Corruption-Robust Algorithms with Uncertainty Weighting for Nonlinear Contextual Bandits and Markov Decision Processes

Chenlu Ye [1] **Wei Xiong** [1] **Quanquan Gu** [2] **Tong Zhang** [1]

Abstract

Despite the significant interest and progress in reinforcement learning (RL) problems with adversarial corruption, current works are either confined to the linear setting or lead to an undesired $\tilde{\mathcal{O}}(\sqrt{T}\zeta)$ regret bound, where T is the number of rounds and ζ is the total amount of corruption. In this paper, we consider contextual bandits with general function approximation and propose a computationally efficient algorithm to achieve a regret of $\tilde{\mathcal{O}}(\sqrt{T}+\zeta)$. The proposed algorithm relies on the recently developed uncertainty-weighted least-squares regression from linear contextual bandits (He et al., 2022) and a new weighted estimator of uncertainty for the general function class. In contrast to the existing analysis for the sum of uncertainty that is heavily based on the linear structure, we develop a novel technique to control the sum of weighted uncertainty, thus establishing the final regret bound. We then generalize our algorithm to the episodic MDP and first achieve an additive dependence on the corruption level ζ in the scenario of general function approximation. Notably, our algorithms achieve regret bounds that either nearly match the lower bound or improve the performance of existing methods for all the corruption levels in both known and unknown ζ cases.

1. Introduction

This paper studies contextual bandits (Langford & Zhang, 2007) and episodic Markov decision processes (MDPs) (Sutton & Barto, 2018) under adversarial corruption and with general function approximation. Distinct from standard contextual bandits or MDPs, we assume that the reward functions and transition kernels are subject to adversarial attacks before they are revealed to the agent at each round. This corruption concept admits model misspecification as a special case, and is complementary to the non-corruption model. It has broad applications such as autonomous vehicles misled by adversarially hacked navigation systems or contaminated traffic signs (Eykholt et al., 2018); recommendation systems tricked by adversarial comments to produce incorrect rankings (Deshpande & Montanari, 2012). Meanwhile, general function approximation has gained considerable attention since practical algorithms usually approximate true value functions or policies by a general function class $\mathcal{F}$ (e.g., neural networks) to handle the numerous states in modern large-scale reinforcement learning.

The goodness of an algorithm for learning a contextual bandit or an MDP is often measured by the notion of *regret*, which is defined as the cumulative suboptimality compared with the best policy. In the standard setting without corruption, statistically and computationally efficient algorithms have been developed for contextual bandits (Foster & Rakhlin, 2020; Zhang, 2022) and MDPs (Wang et al., 2020; Kong et al., 2021) under nonlinear function approximation. However, the performance of these algorithms can severely degrade when the underlying models are corrupted by an adversary (Lykouris et al., 2018). Therefore, designing algorithms that are robust to adversarial corruption has attracted tremendous attention (Bogunovic et al., 2021; Zhao et al., 2021; Lee et al., 2021; Ding et al., 2022; Wei et al., 2022; He et al., 2022). In the linear contextual bandit setting with T rounds, we denote the cumulative amount of corruption by $\zeta := \sum_{t=1}^{T}\zeta_t$. In this setting, Zhao et al. (2021) and Ding et al. (2022) proposed a variant of the OFUL algorithm (Abbasi-Yadkori et al., 2011) which achieved a regret of $\tilde{\mathcal{O}}\left(\sqrt{T}+\zeta\sqrt{T}\right)$, where other dependencies are omitted for simplicity. This *multiplicative* dependence on the corruption level ζ to T was undesirable as the dependence on T degrades whenever $\zeta=\omega(1)$. Wei et al. (2022) developed the COBE+VOFUL algorithm based on the model selection framework[1], which achieves an $\tilde{\mathcal{O}}\left(\sqrt{T}+\zeta\right)$ regret but

[1]The Hong Kong University of Science and Technology; [2]Department of Computer Science, University of California, Los Angeles;. Correspondence to: Tong Zhang <tongzhang@tongzhang-ml.org>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1]The adversary in Wei et al. (2022) corrupts the environment before the agent makes decisions and they define a maximization

with a sub-optimal dependence on the feature dimension. Moreover, COBE+VOFUL was known to be computationally inefficient. The gap to the minimax lower bound was later closed by He et al. (2022), where a novel uncertainty-weighted regression technique was designed to encourage estimators' reliance on the samples with low uncertainties. Nevertheless, similar results have not been established for contextual bandits with nonlinear function approximation.

We tackle the nonlinear contextual bandit with corruption by designing an efficient algorithm with desired additive dependence on the corruption level ζ. Building on recent advances in linear contextual bandits (He et al., 2022), we adapt the uncertainty-weighted regression to general function approximation. However, two challenges arise from this extension: (i) how to design sample-dependent weights with general function approximation? (ii) how to control the uncertainty level in the presence of weighting? To address the first challenge, we design an uncertainty estimator that compares the error on a newly arrived data point with the training error on the historical dataset. For the latter, we develop a novel analysis technique to control the sum of weighted uncertainty when the problem has a low eluder dimension (Russo & Van Roy, 2013). When specialized to the linear setting, the proposed algorithm enjoys a regret bound matching the nearly optimal result in He et al. (2022).

We extend our algorithm to the MDP setting under general function approximation by combining the uncertainty-weighted regression with the $\mathcal{F}$-target Least-Squares Value Iteration ($\mathcal{F}$-LSVI) (Wang et al., 2020). We assume that the adversary could attack both the rewards and transition kernels, where the amount of corruption is measured by the change of the Bellman operator (formally defined in Section 2.2). In addition to the challenges arising in the bandit case, we also encounter stability issues of the exploration bonus in the analysis of backward iterations, meaning that the statistical complexity (i.e., the covering number) of the bonus class could be extremely high. When the function class could be embedded into a (possibly infinite-dimensional) Hilbert space, we show that the notion of an effective dimension could directly control the covering number. Moreover, for the problems with a low eluder dimension, we adopt the sub-sampling technique (Wang et al., 2020; Kong et al., 2021) to select a core set for the uncertainty estimation. The use of weighted regression also brings distinct challenges to the procedure. To the best of our knowledge, the only existing result for MDPs with general function approximation is Wei et al. (2022), whose corruption formulation is most consistent with ours. However, their algorithm utilizes fundamentally different ideas and suffers from a multiplicative dependence on the corruption level ζ. We defer a detailed comparison to Section 1.1.

over the decision set as ζ. The details are discussed in Remark 2.2.

To summarize, our contributions are threefold.

- For contextual bandits under adversarial corruption and with general function approximation, we propose an algorithm based on optimism and uncertainty-weighted regression, which is computationally efficient when the weighted regression problem can be efficiently solved. We characterize the uncertainty of each data point by its diversity from history samples. Moreover, we demonstrate that the total level of uncertainty in the weighted version can be controlled by the eluder dimension (Russo & Van Roy, 2013).
- When extending to MDPs with general function approximation, we apply the uncertainty-weighting technique to the $\mathcal{F}$-LSVI (Wang et al., 2020) and develop an algorithm, which adds a bonus for every step and establishes optimism in the backward iteration. This algorithm is computationally efficient as long as the weighted regression problem is solvable (Wang et al., 2020). To address the stability issue of the bonus class, for a Hilbert space, we show that its covering number is bounded by its effective dimension; for general space, we adapt the sub-sampling technique (Wang et al., 2020; Kong et al., 2021) but with new analysis techniques to handle the sample-dependent weights.
- In terms of regret bounds, for contextual bandits, our algorithm enjoys a regret bound of $\tilde{\mathcal{O}}(\sqrt{T \dim \ln N} + \zeta \dim)$, where T is the number of rounds, $\dim$ is the eluder dimension of the function space, N is the covering number, and ζ is the cumulative corruption. This result nearly matches the bound for linear models in known and unknown ζ cases. For MDPs, the proposed algorithm guarantees an $\tilde{\mathcal{O}}(\sqrt{TH \dim \ln N} + \zeta \dim)$ regret, where H is the episode length, $\dim$ is the sum of the eluder dimension for H function spaces. This is the first result that has an additive dependence on ζ to T for MDPs with general function approximation.

1.1. Related Work

We have discussed most of the related works throughout the introduction, thus deferring a review of additional works to Appendix A. In this section, we will focus on comparing our work with that of Wei et al. (2022), who studied MDPs with general function approximation. We highlight similarities and differences between their work and ours as follows.

In terms of corruption models, both Wei et al. (2022) and this work consider corruptions on the Bellman operator. However, the adversary considered in this paper is slightly stronger than that of Wei et al. (2022), so our results apply to their setting but not vice versa. See Remark 2.2 for details. In terms of algorithmic ideas, we consider general function approximation under the least-squares value iteration

(LSVI) framework, where optimism is achieved by adding exploration bonuses at every observed state-action pair. We then employ a generalized uncertainty-weighted regression to improve the robustness of the proposed algorithm. The proposed algorithm is known to be computationally efficient as long as we can solve the (weighted) regression problems efficiently (Wang et al., 2020; Kong et al., 2021). On the contrary, Wei et al. (2022) consider problems with a low Bellman eluder dimension and propose COBE + GOLF, which only establishes optimism at the *initial* state and the optimistic planning step of GOLF (Jin et al., 2021) is known to be computationally intractable in general. Finally, in terms of the regret bounds, Wei et al. (2022) achieve a regret bound of $\tilde{\mathcal{O}}(\sqrt{T}+\zeta^r)$, where ζ^r quantifies the square norm of corruption before the observation of actions (refer to Remark 2.2 for details). Their bound reduces to $\zeta^r = \mathcal{O}(\sqrt{T\zeta})$ in the worst case, since they assume the boundness of corruption at each round $\zeta_t = \mathcal{O}(1)$. They state that the boundness is without loss of generality since they can reduce the problem to bounded corruption case by projecting the abnormal rewards back to their range. In comparison, our algorithm achieves a regret bound of $\tilde{\mathcal{O}}(\sqrt{T}+\zeta)$ when the corruption level ζ is known. This additive dependence of ζ is desirable. When the corruption level is unknown, we refer readers to Section 4.3 for details. However, we remark that instances covered by the Bellman eluder dimension are broader than the eluder dimension. In this sense, our frameworks do not supersede the results of Wei et al. (2022).

2. Preliminaries

In this section, we formally state the problems considered in the rest of this paper. Before continuing, we first define some notations to facilitate our discussions.

Notations. Given space $\mathcal{X}$ and $\mathcal{A}$, for any function $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, we define $f(x) = \max_{a\in\mathcal{A}} f(x,a)$. We also denote the function space $\mathcal{X} \times \mathcal{A} \to [0,1]$ by $\mathcal{I}^{[0,1]}$. We use the convention that $[n] = \{1, 2, \cdots, n\}$. Sometimes we will use the shorthand notations $z = (x, a)$, $a \wedge b = \min(a, b)$, and $a \vee b = \max(a, b)$.

2.1. Nonlinear Contextual Bandits with Corruption

In a contextual bandit problem with T rounds, the agent observes a *context* $x_t \in \mathcal{X}$ at each round $t \in [T]$ and takes action $a_t \in \mathcal{A}$. After the decision, the environment generates a reward $r_t(x_t, a_t) = f_*(x_t, a_t) + \epsilon_t(x_t, a_t)$, where $\epsilon_t(x_t, a_t)$ is a mean-zero noise. Finally, the agent receives the reward r_t, and the next step begins. To model the adversarial corruption, assume that we have access to a known function class $\mathcal{F} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ but $f_* \notin \mathcal{F}$ due to corruption. Specifically, we introduce the following notion of *cumulative corruption* for contextual bandits.

Definition 2.1 (Cumulative Corruption for Bandits)**.** The cumulative corruption is of level ζ if there exists $f_b \in \mathcal{F}$ such that for any data sequence $\{(x_t, a_t)\}_{t\in[T]}$, we have:[2]

$$\sum_{t=1}^{T} \zeta_t \le \zeta, \qquad \zeta_t = |f_*(x_t, a_t) - f_b(x_t, a_t)|.$$

This corruption model is consistent with that in He et al. (2022) and Bogunovic et al. (2021) when specialized to linear function approximation. However, our definition further allows for general nonlinear function approximation.

Remark 2.2. The adversary in our formulation is slightly stronger than in some prior works such as Lykouris et al. (2018); Zhao et al. (2021); Wei et al. (2022), where the amount of corruption at round t is determined before the decision $a_t \in \mathcal{A}$: $\tilde{\zeta} = \sum_{t=1}^{T} \max_{a\in\mathcal{A}} \zeta_t(x_t, a)$. Hence, $\tilde{\zeta}$ is larger than ζ in Definition 2.1, which is chosen after the observation of action a_t. In addition to the corruption $\tilde{\zeta}$, these prior works consider another notion of cumulative corruption level: $\zeta^r = (T\sum_{t=1}^{T} \max_{a\in\mathcal{A}} \zeta_t^2(x_t, a))^{1/2}$. We will also compare the proposed algorithms with the existing results under this setting in Section 4.

We make the following standard assumptions in the literature of contextual bandit (Abbasi-Yadkori et al., 2011).

Assumption 2.3. Suppose the following conditions hold for contextual bandits:

1. For all $(f, x, a) \in \mathcal{F} \times \mathcal{X} \times \mathcal{A}$, we have $|f(x, a)| \le 1$;
2. At each round t, let $\mathcal{S}_{t-1} = \{x_s, a_s, \epsilon_s\}_{s\in[t-1]}$ denote the history. The noise variable ϵ_t is conditional η-sub-Gaussian, i.e., for all $\lambda \in \mathbb{R}$, $\ln \mathbb{E}[e^{\lambda\epsilon_t}|x_t, a_t, \mathcal{S}_{t-1}] \le \lambda^2\eta^2/2$.

The **learning objective** in contextual bandit is to minimize the regret: $\mathrm{Reg}(T) = \sum_{t=1}^{T}[f_*(x_t) - f_*(x_t, a_t)]$.

2.2. Nonlinear MDPs with Corruption

We consider an episodic MDP, represented by a tuple $\mathcal{M}(\mathcal{X}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{X}$ and $\mathcal{A}$ are the spaces of states and actions, H is episode length, $\mathbb{P} = \{\mathbb{P}^h\}_{h\in[H]}$ is a collection of probability measures, and $r = \{r^h\}_{h\in[H]}$ is the reward function. In each episode, an initial state x^1 is drawn from an unknown but fixed distribution. At each step $h \in [H]$, the agent observes a state $x^h \in \mathcal{X}$, takes an action $a^h \in \mathcal{A}$, receives a reward r^h and transitions to the next state x^{h+1} with probability $\mathbb{P}^h(r^h, x^{h+1}|x^h, a^h)$. We assume that the reward is non-negative and $\sum_{h=1}^{H} r^h(x^h, a^h) \le 1$ almost surely.

Policy and value functions. A deterministic policy π is a set of functions $\{\pi^h : \mathcal{X} \to \mathcal{A}\}_{h\in[H]}$. Given any policy π,

[2] We may also use the definition $r_t(x_t, a_t) = f_b(x_t, a_t) + \zeta_t + \epsilon_t$ with $\sum_{t=1}^{T} |\zeta_t| \le \zeta$.

we define its Q-value and V-value functions starting from step h as follows:

$$Q_\pi^h(x,a) = \sum_{h'=h}^{H} \mathbb{E}_\pi[r^{h'}(x^{h'},a^{h'}) \,|\, x^h = x, a^h = a],$$
$$V_\pi^h(x) = \sum_{h'=h}^{H} \mathbb{E}_\pi[r^{h'}(x^{h'},a^{h'}) \,|\, x^h = x]. \quad (1)$$

It is well known from Sutton & Barto (2018) that there exists an optimal policy π^* such that for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, the optimal functions $V_*^h(x) := V_{\pi^*}^h(x) = \sup_\pi V_\pi^h(x)$ and $Q_*^h(x,a) := Q_{\pi^*}^h(x,a) = \sup_\pi Q_\pi^h(x,a)$ satisfy the following Bellman optimality equation:

$$Q_*^h(x,a) = \mathbb{E}_{r^h, x^{h+1}}\left[r^h + \max_{a' \in \mathcal{A}} Q_*^{h+1}(x^{h+1}, a') \,|\, x, a\right]. \quad (2)$$

Function approximation. We approximate the Q value function by a class of functions $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$ where $\mathcal{F}_h \subset \mathcal{I}^{[0,1]}$ for all $h \in [H]$. We use the convention that $f_{H+1} = 0$ since no reward is collected at step $H+1$. Then, we define the Bellman operator $\mathcal{T}^h$ on space $\mathcal{F}$:

$$(\mathcal{T}^h f^{h+1})(x^h, a^h) := \mathbb{E}_{r^h, x^{h+1}}\left[r^h + f^{h+1}(x^{h+1}) \middle| x^h, a^h\right],$$

and the corresponding Bellman residual:

$$\mathcal{E}^h(f, x^h, a^h) = f^h(x^h, a^h) - (\mathcal{T}^h f^{h+1})(x^h, a^h). \quad (3)$$

We now generalize the concept of corruption to the MDP setting. Since most of the existing approaches for MDPs rely on the condition that Q_*^h satisfies the Bellman optimality equation, we measure the amount of corruption in terms of the change of the Bellman operator, which can occur through attacks on the rewards and probability kernels. Formally, the cumulative corruption for MDPs is defined as follows.

Definition 2.4 (Cumulative corruption)**.** The cumulative corruption is ζ if there exists a complete and compressed operator $\mathcal{T}_b^h : \mathcal{I}^{[0,1]} \to \mathcal{F}^h$ satisfying that $|(\mathcal{T}_b^h g - \mathcal{T}_b^h g')(x,a)| \le \|g - g'\|_\infty$ for any $g, g' \in \mathcal{I}^{[0,1]}$ and $(x,a) \in \mathcal{X} \times \mathcal{A}$, and for all $h \in [H]$ and any sequence $\{x_t^h, a_t^h\}_{t \in [T]} \subset \mathcal{X} \times \mathcal{A}$, we have

$$\sup_{g \in \mathcal{I}^{[0,1]}} |(\mathcal{T}^h g - \mathcal{T}_b^h g)(x_t^h, a_t^h)| \le \zeta_t^h, \quad \sum_{t=1}^{T} \zeta_t^h \le \zeta.$$

One can view $\mathcal{T}_b^h$ as the Bellman operator for the uncorrupted MDP and $\mathcal{T}^h$ for the corrupted MDP. When $H = 1$, the corruption on the Bellman operator $\mathcal{T}^h$ affects only the rewards, thus reducing the problem to that of a corrupted contextual bandit. This corruption on the Bellman operator has also been considered in Wei et al. (2022), and it is a significant extension from the misspecified linear MDP in Jin et al. (2020b). First, instead of assuming that the corruptions in the reward function and transition kernel are uniformly bounded for all rounds, we allow for non-uniform corruption and only require a bound on the cumulative corruption. Furthermore, when specialized to linear function approximation, our corruption on the Bellman operator subsumes the model in Jin et al. (2020b). Finally, we consider general (nonlinear) function approximation, which is a strict generalization of the linear scenario.

