Corruption-Robust Offline Reinforcement Learning with General Function Approximation

Chenlu Ye[*]
The Hong Kong University of Science and Technology
cyeab@connect.ust.hk

Rui Yang[*]
The Hong Kong University of Science and Technology
ryangam@connect.ust.hk

Quanquan Gu
University of California, Los Angeles
qgu@cs.ucla.edu

Tong Zhang
The Hong Kong University of Science and Technology
tongzhang@ust.hk

Abstract

We investigate the problem of corruption robustness in offline reinforcement learning (RL) with general function approximation, where an adversary can corrupt each sample in the offline dataset, and the corruption level $\zeta \geq 0$ quantifies the cumulative corruption amount over n episodes and H steps. Our goal is to find a policy that is robust to such corruption and minimizes the suboptimality gap with respect to the optimal policy for the uncorrupted Markov decision processes (MDPs). Drawing inspiration from the uncertainty-weighting technique from the robust online RL setting [18, 55], we design a new uncertainty weight iteration procedure to efficiently compute on batched samples and propose a corruption-robust algorithm for offline RL. Notably, under the assumption of single policy coverage and the knowledge of ζ, our proposed algorithm achieves a suboptimality bound that is worsened by an additive factor of $\mathcal{O}(\zeta \cdot (\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H))^{1/2}(C(\widehat{\mathcal{F}}, \mu))^{-1/2}n^{-1})$ due to the corruption. Here $\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H)$ is the coverage coefficient that depends on the regularization parameter λ, the confidence set $\widehat{\mathcal{F}}$, and the dataset $\mathcal{Z}_n^H$, and $C(\widehat{\mathcal{F}}, \mu)$ is a coefficient that depends on $\widehat{\mathcal{F}}$ and the underlying data distribution μ. When specialized to linear MDPs, the corruption-dependent error term reduces to $\mathcal{O}(\zeta d n^{-1})$ with d being the dimension of the feature map, which matches the existing lower bound for corrupted linear MDPs. This suggests that our analysis is tight in terms of the corruption-dependent term.

1 Introduction

Offline reinforcement learning (RL) has received tremendous attention recently because it can tackle the limitations of online RL in real-world applications, e.g., healthcare [45] and autonomous driving [35], where collecting online data is risky, expensive and even infeasible. By leveraging a batch of pre-collected datasets, offline RL aims to find the optimal policy that is covered by the dataset without further interaction with the environment. Due to the restriction of the offline dataset, the utilization of pessimism in the face of uncertainty is widespread [1, 2, 53, 40, 16] and plays a central role in

[*] These authors contributed equally to this work.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

providing theoretical guarantees for efficient learning [24, 36, 44, 48, 51, 56, 57, 63, 52]. Notably, these theoretical works demonstrate that a single-policy coverage is sufficient to guarantee sample efficiency.

In this paper, we study offline RL under adversarial corruption and with general function approximation. Adversarial corruption refers to adversarial attacks on the reward functions and transition dynamics on the data at each step before the learner can access the dataset. The learner only knows the cumulative corruption level and cannot tell whether the corruption occurs at each data point. Our corruption formulation subsumes the model misspecification [22] and the fixed fraction of data contamination [60] as special cases. Various real-world problems are under the threat of adversarial corruption, such as chatbots misled by discriminative or unethical conversations [32, 61], and autonomous vehicles tricked by hacked navigation instructions or deliberately contaminated traffic signs [12]. On the other hand, general function approximation (approximating the value function with a nonlinear function class, such as deep neural networks) plays a pivotal role in modern large-scale RL problems, such as large language model [10], robotics [26] and medical treatment [28]. Recent works have established different frameworks to explore the minimal structure condition for the function class that enables sample efficiency [19, 47, 9, 21, 14, 6, 64]. In particular, Wang et al. [47] leverage the concept of eluder dimension [39] and construct the least squares value iteration (LSVI) framework, which establishes optimism at each step for online RL.

For adversarial corruption and general function approximation, a significant amount of research has focused on the online setting. However, offline RL in this setting is still understudied due to restricted coverage conditions and lack of adaptivity. One notable exception is Zhang et al. [60], which assumes ϵ-fraction of the offline dataset is corrupted, and their algorithm suffers from a suboptimal bound on the corruption term. Our work moves a step further and achieves corruption robustness under the LSVI framework [47] in offline RL with general function approximation by generalizing the uncertainty weighting technique [18, 55]. We also propose an algorithm robust to an additional known distribution shift. Due to space limit, we defer it to Appendix C. We summarize our contributions as follows:

- We formally define the corruption level in offline RL. With knowledge of the corruption level, we design an algorithm that draws inspiration from the PEVI algorithm [24] and the uncertainty-weighting technique. The uncertainty for each data point, serving as the bonus function, is quantified by its informativeness with respect to the whole dataset. We propose the uncertainty weight iteration algorithm to calculate the weights efficiently and prove that the output, an approximation of the uncertainty, is sufficient to control the corruption term.
- Theoretically, our proposed algorithm enjoys a suboptimality bound of $\tilde{\mathcal{O}}(H \cdot \mathrm{CC}^{1/4}(\ln N)^{1/2}(C(\widehat{\mathcal{F}},\mu))^{-1/4}n^{-1/2} + \zeta \cdot \mathrm{CC}^{1/2}(C(\widehat{\mathcal{F}},\mu))^{-1/2}n^{-1})$, where H is the episode length, n is the number of episodes, N is the covering number, CC is the coverage coefficient, and $C(\widehat{\mathcal{F}},\mu)$ is the coefficient depicting how well the underlying data distribution μ explores the feature space, and $\widehat{\mathcal{F}}$ is the confidence set. The corruption-dependent term reduces to $\mathcal{O}(\zeta d n^{-1})$ in the linear model of dimension d, thus matching the lower bound for corrupted linear MDPs. It is worth highlighting that our novel analysis enables us to eliminate the uncertainty-related weights from the coverage condition.
- Motivated by our theoretical findings, we present a practical offline RL algorithm with uncertainty weighting and demonstrate its efficacy under diverse data corruption scenarios. Our practical implementation achieves a 104% improvement over the previous state-of-the-art uncertainty-based offline RL algorithm under data corruption, demonstrating its potential for effective deployment in real-world applications.

1.1 Related Work

Corruption-Robust Bandits and RL. There is an emerging body of theoretical literature on bandits and online RL with corruption. The adversarial corruption is first formulated in the multi-armed bandit problem by Lykouris et al. [30], where an adversary corrupts the reward in each round t by ζ_t and the corruption level is measured by $\zeta = \sum_{t=1}^{T}|\zeta_t|$. Then, a lower bound with a linear dependence on ζ is constructed by Gupta et al. [17], indicating that the ideal regret bound should achieve a "parallel" relationship: $\mathrm{Regret}(T) = o(T) + \mathcal{O}(\zeta)$, and the corruption-independent term approximates the non-corrupted bound. When extending to linear contextual bandits, a line of work [4, 8, 13, 27, 62, 25] propose various methods but either derive sub-optimal regrets or require particular assumptions. The gap is later closed by He et al. [18], which achieves the minimax lower bound using a novel sample-dependent weighting technique. Specifically, the weight for each sample is adaptive to its confidence, which is also called uncertainty. Beyond bandits, earlier works on MDPs [5, 20, 23, 29, 33, 37, 38] consider the setting where only the rewards are corrupted, and the transitions remain intact. Wu et al. [50] begin to handle corruption on both rewards and transitions for tabular MDPs. Wei et al. [48] establish a unified framework for RL with unknown corruption under a weak adversary, where the corruption happens before the decision is made in each round. Later, Ye et al. [55] extend the weighting technique [18] to corrupted RL with general function approximation and achieve a linear dependence on the cumulative corruption level ζ. Particularly, Wei et al. [48], Ye et al. [55] both impose corruption on the Bellman operator, which is the same as the corruption model considered in this paper.

Offline RL Against Attacks. The emergence of poisoning attacks in real-world scenarios poses new challenges for offline RL and necessitates improved defenses [49]. There are generally two types of attacks [3], namely test-time attacks and training-time attacks. In test-time attacks, the training data is clean, and the learned policy must contend with an attacker during test time. For example, Yang et al. [53] propose learning conservative and smooth policies robust to different test-time attacks. In contrast, our paper focuses on the training-time attack as another line of work [31, 49, 60], where part of the training data is corrupted maliciously. Wu et al. [49] propose two certification criteria and a new aggregation-based method to improve the learned policy from corrupted data. To the best of our knowledge, [60] is the only theoretical work on corrupted offline RL, which considers that an ϵ-fraction ($\epsilon = \zeta/nH$) of samples are corrupted on both rewards and transitions for linear MDPs and achieves an $\mathcal{O}(\zeta^{1/2}dn^{-1/2})$ suboptimality bound. Notably, distinct from the setting in Zhang et al. [60] that clean data is first collected and then corrupted by an adversary, we consider the setting that data collection and corruption occur at the same time (thus corruption at one step affects the subsequent trajectory). Therefore, our setting is different from that of Zhang et al. [60].

2 Preliminaries

In this section, we formulate the episodic Markov decision process (MDP) with adversarial corruption and under general (nonlinear) function approximation. Before the formal introduction, we introduce some notations to facilitate our presentation.

Notations. Let $[n]$ denote the set $\{1, \dots, n\}$. For spaces $\mathcal{X}$ and $\mathcal{A}$ and a function $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, let $f(x) = \max_{a\in\mathcal{A}} f(x, a)$. Given a semi-definite matrix M and a vector v, we define $\|v\|_M = \sqrt{v^\top Mv}$. For two positive sequences $\{f(n)\}_{n=1}^{\infty}$, $\{g(n)\}_{n=1}^{\infty}$, let $f(n) = \mathcal{O}(g(n))$ if there exists a constant $C > 0$ such that $f(n) \leq Cg(n)$ for all $n \geq 1$, and $f(n) = \Omega(g(n))$ if there exists a constant $C > 0$ such that $f(n) \geq Cg(n)$ for all $n \geq 1$. We use $\tilde{\mathcal{O}}(\cdot)$ to omit polylogarithmic factors. Sometimes we use the shorthand notation $z = (x, a)$.

2.1 Episodic MDPs

We consider an episodic MDP $(\mathcal{X}, \mathcal{A}, H, \mathbb{P}, r)$ with the state space $\mathcal{X}$, action space $\mathcal{A}$, episode length H, transition kernel $\mathbb{P} = \{\mathbb{P}^h\}_{h\in[H]}$, and reward function $r = \{r^h\}_{h\in[H]}$. Suppose that the rewards are bounded: $r^h \geq 0$ for any $h \in [H]$, and $\sum_{h=1}^{H} r^h(x^h, a^h) \leq 1$ almost surely. Given any policy $\pi = \{\pi^h : \mathcal{X} \to \mathcal{A}\}_{h\in[H]}$, we define the Q-value and V-value functions starting from step h as

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

where the expectation $\mathbb{E}_\pi$ is taken with respect to the trajectory under the policy π. There exists an optimal policy π_* and optimal value functions $V_*^h(x) := V_{\pi^*}^h(x) = \sup_\pi V_\pi^h(x)$ and $Q_*^h(x, a) := Q_{\pi^*}^h(x, a) = \sup_\pi Q_\pi^h(x, a)$ that satisfy the Bellman optimality equation:

$$Q_*^h(x, a) = \mathbb{E}_{r^h, x^{h+1}}\big[r^h(s, a) + \max_{a'\in\mathcal{A}} Q_*^{h+1}(x^{h+1}, a') \,|\, x, a\big] := (\mathcal{T}^h Q_*^{h+1})(x, a), \tag{2}$$

where $\mathcal{T}^h$ is called the Bellman operator. Then we define the Bellman residual as

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

2.2 General Function Approximation

We approximate the Q-value functions by a function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$ where $\mathcal{F}_h : \mathcal{X} \times \mathcal{A} \to [0,1]$ for $h \in [H]$, and $f_{H+1} \equiv 0$ since no reward is generated at step $H+1$. Generally, the following assumption is common for the approximation function class.

Assumption 2.1 (Realizability and Completeness)**.** *For all $h \in [H]$, $Q_*^h \in \mathcal{F}^h$. Additionally, for all $g^{h+1}(x^{h+1}) \in [0,1]$, $(\mathcal{T}^h g^{h+1})(x^h, a^h) \in \mathcal{F}^h$.*

The realizability assumption [21] ensures the possibility of learning the true Q-value function by considering the function class $\mathcal{F}$. The Bellman completeness (adopted from Wang et al. [47] and Assumption 18.22 of Zhang [59]) is stronger than that in Jin et al. [21]. The former applies the least squares value iteration (LSVI) algorithm that establishes optimism at each step, while the latter proposes the GOLF algorithm that only establishes optimism at the first step. We use the standard covering number to depict the scale of the function class $\mathcal{F}$.

Definition 2.1 (ϵ-Covering Number)**.** *The ϵ-covering number $N(\epsilon, \mathcal{F}, \rho)$ of a set $\mathcal{F}$ under metric ρ is the smallest cardinality of a subset $\mathcal{F}_0 \subseteq \mathcal{F}$ such that for any $f \in \mathcal{F}$, there exists a $g \in \mathcal{F}_0$ satisfying that $\rho(f, g) \le \epsilon$. We say $\mathcal{F}_0$ is an (ϵ, ρ) cover of $\mathcal{F}$.*

2.3 Offline Data Collection Process

Offline Clean Data. Consider an offline clean dataset with n trajectories $\mathcal{D} = \{(x_i^h, a_i^h, r_i^h)\}_{i,h=1}^{n,H}$. We assume the dataset $\mathcal{D}$ is compliant with an MDP $(\mathcal{X}, \mathcal{A}, H, \mathbb{P}, r)$ with the value functions Q, V and the Bellman operator $\mathcal{T}$: for any policy π,

$$\begin{aligned} &\mathbb{P}\big((\mathcal{T}^h Q_\pi^{h+1})(x_i^h, a_i^h) = r' + V_\pi^{h+1}(x') \big| \{(x_j^h, a_j^h)\}_{j\in[i]}, \{r_j^h, x_j^{h+1}\}_{j\in[i-1]}, Q_\pi^{h+1}\big) \\ &\qquad = \mathbb{P}\big(r^h(x^h, a^h) = r', x^{h+1} = x' \big| x^h = x_i^h, a^h = a_i^h\big), \end{aligned} \tag{4}$$

where the realizability and completeness in Assumption 2.1 hold. The compliance assumption (4) is also made in Jin et al. [20], Zhong et al. [63], which means that $\mathcal{D}$ remains the Markov property and allows $\mathcal{D}$ to be collected by an adaptive behavior policy. The induced distribution of the state-action pair is denoted by $\mu = \{\mu^h\}_{h\in[H]}$.

Adversarial Corruption. During the offline dataset collection process, after observing the state-action pair (x^h, a^h) chosen by the data collector, an adversary corrupts r^h and x^{h+1} at each step h before they are revealed to the collector. For the corrupted dataset $\mathcal{D}$, we define the corrupted value function $Q_\mathcal{D}$, $V_\mathcal{D}$, and the Bellman operator $\mathcal{T}_\mathcal{D}$ satisfying (4). To measure the corruption level, we notice that characterizing the specific modification on each tuple (s, a, s', r) is hard and unnecessary since once one modifies a tuple, the subsequent trajectory changes. Therefore, it is difficult to tell whether the change is caused by the corruption at the current step or a previous step. In fact, we only care about the part of the change that violates the Bellman completeness. Therefore, following the online setting [55, 48], we measure the corruption level by the gap between $\mathcal{T}_\mathcal{D}$ and $\mathcal{T}$ as follows.

Definition 2.2 (Cumulative Corruption)**.** *The cumulative corruption is ζ if at any step $h \in [H]$, for a sequence $\{(x_i^h, a_i^h)\}_{i,h=1}^{n,H} \subset \mathcal{X} \times \mathcal{A}$ chosen by the data collector and a sequence of functions $\{g^h : \mathcal{X} \to [0,1]\}_{h=1}^H$, we have for all $h \in [H]$,*

$$\sum_{i=1}^n |\zeta_i^h| \le \zeta^h, \quad \sum_{h=1}^H \zeta^h := \zeta,$$

where $\zeta_i^h = (\mathcal{T}^h g^{h+1} - \mathcal{T}_\mathcal{D}^h g^{h+1})(x_i^h, a_i^h)$.

The **learning objective** is to find a policy π that minimizes the suboptimality of π given any initial state $x^1 = x$: $\mathrm{SubOpt}(\pi, x) = V_*^1(x) - V_\pi^1(x)$, where $V(\cdot)$ is the value function induced by the uncorrupted MDP.

3 Algorithm

In this section, we first highlight the pivotal role that uncertainty weighting plays in controlling the corruption-related bound. To extend the uncertainty weighting technique to the offline setting, we propose an iteration algorithm. With the proposed algorithm, the theoretical result for the suboptimality is presented.

3.1 Uncertainty-Related Weights

In this subsection, we discuss the choice of weight for a simplified model without state transition ($H = 1$) and use the notation $z_i = (x_i, a_i)$. Given a dataset $\{(z_i, y_i)\}_{i\in[n]}$, we have $y_i = \bar{f}(z_i) + \zeta_i + \epsilon_i$ for $i \in [n]$, where $\bar{f} \in \mathcal{F}$ is the uncorrupted true value, the noise ϵ_i is zero-mean and conditional η-subGaussian, and the corruption level is $\zeta = \sum_{i=1}^n |\zeta_i|$.

We begin with delineating the consequence caused by the adversarial corruption for the traditional least-square regression: $\hat{f} = \min_{f\in\mathcal{F}} \sum_{i=1}^n \big(f(z_i) - y_i\big)^2$. Some calculations lead to the following decomposition:

$$\sum_{i=1}^n (\hat{f}(z_i) - \bar{f}(z_i))^2 = \underbrace{\sum_{i=1}^n \left[(\hat{f}(z_i) - y_i)^2 - (\bar{f}(z_i) - y_i)^2\right]}_{I_1 \le 0} + 2\underbrace{\sum_{i=1}^n (\hat{f}(z_i) - \bar{f}(z_i))\epsilon_i}_{I_2\text{:Noise term}} + 2\underbrace{\sum_{i=1}^n (\hat{f}(z_i) - \bar{f}(z_i))\zeta_i}_{I_3\text{:Corruption term}}.$$

The term $I_1 \le 0$ since $\hat{f}$ is the solution to the least-square regression. The term I_2 is bounded by $\tilde{\mathcal{O}}(\ln N)$ because of the η-subGaussainity of ϵ_i, where N is the covering number of $\mathcal{F}$. The term I_3 is ruined by corruption: $I_3 \le 2\sum_{i=1}^n |\zeta_i| = \mathcal{O}(\zeta)$. Hence, the confidence radius $(\sum_{i=1}^n(\hat{f}(z_i) - \bar{f}(z_i))^2)^{1/2} = \tilde{\mathcal{O}}(\sqrt{\zeta + \ln N})$ will explode whenever the corruption level ζ grows with n.

