beta_k^2)_{k=1}^K$, conditioned on the event that $\bar{P} \in \bigcap_{k\in[K]} B_k$, where $B_k = \{\widetilde{P} \in \mathcal{P} : L_k(\widetilde{P}, \widehat{P}^{(k)}) \le \beta_k^2\}$, we have:*

$$\sum_{k=1}^K W_k \le \alpha + H(d \wedge K(H-1)) + 4\beta_K \sqrt{dK(H-1)} + \sqrt{8KH^2 \log(\frac{2}{\delta})} + KH\zeta \tag{105}$$

Proof. First, we denote

$$\mathcal{F}_t(\beta_k) = \{f \in \mathcal{F} : ||f - \widehat{f}_k||_{D_H^k} \le \beta_k\} = \{\phi(P) : P \in B_k\} \tag{106}$$

and for the convenience of notation, we denote

$$\widetilde{\mathcal{F}}_t = \mathcal{F}_t(\beta_k) \quad \text{for} \quad t \in [(k-1)(H-1)+1, k(H-1)]$$

and

$$\begin{aligned}
&x_1 = (s_1^1, a_1^1, V_2^1), x_2 = (s_2^1, a_2^1, V_3^1), \cdots, x_{H-1} = (s_{H-1}^1, a_{H-1}^1, V_H^1) \\
&x_H = (s_1^2, a_1^2, V_2^2), x_{H+1} = (s_2^2, a_2^2, V_3^2), \cdots, x_{2H} = (s_{H-1}^2, a_{H-1}^2, V_H^2) \\
&\cdots\cdots \\
&x_{(K-1)(H-1)+1} = (s_1^K, a_1^K, V_2^K), x_{(K-1)(H-1)+2} = (s_2^K, a_2^K, V_3^K), \cdots, x_{K(H-1)} = (s_{H-1}^K, a_{H-1}^K, V_H^K)
\end{aligned} \tag{107}$$

According to the definition of W_k, we have

$$\begin{aligned}
&\sum_{k=1}^{K} W_k \leq \sum_{k=1}^{K} \sup_{\widetilde{\mathbb{P}}^k \in B_k} \sum_{h=1}^{H-1} \left(\widetilde{\mathbb{P}}_h^k - \mathbb{P}_h\right) V_{h+1}^k(s_h^k, a_h^k) \\
&= \sum_{k=1}^{K} \sup_{\widetilde{\mathbb{P}}^k \in B_k} \sum_{h=1}^{H-1} \left(f_{\widetilde{\mathbb{P}}^k}(s_h^k, a_h^k, V_{h+1}^k) - \bar{f}(s_h^k, a_h^k, V_{h+1}^k) + \bar{f}(s_h^k, a_h^k, V_{h+1}^k) - \mathbb{P}_h V_{h+1}^k(s_h^k, a_h^k)\right) \\
&\leq \sum_{k=1}^{K} \sup_{\widetilde{\mathbb{P}}^k \in B_k} \sum_{h=1}^{H-1} \left(f_{\widetilde{\mathbb{P}}^k}(s_h^k, a_h^k, V_{h+1}^k) - \bar{f}(s_h^k, a_h^k, V_{h+1}^k)\right) + \sum_{k=1}^{K} \sum_{h=1}^{H-1} \left(\bar{f}(s_h^k, a_h^k, V_{h+1}^k) - \mathbb{P}_h V_{h+1}^k(s_h^k, a_h^k)\right) \\
&\leq \sum_{t=1}^{K(H-1)} w_{\mathcal{F}_t}(x_t) + \sum_{k=1}^{K} \sum_{h=1}^{H-1} \xi_h(s_h^k, a_h^k)
\end{aligned} \tag{108}$$

For the last line of (108), $w_{\widetilde{\mathcal{F}}}(x) = \sup_{f_1, f_2 \in \widetilde{\mathcal{F}}} (f_1(x) - f_2(x))$ represents the width function.

