Corruption-Robust Offline Two-Player Zero-Sum Markov Games

Andi Nika
MPI-SWS

Debmalya Mandal[†]
University of Warwick

Adish Singla
MPI-SWS

Goran Radanovic
MPI-SWS

Abstract

We study data corruption robustness in offline two-player zero-sum Markov games. Given a dataset of realized trajectories of two players, an adversary is allowed to modify an ϵ-fraction of it. The learner's goal is to identify an approximate Nash Equilibrium policy pair from the corrupted data. We consider this problem in linear Markov games under different degrees of data coverage and corruption. We start by providing an information-theoretic lower bound on the suboptimality gap of any learner. Next, we propose robust versions of the Pessimistic Minimax Value Iteration algorithm (Zhong et al., 2022), both under coverage on the corrupted data and under coverage only on the clean data, and show that they achieve (near)-optimal suboptimality gap bounds with respect to ϵ. We note that we are the first to provide such a characterization of the problem of learning approximate Nash Equilibrium policies in offline two-player zero-sum Markov games under data corruption.

1 INTRODUCTION

Some of the most successful applications of Multi-agent Reinforcement Learning (MARL) are in competitive game-playing (Silver et al., 2017; Berner et al., 2019), where we have a model of the environment that we can use for training purposes. Given that many real-world multi-agent applications, such as autonomous driving (Pan et al., 2017) or healthcare (Wang et al., 2018), do not have readily available simulators, there has recently been a growing interest in studying offline settings, where offline data is used to derive agents' policies. Since a dynamic exploration of the environment is impossible, state-of-the-art (SOTA) algorithms use the paradigm of *pessimism in the face of uncertainty* to derive these policies (Jin et al., 2021). Moreover, these works typically assume that the data is coming from a latent distribution with "nice" properties.

In practice, however, datasets may be subject to adversarial attacks that corrupt data points and can significantly impact the performance of the learning process. Such security threats have already been explored in single-agent RL, where prior work has proposed corruption robust algorithms (Zhang et al., 2022). However, these results do not directly translate to MARL due to the intricacies of multi-agent settings. For instance, the learning objective in these settings requires a more complex solution concept of learning a Nash Equilibrium (NE) policy pair for agents instead of simply learning a near-optimal policy for an agent. In this work, we initiate the study of corruption robust algorithms for learning equilibrium policies in offline MARL. More specifically, we focus on two-agent zero-sum Markov games and consider the following research question:

Can we design algorithms that approximately solve offline two-player zero-sum Markov games under data corruption?

To effectively answer this question, we need to account for another crucial factor in offline learning, that is, the quality of the collected data, which drastically affects the quality of the learned policy. It is thus common practice to assume that the collected data *covers* at least some trajectories of interest. It turns out that the necessary coverage assumptions for solving the offline single-player problem are not enough to solve the offline two-player problem (see Figure 1). Thus, stronger assumptions are required, i.e., the so-called *Low Relative Uncertainty* (LRU) assumption. This problem is exacerbated by the presence of corruption in our setting. If good coverage in the clean setting seems natural, supposing that the data has been collected by a *good enough* policy, such an assumption is no longer guaranteed when a potentially malicious adversary intervenes in the data.

Motivated by these observations, we study the problem

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. [†] Work done while the author was with MPI-SWS.

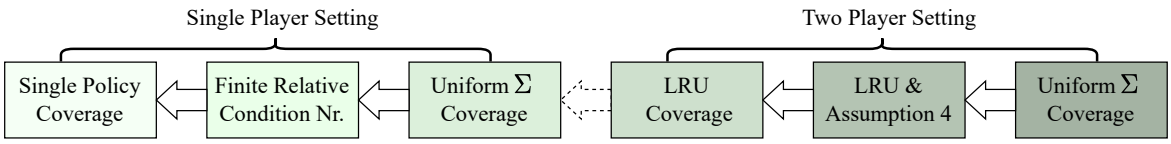

Figure 1: Relationship between coverage assumptions. The minimal coverage requirements are single policy coverage and LRU coverage for the single-player and two-player settings, respectively. Arrows stand for implications. The middle (dashed) arrow denotes the restriction from the two-player to the single-player setting: when fixing the second player's policy, the LRU coverage assumption reduces to the uniform coverage assumption.

of corruption in the offline two-player setting under various assumptions. First, we tackle corruption under the minimal LRU coverage and the more relaxed uniform Σ-coverage (defined in Assumption 2) assumptions on the corrupted data. Furthermore, we also consider the more difficult setting where the minimal LRU coverage holds only on the clean dataset. We build upon recent techniques for the offline two-player zero-sum setting (Zhong et al., 2022) and propose robust versions of their method. We tackle the corruption problem by using two setting-specific robust estimators and carefully designing new bonus terms that capture the additional estimation errors coming from corruption. More concretely, our main results and contributions are summarized below (also, see Table 1):

I. **Lower bound.** First, we formulate the problem of data corruption in offline two-player zero-sum Markov games. We prove an information-theoretic lower bound of $\Omega(Hd\epsilon)$ on the suboptimality gap of any algorithm that uses a corrupted dataset where H is the episode length, d is the dimension, and ϵ is the corruption level.