We also make the following boundedness assumption for the function class.

Assumption 2.5. For any $(h, f, x, a) \in [H] \times \mathcal{F} \times \mathcal{X} \times \mathcal{A}$, we have $f^h(x,a) \in [0,1]$.

Suppose that we play the episodic MDP for T episodes and generate a sequence of policy $\{\pi_t\}_{t \in [T]}$. The **learning objective** is to minimize the cumulative regret: $\mathrm{Reg}(T) = \sum_{t=1}^{T} \left[V_*^1(x_t^1) - V_{\pi_t}^1(x_t^1)\right]$.

2.3. Eluder Dimension and Covering Number

To measure the complexity of a general function class $\mathcal{F}$, Russo & Van Roy (2013) introduced the notion of the eluder dimension. We start with the definition of ϵ-dependence.

Definition 2.6 (ϵ-dependence)**.** A point z is ϵ-dependent on a set $\mathcal{Z}$ with respect to $\mathcal{F}$ if any $f, g \in \mathcal{F}$ such that $\sqrt{\sum_{z_i \in \mathcal{Z}} (f(z_i) - g(z_i))^2} \le \epsilon$ satisfies $|f(z) - g(z)| \le \epsilon$.

Intuitively, the ϵ-dependence means that if any two functions in a given set are relatively consistent on the historical dataset $\mathcal{Z}$, their predictions on z will also be similar. Accordingly, we say that variable z is ϵ-independent of $\mathcal{Z}$ with respect to $\mathcal{F}$ if z is not ϵ-dependent on $\mathcal{Z}$.

Definition 2.7 (Eluder Dimension)**.** For $\epsilon > 0$ and a function class $\mathcal{F}$ defined on $\mathcal{X}$, the ϵ-eluder dimension $\dim_E(\mathcal{F}, \epsilon)$ is the length of the longest sequence of elements in $\mathcal{X}$ such that for some $\epsilon' \ge \epsilon$, each element is ϵ'-independent of its predecessors.

When $\mathcal{F}$ is a set of generalized linear functions, its eluder dimension is linearly dependent on the dimension of the feature map (Russo & Van Roy, 2013). Li et al. (2021) demonstrate that the eluder dimension family is strictly larger than the generalized linear class. Osband & Van Roy (2014) prove that the quadratic function class $\{\phi(x,a)^\top \Lambda \phi(x,a) : \Lambda \in \mathbb{R}^{d \times d}\}$ has an eluder dimension of $\mathcal{O}\left(d^2 \ln(1/\epsilon)\right)$.

For infinite function classes, we require the notion of ϵ-cover and covering number. We refer readers to standard textbooks (e.g., Wainwright (2019); Zhang (2023)) for more details.

Definition 2.8 (ϵ-cover and covering number)**.** Given a function class $\mathcal{F}$, for each $\epsilon > 0$, an ϵ-cover of $\mathcal{F}$ with

respect to $\|\cdot\|_\infty$, denoted by $\mathcal{C}(\mathcal{F}, \epsilon)$, satisfies that for any $f \in \mathcal{F}$, we can find $f' \in \mathcal{C}(\mathcal{F}, \epsilon)$ such that $\|f - f'\|_\infty \leq \epsilon$. The ϵ-covering number $N(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$ is the smallest cardinality of such $\mathcal{C}(\mathcal{F}, \epsilon)$.

3. Algorithms

In this section, we discuss the limitations of existing methods and the technical challenges of extending corruption-robust algorithms to general function approximation, thus motivating our approaches. For notation simplicity, we use the shorthand notation $z = (x, a)$ and $Z_h^t := \{z_h^s\}_{s=1}^t$.

3.1. Bandits with General Function approximation

For nonlinear contextual bandit, the most prominent approach is based on the principle of "Optimism in the Face of Uncertainty" (OFU). In this approach, the core step in each round $t \in [T]$ is to construct an appropriate confidence set $\mathcal{F}_t$ so that the approximator f_b lies in $\mathcal{F}_t$ with high probability. The standard approach is to solve the following least-squares regression problem:

$$\hat{f}_t = \operatorname*{argmin}_{f \in \mathcal{F}_{t-1}} \sum_{s=1}^{t-1} \big(f(z_s) - r_s\big)^2, \tag{4}$$

and construct set $\mathcal{F}_t = \{f \in \mathcal{F}_{t-1} : \sum_{s=1}^{t-1} |f(z_s) - \hat{f}_t(z_s)|^2 \leq \beta_t^2\}$ where $\beta_t^2 = \mathcal{O}(\ln N)$ is set to be the log-covering number (as stated in Proposition 2 of (Russo & Van Roy, 2013)) such that $f_b \in \mathcal{F}_t$ with high probability. However, due to the adversarial corruption, the collected samples are now from $f_* \notin \mathcal{F}$. To ensure that $f_b \in \mathcal{F}_t$ in this case, the confidence radius needs to be enlarged. Specifically, we determine β_t by examining $\sum_{s=1}^{t-1}(f_b(z_s) - \hat{f}_t(z_s))^2$, which suffers from an additional cross-term $\sum_{s=1}^{t-1} |f_b(z_s) - \hat{f}_t(z_s)|\zeta_s$. To control this term, one has to set $\beta_t^2 = \mathcal{O}(\zeta + \ln N)$, leading to a final regret bound of $\mathcal{O}(\zeta\sqrt{T})$. For further understanding, we refer readers to the intuitive explanation for the linear contextual bandit in Appendix B of Wei et al. (2022). The multiplicative relationship between ζ and $\sqrt{T}$ is disastrous, and the regret bound becomes vacuous when $\zeta = \Omega(\sqrt{T})$.

Algorithm 1 CR-Eluer-UCB

Input: $\lambda > 0$, $T, \mathcal{F}$ and $\mathcal{F}_0 = \mathcal{F}$.
for Stage $t = 1, \ldots, T$ **do**
  Observe $x_t \in \mathcal{X}$;
  Find the weighted least-squares solution $\hat{f}_t$ as in (5);
  Find $\beta_t > 0$ and construct the confidence set $\mathcal{F}_t$ as in (6);
  Take the most optimistic function $f_t = \operatorname{argmax}_{f \in \mathcal{F}_t} f(z_t)$ and choose $a_t = \operatorname{argmax}_{a \in \mathcal{A}} f_t(x_t, a)$;
  Receive r_t and set weight σ_t^2 as in (8)
end for

To overcome this issue, we adapt the uncertainty-weighting strategy from He et al. (2022), which considers the linear function class in terms of a feature $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ and employs the weighted ridge regression:

$$\hat{\theta}_t \leftarrow \operatorname*{argmin}_{\theta \in \mathcal{F}_{t-1}} \sum_{s=1}^{t-1} \frac{(\theta^\top \phi(z_s) - r_s)^2}{\max(1, \frac{1}{\alpha}\|\phi(z_s)\|_{\Lambda_s^{-1}})},$$

where $\Lambda_t = \lambda I + \sum_{s=1}^{t-1} \phi(z_s)\phi(z_s)^\top$ and $\alpha > 0$ is a tuning parameter. The weights in this equation are a truncated version of the bonus $\|\phi(z_s)\|_{\Lambda_t^{-1}}$, which can be viewed as the uncertainty of the sample. This means that the estimation of $\hat{f}_t$ will rely more on samples with low uncertainty. The close correlation between weights and bonuses stems from the following considerations: when making decisions, the exploration bonus encourages uncertainty; when making estimations, the weights are used to punish uncertainty.

As a natural extension, we replace (4) with its weighted version:

$$\hat{f}_t = \operatorname*{argmin}_{f \in \mathcal{F}_{t-1}} \sum_{s=1}^{t-1} \frac{\big(f(z_s) - r_s\big)^2}{\sigma_s^2}, \tag{5}$$

and we accordingly set the confidence set as

$$\mathcal{F}_t = \Big\{f \in \mathcal{F}_{t-1} : \sum_{s=1}^{t-1} \frac{\big(f(z_s) - \hat{f}_t(z_s)\big)^2}{\sigma_s^2} \leq \beta_t^2\Big\}. \tag{6}$$

We then compute the most optimistic value function f_t from the confidence set $\mathcal{F}_t$ and follow it greedily. The algorithm is called corruption-robust eluder UCB algorithm (CR-Eluder-UCB), whose pseudo-code is given in Algorithm 1. While the algorithmic framework shares a similar spirit with the weighted ridge regression, the extension is not straightforward because the bonus and weight choices in He et al. (2022) and their theoretical analysis (which will be discussed later) heavily rely on the linear structure. To handle the general function approximation, we first define the following (weighted) uncertainty estimator:[3]

$$D_{\lambda,\sigma,\mathcal{F}_t}(Z_1^t) = 1 \wedge \sup_{f \in \mathcal{F}_t} \frac{|f(z_t) - \hat{f}_t(z_t)|/\sigma_t}{\sqrt{\lambda + \sum_{s=1}^{t-1} |f(z_s) - \hat{f}_t(z_s)|^2/\sigma_s^2}}, \tag{7}$$

where $Z_1^t = \{(x_s, a_s)\}_{s \in [t]}$, and we omit the superscript since bandits only have one step. Intuitively, it measures the degree to which the prediction error on (x_t, a_t) exceeds the historical error evaluated on Z_1^{t-1}, thus serving as an uncertainty estimation of z_t given the historical dataset Z_1^{t-1}. Accordingly, with $\alpha > 0$ as a tuning parameter, we adopt the following weights:

$$\sigma_t^2 = 1 \vee \sup_{f \in \mathcal{F}_t} \frac{|f(z_t) - \hat{f}_t(z_t)|/\alpha}{\sqrt{\lambda + \sum_{s=1}^{t-1} \big(f(z_s) - \hat{f}_t(z_s)\big)^2/\sigma_s^2}}, \tag{8}$$

[3] We remark that Gentile et al. (2022) considers a similar quantity but without weighting.

where $a \vee b := \max(a, b)$. When the problem has a low eluder dimension, regular regression, i.e., $\sigma_s \equiv 1$, has a nonlinear analog to the elliptical potential lemma (Abbasi-Yadkori et al., 2011) for linear problems.

$$\sup_{Z_1^T} \sum_{t=1}^{T} D^2_{\lambda,\sigma,\mathcal{F}}(Z_1^t) = \tilde{\mathcal{O}}\big(\dim_E(\mathcal{F}, \lambda)\big). \tag{9}$$

However, when a sequence of weights $\sigma_s \geq 1$ is introduced, the situation becomes quite different. In linear settings, introducing the weights is straightforward. For any ball $\|\theta\|_2 \leq R$ with radius $R > 0$, the weighted version θ/σ_s also lies in the original ball of parameters. This result heavily relies on the linear structure where the function class is closed under scaling. In contrast, for weighted regression with general nonlinear function approximation, we need to explicitly control the weights in order to control the uncertainty level, which requires new analysis techniques. The generalized result of (8) is Lemma 5.1, which also handles the case of $\sigma_s \geq 1$. This result is essential for the theoretical analysis of nonlinear function approximation.

3.2. MDPs with General Function Approximation

In this subsection, we extend Algorithm 1 to MDPs with general function approximation and adversarial corruption. We adopt the $\mathcal{F}$-LSVI algorithm, which constructs the Q-value function estimations f_t^h backward from $h = H$ to $h = 1$, with the initialization $f_t^{H+1} = 0$. Specifically, suppose that we have constructed f_t^{h+1} and proceed to construct f_t^h. We reduce the problem to the bandit scenario by defining auxiliary variables $y_s^h = r_s^h + f_t^{h+1}(x_s^{h+1})$ for $s \leq t-1$. Then, we solve the following weighted least-squares problem to approximate the Bellman optimality equation (2):

$$\hat{f}_t^h = \operatorname*{argmin}_{f^h \in \mathcal{F}_{t-1}^h} \sum_{s=1}^{t-1} \frac{(f^h(x_s^h, a_s^h) - y_s^h)^2}{(\sigma_s^h)^2}. \tag{10}$$

We create the confidence set $\mathcal{F}_t^h = \{f^h \in \mathcal{F}_{t-1}^h : \lambda + \sum_{s=1}^{t-1}(f^h(z_s^h) - \hat{f}_t^h(z_s^h))^2/(\sigma_s^h)^2 \leq (\beta_t^h)^2\}$ such that $\mathcal{T}_b^h f_t^{h+1} \in \mathcal{F}_t^h$ with a high probability. Then, we choose the most optimistic estimation by adding a bonus function to the least-square solution. Specifically, for any $z \in \mathcal{X} \times \mathcal{A}$,

$$f_t^h(z) = 1 \wedge \big(\hat{f}_t^h(z) + \beta_t^h b_t^h(z)\big), \tag{11}$$

where we also clip the estimation to the valid range. It remains to determine the bonus function b_t^h, which serves as an uncertainty estimator. A natural candidate is

$$\sup_{f^h \in \mathcal{F}_t^h} \frac{|f^h(z) - \hat{f}_t^h(z)|}{\sqrt{\lambda + \sum_{s=1}^{t-1} |f^h(z_s^h) - \hat{f}_t^h(z_s^h)|^2/(\sigma_s^h)^2}}. \tag{12}$$

However, one may not directly use such a bonus. Compared with the bandit setting, since f_t^{h+1} is computed by the least-squares problem in later steps, it depends on $\{x_s^h, a_s^h\}_{s=1}^{t-1}$ because the later state is influenced by the previous decisions. Therefore, concentration inequality cannot be applied directly due to the measurability issue. The standard approach to address the issue is to establish a uniform concentration over an ϵ-covering of the function space f_t^{h+1}, which depends on $\mathcal{F}^{h+1}$ and the space of bonus b_t^{h+1}. The bonus (12) is "unstable" in the sense that the statistical complexity (covering number) of the bonus space could be extremely high because it is obtained from an optimization problem involving Z_1^T whose size can be as large as T in the worst case.

Algorithm 2 CR-LSVI-UCB

Input: $\lambda > 0$, T, $\{\mathcal{F}^h\}$, $\{\mathcal{B}^h(\lambda)\}$ and $\mathcal{F}_0^h = \mathcal{F}^h$
for Episode $t = 1, \ldots, T$ **do**
  Receive the initial state x_t^1
  Let $f_t^{H+1} = 0$
  for Step $h = H, \ldots, 1$ **do**
    Find the weighted least-squares solution $\hat{f}_t^h$ as in (10);
    Set $f_t^h(x^h, a^h)$ as in (11) with $\beta_t^h > 0$, and bonus function $b_t^h(\cdot, \cdot)$ as in (12).
  end for
  Let π_t be the greedy policy of f_t^h for each step $h \in [H]$;
  Play policy π_t and observe trajectory $\{(x_t^h, a_t^h, r_t^h)\}_{h=1}^H$;
  Set weight σ_t^h as in (14) for all $h \in [H]$;
end for

To facilitate our analysis, we assume that we have access to a bonus class $\mathcal{B}^{h+1}(\lambda)$ that has a mild covering number and can approximate (12). While we may encounter an approximation error, we assume that the equality holds for simplicity because it suffices to illustrate the idea that we want to design a corruption-robust algorithm for MDPs with general function approximation. We will provide examples in Appendix E where the approximation error is usually negligible under suitable conditions. In summary, we assume that for any sequence of data $Z_h^{t-1} = \{(x_s^h, a_s^h, r_s^h)\}_{s=1}^{t-1}$, we can find a $b_t^h \in \mathcal{B}^h(\lambda)$ that equals (12). With this bonus space, we are ready to present Corruption-Robust LSVI-UCB (CR-LSVI-UCB) in Algorithm 2. We define the confidence interval quantity for MDPs $D_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t)$ as:[4]

$$1 \wedge \sup_{f^h \in \mathcal{F}_t^h} \frac{|f^h(z_t^h) - \hat{f}_t^h(z_t^h)|/\sigma_t^h}{\sqrt{\lambda + \sum_{s=1}^{t-1} |f^h(z_s^h) - \hat{f}_t^h(z_s^h)|^2/(\sigma_s^h)^2}}. \tag{13}$$

Note that the $D_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t)$ and the bonus b_t^h mainly differ in a factor of weight in the iteration t. Therefore, they are almost identical for the regular regression. However, as we employ weighted regression in the algorithm, they are not identical. Specifically, $b_t^h(z)$ is the bonus used for the agent to ensure optimism, while $D_t^h(z)$ is the ratio between the weighted prediction error and weighted in-sample error,

[4] We note that a similar treatment of nonlinear LSVI-UCB appeared in Zhang (2023) but without weighting.

which is used in the theoretical analysis to bound the regret (see Eqn. (15) for details).

Similarly, with α as the tuning parameter, the sample-dependent weights are chosen to be

$$(\sigma_t^h)^2 = 1 \vee \sup_{f^h \in \mathcal{F}_t^h} \frac{|f^h(z_t^h) - \hat{f}_t^h(z_t^h)|/\alpha}{\sqrt{\lambda + \sum_{s=1}^{t-1}(f^h(z_s^h) - \hat{f}_t^h(z_s^h))^2/(\sigma_s^h)^2}}. \tag{14}$$

4. Main Results

In this section, we establish the main theoretical results.

4.1. Bandits with General Function Approximation

Theorem 4.1. *Suppose that Assumption 2.3 holds. For any cumulative corruption $\zeta > 0$ and $\delta \in (0,1)$, we take the covering parameter $\gamma = 1/(T\zeta)$, the eluder parameter $\lambda = \ln(N(\gamma, \mathcal{F}, \|\cdot\|_\infty))$, the weighting parameter $\alpha = \sqrt{\ln(N(\gamma, \mathcal{F}, \|\cdot\|_\infty))}/\zeta$ and the confidence radius*

$$\beta = c_\beta(\alpha\zeta + \sqrt{\ln(N(\gamma, \mathcal{F}, \|\cdot\|_\infty)/\delta)} + \sqrt{c_0}),$$

where $c_0 = \sqrt{\eta^2 \ln(2/\delta)}$ and $c_\beta > 0$ is an absolute constant. Then, with probability at least $1-\delta$, the cumulative regret with T rounds is bounded by

$$\tilde{\mathcal{O}}\Big(\sqrt{T \dim_E\left(\mathcal{F}, \frac{\lambda}{T}\right) \ln(N(\gamma, \mathcal{F}, \|\cdot\|_\infty))} + \zeta \dim_E\left(\mathcal{F}, \frac{\lambda}{T}\right)\Big).$$

Interpretation. This theorem asserts that for contextual bandits under adversarial corruption and with general function approximation, Algorithm 1 achieves an additive dependence on the cumulative corruption ζ as desired. To interpret the result, we consider linear contextual bandits with corruption, where $\dim_E(\mathcal{F}, \lambda/T) = d$ and $\ln(N(\gamma, \mathcal{F}, \|\cdot\|_\infty)) = \tilde{\mathcal{O}}(d)$. This implies a regret bound of $\tilde{\mathcal{O}}(d\sqrt{T} + d\zeta)$, matching that of He et al. (2022). According to the lower bound in He et al. (2022) and two existing lower bound results (Lattimore & Szepesvári, 2020; Bogunovic et al., 2021), our regret is minimax optimal up to logarithmic factors. Particularly when $\zeta = \mathcal{O}(\sqrt{T})$ and $\dim_E(\mathcal{F}, \lambda/T) = \tilde{\mathcal{O}}(\ln(N(\gamma, \mathcal{F}, \|\cdot\|_\infty))$, the first term dominates and matches the regret bound in the uncorrupted setting. Moreover, our regret is sublinear when $\zeta = o(T)$.