By using Lemma G.8, the first term of the above equation can be upper bounded by

$$\sum_{t=1}^{K(H-1)} w_{\mathcal{F}_t}(x_t) \leq \alpha + H(d \wedge K(H-1)) + 4\beta_K \sqrt{dK(H-1)} \tag{109}$$

For the second term, by using Assumption 4 and Azuma-Hoeffding's inequality, we have with probability at least $1-\delta$,

$$\sum_{k=1}^{K} \sum_{h=1}^{H-1} |\xi_h(s_h^k, a_h^k)| \leq \sqrt{8KH^2 \log(\frac{2}{\delta})} + KH\zeta \tag{110}$$

By combining (108), (109), (110), we have

$$\sum_{k=1}^{K} W_k \leq \alpha + H(d \wedge K(H-1)) + 4\beta_K \sqrt{dK(H-1)} + \sqrt{8KH^2 \log(\frac{2}{\delta})} + KH\zeta \tag{111}$$

□

Now we are able to analyze the regret bound of Robust-UCRL-VTR. We define the regret of the algorithm as

$$R_K = \sum_{k=1}^{K} (V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)).$$

Theorem D.8. *(Regret bound of robust model-based methods)*
Let Assumption 3 and 4 hold and $\alpha \in (0,1)$. For each $k \in [K]$, let β_k be

$$\beta_k = 3\sqrt{kH}\zeta + 5\sqrt{C'H^2 \cdot \log\left(\frac{4KH\mathcal{N}(\mathcal{F}, \alpha)}{\delta}\right)} + 4\sqrt{\alpha k H^2} \tag{112}$$

then with probability at least $1-\delta$, the total regret of Algorithm 4 is at most $\widetilde{O}\left(\sqrt{d_E}KH\zeta \log(1/\delta) + \sqrt{d_E^2 KH^3} \log(1/\delta)\right)$, where d_E represents the eluder dimension of the function class.

Proof. First, for any $k \in [K]$, and $h \in [H-1]$, $\zeta_{h+1}^k \in [-H, H]$, and $\{\zeta_{h+1}^k\}_{(k,h) \in [K] \times [H-1]}$ is a martingale difference sequence. Thus, with probability at least $1 - \delta/2$, $\sum_{k=1}^{K} \sum_{h=1}^{H-1} \zeta_{h+1}^k \leq \sqrt{2KH^3 \log(\frac{2}{\delta})}$. In the same way,

with probability at least $1-\delta/2$, $\sum_{k=1}^{K}\sum_{h=1}^{H-1}\zeta_{h+1}^{k*} \le \sqrt{2KH^3\log(\frac{2}{\delta})}$. Then conditioned on the event in Lemma D.3, we can obtain the regret bound by applying Lemma D.5, D.6, and D.7:

$$\begin{aligned}
R_K &= \sum_{k=1}^{K}(V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)) \\
&\le \sum_{k=1}^{K}(V_1^*(s_1^k) - V_1^k(s_1^k)) + \sum_{k=1}^{K}(V_1^k(s_1^k) - V_1^{\pi_k}(s_1^k)) \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H-1}\zeta_{h+1}^{k*} + \sum_{k=1}^{K}\sum_{h=1}^{H}\xi_h^{k*} + \sum_{k=1}^{K}\sum_{h=1}^{H-1}\zeta_{h+1}^{k} + \sum_{k=1}^{K}W_k \\
&\le 2\sqrt{KH^3\log(\frac{2}{\delta})} + \alpha + H(d\wedge K(H-1)) + 4\beta_K\sqrt{dK(H-1)} + 2\left(\sqrt{8KH^2\log(\frac{2}{\delta})} + KH\zeta\right) \\
&\le \alpha + H(d\wedge K(H-1)) + 4\beta_K\sqrt{dK(H-1)} + 2KH\zeta + 8\sqrt{KH^3\log(\frac{2}{\delta})}
\end{aligned} \tag{113}$$

By applying the definition of β_K in (112), we complete our proof.

□

E Proof of Theorem 6.1

In this section, we present the complete proof of Theorem 6.1. Prior to providing the proof of the theorem itself (Theorem E.2), we first establish the groundwork by introducing the following lemmas.