To control the corruption term, motivated by the uncertainty-weighting technique from online settings [55, 18, 65], we apply the weighted regression: $\hat{f} = \min_{f\in\mathcal{F}} \sum_{i=1}^n \big(f(z_i) - y_i\big)^2/\sigma_i^2$, where ideally, we desire the following uncertainty-related weights:

$$\sigma_i^2 = \max\left(1, \frac{1}{\alpha}\underbrace{\sup_{f,f'\in\mathcal{F}} \frac{|f(z_i) - f'(z_i)|}{\sqrt{\lambda + \sum_{j=1}^n (f(z_j) - f'(z_j))^2/\sigma_j^2}}}_{\text{Uncertainty}}\right), \quad i = 1, \dots, n, \tag{5}$$

where $\alpha, \lambda > 0$ are pre-determined parameters. The uncertainty quantity in the above equation is the supremum of the ratio between the prediction error $|f(z_i) - f'(z_i)|$ and the training error $\sqrt{\sum_{j=1}^n (f(z_j) - f'(z_j))^2/\sigma_j^2}$ over $f, f' \in \mathcal{F}$. Intuitively, the quantity depicts the relative information of a sample z_i against the whole training set $\{z_1, \dots, z_n\}$. We can use the linear function class as a special example to explain it. When the function space $\mathcal{F}^h$ is embedded into a d-dimensional vector space: $\mathcal{F}^h = \{\langle w(f), \phi(\cdot)\rangle : z \to \mathcal{R}\}$, the uncertainty quantity becomes

$$\begin{aligned}\sup_{f,f'\in\mathcal{F}} \frac{|\langle w(f) - w(f'), \phi(z_i)\rangle|}{\sqrt{\lambda + \sum_{j=1}^n \big(\langle w(f) - w(f'), \phi(z_j)\rangle\big)^2/\sigma_j^2}} &\le \sup_{f,f'\in\mathcal{F}} \frac{|\langle w(f) - w(f'), \phi(z_i)\rangle|}{\sqrt{\big(w(f) - w(f')\big)^\top \Lambda \big(w(f) - w(f')\big)}} \\ &\le \sqrt{\phi^\top(z_i)\Lambda^{-1}\phi(z_i)},\end{aligned}$$

where $\Lambda = \sum_{j=1}^n \phi(z_j)\phi^\top(z_j)/\sigma_j^2$. Moreover, $\big(\phi^\top(z_i)\Lambda^{-1}\phi(z_i)\big)^{-1}$ represents the effective number of samples in the $\{z_i\}_{i=1}^n$ along the $\phi(z_i)$'s direction. We discuss in Lemma B.3 that under mild conditions the linear and nonlinear uncertainty quantities are almost equivalent.

However, since the uncertainty also depends on weights, it is impossible to determine all the weights $\{\sigma_i\}_{i\in[n]}$ simultaneously. Compared with the online setting where the weight in each round can be determined sequentially (iteratively in rounds), we face two challenges in the offline setting: (a) how to compute uncertainty-related weights iteratively? (b) will an approximate solution to the uncertainty play an equivalent role in controlling the corruption term?

To solve the first challenge, we propose the weight iteration algorithm in Algorithm 1. Moreover, we demonstrate the convergence of this algorithm by the monotone convergence theorem in the following lemma, which ensures that the output weights are sufficiently close to desired ones (5). The proof is provided in Appendix B.1.

Lemma 3.1. *There exists a T such that the output of Algorithm 1 $\{\sigma_i := \sigma_i^{T+1}\}_{i=1}^n$ satisfy:*

$$\sigma_i^2 \ge \max\big(1, \psi(z_i)/2\big), \quad \sigma_i^2 \le \max\big(1, \psi(z_i)\big), \tag{6}$$

where $\psi(z_i) = \sup_{f,f'\in\mathcal{F}} \frac{|f(z_i) - f'(z_i)|/\alpha}{\sqrt{\lambda + \sum_{j=1}^n (f(z_j) - f'(z_j))^2/\sigma_j^2}}$.

Algorithm 1 Uncertainty Weight Iteration

```
**Input:** $\{(x_i, a_i)\}_{i=1}^n, \mathcal{F}, \alpha > 0$
**Initialization:** $t = 0,\ \sigma_i^0 = 1, i = 1, \ldots, n$
**repeat**
    $t \leftarrow t + 1$
    $(\sigma_i^t)^2 \leftarrow \max\left(1, \sup_{f,f'\in\mathcal{F}} \frac{|f(x_i,a_i)-f'(x_i,a_i)|/\alpha}{\sqrt{\lambda+\sum_{j=1}^n (f(x_j,a_j)-f'(x_j,a_j))^2/(\sigma_j^{t-1})^2}}\right), i = 1, \ldots, n$
**until** $\max_{i\in[n]} \left(\sigma_i^t/\sigma_i^{t-1}\right)^2 \leq 2$
**Output:** $\{\sigma_i^t\}_{i=1}^n$
```

For the second challenge, the weighted version $L_n := \sum_{i=1}^n (\hat{f}(z_i) - \bar{f}(z_i))^2/\sigma_i^2$ can also be decomposed into three terms correspondingly. We can demonstrate that an approximate choice of weights satisfying (6) is sufficient to control the corruption term as

$$\sum_{i=1}^n \frac{(\hat{f}(z_i) - \bar{f}(z_i))\zeta_i}{\sigma_i^2} = \sum_{i=1}^n \frac{|\hat{f}(z_i) - \bar{f}(z_i)|\zeta_i \cdot \sqrt{\lambda + L_n}}{\sigma_i^2 \sqrt{\lambda + \sum_{j=1}^n (\hat{f}(z_j) - \bar{f}(z_j))^2/\sigma_j^2}} \leq 2\alpha\zeta\sqrt{\lambda + L_n}.$$

Since the corruption-unrelated terms (corresponding to I_1, I_2) can still be bounded by $\tilde{\mathcal{O}}(\ln N)$, we have $L_n = \tilde{\mathcal{O}}(\ln N + 2\alpha\zeta\sqrt{L_n})$, leading to an $\tilde{\mathcal{O}}(\alpha\zeta + \sqrt{\ln N})$ confidence radius. Therefore, with a sufficiently small α, the effect of corruption can be countered.

3.2 Corruption-Robust Algorithm

Algorithm 2 CR-PEVI

```
**Input:** $\mathcal{D} = \{(x_i^h, a_i^h, r_i^h)\}_{i,h=1}^{n,H}, \mathcal{F}$
**Initialization:** Set $f_n^{H+1}(\cdot) \leftarrow 0$
**for** step $h = H, H-1, \ldots, 1$ **do**
    Choosing weights $\{\sigma_i^h\}_{i=1}^n$ by proceeding Algorithm 1 with inputs $\{(x_i^h, a_i^h)\}_{i=1}^n, \mathcal{F}^h, \alpha$
    Find the weighted least-squares solution in (7)
    Find $\beta^h$ and construct confidence set
        $$\widehat{\mathcal{F}}^h = \left\{f \in \mathcal{F}^h : \lambda + \sum_{i=1}^n (f(x_i^h, a_i^h) - \hat{f}^h(x_i^h, a_i^h))^2/(\sigma_i^h)^2 \leq (\beta^h)^2\right\}$$
    Construct bonus function as (8)
    Let $f_n^h(\cdot,\cdot) = \max\left(0, \hat{f}^h(\cdot,\cdot) - \beta^h b^h(\cdot,\cdot)\right)$
    Set $\hat{\pi}^h(\cdot) = \mathrm{argmax}_{a\in\mathcal{A}} f_n^h(\cdot, a)$
**end for**
**Output:** $\{\hat{\pi}^h\}_{h=1}^H$
```

Now, for the offline RL with general function approximation, we integrate the uncertainty weight iteration algorithm with the pessimistic value iteration (PEVI) algorithm [24], and propose a Corruption-Robust PEVI (CR-PEVI) in Algorithm 2. Our algorithm employs backward induction from step H to 1. Set estimated value function $f_n^{H+1}(\cdot) = 0$. At each step $h \in [H]$, having obtained f_n^{h+1}, we calculate f_n^h by solving the following weighted least-square regression:

$$\hat{f}^h = \underset{f^h\in\mathcal{F}^h}{\mathrm{argmin}} \sum_{i=1}^n \frac{(f(x_i^h, a_i^h) - r_i^h - f_n^{h+1}(x_i^{h+1}))^2}{(\sigma_i^h)^2}, \tag{7}$$

where the weights are obtained via Algorithm 1. As opposed to online RL where the necessity of exploration stimulates optimistic estimation, the literature on offline RL [24, 63] is more inclined to pessimism due to the limitation of offline data coverage. Hence, we construct a confidence set $\widehat{\mathcal{F}}^h = \{f \in \widehat{\mathcal{F}}^h : \lambda + \sum_{i=1}^n (f(x_i^h, a_i^h) - \hat{f}^h(x_i^h, a_i^h))^2/(\sigma_i^h)^2 \leq (\beta^h)^2\}$ such that the uncorrupted Bellman operator $\mathcal{T}^h$ converts the value function f_n^{h+1} into the function class $\widehat{\mathcal{F}}^h$ (i.e., $\mathcal{T}^h f_n^{h+1} \in$

$\widehat{\mathcal{F}}^h$) with high probability. For the bonus function, we follow [55] and choose it as

$$b_h(x,a) = \sup_{f,f' \in \widehat{\mathcal{F}}^h} \frac{|f(x,a) - f'(x,a)|}{\sqrt{\lambda + \sum_{i=1}^n (f(x_i^h, a_i^h) - f'(x_i^h, a_i^h))^2/(\sigma_i^h)^2}}, \tag{8}$$

which is seldom used in practical algorithms due to its unstability. Specifically, the covering number of the space containing (8) may be uncontrollable. According to Appendix E in [55], the issue of the covering number can be addressed under mild conditions by some techniques. Therefore, to maintain readability and consistency in this paper, we assume the corresponding bonus function space $\mathcal{B}^{h+1}$ of (8) has a bounded covering number. Then we introduce pessimism by subtracting b^h from the estimated value function: $f_n^h(x,a) = \max(0, \hat{f}^h(x,a) - \beta^h b^h(x,a))$.

4 Theoretical Analysis

4.1 Coverage Condition

No guarantee for the suboptimality can be provided with insufficient data coverage. Based on pessimism, Jin et al. [20], Rashidinejad et al. [36] have demonstrated that the coverage over the optimal policy is sufficient for sample-efficient offline RL. The following condition covers the optimal policy under general function approximation.

Definition 4.1 (Coverage Coefficient)**.** *Consider the offline dataset $\{x_i^h, a_i^h\}_{i,h=1}^{n,H}$. For any initial state $x^1 = x \in \mathcal{X}$, the coverage coefficient is:*

$$\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) = \max_{h\in[H]} \mathbb{E}_{\pi_*}\left[\sup_{f,f'\in\widehat{\mathcal{F}}^h} \frac{n(f(x^h,a^h) - f'(x^h,a^h))^2}{\lambda + \sum_{i=1}^n (f(x_i^h,a_i^h) - f'(x_i^h,a_i^h))^2} \,\middle|\, x^1 = x \right], \tag{9}$$

where $\mathbb{E}_{\pi_*}$ is taken with respect to the trajectory induced by π_* in the underlying uncorrupted MDP.

In the face of corruption, we require single-policy coverage over the uncorrupted trajectory. This coefficient depicts the expected uncertainty of sample (x^h, a^h) induced by the optimal policy π_* compared to the n training samples. We use the linear MDP to interpret this condition, where the function space $\mathcal{F}^h$ is embedded into a d-dimensional vector space: $\mathcal{F}^h = \{\langle w(f), \phi(\cdot)\rangle : z \to \mathbb{R}\}$, where z denotes the state-action pair (x,a). With the notation $\Lambda^h = \lambda I + \sum_{i=1}^n \phi(z_i^h)\phi(z_i^h)^\top$, we can demonstrate that if the sufficient "coverage" in Jin et al. [24] holds: there exists a constant $c^\dagger$ such that $\Lambda^h \succeq I + c^\dagger n \mathbb{E}_{\pi_*}[\phi(z^h)\phi(z^h)^\top \mid x^1 = x]$ for all $h \in [H]$, our coverage coefficient is bounded: $\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) \le d/c^\dagger < \infty$. Therefore, the sufficient "coverage" condition implies our coverage condition in the linear setting. We will discuss the connection formally in Lemma B.1.

4.2 Main Result

Before presenting the main theorem, we introduce a new general version of the well-explored dataset condition, which is the key to eliminating the uncertainty-related weights from the suboptimality bound and deriving the final result.

Assumption 4.1 (Well-Explored Dataset)**.** *There exists a constant $C(\widehat{\mathcal{F}}, \mu) > 8n^{-1}$ such that for a $(n^{-1}, \|\cdot\|_\infty)$ cover of $\widehat{\mathcal{F}}^h$, denoted by $\bar{\mathcal{F}}^h$, and for any two distinct $f, f' \in \bar{\mathcal{F}}^h$,*

$$\min_{h\in[H]} \sum_{z^h \in \mathcal{X}\times\mathcal{A}} \mu^h(z^h)\big(f(z^h) - f'(z^h)\big)^2 \ge C(\widehat{\mathcal{F}}, \mu). \tag{10}$$

We interpret this condition with the linear model, where the condition (10) becomes: for any two distinct $f, f' \in \bar{\mathcal{F}}^h$, $\min_{h\in[H]} \big(w(f) - w(f')\big)^\top \bar{\Lambda}^h \big(w(f) - w(f')\big) \ge C(\widehat{\mathcal{F}}, \mu)$. As proved in Lemma B.2, with high probability, the above condition holds with $C(\widehat{\mathcal{F}}, \mu) = \Theta(d^{-1})$ when the n trajectories of $\mathcal{D}$ are independent, and the data distribution μ^h satisfies the minimum eigenvalue condition:

$$\sigma_{\min}\big(\mathbb{E}_{z^h\sim\mu^h}[\phi(z^h)\phi(z^h)^\top]\big) = \bar{c}/d, \tag{11}$$

where $\bar{c} > 0$ is an absolute constant. This is a widely-adopted assumption in the literature [11, 46, 63]. Note that $\Theta(d^{-1})$ is the largest possible minimum eigenvalue since for any data distribution $\tilde{\mu}^h$, $\sigma_{\min}(\mathbb{E}_{z^h\sim\mu^h}[\phi(z^h)\phi(z^h)^\top]) \le d^{-1}$ by using $\|\phi(z^h)\| \le 1$ for any $z^h \in \mathcal{X}\times\mathcal{A}$.

Theorem 1. *If the coverage coefficient in Definition 4.1 is finite, under Assumption 4.1, for corruption $\zeta = \sum_{h=1}^{H} \zeta^h$ and $\delta > 0$, we choose the covering parameter $\gamma = 1/(n \max_h \beta^h \zeta^h)$, $\lambda = \ln(N_n(\gamma))$, the weighting parameter $\alpha = H\sqrt{\ln N_n(\gamma)}/\zeta$, and the confidence radius*

$$\beta^h = c_\beta\big(\alpha\zeta^h + \sqrt{\ln(HN_n(\gamma)/\delta)}\big), \quad \text{for } h = H, \ldots, 1,$$

where $N_n(\gamma) = \max_h N(\gamma/n, \mathcal{F}^h) \cdot N(\gamma/n, \mathcal{F}^{h+1}) \cdot N(\gamma/n, \mathcal{B}^{h+1}(\lambda))$. Then, with probability at least $1 - 2\delta$, the sub-optimality of Algorithm 2 is bounded by

$$\mathrm{SubOpt}(\hat{\pi}, x) = \tilde{\mathcal{O}}\Big(\frac{H(\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H))^{1/4} \cdot (\ln N_n(\gamma))^{1/2}}{n^{1/2}(C(\widehat{\mathcal{F}}, \mu))^{1/4}} + \frac{\zeta(\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H))^{1/2}}{n(C(\widehat{\mathcal{F}}, \mu))^{1/2}}\Big).$$

When $\zeta = \mathcal{O}(\sqrt{n})$, our algorithm achieves the same order of suboptimality as the uncorrupted case. Whenever $\zeta = o(n)$, our algorithm is sample-efficient. Moreover, when specialized to the linear MDP with dimension d, where $\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) = \mathcal{O}(d/c^\dagger)$, $C(\widehat{\mathcal{F}}, \mu) = \Theta(d^{-1})$ and $\ln N_n(\gamma) = \tilde{\mathcal{O}}(d^2)$, the suboptimality bound in Theorem 1 becomes $\tilde{\mathcal{O}}(Hd^{3/2}n^{-1/2} + d\zeta n^{-1})$. The corruption-independent term $\tilde{\mathcal{O}}(Hd^{3/2}n^{-1/2})$ matches that of PEVI [24]. The corruption-dependent term nearly matches the lower bound, as will be discussed later.

Remark 4.1. *Although the theory requires a known corruption level ζ, in the experiments, we treat the uncertainty ratio $\alpha = O(1/\zeta)$ as a tuning hyperparameter. The use of independent and identically distributed (i.i.d.) trajectories in our experiments renders the hyperparameter tuning process straightforward and conducive to optimizing the performance. Additionally, we can offer a choice of $\alpha = \Theta(1/\sqrt{n})$. This choice finds support in the online setting [55, 18], where this specific choice of α ensures that suboptimality remains in the order of uncorrupted error bound, even when $\zeta = O(\sqrt{n})$.*

Proof sketch. The detailed proof of Theorem 1 is provided in Appendix A. Here we present a brief proof sketch for the suboptimality bound, which is accomplished by three steps: (1) by Lemma A.1, if the uncorrupted Bellman backup $\mathcal{T}^h f_n \in \widehat{\mathcal{F}}^h$ for each $h \in [H]$, we can bound the suboptimality by the sum of the bonus $\sum_{h=1}^{H} \beta^h \mathbb{E}_{\pi_*}[b^h(x^h, a^h) \,|\, x^1 = x]$; (2) by Lemma A.2, we demonstrate that an approximate uncertainty weight satisfying (6) is the key to bound the weighted Bellman error $(\sum_{i=1}^{n}((\hat{f}^h - (\mathcal{T}^h f_n^{h+1}))(x_i^h, a_i^h))^2/(\sigma_i^h)^2)^{1/2}$ by $\beta^h = c_\beta(\alpha\zeta^h + \sqrt{\ln(HN_n(\gamma)/\delta)})$; and (3) combining the results in the first two steps, we can obtain the suboptimality bounded by:

$$\mathrm{SubOpt}(\hat{\pi}, x) = \tilde{\mathcal{O}}\big(H(\mathrm{CC}^\sigma(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H))^{1/2} \cdot (\ln N_n(\gamma))^{1/2} \cdot n^{-1/2} + \zeta \cdot \mathrm{CC}^\sigma(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) \cdot n^{-1}\big).$$

Here $\mathrm{CC}^\sigma(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H)$ denotes the weighted coverage coefficient as follows

$$\mathrm{CC}^\sigma(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) = \max_{h \in [H]} \mathbb{E}_{\pi_*}\Big[\sup_{f, f' \in \widehat{\mathcal{F}}^h} \frac{n(f(x^h, a^h) - f'(x^h, a^h))^2/\sigma^h(x^h, a^h)^2}{\lambda + \sum_{i=1}^{n}(f(x_i^h, a_i^h) - f'(x_i^h, a_i^h))^2/(\sigma_i^h)^2} \,\Big|\, x^1 = x\Big], \tag{12}$$

where the weight for the trajectory induced by the optimal policy π_* is

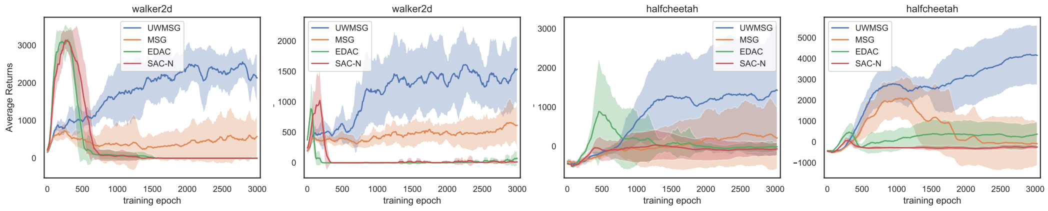
$$(\sigma^h(x^h, a^h))^2 = \max\Bigg(1, \sup_{f, f' \in \widehat{\mathcal{F}}^h} \frac{|f(x^h, a^h) - f'(x^h, a^h)|/\alpha}{\sqrt{\lambda + \sum_{i=1}^{n}(f(x_i^h, a_i^h) - f'(x_i^h, a_i^h))^2/(\sigma_i^h)^2}}\Bigg). \tag{13}$$

To control the weighted coverage coefficient CC^σ by its unweighted counterpart CC, which is a challenging task due to the intricate form of uncertainty-related weights, a novel technique has been employed. The main idea behind this technique is to transform CC^σ into a modified version of CC multiplied by a term that can be bounded by leveraging a cover of the function class. The connection between CC and CC^σ is shown in the following lemma.