4.2. MDPs with General Function Approximation

Theorem 4.2. *Suppose that Assumption 2.5 holds. For any cumulative corruption $\zeta > 0$ and $\delta \in (0,1)$, we take the covering parameter $\gamma = 1/(T\beta\zeta)$, the eluder parameter $\lambda = \ln(N_T(\gamma))$, the weighting parameter $\alpha = \sqrt{\ln N_T(\gamma)}/\zeta$ and the confidence radius $\beta_t^h = \beta(\delta) = c_\beta(\alpha\zeta + \sqrt{\ln(HN_T(\gamma)/\delta)})$, where*

$$N_T(\gamma) = \max_h N\Big(\frac{\gamma}{T}, \mathcal{F}^h\Big) \cdot N\Big(\frac{\gamma}{T}, \mathcal{F}^{h+1}\Big) \cdot N\Big(\frac{\gamma}{T}, \mathcal{B}^{h+1}(\lambda)\Big),$$

and we remark that the coverings are all with respect to the $\|\cdot\|_\infty$. Then, with probability at least $1-\delta$, the cumulative regret with T rounds is bounded by

$$\tilde{\mathcal{O}}\Big(\sqrt{TH \ln(N_T(\gamma)) \dim_E\left(\mathcal{F}, \frac{\lambda}{T}\right)} + \zeta \dim_E\left(\mathcal{F}, \frac{\lambda}{T}\right)\Big). \tag{15}$$

Interpretation. This theorem guarantees that for corrupted models in MDPs with general function approximation, Algorithm 2 achieves an additive dependence on ζ. In particular, for linear MDPs, we can obtain that $\ln N_T(\gamma) = \tilde{\mathcal{O}}(d^2)$ and

$$\dim_E\left(\mathcal{F}, \frac{\lambda}{T}\right) = \sum_{h=1}^{H} \dim_E\left(\mathcal{F}^h, \frac{\lambda}{T}\right) = \tilde{\mathcal{O}}(Hd).$$

It follows from Theorem 4.2 that $\mathrm{Reg}(T) = \tilde{\mathcal{O}}(\sqrt{TH^2d^3} + \zeta Hd)$, where the first term matches the bound of LSVI-UCB in the non-corrupted setting (Jin et al., 2020b). We compare our result with regret for misspecified linear MDPs in Jin et al. (2020b), which is a special case of our corrupted setting. By taking $\zeta = T\zeta'$, where ζ' is the uniform bound for misspecification, our algorithm achieves a regret bound of $\tilde{\mathcal{O}}(\sqrt{TH^2d^3} + \zeta' dHT)$, which is consistent with that in Jin et al. (2020b). We note that our corruption-independent term suffers from a $\sqrt{d}$ amplification due to the uniform concentration used for handling temporal dependency, which also happens for other LSVI-based works (Jin et al., 2020b; Wang et al., 2020; Kong et al., 2021).

4.3. Unknown Corruption Level

This section discusses the solution for unknown cumulative corruption ζ by following He et al. (2022). We estimate ζ by a tuning parameter $\bar{\zeta}$ and replace ζ in the threshold parameter α by $\bar{\zeta}$ for the bandit and MDP models. The bound for MDPs is shown in the following theorem.

Theorem 4.3. *Set $\beta_t^h = c_\beta\sqrt{\ln(HN_T(\gamma)/\delta)}$ and $\alpha = \sqrt{\ln N_T(\gamma)}/\bar{\zeta}$, and keep other conditions the same as Theorem 4.2. If $0 \le \zeta \le \bar{\zeta}$, with probability at least $1-\delta$, we have the same bound in (15) except that we replace ζ with $\bar{\zeta}$. If $\zeta > \bar{\zeta}$, we have $\mathrm{Reg}(T) = \tilde{\mathcal{O}}(T)$.*

The detailed proof is provided in Appendix F, and we can have similar results for bandits.

Comparison under a weak adversary. We compare our result with the COBE framework (Wei et al., 2022) under the weak adversary scenario introduced in Remark 2.2. The key property of COBE is that under a weak adversary, it can convert any algorithm **Alg** with a known corruption level ζ to a new COBE+**Alg** that achieves the same regret bound in the unknown ζ case.

For general function approximation, COBE + GOLF achieves a regret bound of $\mathcal{O}(\sqrt{T} + \zeta^r)$, which degen-

erates to $\tilde{\mathcal{O}}(\sqrt{T\zeta})$ in the worst case. Thus, by choosing $\bar{\zeta} = \Theta(\sqrt{T})$, our regret bound outperforms theirs for all $\zeta = \mathcal{O}(\sqrt{T})$. Moreover, in this case, our regret bound is order-optimal due to the lower bound for the uncorrupted linear MDP. When specialized to linear setting, their proposed COBE + VARLin attains a regret bound of $\mathcal{O}\big(Hd^{4.5}\sqrt{T} + Hd^4\zeta\big)$, which is better than ours when $\zeta > \sqrt{T}$ in terms of the order of T but is also of a worse dependence on the feature dimension d.

Particularly, we can take the base algorithm **Alg** in COBE as Algorithm 2. By combining Theorem 3 of Wei et al. (2022) and Theorem 4.2, we achieve a regret bound of $\tilde{\mathcal{O}}(\sqrt{TH \dim \ln N} + \zeta \dim)$ in the unknown ζ case where N denotes the covering number and $\dim$ denotes the sum of eluder dimension. This bound reduces to $\mathcal{O}\big(d^{1.5}H\sqrt{T} + dH\zeta\big)$ in the linear setting, which is better than that of COBE + VARLin. Nevertheless, this bound only applies to the weak adversary. While our results also apply to the strong adversary as mentioned in Remark 2.2.

5. Proof Sketch

This section provides an overview of the proof to highlight the technical challenges and novelties. We focus on MDPs and defer the details of contextual bandits to Appendix B because the main ideas are similar. The proof strategy of Theorem 4.2 follows three steps.

Step I: Regret under optimism. If for all $(t,h) \in [T]\times[H]$, the Bellman backup $\mathcal{T}_b^h f_t^{h+1}$ belongs to the confidence set $\mathcal{F}_t^h$, we can show that $\sum_{t=1}^T V_*^1(x_t^1) \le \sum_{t=1}^T f_t^1(x_t^1) + H\zeta$. This optimism allows the regret to be upper bounded as

$$\begin{aligned}\mathrm{Reg}(T) &\le \sum_{t=1}^T [f_t^1(x_t^1) - V_{\pi_t}^1(x_t^1)] + H\zeta \\ &= \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{\pi_t} \mathcal{E}^h(f_t, x_t^h, a_t^h) + H\zeta,\end{aligned}$$

where the equality is from Lemma G.1 and the subscript denotes the distribution induced by executing π_t. This is the main technical reason why we consider the optimism-based algorithm . The details are shown in Appendix C.1.

Step II: Sharper confidence radius for optimism. To ensure that the optimism in step I is achieved with a high probability, we determine the confidence radius by bounding the in-sample gap between $\mathcal{T}_b^h f_t^{h+1}$ and $\hat{f}_t^h$, i.e., $\sum_{s=1}^{t-1}((\hat{f}_t^h - (\mathcal{T}_b^h f_t^{h+1}))(x_s^h, a_s^h))^2/(\sigma_s^h)^2$. This can be done by standard martingale concentration inequalities and a union bound. The key observation here is that with our uncertainty-based weight design, the cross-term can be controlled as follows:

$$\sum_{s=1}^t \frac{(\hat{f}_t^h(x_s^h, a_s^h) - (\mathcal{T}_b^h f_t^{h+1})(x_s^h, a_s^h))\zeta_s^h}{(\sigma_s^h)^2} \le 2\alpha\zeta \sup_{s<t} \beta_s^h.$$

In Lemma C.3, we demonstrate that the optimism is achieved with $(\beta_t^h)^2 = \mathcal{O}(\ln N_T^h(\lambda))$. In addition to the covering number, the regular regression also requires an $\Omega(\zeta)$ confidence radius as discussed in Section 3.1. This sharper bound for the estimation error illustrates the power of weighted regression and is the key to our improvement. The detailed proof is presented in Appendix C.2.

Step III: Bound the sum of bonus. If optimism is achieved, we can further upper bound the regret by

$$\begin{aligned}\mathrm{Reg}(T) \le \tilde{\mathcal{O}}\Bigg(&\Big[TH\beta \sum_{h=1}^H \sup_{Z_h^T} \sum_{t=1}^T D^2_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t)\Big]^{1/2} \\ &+ \zeta \sum_{h=1}^H \Big(1 + \sup_{Z_h^T} \sum_{t=1}^T D^2_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t)\Big)\Bigg).\end{aligned} \tag{16}$$

It remains to handle the sum of weighted uncertainty estimators $D^2_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t)$. As previously discussed at the end of Section 3.1, unlike the linear setting where the weighted function still lies in the original function class, the general function class is not closed under scaling. Therefore, we must specifically deal with the weights in our analysis. To solve this problem for uncertainty-based weights, we use a novel weight-level control technique. The key insight is that we can divide the samples into different classes based on their uncertainty levels. For samples in each class, through a refined analysis, we can demonstrate that their weights are roughly the same order, thus effectively canceling each other out. This is summarized in the following lemma.

Lemma 5.1. *For a function space $\mathcal{G}$ and any given sequence $Z^T = \{(x_t, a_t)\} \subset \mathcal{X} \times \mathcal{A}$. Under Algorithm 1 and 2, taking $D_{\lambda,\sigma,\mathcal{G}}(Z^T)$ in (7), the weight $\{\sigma_t\}_{t\in[T]}$ in (8), $\alpha = \sqrt{\ln N}/\zeta$ and $\lambda = \ln N$, we obtain*

$$\sup_{Z_1^T} \sum_{t=1}^T (D_{\lambda,\sigma,\mathcal{G}_t}(Z_1^t))^2 \le (\sqrt{8c_0} + 3) \dim_E\left(\mathcal{G}, \frac{\lambda}{T}\right) \log\left(\frac{T}{\lambda}\right) \ln T,$$

where c_0 is an absolute constant such that $\lambda + \beta_t^2 \le c_0 \ln N$, and we denote $\log_2(\cdot)$ by $\log(\cdot)$ and $N(\gamma, \mathcal{G}, \|\cdot\|_\infty)$ by N.

The detailed proofs of (16) and Lemma 5.1 are provided in Appendix C.3 and D respectively.

6. Conclusions

In this paper, we study contextual bandits and MDPs in the presence of adversarial corruption and with general function approximation. We propose a CR-Eluder-UCB for contextual bandits and a CR-LSVI-UCB for episodic MDPs,

respectively. The proposed algorithms are computationally efficient when the weighted regression problems can be efficiently solved (Wang et al., 2020) and are based on the uncertainty-weighted least-squares regression (He et al., 2022). Accordingly, we design a new uncertainty-weighted estimator for general function classes and develop novel techniques to control the sum of the weighted uncertainty level. This leads to regret bounds that depend additively on the corruption level. Specifically, for the contextual bandit problem, the CR-Eluder-UCB algorithm achieves regret bounds that nearly match the lower bound when specialized to the linear setting for both known and unknown corruption levels. Moreover, for nonlinear MDPs with general function approximation, CR-LSVI-UCB is the first algorithm that establishes an additive dependence on ζ.

We hope this work provides valuable insights into the broad applicability of the weighted least-squares regression technique in achieving improved robustness in the general function approximation setting. For future research, it would be interesting to explore the potential of combining uncertainty weighting with variance weighting to enhance further the regret bound for nonlinear MDPs.

7. Acknowledegement

The authors would like to thank Chen-Yu Wei for the helpful discussions. Chenlu Ye, Wei Xiong and Tong Zhang acknowledge the funding supported by the GRF 16310222 and GRF 16201320.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Bogunovic, I., Losalka, A., Krause, A., and Scarlett, J. Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pp. 991–999. PMLR, 2021.

Chen, L. and Luo, H. Finding the stochastic shortest path with low regret: The adversarial cost and unknown transition case. In *International Conference on Machine Learning*, pp. 1651–1660. PMLR, 2021.

Deshpande, Y. and Montanari, A. Linear bandits in high dimension and recommendation systems. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1750–1754. IEEE, 2012.

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

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1625–1634, 2018.

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

Foster, D. J., Gentile, C., Mohri, M., and Zimmert, J. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33:11478–11489, 2020.

Gentile, C., Wang, Z., and Zhang, T. Achieving minimax rates in pool-based batch active learning. *arXiv preprint arXiv:2202.05448*, 2022.

Gupta, A., Koren, T., and Talwar, K. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pp. 1562–1578. PMLR, 2019.

He, J., Zhou, D., Zhang, T., and Gu, Q. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *arXiv preprint arXiv:2205.06811*, 2022.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1704–1713. PMLR, 06–11 Aug 2017.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pp. 4860–4869. PMLR, 2020a.

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

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34, 2021.

Jin, T. and Luo, H. Simultaneously learning stochastic and adversarial episodic mdps with known transition. *Advances in neural information processing systems*, 33: 16557–16566, 2020.

Kong, D., Salakhutdinov, R., Wang, R., and Yang, L. F. Online sub-sampling for reinforcement learning

with general function approximation. *arXiv preprint arXiv:2106.07203*, 2021.

Langford, J. and Zhang, T. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, 20, 2007.

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

Lee, C.-W., Luo, H., Wei, C.-Y., Zhang, M., and Zhang, X. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning*, pp. 6142–6151. PMLR, 2021.

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

Li, Y., Lou, E. Y., and Shan, L. Stochastic linear optimization with adversarial corruption. *arXiv preprint arXiv:1909.02109*, 2019.

Luo, H., Wei, C.-Y., and Lee, C.-W. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34:22931–22942, 2021.

Lykouris, T., Mirrokni, V., and Paes Leme, R. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122, 2018.

Neu, G., György, A., Szepesvári, C., et al. The online loop-free stochastic shortest-path problem. In *COLT*, volume 2010, pp. 231–243. Citeseer, 2010.

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

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pp. 5478–5486. PMLR, 2019.

Rosenberg, A. and Mansour, Y. Stochastic shortest path with adversarially changing costs. *arXiv preprint arXiv:2006.11561*, 2020.

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

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

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

Wei, C.-Y., Dann, C., and Zimmert, J. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pp. 1043–1096. PMLR, 2022.

Wu, T., Yang, Y., Du, S., and Wang, L. On reinforcement learning with adversarial corruption and its application to block mdp. In *International Conference on Machine Learning*, pp. 11296–11306. PMLR, 2021.

Zhang, T. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.

Zhang, T. *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press, 2023. URL http://tongzhang-ml.org/lt-book.html. in press.

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

A. Additional Related Work

Corruption-robust bandits. Lykouris et al. (2018) first studied the multi-armed bandit problem with corruption where the reward was corrupted by ζ_t at round t and the corruption level was defined as $\zeta = \sum_{t=1}^{T} \zeta_t$. Lykouris et al. (2018) proposed an algorithm whose regret bound was of $\tilde{\mathcal{O}}(\sqrt{T}\zeta)$. Gupta et al. (2019) established a lower bound, showing that linear dependence on ζ was near-optimal. Therefore, the main goal was to design a corruption-robust algorithm with a regret bound of $\mathrm{Reg}(T) = o(T) + \mathcal{O}(\zeta)$. In particular, when ζ was non-dominating, we expected the first term to approach the non-corruption counterpart. Beyond the multi-armed bandit, Li et al. (2019) and Bogunovic et al. (2021) considered the stochastic bandit with linear function approximation. However, their algorithms heavily relied on arm-elimination techniques, thus confined to the stochastic and finite-arm scenarios. Bogunovic et al. (2021) and Lee et al. (2021) studied the corruption-robust linear contextual bandit with additional assumptions, including a diversity assumption on the context and a linearity assumption on the corruption. Zhao et al. (2021) and Ding et al. (2022) proposed a variant of OFUL (Abbasi-Yadkori et al., 2011) but the regret bounds were sub-optimal. Foster et al. (2020) considered an algorithm based on online regression oracle but the result was also sub-optimal. More recently, Wei et al. (2022) developed a general framework to handle unknown corruption case based on model selection. The COBE + OFUL algorithm of Wei et al. (2022) achieved a regret of $\tilde{\mathcal{O}}\left(\sqrt{T} + \zeta^r\right)$ where ζ^r is a different notion of corruption level. However, in the worse case, we had $\zeta^r = \mathcal{O}(\sqrt{T\zeta})$, which was still a multiplicative dependence on ζ. Wei et al. (2022) also proposed a COBE + VOFUL, which enjoyed an additive dependence of the corruption level ζ but with a sub-optimal dependence on the feature dimension. Also, COBE + VOFUL was known to be computationally inefficient. The suboptimality was fully addressed in He et al. (2022) for linear contextual bandits, which showed that a computationally-efficient algorithm could achieve the minimax lower bound in linear contextual bandits. The idea was to adaptively use the history samples in parameter estimation by assigning sample-dependent weights in the regression subroutine. However, whether we can design a computationally-efficient algorithm to achieve a regret of $\tilde{\mathcal{O}}\left(\sqrt{T} + \zeta\right)$ remains open for nonlinear contextual bandits.

Corruption-robust MDPs. Most existing works studied the adversarial reward setting where an adversary corrupted the reward function at each round, but the transition kernel remained fixed. See, e.g., Neu et al. (2010); Rosenberg & Mansour (2019; 2020); Jin et al. (2020a); Luo et al. (2021); Chen & Luo (2021) and reference therein. Notably, Jin et al. (2020a) and Jin & Luo (2020) developed algorithms that achieved near-minimax optimal regret bound in the adversarial reward case while preserving refined instance-dependent bounds in the static case. Therefore, the adversarial reward setting was relatively well understood. When the transition kernel could also be corrupted, Wu et al. (2021) designed an algorithm for the tabular MDP whose regret scales optimally with respect to the corruption level ζ. Wei et al. (2022) considered corrupted MDPs with general function approximation under a weak adversary, that determined the corruption at one round before the agent made the decision. The corruption in Wei et al. (2022) was also imposed on the Bellman operator, so was consistent with ours. They proposed COBE + GOLF based on model selection and GOLF (Jin et al., 2021), which achieved a regret bound of $\tilde{\mathcal{O}}(\sqrt{T} + \zeta^r)$, in the unknown corruption level case under the weak adversary. However, we remark that instances covered by the Bellman eluder dimension are broader than eluder dimension. In this sense, our frameworks do not supersede the results of Wei et al. (2022).