Lemma E.1. *(Estimation error of Single-epoch-Algorithm)*

For the Single-epoch-Algorithm (Algorithm 5), with probability at least $1-\delta$, we have

$$|\overline{V_1}(s_1) - V_1^{\pi_{ave}}(s_1)| \le \sqrt{\frac{8H^2\log(\frac{2}{\delta})}{K}}$$

where $\pi_{ave} = Unif\{\pi^1, \pi^2, \cdots, \pi^K\}$.

Proof. For $k = 1, 2, \cdots, K$, $R_k = \sum_{h=1}^{H} r_h(s_h^k, a_h^k)$, where $\{(s_h^k, a_h^k)\}_{h\in[H]}$ is sampled under policy π^k. Next we define $Z_0 = 0$, $Z_l = \sum_{k=1}^{l} R_k - \sum_{k=1}^{l} V^{\pi^k}$, $l = 1, 2, \cdots, K$. Then we have

$$\begin{aligned}
\mathbb{E}[Z_l|\mathcal{F}_{l-1}] &= \sum_{k=1}^{l-1} R_k + \mathbb{E}[R_l|\mathcal{F}_{l-1}] - \sum_{k=1}^{l} V^{\pi^k} \\
&= \sum_{k=1}^{l-1} R_k + V^{\pi^l} - \sum_{k=1}^{l} V^{\pi^k} \\
&= \sum_{k=1}^{l-1} R_k - \sum_{k=1}^{l-1} V^{\pi^k} = Z_{l-1}
\end{aligned} \tag{114}$$

This shows that $\{Z_l\}_{l=1}^K$ is a martingale. Moreover, $|Z_l - Z_{l-1}| \le 2H$, $\forall l \in [K]$. By Azuma-Hoeffding's inequality, we have for any $\epsilon \ge 0$,

$$\mathbb{P}\left(|\frac{1}{K}\sum_{k=1}^{K} R_k - \frac{1}{K}\sum_{k=1}^{K} V^{\pi_k}| \ge \frac{\epsilon}{K}\right) \le 2\exp\left\{-\frac{\epsilon^2}{8KH^2}\right\}$$

By using the fact that $\overline{V_1}(s_1) = \frac{1}{K}\sum_{k=1}^{K} R_k$, and $V_1^{\pi_{\text{ave}}}(s_1) = \frac{1}{K}\sum_{k=1}^{K} V^{\pi_k}$, we complete our proof. □

For the analysis of Algorithm 2, we need the following high probability events to represent that we get a good policy $\pi^{(i)}$ in epochs with correct misspecified parameter setting (i.e. $\zeta^{(i)} \geq \zeta$):

For each epoch $i \in \left\{0, 1, 2, \cdots, \lfloor \log_2\left(\frac{1}{\zeta}\right)\rfloor \wedge \lfloor \log_2\left(\sqrt{3K+1}\right)\rfloor\right\}$, we define:

$$\mathcal{E}_i(\delta) = \left\{V_1^*(s_1) \geq V_1^{\pi^{(i)}}(s_1) \geq V_1^*(s_1) - L(d,H,\delta)\cdot\left(\frac{d^\alpha H^\beta}{\sqrt{K^{(i)}}} + d^\alpha H^\beta \zeta^{(i)}\right)\right\}$$

where $\pi^{(i)}$ is the uniform mixture of policies gained from the i-th epoch, and $L(d,H,\delta)$ is a function of logarithmic order on d, H, δ.

We know that for any $i \in \left\{0, 1, 2, \cdots, \lfloor \log_2\left(\frac{1}{\zeta}\right)\rfloor \wedge \lfloor \log_2\left(\sqrt{3K+1}\right)\rfloor\right\}$, $\zeta^{(i)} = \frac{1}{2^i} \geq \zeta$, then according to the property of input base algorithm, i.e., it has a regret bound of $\widetilde{O}\left(d^\alpha H^\beta(\sqrt{K} + K\cdot\zeta)\right)$ if the input misspecified parameter is ζ, we know that $\mathcal{E}_i(\delta)$ happens with probability at least $1-\delta$. Next, we define the intersection of all these events:

$$\mathcal{E}(\zeta,\delta) = \bigcap_{i=0}^{\lfloor \log_2\left(\frac{1}{\zeta}\right)\rfloor \wedge \lfloor \log_2\left(\sqrt{3K+1}\right)\rfloor} \mathcal{E}_i\left(\frac{\delta}{2(\lceil \log_2(\frac{1}{\zeta})\rceil \wedge \lfloor \log_2\left(\sqrt{3K+1}\right)\rfloor)}\right) \tag{115}$$