II. **Uniform Σ and corrupted covariates.** Next, we consider the corruption problem under the Uniform Σ-coverage assumption on the corrupted data. We propose R-PMVI that uses a Robust Least Squares estimator (Zhang et al., 2022) and show that it incurs a near-optimal bound on the error coming from corruption with high probability, showing an improvement on the single-player bounds of Zhang et al. (2022) by a factor of H.

III. **LRU and clean covariates.** Furthermore, we consider the LRU coverage on the corrupted data. We additionally assume that corruption is done only on the reward and next state part of the data tuples, i.e., assuming clean covariates. In this setting, we propose S-PMVI that uses the Spectrally Regularized Alternating Minimization algorithm (Chen et al., 2022a) oracle as a robust estimator and provide a high probability near-optimal bound on its suboptimality gap. We note that, under no corruption, we recover the SOTA bounds of Zhong et al. (2022), while obtaining similar rates of ϵ as in the single agent SOTA (Zhang et al., 2022).

IV. **LRU on the clean data.** Finally, we consider the most restrictive setting where no guarantees of coverage on the corrupted data are given but only make the necessary LRU assumption on the clean data. In this setting, we propose a new bonus term for the S-PMVI algorithm, that takes into account the additional error terms. Our method yields $O(d^{3/2}\sqrt{\epsilon})$ bounds on the suboptimality gap, similar to the single-player setting. Moreover, under an additional mild assumption on the feature space, we are able to recover the optimal $O(\epsilon)$ rate, at the cost of an additional $O(d^{3/2})$ factor.

Coverage	Covariates	Algorithm	Suboptimality Gap	Result
Uniform Σ on corrupted data	Corrupted	R-PMVI	$\widetilde{O}\left(H^2 d\epsilon + H^{3/2} f(d) K^{-1/2}\right)$	[Theorem 2]
LRU on corrupted data	Clean	S-PMVI	$\widetilde{O}\left(H^2 d\epsilon + H^2 K^{-1/2} d^{3/2}\right)$	[Theorem 3]
LRU on clean data	Corrupted	S-PMVI	$\widetilde{O}\left(H^2 d^{3/2}\sqrt{\epsilon} + H^2 K^{-1/2} d^{3/2}\right)$	[Theorem 4]
LRU on clean data & A.4	Corrupted	S-PMVI	$\widetilde{O}\left(H^2 d^3 \epsilon + H^2 K^{-1/2} d^3\right)$	[Theorem 5]

Table 1: Summary of our results under Low Relative Uncertainty and uniform Σ-coverage assumptions (see Assumptions 2 and 3 for definitions) on clean or corrupted data, and different corruption levels of the feature covariance matrix. Here ϵ denotes the corruption level, K denotes the number of trajectories contained in the data, and $f(x)$ denotes a polynomial function of x. The Covariates column refers to whether the state-action part of the data tuple is corrupted or not. We have omitted linear dependence on noise variance γ^2 of the rewards for ease of presentation. We point the reader to the relevant results for a detailed description. We note that the universal lower bound for the offline two-player zero-sum MG setting is $\Omega(Hd\epsilon)$.

Furthermore, we observe that convergence to Nash equilibria, without knowledge of ϵ and uniform coverage, is impossible in the offline two-player setting, which is a direct implication of the single-player setting (Zhang et al., 2022). Finally, we provide a comprehensive discussion on the relationship between used coverage assumptions and other similar assumptions found in the offline MARL literature and position them relative to each other.