Lemma 4.1. *Under Assumption 4.1 and choose $\beta^h = C_\beta\sqrt{\ln N}$ (where $C_\beta > 0$ contains the logarithmic terms that are omitted) and λ given in Theorem 1, we have*

$$\mathrm{CC}^\sigma(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) = \tilde{\mathcal{O}}\big((\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H))^{1/2} \cdot (C(\widehat{\mathcal{F}}, \mu))^{-1/2}\big).$$

Intuitively, because the weights themselves are closely related to the uncertainty, the weighted coverage coefficient CC^σ where the high-uncertain samples are down-weighted can be upper-bounded by the square root of the unweighted coefficient CC. See Appendix B.3 for details.



(a) (b) (c) (d)

Figure 1: Performance on the Walker2d and the Halfcheetah tasks under (a) random reward, (b) random dynamics, (c) adversarial reward, and (d) adversarial dynamics attacks.

Lower Bound. We construct a lower bound for linear MDPs with adversarial corruption ζ to show that the $\mathcal{O}(d\zeta n^{-1})$ corruption term of our suboptimality is optimal. The construction of the lower bound is adapted from Zhang et al. [60], where an ϵ-constant fraction of the dataset is contaminated. We present the proof of Theorem 2 in Appendix B.4.

Theorem 2 (Minimax Lower Bound for Linear MDPs)**.** *Under linear MDPs with corruption (Definition 2.2), for any fixed data-collecting distribution ν, any algorithm $L : \mathcal{D} \to \Pi$ with the knowledge of ζ cannot find a better policy than $\mathcal{O}(d\zeta n^{-1})$-optimal policy with probability more than $1/4$:*

$$\min_{L,\nu} \max_{MDP,\mathcal{D}} \mathrm{SubOpt}(\hat{\pi}_L) = \Omega\big(d\zeta/n\big),$$

where $\mathcal{D}$ is the corrupted dataset initially generated from the MDP and then corrupted by an adversary, and $\hat{\pi}_L$ is the policy generated by the algorithm L.

5 Experiments

Based on our theoretical results, we propose a practical implementation for CR-PEVI and verify its effectiveness on simulation tasks with corrupted offline data.

Practical Implementation. To make our algorithm more practical, we use neural networks to estimate the Q function (i.e., f in our theory) and the weight function σ. In linear MDPs, under a sufficiently broad function approximation class $\widehat{\mathcal{F}}^h$ and sufficiently small parameter λ in Eq. (8), the bonus function can be simplified as the bootstrapped uncertainty of Q functions, which in turn can be estimated via the standard deviation of an ensemble of Q networks. We defer the detailed discussion to Appendix B.5. Following the state-of-the-art uncertainty-based offline RL algorithm Model Standard-deviation Gradients (MSG) [16], we learn a group of Q networks $Q_{w_i}, i = 1, \dots, K$ with independent targets and optimize a policy π_θ with a lower-confidence bound (LCB) objective [16, 2]. Specifically, Q_{w_i} is learned to minimize a weighted regression objective similar to Eq.(7). The weight function σ is estimated via bootstrapped uncertainty: $\sigma(x, a) = \max(\sqrt{\mathbb{V}_{i=1,\dots,K}\left[Q_{w_i}(x, a)\right]}, 1)$, where $\mathbb{V}_{i=1,\dots,K}\left[Q_{w_i}\right]$ is the variance between the group of Q functions. This uncertainty estimation method has also been adopted by prior works [2, 16, 54]. We refer to our practical algorithm as Uncertainty Weighted MSG (UWMSG) and defer details to Appendix D.

Experimental Setup. We assess the performance of our approach using continuous control tasks from [15] and introduce both random and adversarial attacks on either the rewards or dynamics for the offline datasets. Details about the four types of data corruption and their cumulative corruption levels are deferred to Appendix D. The ensemble size K is set to 10 for all experiments. For evaluation, we report average returns with standard deviations over 10 random seeds. More implementation details are also provided in Appendix D.

Experimental Results. We compare UWMSG with the state-of-the-art uncertainty-base offline RL methods, MSG [16], EDAC [1], and SAC-N [1] under four types of data corruption. In particular, MSG can be considered as UWMSG with a constant weighting function $\sigma(s, a) = 1$. As demonstrated in Table 1 and Figure 1, our empirical results find that (1) current offline RL methods are susceptible to data corruption, e.g., MSG, EDAC, SAC-N achieve poor performance under adversarial attacks, and (2) our proposed UWMSG significantly improves performance under different data corruption scenarios, with an average improvement of 104% over MSG. More results can be found in Appendix F.

Table 1: Comparison of different offline RL algorithms under different environments and different data corruption types.

Environment	Attack Type	UWMSG	MSG	EDAC	SAC-N
Halfcheetah	Random Reward	7299.9 ± 169.0	4339.5 ± 3958.7	7128.2 ± 120.5	**7357.5 ± 165.2**
	Random Dynamics	**1425.0 ± 1659.5**	212.9 ± 793.3	-12.3 ± 131.8	-66.2 ± 169.9
	Adversarial Reward	**1016.7 ± 503.9**	243.1 ± 338.6	-127.2 ± 30.3	-55.7 ± 24.0
	Adversarial Dynamics	**4144.3 ± 1437.6**	-87.3 ± 1055.6	374.0 ± 589.5	-246.8 ± 104.9
Walker2d	Random Reward	**2189.7 ± 603.0**	539.1 ± 534.6	-3.7 ± 1.2	-3.8 ± 1.4
	Random Dynamics	**2278.9 ± 706.4**	2122.8 ± 821.5	-3.6 ± 0.8	-3.1 ± 0.7
	Adversarial Reward	**1433.4 ± 592.1**	605.8 ± 310.7	61.0 ± 120.6	9.5 ± 21.6
	Adversarial Dynamics	**946.0 ± 300.4**	506.0 ± 175.5	-4.8 ± 0.3	-5.1 ± 0.7
Hopper	Random Reward	**2021.1 ± 888.6**	1599.6 ± 814.5	107.7 ± 65.5	178.8 ± 114.0
	Random Dynamics	**2116.2 ± 618.6**	1552.7 ± 532.8	5.9 ± 1.2	3.9 ± 2.2
	Adversarial Reward	**751.4 ± 72.5**	651.0 ± 58.0	29.2 ± 18.7	111.8 ± 62.7
	Adversarial Dynamics	**931.2 ± 227.8**	717.7 ± 204.3	5.9 ± 1.2	3.9 ± 2.2
Average		**2212.8**	1083.6	630.0	607.1

In summary, the experimental results validate the theoretical impact of data corruption for value-based offline RL algorithms. Our practical implementation algorithm demonstrates superior efficacy under different data corruption types, thereby highlighting its potential for real-world applications.

6 Conclusion

This work investigates the adversarially corrupted offline RL with general function approximation. We propose the uncertainty weight iteration and a weighted version of PEVI. Under a partial coverage condition and a well-explored dataset, our algorithm achieves a suboptimality bound, where the corruption-independent term recovers the uncorrupted bound, and the corruption-related term nearly matches the lower bound in linear models. Furthermore, our experiments demonstrate promising results showing that our practical implementation, UWMSG, significantly enhances the performance of the state-of-the-art offline RL algorithm under reward and dynamics data corruptions.

Our work suggests several potential future directions. First, it remains unsolved whether one can design robust algorithms to handle additional distribution shifts of the state-action pairs. Second, when the corruption level is unknown, how to design a theoretically robust offline RL algorithm and overcome the lack of adaptivity in the offline setting requires further research. Finally, we hope that our work sheds light on applying uncertainty weights to improve robustness in deep offline RL works and even practical applications.

Acknowledgements

We thank Wei Xiong for valuable discussions and feedback on an early draft of this work. We also thank the anonymous reviewers for their helpful comments.

References

[1] An, G., Moon, S., Kim, J.-H., and Song, H. O. (2021). Uncertainty-based offline reinforcement learning with diversified q-ensemble. *Advances in neural information processing systems*, 34:7436–7447.

[2] Bai, C., Wang, L., Yang, Z., Deng, Z., Garg, A., Liu, P., and Wang, Z. (2022). Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning. *arXiv preprint arXiv:2202.11566*.

[3] Behzadan, V. and Munir, A. (2018). Mitigation of policy manipulation attacks on deep q-networks with parameter-space noise. In *Computer Safety, Reliability, and Security: SAFECOMP 2018 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings 37*, pages 406–417. Springer.

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

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

[6] Chen, Z., Li, C. J., Yuan, A., Gu, Q., and Jordan, M. I. (2023). A general framework for sample-efficient function approximation in reinforcement learning. In *International Conference on Learning Representations*.

[7] Deng, D., Chen, G., Yu, Y., Liu, F., and Heng, P.-A. (2023). Uncertainty estimation by fisher information-based evidential deep learning. *arXiv preprint arXiv:2303.02045*.

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

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

[10] Du, Y., Watkins, O., Wang, Z., Colas, C., Darrell, T., Abbeel, P., Gupta, A., and Andreas, J. (2023). Guiding pretraining in reinforcement learning with large language models. *arXiv preprint arXiv:2302.06692*.

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

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

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

[14] Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. (2021). The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*.

[15] Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. (2020). D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*.

[16] Ghasemipour, K., Gu, S. S., and Nachum, O. (2022). Why so pessimistic? estimating uncertainties for offline rl through ensembles, and why their independence matters. *Advances in Neural Information Processing Systems*, 35:18267–18281.

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

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

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

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

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

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

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

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

[25] Kang, Y., Hsieh, C.-J., and Lee, T. (2023). Robust lipschitz bandits to adversarial corruptions. *arXiv preprint arXiv:2305.18543*.

[26] Kober, J., Bagnell, J. A., and Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274.

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

[28] Liu, Y., Logan, B., Liu, N., Xu, Z., Tang, J., and Wang, Y. (2017). Deep reinforcement learning for dynamic treatment regimes on medical registry data. In *2017 IEEE international conference on healthcare informatics (ICHI)*, pages 380–385. IEEE.

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

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

[31] Ma, Y., Zhang, X., Sun, W., and Zhu, J. (2019). Policy poisoning in batch reinforcement learning and control. *Advances in Neural Information Processing Systems*, 32.

[32] Neff, G. (2016). Talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*.

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

[34] Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. (2016). Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29.

[35] Pan, Y., Cheng, C.-A., Saigol, K., Lee, K., Yan, X., Theodorou, E., and Boots, B. (2017). Agile autonomous driving using end-to-end deep imitation learning. *arXiv preprint arXiv:1709.07174*.

[36] Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716.

[37] Rosenberg, A. and Mansour, Y. (2019). Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pages 5478–5486. PMLR.

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

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

[40] Sun, H., Han, L., Yang, R., Ma, X., Guo, J., and Zhou, B. (2022a). Exploit reward shifting in value-based deep-rl: Optimistic curiosity-based exploration and conservative exploitation via linear reward shaping. *Advances in Neural Information Processing Systems*, 35:37719–37734.

[41] Sun, H., van Breugel, B., Crabbe, J., Seedat, N., and van der Schaar, M. (2022b). Daux: a density-based approach for uncertainty explanations. *arXiv preprint arXiv:2207.05161*.

[42] Tarasov, D., Nikulin, A., Akimov, D., Kurenkov, V., and Kolesnikov, S. (2022). CORL: Research-oriented deep offline reinforcement learning library. In *3rd Offline RL Workshop: Offline RL as a "Launchpad"*.

[43] Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434.

[44] Uehara, M. and Sun, W. (2021). Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*.

[45] Wang, L., Zhang, W., He, X., and Zha, H. (2018). Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2447–2456.

[46] Wang, R., Foster, D. P., and Kakade, S. M. (2020a). What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*.

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

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

[49] Wu, F., Li, L., Zhang, H., Kailkhura, B., Kenthapadi, K., Zhao, D., and Li, B. (2022). Copa: Certifying robust policies for offline reinforcement learning against poisoning attacks. In *International Conference on Learning Representations*.

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

[51] Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. (2021). Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694.

[52] Xiong, W., Zhong, H., Shi, C., Shen, C., Wang, L., and Zhang, T. (2022). Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *arXiv preprint arXiv:2205.15512*.

[53] Yang, R., Bai, C., Ma, X., Wang, Z., Zhang, C., and Han, L. (2022). Rorl: Robust offline reinforcement learning via conservative smoothing. In *Advances in Neural Information Processing Systems*.

[54] Yang, R., Yong, L., Ma, X., Hu, H., Zhang, C., and Zhang, T. (2023). What is essential for unseen goal generalization of offline goal-conditioned rl? In *International Conference on Machine Learning*, pages 39543–39571. PMLR.

[55] Ye, C., Xiong, W., Gu, Q., and Zhang, T. (2022). Corruption-robust algorithms with uncertainty weighting for nonlinear contextual bandits and markov decision processes.

[56] Yin, M. and Wang, Y.-X. (2021). Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34:4065–4078.

[57] Zanette, A., Wainwright, M. J., and Brunskill, E. (2021). Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640.

[58] Zhang, H., Chen, H., Xiao, C., Li, B., Liu, M., Boning, D., and Hsieh, C.-J. (2020a). Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33:21024–21037.

[59] Zhang, T. (2023). *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press. in press, also available as `http://tongzhang-ml.org/lt-book.html`.

[60] Zhang, X., Chen, Y., Zhu, X., and Sun, W. (2022). Corruption-robust offline reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 5757–5773. PMLR.

[61] Zhang, X., Ma, Y., Singla, A., and Zhu, X. (2020b). Adaptive reward-poisoning attacks against reinforcement learning. In *International Conference on Machine Learning*, pages 11225–11234. PMLR.

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

[63] Zhong, H., Xiong, W., Tan, J., Wang, L., Zhang, T., Wang, Z., and Yang, Z. (2022a). Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. In *International Conference on Machine Learning*, pages 27117–27142. PMLR.

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

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

A Proof of Theorem 1

A.1 Step I: Suboptimality Decomposition

Lemma A.1 (Regret Decomposition)**.** *Assuming that $\mathcal{T}^h f_n \in \widehat{\mathcal{F}}^h$ for all $h \in [H]$, we have*

$$\mathrm{SubOpt}(\hat{\pi}, x) \le 2\sum_{h=1}^{H} \beta^h \mathbb{E}_{\pi_*}\big[b^h(x^h, a^h) \,\big|\, x^1 = x\big].$$

Proof. By invoking Lemma G.1, we can decompose the suboptimality as follows:

$$\begin{aligned}
\mathrm{SubOpt}(\hat{\pi}, x) &= V_*^1(x) - V_{\hat{\pi}}^1(x) \\
&= \sum_{h=1}^{H} \mathbb{E}_{\pi_*}\big[f_n^h(x^h, \pi_*(x^h)) - f_n^h(x^h, \hat{\pi}(x^h)) \,\big|\, x^1 = x\big] \\
&\quad - \sum_{h=1}^{H} \mathbb{E}_{\pi_*}\big[\mathcal{E}^h(f_n, x^h, a^h) \,\big|\, x^1 = x\big] + \sum_{h=1}^{H} \mathbb{E}_{\hat{\pi}}\big[\mathcal{E}^h(f_n, x^h, a^h) \,\big|\, x^1 = x\big] \\
&\le \underbrace{- \sum_{h=1}^{H} \mathbb{E}_{\pi_*}\big[\mathcal{E}^h(f_n, x^h, a^h) \,\big|\, x^1 = x\big]}_{(a)} + \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\hat{\pi}}\big[\mathcal{E}^h(f_n, x^h, a^h) \,\big|\, x^1 = x\big]}_{(b)}, \qquad (14)
\end{aligned}$$

where the inequality is due to $\hat{\pi}(x^h) = \mathrm{argmax}_{a \in \mathcal{A}} f_n^h(x^h, a)$. Then, we handle the above two terms respectively. We first tackle term (b). Supposing that $\mathcal{T}^h f_n^{h+1} \in \widehat{\mathcal{F}}^h$, we get from the definition of the confidence set in Algorithm 2 that

$$\Big(\lambda + \sum_{i=1}^{n} \frac{((\mathcal{T}^h f_n^{h+1})(x_i^h, a_i^h) - \hat{f}^h(x_i^h, a_i^h))^2}{(\sigma_i^h)^2}\Big)^{1/2} \le \beta^h. \qquad (15)$$

Therefore, for any $(x^h, a^h) \in \mathcal{X} \times \mathcal{A}$, we have

$$\begin{aligned}
&\big|\hat{f}^h(x^h, a^h) - (\mathcal{T}^h f_n^{h+1})(x^h, a^h)\big| \\
&\quad = \Big(\lambda + \sum_{i=1}^{n} \frac{((\mathcal{T}^h f_n^{h+1})(x_i^h, a_i^h) - \hat{f}^h(x_i^h, a_i^h))^2}{(\sigma_i^h)^2}\Big)^{1/2} \cdot \frac{\big|\hat{f}^h(x^h, a^h) - (\mathcal{T}^h f_n^{h+1})(x^h, a^h)\big|}{\Big(\lambda + \sum_{i=1}^{n} ((\mathcal{T}^h f_n^{h+1})(x_i^h, a_i^h) - \hat{f}^h(x_i^h, a_i^h))^2/(\sigma_i^h)^2\Big)^{1/2}} \\
&\quad \le \beta^h b^h(x^h, a^h),
\end{aligned}$$

where the inequality uses (15) and the definition of the bonus b^h (8). Then, combining the above result with $f_n^h(x, a) = \max(0, \hat{f}^h(x, a) - \beta^h b^h(x, a))$ and $(\mathcal{T}^h f_n^{h+1})(x^h, a^h) \ge 0$, we get

$$-2\beta^h b^h(x^h, a^h) \le f_n^h(x^h, a^h) - (\mathcal{T}^h f_n^{h+1})(x^h, a^h) \le 0.$$

Hence, the term (b) is bounded by

$$\sum_{h=1}^{H} \mathbb{E}_{\hat{\pi}}\big[\mathcal{E}^h(f_n, x^h, a^h) \,\big|\, x^1 = x\big] \le 0. \qquad (16)$$

Moreover, the term (a) is bounded by

$$- \sum_{h=1}^{H} \mathbb{E}_{\pi_*}\big[\mathcal{E}^h(f_n, x^h, a^h) \,\big|\, x^1 = x\big] \le 2\sum_{h=1}^{H} \beta^h \mathbb{E}_{\pi_*}\big[b^h(x^h, a^h) \,\big|\, x^1 = x\big]. \qquad (17)$$

Finally, by taking (17) and (16) back into (14), we conclude the proof. □

A.2 Step II: Sharper confidence radius for Pessimism

Lemma A.2 (Confidence Radius)**.** *In Algorithm 2, for all $h \in [H]$ we have $\mathcal{T}^h f_n^{h+1} \in \widehat{\mathcal{F}}^h$ with probability at least $1-\delta$, where we choose*

$$\beta^h = 24\alpha\zeta^h + \Big(12\lambda + 12\ln(2HN_n^h(\gamma)/\delta) + 12(5\beta^{h+1}\gamma)^2 n + 60\beta^{h+1}\gamma\sqrt{nC_1^h(n,\zeta)}\Big)^{1/2},$$

where

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

we use the notation $C_1^h(n,\zeta) = 2(\sum_{i=1}^n(\zeta_i^h)^2 + 2n\eta^2 + 3\eta^2\ln(2/\delta))$, and $\zeta^h = \sum_{i=1}^n \zeta_i^h$.