B. Proof for Contextual Bandits

This section presents the proof of Theorem 4.1.

B.1. Step I: Regret under Optimism

Lemma B.1. *Assume that $f_b \in \mathcal{F}_t$ for all $t \in [T]$. Then, we have*

$$\mathrm{Reg}(T) \le 2\zeta + \sum_{t=1}^{T} \left(f_t(x_t, a_t) - f_b(x_t, a_t) \right).$$

Proof. According to Algorithm 1, for all $t \in [T]$, since $f_b \in \mathcal{F}_t$, we have for $a_t^b = \mathrm{argmax}_{a \in \mathcal{A}} f_b(x_t, a)$,

$$f_b(x_t, a_t^b) \le f_t(x_t, a_t^b) \le f_t(x_t, a_t), \tag{17}$$

where the first inequality is due to the optimism of f_t, and the second inequality uses $a_t = \mathrm{argmax}_{a \in \mathcal{A}} f_t(x_t, a)$. Therefore,

the regret is bounded as follows:

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^{T}\left(\max_{a\in\mathcal{A}} f_*(x_t,a) - f_*(x_t,a_t)\right)\\
&\le \sum_{t=1}^{T}\left(f_b(x_t,a_t^*) - f_b(x_t,a_t) + 2\zeta_t\right)\\
&\le 2\zeta + \sum_{t=1}^{T}\left(f_b(x_t,a_t^b) - f_b(x_t,a_t)\right)\\
&\le 2\zeta + \sum_{t=1}^{T}\left(f_t(x_t,a_t) - f_b(x_t,a_t)\right),
\end{aligned}$$

where the first inequality uses Definition 2.1 and $\mathrm{argmax}_{a\in\mathcal{A}} f_*(x_t,a)$ is denoted by a_t^*, the second inequality is due to $a_t^b = \mathrm{argmax}_{a\in\mathcal{A}} f_b(x_t,a)$, and the last inequality is deduced from (17). □

B.2. Step II: Sharper confidence Radius for Optimism

Lemma B.2. *We have $f_b \in \mathcal{F}_t$ for all $t\in[T]$ with probability at least $1-\delta$ by taking $\lambda \le \beta^2$ and*

$$\beta_t = \tilde{c}_\beta\left(\eta\sqrt{\ln N(\gamma,\mathcal{F},\|\cdot\|_\infty)/\delta} + \alpha\zeta + \gamma\sqrt{t} + \sqrt{\gamma\sqrt{tC_1(t,\zeta)}}\right),$$

where $C_1(t,\zeta) = 2(\zeta^2 + 2t\eta^2 + 3\eta^2\ln(2/\delta))$, and $\tilde{c}_\beta > 0$ is an absolute constant.

Proof. By invoking Lemma G.4 with $\epsilon' = 0$, we obtain that with probability at least $1-\delta$, for all $t\in[T]$:

$$\begin{aligned}
\sum_{s=1}^{t-1}(\hat{f}_t(z_s) - f_b(z_s))^2/\sigma_s^2 \le& 10\eta^2\ln(2N(\gamma,\mathcal{F},\|\cdot\|_\infty)/\delta) + 5\sum_{s=1}^{t-1}|\hat{f}_t(z_s) - f_b(z_s)|\zeta_s/\sigma_s^2\\
&+ 10\gamma(\gamma t + \sqrt{tC_1(t,\zeta)}),
\end{aligned} \tag{18}$$

where $C_1(t,\zeta) = 2(\zeta^2 + 2t\eta^2 + 3\eta^2\ln(2/\delta))$.

According to the definition of σ_s in (8), we have for all $s \le t-1$,

$$\begin{aligned}
&|\hat{f}_t(z_s) - f_b(z_s)|/\sigma_s^2\\
&\quad\le |\hat{f}_t(z_s) - \hat{f}_s(z_s))|/\sigma_s^2 + |f_b(z_s) - \hat{f}_s(z_s))|/\sigma_s^2\\
&\quad\le \alpha\left(\sqrt{\lambda + \sum_{i=1}^{s-1}(\hat{f}_t(z_i) - \hat{f}_s(z_i))^2/\sigma_i^2} + \sqrt{\lambda + \sum_{i=1}^{s-1}(f_b(z_i) - \hat{f}_s(z_i))^2/\sigma_i^2}\right)\\
&\quad\le 2\alpha\sqrt{\lambda + \beta_s^2},
\end{aligned} \tag{19}$$

where the last inequality is due to $\hat{f}_t \in \mathcal{F}_{t-1} \subseteq \mathcal{F}_s$. Substituting (19) back into (18), we have

$$\begin{aligned}
&\left(\sum_{s=1}^{t-1}(\hat{f}_t(z_s) - f_b(z_s))^2/\sigma_s^2\right)^{1/2}\\
&\le \left(10\eta^2\ln(2N(\gamma,\mathcal{F},\|\cdot\|_\infty)/\delta) + 10\alpha\zeta\sup_{s<t}\sqrt{\lambda+\beta_s^2} + 10\gamma^2 t + 10\gamma\sqrt{tC_1(t,\zeta)}\right)^{1/2}\\
&\le \eta\sqrt{10\ln(2N(\gamma,\mathcal{F},\|\cdot\|_\infty)/\delta)} + \sqrt{10\alpha\zeta\sup_{s<t}\sqrt{\lambda+\beta_s^2}} + \gamma\sqrt{10t} + \sqrt{10\gamma\sqrt{tC_1(t,\zeta)}}\\
&\le \eta\sqrt{10\ln(2N(\gamma,\mathcal{F},\|\cdot\|_\infty)/\delta)} + 10\alpha\zeta + \sup_{s<t}\sqrt{\lambda+\beta_s^2}/4 + \gamma\sqrt{10t} + \sqrt{10\gamma\sqrt{tC_1(t,\zeta)}}\\
&\le \beta,
\end{aligned}$$

where the second inequality applies $\|x\|_2 \leq \|x\|_1$ for any vector x, the third inequality applies $\sqrt{2a \cdot b/2} \leq a + b/4$, and the last inequality holds since $\lambda \leq \beta^2$ and the upper bound of all β_s is β.

□

B.3. Step III: Bound the Sum of Bonus

We define event $E(T) = \{f_b \in \mathcal{F}_t,\ \forall\, t \in [T]\}$. Lemma (B.2) shows that $\mathbb{P}(E(T)) \geq 1 - \delta$. Therefore, by invoking Lemma B.1, we have with probability at least $1 - \delta$,

$$\begin{aligned}
\mathrm{Reg}(T) &\leq 2\zeta + \sum_{t=1}^{T} \left(f_t(z_t) - f_b(z_t)\right) && (20)\\
&\leq 2\zeta + \sum_{t=1}^{T} \min\left(2, |f_t(z_t) - f_b(z_t)|\right)\\
&= 2\zeta + \underbrace{\sum_{t:\sigma_t=1} \min\left(2, |f_t(z_t) - f_b(z_t)|\right)}_{p_1} + \underbrace{\sum_{t:\sigma_t>1} \min\left(2, |f_t(z_t) - f_b(z_t)|\right)}_{p_2}. && (21)
\end{aligned}$$

Now, we bound these two terms as follows. For term p_1, we obtain

$$\begin{aligned}
p_1 &= \sum_{t:\sigma_t=1} \min\left(2, |f_t(z_t) - f_b(z_t)|\right)\\
&\leq \sum_{t:\sigma_t=1} \min\left(2, \sqrt{5}\beta \cdot \frac{|f_t(z_t) - f_b(z_t)|/\sigma_t}{\sqrt{\lambda + \sum_{i=1}^{t-1} \left(f_t(z_i) - f_b(z_i)\right)^2 / \sigma_i^2}}\right)\\
&\leq 3\beta \sum_{t:\sigma_t=1} \min\left(1, \frac{|f_t(z_t) - f_b(z_t)|/\sigma_t}{\sqrt{\lambda + \sum_{i=1}^{t-1} \left(f_t(z_i) - f_b(z_i)\right)^2 / \sigma_i^2}}\right)\\
&\leq 3\beta \sum_{t=1}^{T} D_{\lambda,\sigma,\mathcal{F}_t}(Z_1^t)\\
&\leq 3\beta \sqrt{T \sup_{Z_1^T} \sum_{t=1}^{T} (D_{\lambda,\sigma,\mathcal{F}_t}(Z_1^t))^2}, && (22)
\end{aligned}$$

where we obtain the first inequality since $\lambda \leq \beta^2$ and

$$\begin{aligned}
&\lambda + \sum_{i=1}^{t-1} \left(f_t(z_i) - f_b(z_i)\right)^2 / \sigma_i^2\\
&\leq \lambda + 2\left(\sum_{i=1}^{t-1} \left(f_t(z_i) - \hat{f}_t(z_i)\right)^2 / \sigma_i^2 + \sum_{i=1}^{t-1} \left(f_b(z_i) - \hat{f}_t(z_i)\right)^2 / \sigma_i^2\right)\\
&\leq 5\beta^2, && (23)
\end{aligned}$$

and the second inequality holds because $\beta \geq 1$. Then, the term p_2 is bounded as follows:

$$\begin{aligned}
p_2 &= \sum_{t:\sigma_t>1} \min\left(2, |f_t(z_t) - f_b(z_t)|\right) \\
&= \sum_{t:\sigma_t>1} \min\left(2, \frac{|f_t(z_t) - f_b(z_t)|}{\sqrt{\lambda + \sum_{s=1}^{t-1} \left(f_t(z_s) - f_b(z_s)\right)^2 / \sigma_s^2}} \cdot \sqrt{\lambda + \sum_{i=1}^{t-1} \left(f_t(z_i) - f_b(z_i)\right)^2 / \sigma_i^2}\right) \\
&\leq \sum_{t:\sigma_t>1} \min\left(2, \frac{|f_t(z_t) - f_b(z_t)|/\sigma_t}{\sqrt{\lambda + \sum_{s=1}^{t-1} \left(f_t(z_s) - f_b(z_s)\right)^2 / \sigma_s^2}} \cdot \sqrt{5}\sigma_t\beta\right) \\
&\leq \sum_{t:\sigma_t>1} \min\left(2, \left(\sup_{f\in\mathcal{F}_t} \frac{|f(z_t) - \hat{f}_t(z_t)|/\sigma_t}{\sqrt{\lambda + \sum_{s=1}^{t-1}(f(z_s) - \hat{f}_t(z_s))^2/\sigma_s^2}}\right)^2 \cdot 3\beta/\alpha\right) \\
&\leq 3\beta/\alpha \sum_{t=1}^{T} (D_{\lambda,\sigma,\mathcal{F}_t}(Z_1^t))^2 \\
&\leq 3\beta/\alpha \sup_{Z_1^T} \sum_{t=1}^{T} (D_{\lambda,\sigma,\mathcal{F}_t}(Z_1^t))^2, \qquad (24)
\end{aligned}$$

where the first inequality invokes (23), the second inequality is deduced since for $\sigma_t > 1$,

$$\sigma_t = \frac{1}{\alpha} \sup_{f\in\mathcal{F}_t} \frac{|f(z_t) - \hat{f}_t(z_t)|/\sigma_t}{\sqrt{\lambda + \sum_{s=1}^{t-1} \left(f(z_s) - \hat{f}_t(z_s)\right)^2 / \sigma_s^2}},$$

and the third inequality applies the definition of $D_{\lambda,\sigma,\mathcal{F}_t}(Z_1^t)$ in (7). Let $\dim_\sigma$ denote $\sup_{Z_1^T} \sum_{t=1}^{T} (D_{\lambda,\sigma,\mathcal{F}_t}(Z_1^t))^2$ for simplicity. Hence, by substituting the results in (22) and (24) back into (21), we have with probability at least $1-\delta$,

$$\begin{aligned}
\mathrm{Reg}(T) &\leq 2\zeta + 3\beta\sqrt{T\dim_\sigma} + 3\beta/\alpha \dim_\sigma \\
&\leq 3c_\beta\alpha\zeta\sqrt{T\dim_\sigma} + 3c_\beta\sqrt{T\dim_\sigma \ln(N(\gamma,\mathcal{F},\|\cdot\|_\infty)/\delta)} + 3c_\beta\sqrt{c_0 T\dim_\sigma} \\
&\quad + 3\zeta\dim_\sigma + 3\dim_\sigma \sqrt{\ln(N(\gamma,\mathcal{F},\|\cdot\|_\infty)/\delta)}/\alpha + 3\sqrt{c_0}\dim_\sigma/\alpha \\
&= \tilde{\mathcal{O}}\left(\sqrt{T\dim_\sigma \ln N(\gamma,\mathcal{F},\|\cdot\|_\infty)} + \zeta\dim_\sigma\right),
\end{aligned}$$

where the last equality is obtained by setting $\alpha = \sqrt{\ln N(\gamma,\mathcal{F},\|\cdot\|_\infty)}/\zeta$.

Ultimately, invoking Lemma 5.1 results in the desired bound.

C. Proof for the MDP Case

This section contains the proof of Theorem 4.2.

C.1. Step I: Regret under Optimism

Lemma C.1. *Assuming that $\mathcal{T}_b^h f_t^{h+1} \in \mathcal{F}_t^h$ (where $\mathcal{F}_t^h$ is defined in Section 3.2) for all $(t,h) \in [T]\times[H]$, we have*

$$\mathrm{Reg}(T) \leq H\zeta + \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi_t}\mathcal{E}^h(f_t, x_t^h, a_t^h).$$

Proof. We use the notation $f_t = \{f_t^h\}_{h=1}^{H+1}$ and $f_t^h(x) = \max_{a\in\mathcal{A}} f_t^h(x,a)$ for any $x \in \mathcal{X}$. Since $\mathcal{T}_b^h f_t^{h+1} \in \mathcal{F}_t^h$ for all

$(t,h) \in [T] \times [H]$, we can invoke Lemma C.2 to obtain that

$$\begin{aligned}\mathrm{Reg}(T) &= \sum_{t=1}^{T}[V_*^1(x_t^1) - V_{\pi_t}^1(x_t^1)] \\ &\le H\zeta + \sum_{t=1}^{T}[f_t^1(x_t^1) - V_{\pi_t}^1(x_t^1)] \\ &\le H\zeta + \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi_t}\mathcal{E}^h(f_t, x_t^h, a_t^h),\end{aligned}$$

where the first inequality uses (25), the second inequality is obtained by Lemma G.1. □

Lemma C.2. *Assume that for all $(t,h) \in [T] \times [H]$, we have $\mathcal{T}_b^h f_t^{h+1} \in \mathcal{F}_t^h$. Then, the f_t^h satisfies the following inequalities:*

$$\sum_{t=1}^{T} V_*^1(x_t^1) \le \sum_{t=1}^{T} f_t^1(x_t^1) + H\zeta, \tag{25}$$

$$\left|f_t^h(x_t^h, a_t^h) - (\mathcal{T}^h f_t^{h+1})(x_t^h, a_t^h)\right| \le 2\beta_t^h b_t^h(x_t^h, a_t^h) + \zeta_t^h. \tag{26}$$

Proof. From $\mathcal{T}_b^h f_t^{h+1} \in \mathcal{F}_t^h$, we have

$$\left(\sum_{s=1}^{t-1}\frac{\left(\hat{f}_t^h(x_s^h, a_s^h) - (\mathcal{T}_b^h f_t^{h+1})(x_s^h, a_s^h)\right)^2}{(\sigma_s^h)^2} + \lambda\right)^{1/2} \le \beta_t^h.$$

Combining the inequality above and the definition of $b_t^h(\cdot)$ in (12), we obtain

$$\left|\hat{f}_t^h(x^h, a^h) - (\mathcal{T}_b^h f_t^{h+1})(x^h, a^h)\right| \le \beta_t^h b_t^h(x^h, a^h).$$

Hence, by taking $g_t^{h+1} = f_t^{h+1}$ for all $(t,h) \in [T] \times [H]$ in Definition 2.4, we have $|(\mathcal{T}^h f_t^{h+1} - \mathcal{T}_b^h f_t^{h+1})(x_t^h, a_t^h)| \le \zeta_t^h$. It follows that

$$-\zeta_t^h \le f_t^h(x_t^h, a_t^h) - (\mathcal{T}^h f_t^{h+1})(x_t^h, a_t^h) \le 2\beta_t^h b_t^h(x_t^h, a_t^h) + \zeta_t^h, \tag{27}$$

which validate (26). Then, we will demonstrate (25) by induction. When $h = H+1$, we know that $Q_*^{H+1} = f_t^{H+1} = 0$. Assume that at step $h+1$, we have

$$\sum_{t=1}^{T} Q_*^{h+1}(x_t^{h+1}, a_t^{h+1}) \le \sum_{t=1}^{T} f_t^{h+1}(x_t^{h+1}, a_t^{h+1}) + (H-h)\zeta.$$

Then, at step h, we obtain that

$$\begin{aligned}\sum_{t=1}^{T}\left(Q_*^h(x_t^h, a_t^h) - f_t^h(x_t^h, a_t^h)\right) &\le \sum_{t=1}^{T}\left((\mathcal{T}^h Q_*^{h+1})(x_t^h, a_t^h) - (\mathcal{T}^h f_t^{h+1})(x_t^h, a_t^h) + \zeta_t^h\right) \\ &\le \sum_{t=1}^{T}\mathbb{E}\left[V_*^{h+1}(x_t^{h+1}) - f_t^{h+1}(x_t^{h+1})\right] + \zeta \\ &\le (H-h+1)\zeta,\end{aligned}$$

where the first inequality uses (27), the second inequality invokes the definition of Bellman operator and cumulative corruption, and the last inequality is due to the induction hypothesis at step $h+1$.

Then, it follows that

$$\begin{aligned}\sum_{t=1}^{T} V_*^h(x_t^h) &= \sum_{t=1}^{T} Q_*^h(x_t^h, \pi_*^h(x_t^h)) \\ &\le \sum_{t=1}^{T} f_t^h(x_t^h, \pi_*^h(x_t^h)) + (H-h)\zeta \\ &\le \sum_{t=1}^{T} f_t^h(x_t^h) + (H-h)\zeta,\end{aligned}$$

where the last inequality is because of the greedy step for the optimistic function f_t^h in Algorithm 2. □

C.2. Step II: Confidence Radius for Optimism

In this step, we derive the confidence radius for f_t^h in CR-LSVI-UCB (Algorithm 2).