2 PRELIMINARIES

Notation. We begin this section by introducing the notation to be used throughout the paper. As usual, I denotes the identity matrix, $\|\cdot\|_2$ denotes the Euclidean norm, $\|\cdot\|_A$ denotes the Mahalanobis norm given square matrix A, and $\Delta(\mathcal{X})$ denotes the probability simplex on set $\mathcal{X}$. When $\widetilde{O}$ is used, any polylogarithmic terms are omitted. Furthermore, $[H]$ denotes the set of natural numbers up to and including H, $\langle\cdot,\cdot\rangle$ denotes the inner product, $\mathbf{1}\{\cdot\}$ denotes the indicator function, and $\succeq$ denotes the Loewner order, where $A \succeq B$ is equivalent to $A - B$ being positive semi-definite. Finally, we define $\Pi_h(x) = \min\{h, \max\{x, 0\}\}$.

2.1 Two-player Zero-sum Markov Games

Let $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, p, r, H)$ be a finite-horizon zero-sum Markov game between two players, one of which is trying to maximize the total reward and the other is trying to minimize it. Here $\mathcal{S}$ denotes the state space with S states[1], $\mathcal{A}$ is the action space of the first player with A actions, $\mathcal{B}$ is the action space of the second player with B actions, $p = (p_1, \dots, p_H)$ where $p_h \in \mathbb{R}^{SAB \times S}$ $\forall h \in [H]$ denote the transition kernels, $r = (r_1, \dots, r_H)$ where $r_h \in \Delta(\mathbb{R})^{SAB}$ $\forall h \in [H]$ are γ^2-subGaussian rewards, and H is the horizon length. At each time step h and state s_h the max player selects action $a_h \in \mathcal{A}$ and the min player selects action $b_h \in \mathcal{B}$ and both observe reward $r_h(s_h, a_h, b_h)$. Then, the next state s_{h+1} is sampled from $p_h(\cdot|s_h, a_h, b_h)$.

A strategy pair (π, ν) is comprised of the strategy of the first player $\pi = (\pi_1, \dots, \pi_H)$, $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ $\forall h \in [H]$ and that of the second player $\nu = (\nu_1, \dots, \nu_H)$, $\nu_h : \mathcal{S} \to \Delta(\mathcal{B})$ $\forall h \in [H]$. Given $h \in [H]$, we define the state-value function and state-action value function as

$$V_h^{\pi,\nu}(s_h) = \mathbb{E}\left[\sum_{t=h}^{H} r_h(s_t, a_t, b_t) | \pi, \nu, s_h\right] ,$$

$$Q_h^{\pi,\nu}(s_h, a_h, b_h) = \mathbb{E}\left[\sum_{t=h}^{H} r_h(s_t, a_t, b_t) | \pi, \nu, s_h, a_h, b_h\right] .$$

[1]We only introduce the state space cardinality notation for convenience. Note that in linear Markov games, S may be intractably large.

2.2 Nash Equilibria and Performance Metrics

Let ν be a fixed strategy of the second player. Then, an optimal policy with respect to the MDP induced by ν is called the best response of the first player and we denote it by $\text{br}(\nu)$. Similarly, for a fixed strategy π of the first player, the best response of the second player is denoted by $\text{br}(\pi)$. Further, for any π, ν and $h \le H$, we define

$$V_h^{\pi,*}(s_h) = V_h^{\pi,\text{br}(\pi)}(s_h) = \inf_{\nu} V_h^{\pi,\nu}(s_h) ,$$

$$V_h^{*,\nu}(s_h) = V_h^{\text{br}(\nu),\nu}(s_h) = \sup_{\mu} V_h^{\mu,\nu}(s_h) .$$

Definition 1. *A Nash Equilibrium (NE) is a strategy pair (π^*, ν^*) such that, for all $h \in [H]$ and $s \in \mathcal{S}$:*

$$\sup_{\pi}\inf_{\nu} V_h^{\pi,\nu}(s) = V_h^{\pi^*,\nu^*}(s) = \inf_{\nu}\sup_{\pi} V_h^{\pi,\nu}(s) .$$

It is well-known that the NE of a zero-sum Markov game with a unique value function exists (Shapley, 1953). We define $V_h^*(s_h) = V_h^{\pi^*,\nu^*}(s_h)$, for all $h \le H$. Then, for all strategy pairs (π, ν), the weak duality is written as $V_h^{\pi,*}(s_h) \le V_h^*(s_h) \le V_h^{*,\nu}(s_h), \forall h \le H$. Consequently, the suboptimality gap of (π, ν) is

$$\text{SubOpt}(\pi, \nu, s) = V_1^{*,\nu}(s) - V_1^{\pi,*}(s) .$$

Note that the duality is always non-negative, and it is zero only when (π, ν) is a NE strategy. We say that a strategy (π, ν) is a η-approximate NE if we have $\text{SubOpt}(\pi, \nu, s) \le \eta$ for all $s \in \mathcal{S}$.