Proof. At each step $h \in [H]$, let $\mathcal{F}_\gamma^{h+1}$ be 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 $\bar{\mathcal{F}}_\gamma^{h+1} = \mathcal{F}_\gamma^{h+1} \oplus \beta_{\mathcal{T}}^{h+1}\mathcal{B}_\gamma^{h+1}$ as a $((1+\beta^{h+1})\gamma, \|\cdot\|_\infty)$ cover of $f_n^{h+1}(\cdot)$. Given f_n^{h+1}, we have $\bar{f}^{h+1} \in \bar{\mathcal{F}}_\gamma^{h+1}$ so that $\|\bar{f}^{h+1} - f_n^{h+1}\|_\infty \le \bar{\epsilon} = (1+\beta^{h+1})\gamma$. Then, we define $y_i^h = r_i^h + f_n^{h+1}(x_i^{h+1})$, $\bar{y}_i^h = r_i^h + \bar{f}^{h+1}(x_i^{h+1})$ and

$$\tilde{f}^h = \operatorname*{argmin}_{f^h \in \mathcal{F}^h} \sum_{i=1}^n (f^h(x_i^h, a_i^h) - \bar{y}_i^h)^2.$$

We know from the definition of the covers that

$$\begin{aligned}\left(\sum_{i=1}^n(\hat{f}^h(x_i^h,a_i^h) - \bar{y}_i^h)^2\right)^{1/2} &\le \left(\sum_{i=1}^n(\hat{f}^h(x_i^h,a_i^h) - y_i^h)^2\right)^{1/2} + \sqrt{n}\bar{\epsilon}\\ &\le \left(\sum_{i=1}^n(\tilde{f}^h(x_i^h,a_i^h) - y_i^h)^2\right)^{1/2} + \sqrt{n}\bar{\epsilon} \le \left(\sum_{i=1}^n(\tilde{f}^h(x_i^h,a_i^h) - \bar{y}_i^h)^2\right)^{1/2} + 2\sqrt{n}\bar{\epsilon}, \quad (18)\end{aligned}$$

where the first and third inequality is due to $\|\bar{f}^{h+1} - f_n^{h+1}\|_\infty \le \bar{\epsilon}$, and the second inequality comes from the fact that $\hat{f}^h$ is the ERM solution to the least squares problem. Then, we can invoke Lemma G.3 by taking f_* as $\mathbb{E}[\bar{y}_i^h | x_i^h, a_i^h] = (\mathcal{T}_{\mathcal{D}}^h \bar{f}^{h+1})(x_i^h, a_i^h)$ and f_b as $\mathcal{T}^h\bar{f}^{h+1}$. With probability at least $1-\delta$, we obtain:

$$\begin{aligned}&\sum_{i=1}^n \frac{\Big(\hat{f}^h(x_i^h,a_i^h) - (\mathcal{T}^h\bar{f}^{h+1})(x_i^h,a_i^h)\Big)^2}{(\sigma_i^h)^2}\\ &\quad\le 10\ln(2HN_n^h(\gamma)/\delta) + 5\underbrace{\sum_{i=1}^n \frac{|\hat{f}^h(x_i^h,a_i^h) - (\mathcal{T}^h\bar{f}^{h+1})(x_i^h,a_i^h)|\cdot|\zeta_i^h|}{(\sigma_i^h)^2}}_{(a)}\\ &\quad + 10(\gamma+2\bar{\epsilon})\cdot\Big((\gamma+2\bar{\epsilon})n + \sqrt{nC_1(n,\zeta)}\Big), \qquad (19)\end{aligned}$$

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

According to Lemma 3.1, the term (a) can be controlled by the weight design:

$$\begin{aligned}&\sum_{i=1}^n |\zeta_i^h|\cdot\frac{|\hat{f}^h(x_i^h,a_i^h) - (\mathcal{T}^h\bar{f}^{h+1})(x_i^h,a_i^h)|}{(\sigma_i^h)^2}\\ &\quad\le \zeta^h \sup_i \frac{|\hat{f}^h(x_i^h,a_i^h) - (\mathcal{T}^h\bar{f}^{h+1})(x_i^h,a_i^h)|}{(\sigma_i^h)^2}\\ &\quad\le 2\alpha\zeta^h\sqrt{\lambda + \sum_{i=1}^n\frac{\big(\hat{f}^h(x_i^h,a_i^h) - (\mathcal{T}^h\bar{f}^{h+1})(x_i^h,a_i^h)\big)^2}{(\sigma_i^h)^2}},\end{aligned}$$

where the second inequality is obtained since σ_i^h satisfies (6). Taking this result back into (19), we finally get for all $h \in [H]$,

$$
\begin{aligned}
&\sum_{i=1}^n \frac{\big(\hat{f}^h(x_i^h, a_i^h) - (\mathcal{T}^h \bar{f}^{h+1})(x_i^h, a_i^h)\big)^2}{(\sigma_i^h)^2} \\
&\qquad \le 10 \ln(2HN_n^h(\gamma)/\delta) + 10\alpha\zeta^h\beta^h + 5\gamma\zeta + 10(\gamma + 2\bar{\epsilon}) \cdot ((\gamma + 2\bar{\epsilon})n + \sqrt{nC_1(t, \zeta)}) \\
&\qquad = 10 \ln(2HN_n^h(\gamma)/\delta) + 10\alpha\zeta^h\beta^h + 5\gamma\zeta + 10(2\beta^{h+1} + 3)^2\gamma^2 n + 10(2\beta^{h+1} + 3)\gamma\sqrt{nC_1(n, \zeta)},
\end{aligned}
$$

where the last equality uses $\bar{\epsilon} = (1 + \beta^{h+1})\gamma$. Therefore, it follows that with probability at least $1 - \delta$,

$$
\begin{aligned}
&\left(\sum_{i=1}^n \frac{\Big(\hat{f}_n^h(x_i^h, a_i^h) - (\mathcal{T}^h f_n^{h+1})(x_i^h, a_i^h)\Big)^2}{(\sigma_i^h)^2} + \lambda \right)^{1/2} \\
&\qquad \le \left(\sum_{i=1}^n \frac{\Big(\hat{f}_n^h(x_i^h, a_i^h) - (\mathcal{T}^h \bar{f}_n^{h+1})(x_i^h, a_i^h)\Big)^2}{(\sigma_i^h)^2} \right)^{1/2} + \sqrt{n}\bar{\epsilon} + \sqrt{\lambda} \\
&\qquad \le \Big(10 \ln(2HN_n^h(\gamma)/\delta) + 10\alpha\zeta^h\beta^h + 5\gamma\zeta + 10(2\beta^{h+1} + 3)^2\gamma^2 n \\
&\qquad\qquad + 10(2\beta^{h+1} + 3)\gamma\sqrt{nC_1(n, \zeta)} \Big)^{1/2} + (\beta^{h+1} + 1)\gamma\sqrt{n} + \sqrt{\lambda} \\
&\qquad \le \Big(12\lambda + 12 \ln(2HN_n^h(\gamma)/\delta) + 24\alpha\zeta^h\beta^h + 12(5\beta^{h+1}\gamma)^2 n + 60\beta^{h+1}\gamma\sqrt{nC_1(n, \zeta)} \Big)^{1/2} \le \beta^h.
\end{aligned}
$$

Therefore, we complete the proof. □

A.3 Step III: Bound the Suboptimality

Proof of Theorem 1. We know from Lemma A.1 and Lemma A.2 that with probability at least $1 - \delta$,

$$
\begin{aligned}
\mathrm{SubOpt}(\hat{\pi}, x) &\le 2 \sum_{h=1}^H \beta^h \mathbb{E}_{\pi_*} \big[b^h(x^h, a^h) \,\big|\, x^1 = x \big] \\
&= 2 \sum_{h=1}^H \beta^h \mathbb{E}_{\pi_*} \bigg[\frac{b^h(x^h, a^h)}{\sigma^h(x^h, a^h)} \cdot \mathbb{1}(\sigma^h(x^h, a^h) = 1) \\
&\qquad + \frac{b^h(x^h, a^h)}{\sigma^h(x^h, a^h)} \cdot \sigma^h(x^h, a^h) \cdot \mathbb{1}(\sigma^h(x^h, a^h) > 1) \,\bigg|\, x^1 = x \bigg] \\
&\le 2 \sum_{h=1}^H \beta^h \mathbb{E}_{\pi_*} \bigg[\frac{b^h(x^h, a^h)}{\sigma^h(x^h, a^h)} + \Big(\frac{b^h(x^h, a^h)}{\sigma^h(x^h, a^h)} \Big)^2 \cdot \frac{1}{\alpha} \,\bigg|\, x^1 = x \bigg], \qquad (20)
\end{aligned}
$$

where the last inequality is deduced since from the definition of $\sigma^h(x^h, a^h)$ in (13), we get for $\sigma^h(x^h, a^h) > 1$ that

$$
\sigma^h(x^h, a^h) = \frac{1}{\sigma^h(x^h, a^h)} \cdot \sup_{f, f' \in \widehat{\mathcal{F}}^h} \frac{|f(x^h, a^h) - f'(x^h, a^h)|/\alpha}{\sqrt{\lambda + \sum_{i=1}^n (f(x_i^h, a_i^h) - f'(x_i^h, a_i^h))^2/(\sigma_i^h)^2}} = \frac{b^h(x^h, a^h)}{\alpha\sigma^h(x^h, a^h)}.
$$

From the definition of the weighted coverage condition (12), we have

$$
\mathbb{E}_{\pi_*} \bigg[\Big(\frac{b^h(x^h, a^h)}{\sigma^h(x^h, a^h)} \Big)^2 \,\bigg|\, x^1 = x \bigg] = \frac{\mathrm{CC}^\sigma(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H)}{n}.
$$

Combining the above equation and $\mathbb{E}X \leq \sqrt{\mathbb{E}X^2}$, we can bound (20) by

$$\begin{aligned}
&\sum_{h=1}^{H} \beta^h \mathbb{E}_{\pi_*}\bigg[\frac{b^h(x^h,a^h)}{\sigma^h(x^h,a^h)} + \Big(\frac{b^h(x^h,a^h)}{\sigma^h(x^h,a^h)}\Big)^2 \cdot \frac{1}{\alpha}\,\bigg|\, x^1 = x\bigg] \\
&\quad \leq \sum_{h=1}^{H} \beta^h \bigg[\sqrt{\frac{\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}_n^H)}{n}} + \frac{\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}_n^H)}{n}\cdot\frac{1}{\alpha}\bigg] \\
&\quad \leq \sqrt{\frac{\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}_n^H)}{n}} \sum_{h=1}^{H}\beta^h + \frac{\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}_n^H)}{n}\cdot\frac{\sum_{h=1}^{H}\beta^h}{\alpha} \\
&\quad \leq \sqrt{\frac{\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}_n^H)}{n}}\cdot c_\beta\big(\alpha\zeta + H\sqrt{\ln N}\big) + \frac{\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}_n^H)}{n}\cdot c_\beta\Big(\zeta + \frac{H\sqrt{\ln N}}{\alpha}\Big).
\end{aligned}$$

By choosing $\alpha = H\sqrt{\ln N}/\zeta$, we can obtain the result:

$$\mathrm{SubOpt}(\hat{\pi},x) = \tilde{\mathcal{O}}\bigg(\frac{3H\sqrt{\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}_n^H)\cdot\ln N}}{\sqrt{n}} + \frac{\zeta\cdot\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}_n^H)}{n}\bigg).$$

Ultimately, we can invoke Lemma 4.1 to obtain that with probability at least $1-2\delta$,

$$\mathrm{SubOpt}(\hat{\pi},x) = \tilde{\mathcal{O}}\bigg(\frac{H(\mathrm{CC}(\lambda,\widehat{\mathcal{F}},\mathcal{Z}_n^H))^{1/4}\cdot(\ln N_n(\gamma))^{1/2}}{n^{1/2}(C(\widehat{\mathcal{F}},\mu))^{1/4}} + \frac{\zeta(\mathrm{CC}(\lambda,\widehat{\mathcal{F}},\mathcal{Z}_n^H))^{1/2}}{n(C(\widehat{\mathcal{F}},\mu))^{1/2}}\bigg),$$

which concludes the proof. □

B Proofs of Auxiliary Results

B.1 Proof of Lemma 3.1

To begin with, we demonstrate that the uncertainty weight iteration (Algorithm 1) converges, thus satisfying the approximate condition in (6).

Proof of Lemma 3.1. We will demonstrate this result via the convergence of monotone real number sequences. To begin with, we prove the monotonicity of $\{(\sigma_i^t)^2\}_{t=0}^\infty$ by induction. When $\tau = 0$, we know that $(\sigma_i^1)^2 \geq 1 = \sigma_i^0$ for $i \in [n]$. When $\tau = t$, assume that $(\sigma_i^t)^2 \geq (\sigma_i^{t-1})^2$ for all $i \in [n]$, which implies that

$$\begin{aligned}
&\sup_{f,f'\in\mathcal{F}} \frac{|f(x_i,a_i) - f'(x_i,a_i)|/(\alpha)}{\sqrt{\lambda + \sum_{j=1}^n (f(x_j,a_j) - f'(x_j,a_j))^2/(\sigma_j^t)^2}} \\
&\geq \sup_{f,f'\in\mathcal{F}} \frac{|f(x_i,a_i) - f'(x_i,a_i)|/(\alpha)}{\sqrt{\lambda + \sum_{j=1}^n (f(x_j,a_j) - f'(x_j,a_j))^2/(\sigma_j^{t-1})^2}}.
\end{aligned}$$

Therefore, we get $(\sigma_i^{t+1})^2 \geq (\sigma_i^t)^2$ for all $i \in [n]$. Then, when $t = T$, we deduce from $(\sigma_i^{T+1})^2 \geq (\sigma_i^T)^2$ that for each $i \in [n]$,

$$(\sigma_i^T)^2 \leq (\sigma_i^{T+1})^2 \leq \max\bigg(1, \sup_{f,f'\in\mathcal{F}} \frac{|f(x_i,a_i) - f'(x_i,a_i)|/\alpha}{\sqrt{\lambda + \sum_{j=1}^n (f(x_j,a_j) - f'(x_j,a_j))^2/(\sigma_j^N)^2}}\bigg),$$

which implies the second inequality of (6).

Then, we obtain the upper bounds of each sequence: for $i \in [n]$ and any $t \geq 0$

$$\sup_{f,f'\in\mathcal{F}} \frac{|f(x_i,a_i) - f'(x_i,a_i)|/\alpha}{\sqrt{\lambda + \sum_{j=1}^n (f(x_j,a_j) - f'(x_j,a_j))^2/(\sigma_j^t)^2}} \leq \frac{1}{\alpha\sqrt{\lambda}},$$

where we use $f(\cdot) \in [0,1]$ for any $f \in \mathcal{F}$. Thus, each $\{(\sigma_i^t)^2\}_{t=0}^{\infty}$ has a $\max(1, 1/(\alpha\sqrt{\lambda}))$ upper bound.

According to the convergence of monotone real number sequences, the sequence $\{(\sigma_i^t)^2\}_{t=1}^{\infty}$ converges for all $i \in [n]$, which implies that $\{\log((\sigma_i^t)^2)\}_{t=1}^{\infty}$ converges. We know from the definition of convergence that for any $\mu > 0$, there exists an $N(\mu)$ such that for any $t \geq N$,

$$\log\big((\sigma_i^{t+1})^2\big) - \log\big((\sigma_i^t)^2\big) \leq \log(1+\mu),$$

which implies that

$$\frac{(\sigma_i^{t+1})^2}{(\sigma_i^n)^2} \leq 1+\mu. \tag{21}$$

For any $t \geq N$, if $\sigma_i^{t+1} = 1$, we have from the monotonicity that $\sigma_i^t = 1$, thus satisfying the first inequality of (6). If $\sigma_i^{n+1} > 1$, we deduce from (21) that

$$(\sigma_i^t)^2 \geq \frac{1}{1+\mu}(\sigma_i^{t+1})^2 = \frac{1}{1+\mu} \sup_{f,f' \in \mathcal{F}} \frac{|f(x_i,a_i) - f'(x_i,a_i)|/(\alpha)}{\sqrt{\lambda + \sum_{j=1}^n (f(x_j,a_j) - f'(x_j,a_j))^2/(\sigma_j^t)^2}}.$$

Hence, the inequality is proved by taking $\mu = 1$ and stop the iteration at round $t = N+1$. □

B.2 Connection between Coverage Coefficients

In this part, we state that in the linear MDP, the coverage condition in Jin et al. [24] implies our coverage assumption. Recall the coverage coefficient defined in (9):

$$\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) = \max_{h \in [H]} \mathbb{E}_{\pi_*}\bigg[\sup_{f,f' \in \widehat{\mathcal{F}}^h} \frac{n(f(x^h,a^h) - f'(x^h,a^h))^2}{\lambda + \sum_{i=1}^n (f(x_i^h,a_i^h) - f'(x_i^h,a_i^h))^2} \,\bigg|\, x^1 = x\bigg].$$

When the function space $\mathcal{F}^h$ is embedded into a d-dimensional vector space: $\mathcal{F}^h = \{\langle w(f), \phi(\cdot)\rangle : z \to \mathbb{R}\}$, where z denotes the state-action pair (x,a). Then, we define $\Lambda^h = \lambda I + \sum_{i=1}^n \phi(z_i^h)\phi(z_i^h)^\top$.