Lemma C.3. *In Algorithm 2 under Assumption 2.5, for all $(t,h) \in [T] \times [H]$, we have $\mathcal{T}_b^h f_t^{h+1} \in \mathcal{F}_t^h$ with probability at least $1-\delta$, where we set for each $t > 0$ that $\beta_t^{H+1} = 0$ and from $h = H$ to $h = 1$,*

$$\begin{aligned}(\beta_t^h)^2 \ge& 12\lambda + 12\ln(2HN_T^h(\gamma)/\delta) + 12\alpha\zeta \sup_{s<t} \beta_{s-1}^h + 12(5\sup_s \beta_s^{h+1}\gamma)^2 T \\ &+ 60\sup_s \beta_s^{h+1}\gamma\sqrt{TC_1(t,\zeta)},\end{aligned}$$

where $N_T^h(\gamma) = N(\gamma, \mathcal{F}^h, \|\cdot\|_\infty) \cdot N(\gamma, \mathcal{F}^{h+1}, \|\cdot\|_\infty) \cdot N(\gamma, \mathcal{B}^{h+1}(\lambda), \|\cdot\|_\infty)$ and $C_1(t,\zeta) = 2(\zeta^2 + 2t + 3\ln(2/\delta))$.

Proof. We make the statement that with probability at least $1-\delta$, we have for all $(t,h) \in [T] \times [H]$ and any $\tau \ge t$,

$$\left(\sum_{s=1}^{t-1} \frac{\left(\hat{f}_t^h(x_s^h, a_s^h) - (\mathcal{T}_b^h f_\tau^{h+1})(x_s^h, a_s^h)\right)^2}{(\sigma_s^h)^2} + \lambda\right)^{1/2} \le \beta_t^h, \tag{28}$$

which will be proved by induction. Firstly, the statement holds for $t = 1$.

Then, for $t > 1$, suppose that (28) holds for all $s \le t-1$, which means that for all $(s,h) \in [t-1] \times [H]$ and any $\tau \ge s$,

$$\left(\sum_{i=1}^{s-1} \frac{\left(\hat{f}_s^h(x_i^h, a_i^h) - (\mathcal{T}_b^h f_\tau^{h+1})(x_i^h, a_i^h)\right)^2}{(\sigma_i^h)^2} + \lambda\right)^{1/2} \le \beta_s^h.$$

Now, for round t, fix $h \in [H]$ and $\tau \ge t$. Define $\mathcal{F}_\gamma^{h+1}$ as a $\gamma\|\cdot\|_\infty$ cover of $\mathcal{F}^{h+1}$, and $\mathcal{B}_\gamma^{h+1}$ as an $\gamma\|\cdot\|_\infty$ cover of $\mathcal{B}^{h+1}(\lambda)$. Then, we construct $\bar{\mathcal{F}}_\gamma^{h+1} = \mathcal{F}_\gamma^{h+1} \oplus \beta_\tau^{h+1}\mathcal{B}_\gamma^{h+1}$ as a $(1+\beta_\tau^{h+1})\gamma\|\cdot\|_\infty$ cover of $\{f_\tau^{h+1}(\cdot)\}$. Thus, given f_τ^{h+1}, let $\bar{f}_\tau^{h+1} \in \bar{\mathcal{F}}_\gamma^{h+1}$ so that $\|\bar{f}_\tau^{h+1} - f_\tau^{h+1}\|_\infty \le \bar{\epsilon} = (1+\beta_\tau^{h+1})\gamma$. Define $\bar{y}_s^h = r_s^h + \bar{f}_\tau^{h+1}(x_s^{h+1})$ and

$$\tilde{f}_t^h = \operatorname*{argmin}_{f^h \in \mathcal{F}_{t-1}^h} \sum_{s=1}^{t-1} (f^h(x_s^h, a_s^h) - \bar{y}_s^h)^2.$$

Then, we find that

$$\begin{aligned}\left(\sum_{s=1}^{t-1} (\hat{f}_t^h(x_s^h, a_s^h) - \bar{y}_s^h)^2\right)^{1/2} &\le \left(\sum_{s=1}^{t-1} (\hat{f}_t^h(x_s^h, a_s^h) - y_s^h)^2\right)^{1/2} + \sqrt{t}\bar{\epsilon} \\ \le \left(\sum_{s=1}^{t-1} (\tilde{f}_t^h(x_s^h, a_s^h) - y_s^h)^2\right)^{1/2} &+ \sqrt{t}\bar{\epsilon} \le \left(\sum_{s=1}^{t-1} (\tilde{f}_t^h(x_s^h, a_s^h) - \bar{y}_s^h)^2\right)^{1/2} + 2\sqrt{t}\bar{\epsilon},\end{aligned} \tag{29}$$

where the first and third inequality is obtained by applying $\|\bar{f}_\tau^{h+1} - f_\tau^{h+1}\|_\infty \le \bar{\epsilon}$, and the second inequality uses the fact that $\hat{f}_t^h$ is the ERM solution of the least squares problem. Therefore, we replace f_* by $\mathbb{E}[\bar{y}_s^h | x_s^h, a_s^h] = (\mathcal{T}^h \bar{f}_\tau^{h+1})(x_s^h, a_s^h)$ and f_b by $\mathcal{T}_b^h \bar{f}_\tau^{h+1}$ in Lemma G.4 so that from the corruption in Definition 2.4 for corruption,

$$|(f_b - f_*)(x_s^h, a_s^h)| \le |(\mathcal{T}_b^h \bar{f}_\tau^{h+1} - \mathcal{T}^h \bar{f}_\tau^{h+1})(x_s^h, a_s^h)| \le \zeta_s^h.$$

Then, since $\hat{f}_t^h$ is the approximate ERM in (29), we can invoke Lemma G.4 with the f_* and f_b, $\epsilon' = 2\bar{\epsilon}$, $\eta = 1$ and $\zeta_s = \zeta_s^h$. Then, by taking a union bound over $\bar{f}_\tau^{h+1} \in \bar{\mathcal{F}}_\gamma^{h+1}$ and $h \in [H]$, we can verify that with probability at least $1 - \delta$, the following inequality holds for all $t \in [T]$:

$$\begin{aligned}
&\sum_{s=1}^{t-1} \frac{\left(\hat{f}_t^h(x_s^h, a_s^h) - (\mathcal{T}_b^h \bar{f}_\tau^{h+1})(x_s^h, a_s^h)\right)^2}{(\sigma_s^h)^2} \\
&\quad \le 10 \ln(2HN_T^h(\gamma)/\delta) + 5 \sum_{s=1}^{t-1} |\hat{f}_t^h(x_s^h, a_s^h) - (\mathcal{T}_b^h \bar{f}_\tau^{h+1})(x_s^h, a_s^h)| \cdot \zeta_s^h / (\sigma_s^h)^2 \\
&\qquad + 10(\gamma + 2\bar{\epsilon}) \cdot \left((\gamma + 2\bar{\epsilon})t + \sqrt{tC_1(t, \zeta)}\right),
\end{aligned} \tag{30}$$

where $C_1(t, \zeta) = 2(\zeta^2 + 2t + 3\ln(2/\delta))$. Further, for all $s \le t - 1$, the weights' definition in (14) indicates that

$$\begin{aligned}
&|\hat{f}_t^h(x_s^h, a_s^h) - (\mathcal{T}_b^h \bar{f}_\tau^{h+1})(x_s^h, a_s^h)|/(\sigma_s^h)^2 \\
&\quad \le |\hat{f}_t^h(x_s^h, a_s^h) - (\mathcal{T}_b^h f_\tau^{h+1})(x_s^h, a_s^h)|/(\sigma_s^h)^2 + \gamma \\
&\quad \le |\hat{f}_t^h(x_s^h, a_s^h) - \hat{f}_s^h(x_s^h, a_s^h)|/(\sigma_s^h)^2 + |(\mathcal{T}_b^h f_\tau^{h+1})(x_s^h, a_s^h) - \hat{f}_s^h(x_s^h, a_s^h)|/(\sigma_s^h)^2 + \gamma \\
&\quad \le 2\alpha\beta_s^h + \gamma,
\end{aligned}$$

where the last inequality is due to $\hat{f}_t^h \in \mathcal{F}_{t-1}^h \subset \mathcal{F}_s^h$ and the induction hypothesis that $\mathcal{T}_b^h f_\tau^{h+1} \in \mathcal{F}_s^h$ for $\tau \ge s$. Then, combining the inequality above and (30), we deduce that

$$\begin{aligned}
&\sum_{s=1}^{t-1} \frac{\left(\hat{f}_t^h(x_s^h, a_s^h) - (\mathcal{T}_b^h \bar{f}_\tau^{h+1})(x_s^h, a_s^h)\right)^2}{(\sigma_s^h)^2} \\
&\quad \le 10 \ln(2HN_T^h(\gamma)/\delta) + 10\alpha\zeta \sup_{s<t} \beta_s^h + 5\gamma\zeta + 10(\gamma + 2\bar{\epsilon}) \cdot ((\gamma + 2\bar{\epsilon})t + \sqrt{tC_1(t, \zeta)}).
\end{aligned} \tag{31}$$

Therefore, it follows that with probability at least $1 - \delta$,

$$\begin{aligned}
&\left(\sum_{s=1}^{t-1} \frac{\left(\hat{f}_t^h(x_s^h, a_s^h) - (\mathcal{T}_b^h f_\tau^{h+1})(x_s^h, a_s^h)\right)^2}{(\sigma_s^h)^2} + \lambda\right)^{1/2} \\
&\quad \le \left(\sum_{s=1}^{t-1} \frac{\left(\hat{f}_t^h(x_s^h, a_s^h) - (\mathcal{T}_b^h \bar{f}_\tau^{h+1})(x_s^h, a_s^h)\right)^2}{(\sigma_s^h)^2}\right)^{1/2} + \sqrt{t}\bar{\epsilon} + \sqrt{\lambda} \\
&\quad \le \Bigg(10 \ln(2HN_T^h(\gamma)/\delta) + 10\alpha\zeta \sup_{s<t} \beta_s^h + 5\gamma\zeta + 10(2\beta_\tau^{h+1} + 3)^2\gamma^2 T \\
&\qquad + 10(2\beta_\tau^{h+1} + 3)\gamma\sqrt{TC_1(t, \zeta)}\Bigg)^{1/2} + (\beta_\tau^{h+1} + 1)\gamma\sqrt{T} + \sqrt{\lambda} \\
&\quad \le \Bigg(12\lambda + 12 \ln(2HN_T^h(\gamma)/\delta) + 12\alpha\zeta \sup_{s<t} \beta_{s-1}^h + 12(5 \sup_s \beta_s^{h+1}\gamma)^2 T \\
&\qquad + 60 \sup_s \beta_s^{h+1}\gamma\sqrt{TC_1(t, \zeta)}\Bigg)^{1/2} \le \beta_t^h,
\end{aligned}$$

where the first inequality uses the triangle inequality, the second inequality applies (31), and the second last inequality uses Cauchy-Schwarz inequality. Therefore, we validate the statement in (28). For all $(t,h) \in [T] \times [H]$, by taking $\tau = t$ in (28), we finally complete the proof. □

Then, we demonstrate the following lemma with techniques similar to Lemma B.2.

C.3. Step III: Bound the Sum of Bonuses

Ultimately, we bound the cumulative regret for MDPs with corruption in Theorem 4.2 as follows.

Recall the definition of event $E(T)$ in Lemma C.1: for all $(t,h) \in [T] \times [H]$, f_t^h satisfies the following inequalities:

$$\sum_{t=1}^{T} V_*^1(x_t^1) \le \sum_{t=1}^{T} f_t^1(x_t^1) + H\zeta,$$
$$\left| f_t^h(x_t^h, a_t^h) - (\mathcal{T}^h f_t^{h+1})(x_t^h, a_t^h) \right| \le 2\beta_t^h b_t^h(x_t^h, a_t^h) + \zeta_t^h.$$

Since taking $\beta_t^h = \beta$ for all $(t,h) \in [T] \times [H]$ satisfies Lemma C.3, event $E(T)$ holds with probability at least $1-\delta$. Therefore, assuming $E(T)$ happens, we obtain by applying Lemma C.1 that

$$\begin{aligned}
\mathrm{Reg}(T) &\le H\zeta + \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi_t} \mathcal{E}^h(f_t, x_t^h, a_t^h) \\
&\le 2H\zeta + 2 \underbrace{\sum_{(t,h):\sigma_t^h = 1} \mathbb{E}_{\pi_t} \min(1, \beta_t^h b_t^h(x_t^h, a_t^h))}_{p_1} \\
&\quad + 2 \underbrace{\sum_{(t,h):\sigma_t^h > 1} \mathbb{E}_{\pi_t} \min(1, \beta_t^h b_t^h(x_t^h, a_t^h))}_{p_2},
\end{aligned} \tag{32}$$

where the second inequality applies (26) and $(f_t^h - \mathcal{T}_b^h f_t^{h+1})(x_t^h, a_t^h) \le 2$. Then, we bound the two terms above respectively. For the first term, we deduce that

$$\begin{aligned}
p_1 &\le \sum_{(t,h):\sigma_t^h = 1} \mathbb{E}_{\pi_t} \max(1, \beta_t^h) \cdot \min(1, b_t^h(x_t^h, a_t^h)) \\
&\le \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H} \max(1, (\beta_t^h)^2)} \cdot \mathbb{E}_{\pi_t} \sqrt{\sum_{(t,h):\sigma_t^h = 1} \min(1, (b_t^h(x_t^h, a_t^h))^2)} \\
&\le \sqrt{TH}(1+\beta) \sqrt{\sum_{h=1}^{H} \sup_{Z_h^T} \sum_{t=1}^{T} (D_{\lambda, \sigma^h, \mathcal{F}_t^h}(Z_h^t))^2},
\end{aligned}$$

where the first inequality is due to the fact that $\min(a_1 a_2, b_1 b_2) \le \max(a_1, b_1) \cdot \min(a_2, b_2)$, the second inequality is obtained by using Cauchy-Schwarz inequality, and the last inequality utilizes the definition of $D_{\lambda, \sigma^h, \mathcal{F}_t^h}(Z_h^t)$ in (13) and the selection of confidence radius: $\beta_t^h = \beta$.

Then, for $\sigma_t^h > 1$, according to the definition of σ_t^h in (14), we have $(\sigma_t^h)^2 = 1/\alpha \cdot b_t^h(x_t^h, a_t^h)$. Thus, we can bound the

second term as

$$
\begin{aligned}
p_2 &\leq \sum_{(t,h):\sigma_t^h>1} \mathbb{E}_{\pi_t} \min(1, \beta_t^h (\sigma_t^h)^2 \cdot b_t^h(x_t^h, a_t^h)/(\sigma_t^h)^2) \\
&\leq \sum_{(t,h):\sigma_t^h>1} \mathbb{E}_{\pi_t} \min(1, \beta_t^h/\alpha \cdot (b_t^h(x_t^h, a_t^h))^2/(\sigma_t^h)^2) \\
&\leq \beta/\alpha \cdot \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi_t} \min(1, (b_t^h(x_t^h, a_t^h))^2/(\sigma_t^h)^2) \\
&\leq \beta/\alpha \cdot \sum_{h=1}^{H}\sum_{t=1}^{T} \mathbb{E}_{\pi_t} (D_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t))^2 \\
&\leq \beta/\alpha \cdot \sum_{h=1}^{H} \sup_{Z_h^T} \sum_{t=1}^{T} (D_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t))^2,
\end{aligned}
$$

where the $D_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t)$ is formulated in Definition 13. Combining these results, we get

$$
\begin{aligned}
\mathrm{Reg}(T) &\leq 2H\zeta + \sqrt{TH}(1+\beta)\sqrt{\sum_{h=1}^{H} \sup_{Z_h^T} \sum_{t=1}^{T} (D_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t))^2} + \beta/\alpha \cdot \sum_{h=1}^{H} \sup_{Z_h^T} \sum_{t=1}^{T} (D_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t))^2 \\
&= \tilde{\mathcal{O}}\Bigg(\left(H + \sum_{h=1}^{H} \sup_{Z_h^T} \sum_{t=1}^{T} (D_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t))^2\right)\zeta + \sqrt{TH\ln(N_T(\gamma)) \sum_{h=1}^{H} \sup_{Z_h^T} \sum_{t=1}^{T} (D_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t))^2} \\
&\quad + \alpha\zeta\sqrt{TH\sum_{h=1}^{H} \sup_{Z_h^T} \sum_{t=1}^{T} (D_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t))^2} + \sqrt{\ln(N_T(\gamma))} \sum_{h=1}^{H} \sup_{Z_h^T} \sum_{t=1}^{T} (D_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t))^2)/\alpha\Bigg) \\
&= \tilde{\mathcal{O}}\left(\sqrt{TH\ln(N_T(\gamma)) \sum_{h=1}^{H} \sup_{Z_h^T} \sum_{t=1}^{T} (D_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t))^2} + \zeta \sum_{h=1}^{H} \sup_{Z_h^T} \sum_{t=1}^{T} (D_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t))^2\right),
\end{aligned}
$$

where the first inequality is deduced by taking the bounds of terms p_1 and p_2 back into (32), the first equality uses the choice of $\beta = \mathcal{O}(\alpha\zeta + \sqrt{\ln(H\ln(N_T(\gamma))/\delta)})$, and the last equation is obtained by setting $\alpha = \sqrt{\ln(N_T(\gamma))}/\zeta$.

Then, it suffices to replace weighted eluder dimension $\sup_{Z_h^T} \sum_{t=1}^{T} (D_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t))^2$ with the eluder dimension $\dim_E(\mathcal{F}, \epsilon)$ in Definition 2.7. Because $\mathcal{F}$ is factorized as $\prod_{h=1}^{H} \mathcal{F}^h$, we get

$$
\dim_E(\mathcal{F}, \epsilon) = \sum_{h=1}^{H} \dim_E(\mathcal{F}^h, \epsilon).
$$

By invoking Lemma 5.1 for each function space $\mathcal{F}^h$, we obtain

$$
\sup_{Z_h^T} \sum_{t=1}^{T} (D_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t))^2 \leq (\sqrt{8c_0} + 3)\dim_E(\mathcal{F}^h, \lambda/T)\log(T/\lambda)\ln T,
$$

which indicates that

$$
\sum_{h=1}^{H} \sup_{Z_h^T} \sum_{t=1}^{T} (D_{\lambda,\sigma^h,\mathcal{F}_t^h}(Z_h^t))^2 \leq (\sqrt{8c_0} + 3)\dim_E(\mathcal{F}, \lambda/T)\log(T/\lambda)\ln T.
$$

Therefore, it follows that

$$
\mathrm{Reg}(T) = \tilde{\mathcal{O}}\left(\sqrt{TH\ln(N_T(\gamma))\dim_E(\mathcal{F}, \lambda/T)} + \zeta\left(H + \dim_E(\mathcal{F}, \lambda/T)\right)\right).
$$

D. Relationship between Confidence and Eluder Dimension

For simplicity, given a data sequence $\mathcal{Z} = \{z_t\}_{t\in[T]}$, a positive weight sequence $\{\sigma_t\}_{t\in[T]}$ and functions f, f' defined on a space containing $\mathcal{Z}$, we denote the first t elements in $\mathcal{Z}$ by $\mathcal{Z}_t = \{z_s\}_{s\in[t]}$ and define the quadratic error and its weighted version as follows:

$$\|f - f'\|_{\mathcal{Z}} = \left(\sum_{z_i\in\mathcal{Z}} (f(z_i) - f'(z_i))^2\right)^{1/2},$$
$$\|f - f'\|_{\mathcal{Z},\sigma} = \left(\sum_{z_i\in\mathcal{Z}} \frac{(f(z_i) - f'(z_i))^2}{\sigma_i^2}\right)^{1/2}.$$

The following lemma reveals the relationship between the sum of eluder-like confidence quantity and the eluder dimension.