2.3 Linear Markov Games

We consider linear two-player zero-sum Markov games $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, p, r, H)$. Here, we formally state the linearity assumption for Markov games, standard in the literature (Xie et al., 2020).

Definition 2 (Linear Markov games)**.** *For each $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ we have*

$$r_h(s, a, b) = \phi(s, a, b)^\top \theta_h + \zeta \quad \textit{and}$$

$$p_h(\cdot|s, a, b) = \phi(s, a, b)^\top \mu_h(\cdot) ,$$

where $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}^d$ is a known feature map, $\theta_h \in \mathbb{R}^d$ is an unknown vector, ζ is zero-mean γ^2-subGaussian noise, and $\mu_h = (\mu_h^{(i)})_{i \in [d]}$ is a vector of d unknown signed measures on $\mathcal{S}$. We assume that $\|\phi(\cdot,\cdot,\cdot)\|_2 \le 1$, $\|\theta_h\|_2 \le \sqrt{d}$, and $\|\mu_h(\mathcal{S})\|_2 \le \sqrt{d}$ for all $h \in [H]$.

As previously observed (Zhong et al., 2022), given a policy pair (π, ν) and time-step h, there exists a d-dimensional weight vector $\omega_h^{\pi,\nu}$, with $\|\omega_h^{\pi,\nu}\|_2 \le H\sqrt{d}$, such that $Q_h^{\pi,\nu}(s, a, b) = \phi(s, a, b)^\top \omega_h^{\pi,\nu}$ for any (s, a, b) tuple.

2.4 Offline Data Collection

In offline RL, the objective is to learn an optimal policy from data that has already been collected beforehand (Levine et al., 2020). Similarly, in offline Markov Games, the objective is to learn an approximate NE strategy based on a given dataset. Formally, we are given a dataset $D = \{(s_h^\tau, a_h^\tau, b_h^\tau, r_h^\tau, s_{h+1}^\tau)\}_{h\in[H]}^{\tau\in[K]}$ of K trajectories, gathered from a behavioral policy $\rho = (\rho_1, \ldots, \rho_H), \rho_h : \mathcal{S} \to \Delta(\mathcal{A} \times \mathcal{B}), \forall h \in [H]$.[2]

Given $h \in [H]$, a strategy pair (π, ν), and a tuple (s, a, b), we denote by $d_h^{\pi,\nu}(s, a, b)$ the probability that the state-action tuple (s, a, b) is traversed in time step h by (π, ν), i.e., $d_h^{\pi,\nu}(s, a, b) = \mathbb{P}(s_h = s, a_h = a, b_h = b|\pi, \nu)$. If $d_h^{\pi,\nu}(s, a, b) > 0$, for all $h \in [H]$, we say that state-action tuple (s, a, b) is covered by policy pair (π, ν).

Next, we formally define the compliance of a given offline dataset with an underlying Markov game, which basically implies that the clean collected data follows the same dynamics as the environment. We will assume later on that the clean dataset is in compliance with the underlying Markov game.

Definition 3 (Compliance of dataset)**.** *Given a Markov game $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, p, r, H)$ and a dataset $D = \{(s_h^\tau, a_h^\tau, b_h^\tau, r_h^\tau, s_{h+1}^\tau)\}_{h\in[H]}^{\tau\in[K]}$, we say the dataset D is compliant with $\mathcal{G}$ if, for all $h \in [H]$ and $s \in \mathcal{S}$,*

$$\mathbb{P}\left(r_h^\tau = r, s_{h+1}^\tau = s|\{(s_h^i, a_h^i, b_h^i)\}_{i=1}^\tau, \{(r_h^i, s_{h+1}^i)\}_{i=1}^{\tau-1}\right) = \mathbb{P}_h\left(r_h = r, s_{h+1} = s|s_h = s_h^\tau, a_h = a_h^\tau, b_h = b_h^\tau\right) ,$$

where $\mathbb{P}$ is with respect to D and $\mathbb{P}_h$ is the probability measure taken with respect to the underlying Markov game $\mathcal{G}$.

2.5 Corruption Robust Estimation

The standard assumption in statistical estimation is that the samples we are given come from a fixed distribution, allowing one to directly use probability laws to obtain unbiased estimates of interest. However, it is usually the case that outliers are present in the data that do not belong to the underlying distribution, or that an adversary can arbitrarily corrupt the data. Only one outlier is enough to arbitrarily shift the empirical mean of the data, and therefore acquiring robust estimators for the moments of the distribution is important. We will use the Huber contamination model, akin to the corruption model in single-player offline RL (Zhang et al., 2022).