The coverage condition in Jin et al. [24] assumes that there exists a constant $c^\dagger$ such that for all $h \in [H]$,

$$\Lambda^h \succeq I + c^\dagger n \mathbb{E}_{\pi_*}\big[\phi(z^h)\phi(z^h)^\top \,\big|\, x^1 = x\big]. \tag{22}$$

Lemma B.1. *In the linear setting with dimension d, if the coverage condition in* (22) *holds with a finite constant* $c^\dagger > 0$*, the coverage coefficient in* (9) *is also finite:*

$$\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) \leq \frac{d}{c^\dagger} < \infty.$$

Proof. We deduce that

$$\begin{aligned}
&\sup_{f,f' \in \widehat{\mathcal{F}}^h} \frac{n\big(\langle w(f) - w(f'), \phi(z)\rangle\big)^2}{\lambda + \sum_{i=1}^n \big(\langle w(f) - w(f'), \phi(z_i^h)\rangle\big)^2} \\
&\quad\leq \sup_{f,f' \in \widehat{\mathcal{F}}^h} \frac{n\big(\langle w(f) - w(f'), \phi(z)\rangle\big)^2}{\big(w(f) - w(f')\big)^\top \Lambda^h \big(w(f) - w(f')\big)} \\
&\quad\leq n\phi(z)^\top (\Lambda^h)^{-1}\phi(z),
\end{aligned}$$

where the second inequality uses

$$\langle w(f) - w(f'), \phi(z)\rangle \leq \sqrt{\big(w(f) - w(f')\big)^\top \Lambda^h \big(w(f) - w(f')\big)} \cdot \sqrt{\phi(z)^\top (\Lambda^h)^{-1}\phi(z)}.$$

Thus, the coverage coefficient (9) for the linear model is bounded by

$$\begin{aligned}
\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) &\leq \max_{h \in [H]} \mathbb{E}_{\pi_*}\bigg[\sup_{f,f' \in \widehat{\mathcal{F}}^h} \frac{n\big(\langle w(f) - w(f'), \phi(z)\rangle\big)^2}{\lambda + \sum_{i=1}^n \big(\langle w(f) - w(f'), \phi(z_i^h)\rangle\big)^2} \,\bigg|\, x^1 = x\bigg] \\
&\leq \max_{h \in [H]} \mathbb{E}_{\pi_*}\big[n\phi(z^h)^\top (\Lambda^h)^{-1}\phi(z^h) \,\big|\, x^1 = x\big].
\end{aligned}$$

If the sufficient "coverage" in (22) holds, then, we have

$$\begin{aligned}\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) &\le \frac{1}{c^\dagger} \max_{h\in[H]} \mathrm{Tr}\Big\{(\Lambda^h)^{-1} \cdot \big(I + c^\dagger n \mathbb{E}_{\pi_*}\big[\phi(z^h)\phi(z^h)^\top \,\big|\, x^1 = x\big]\big)\Big\} \\ &\le \frac{d}{c^\dagger} < \infty,\end{aligned}$$

which concludes the proof. □

B.3 Connections between Weighted and Unweighted Coverage Coefficients

We use the shorthand notation $z = (x, a)$ for any $(x, a) \in \mathcal{X} \times \mathcal{A}$. Recall the definition of weighted coverage coefficient in (12):

$$\mathrm{CC}^\sigma(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) = \max_{h\in[H]} \mathbb{E}_{\pi_*}\bigg[\sup_{f,f'\in\widehat{\mathcal{F}}^h} \frac{n(f(z^h) - f'(z^h))^2/(\sigma^h(z^h))^2}{\lambda + \sum_{i=1}^n (f(z_i^h) - f'(z_i^h))^2/(\sigma_i^h)^2} \,\bigg|\, x^1 = x\bigg],$$

where

$$(\sigma^h(z^h))^2 = \max\bigg(1, \sup_{f,f'\in\widehat{\mathcal{F}}^h} \frac{|f(z^h) - f'(z^h)|/\alpha}{\sqrt{\lambda + \sum_{i=1}^n (f(z_i^h) - f'(z_i^h))^2/(\sigma_i^h)^2}}\bigg).$$

In the sequel, we explore the relationship between the weighted coverage coefficient $\mathrm{CC}^\sigma(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H)$ and the unweighted coverage coefficient $\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H)$ defined in (9):

$$\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) = \max_{h\in[H]} \mathbb{E}_{\pi_*}\bigg[\sup_{f,f'\in\widehat{\mathcal{F}}^h} \frac{n(f(z^h) - f'(z^h))^2}{\lambda + \sum_{i=1}^n (f(z_i^h) - f'(z_i^h))^2} \,\bigg|\, x^1 = x\bigg].$$

Let μ^h be the empirical data distribution:

$$\sum_{i=1}^n (f(z_i^h) - f'(z_i^h))^2/(\sigma_i^h)^2 = n \sum_{z^h \in \mathcal{X}\times\mathcal{A}} \mu(z^h) \cdot \frac{(f(z^h) - f'(z^h))^2}{(\sigma^h(z^h))^2}.$$

Now, we present the proof of Lemma 4.1.

Proof of Lemma 4.1. According to the proof of Lemma 3.1, for each $i \in [n], h \in [H]$, the weight $(\sigma_i^h)^2$ is upper bounded by

$$(\sigma_i^h)^2 \le \max(1, 1/(\alpha\sqrt{\lambda})).$$

If $\alpha\sqrt{\lambda} > 1$, we have

$$\mathrm{CC}^\sigma(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) = \mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H).$$

If $\alpha\sqrt{\lambda} \le 1$, we have

$$\begin{aligned}\mathrm{CC}^\sigma(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) &= \max_{h\in[H]} \sum_{z^h} d_{\pi_*}^h(z^h) \sup_{f,f'\in\widehat{\mathcal{F}}^h} \frac{n(f(z^h) - f'(z^h))^2/(\sigma^h(z^h))^2}{\lambda + n\sum_{\bar{z}^h} \mu(\bar{z}^h)(f(\bar{z}^h) - f'(\bar{z}^h))^2/(\sigma^h(\bar{z}^h))^2} \\ &\le \max_{h\in[H]} \sum_{z^h} d_{\pi_*}^h(z^h) \sup_{f,f'\in\widehat{\mathcal{F}}^h} \frac{(f(z^h) - f'(z^h))^2/(\sigma^h(z^h))^2}{\lambda/n + \alpha\sqrt{\lambda}\sum_{\bar{z}^h} \mu(\bar{z}^h)(f(\bar{z}^h) - f'(\bar{z}^h))^2} \\ &\le \frac{1}{\alpha\sqrt{\lambda}} \max_{h\in[H]} \sum_{z^h} d_{\pi_*}^h(z^h) \sup_{f,f'\in\widehat{\mathcal{F}}^h} \frac{(f(z^h) - f'(z^h))^2/(\sigma^h(z^h))^2}{\lambda/n + \sum_{\bar{z}^h} \mu(\bar{z}^h)(f(\bar{z}^h) - f'(\bar{z}^h))^2},\end{aligned} \tag{23}$$

where we omit the condition $\,|\, x^1 = x$ from the distribution $d_{\pi_*}^h(\cdot \,|\, x^1 = x)$ when there is no confusion, and the last inequality uses $\alpha\sqrt{\lambda} \le 1$. For any $z^h \sim d_{\pi_*}^h(z^h)$, we take the $f_{z^h}, f'_{z^h} \in \widehat{\mathcal{F}}^h$ that maximize the term:

$$\frac{(f_{z^h}(z^h) - f'_{z^h}(z^h))^2/(\sigma^h(z^h))^2}{\lambda/n + \sum_{\bar{z}^h} \mu(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2}. \tag{24}$$

Then, we can write (23) as

$$\mathrm{CC}^{\sigma}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) \leq \frac{1}{\alpha\sqrt{\lambda}} \max_{h\in[H]} \sum_{z^h} d_{\pi_*}^h(z^h) \frac{(f_{z^h}(z^h) - f'_{z^h}(z^h))^2/(\sigma^h(z^h))^2}{\lambda/n + \sum_{\bar{z}^h} \mu(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2}. \tag{25}$$

To further bound (25), we need to lower bound $(\sigma^h(z^h))^2$ in (13) by

$$\begin{aligned}\sup_{f,f'\in\widehat{\mathcal{F}}^h} \frac{|f(z^h) - f'(z^h)|/\alpha}{\sqrt{\lambda + \sum_{i=1}^n (f(z_i^h) - f'(z_i^h))^2/(\sigma_i^h)^2}} &\geq \frac{|f_{z^h}(z^h) - f'_{z^h}(z^h)|/\alpha}{\sqrt{\lambda + \sum_{i=1}^n (f_{z^h}(z_i^h) - f'_{z^h}(z_i^h))^2/(\sigma_i^h)^2}} \\ &\geq \frac{|f_{z^h}(z^h) - f'_{z^h}(z^h)|}{2\alpha\beta^h},\end{aligned}$$

where the last inequality is due to the confidence interval:

$$\begin{aligned}&\lambda + \sum_{i=1}^n \frac{(f_{z^h}(z_i^h) - f'_{z^h}(z_i^h))^2}{(\sigma_i^h)^2} \\ &= \lambda + \sum_{i=1}^n \frac{(f_{z^h}(z_i^h) - \hat{f}(z_i^h) + \hat{f}(z_i^h) - f'_{z^h}(z_i^h))^2}{(\sigma_i^h)^2} \\ &\leq 2\Big[\lambda + \sum_{i=1}^n \frac{(f_{z^h}(z_i^h) - \hat{f}(z_i^h))^2}{(\sigma_i^h)^2} + \lambda + \sum_{i=1}^n \frac{\hat{f}(z_i^h) - f'_{z^h}(z_i^h))^2}{(\sigma_i^h)^2}\Big] \leq 4(\beta^h)^2.\end{aligned}$$

Hence, by taking this lower bound into (25) and taking $\beta^h = C_\beta\sqrt{\ln N} = C_\beta\sqrt{\lambda}$, we get

$$\begin{aligned}&\mathrm{CC}^{\sigma}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) \\ &\leq \frac{2\alpha\beta^h}{\alpha\sqrt{\lambda}} \max_{h\in[H]} \sum_{z^h} d_{\pi_*}^h(z^h) \frac{|f_{z^h}(z^h) - f'_{z^h}(z^h)|}{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2} \\ &= 2C_\beta \max_{h\in[H]} \underbrace{\sum_{z^h} d_{\pi_*}^h(z^h) \frac{|f_{z^h}(z^h) - f'_{z^h}(z^h)|}{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2} \cdot \mathbb{1}\big(\|f_{z^h} - f'_{z^h}\|_\infty \leq n^{-1}\big)}_{(a)} \\ &\quad + 2C_\beta \max_{h\in[H]} \sum_{z^h} d_{\pi_*}^h(z^h) \frac{|f_{z^h}(z^h) - f'_{z^h}(z^h)|}{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2} \cdot \mathbb{1}\big(\|f_{z^h} - f'_{z^h}\|_\infty > n^{-1}\big).\end{aligned} \tag{26}$$

The term (a) can be bounded by

$$(a) \leq \sum_{z^h} d_{\pi_*}^h(z^h) \frac{n^{-1}}{\lambda n^{-1}} = \frac{1}{\lambda},$$

which is taken back into (26) to obtain

$$
\begin{aligned}
&\mathrm{CC}^{\sigma}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) \\
&\leq \frac{2C_\beta}{\lambda} + 2C_\beta \max_{h\in[H]} \sum_{z^h} d^h_{\pi_*}(z^h) \frac{|f_{z^h}(z^h) - f'_{z^h}(z^h)|}{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2} \cdot \mathbb{1}\big(\|f_{z^h} - f'_{z^h}\|_\infty > n^{-1}\big) \\
&\leq \frac{2C_\beta}{\lambda} + 2C_\beta \max_{h\in[H]} \sum_{z^h} \frac{\sqrt{d^h_{\pi_*}(z^h)}}{\sqrt{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2}} \\
&\quad \cdot \frac{\sqrt{d^h_{\pi_*}(z^h)}|f_{z^h}(z^h) - f'_{z^h}(z^h)|}{\sqrt{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2}} \cdot \mathbb{1}\big(\|f_{z^h} - f'_{z^h}\|_\infty > n^{-1}\big) \\
&\leq \frac{2C_\beta}{\lambda} + 2C_\beta \max_{h\in[H]} \underbrace{\Big(\sum_{z^h} \frac{d^h_{\pi_*}(z^h)}{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2} \cdot \mathbb{1}\big(\|f_{z^h} - f'_{z^h}\|_\infty > n^{-1}\big)\Big)^{1/2}}_{I_1} \\
&\quad \cdot \underbrace{\Big(\max_{h\in[H]} \sum_{z^h} d^h_{\pi_*}(z^h) \frac{(f_{z^h}(z^h) - f'_{z^h}(z^h))^2}{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2}\Big)^{1/2}}_{I_2}, \qquad (27)
\end{aligned}
$$

where the second equality is by the Cauchy-Schwarz inequality. Then, we handle the terms I_1 and I_2 separately. For the term I_2, since f_{z^h}, f'_{z^h} maximize the term (24), they also maximize the term

$$
\frac{(f(z^h) - f'(z^h))^2}{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f(\bar{z}^h) - f'(\bar{z}^h))^2},
$$

which indicates that I_2 can be written as

$$
\begin{aligned}
I_2 &= \max_{h\in[H]} \sum_{z^h} d^h_{\pi_*}(z^h) \sup_{f,f'\in\widehat{\mathcal{F}}^h} \frac{n(f(z^h) - f'(z^h))^2}{\lambda + \sum_{i=1}^n (f(z_i^h) - f'(z_i^h))^2} \\
&\leq \mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H).
\end{aligned}
$$

To bound the term I_1, for any distinct $f, f' \in \widehat{\mathcal{F}}^h$ and $\|f - f'\|_\infty > n^{-1}$, there exists two distinct functions $\bar{f}, \bar{f}' \in \bar{\mathcal{F}}^h$ such that $\|f - \bar{f}\|_\infty \leq n^{-1}$ and $\|f' - \bar{f}'\|_\infty \leq n^{-1}$. Then, we get

$$
\begin{aligned}
&\Big|\sum_{z^h} \mu^h(z)\big(f(z^h) - f'(z^h)\big)^2 - \sum_{z^h} \mu^h(z)\big(\bar{f}(z^h) - \bar{f}'(z^h)\big)^2\Big| \\
&\qquad \leq \Big|\sum_{z^h} \mu^h(z)\big(f(z^h) - \bar{f}(z^h) - (f'(z^h) - \bar{f}'(z^h))\big)\big(f(z^h) - f'(z^h) + \bar{f}(z^h) - \bar{f}'(z^h)\big)\Big| \\
&\qquad \leq \sum_{z^h} \mu^h(z^h) \cdot \frac{2}{n} \cdot 2 = \frac{4}{n},
\end{aligned}
$$

where the last inequality is because $\bar{f}, \bar{f}'$ satisfy (10) and $f \in [0,1]$ for any $f \in \mathcal{F}^h$. It follows that

$$
\begin{aligned}
\sum_{z^h} \mu^h(z)\big(f(z^h) - f'(z^h)\big)^2 &\geq \sum_{z^h} \mu^h(z)\big(\bar{f}(z^h) - \bar{f}'(z^h)\big)^2 - \frac{4}{n} \\
&\geq C(\widehat{\mathcal{F}}, \mu) - \frac{4}{n} \geq \frac{C(\widehat{\mathcal{F}}, \mu)}{2}. \qquad (28)
\end{aligned}
$$

Therefore, we apply the above result with $f = f_{z^h},\ f' = f'_{z^h}$ to bound the term I_1 as

$$
I_1 \leq \sum_{z^h} \frac{d^h_{\pi_*}(z^h)}{C(\widehat{\mathcal{F}}, \mu)/2} = \frac{2}{C(\widehat{\mathcal{F}}, \mu)},
$$

where the last equality uses $\sum_{z^h} d_{\pi_*}^h(z^h) = 1$. Therefore, by taking the bound of terms I_1, I_2 into (27), we have

$$\mathrm{CC}^\sigma(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) = \frac{2C_\beta}{\lambda} + 2C_\beta \sqrt{\frac{2\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H)}{C(\widehat{\mathcal{F}}, \mu)}} = \tilde{\mathcal{O}}\Big(\sqrt{\frac{\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H)}{C(\widehat{\mathcal{F}}, \mu)}}\Big).$$

Finally, we will analyze the relationship between $\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H)$ and $C(\widehat{\mathcal{F}}, \mu)$. Let $f_{z^h}, f'_{z^h} \in \widehat{\mathcal{F}}^h$ be the maximizer of

$$\frac{n(f(z^h) - f'(z^h))^2}{\lambda + \sum_{i=1}^n (f(z_i^h) - f'(z_i^h))^2}.$$

Then, we have

$$\begin{aligned}
&\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) \\
&\quad = \max_{h \in [H]} \sum_{z^h} d_{\pi_*}^h(z^h) \frac{(f_{z^h}(z^h) - f'_{z^h}(z^h))^2}{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2} \cdot \mathbb{1}\big(\|f_{z^h} - f'_{z^h}\|_\infty \le n^{-1}\big) \\
&\qquad + \max_{h \in [H]} \sum_{z^h} d_{\pi_*}^h(z^h) \frac{(f_{z^h}(z^h) - f'_{z^h}(z^h))^2}{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2} \cdot \mathbb{1}\big(\|f_{z^h} - f'_{z^h}\|_\infty > n^{-1}\big) \\
&\quad \le \frac{1}{\lambda n} + \sum_{z^h} \frac{d_{\pi_*}^h(z^h)}{C(\widehat{\mathcal{F}}, \mu)/2} \\
&\quad = O((C(\widehat{\mathcal{F}}, \mu))^{-1}),
\end{aligned}$$

where the inequality holds due to (28). It follows that

$$\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) = O\Big(\sqrt{\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H)/C(\widehat{\mathcal{F}}, \mu)}\Big),$$

which concludes the proof. □

Interpretation of Assumption 4.1 in linear MDPs. Now, we illustrate the condition (10) with the linear model. When the function space $\mathcal{F}^h$ can be embedded into a d-dimensional vector space: $\mathcal{F}^h = \{\langle w(f), \phi(\cdot)\rangle : z \to \mathbb{R}\}$, the condition (10) becomes: for any two distinct $f, f' \in \bar{\mathcal{F}}^h$,

$$\sum_{z^h} \mu^h(z)\big(f(z^h) - f'(z^h)\big)^2 = \big(w(f) - w(f')\big)^\top \bar{\Lambda}^h \big(w(f) - w(f')\big) \ge C(\widehat{\mathcal{F}}, \mu). \tag{29}$$

where μ^h is the data empirical distribution.

In the following lemma, we demonstrate that the above condition holds as long as the learner has excess to a well-explored dataset (30), which is a wildly-adopted assumption in the literature of offline linear MDPs [11, 46, 63]. Note that d^{-1} is the largest possible order of the minimum eigenvalue since for any data distribution $\tilde{\mu}^h$, $\sigma_{\min}(\mathbb{E}_{z^h \sim \mu^h}[\phi(z)\phi(z)^\top]) \le d^{-1}$ by using $\|\phi(z)\| \le 1$ for any $z \in \mathcal{X} \times \mathcal{A}$.