By taking the union bound, we have

$$\mathbb{P}\left(\mathcal{E}(\zeta,\delta)\right) \geq 1 - \delta/2$$

Moreover, we define the following events to represent we get a good estimator of value function for each epoch $i \in \left\{0, 1, 2, \cdots, \lfloor \log_2\left(\sqrt{3K+1}\right)\rfloor\right\}$.

$$\mathcal{G}_i(\delta) = \left\{|\overline{V_1}^{(i)}(s_1) - V_1^{\pi^{(i)}}(s_1)| \leq \sqrt{\frac{8H^2\log(\frac{2}{\delta})}{K^{(i)}}}\right\}$$

Although the last few epochs are executed with the same policy, this process can still be regared as a martingale, and Lemma E.1 still holds. From Lemma E.1, we know that $\mathcal{G}_i(\delta)$ happens with probability at least $1-\delta$. Next, we define the intersection of all these events:

$$\mathcal{G}(\delta) = \bigcap_{i=0}^{\lfloor \log_2\left(\sqrt{3K+1}\right)\rfloor} \mathcal{G}_i\left(\frac{\delta}{2\lceil \log_2\left(\sqrt{3K+1}\right)\rceil}\right) \tag{116}$$

By taking the union bound, we have

$$\mathbb{P}\left(\mathcal{G}(\delta)\right) \geq 1 - \delta/2$$

Therefore,

$$\mathbb{P}\left(\mathcal{E}(\zeta,\delta)\bigcap\mathcal{G}(\delta)\right) \geq 1-\delta \tag{117}$$

Theorem E.2. ***(Regret bound under locally-bounded misspecified MDP with unknown misspecified parameter ζ)***

Suppose the input base algorithm Alg. that needs to know the locally-bounded misspecified parameter ζ has a regret bound of $\widetilde{O}\left(d^\alpha H^\beta(\sqrt{K} + K\cdot\zeta)\right)$, then conditioned on the high probability event $\mathcal{E}(\zeta,\delta)\bigcap\mathcal{G}(\delta)$ in (117), the total regret of our meta-algorithm (Algorithm 2) is still $\widetilde{O}\left(d^\alpha H^\beta(\sqrt{K} + K\cdot\zeta)\right)$.

Proof. Conditioned on the event $\mathcal{E}(\zeta,\delta)\bigcap\mathcal{G}(\delta)$ (The definition is in (115) and (116)), we have a claim here: for all i such that $\zeta^{(i)} \geq \zeta$,

$$|\overline{V_1}^{(i)} - \overline{V_1}^{(i-1)}| \leq C(d,H,\delta)\cdot\zeta^{(i)} \tag{118}$$

where

$$C(d,H,\delta)=3\sqrt{8H^2\log\left(\frac{2\lceil\log_2\left(\sqrt{3K+1}\right)\rceil}{\delta}\right)}+6L\left(d,H,\frac{2\lceil\log_2\left(\sqrt{3K+1}\right)\rceil}{\delta}\right)\cdot d^\alpha H^\beta \tag{119}$$

is a function of (d,H,δ), which has an order of $\widetilde{O}(d^\alpha H^\beta)$.

This is because

$$\begin{aligned}
&|\overline{V_1}^{(i)}-\overline{V_1}^{(i-1)}|\le|\overline{V_1}^{(i)}-V_1^{\pi^{(i)}}|+|V_1^{\pi^{(i)}}-V_1^{\pi^{(i-1)}}|+|V_1^{\pi^{(i-1)}}-\overline{V_1}^{(i-1)}|\\
&\le\sqrt{\frac{8H^2\log(\frac{2\lceil\log_2\left(\sqrt{3K+1}\right)\rceil}{\delta})}{K^{(i)}}}\\
&+L\left(d,H,\frac{2\lceil\log_2\left(\sqrt{3K+1}\right)\rceil}{\delta}\right)\cdot\left(\frac{d^\alpha H^\beta}{\sqrt{K^{(i)}}}+d^\alpha H^\beta\cdot\zeta^{(i)}+\frac{d^\alpha H^\beta}{\sqrt{K^{(i-1)}}}+d^\alpha H^\beta\cdot\zeta^{(i-1)}\right)\\
&+\sqrt{\frac{8H^2\log(\frac{2\lceil\log_2\left(\sqrt{3K+1}\right)\rceil}{\delta})}{K^{(i-1)}}}\\
&\le C(d,H,\delta)\cdot\zeta^{(i)}
\end{aligned}\tag{120}$$