Assumption 1 (ϵ-contamination in offline Markov games)**.** *Given $\epsilon \in [0, 1]$ and a set of clean tuples $\widetilde{D} = \{(\widetilde{s}_i, \widetilde{a}_i, \widetilde{b}_i, \widetilde{r}_i, \widetilde{s}_i')\}_{i=1}^N$, an adversary is allowed to inspect the tuples and replace any ϵN of them with arbitrary contaminated tuples $(s, a, b, r, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times \mathcal{R} \times \mathcal{S}$. The resulting set $D = \{(s_i, a_i, b_i, r_i, s_i')\}_{i=1}^N$ is then revealed to the learner.*

We say that a set of samples is ϵ-corrupted if it is generated by the above process. Given an ϵ-corrupted set of data points, the goal of robust statistics is to compute accurate estimates of the first and second moments. In (Diakonikolas et al., 2017), the authors provide efficient and nearly sample-optimal filtering algorithms for mean and covariance estimation.

For the linear regression problem, different robust estimators require different coverage assumptions. For instance, SCRAM (Chen et al., 2022a) does not make strong coverage assumptions but assumes clean covariates, while Robust Least Squares (RLS) (Bakshi and Prasad, 2021; Pensia et al., 2020; Zhang et al., 2022) needs stronger coverage while allowing for corrupted covariates. We will use both aforementioned methods as robust estimator oracles under different coverage assumptions and corruption models.

3 PROBLEM FORMULATION

We assume that an unknown experimenter collects a dataset $\widetilde{D} = \{(\tilde{s}_h^\tau, \tilde{a}_h^\tau, \tilde{b}_h^\tau, \tilde{r}_h^\tau, \tilde{s}_{h+1}^\tau)\}_{\tau=1,h=1}^{K,H}$ of K trajectories, in compliance with $\mathcal{G}$. We assume that $\{(\tilde{s}_h^\tau, \tilde{a}_h^\tau, \tilde{b}_h^\tau, \tilde{r}_h^\tau, \tilde{s}_{h+1}^\tau)\}_{\tau=1,h=1}^{K,H}$ is collected from behavioral policy $\rho = (\rho^1, \rho^2)$. Formally, for any $h \in [H]$, we have $(\widetilde{s}_h, \widetilde{a}_h, \widetilde{b}_h) \sim d_h^\rho$ and $\widetilde{s}_{h+1} \sim p_h(\cdot|\widetilde{s}_h, \widetilde{a}_h, \widetilde{b}_h)$ for any $(\widetilde{s}_h, \widetilde{a}_h, \widetilde{b}_h, \widetilde{r}_h, \widetilde{s}_{h+1}) \in \widetilde{D}$. Subsequently, an adversary contaminates an ϵ-fraction of all tuples of $\widetilde{D}$ and provides us with the corrupted dataset $D = \{(s_h^\tau, a_h^\tau, b_h^\tau, r_h^\tau, s_{h+1}^\tau)\}_{\tau=1,h=1}^{K,H}$.

The learner's objective is, given access to an ϵ-corrupted dataset D, to be able to compute an approximate Nash equilibrium policy pair. In the clean offline setting, this is usually done by leveraging pessimism in order to penalize samples that were not sufficiently covered. However, if the dataset does not cover state actions of interest, i.e., those traversed by equilibrium policies, then learning becomes impossible, even in the clean setting. First, we formally define what we mean by coverage.

Definition 4. *A strategy pair (π, ν) is covered by the dataset generated by behavioral policy ρ if and only if every tuple (s, a, b) that is covered by (π, ν) is also covered by ρ. In other words, we have*

$$\frac{d_h^{\pi,\nu}(s, a, b)}{d_h^\rho(s, a, b)} < \infty, \forall (s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}, h \in [H] .$$

[2] The behavioral policy can be thought of as a product of a min and max policy.

For the offline single-player setting, it has been established that coverage of the optimal (or any target policy used as reference) policy is necessary to achieve convergence. For the two-player zero-sum Markov game setting, Zhong et al. (2022) show that coverage of a Nash equilibrium policy pair and its neighbors across each player is necessary for learning, as we will see in the next section.[3] We will consider the corruption problem under different coverage assumptions, starting from strong assumptions on the corrupted data, and ending with the setting where minimal coverage assumptions hold only on the clean dataset.