Lemma B.2. *In the linear setting, if we assume that the data distributions μ^h satisfy the following minimum eigenvalue condition: there exists an absolute constant $\bar{c} > 0$ such that*

$$\sigma_{\min}\Big(\mathbb{E}_{z^h \sim \mu^h}[\phi(z)\phi(z)^\top]\Big) = \frac{\bar{c}}{d}, \tag{30}$$

the dataset $\mathcal{D}$ consists of $n \ge 128 d^2 \bar{c}^{-2} \log(d/(2\delta)$ independent trajectories, then, the condition (29) *with $C(\widehat{\mathcal{F}}, \mu) = a\bar{c}/(2d)$ will holds with probability at least $1 - \delta$.*

Proof. To begin with, we aim to prove that the empirical matrix $\bar{\Lambda}^h = n^{-1} \sum_{i=1}^n \phi(z_i^h)\phi(z_i^h)^\top$ is positive definite with high probability. Since $\|\phi(z)\| \le 1$ for any $z \in \mathcal{X} \times \mathcal{A}$, we have for each $i \in [n]$,

$$\big(\phi(z_i^h)\phi(z_i^h)^\top - \mathbb{E}_{z^h \sim \mu^h}[\phi(z)\phi(z)^\top]\big)^2 \preceq (2I)^2.$$

By invoking the matrix Hoeffding's concentration in Lemma G.4 with $X_i = \phi(z_i^h)\phi(z_i^h)^\top - \mathbb{E}_{z^h \sim \mu^h}[\phi(z)\phi(z)^\top]$, $A_i = 2I$, $\sigma^2 = 4n$, we obtain

$$\mathbb{P}\Big(\Big\| \sum_{i=1}^n \big(\phi(z_i^h)\phi(z_i^h)^\top - \mathbb{E}_{z^h\sim\mu^h}[\phi(z)\phi(z)^\top]\big)\Big\|_{\mathrm{op}}\Big) \geq t\Big) \leq 2d \cdot e^{-t^2/(32n)}.$$

For any $\delta > 0$, by taking $t = \sqrt{32n\log(d/\delta)}$, we have with probability at least $1-\delta$,

$$\Big\|\frac{1}{n}\sum_{i=1}^n \big(\phi(z_i^h)\phi(z_i^h)^\top - \mathbb{E}_{z^h\sim\mu^h}[\phi(z)\phi(z)^\top]\big)\Big\|_{\mathrm{op}} \leq \sqrt{\frac{32\log(d/(2\delta))}{n}}.$$

Hence, whenever $n \geq 128 d^2 \bar{c}^{-2}\log(d/(2\delta))$, by combing the above result with (30), we have with probability at least $1-\delta$,

$$\sigma_{\min}\big(\bar{\Lambda}^h\big) = \sigma_{\min}\Big(\frac{1}{n}\sum_{i=1}^n \phi(z_i^h)\phi(z_i^h)^\top\Big) \geq \frac{\bar{c}}{2d}. \tag{31}$$

Then, since the cardinality of the cover $\bar{\mathcal{F}}^h$ is $\tilde{\mathcal{O}}(d)$, we can define $a = \min_{f,f'\bar{\mathcal{F}}^h} \|w(f) - w(f')\|^2$. Thus, by using (31), the condition (29) is inferred: for any $f, f'\bar{\mathcal{F}}^h$ with probability at least $1-\delta$,

$$\big(w(f) - w(f')\big)^\top \bar{\Lambda}^h \big(w(f) - w(f')\big) \geq \frac{\bar{c}}{2d}\big\|w(f) - w(f')\big\|^2 \geq \frac{a\bar{c}}{2d} = C(\widehat{\mathcal{F}}, \mu).$$

□

B.4 Lower Bound for Linear MDPs with Corruption

Proof of Theorem 2. For any dimension d, step horizon H and corruption level ζ, we construct a tabular MDP with action number $A > 2$ and state number $S \leq (A/2)^{H/2}$ such that $SA = d$ by following the proof of Theorem 3.1 in [60]. Particularly, since we assume that $V^h(x) \in [0,1]$ for any $h \in [H]$ and $x \in \mathcal{X}$, we set all the rewards equaling to 1 to $1/H$. Then, by taking the corruption fraction as $\epsilon = \zeta/n$, an ϵ-contamination adversary can perturb all the chosen rewards from $1/H$ to 0. Thus, the learner must suffer from at least $\Omega(SA\epsilon) = \Omega(SA\zeta n^{-1})$ suboptimality with probability at least $1/4$. □

B.5 Relationship between the Bootstrapped Uncertainty and Bonus

In the sequel, we discuss the relationship between the bootstrapped uncertainty and the bonus function by considering linear function approximation. Using z to denote the feature variable of state-action pair (x, a), we estimate the Q-value function Q_ϕ by $Q_{w^h}^h(z) = (w^h)^\top\phi(z)$ to minimize the Bellman error target with weights:

$$\hat{w}^h = \operatorname*{argmin}_{w\in\mathbb{R}^d} \sum_{z_i^h\in\mathcal{D}} \frac{\big((\widehat{\mathcal{T}}^h Q_w)(z_i^h) - w^\top\phi(z_i^h)\big)^2}{\sigma(z_i^h)}, \tag{32}$$

where $\{\sigma(z_i^h) > 0\}$ is a group of predetermined weights and for any $z_i^h \in \mathcal{D}$,

$$(\widehat{\mathcal{T}}^h Q_w)(z_i^h) = r^h(z_i^h) + V_w^{h+1}(z_i^{h+1}).$$

Additionally, we define the noise in this weighted least square problem as $\epsilon = \widehat{\mathcal{T}}^h Q_w(x,a) - Q_w^h(z)$.

Bonus functions. In the traditional linear MDP, we often use the following term as the bonus function:

$$b_L^h(z) = \sqrt{\phi(z)^\top(\Lambda^h)^{-1}\phi(z)}, \tag{33}$$

where $\Lambda^h = \sum_{i=1}^n \phi(z_i^h)\phi(z_i^h)^\top/(\sigma_i^h)^2$, and we use the shorthand notation for any matrix A and vector x: $\|x\|_A = \sqrt{x^\top Ax}$.

The bonus function in the general form (8) turns into the following form under the linear setting:

$$b^h(z) = \sup_{f,f' \in \widehat{\mathcal{F}}^h} \frac{\left|\big(w(f) - w(f')\big)^\top \phi(z)\right|}{\sqrt{\lambda + \sum_{i=1}^n \big((w(f) - w(f'))^\top \phi(z_i^h)/\sigma_i^h\big)^2}}.$$

We demonstrate that the linear and general forms of bonus functions are almost equivalent under mild conditions.

Lemma B.3. *Under the linear MDP, if the function space is broad enough such that for any $z \in \mathcal{X} \times \mathcal{A}$, there exists $f, f' \in \widehat{\mathcal{F}}^h$ satisfying that $w(f) - w(f')$ and $(\Lambda^h)^{-1}\phi(z)$ are in the same direction and not too close, i.e., for some $\alpha > 0$,*

$$w(f) - w(f') = \alpha \cdot (\Lambda^h)^{-1}\phi(z), \quad \textit{and} \quad \|w(f) - w(f')\|_{\Lambda^h}^2 \geq \lambda,$$

then, we have for any $z \in \mathcal{X} \times \mathcal{A}$,

$$\frac{b_L^h(z)}{\lambda^{1/4} + 1} \leq b^h(z) \leq b_L^h(z).$$

Proof. First, we will prove $b^h(z) \leq b_L^h(z)$. By the definition of b^h, we have

$$\begin{aligned}
b^h(z) = &\sup_{f,f' \in \widehat{\mathcal{F}}^h} \frac{\left|\big(w(f) - w(f')\big)^\top \phi(z)\right|}{\sqrt{\lambda + \sum_{i=1}^n \big((w(f) - w(f'))^\top (\phi(z_i^h)/\sigma_i^h\big)^2}} \\
\leq &\sup_{f,f' \in \widehat{\mathcal{F}}^h} \frac{\left|\big(w(f) - w(f')\big)^\top \phi(z)\right|}{\sqrt{\big(w(f) - w(f')\big)^\top \Lambda^h \big(w(f) - w(f')\big)}} \\
\leq &\sqrt{\phi(z)^\top (\Lambda^h)^{-1} \phi(z)} = b_L^h(z).
\end{aligned}$$

Then, we will prove $b^h(z) \geq b_L^h(z)$. By the assumption, for any $z \in \mathcal{X} \times \mathcal{A}$, there exists $f_1, f_2 \in \widehat{\mathcal{F}}^h$ such that for some $\alpha > 0$,

$$w(f_1) - w(f_2) = \alpha \cdot (\Lambda^h)^{-1}\phi(z),$$

which implies that

$$(\Lambda^h)^{1/2}(w(f_1) - w(f_2)) = \alpha \cdot (\Lambda^h)^{-1/2}\phi(z).$$

Then, we have

$$\begin{aligned}
b^h(z) = &\sup_{f,f' \in \widehat{\mathcal{F}}^h} \frac{\left|\big(w(f) - w(f')\big)^\top \phi(z)\right|}{\sqrt{\lambda + \big(w(f) - w(f')\big)^\top \Lambda^h \big(w(f) - w(f')\big)}} \\
\geq &\frac{\left|\big(w(f_1) - w(f_2)\big)^\top \phi(z)\right|}{\sqrt{\lambda + \big(w(f_1) - w(f_2)\big)^\top \Lambda^h \big(w(f_1) - w(f_2)\big)}} \\
= &\frac{\left\|(\Lambda^h)^{1/2}(w(f_1) - w(f_2))\right\| \cdot \left\|(\Lambda^h)^{-1/2}\phi(z)\right\|}{\|w(f) - w(f')\|_{\Lambda^h}} \cdot \frac{\|w(f) - w(f')\|_{\Lambda^h}}{\sqrt{\lambda + \|w(f) - w(f')\|_{\Lambda^h}^2}} \\
= &\frac{\left\|(\Lambda^h)^{1/2}(w(f_1) - w(f_2))\right\| \cdot \left\|(\Lambda^h)^{-1/2}\phi(z)\right\|}{\|w(f) - w(f')\|_{\Lambda^h}} \cdot \frac{1}{\sqrt{\lambda/\|w(f) - w(f')\|_{\Lambda^h}^2 + 1}}. \qquad (34)
\end{aligned}$$

Since $\|w(f) - w(f')\|_{\Lambda^h}^2 \geq \sqrt{\lambda}$, we have

$$\sqrt{\frac{\lambda}{\|w(f) - w(f')\|_{\Lambda^h}^2} + 1} \leq \sqrt{\lambda^{1/2} + 1} \leq \lambda^{1/4} + 1.$$

Taking this result back into (34) leads to

$$
\begin{aligned}
b^h(a) &\geq \left\|(\Lambda^h)^{-1/2}\phi(z)\right\| \cdot \frac{1}{\sqrt{\lambda/\|w(f)-w(f')\|_{\Lambda^h}^2+1}} \\
&\geq \frac{\sqrt{\phi(z)^\top(\Lambda^h)^{-1}\phi(z)}}{\lambda^{1/4}+1} = \frac{b_L^h(z)}{\lambda^{1/4}+1}.
\end{aligned}
$$

Therefore, we conclude the proof. □

Connection between the bootstrapped uncertainty and bonus functions. We begin with illustrating the equivalence between the bootstrapped uncertainty and the linear form of the bonus in the following lemma. Since we actually compute the uncertainty weights for only single iteration, we let $\sigma_i^h = 1$.

Lemma B.4. *For any $z \in \mathcal{X} \times \mathcal{A}$,*

$$
\mathit{Var}_{\hat{w}^h}(Q^h_{\hat{w}^h}(z)) = \mathit{Var}_{\hat{w}^h}(z^\top \hat{w}^h) = z^\top(\Lambda^h)^{-1}z.
$$

Proof. Let $y_i^h = r^h(z_i^h) + V_w^{h+1}(z_i^{h+1})$. Under the assumption that $\epsilon_i^h \sim N(0,1)$, since the closed form solution to the problem is

$$
\hat{w}^h = (\Lambda^h)^{-1}\sum_{z_i^h\in\mathcal{D}} y_i^h z_i^h = (\Lambda^h)^{-1}\sum_{z_i^h\in\mathcal{D}}(Q_\phi^h(z_i^h)+\epsilon_i^h)z_i^h,
$$

we obtain that

$$
\hat{w}^h|\mathcal{D} \sim N(\mu^h, (\Lambda^h)^{-1}),
$$

where

$$
\mu^h = (\Lambda^h)^{-1}\sum_{z_i^h\in\mathcal{D}} Q_\phi^h(z_i^h)z_i^h, \quad \Lambda^h = \sum_{i=1}^n z_i^h(z_i^h)^\top.
$$

Then, it follows that for any $z \in \mathcal{A} \times \mathcal{A}$,

$$
\mathrm{Var}_{\hat{w}^h}(Q^h_{\hat{w}^h}(z)|\mathcal{D}) = \mathrm{Var}_{\hat{w}^h}(z^\top\hat{w}^h|\mathcal{D}) = z^\top(\Lambda^h)^{-1}z.
$$

Hence, we complete the proof. □

In Lemma B.4, we find that the standard deviation of the Q-value function is equivalent to the linear LCB-bonus $b_L^h(z) = \sqrt{z^\top(\Lambda^h)^{-1}z}$. Moreover, recall that the bootstrapped uncertainty $\mathbb{V}_{j=1,\ldots,N}\left[Q_{w_j}\right]$ is the standard deviation of the bootstrapped Q-value functions. Therefore, according to Osband et al. [34] our proposed bootstrapped uncertainty can serve as an estimation for the bonus function b_L^h under the linear MDP setting. Theoretically, we can use the uncertainty weight iteration (Algorithm 1) to construct the weighted bootstrap uncertainty.

By combining Lemma B.3 and Lemma B.4, we conclude that by taking a sufficiently broad function approximation class $\widehat{\mathcal{F}}^h$ and sufficiently small parameter λ, the proposed bootstrapped uncertainty is an estimation of the general form of the bonus in linear MDPs. More importantly, in the experiments, the estimation of the uncertainty is simplified and shares the spirit with the theoretical analysis due to two reasons: 1) due to the complexity of the nonlinear version of uncertainty, which is expressed as

$$
\sup_{f,f'\in\mathcal{F}} \frac{|f(z_i)-f'(z_i)|}{\sqrt{\lambda+\sum_{j=1}^n(f(z_j)-f'(z_j))^2/\sigma_j^2}},
$$

we simplify the estimation by using the bootstrap uncertainty, which is an unbiased estimation of uncertainty in the linear version, and 2) combining the uncertainty weight iteration and the bootstrap uncertainty estimation is cumbersome, so we only iterate once during simulations.

C Results for Distribution Shift

In this section, we consider an MDP$(\mathcal{X}, \mathcal{A}, H, \mathbb{P}, r)$ and an offline dataset $\mathcal{D} = \{(x_i^h, a_i^h)\}_{i,h=1}^H$ with adversarial corruption and distribution shift. Specifically, for each trajectory $\{(x_i^h, a_i^h)\}_{h=1}^H$, we define $\rho_i > 0$ to measure the ditribution shift of this trajectory. For example, when the learner's goal varies from the training data, the goal shift can be embedded into the initial state x^1 and captured by $\rho(x^1)$. Hence, we define a new notion of corruption level ζ, capturing both adversarial corruption and distribution shift.

Definition C.1 (Cumulative Corruption). *The cumulative corruption is ζ if at any step $h \in [H]$, for any sequence $\{x_i^h, a_i^h\}_{i,h=1}^{n,H} \subset \mathcal{X} \times \mathcal{A}$, $\{\rho_i\}_{i=1}^n \subset \mathbb{R}^+$ and function $\{g^h : \mathcal{X} \to [0,1]\}_{h=1}^H$, we have*

$$\zeta = \sum_{h=1}^H \sum_{i=1}^n \rho_i |\zeta_i^h|, \quad \zeta_i^h = (\mathcal{T}^h g - \mathcal{T}_{\mathcal{D}}^h g)(x_i^h, a_i^h).$$

Algorithm 3 Uncertainty Weight Iteration with Distribution Shift

```
Input: $\{(x_i, a_i, \rho_i)\}_{i=1}^n, \mathcal{F}, \alpha > 0$
Initialization: $t = 0,\ \sigma_i^0 = 1, i = 1, \dots, n$
repeat
    $t \leftarrow t + 1$
    $(\sigma_i^t)^2 \leftarrow \max\left(1, \sup_{f,f' \in \mathcal{F}} \frac{|f(x_i,a_i) - f'(x_i,a_i)|/(\alpha\rho_i)}{\sqrt{\lambda + \sum_{j=1}^n (f(x_j,a_j) - f'(x_j,a_j))^2/(\sigma_j^{t-1})^2}}\right), i = 1, \dots, n$
until $\max_{i \in [n]} \left(\sigma_i^t / \sigma_i^{t-1}\right)^2 \le 2$
Output: $\{\sigma_i^t\}_{i=1}^n$
```

The main challenge in handling distribution shifts is the new weight design. We propose uncertainty weight iteration with distribution shift in Algorithm 3, where we put ρ_i on the denominator. Similarly, we can follow Lemma 3.1 to demonstrate that the iteration converges and the output weights $\{\sigma_i := \sigma_i^N\}_{i=1}^n$ satisfy

$$\sigma_i^2 \ge \max\Big(1, \frac{1}{2} \sup_{f,f' \in \mathcal{F}} \frac{|f(x_i,a_i) - f'(x_i,a_i)|/(\alpha\rho_i)}{\sqrt{\lambda + \sum_{j=1}^n (f(x_j,a_j) - f'(x_j,a_j))^2/\sigma_j^2}}\Big). \tag{35}$$

Therefore, just by replacing the weight iteration (Algorithm 1) in CR-PEVI Algorithm 2 with Algorithm 3, we obtain an algorithm robust to both corruption and distribution shift, named as **CORDS-PEVI**. Because CORDS-PEVI highly repeats Algorithm 2, we do not present the pseudo-code of the algorithm.

Then, the suboptimality bound achieved by CORDS-PEVI is presented in the following theorem.

Theorem 3. *If the coverage coefficient in Definition 4.1 is finite and Assumption 4.1 holds, under CORDS-PEVI, for any cumulative corruption ζ and $\delta > 0$, we choose the covering parameter $\gamma = 1/(n \max_h \beta^h \zeta^h)$, the eluder parameter $\lambda = \ln(N_n(\gamma))$, the weighting parameter $\alpha = H\sqrt{\ln N_n(\gamma)}/\zeta$, and the confidence radius*

$$\beta^h = c_\beta \left(\alpha\zeta^h + \sqrt{\ln(HN_n(\gamma)/\delta)}\right) \textit{ for } h = H, \dots, 1,$$

where

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

Then, with probability at least $1 - 3\delta$, the sub-optimality is bounded by

$$\mathrm{SubOpt}(\hat{\pi}, x) = \tilde{\mathcal{O}}\Big(\frac{H(\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H))^{1/4} \cdot (\ln N_n(\gamma))^{1/2}}{n^{1/2}(C(\widehat{\mathcal{F}}, \mu))^{1/4}} + \frac{\zeta(\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H))^{1/2}}{n(C(\widehat{\mathcal{F}}, \mu))^{1/2}}\Big).$$

C.1 Analysis of the Result

The main difference in the analysis between the model with and without a distribution shift is the bound for the confidence radius.

C.1.1 Sharp bound of the confidence radius.

Lemma C.1 (Confidence Radius)**.** *In CORDS-PEVI, for all $h \in [H]$ we have $\mathcal{T}^h f_n^{h+1} \in \widehat{\mathcal{F}}^h$ with probability at least $1-\delta$, where*

$$\beta^h = 12\alpha\zeta^h + \big(12\lambda + 12\ln(2HN_n^h(\gamma)/\delta) + 12(5\beta^{h+1}\gamma)^2 n + 60\beta^{h+1}\gamma\sqrt{nC_1^h(n,\zeta)}\big)^{1/2},$$

we use the notation

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

and $C_1^h(n,\zeta) = 2(\sum_{i=1}^n (\zeta_i^h)^2 + 2n\eta^2 + 3\eta^2\ln(2/\delta))$, $\zeta^h = \sum_{i=1}^n \rho_i\zeta_i^h$.