The above inequality is derived by using (115), (116), and the fact that $K^{(i)}=\frac{1}{(\zeta^{(i)})^2}$

Then we discuss following two cases with respect to ζ.

Case 1 $0<\zeta<\frac{1}{2^{\lfloor\log_2(\sqrt{3K+1})\rfloor}}$ In this case, for all $i\in\left\{0,1,2,\cdots,\lfloor\log_2\left(\sqrt{3K+1}\right)\rfloor\right\}$, $\zeta^{(i)}\ge\zeta$. This means that the algorithm will not violate the condition $|\overline{V_1}^{(i)}-\overline{V_1}^{(i-1)}|\le C(d,H,\delta)\cdot\zeta^{(i)}$ for all $i\in\left\{1,2,\cdots,\lfloor\log_2\left(\sqrt{3K+1}\right)\rfloor\right\}$. Then

$$\begin{aligned}
\text{Regret}(K)&=\sum_{i=0}^{\lfloor\log_2(\sqrt{3K+1})\rfloor}\widetilde{O}\left(d^\alpha H^\beta(\sqrt{K^{(i)}}+K^{(i)}\cdot\zeta^{(i)})\right)\\
&\le\sum_{i=0}^{\lfloor\log_2(\sqrt{3K+1})\rfloor}\widetilde{O}\left(d^\alpha H^\beta\sqrt{K^{(i)}}\right)\quad\left(\zeta^{(i)}=\sqrt{\frac{1}{K^{(i)}}}\right)\\
&=\widetilde{O}(d^\alpha H^\beta)\cdot\sum_{i=0}^{\lfloor\log_2(\sqrt{3K+1})\rfloor}2^i\\
&\le\widetilde{O}(d^\alpha H^\beta)\cdot\left(\sqrt{3K+1}\right)\\
&=\widetilde{O}\left(\sqrt{K}d^\alpha H^\beta\right)
\end{aligned}\tag{121}$$

Case 2 $\frac{1}{2^{\lfloor\log_2(\sqrt{3K+1})\rfloor}}\le\zeta\le1$ We have for any $i\ge1$ such that $\zeta^{(i)}\ge\zeta$, $|\overline{V_1}^{(i)}-\overline{V_1}^{(i-1)}|\le C(d,H,\delta)\cdot\zeta^{(i)}$. We denote j to be the first epoch number that violates the condition. This means that

$$|\overline{V_1}^{(j)}-\overline{V_1}^{(j-1)}|>C(d,H,\delta)\cdot\zeta^{(j)}\tag{122}$$

while

$$|\overline{V_1}^{(i)}-\overline{V_1}^{(i-1)}|\le C(d,H,\delta)\cdot\zeta^{(i)},\quad\forall i=1,\cdots,j-1\tag{123}$$

According to our claim (118), we know that $\zeta^{(j)}<\zeta$. Moreover, according to our exponentially decreasing $\{\zeta^{(i)}\}$,

there must exist a $\zeta^{(s)}$, such that $\zeta \leq \zeta^{(s)} < 2\zeta$. For the gap between $\overline{V_1}^{(j-1)}$ and $\overline{V_1}^{(s)}$, from (123) we have

$$\begin{aligned}|\overline{V_1}^{(j-1)} - \overline{V_1}^{(s)}| &\leq |\overline{V_1}^{(j-1)} - \overline{V_1}^{(j-2)}| + |\overline{V_1}^{(j-2)} - \overline{V_1}^{(j-3)}| + \cdots + |\overline{V_1}^{(s+1)} - \overline{V_1}^{(s)}| \\ &\leq C(d,H,\delta) \cdot \left(\frac{1}{2^{j-1}} + \frac{1}{2^{j-2}} + \cdots + \frac{1}{2^{s+1}}\right) \\ &\leq C(d,H,\delta) \cdot \frac{1}{2^s} = C(d,H,\delta) \cdot \zeta^{(s)}\end{aligned} \tag{124}$$