4 RESULTS UNDER COVERAGE ON CORRUPTED DATA

Recall from Section 2.3 that, given (π, ν), $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, and $h \in [H]$, Definition 2 implies that $Q_h^{\pi,\nu}(s, a, b) = \phi(s, a, b)^\top \omega_h^{\pi,\nu}$, for some $\omega_h^{\pi,\nu} \in \mathbb{R}^d$. Thus, learning the action value function Q_h^* that corresponds to a Nash equilibrium (NE) pair reduces to learning the optimal weights, which we denote by ω_h^*. For that, we will rely on *Pessimistic Minimax Value Iteration* (PMVI) (Zhong et al., 2022), which computes an approximate NE pair using offline data. However, PMVI computes estimates of ω_h^* by solving regularized least-squares on the Bellman operator. We cannot directly use those estimates since our data contains corrupted samples. Instead, we use a robust estimator to tackle the corruption problem. Let R-EST denote a generic robust estimator.

First, we randomly split the data into H batches D_h and set $\underline{V}_{H+1}(\cdot) = \overline{V}_{H+1}(\cdot) = 0$. Then, for each time step $h = H, H-1, \ldots, 1$, we obtain

$$\underline{\omega}_h \leftarrow \text{R-EST}\Big(\{\phi(s_h^\tau, a_h^\tau, b_h^\tau), r_h^\tau + \underline{V}_{h+1}(s_{h+1}^\tau)\}_{\tau=1}^K\Big) ,$$
$$\overline{\omega}_h \leftarrow \text{R-EST}\Big(\{\phi(s_h^\tau, a_h^\tau, b_h^\tau), r_h^\tau + \overline{V}_{h+1}(s_{h+1}^\tau)\}_{\tau=1}^K\Big) .$$

Once we have estimates of the optimal weights, the algorithm proceeds to construct estimates of the Q-values for both players, which depend on carefully constructed bonus terms Γ_h – we provide different bonus terms depending on the robust estimators we use and other characteristics of corruption.

Next, we compute a Nash equilibrium corresponding to payoffs $\underline{Q}(\cdot, \cdot, \cdot)$ and $\overline{Q}(\cdot, \cdot, \cdot)$, and obtain $(\widehat{\pi}_h, \nu'_h)$ and $(\pi', \widehat{\nu}_h)$, as solutions corresponding to $\underline{Q}(\cdot, \cdot, \cdot)$ and $\overline{Q}(\cdot, \cdot, \cdot)$, respectively.

Finally, the algorithm estimates the value functions for both players based on the computed policy pairs. After H steps, the algorithm terminates and outputs the estimated pairs of strategies $(\widehat{\pi}, \widehat{\nu})$. The pseudocode of the described method is given in Algorithm 1.

[3] Formal definitions are given in the next section.

Algorithm 1 Robust PMVI

Input: Dataset D, failure probability δ, robust estimator R-EST, bonus functions $\Gamma_h(\cdot, \cdot, \cdot)$.
Initialize: Randomly split dataset D into H subsets D_h of cardinality K; set $\underline{V}_{H+1}(s) = \overline{V}_{H+1}(s) = 0$, for all $s \in \mathcal{S}$.
for $h = H, H-1, \ldots, 1$ **do**:
  Compute estimates $\underline{\omega}_h$ and $\overline{\omega}_h$ via R-EST.
  $\underline{Q}_h(\cdot, \cdot, \cdot) \leftarrow \Pi_{H-h+1}\left(\phi(\cdot, \cdot, \cdot)^\top \underline{\omega}_h - \Gamma_h(\cdot, \cdot, \cdot)\right)$.
  $\overline{Q}_h(\cdot, \cdot, \cdot) \leftarrow \Pi_{H-h+1}\left(\phi(\cdot, \cdot, \cdot)^\top \underline{\omega}_h + \Gamma_h(\cdot, \cdot, \cdot)\right)$.
  Compute $(\widehat{\pi}_h, \nu'_h)$ and $(\pi'_h, \widehat{\nu}_h)$ as NE solutions to $\underline{Q}_h(\cdot, \cdot, \cdot)$ and $\overline{Q}_h(\cdot, \cdot, \cdot)$, respectively.
  $\underline{V}_h(\cdot) \leftarrow \mathbb{E}_{a\sim\widehat{\pi}_h, b\sim\nu'_h}[\underline{Q}_h(\cdot, a, b)]$.
  $\overline{V}_h(\cdot) \leftarrow \mathbb{E}_{a\sim\pi'_h, b\sim\widehat{\nu}_h}[\overline{Q}_h(\cdot, a, b)]$.
end for
Output: $\widehat{\pi} = (\widehat{\pi}_h)_{h=1}^H, \widehat{\nu} = (\widehat{\nu}_h)_{h=1}^H$.