Proof of Lemma 5.1. Let $\mathcal{Z} = \{z_t\}_{t=1}^T \subset \mathcal{X} \times \mathcal{A}$.

Step I: Matched levels. To begin with, we divide the sequence $\mathcal{Z}$ into $\log(T/\lambda) + 1$ disjoint sets. For each $z_t \in \mathcal{Z}$, Let $f \in \mathcal{F}_t$ be the function that maximizes

$$\frac{(f(z_t) - \hat{f}_t(z_t))^2/\sigma_t^2}{\lambda + \|f - \hat{f}_t\|^2_{\mathcal{Z}_{t-1},\sigma}},$$

where we define $\mathcal{Z}_{t-1} = \{z_s\}_{s\in[t-1]}$ and $L(z) = (f(z) - \hat{f}_t(z))^2$ for such f. Since $L(z) \in [0, 1]$, we decompose $\mathcal{Z}$ into $\log(T/\lambda) + 1$ disjoint subsequences:

$$\mathcal{Z} = \cup_{\iota=0}^{\log(T/\lambda)} \mathcal{Z}^\iota,$$

where we define for $\iota = 0, \ldots, \log(T/\lambda) - 1$,

$$\mathcal{Z}^\iota = \{z_t \in \mathcal{Z} \,|\, L(z_t) \in (2^{-\iota-1}, 2^{-\iota}]\},$$

and

$$\mathcal{Z}^{\log(T/\lambda)} = \{z_t \in \mathcal{Z} \,|\, L(z_t) \in [0, \lambda/T]\}.$$

Correspondingly, we can also divide $\mathbb{R}^+$ into $\log(T/\lambda) + 2$ disjoint subsets:

$$\mathbb{R}^+ = \cup_{\iota=-1}^{\log(T/\lambda)} \mathcal{R}^\iota,$$

where we define $\mathcal{R}^\iota = [2^{\iota/2} \ln N, 2^{(\iota+1)/2} \ln N)$ for $\iota = 0, \ldots, \log(T/\lambda) - 1$, $\mathcal{R}^{\log(T/\lambda)} = [\sqrt{T/\lambda} \ln N, +\infty)$, and $\mathcal{R}^{-1} = [0, \ln N)$. Since $\zeta \in \mathbb{R}^+$, there exists an $\iota_0 \in \{-1, 0, \ldots, \log(T/\lambda)\}$ such that $\zeta \in \mathcal{R}^{\iota_0}$.

Step II: Control weights in each level. Now, for any $z_t \in \mathcal{Z}^{\log(T/\lambda)}$, we have

$$\sup_{f\in\mathcal{F}_t} \frac{(f(z_t) - \hat{f}_t(z_t))^2/\sigma_t^2}{\lambda + \|f - \hat{f}_t\|^2_{\mathcal{Z}_{t-1},\sigma}} \le \frac{L(z_t)}{\lambda} \le \frac{1}{T},$$

which implies that

$$\sum_{z_t\in\mathcal{Z}^{\log(T/\lambda)}} (D_{\lambda,\sigma,\mathcal{F}}(Z_1^t))^2 \le 1. \tag{33}$$

Moreover, for $0 \le \iota \le \log(T/\lambda) - 1$, we need to control the upper and lower bound of weights $\{\sigma_t : z_t \in \mathcal{Z}^\iota\}$. For convenience, define for $t \in [T]$,

$$\psi_t = \frac{1}{\alpha} \sup_{f\in\mathcal{F}_t} \frac{|f(z_t) - \hat{f}_t(z_t)|}{\sqrt{\lambda + \|f - \hat{f}_t\|^2_{\mathcal{Z}_{t-1},\sigma}}}.$$

Besides, we observe that

$$\operatorname*{argmax}_{f\in\mathcal{F}_t} \frac{(f(z_t)-\hat{f}_t(z_t))^2/\sigma_t^2}{\lambda+\|f-\hat{f}_t\|^2_{\mathcal{Z}_{t-1},\sigma}} = \operatorname*{argmax}_{f\in\mathcal{F}_t} \frac{|f(z_t)-\hat{f}_t(z_t)|}{\sqrt{\lambda+\|f-\hat{f}_t\|^2_{\mathcal{Z}_{t-1},\sigma}}},$$

so we can take the f as the solution of the two terms above for ψ_t. Hence, we have $|f(z_t)-\hat{f}_t(z_t)| = \sqrt{L(z_t)}$.

Now, consider two situations for ι. When $\iota > \iota_0$, we have $\zeta < 2^{(\iota_0+1)/2}\ln N \le 2^{\iota/2}\ln N$. For any $z_t \in \mathcal{Z}^\iota$, we know that $L(z_t) \le 2^{-\iota}$. Hence, it follows that

$$\psi_t \le \frac{1}{\alpha}\cdot\frac{2^{-\iota/2}}{\sqrt{\lambda}} \le \frac{\zeta}{\sqrt{\ln N}}\cdot\frac{2^{-\iota/2}}{\sqrt{\ln N}} \le 1,$$

where the second inequality is deduced by taking $\alpha = \sqrt{\ln N}/\zeta$ and $\lambda = \ln N$. Therefore, since $\sigma_t^2 = \max(1,\psi_t)$, we obtain $\sigma_t = 1$ for all $z_t \in \mathcal{Z}^\iota$.

When $\iota \le \iota_0$, we have $\zeta \ge 2^{\iota_0/2}\ln N \ge 2^{\iota/2}\ln N$, and $2^{-\iota-1} \le L(z_t) \le 2^{-\iota}$ for all $z_t \in \mathcal{Z}^\iota$. Then, we can verify that

$$\psi_t \le \frac{\zeta}{\sqrt{\ln N}}\cdot\frac{2^{-\iota/2}}{\sqrt{\ln N}} = \frac{\zeta}{2^{\iota/2}\ln N},$$
$$\psi_t \ge \frac{\zeta}{\sqrt{\ln N}}\cdot\frac{2^{-(\iota+1)/2}}{\sqrt{c_0\ln N}} = \frac{\zeta}{\sqrt{2c_0}2^{\iota/2}\ln N},$$

where the inequality of the second row applies

$$\lambda + \|f-\hat{f}_t\|^2_{\mathcal{Z}_{t-1},\sigma} \le \lambda + \beta_t^2 \le c_0\ln N,$$

where $c_0 > 1$ is an absolute constant. Then, since $\zeta/(2^{\iota/2}\ln N) \ge 1$, we know that $\zeta/(\sqrt{2c_0}2^{\iota/2}\ln N) \le \max(1,\psi_t) \le \zeta/(2^{\iota/2}\ln N)$. Therefore, we get for all $z_t \in \mathcal{Z}^\iota$,

$$\sigma_t^2 \in [\zeta/(\sqrt{2c_0}2^{\iota/2}\ln N), \zeta/(2^{\iota/2}\ln N)].$$

Step III: Bound the sum. Now, we bound $\sum_{z_t\in\mathcal{Z}^\iota}(D_{\lambda,\sigma,\mathcal{F}}(Z_1^t))^2$ for $\iota \in [0,\log(T/\lambda)-1]$. For each $\iota = 0,\dots,\log(T/\lambda)-1$, we decompose $\mathcal{Z}^\iota$ into $N^\iota+1$ disjoint subsets: $\mathcal{Z}^\iota = \cup_{j=1}^{N^\iota+1}\mathcal{Z}_j^\iota$, where we define $N^\iota = |\mathcal{Z}^\iota|/\dim_E(\mathcal{F},2^{-\iota-1})$. With a slight abuse of notation, let $\mathcal{Z}^\iota = \{z_i\}_{i\in[|\mathcal{Z}^\iota|]}$, where the elements are arranged in the same order as the original set $\mathcal{Z}$. Initially, let $\mathcal{Z}_j^\iota = \{\}$ for all $j \in [N^\iota+1]$. Then, from $i=1$ to $|\mathcal{Z}^\iota|$, we find the smallest $j \in [N^\iota]$ such that z_i is $2^{-\iota-1}$-independent of $\mathcal{Z}_j^\iota$ with respect to $\mathcal{F}$. If such a j does not exist, set $j = N^\iota+1$. Then, we denote the choice of j for each z_i by $j(z_i)$. According to the design of the procedure, it is obvious that for all $z_i \in \mathcal{Z}^\iota$, z_i is $2^{-\iota-1}$-dependent on each of $\mathcal{Z}_{1,i}^\iota,\dots,\mathcal{Z}_{j(z_i)-1,i}^\iota$, where $\mathcal{Z}_{k,i}^\iota = \mathcal{Z}_k^\iota \cap \{z_1,\dots,z_{i-1}\}$ for $k=1,\dots,j(z_i)-1$.

For any $z_i \in \mathcal{Z}^\iota$ indexed by t in $\mathcal{Z}$, by taking the function f that maximizes

$$\frac{(f(z_t)-\hat{f}_t(z_t))^2/\sigma_t^2}{\lambda+\|f-\hat{f}_t\|^2_{\mathcal{Z}_{t-1},\sigma}},$$

we have $(f(z_i)-\hat{f}_t(z_i))^2 \ge 2^{-\iota-1}$. Additionally, because z_i is $2^{-\iota-1}$-dependent on each of $\mathcal{Z}_{1,i}^\iota,\dots,\mathcal{Z}_{j(z_i)-1,i}^\iota$, we get for each $k=1,\dots,j(z_i)-1$,

$$\|f-\hat{f}_t\|^2_{\mathcal{Z}_{k,i}^\iota} \ge 2^{-\iota-1}.$$

It follows that

$$\frac{(f(z_t)-\hat{f}_t(z_t))^2/\sigma_t^2}{\lambda+\|f-\hat{f}_t\|^2_{\mathcal{Z}_{t-1},\sigma}} \le \frac{2^{-\iota}/\sigma_t^2}{\lambda+\sum_{k=1}^{j(z_i)-1}\|f-\hat{f}_t\|^2_{\mathcal{Z}_{k,i}^\iota,\sigma}}.$$

When $\iota > \iota_0$, we have $\sigma_t = 1$ for all $z_t \in \mathcal{Z}^\iota$. Thus, it follows that

$$\frac{(f(z_t)-\hat{f}_t(z_t))^2/\sigma_t^2}{\lambda+\|f-\hat{f}_t\|^2_{\mathcal{Z}_{t-1},\sigma}} \le \frac{2^{-\iota}}{\lambda+(j(z_i)-1)2^{-\iota-1}} = \frac{2}{j(z_i)-1+\lambda 2^{\iota+1}}.$$

Summing over all $z_t \in \mathcal{Z}^\iota$, we obtain

$$\begin{aligned}\sum_{z_t\in\mathcal{Z}^\iota}(D_{\lambda,\sigma,\mathcal{F}}(Z_1^t))^2 &\le \sum_{j=1}^{N^\iota}\sum_{z_i\in\mathcal{Z}_j^\iota}\frac{2}{j-1+\lambda 2^{\iota+1}}+\sum_{z_i\in\mathcal{Z}_{N^\iota+1}^\iota}\frac{2}{N^\iota}\\ &\le \sum_{j=1}^{N^\iota}\frac{2|\mathcal{Z}_j^\iota|}{j}+\frac{2|\mathcal{Z}_{N^\iota+1}^\iota|}{N^\iota}\\ &\le 2\dim_E(\mathcal{F},2^{-\iota-1})\ln N^\iota+2|\mathcal{Z}^\iota|\cdot\frac{\dim_E(\mathcal{F},2^{-\iota-1})}{|\mathcal{Z}^\iota|}\\ &\le 4\dim_E(\mathcal{F},2^{-\iota-1})\ln N^\iota, \end{aligned}\tag{34}$$

where the third inequality is deduced since by the definition of eluder dimension, we have $|\mathcal{Z}_j^\iota|\le\dim_E(\mathcal{F},2^{-\iota-1})$ for all $j\in[N^\iota]$.

When $\iota\le\iota_0$, we have $\sigma_s^2\in[\zeta/(\sqrt{2c_0}2^{\iota/2}\ln N),\zeta/(2^{\iota/2}\ln N)]$ for all $z_s\in\mathcal{Z}^\iota$. Therefore, their weights are roughly of the same order. Then, we can verify that

$$\begin{aligned}\frac{(f(z_t)-\hat{f}_t(z_t))^2/\sigma_t^2}{\lambda+\|f-\hat{f}_t\|^2_{\mathcal{Z}_{t-1},\sigma}} &\le \frac{(f(z_t)-\hat{f}_t(z_t))^2/\sigma_t^2}{\lambda+\|f-\hat{f}_t\|^2_{\mathcal{Z}_{t-1}\cap\mathcal{Z}^\iota,\sigma}}\\ &\le \frac{2^{-\iota}\sqrt{2c_0}2^{\iota/2}\ln N/\zeta}{\lambda+(j(z_i)-1)2^{-\iota-1}\cdot 2^{\iota/2}\ln N/\zeta}\\ &\le \frac{\sqrt{8c_0}}{j(z_i)-1+\lambda 2^{\iota/2+1}\zeta/\ln N}\\ &\le \frac{\sqrt{8c_0}}{j(z_i)-1+\lambda 2^{\iota+1}},\end{aligned}$$

where the last inequality uses $\zeta\ge 2^{\iota/2}\ln N$. Summing over all $z_t\in\mathcal{Z}^\iota$, we have

$$\begin{aligned}\sum_{z_t\in\mathcal{Z}^\iota}(D_{\lambda,\sigma,\mathcal{F}}(Z_1^t))^2 &\le \sum_{j=1}^{N^\iota}\sum_{z_i\in\mathcal{Z}_j^\iota}\frac{\sqrt{8c_0}}{j-1+\lambda 2^{\iota+1}}+\sum_{z_i\in\mathcal{Z}_{N^\iota+1}^\iota}\frac{2}{N^\iota}\\ &\le (\sqrt{8c_0}+2)\dim_E(\mathcal{F},2^{-\iota-1})\ln N^\iota.\end{aligned}\tag{35}$$

Eventually, combining (33), (34) and (35), we obtain

$$\begin{aligned}&\sum_{t=1}^T(D_{\lambda,\sigma,\mathcal{F}}(Z_1^t))^2\\ &=\sum_{\iota=0}^{\log(T/\lambda)}\sum_{z_t\in\mathcal{Z}^\iota}(D_{\lambda,\sigma,\mathcal{F}}(Z_1^t))^2\\ &\le\sum_{\iota=0}^{\iota_0}(\sqrt{8c_0}+2)\dim_E(\mathcal{F},2^{-\iota-1})\ln N^\iota+\sum_{\iota=\iota_0+1}^{\log(T/\lambda)-1}4\dim_E(\mathcal{F},2^{-\iota-1})\ln N^\iota+1\\ &\le(\sqrt{8c_0}+3)\dim_E(\mathcal{F},\lambda/T)\log(T/\lambda)\ln T,\end{aligned}$$

where the last inequality uses the monotonicity of the eluder dimension. Note that if $\iota_0=-1$, let the sum from 0 to -1 be 0. Ultimately, we accomplish the proof due to the arbitrariness of Z_1^T. □

E. Stable Bonus Funtion Space

In this section, we discuss how to stabilize the bonus function space to address the issue of covering numbers as stated in Section 3.2. We first consider a case where $\mathcal{F}$ is an RKHS so that the covering number can be directly controlled. Then, we adopt the sensitivity subsampling technique from Wang et al. (2020) for the general cases.

E.1. RKHS

In this section, we consider that for each $h \in [H]$, the function class $\mathcal{F}^h$ is embedded into some reproducing kernel Hilbert space (RKHS) $\mathcal{H}^h$:

$$\mathcal{F}^h \subset \mathcal{H}^h = \{\langle w(f), \phi(\cdot)\rangle : z \to \mathbb{R}\},$$

where we use z to denote state-action pair (x, a). We assume that $\mathcal{H}^h$ has a representation $\{\phi_j(z)\}_{j=1}^{\infty}$ in 2-norm. For any $\epsilon > 0$, define the ϵ-scale sensitive dimension as follows:

$$d(\epsilon) = \min\left\{|\mathcal{S}| : \sup_{z\in\mathcal{Z}} \sum_{j\notin\mathcal{S}} (\phi_j(z))^2 \le \epsilon\right\}.$$

Then, we can write the bonus function as

$$b_t^h(z) = \sup_{f\in\mathcal{F}_t^h} \frac{|\langle w(f) - w(\hat{f}), \phi(z)\rangle|}{\sqrt{\lambda + (w(f) - w(\hat{f}))^\top \Lambda_t (w(f) - w(\hat{f}))}},$$

where $\Lambda_t = \sum_{s=1}^{t-1} \phi(z_s)\phi(z_s)^\top/(\sigma_s^h)^2$. Then, the log-covering number of the bonus space $\mathcal{B}^h(\lambda)$ is bounded as follows.