Then we can bound the gap between $V_1^{\pi^{(j-1)}}$ and $V_1^{\pi^{(s)}}$

$$\begin{aligned}|V_1^{\pi^{(j-1)}} - V_1^{\pi^{(s)}}| &\leq |V_1^{\pi^{(j-1)}} - \overline{V_1}^{(j-1)}| + |\overline{V_1}^{(j-1)} - \overline{V_1}^{(s)}| + |\overline{V_1}^{(s)} - V_1^{\pi^{(s)}}| \\ &\leq C(d,H,\delta) \cdot \zeta^{(j-1)} + C(d,H,\delta) \cdot \zeta^{(s)} + C(d,H,\delta) \cdot \zeta^{(s)} \\ &\leq 3C(d,H,\delta) \cdot \zeta^{(s)}\end{aligned} \tag{125}$$

Similarly, for any $s+1 \leq i \leq j-1$, we have

$$|V_1^{\pi^{(i)}} - V_1^{\pi^{(s)}}| \leq 3C(d,H,\delta) \cdot \zeta^{(s)}$$

Therefore, for any $s+1 \leq i \leq j-1$, we have

$$\begin{aligned}V_1^* - V_1^{\pi^{(i)}} &= V_1^* - V_1^{\pi^{(s)}} + V_1^{\pi^{(s)}} - V_1^{\pi^{(i)}} \\ &\leq C(d,H,\delta) \cdot \zeta^{(s)} + 3C(d,H,\delta) \cdot \zeta^{(s)} \\ &= 4C(d,H,\delta) \cdot \zeta^{(s)}\end{aligned} \tag{126}$$

Next, we will give the regret bound in this case.

$$\begin{aligned}\text{Regret}(K) &= \sum_{i=0}^{s} \widetilde{O}\left(d^\alpha H^\beta(\sqrt{K^{(i)}} + K^{(i)} \cdot \zeta^{(i)})\right) + \sum_{i=s+1}^{j-1} K^{(i)}(V_1^* - V_1^{\pi^{(i)}}) + (K - \sum_{i=0}^{j-1} K^{(i)})(V_1^* - V_1^{\pi^{(j-1)}}) \\ &\leq \widetilde{O}(d^\alpha H^\beta) \cdot \sum_{i=0}^{s} 2^i + \left[\sum_{i=s+1}^{j-1} K^{(i)} + (K - \sum_{i=0}^{j-1} K^{(i)})\right] \cdot 4C(d,H,\delta) \cdot \zeta^{(s)} \quad \text{(By (126))} \\ &\leq \widetilde{O}(d^\alpha H^\beta) \cdot \sum_{i=0}^{\lfloor \log_2(\sqrt{3K+1}) \rfloor} 2^i + 4K \cdot C(d,H,\delta) \cdot \zeta^{(s)} \\ &\leq \widetilde{O}(d^\alpha H^\beta) \cdot \left(\sqrt{3K+1}\right) + 4K \cdot C(d,H,\delta) \cdot \zeta^{(s)} \\ &\leq \widetilde{O}\left(\sqrt{K} d^\alpha H^\beta\right) + 4K \cdot C(d,H,\delta) \cdot 2\zeta \leq \widetilde{O}\left(d^\alpha H^\beta(\sqrt{K} + K \cdot \zeta)\right)\end{aligned} \tag{127}$$

This completes the proof.