Proof. We use similar methods as the proof of Lemma A.2. At each step $h \in [H]$, by notating $\mathcal{F}_\gamma^{h+1}$ as a $(\gamma, \|\cdot\|_\infty)$ cover of $\mathcal{F}^{h+1}$, and $\mathcal{B}_\gamma^{h+1}$ as a $(\gamma, \|\cdot\|_\infty)$ cover of $\mathcal{B}^{h+1}(\lambda)$, 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 $((1+\beta^{h+1})\gamma, \|\cdot\|_\infty)$ cover of $f_n^{h+1}(\cdot)$. For the f_n^{h+1}, there exists a $\bar{f}^{h+1} \in \bar{\mathcal{F}}_\gamma^{h+1}$ such that $\|\bar{f}^{h+1} - f_n^{h+1}\|_\infty \le \bar{\epsilon} = (1+\beta^{h+1})\gamma$. Then, we define $y_i^h = r_i^h + f_n^{h+1}(x_i^{h+1})$, $\bar{y}_i^h = r_i^h + \bar{f}^{h+1}(x_i^{h+1})$ and

$$\tilde{f}^h = \operatorname*{argmin}_{f^h \in \mathcal{F}^h} \sum_{i=1}^n (f^h(x_i^h, a_i^h) - \bar{y}_i^h)^2.$$

Since (18) also holds true, we can invoke Lemma G.3 by taking f_* as $\mathbb{E}[\bar{y}_i^h | x_i^h, a_i^h] = (\mathcal{T}_{\mathcal{D}}^h \bar{f}^{h+1})(x_i^h, a_i^h)$ and f_b as $\mathcal{T}^h \bar{f}^{h+1}$. With probability at least $1-\delta$, we obtain:

$$\begin{aligned}
&\sum_{i=1}^n \frac{\Big(\hat{f}^h(x_i^h, a_i^h) - (\mathcal{T}^h \bar{f}^{h+1})(x_i^h, a_i^h)\Big)^2}{(\sigma_i^h)^2} \\
&\qquad \le 10\ln(2HN_n^h(\gamma)/\delta) + 5\underbrace{\sum_{i=1}^n \frac{|\hat{f}^h(x_i^h, a_i^h) - (\mathcal{T}^h \bar{f}^{h+1})(x_i^h, a_i^h)| \cdot |\zeta_i^h|}{(\sigma_i^h)^2}}_{(a)} \\
&\qquad + 10(\gamma + 2\bar{\epsilon}) \cdot \Big((\gamma + 2\bar{\epsilon})n + \sqrt{nC_1(n,\zeta)}\Big),
\end{aligned} \tag{36}$$

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

From the weight design and Lemma 3.1, term (a) is bounded by

$$\begin{aligned}
\sum_{i=1}^n |\zeta_i^h| \cdot \frac{|\hat{f}^h(x_i^h, a_i^h) - (\mathcal{T}^h \bar{f}^{h+1})(x_i^h, a_i^h)|}{(\sigma_i^h)^2} &\le \sum_{i=1}^n \rho_i |\zeta_i^h| \cdot \frac{|\hat{f}^h(x_i^h, a_i^h) - (\mathcal{T}^h \bar{f}^{h+1})(x_i^h, a_i^h)|}{\rho_i(\sigma_i^h)^2} \\
&\le \zeta^h \sup_i \frac{|\hat{f}^h(x_i^h, a_i^h) - (\mathcal{T}^h \bar{f}^{h+1})(x_i^h, a_i^h)|}{\rho_i(\sigma_i^h)^2} \\
&\le 2\alpha\zeta^h \sqrt{\lambda + \sum_{i=1}^n \frac{\Big(\hat{f}^h(x_i^h, a_i^h) - (\mathcal{T}^h \bar{f}^{h+1})(x_i^h, a_i^h)\Big)^2}{(\sigma_i^h)^2}}
\end{aligned}$$

where the last inequality is obtained since σ_i^h satisfies (35). Taking this result back into (36), we finally get for all $h \in [H]$,

$$\begin{aligned}
&\sum_{i=1}^n \frac{\big(\hat{f}^h(x_i^h, a_i^h) - (\mathcal{T}^h \bar{f}^{h+1})(x_i^h, a_i^h)\big)^2}{(\sigma_i^h)^2} \\
&\qquad \le 10\ln(2HN_n^h(\gamma)/\delta) + 10\alpha\zeta^h\beta^h + 5\gamma\zeta + 10(\gamma + 2\bar{\epsilon}) \cdot ((\gamma + 2\bar{\epsilon})n + \sqrt{nC_1(t,\zeta)}) \\
&\qquad = 10\ln(2HN_n^h(\gamma)/\delta) + 10\alpha\zeta^h\beta^h + 5\gamma\zeta + 10(2\beta^{h+1}+3)^2\gamma^2 n + 10(2\beta^{h+1}+3)\gamma\sqrt{nC_1(n,\zeta)},
\end{aligned}$$

where the last euqlaity uses $\bar{\epsilon} = (1 + \beta^{h+1})\gamma$. Therefore, it follows that with probability at least $1 - \delta$,

$$
\begin{aligned}
&\left(\sum_{i=1}^{n} \frac{\left(\hat{f}_n^h(x_i^h, a_i^h) - (\mathcal{T}^h f_n^{h+1})(x_i^h, a_i^h)\right)^2}{(\sigma_i^h)^2} + \lambda\right)^{1/2} \\
&\quad \le \left(\sum_{i=1}^{n} \frac{\left(\hat{f}_n^h(x_i^h, a_i^h) - (\mathcal{T}^h \bar{f}_n^{h+1})(x_i^h, a_i^h)\right)^2}{(\sigma_i^h)^2}\right)^{1/2} + \sqrt{n}\bar{\epsilon} + \sqrt{\lambda} \\
&\quad \le \big(10\ln(2HN_n^h(\gamma)/\delta) + 10\alpha\zeta^h\beta^h + 5\gamma\zeta + 10(2\beta^{h+1} + 3)^2\gamma^2 n \\
&\qquad + 10(2\beta^{h+1} + 3)\gamma\sqrt{nC_1(n, \zeta)}\big)^{1/2} + (\beta^{h+1} + 1)\gamma\sqrt{n} + \sqrt{\lambda} \\
&\quad \le \Big(12\lambda + 12\ln(2HN_n^h(\gamma)/\delta) + 24\alpha\zeta^h\beta^h + 12(5\beta^{h+1}\gamma)^2 n + 60\beta^{h+1}\gamma\sqrt{nC_1(n, \zeta)}\Big)^{1/2} \\
&\quad \le \beta^h,
\end{aligned}
$$

which finishes the proof. □

C.1.2 Connections between weighted and unweighted coefficient.

We define the coverage coefficient incorporating the uncertainty weights under distribution shift as

$$
\mathrm{CC}^{\sigma}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) = \max_{h \in [H]} \mathbb{E}_{\pi_*}\left[\sup_{f, f' \in \widehat{\mathcal{F}}^h} \frac{n(f(x^h, a^h) - f'(x^h, a^h))^2/\sigma^h(x^h, a^h)^2}{\lambda + \sum_{i=1}^{n}(f(x_i^h, a_i^h) - f'(x_i^h, a_i^h))^2/(\sigma_i^h)^2} \,\middle|\, x^1 = x\right], \tag{37}
$$

where the weight for the trajectory induced by the optimal policy is

$$
(\sigma_*^h(x^h, a^h))^2 = \max\left(1, \sup_{f, f' \in \widehat{\mathcal{F}}^h} \frac{|f(x^h, a^h) - f'(x^h, a^h)|/\alpha}{\sqrt{\lambda + \sum_{i=1}^{n}(f(x_i^h, a_i^h) - f'(x_i^h, a_i^h))^2/(\sigma_i^h)^2}}\right), \tag{38}
$$

where we do not consider the distribution shift for the expected trajectory induced by the optimal policy π_*.

Lemma C.2. *Let $\rho_{\min} = \min_{i \in [n]} \rho_i$. Under CORDS-PEVI, Assumption 4.1, and the $\beta^h = C_\beta\sqrt{\ln N}$ (where $C_\beta > 0$ contains the logarithmic terms that can be omitted) and λ given in Theorem 3, we have*

$$
\mathrm{CC}^{\sigma}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) = \tilde{O}\big((\mathrm{CC}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H))^{1/2} \cdot (C(\widehat{\mathcal{F}}, \mu))^{-1/2}\big).
$$

Proof. We adopt the same approaches in the proof of Lemma 4.1. For each $i \in [n], h \in [H]$, since the weight $(\sigma_i^h)^2$ yielded by Algorithm 3 is upper bounded by $1/(\alpha\sqrt{\lambda}\rho_{\min})$, we get

$$
\begin{aligned}
\mathrm{CC}^{\sigma}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) &= \max_{h \in [H]} \sum_{z^h} d_{\pi_*}^h(z^h) \sup_{f, f' \in \widehat{\mathcal{F}}^h} \frac{n(f(z^h) - f'(z^h))^2/(\sigma_*^h(z^h))^2}{\lambda + n\sum_{\bar{z}^h} \mu(z^h)(f(\bar{z}^h) - f'(\bar{z}^h))^2/(\sigma^h(\bar{z}^h))^2} \\
&\le \max_{h \in [H]} \sum_{z^h} d_{\pi_*}^h(z^h) \sup_{f, f' \in \widehat{\mathcal{F}}^h} \frac{(f(z^h) - f'(z^h))^2/(\sigma_*^h(z^h))^2}{\lambda/n + \alpha\sqrt{\lambda}\sum_{\bar{z}^h} \mu(\bar{z}^h)(f(\bar{z}^h) - f'(\bar{z}^h))^2} \\
&\le \frac{1}{\alpha\rho_{\min}\sqrt{\lambda}} \max_{h \in [H]} \sum_{z^h} d_{\pi_*}^h(z^h) \sup_{f, f' \in \widehat{\mathcal{F}}^h} \frac{(f(z^h) - f'(z^h))^2/(\sigma_*^h(z^h))^2}{\lambda/n + \sum_{\bar{z}^h} \mu(\bar{z}^h)(f(\bar{z}^h) - f'(\bar{z}^h))^2},
\end{aligned} \tag{39}
$$

where the last inequality uses $\alpha\rho_{\min}\sqrt{\lambda} \le 1$. For any $z^h \sim d_{\pi_*}^h(z^h)$, take the $f_{z^h}, f'_{z^h} \in \widehat{\mathcal{F}}^h$ that maximize the term:

$$
\frac{(f_{z^h}(z^h) - f'_{z^h}(z^h))^2/(\sigma_*^h(z^h))^2}{\lambda/n + \sum_{\bar{z}^h} \mu(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2}. \tag{40}
$$

Then, (39) is written as

$$\begin{aligned}
\mathrm{CC}^{\sigma}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) &\leq \frac{1}{\alpha\rho_{\min}\sqrt{\lambda}} \max_{h\in[H]} \sum_{z^h} d^h_{\pi_*}(z^h) \frac{(f_{z^h}(z^h) - f'_{z^h}(z^h))^2/(\sigma_*^h(z^h))^2}{\lambda/n + \sum_{\bar{z}^h} \mu(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2} \\
&= \frac{1}{\alpha\rho_{\min}\sqrt{\lambda}} \max_{h\in[H]} \sum_{z^h} d^h_{\pi_*}(z^h) \frac{(f_{z^h}(z^h) - f'_{z^h}(z^h))^2/(\sigma_*^h(z^h))^2}{\lambda/n + \sum_{\bar{z}^h} \mu(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2}. \qquad (41)
\end{aligned}$$

To bound the above term, we need to lower bound $(\sigma_*^h(z^h))^2$ in (38) by

$$\begin{aligned}
\sup_{f,f'\in\widehat{\mathcal{F}}^h} \frac{|f(z^h) - f'(z^h)|/\alpha}{\sqrt{\lambda + \sum_{i=1}^n (f(z_i^h) - f'(z_i^h))^2/(\sigma_i^h)^2}} &\geq \frac{|f_{z^h}(z^h) - f'_{z^h}(z^h)|/\alpha}{\sqrt{\lambda + \sum_{i=1}^n (f_{z^h}(z_i^h) - f'_{z^h}(z_i^h))^2/(\sigma_i^h)^2}} \\
&\geq \frac{|f_{z^h}(z^h) - f'_{z^h}(z^h)|}{2\alpha\beta^h},
\end{aligned}$$

where the last inequality uses $f_{z^h}, f_{z^h} \in \widehat{\mathcal{F}}^h$:

$$\begin{aligned}
\lambda + \sum_{i=1}^n \frac{(f_{z^h}(z_i^h) - f'_{z^h}(z_i^h))^2}{(\sigma_i^h)^2} &= \lambda + \sum_{i=1}^n \frac{(f_{z^h}(z_i^h) - \hat{f}(z_i^h) + \hat{f}(z_i^h) - f'_{z^h}(z_i^h))^2}{(\sigma_i^h)^2} \\
&\leq 2\Big[\lambda + \sum_{i=1}^n \frac{(f_{z^h}(z_i^h) - \hat{f}(z_i^h))^2}{(\sigma_i^h)^2} + \lambda + \sum_{i=1}^n \frac{\hat{f}(z_i^h) - f'_{z^h}(z_i^h))^2}{(\sigma_i^h)^2}\Big] \leq 4(\beta^h)^2.
\end{aligned}$$

Thus, substituting this lower bound into (41) and taking $\beta^h = C_\beta\sqrt{\ln N} = C_\beta\sqrt{\lambda}$, we get

$$\begin{aligned}
\mathrm{CC}^{\sigma}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) &\leq \frac{2\alpha\beta^h}{\alpha\rho_{\min}\sqrt{\lambda}} \max_{h\in[H]} \sum_{z^h} d^h_{\pi_*}(z^h) \frac{|f_{z^h}(z^h) - f'_{z^h}(z^h)|}{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2} \\
&= \frac{2C_\beta}{\rho_{\min}} \max_{h\in[H]} \sum_{z^h} d^h_{\pi_*}(z^h) \frac{|f_{z^h}(z^h) - f'_{z^h}(z^h)| \cdot \mathbb{1}\big(\|f_{z^h} - f'_{z^h}\|_\infty \leq n^{-1}\big)}{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2} \\
&\quad + \frac{2C_\beta}{\rho_{\min}} \max_{h\in[H]} \sum_{z^h} d^h_{\pi_*}(z^h) \frac{|f_{z^h}(z^h) - f'_{z^h}(z^h)| \cdot \mathbb{1}\big(\|f_{z^h} - f'_{z^h}\|_\infty > n^{-1}\big)}{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2} \\
&\leq \frac{2C_\beta}{\lambda\rho_{\min}} + \frac{2C_\beta}{\rho_{\min}} \max_{h\in[H]} \sum_{z^h} d^h_{\pi_*}(z^h) \frac{|f_{z^h}(z^h) - f'_{z^h}(z^h)| \cdot \mathbb{1}\big(\|f_{z^h} - f'_{z^h}\|_\infty > n^{-1}\big)}{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2}.
\end{aligned}$$

Then, we can invoke the Cauchy-Schwarz inequality to split the above term into two parts:

$$\begin{aligned}
\mathrm{CC}^{\sigma}(\lambda, \widehat{\mathcal{F}}, \mathcal{Z}_n^H) &\leq \frac{2C_\beta}{\lambda\rho_{\min}} + \frac{2C_\beta}{\rho_{\min}} \max_{h\in[H]} \sum_{z^h} \frac{\sqrt{d^h_{\pi_*}(z^h)}}{\sqrt{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2}} \\
&\qquad \cdot \frac{\sqrt{d^h_{\pi_*}(z^h)}|f_{z^h}(z^h) - f'_{z^h}(z^h)|}{\sqrt{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2}} \cdot \mathbb{1}\big(\|f_{z^h} - f'_{z^h}\|_\infty > n^{-1}\big) \\
&\leq \frac{2C_\beta}{\lambda\rho_{\min}} + \frac{2C_\beta}{\rho_{\min}} \max_{h\in[H]} \underbrace{\Big(\sum_{z^h} \frac{d^h_{\pi_*}(z^h) \cdot \mathbb{1}\big(\|f_{z^h} - f'_{z^h}\|_\infty > n^{-1}\big)}{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2}\Big)}_{I_1}{}^{1/2} \\
&\qquad \cdot \underbrace{\Big(\max_{h\in[H]} \sum_{z^h} d^h_{\pi_*}(z^h) \frac{(f_{z^h}(z^h) - f'_{z^h}(z^h))^2}{\lambda/n + \sum_{\bar{z}^h} \mu^h(\bar{z}^h)(f_{z^h}(\bar{z}^h) - f'_{z^h}(\bar{z}^h))^2}\Big)}_{I_2}{}^{1/2}, \qquad (42)
\end{aligned}$$

Then, the terms I_1 and I_2 can be handled in the same way as the proof of Lemma 4.1. For the term I_1, we get

$$I_1 \leq \sum_{z^h} \frac{d^h_{\pi_*}(z^h)}{C(\widehat{\mathcal{F}}, \mu)/2} = \frac{2}{C(\widehat{\mathcal{F}}, \mu)}.$$

For the term I_2, we have

$$I_2 = \max_{h\in[H]} \sum_{z^h} d^h_{\pi_*}(z^h) \sup_{f,f'\in\widehat{\mathcal{F}}^h} \frac{n(f(z^h)-f'(z^h))^2}{\lambda+\sum_{i=1}^n (f(z^h_i)-f'(z^h_i))^2} = \mathrm{CC}(\lambda,\widehat{\mathcal{F}},\mathcal{Z}^H_n).$$

Therefore, by taking the bound of the terms I_1, I_2 into (42), we have

$$\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}^H_n) = \frac{2C_\beta}{\lambda\rho_{\min}} + \frac{2C_\beta}{\rho_{\min}}\sqrt{\frac{2\mathrm{CC}(\lambda,\widehat{\mathcal{F}},\mathcal{Z}^H_n)}{C(\widehat{\mathcal{F}},\mu)}} = \tilde{\mathcal{O}}\Big(\sqrt{\frac{\mathrm{CC}(\lambda,\widehat{\mathcal{F}},\mathcal{Z}^H_n)}{C(\widehat{\mathcal{F}},\mu)}}\Big),$$

which concludes the proof. □

C.1.3 The suboptimality bound.

Having the guarantee for the confidence radius and the connection between weighted and unweighted coverage coefficient, we can follow the three steps in the proof of Theorem 1 to bound the suboptimality.