Lemma E.1. *Suppose that $\|w\| \le 1$, $\|\phi(x, a)\| \le 1$ for all $(x, a) \in \mathcal{X} \times \mathcal{A}$, $\|\Lambda_t\|_{\text{op}} \le \rho$ for all $t \in [T]$ and $\lambda > 1$. Let $N(\gamma, \mathcal{B}^h(\lambda), \|\cdot\|_\infty)$ be the infinity γ-covering number of $\mathcal{B}^h(\lambda)$. Then, we have*

$$\ln N(\gamma, \mathcal{B}^h(\lambda), \|\cdot\|_\infty) = \tilde{\mathcal{O}}\left((d((\gamma^2/46\rho)^2))^2\right).$$

Proof. For any two functions $b_1, b_2 \in \mathcal{B}^h(\lambda)$ with parameters w_1, w_2 and Λ_1, Λ_2. Since sup_w is a contraction map, we have

$$\begin{aligned}
\|b_1 - b_2\|_\infty &\le \sup_{z,\|w\|\le 1} \left| \frac{|\langle w - w_1, \phi(z)\rangle|}{\sqrt{\lambda + (w - w_1)^\top \Lambda_1 (w - w_1)}} - \frac{|\langle w - w_2, \phi(z)\rangle|}{\sqrt{\lambda + (w - w_2)^\top \Lambda_2 (w - w_2)}} \right| \\
&\le \sup_{\|\phi|\le 1, \|w\|\le 1} \left\{ \left| \frac{|\langle w - w_1, \phi\rangle|}{\sqrt{\lambda + (w - w_1)^\top \Lambda_1 (w - w_1)}} - \frac{|\langle w - w_2, \phi\rangle|}{\sqrt{\lambda + (w - w_1)^\top \Lambda_1 (w - w_1)}} \right| \right. \\
&\qquad \left. + \left| \frac{|\langle w - w_2, \phi\rangle|}{\sqrt{\lambda + (w - w_1)^\top \Lambda_1 (w - w_1)}} - \frac{|\langle w - w_2, \phi\rangle|}{\sqrt{\lambda + (w - w_2)^\top \Lambda_2 (w - w_2)}} \right| \right\} \\
&\le \sup_{\|\phi\|\le 1} |\langle w_1 - w_2, \phi\rangle| \\
&\qquad + \sup_{\|\phi|\le 1, \|w\|\le 1} |\langle w - w_2, \phi\rangle| \cdot \left| \sqrt{\lambda + (w - w_2)^\top \Lambda_2 (w - w_2)} - \sqrt{\lambda + (w - w_1)^\top \Lambda_1 (w - w_1)} \right| \\
&\le \|w_1 - w_2\| + 2 \sqrt{\underbrace{\sup_{\|w\|\le 1} \left|(w - w_2)^\top \Lambda_2 (w - w_2) - (w - w_1)^\top \Lambda_1 (w - w_1)\right|}_{(a)}}.
\end{aligned}$$

Then, we can bound (a) by

$$\begin{aligned}
&\sup_{\|w\|\le 1} \left|(w - w_2)^\top \Lambda_2 (w - w_2) - (w - w_1)^\top \Lambda_1 (w - w_1)\right| \\
&\quad \le \sup_{\|w\|\le 1} \left|(w - w_2)^\top (\Lambda_2 - \Lambda_1)(w - w_2) + (w_1 - w_2)^\top \Lambda_1 (w_1 - w_2)\right. \\
&\qquad\qquad \left. + (w - w_1)^\top \Lambda_1 (w_1 - w_2) + (w_1 - w_2)^\top \Lambda_1 (w - w_1)\right| \\
&\quad \le \|\Lambda_2 - \Lambda_1\|_{\text{op}} + \rho\|w_1 - w_2\|^2 + 4\rho\|w_1 - w_2\|,
\end{aligned}$$

where the last inequality is obtained due to $\|\Lambda\|_{op} \le \rho$. Taking this result back, we get

$$\|b_1 - b_2\|_\infty \le \|w_1 - w_2\| + 2\sqrt{\|\Lambda_2 - \Lambda_1\|_{\text{op}} + \rho\|w_1 - w_2\|^2 + 4\rho\|w_1 - w_2\|}. \tag{36}$$

Let $\mathcal{C}_w$ be the $\gamma^2/23\rho$ cover of $\{w \in \mathcal{H}^h : \|w\| \le 1\}$ and $\mathcal{C}_\Lambda$ be the $\gamma^2/23$ cover of $\{\Lambda \in \mathcal{G} : \|\Lambda\|_{\text{op}} \le \rho\}$, where $\mathcal{G}$ is defined in Lemma E.2. By invoking Lemma E.2 and E.3, we have

$$\ln|\mathcal{C}_w| \le d((\gamma^2/46\rho)^2)\ln(1+92\rho/\gamma^2)$$
$$\ln|\mathcal{C}_\Lambda| \le (d((\gamma^2/46\rho)^2))^2\ln(1+92\rho\sqrt{d((\gamma^2/46\rho)^2)}/\gamma^2).$$

From (36), for any $b_1 \in \mathcal{B}^h(\lambda)$ with $w(f)$ and $\Lambda = \sum_{s=1}^{t-1}\phi(z_s)\phi(z_s)^\top/(\sigma_s^h)^2$, let $\tilde{\phi}(z_s) = \phi(z_s)/\sigma_s^h$, so Λ can be written as $\Lambda = \sum_{s=1}^{t-1}\tilde{\phi}(z_s)\tilde{\phi}(z_s)^\top$ with $\|\tilde{\phi}\| \le 1$, which indicates that $\Lambda \in \mathcal{G}$. Hence, there always exist a b_2 parameterized by $\tilde{w} \in \mathcal{C}_w$ and $A \in \mathcal{C}_\Lambda$ such that $\|b_1 - b_2\|_\infty \le \gamma$. Therefore, we finally obtain

$$\begin{aligned}\ln N(\gamma, \mathcal{B}^h(\lambda), \|\cdot\|_\infty) &\le \ln(|\mathcal{C}_w|\cdot|\mathcal{C}_\Lambda|)\\ &\le d((\gamma^2/46\rho)^2)\ln(1+92\rho/\gamma^2) + (d((\gamma^2/46\rho)^2))^2\ln(1+92\rho\sqrt{d((\gamma^2/46\rho)^2)}/\gamma^2).\end{aligned}$$

□

Lemma E.2 (Covering number of Covariance Matrix)**.** *Consider the space*

$$\mathcal{G} = \left\{\sum_{s=1}^{n}\phi(z_s)\phi(z_s)^\top : \phi \in \mathcal{H}^h, \|\phi\| \le 1, n \in \mathbb{N}\right\}.$$

For any $\epsilon > 0$, let $N(\gamma, \mathcal{G}, R)$ denote the γ-covering number of the ball $\mathcal{B}_\mathcal{G}(R) = \{\Lambda \in \mathcal{G} : \|\Lambda\|_{\text{op}} \le R\}$ with respect to the operator norm. Then, its log-covering number with respect to operator norm is bounded by

$$N(\gamma, \mathcal{G}, R) \le (d((\gamma/2R)^2))^2\ln(1+4R\sqrt{d((\gamma/2R)^2)}/\gamma).$$

Proof. First of all, we consider the log-covering number of a RKHS ball. For any $\epsilon > 0$, let $N(\gamma, \mathcal{H}^h, R)$ denote the log-γ-covering number of the ball $\mathcal{B}_{\mathcal{H}^h}(R) = \{\phi \in \mathcal{H}^h : \|\phi\| \le R\}$ with respect to the norm $\|\cdot\|_\mathcal{H}$ on the Hilbert space. We claim that its log-covering number is bounded by

$$N_\phi(\gamma, \mathcal{H}^h, R) \le d((\gamma/2R)^2)\ln(1+4R/\gamma).$$

Recall that $\{\phi_j\}_{j=1}^\infty$ are representations of $\mathcal{H}^h$. There exists an index set $\mathcal{S}$ with the smallest cardinality such that

$$\sup_{z\in\mathcal{Z}}\sum_{j\notin\mathcal{S}}(\phi_j)^2 \le (\frac{\gamma}{2R})^2 := \gamma'.$$

Hence, we get $|\mathcal{S}| = d(\gamma')$. Then, any $\phi \in \mathcal{H}^h$ can be written as $\phi = \sum_{j=1}^\infty \alpha_j\phi_j e_j$, where e_j is the infinite-dimensional vector with only index j equaling 1, we define $\Pi_\mathcal{S} : \mathcal{H}^h \to \mathcal{H}^h$ as the projection map onto the subspace spanned by $\{\phi_j e_j\}_{j\in\mathcal{S}}$:

$$\Pi_\mathcal{S}(\phi) = \sum_{j\in\mathcal{S}}\alpha_j\phi_j e_j.$$

For any $\phi \in \mathcal{B}_{\mathcal{H}^h}(R)$, notice that $\|\alpha\| \le R$, so we have

$$\|\phi - \Pi_\mathcal{S}(\phi)\| \le \sqrt{\sum_{j\notin\mathcal{S}}\alpha_j^2\phi_j^2} \le R\sqrt{\sum_{j\notin\mathcal{S}}\phi_j^2} \le R\sqrt{\gamma'} = \gamma/2.$$

Then, Let $\mathcal{C}(\gamma/2, d(\gamma'), R)$ be the $\gamma/2$-cover of $\{\alpha \in \mathbb{R}^{d(\gamma')} : \|\alpha\| \le R\}$ with respect to the Euclidean norm. Thus, there exists $\tilde{\alpha} \in \mathcal{C}(\gamma/2, d(\gamma'), R)$ such that $\sum_{j\in\mathcal{S}}(\alpha_j - \tilde{\alpha}_j)^2 \le (\gamma/2)^2$. It follows that for any $z \in \mathcal{Z}$,

$$\begin{aligned}\left\|\Pi_\mathcal{S}(\phi) - \sum_{j\in\mathcal{S}}\tilde{\alpha}_j e_j\phi_j\right\| &= \left\|\sum_{j\in\mathcal{S}}(\alpha_j - \tilde{\alpha}_j)e_j\phi_j\right\|\\ &\le \sqrt{\sum_{j\in\mathcal{S}}(\alpha_j - \tilde{\alpha}_j)^2\cdot\phi_j^2}\\ &\le \gamma/2.\end{aligned}$$

Therefore, the following result is obtained:

$$\left\|\phi - \sum_{j\in\mathcal{S}} \tilde{\alpha}_j \phi_j e_j\right\| \le \|\phi - \Pi_{\mathcal{S}}(\phi)\| + \left\|\Pi_{\mathcal{S}}(\phi) - \sum_{j\in\mathcal{S}} \tilde{\alpha}_j \phi_j e_j\right\| \le \gamma,$$

which indicates that

$$N_\phi(\gamma, \mathcal{H}^h, R) \le |\mathcal{C}(\gamma/2, d(\gamma'), R)| \le \ln\left((1+4R/\gamma)^{d(\gamma')}\right) = d(\gamma')\ln(1+4R/\gamma),$$

where the second inequality uses Lemma 5.2 in Vershynin (2010).

Further, let $\mathcal{C}(\gamma/2, d^2(\gamma'), R)$ be the $\gamma/2$-cover of $\{A \in \mathbb{R}^{d(\gamma')\times d(\gamma')} : \|A\|_{\text{op}} \le R\}$ with respect to the Frobenius norm. For a covariance matrix $\Lambda \in \mathcal{G}$, there exists a matrix $A \in \mathcal{C}(\gamma/2, d^2(\gamma'), R)$ with $\|A\|_{\text{op}} \le t$ such that

$$\|\Lambda - A\|_{\text{op}} \le \|\Lambda - A\|_F \le \gamma,$$

where the second inequality is obtained from the claim in the vector case. For any matrix $A \in \mathbb{R}^{d(\gamma')\times d(\gamma')}$ with operator norm bounded by R, its Frobenius norm is bounded by $R\sqrt{d(\gamma')}$. Therefore, we can bound the log-covering number for the matrix space as

$$N(\gamma, \mathcal{G}, R) \le d^2(\gamma')\ln(1+4R\sqrt{d(\gamma')}/\gamma).$$

□

Lemma E.3 (Covering Number of Hilbert Inner Product)**.** *For any $\epsilon > 0$, let $N_w(\gamma, \mathcal{H}^h, R)$ be the log-γ-covering number of the Ball $\{w \in \mathcal{H}^h : \|w\| \le R\}$. Then, supposing that $\|\phi\| \le 1$, we have*

$$N_w(\gamma, \mathcal{H}^h, R) \le d((\gamma/2R)^2)\ln(1+4R/\gamma)$$

Proof. Consider the representations $\{\phi_j\}_{j=1}^\infty$ of $\mathcal{H}$ and the core index set $\mathcal{S}$ such that $\sup_{z\in\mathcal{Z}} \sum_{j\notin\mathcal{S}} (\phi_j)^2 \le (\frac{\gamma}{2R})^2 = \gamma'$. For any $f \in \mathcal{F}^h$, we can write it as $f(z) = \sum_{j=1}^\infty w_j\phi_j(z)$. Define $\Pi_{\mathcal{S}} : \mathcal{H}^h \to \mathcal{H}^h$ as the projection map onto the subspace spanned by $\{\phi_j e_j\}_{j\in\mathcal{S}}$:

$$\Pi_{\mathcal{S}}(f) = \sum_{j\in\mathcal{S}} w_j\phi_j.$$

For any $f \in \mathcal{B}_{\mathcal{H}^h}(R) := \{f \in \mathcal{H}^h : \|f\|_{\mathcal{H}^h} \le R\}$, we have $\|w\| \le R$, which implies that

$$\|f - \Pi_{\mathcal{S}}(f)\|_{\mathcal{H}^h} = \left\|\sum_{j\notin\mathcal{S}} w_j\phi_j\right\|_{\mathcal{H}} \le \sqrt{\sum_{j\notin\mathcal{S}} w_j^2} \cdot \sup_{z\in\mathcal{Z}} \sqrt{\sum_{j\notin\mathcal{S}} (\phi_j(z))^2} \le R\sqrt{\gamma'} = \gamma/2.$$

Now, use $\mathcal{C}(\gamma/2, d(\gamma'), R)$ to denote the $\gamma/2$-cover of $\{w \in \mathbb{R}^{d(\gamma')} : \|w\| \le R\}$ with respect to the Euclidean norm. There exists $\tilde{w} \in \mathcal{C}(\gamma/2, d(\gamma'), R)$ such that $\sum_{j\in\mathcal{S}} (w_j - \tilde{w}_j)^2 \le (\gamma/2)^2$. Therefore, for any $z \in \mathcal{Z}$,

$$\begin{aligned}
\left\|\Pi_{\mathcal{S}}(f) - \sum_{j\in\mathcal{S}} \tilde{w}_j\phi_j\right\|_{\mathcal{H}^h} &= \left\|\sum_{j\in\mathcal{S}} (w_j - \tilde{w}_j)\phi_j\right\|_{\mathcal{H}^h} \\
&\le \sqrt{\sum_{j\in\mathcal{S}} (w_j - \tilde{w}_j)^2} \cdot \sup_{z\in\mathcal{Z}} \sqrt{\sum_{j\notin\mathcal{S}} (\phi_j(z))^2} \\
&\le \gamma/2.
\end{aligned}$$

Therefore, the following result is obtained:

$$\left\|f - \sum_{j\in\mathcal{S}} \tilde{w}_j\phi_j\right\|_{\mathcal{H}^h} \le \|f - \Pi_{\mathcal{S}}(f)\|_{\mathcal{H}^h} + \left\|\Pi_{\mathcal{S}}(f) - \sum_{j\in\mathcal{S}} \tilde{w}_j\phi_j\right\|_{\mathcal{H}^h} \le \gamma.$$

Finally, for any parameter $w = [w_1, w_2, \ldots]$ inducing the function $f = \sum_{j=1}^{\infty} w_j \phi_j$, there exists a vector $\tilde{w} = [\tilde{w}_j]_{j \in \mathcal{S}}$ such that

$$\|w - \tilde{w}\| = \left\| f - \sum_{j \in \mathcal{S}} \tilde{w}_j \phi_j \right\|_{\mathcal{H}^h} \leq \gamma.$$

Then, by invoking Lemma 5.2 in Vershynin (2010), we attain the result:

$$N_w(\gamma, \mathcal{H}^h, R) \leq |\mathcal{C}(\gamma/2, d(\gamma'), R)| \leq \ln(1 + 4R/\gamma)^{d(\gamma')} = d(\gamma') \ln(1 + 4R/\gamma).$$

□

E.2. Stabilization with Subsampling

The main idea of the sub-sampling procedure in Wang et al. (2020) is to compute the bonus by a sub-sampled dataset. This subset enjoys a log-covering number that is upper bounded by the eluder dimension because the number of distinct elements of the subsampled dataset is much smaller than the original one. Meanwhile, the uncertainty estimations are roughly preserved using the sub-sampled dataset. To achieve the goal, we measure the importance of the data points in the dataset and only maintain those critical for the bonus construction (by importance sampling).

However, the extension is not straightforward because the uncertainty-weighted regression brings distinct challenges. Controlling the uncertainty level in the face of weights also plays a central role in our analysis. Before continuing, we additionally assume that the covering number of state-action pairs $\mathcal{X} \times \mathcal{A}$ is bounded, which is also required by Wang et al. (2020); Kong et al. (2021).

Assumption E.4. There exists a γ-cover $\mathcal{C}(\mathcal{X} \times \mathcal{A}, \gamma)$ with bounded size $N(\mathcal{X} \times \mathcal{A}, \gamma)$ such that for any $(x, a) \in \mathcal{X} \times \mathcal{A}$, there exists $(x', a') \in \mathcal{X} \times \mathcal{A}$ satisfying that $\sup_{f \in \mathcal{F}} |f(x, a) - f(x', a')| \leq \epsilon$.

For simplicity, we denote the space $\mathcal{X} \times \mathcal{A}$ by $\mathcal{Z}$. In the sub-sampling algorithm, a core set is generated according to each element's sensitivity, the error on one sample compared to the whole-sample error. We first define

$$\text{Sensitivity}_{\lambda,\sigma,\mathcal{F},\mathcal{Z}}(z) = \min \left\{ 1, \sup_{f,g \in \mathcal{F}} \frac{(f(z) - g(z))^2/\sigma_z^2}{\lambda + \|f - g\|_{\mathcal{Z},\sigma}^2} \right\}.$$

This sensitivity resembles the quantity considered in (2.7) except that the former is compared to the whole sample set, and the latter is compared to the history samples. Thus, we can bound the sum of sensitivities by invoking Lemma 5.1.

Lemma E.5. *For a given set of state-action pairs Z_1^t and the corresponding sequence of weights, we have*

$$\sum_{z \in Z_1^t} \text{Sensitivity}_{\lambda,\sigma,\mathcal{F}_t^h,Z_1^t}(z) \leq \sum_{s=1}^{t} D_{\lambda,\sigma,\mathcal{F}}(Z_1^s)^2 = \tilde{\mathcal{O}}(\dim_E(\mathcal{F}, \lambda/T)).$$

This lemma plays the same role as Lemma 1 in Wang et al. (2020). Let $\hat{\mathcal{Z}}$ denote the core set generated during Algorithm 3 in Wang et al. (2020). We construct the confidence set based on $\hat{\mathcal{Z}}$:

$$\hat{\mathcal{F}}_t^h = \left\{ f^h \in \mathcal{F}_{t-1}^h : \lambda + \|f^h - \bar{f}\|_{\hat{\mathcal{Z}},\sigma}^2 \leq 3\beta^2 + 2 \right\},$$

and the bonus function $\hat{b}_t^h(\cdot) \in \hat{\mathcal{B}}^h(\lambda)$:

$$\hat{b}_t^h(z) = \sup_{f^h \in \hat{\mathcal{F}}_t^h} \frac{|f^h(z) - \hat{f}_t^h(z)|}{\sqrt{\lambda + \|f^h - \bar{f}\|_{\hat{\mathcal{Z}},\sigma}^2}},$$

where $\bar{f} \in \mathcal{F}_\gamma^h$ is close to $\hat{f}_t^h$ in the sense that $\|\bar{f} - \hat{f}_t^h\|_\infty \leq \gamma$. Combining Lemma E.5 and Lemma 2 to 4 in Wang et al. (2020), we obtain the following lemma.