□

F Comparison with Transfer Error in Policy-Based Methods

In those policy-based methods (Agarwal et al., 2020a; Feng et al., 2021; Zanette et al., 2021; Li et al., 2023), they use a notion called transfer error to measure the model misspecification. They assume that the minimizer θ^* of the misspecification error with respect to state-action function with some policy π, Q^π, under the distribution of policy cover has a bounded transfer error when transferred to an arbitrary distribution d^π induced by a policy π. Formally, they define: $\theta^* = \text{argmin}_{||\theta|| \leq W} \mathbb{E}_{(s,a) \sim \rho_{\text{cover}}} \left[\phi(s,a)^\top \theta - Q^\pi(s,a)\right]^2$ then assume that for any policy π,

$$\mathbb{E}_{(s,a) \sim d^\pi} \left[\phi(s,a)^\top \theta^* - Q^\pi(s,a)\right]^2 \leq \zeta^2$$

While a direct comparison between the bounded transfer error assumption and our Assumption 2 and 4 is not feasible, they share a common characteristic. Both assumptions measure model misspecification error based on the average sense of the policy-induced distribution, rather than considering the maximum misspecification error across all state-action pairs. We consider this shared attribute to be a crucial step in establishing a connection between value-based (or model-based) and policy-based approaches regarding model misspecification.

G Auxiliary Lemmas

In this section, we provide the necessary auxiliary lemmas that we will utilize in our proof.

Notations $\mathcal{N}_\epsilon$ denotes the ϵ-covering number of the class $\mathcal{V}$ with respect to distance $d(V_1, V_2) := \sup_{s\in\mathcal{S}}[V_1(s) - V_2(s)]$.

Lemma G.1. *Let $\Lambda_t = \lambda I + \sum_{i=1}^{t} \phi_i \phi_i^\top$ where $\phi_i \in \mathbb{R}^d$ and $\lambda > 0$. Then:*

$$\sum_{i=1}^{t} \phi_i^\top (\Lambda_t)^{-1} \phi_i \leq d$$

Lemma G.2. *(Y. Abbasi-Yadkori et al.,2011). Let $\{\phi_t\}_{t\geq 0}$ be a bounded sequence in $\mathbb{R}^d$ satisfying $\sup_{t\geq 0} ||\phi_t|| \geq 1$. Let $\Lambda_0 \in \mathbb{R}^{d\times d}$ be a positive definite matrix. For any $t \geq 0$, we define $\Lambda_t = \Lambda_0 + \sum_{j=1}^{t} \phi_j \phi_j^\top$. Then if the smallest eigenvalue of Λ_0 satisfies $\lambda_{min}(\Lambda_0) \geq 1$, we have*

$$\log\left[\frac{det(\Lambda_t)}{det(\Lambda_0)}\right] \leq \sum_{j=1}^{t} \phi_j^\top \Lambda_{j-1}^{-1} \phi_j \leq 2\log\left[\frac{det(\Lambda_t)}{det(\Lambda_0)}\right]$$

.

Lemma G.3. *(Lemma D.4 in (Jin et al., 2020)). Let $\{s_\tau\}_{\tau=1}^{\infty}$ be a stochastic process on state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}_\tau\}_{\tau=0}^{\infty}$. Let $\{\phi_\tau\}_{\tau=0}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process where $\phi_\tau \in \mathcal{F}_{\tau-1}$, and $||\phi_\tau|| \leq 1$. Let $\Lambda_k = \lambda I_d + \sum_{\tau=1}^{k-1} \phi_\tau \phi_\tau^\top$. Then with probablility at least $1-\delta$, for all $k \geq 0$ and $V \in \mathcal{V}$ such that $\sup_{s\in\mathcal{S}} |V(s)| \leq H$, we have*

$$\left|\left|\sum_{\tau=1}^{k} \phi_\tau \left(V(s_\tau) - \mathbb{E}[V(s_\tau)|\mathcal{F}_{\tau-1}]\right)\right|\right|_{\Lambda_k^{-1}}^2 \leq 4H^2\left(\frac{d}{2}\log\left(\frac{k+\lambda}{\lambda}\right) + \log\left(\frac{\mathcal{N}_\epsilon}{\delta}\right)\right) + \frac{8k^2\epsilon^2}{\lambda}$$

Lemma G.4. *For any $\epsilon > 0$, the ϵ-covering number of Euclidean ball in $\mathbb{R}^d$ with radius $R > 0$ is upper bounded by $(1 + 2R/\epsilon)^d$.*