Proof of Theorem 3. Since when $\sigma^h_*(x^h,a^h) > 1$, we have

$$\sigma^h_*(x^h,a^h) = \frac{1}{\sigma^h_*(x^h,a^h)} \cdot \sup_{f,f'\in\widehat{\mathcal{F}}^h} \frac{|f(x^h,a^h)-f'(x^h,a^h)|/\alpha}{\sqrt{\lambda+\sum_{i=1}^n (f(x^h_i,a^h_i)-f'(x^h_i,a^h_i))^2/(\sigma^h_i)^2}} = \frac{b^h(x^h,a^h)}{\alpha\sigma^h_*(x^h,a^h)},$$

from which it follows that

$$\begin{aligned}
&\sum_{h=1}^H \beta^h \mathbb{E}_{\tilde{\pi}_*}\big[b^h(x^h,a^h)\,\big|\,x^1=x\big] \\
&\quad\le \sum_{h=1}^H \beta^h \mathbb{E}_{\tilde{\pi}_*}\left[\frac{b^h(x^h,a^h)}{\sigma^h_*(x^h,a^h)} + \Big(\frac{b^h(x^h,a^h)}{\sigma^h_*(x^h,a^h)}\Big)^2\cdot\frac{1}{\alpha}\,\middle|\,x^1=x\right] \\
&\quad\le \sum_{h=1}^H \left[\beta^h\cdot\sqrt{\frac{\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}^H_n)}{n}} + \frac{\beta^h}{\alpha}\cdot\frac{\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}^H_n)}{n}\right],
\end{aligned}$$

where the last inequality uses $\mathbb{E}X \le \sqrt{\mathbb{E}X^2}$ and

$$\mathbb{E}_{\pi_*}\Big[\Big(\frac{b^h(x^h,a^h)}{\sigma^h_*(x^h,a^h)}\Big)^2\,\Big|\,x^1=x\Big] = \frac{\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}^H_n)}{n}.$$

We further get with probability at least $1-\delta$,

$$\begin{aligned}
&\sqrt{\frac{\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}^H_n)}{n}}\sum_{h=1}^H \beta^h + \frac{\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}^H_n)}{n}\cdot\frac{\sum_{h=1}^H\beta^h}{\alpha} \\
&\quad\le \sqrt{\frac{\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}^H_n)}{n}}\cdot c_\beta\big(\alpha\zeta + H\sqrt{\ln N}\big) + \frac{\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}^H_n)}{n}\cdot c_\beta\Big(\zeta + \frac{H\sqrt{\ln N}}{\alpha}\Big) \\
&\quad= \tilde{\mathcal{O}}\Big(\frac{3H\sqrt{\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}^H_n)\cdot\ln N}}{\sqrt{n}} + \frac{3\zeta\cdot\mathrm{CC}^\sigma(\lambda,\widehat{\mathcal{F}},\mathcal{Z}^H_n)}{n}\Big)
\end{aligned}$$

where we choose $\alpha = H\sqrt{\ln N}/\zeta$. By invoking Lemma C.2, we obtain that with probability at least $1-2\delta$,

$$\mathrm{SubOpt}(\hat{\pi},x) = \tilde{\mathcal{O}}\Big(\frac{H(\mathrm{CC}(\lambda,\widehat{\mathcal{F}},\mathcal{Z}^H_n))^{1/4}\cdot(\ln N_n(\gamma))^{1/2}}{n^{1/2}(C(\widehat{\mathcal{F}},\mu))^{1/4}} + \frac{\zeta(\mathrm{CC}(\lambda,\widehat{\mathcal{F}},\mathcal{Z}^H_n))^{1/2}}{n(C(\widehat{\mathcal{F}},\mu))^{1/2}}\Big).$$

Ultimately, we conclude the proof. □

D Implementation Details

Implementation of UWMSG. Following prior uncertainty-based offline RL algorithm Model Standard-deviation Gradients (MSG) [16], we learn a group of Q networks $Q_{w_i}, i = 1, \dots, K$ with independent targets and optimize a policy π_θ with a lower-confidence bound (LCB) objective [16, 2]. Specifically, Q_{w_i} is learned to minimize the following weighted regression objective with samples from the offline dataset $\mathcal{D}$:

$$\min_{w_i} \mathbb{E}_{(x,a,r,x')\sim\mathcal{D}} \Big[\frac{\big(\widehat{\mathcal{T}} Q_{w_i}(x,a) - Q_{w_i}(x,a)\big)^2}{\sigma(x,a)^2} \Big]. \tag{43}$$

The weight function σ is estimated via bootstrapped uncertainty: $\sigma(x,a) = \text{clip}(\mu \times \sqrt{\mathbb{V}_{i=1,\dots,K}\left[Q_{w_i}(x,a)\right]}, 1, M)$, where $\mathbb{V}_{i=1,\dots,K}\left[Q_{w_i}\right]$ refers to the variance between the group of Q functions, and M is used to control the maximum value of the weighting function. Note that $\sigma(x,a)$ is detached from the gradients, and the update is exclusively on w_i. We introduce the uncertainty ratio μ for $\sigma(x,a)$ to tune the weight function. The independent target $\widehat{\mathcal{T}} Q_{w_i}$ for Q_{w_i} is defined as follows:

$$\widehat{\mathcal{T}} Q_{w_i}(x,a) := r(x,a) + \gamma \mathbb{E}_{a'\sim\pi_\theta(\cdot|x')} \left[Q_{w_i'}(x',a')\right], \tag{44}$$

where $Q_{w_i'}$ is the target network for Q_{w_i}. In empirical offline RL, it is a common practice [1, 2, 16] to utilize the discounted form of the Q function rather than the episodic version. The policy π_θ optimizes the same pessimistic objective as MSG:

$$\min_\theta \mathbb{E}_{x\sim\mathcal{D}, a\sim\pi_\theta(\cdot|x)} \Big[\mathbb{E}_{i=1,\dots,K}\left[Q_{w_i}(x,a)\right] - \beta \cdot \sqrt{\mathbb{V}_{i=1,\dots,K}\left[Q_{w_i}(x,a)\right]}\Big], \tag{45}$$

Although the weighting function $\sigma(x,a)$ shares some similarities with the pessimistic bonus, they differ in the following aspects: (1) $\sigma(x,a)$ **measures the intrinsic variance of the corrupted data, while the pessimistic bonus penalizes out-of-distribution (OOD) actions produced by** π_θ**; (2)** $\sigma(x,a)$ **weights the** Q **learning objective and is detached from gradients, whereas the pessimistic bonus requires gradients for** π_θ.

Training and Evaluation Details. We use 3-layer MLPs with 256 neurons in each layer for both Q and policy networks. The ensemble size is set to $K = 10$ for all the experiments. The hyperparameters, such as learning rate and optimizer, are listed in Table 2. We train each algorithm for 3000 epochs, where one epoch contains 1000 updates. Regarding the offline datasets, we use 'halfcheetah-medium-v2', 'walker2d-medium-replay-v2', and 'hopper-medium-replay-v2' datasets and refer to them as 'halfcheetah', 'walker2d', and 'hopper' in our paper. Our implementation is based on SAC-N [1]. Therefore, there are additional entropy regularization terms for both $\widehat{\mathcal{T}} Q_{w_i}(s,a)$ in Eq (44) and the policy objective in Eq (45). For hyperparameters related to corruption and uncertainty weighting, we list them in Table 3. Since tasks vary in their ability to resist corruption, their hyperparameters are tuned separately. We use the same LCB ratio β for MSG and UWMSG, which is searched within $\{4.0, 6.0\}$. The uncertainty ratio μ for UWMSG is search from $\{0.2, 0.3, 0.5, 0.7, 1.0\}$.

To evaluate algorithms, we run the deterministic policy of each agent for 1000 steps and report their average cumulative returns with standard deviations over 10 random seeds. Our code is based on [42] and is available at https://github.com/YangRui2015/UWMSG.

Data Corruption Details. We implement both random and adversarial corruption on either rewards or dynamics. The four types of data corruption are listed below:

- Random reward attack: randomly sample $c\%$ transitions (x,a,r,x') from D, and modify the reward $\hat{r} \sim \text{Uniform}[-\epsilon, \epsilon]$, where c is the corruption rate and ϵ is the corruption scale.
- Random dynamics attack: randomly sample $c\%$ transitions (x,a,r,x'), and modify the next-step state $\hat{x}' = x' + \delta \cdot \text{std}, \delta \sim \text{Uniform}[-\epsilon, \epsilon]^d$, where d is dimension of states and std is the d-dimensional standard deviation of all states in the offline dataset.
- Adversarial reward attack: randomly sample $c\%$ transitions (x,a,r,x'), and modify the reward as: $\hat{r} = -\epsilon \times r$.

- Adversarial dynamics attack: pretrain a group of Q_p functions and a policy function π_p, then randomly sample $c\%$ transitions (x, a, r, x'), and modify the next-step states $\hat{x}' = \min_{\hat{x}' \in \mathbb{B}_d(x', \epsilon)} Q_p(x', \pi_p(\hat{x}'))$, where $\mathbb{B}_d(x', \epsilon) = \{|\hat{x}' - x'| \leq \epsilon \cdot \text{std}\}$ regularizes the maximum difference for each state dimension. The optimization is implemented through gradient descent similar to prior works [58, 53].

For the implementation of an adversarial dynamics attack, the optimization is performed through 10-step gradient descent with learning rate $\frac{\epsilon}{10}$. After each gradient descent step, the states are clipped within $\mathbb{B}_d(x', \epsilon)$. The pretraining algorithm used is MSG for the halfcheetah and walker2d tasks, while EDAC is employed for the hopper task due to its significantly better performance on this task compared to MSG in the absence of corruption. Finally, the corrupted data is saved and will be loaded for future training. To control the cumulative corruption ζ under continuous state-action spaces, we incorporate random or adversarial noise with predefined corruption ranges and corruption scales into the rewards and next-step states. This is because the fact that $\zeta_i \leq |r - r_{\mathcal{D}}| + D_{TV}(P \| P_{\mathcal{D}}) \leq |r - r_{\mathcal{D}}| + \sqrt{\frac{1}{2} D_{KL}(P \| P_{\mathcal{D}})}$. When we consider P and $P_{\mathcal{D}}$ are both Diagonal Gaussian distributions with the same constant variance, $\zeta_i \leq |r - r_{\mathcal{D}}| + \sqrt{\frac{1}{2}} \|\mu - \mu_{\mathcal{D}}\| + \text{const}$. When we corrupt only one element in rewards and dynamics, the empirical cumulative corruption can be approximated as the multiplication of the number of corrupted samples and the corruption scale: $\zeta = |D| \times c\% \times \epsilon$, where $|D|$ represents the size of the dataset, and ϵ represents the corruption scale. Note that this approximation may not hold for our reward corruption, but we use the same calculation for simplicity.

Table 2: Hyper-parameters for UWMSG and MSG.

Hyper-parameters	Value
Ensemble size K	10
Policy network	FC(256,256,256) with ReLU
Q-network	FC(256,256,256) with ReLU
LCB ratio β	$\{4.0, 6.0\}$
Uncertainty ratio μ	$\{0.2, 0.3, 0.5, 0.7, 1.0\}$
Maximum value of uncertainty weight M	10
Target network smoothing coefficient τ	5e-3
Discount factor γ	0.99
Policy learning rate	3e-4
Q network learning rate	3e-4
Optimizer	Adam
Automatic Entropy Tuning	True
batch size	256

Table 3: Data corruption settings and the hyperparameters used for uncertainty-weighting in UWMSG.

Attack type	Attack object	Environment	Corruption rate $c\%$	Corruption scale ϵ	Cumulative corruption ζ	LCB ratio β	Uncertainty ratio μ
Random	Reward	halfcheetah	20%	30.0	5.99×10^6	4.0	0.7
		walker2d	30%	30.0	2.72×10^6	4.0	0.3
		hopper	20%	30.0	2.41×10^6	6.0	0.7
	Dynamics	halfcheetah	20%	2.0	4.00×10^5	4.0	0.5
		walker2d	10%	0.5	1.51×10^4	6.0	0.5
		hopper	10%	0.5	2.01×10^4	6.0	0.7
Adversarial	Reward	halfcheetah	20%	3.0	5.99×10^5	4.0	0.7
		walker2d	20%	3.0	1.81×10^5	4.0	0.5
		hopper	10%	5.0	2.01×10^5	6.0	0.7
	Dynamics	halfcheetah	30%	1.2	3.60×10^5	4.0	0.2
		walker2d	10%	0.3	9.05×10^3	4.0	0.5
		hopper	10%	0.5	2.01×10^4	6.0	1.0

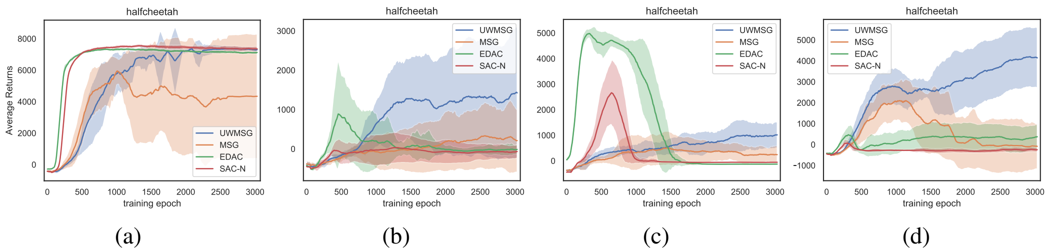
E Comparison with Uncertainty-weighted Actor Critic (UWAC)

Our practical implementation algorithm, UWMSG, shares some similarities with UWAC in terms of utilizing uncertainty weighting technique for offline RL. However, there are three key differences between our approaches:

1. We focus on offline RL with data corruption, rather than the general offline RL setting explored by UWAC. Therefore, in our setting, the uncertainty arises from both corrupted datasets and OOD actions.
2. While UWAC penalizes OOD actions in the Q objective through minimizing $\mathbb{E}_{(x,a,r,x')\sim\mathcal{D},a'\sim\pi_\theta(\cdot|x')}\left[\frac{\left(\widehat{\mathcal{T}}Q(x,a)-Q(x,a)\right)^2}{Var[Q(x',a')]}\right]$, with the aim of reducing the importance of OOD actions, our uncertainty weighting focuses on penalizing in-dataset (x, a) pairs.
3. Another distinction lies in the uncertainty estimation methods employed. UWAC uses dropout uncertainty, while we utilize bootstrapped uncertainty in our work. However, it is worth noting that our approach is not limited to a specific type of uncertainty estimation. In the future, more advanced uncertainty estimation methods (e.g., [41, 7]) can be applied to potentially enhance the performance of UWMSG.

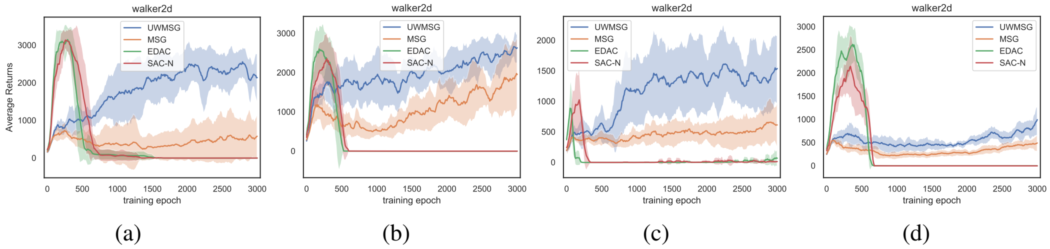
F Additional Results

Learning Curves All learning curves are shown in Figure 2, Figure 3, and Figure 4. We can find that (1) current offline RL methods are susceptible to data corruption, e.g., MSG, EDAC, SAC-N achieve poor and unstable performance under adversarial attacks, and (2) our proposed UWMSG method significantly improves performance under different data corruption scenarios. Moreover, we posit that the reason for the observed initial increase and subsequent significant decrease of EDAC and SAC-N performance in some cases may be attributed to the characteristics of Temporal Difference (TD) learning. Specifically, the effect of corruption needs to accumulate over time, which may necessitate an extended training period to destroy the performance. In contrast, our algorithm UWMSG does not suffer from this problem and exhibits stable performance.



(a) (b) (c) (d)

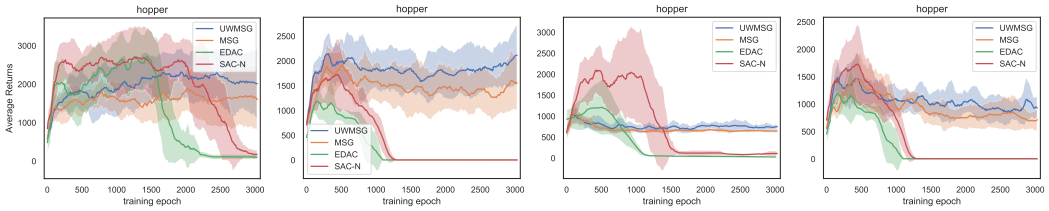
Figure 2: Comparison on the halfcheetah task under (a) random reward, (b) random dynamics, (c) adversarial reward, and (d) adversarial dynamics attacks.



(a) (b) (c) (d)

Figure 3: Comparison on the walker2d task under (a) random reward, (b) random dynamics, (c) adversarial reward, and (d) adversarial dynamics attacks.

Varying Corruption Level We evaluate the performance of UWMSG under varying levels of corruption in Figure 5. This is achieved by maintaining a consistent corruption scale in Table 3 while adjusting the corruption rate. As depicted in the figure, as the cumulative corruption level rises, the overall performance of UWMSG progressively declines. These findings align with our theoretical analysis. Besides, the results indicate that dynamics corruption poses a greater challenge compared to reward corruption, leading to a larger drop in performance with a smaller corruption level.



(a) (b) (c) (d)

Figure 4: Comparison on the hopper task under (a) random reward, (b) random dynamics, (c) adversarial reward, and (d) adversarial dynamics attacks.

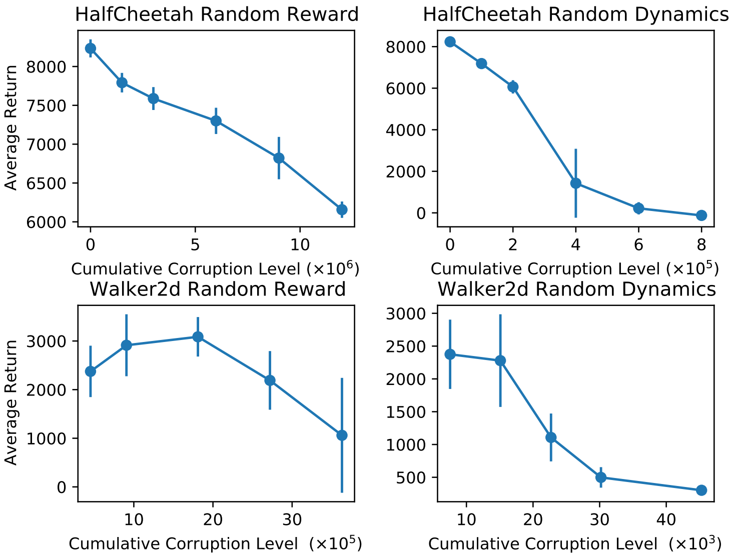


Figure 5: Performance of UWMSG under varying levels of corruption. Results are averaged over 5 random seeds.

G Technical Lemmas

Lemma G.1 (Lemma 3.1 of Jin et al. [24]). *Let $\hat{\pi} = \{\hat{\pi}^h\}_{h=1}^H$ be the greedy policy such that for any x, $\hat{\pi}^h(x) = \operatorname{argmax}_{a\in\mathcal{A}} f_n^h(x,a)$. For any initial state $x \in \mathcal{X}$,*

$$\begin{aligned}\mathrm{SubOpt}(\hat{\pi}, x) = &\sum_{h=1}^{H} \mathbb{E}_{\pi_*}\left[f_n^h(x^h, \pi_*(x^h)) - f_n^h(x^h, \hat{\pi}(x^h)) \,\middle|\, x^1 = x\right] \\ &- \sum_{h=1}^{H} \mathbb{E}_{\pi_*}\left[\mathcal{E}^h(f_n, x^h, a^h) \,\middle|\, x^1 = x\right] + \sum_{h=1}^{H} \mathbb{E}_{\hat{\pi}}\left[\mathcal{E}^h(f_n, x^h, a^h) \,\middle|\, x^1 = x\right],\end{aligned}$$

where $\mathcal{E}^h(f, x^h, a^h) = f^h(x^h, a^h) - (\mathcal{T}^h f^{h+1})(x^h, a^h)$ is the Bellman residual.

Lemma G.2. *Let $\{\epsilon_s\}$ be a sequence of zero-mean conditional η-sub-Gaussian random variables: $\ln \mathbb{E}[e^{\lambda\epsilon_s}|\mathcal{S}_{s-1}] \le \lambda^2\eta^2/2$, where $\mathcal{S}_{s-1}$ 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. The proof can is presented in Lemma G.2 of Ye et al. [55]. □

Lemma G.3 (Lemma G.4 of Ye et al. [55])**.** *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 $t \in [T]$:

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

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

Proof. The proof can be seen in Lemma G.4 of Ye et al. [55]. □

Lemma G.4 (Theorem 1.3 of Tropp [43])**.** *For a finite sequence $\{X_i\}_{i \in [n]}$ of independent, random and self-adjoint matrices with dimension d, let $\{A_i\}_{i \in [n]}$ be a sequence of fixed self-adjoint matrices. If each random matrix satisfies*

$$\mathbb{E}X_i = 0 \quad X_i^2 \preceq A_i^2 \quad \textit{almost surely},$$

then, for all $t \ge 0$,

$$\mathbb{P}\Big(\lambda_{\max}\big(\sum_{i=1}^{n} X_i\big) \ge t\Big) \le d \cdot e^{-t^2/(8\sigma^2)},$$

where $\sigma^2 = \|\sum_{i=1}^n A_i^2\|_{\mathrm{op}}$.