In the following sections, we will introduce different estimators that compute the Bellman operator weights under different coverage assumptions. Depending on which estimator we use, we also get the corresponding convergence guarantees. Intuitively, the error of estimating the Bellman operator is inflated by an extra term coming from data corruption. To have a sense of the strength of such guarantees in terms of ϵ, we now present our first contribution, a result that states an algorithm-independent minimax lower bound for the corruption problem in the offline two-player zero-sum setting.

Theorem 1. *For every algorithm L, there exists a Markov game $\mathcal{G}$, an instance of the corrupted dataset, corruption level ϵ, and a data collecting distribution ρ, such that, with probability at least $1/4$, L will find a no-better than $\Omega(Hd\epsilon)$-approximate NE policy pair $(\widetilde{\pi}, \widetilde{\nu})$. That is, with a probability of at least $1/4$, we have, for every $s \in \mathcal{S}$:*

$$SubOpt(\widetilde{\pi}, \widetilde{\nu}, s) = \Omega(Hd\epsilon) .$$

4.1 Uniform Σ-Coverage and Corrupted Covariates

We start by considering the strongest coverage assumption first. Throughout Section 4.1, we assume that the corrupted data D satisfies the following.

Assumption 2 (Uniform Σ-coverage)**.** *For all $h \in [H]$ and $s \in \mathcal{S}$, $\mathbb{E}_\rho[\Sigma_h | s_1 = s] \succeq \kappa I$ for some $\kappa > 0$, where $\Sigma_h = \phi(s_h, a_h, b_h)\phi(s_h, a_h, b_h)^\top$.*

Assumption 2 implies that every state-action tuple in the support of the behavioral policy ρ is covered by the offline data. In Section 6 we further discuss the strength of this assumption relative to other coverage assumptions studied in the literature.

Inspired by the Robust Value Iteration algorithm (Zhang et al., 2022) that uses a Robust Least Squares (RLS) oracle as a subroutine for estimating the weights of the value function, we propose RLS-PMVI, which uses a similar oracle that relies on Assumption 2. However, it allows for an arbitrary corruption model, where an ϵ-fraction of the dataset can be arbitrarily corrupted, that is, any component of the data samples can be modified arbitrarily. RLS-PMVI returns an approximate NE policy pair that incurs the following bounds on the gap.

Theorem 2. *Suppose that Assumption 2 holds on an ϵ-corrupted dataset D corresponding to a linear Markov game. Then, given $\delta > 0$, with probability at least $1-\delta$, RLS-PMVI with bonus term $\Gamma_h(s,a,b) = 0$ achieves suboptimality gap upper bounded by*

$$\widetilde{O}\left(\sqrt{\frac{H(H+\gamma)^2 poly(d)}{\kappa^2 K}} + \frac{H(H+\gamma)}{\kappa}\epsilon\right) .$$

The proof of Theorem 2 is based on similar ideas to those used in the single-player setting. We provide the full proof in Appendix for completion. The following remarks are in order.

Remark 1. *Note that the order of ϵ is optimal. Furthermore, since the bonus term is 0, the algorithm does not require knowledge of ϵ. Thus, if we have uniform coverage, under a stricter corruption model, i.e., the features can also be corrupted, and without knowledge of ϵ, RLS-PMVI incurs a suboptimality gap with optimal dependence on ϵ.*

Remark 2. *Note that the coverage constant κ is in the order of $1/d$, since $\|\phi\|_2 \leq 1$. Thus, the bound corresponding to the corruption error becomes $O(H^2 d\epsilon)$.*

4.2 LRU Coverage and Clean Covariates

As mentioned in the previous section, Assumption 2 is a very strong requirement to make on the data. Thus, we now focus our attention on a relaxed scenario, where the observed data covers only policies of interest. In the two-player zero-sum setting, it turns out that coverage of the NE policy pair alone is not enough to efficiently compute an approximate solution. We state such an assumption below. Given dataset $\mathcal{D}$ of K trajectories, let us denote by

$$\Lambda_h = \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau, b_h^\tau)\phi(s_h^\tau, a_h^\tau, b_h^\tau)^\top + I \qquad (1)$$

the regularized sample covariance with respect to $\mathcal{D}$.

Assumption 3 (Low relative uncertainty)**.** *There exists a constant $c_1 > 0$ such that, for all $x \in \mathcal{S}$:*

$$\Lambda_h \succeq I + c_1 K \max\left\{\sup_{\nu} \mathbb{E}_{\pi^*,\nu}[\phi_h\phi_h^\top | s_1 = x], \sup_{\pi} \mathbb{E}_{\pi,\nu^*}[\phi_h\phi_h^\top | s_1 = x]\right\} ,$$

where $\phi_h = \phi(s_h, a_h, b_h)$, for any $h \in [H]$.