Lemma G.5. *(Lemma D.6 in [Jin et al., 2020]). Let $\mathcal{V}$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the following form*

$$V(\cdot) = \min\{\max_{a\in\mathcal{A}}\{w^\top \phi(\cdot, a) + \beta\sqrt{\phi(\cdot, a)^\top \Lambda^{-1} \phi(\cdot, a)}\}, H\}$$

where the parameters $(\omega, \beta, \mathbf{\Lambda})$ satisfy $||\mathbf{w}|| \leq \mathbf{L}$, $\beta \in [0, B]$, and the minimum eigenvalue satisfies $\lambda_{min}(\Lambda) \geq \lambda$. Assume for all (s, a), we have $||\phi(s, a)|| \leq 1$, and let $\mathcal{N}_\epsilon$ be the ϵ-covering number of $\mathcal{V}$ with respect to distance $dist(V, V') = \sup_s |V(s) - V'(s)|$. Then

$$\log \mathcal{N}_\epsilon \leq d\log(1 + 4L/\epsilon) + d^2 \log[1 + 8d^{1/2}B^2/(\lambda\epsilon^2)]$$

Lemma G.6. *(Freedman (1975)). Consider a real-valued martingale $\{Y_k : k = 0, 1, 2, \cdots\}$ with difference sequence $\{X_k : k = 0, 1, 2, \cdots\}$, which is adapted to the filtration $\{\mathcal{F}_k : k = 0, 1, 2, \cdots\}$. Assume that the difference sequence is uniformly bounded:*

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

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

$$\sum_{k=1}^{n} \mathbb{E}\left[X_k^2 | \mathcal{F}_{k-1}\right] \leq \sigma^2$$

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

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

Lemma G.7. *(Lemma 4 in Russo and Van Roy (2013)).*

Consider random variables $(\mathcal{Z}_n | n \in \mathbb{N})$ adapted to the filtration $(\mathcal{F}_n : n = 0, 1, \cdots)$. Assume $\mathbb{E}[\exp\{\lambda \mathcal{Z}_i\}]$ is finite for all λ. We define the conditional mean $\mu_i = \mathbb{E}[\mathcal{Z}_i | \mathcal{F}_{i-1}]$ and the conditional cumulant generating function of the centered random variable $[\mathcal{Z}_i - \mu_i]$ by $\phi_i(\lambda) = \log \mathbb{E}\left[\exp\{\lambda([\mathcal{Z}_i - \mu_i])\} | \mathcal{F}_{i-1}\right]$. Then we have:

For all $x \geq 0$, and $\lambda \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} \lambda \mathcal{Z}_i \leq x + \sum_{i=1}^{n} \left[\lambda \mu_i + \phi_i(\lambda)\right], \forall n \in \mathbb{N}\right) \geq 1 - e^{-x}$$

Lemma G.8. *(Lemma 2 in Russo and Van Roy (2013)).*

Let $\mathcal{F} \subset B_\infty(\mathcal{X}, C)$ be a set of functions bounded by $C > 0$. We define the width of a subset $\widetilde{\mathcal{F}} \subset \mathcal{F}$ at $x \in \mathcal{X}$ by $w_{\widetilde{\mathcal{F}}}(x) = \sup_{f_1, f_2 \in \widetilde{\mathcal{F}}} (f_1(x) - f_2(x))$. If $(\beta_t \geq 0 | t \in \mathbb{N})$ is a non-decreasing sequence, and $\{x_t\}_{t \geq 1}$ be the sequences in $\mathcal{X}$. For all $t \in \mathbb{N}$, $\mathcal{F}_t := \left\{f \in \mathcal{F} : \sup_{f_1, f_2 \in \mathcal{F}} ||f_1 - f_2||_{2, E_t} \leq 2\sqrt{\beta_t}\right\}$, where the empirical 2-norm $||\cdot||_{2, E_t}$ is defined by $||g||_{2, E_t}^2 = \sum_{k=1}^{t-1} g^2(x_k)$. Then for all $T \in \mathbb{N}$, we have

$$\sum_{t=1}^{T} w_{\mathcal{F}_t}(x_t) \leq \alpha + C(d \wedge T) + 4\sqrt{d \beta_T T} \tag{128}$$

where $d = \dim_E(\mathcal{F}, \alpha)$.