Lemma E.6. *If $|\mathcal{Z}| \leq TH$, the following results hold:*

1. *With probability at least* $1-\delta/(16TH)$*, for all* $h \in [H]$*, we have*

$$b_t^h(z) \le \sqrt{2}\hat{b}_t^h(z).$$

2. *If* T *is sufficiently large,*

$$\ln|\hat{\mathcal{B}}^h(\lambda)| = \tilde{\mathcal{O}}\left(\dim_E(\mathcal{F}^h, \delta/(TH)^3)\cdot \ln(N(\mathcal{F}^h, \delta/(TH)^2)/\delta)\cdot \log(N(\mathcal{X}\times\mathcal{A}, \delta/(TH)))\right).$$

Then, by replacing the bonus $b_t^h(z)$ with $\hat{b}_t^h(z)$ in Algorithm 2, we can bound the regret in Theorem 4.2 by

$$\mathrm{Reg}(T) = \tilde{\mathcal{O}}\left(\sqrt{TH\iota} + \zeta \dim_E(\mathcal{F}, \lambda/T)\right),$$

where

$$\begin{aligned}\iota &= \ln|\hat{\mathcal{B}}^h(\lambda)| \dim_E(\mathcal{F}, \lambda/T)\\ &= \dim_E(\mathcal{F}^h, \delta/(TH)^3)^2 \cdot \ln(N(\mathcal{F}^h, \delta/(TH)^2)/\delta)\cdot \log(N(\mathcal{X}\times\mathcal{A}, \delta/(TH))).\end{aligned}$$

F. Unknown Corruption Level

Proof of Theorem 4.3. We first consider the case when $\zeta \le \bar{\zeta}$. Since only the predetermined parameters α and β are modified by replacing ζ with $\bar{\zeta}$ and ζ is upper bounded by $\bar{\zeta}$, it is easy to obtain the regret bound $\mathrm{Reg}(T) = \tilde{\mathcal{O}}(\sqrt{TH\ln(N_T(\gamma))\dim_E(\mathcal{F}, \lambda/T)} + \bar{\zeta}\dim_E(\mathcal{F}, \lambda/T))$ by following the analysis of Theorem 4.2. Additionally, Lemma 5.1 can also be proved by discussing the value of $\bar{\zeta}$ in the first step.

Then, for the case when $\zeta > \bar{\zeta}$, we simply take the trivial bound T. □

Theorem F.1. *Select* $\alpha = \sqrt{\ln N(\gamma, \mathcal{F}, \|\cdot\|_\infty)}/\bar{\zeta}$*,* $\beta_t^h = c_\beta(\sqrt{\ln(N(\gamma, \mathcal{F}, \|\cdot\|_\infty)/\delta)} + \sqrt{c_0})$ *and keep the other conditions the same as Theorem 4.2. If* $0 \le \zeta \le \bar{\zeta}$*, with probability at least* $1-\delta$*, we have* $\mathrm{Reg}(T) = \tilde{\mathcal{O}}(\sqrt{T\dim_E(\mathcal{F}, \lambda/T)\ln(N(\gamma, \mathcal{F}, \|\cdot\|_\infty))} + \bar{\zeta}\dim_E(\mathcal{F}, \lambda/T))$*. If* $\zeta > \bar{\zeta}$*, we have* $\mathrm{Reg}(T) = \tilde{\mathcal{O}}(T)$*.*

G. Technical Lemmas

Lemma G.1 (Value-decomposition Lemma (Jiang et al., 2017))**.** *Consider any candidate value function* $f = \{f^h(x^h, a^h) : \mathcal{X}\times\mathcal{A} \to \mathbb{R}\}$*, with* $f^{H+1}(\cdot) = 0$*. Let* π_f *be its greedy policy. We have*

$$f^1(x^1) - V^1_{\pi_f}(x^1) = \mathbb{E}_{\pi_f}\left[\sum_{h=1}^H \mathcal{E}^h(f, x^h, a^h)\,\middle|\, x^1\right],$$

where $\mathbb{E}_{\pi_f}[\cdot|x^h]$ *means taking expectation over the trajectory* $\{(x^{h'}, a^{h'})\}_{h'=h}^H$ *induced by policy* π_f *starting from the state* x^h*.*

Lemma G.2. *Let* $\{\epsilon_s\}$ *be a sequence of zero-mean conditional* σ*-sub-Gaussian random variables:* $\ln\mathbb{E}[e^{\lambda\epsilon_i}|\mathcal{S}_{i-1}] \le \lambda^2\sigma^2/2$*, where* $\mathcal{S}_{i-1}$ *represents the history data. We have for* $t \ge 1$*, with probability at least* $1-\delta$*,*

$$\sum_{s=1}^t \epsilon_i^2 \le 2t\sigma^2 + 3\sigma^2\ln(1/\delta).$$

Proof. By invoking the logarithmic moment generating function estimate in Theorem 2.29 from Zhang (2023), we know that for $\lambda \ge 0$,

$$\ln\mathbb{E}\left[\exp\left(\lambda\epsilon_i^2\right)|\,\mathcal{S}_{i-1}\right] \le \lambda\sigma^2 + \frac{(\lambda\sigma^2)^2}{1-2\lambda\sigma^2}. \tag{37}$$

Then, by using iterated expectations due to the tower property of conditional expectation, we get

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{t}\epsilon_i^2\right)\right] &= \mathbb{E}\left\{\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{t-1}\epsilon_i^2+\epsilon_t^2\right)\Big|\,\mathcal{S}_{t-1}\right]\right\} \\
&= \mathbb{E}\left\{\exp\left(\lambda\sum_{i=1}^{t-1}\epsilon_i^2\right)\cdot\mathbb{E}\left[\exp\left(\epsilon_t^2\right)\Big|\,\mathcal{S}_{t-1}\right]\right\} \\
&\le \exp\left(\lambda\sigma^2+\frac{(\lambda\sigma^2)^2}{1-2\lambda\sigma^2}\right)\cdot\mathbb{E}\left\{\exp\left(\lambda\sum_{i=1}^{t-1}\epsilon_i^2\right)\right\} \\
\ldots &\le \exp\left(\lambda t\sigma^2+\frac{(\lambda t\sigma^2)^2}{1-2\lambda\sigma^2}\right),
\end{aligned}$$

where the first ineqaulity uses (37). Now, we can apply the second ineqaulity of Lemma 2.9 from (Zhang, 2023) with $\mu = t\sigma^2, \alpha = 2t\sigma^4, \beta = 2\sigma^2$ and $\epsilon = 2\sigma^2\sqrt{ut}$ to obtain

$$\inf_{\lambda\ge 0}\left\{-\lambda\big(t\sigma^2+2\sqrt{ut\sigma^4}+2u\sigma^2\big)+\ln\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{t}\epsilon_i^2\right)\right]\right\}\le -u. \tag{38}$$

Thus, it follows that

$$\begin{aligned}
&\mathbb{P}\left(\sum_{s=1}^{t}\epsilon_i^2\le t\sigma^2+2\sqrt{ut\sigma^4}+2u\sigma^2\right) \\
&\le \inf_{\lambda\ge 0}\frac{\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{t}\epsilon_i^2\right)\right]}{\exp\left(\lambda(t\sigma^2+2\sqrt{ut\sigma^4}+2u\sigma^2)\right)} \\
&= \inf_{\lambda\ge 0}\exp\left(-\lambda\big(t\sigma^2+2\sqrt{ut\sigma^4}+2u\sigma^2\big)+\ln\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{t}\epsilon_i^2\right)\right]\right) \\
&\le e^{-u},
\end{aligned}$$

where the first inequality applies Markov's Inequality, and the second inequality uses (38) and the monotonicity of the exponential function.

Taking $u = \ln(1/\delta)$ for $\delta > 0$, we obtain that with probability at least $1-\delta$

$$\begin{aligned}
\sum_{s=1}^{t}\epsilon_i^2 &\le t\sigma^2+2\sqrt{t\ln(1/\delta)\sigma^4}+2\ln(1/\delta)\sigma^2 \\
&\le 2t\sigma^2+3\sigma^2\ln(1/\delta),
\end{aligned}$$

where the second inequality is deduced since $2\sqrt{t\ln(1/\delta)\sigma^4}\le t\sigma^2+\ln(1/\delta)\sigma^2$. □

Lemma G.3. *Consider a sequence of random variables $\{Z_t\}_{t\in\mathbb{N}}$ adapted to the filtration $\{\mathcal{S}_t\}_{t\in\mathbb{N}}$ and a function $f\in\mathcal{F}$. For any $\lambda > 0$, with probability at least $1-\delta$, for all $t\ge 1$, we have*

$$-\sum_{s=1}^{t}f(Z_s)-\frac{1}{\lambda}\sum_{s=1}^{t}\ln\mathbb{E}[e^{-\lambda f(Z_s)}|\mathcal{S}_{s-1}]\le\frac{1}{\lambda\delta}$$

Proof. The proof of this lemma is presented in Lemma 4 of Russo & Van Roy (2013). □

Then, we demonstrate the convergence of training error for both ERM and approximate ERM solutions in the following lemma.

Lemma G.4. *Consider a function space $\mathcal{F} : \mathcal{Z} \to \mathbb{R}$ and filtered sequence $\{z_t, \epsilon_t\}$ in $\mathcal{X} \times \mathbb{R}$ so that ϵ_t is conditional zero-mean η-sub-Gaussian noise. For $f_*(\cdot) : \mathcal{Z} \to \mathbb{R}$, suppose that $y_t = f_*(z_t) + \epsilon_t$ and there exists a function $f_b \in \mathcal{F}$ such that for any $t \in [T]$, $\sum_{s=1}^t |f_*(z_s) - f_b(z_s)| := \sum_{s=1}^t \zeta_s \le \zeta$. If $\hat{f}_t$ is an (approximate) ERM solution for some $\epsilon' \ge 0$:*

$$\left(\sum_{s=1}^t (\hat{f}_t(z_s) - y_s)^2/\sigma_s^2\right)^{1/2} \le \min_{f \in \mathcal{F}_{t-1}} \left(\sum_{s=1}^t (f(z_s) - y_s)^2/\sigma_s^2\right)^{1/2} + \sqrt{t}\epsilon',$$

with probability at least $1 - \delta$, we have for all $t \in [T]$:

$$\begin{aligned}\sum_{s=1}^t (\hat{f}_t(z_s) - f_b(z_s))^2/\sigma_s^2 \le & 10\eta^2 \ln(2N(\gamma, \mathcal{F}, \|\cdot\|_\infty)/\delta) + 5\sum_{s=1}^t |\hat{f}_t(z_s) - f_b(z_s)|\zeta_s/\sigma_s^2 \\ & + 10(\gamma + \epsilon')((\gamma + \epsilon')t + \sqrt{tC_1(t, \zeta)}),\end{aligned}$$

where $C_1(t, \zeta) = 2(\zeta^2 + 2t\eta^2 + 3\eta^2 \ln(2/\delta))$.

Proof. For $f \in \mathcal{F}$, define

$$\phi(f, z_t) = -a\left[(f(z_t) - y_t)^2 - (f_b(z_t) - y_t)^2\right]/\sigma_t^2,$$

where $a = \eta^{-2}/4$. Let $\mathcal{F}_\gamma$ be an γ-cover of $\mathcal{F}$ in the sup-norm. Denote the cardinality of $\mathcal{F}_\gamma$ by $N = N(\gamma, \mathcal{F}, \|\cdot\|_\infty)$.

Since ϵ_t is conditional η-sub-Gaussian and $\phi(f, z_t)$ can be written as

$$\phi(f, z_t) = 2a(f(z_t) - f_b(z_t))/\sigma_t^2 \cdot \epsilon_t - a(f(z_t) - f_b(z_t))^2/\sigma_t^2 + 2a(f(z_t) - f_b(z_t))\zeta_t/\sigma_t^2,$$

the $\phi(f, z_t)$ is conditional $2a(f(z_t) - f_b(z_t))\eta/\sigma_t^2$-sub-Gaussian with mean

$$\mu = -a(f(z_t) - f_b(z_t))^2/\sigma_t^2 + 2a(f(z_t) - f_b(z_t))\zeta_t/\sigma_t^2,$$

where $a = \eta^{-2}/4$. We know that if a variable X is σ-sub-Gaussian with mean μ conditional on $\mathcal{S}$, the property of sub-Gaussianity implies that

$$\ln \mathbb{E}\big[\exp(s(X - \mu)) \,\big|\, \mathcal{S}\big] \le \frac{\sigma^2 s^2}{2}.$$

By taking $s = 1$ in the inequality above, we get

$$\ln \mathbb{E}_{y_t}[\exp(\phi(f, z_t) - \mu) \,|\, z_t, \mathcal{S}_{t-1}] \le \frac{4a^2(f(z_t) - f_b(z_t))^2\eta^2}{2\sigma_t^4} = \frac{(f(z_t) - f_b(z_t))^2}{8\eta^2\sigma_t^4}.$$

It follows that

$$\begin{aligned}\ln \mathbb{E}_{y_t}[\exp(\phi(f, z_t)) \,|\, z_t, \mathcal{S}_{t-1}] &\le \frac{(f(z_t) - f_b(z_t))^2}{8\eta^2\sigma_t^4} - \frac{(f(z_t) - f_b(z_t))^2}{4\eta^2\sigma_t^2} + \frac{(f(z_t) - f_b(z_t))\zeta_t}{2\eta^2\sigma_t^2} \\ &\le \frac{(f(z_t) - f_b(z_t))^2}{8\eta^2\sigma_t^2} - \frac{(f(z_t) - f_b(z_t))^2}{4\eta^2\sigma_t^2} + \frac{(f(z_t) - f_b(z_t))\zeta_t}{2\eta^2\sigma_t^2} \\ &\le -\frac{(f(z_t) - f_b(z_t))^2}{8\eta^2\sigma_t^2} + \frac{(f(z_t) - f_b(z_t))\zeta_t}{2\eta^2\sigma_t^2},\end{aligned}$$

where the second inequality uses $\sigma_t^2 \ge 1$.

Invoking Lemma G.3 with $\lambda = 1$, we have for all $f \in \mathcal{F}_\gamma$ and $t \in [T]$, with probability at least $1 - \delta/2$,

$$\sum_{s=1}^t \phi(f, z_s) \le -\sum_{s=1}^t \frac{(f(z_s) - f_b(z_s))^2}{8\eta^2\sigma_s^2} + \sum_{s=1}^t \frac{(f(z_s) - f_b(z_s))\zeta_s}{2\eta^2\sigma_s^2} + \ln(2N/\delta). \tag{39}$$

Additionally, for all $t \in [T]$, we have with probability at least $1 - \delta/2$,

$$\begin{aligned}
\sum_{s=1}^{t}(f_b(z_s) - y_s)^2 &= \sum_{s=1}^{t}(f_b(z_s) - f_*(z_s) + f_*(z_s) - y_s)^2 \\
&\leq 2\sum_{s=1}^{t-1}\left((f_b(z_s) - f_*(z_s))^2 + (f_*(z_s) - y_s)^2\right) \\
&\leq 2\left(\sum_{s=1}^{t-1}\zeta_s^2 + \sum_{s=1}^{t-1}\epsilon_s^2\right) \\
&\leq 2\left(\zeta^2 + 2t\eta^2 + 3\eta^2\ln(2/\delta)\right) := C_1(t,\zeta),
\end{aligned} \tag{40}$$

where the first inequality is obtained since Cauchy-Schwarz inequality, the second inequality uses $|f_b(z_s) - f_*(z_s)| \leq \zeta_s$, and the last inequality comes from Lemma G.2. To simplify notations, we use $C_1(t,\zeta)$ to denote $2(\zeta^2 + 2t\eta^2 + 3\eta^2\ln(2/\delta))$. Now, given $\hat{f}_t$, there exists $f \in \mathcal{F}_\gamma$ so that $\|\hat{f}_t - f\|_\infty \leq \gamma$. With probability at least $1 - \delta/2$,

$$\begin{aligned}
&\sum_{s=1}^{t}\left[(f(z_s) - y_s)^2 - (f_b(z_s) - y_s)^2\right]/\sigma_s^2 \\
&\leq \left(\sqrt{\sum_{s=1}^{t}(\hat{f}_t(z_s) - y_s)^2/\sigma_s^2} + \sqrt{t}\gamma\right)^2 - \sum_{s=1}^{t}(f_b(z_s) - y_s)^2/\sigma_s^2 \\
&\leq \left(\sqrt{\sum_{s=1}^{t}(f_b(z_s) - y_s)^2/\sigma_s^2} + \sqrt{t}(\gamma + \epsilon')\right)^2 - \sum_{s=1}^{t}(f_b(z_s) - y_s)^2/\sigma_s^2 \\
&\leq (\gamma + \epsilon')^2 t + 2(\gamma + \epsilon')\sqrt{tC_1(t,\zeta)},
\end{aligned} \tag{41}$$

where the first inequality uses $|f(z_s) - \hat{f}_t(z_s)| \leq \gamma$ and triangle inequality for all s, the second inequality is because of the condition that $\hat{f}_t$ is an approximate ERM solution, and the last inequality utilizes $\sigma_s \geq 1$ and (40). Finally, with probability at least $1 - \delta$, we get

$$\begin{aligned}
&\left(\sum_{s=1}^{t}(\hat{f}_t(z_s) - f_b(z_s))^2/\sigma_s^2\right)^{1/2} \\
&\leq \sqrt{\gamma^2 t} + \left(\sum_{s=1}^{t}(f(z_s) - f_b(z_s))^2/\sigma_s^2\right)^{1/2} \\
&\leq \sqrt{\gamma^2 t} + \left(4\sum_{s=1}^{t}(f(z_s) - f_b(z_s))\zeta_s/\sigma_s^2 + 8\eta^2\ln(2N/\delta) - 8\eta^2\sum_{s=1}^{t}\phi(f, z_s)\right)^{1/2} \\
&\leq \sqrt{\gamma^2 t} + \left(4\sum_{s=1}^{t}|\hat{f}_t(z_s) - f_b(z_s)|\zeta_s/\sigma_s^2 + 4\gamma\zeta + 8\eta^2\ln(2N/\delta) + 2(\gamma + \epsilon')^2 t\right. \\
&\quad \left. + 4(\gamma + \epsilon')\sqrt{t}C_1'(t,\zeta)\right)^{1/2} \\
&\leq \left(10\eta^2\ln(2N/\delta) + 5\sum_{s=1}^{t}|\hat{f}_t(z_s) - f_b(z_s)|\zeta_s/\sigma_s^2 + 5\gamma\zeta + 8(\gamma + \epsilon')^2 t + 5(\gamma + \epsilon')\sqrt{tC_1(t,\zeta)}\right)^{1/2},
\end{aligned}$$

where the second inequality is deduced from (39), the third inequality is by the definition of σ_s in (14) and the last inequality uses Cauchy-Schwarz inequality. □