As shown in (Zhong et al., 2022) (and (Cui and Du, 2022) for tabular settings), Assumption 3 (and Assumption 6 for tabular settings[4]) is necessary for learning NE policies in two-player settings. Thus, the focus of Section 4.2 will be on data that satisfies such an assumption. An immediate problem that naturally follows is choosing the right robust estimator. The RLS estimator, described in the previous section, provides nice guarantees but it relies on Assumption 2. If we are to assume only LRU coverage on the data, then we need a different estimator.

To that end, we utilize a result by Chen et al. (2022a) that does not require Assumption 2 to hold. Their algorithm, SCRAM, utilizes an alternating minimization scheme to compute first-order stationary points. However, the utilization of such an oracle relies on the assumption that the covariates of the dataset are not corrupted. Translated into our setting, this assumption requires the features $\phi(s,a,b)$, and, as a consequence, tuples $(s,a,b) \in D$ to be clean, i.e., only an ϵ-fraction of the rewards and the next states are allowed to be arbitrarily corrupted. Below, we state the conditions of SCRAM and its guarantee. A detailed version can be found in Appendix.

Lemma 1 (Chen et al. (2022a))**.** *Given a dataset $D = \{x_i, y_i\}_{i\in[K]}$, which is an ϵ-corrupted version of dataset $\widetilde{D} = \{x_i, \widetilde{y}_i\}$, where $\widetilde{y}_i = \langle \omega_i^*, x_i\rangle + \xi_i$, $\epsilon < 0.499$, $\|x_i\|_2 \leq 1$ and ξ_i are conditionally zero-mean γ^2 sub-Gaussian, then SCRAM returns estimators $(\omega_k)_{k\in[K]}$, such that*

$$\|\omega - \omega^*\|_\Sigma \leq O(\epsilon\gamma + \gamma d^{1/2} K^{-1/4}) , \qquad (2)$$

omitting poly-log factors, where $\Sigma = (1/K)\sum_{k=1}^{K} x_k x_k^\top$ is the sample covariance.[5]

In Section 4.2, we use SCRAM as a robust estimator with bonus term defined as

$$\Gamma_h(s,a,b) = \left(\sqrt{K}\mathcal{E} + 2H\sqrt{d}\right)\|\phi(s,a,b)\|_{\Lambda_h^{-1}} ,$$

where Λ_h is defined in Equation 1 and $\mathcal{E}$ denotes the upper bound in Equation (2).

[4] See Section 6.

[5] For explicit bounds, see Appendix.

We are now ready to state the main result of Section 4.2, which gives an upper bound on the suboptimality gap of the policy pair returned by SCRAM-PMVI.

Theorem 3. *Suppose that Assumption 3 holds with given constant c_1. Let $\delta > 0$, $\epsilon < 1/2$ and let D be the ϵ-corrupted version of the dataset $\widetilde{D}$ comprised of K trajectories of length H, where $K \geq \log(\min(K, d))/\epsilon$ and $(\widetilde{s}_h^\tau, \widetilde{a}_h^\tau, \widetilde{b}_h^\tau) = (s_h^\tau, a_h^\tau, b_h^\tau)$, for all $\tau \in [K]$, $h \in [H]$. Then, with probability at least $1 - \delta$, SCRAM-PMVI outputs $(\widehat{\pi}, \widehat{\nu})$ that satisfy, for every $s \in \mathcal{S}$:*

$$\mathit{SubOpt}(\widehat{\pi}, \widehat{\nu}, s) \leq \widetilde{O}\left(\frac{1}{\sqrt{c_1}}(\gamma + H)H\sqrt{d}\epsilon + \frac{H^2 d}{\sqrt{c_1 K}}\right) .$$

Note the $\sqrt{d}$ factor in the term coming from the corruption error. This might seem contradictory to our lower bound of Theorem 1 at first. However, there is a hidden dependence on d in the c_1 constant, since c_1 cannot be arbitrarily large. By Assumption 3, first one can easily see that $0 < c_1 \leq 1$. Moreover, as shown in the proof of Proposition 1 (see Appendix), c_1 can be $O(1/d)$, in which case the bounds would be written as

$$\mathrm{SubOpt}(\widehat{\pi}, \widehat{\nu}, s) \leq \widetilde{O}\left((\gamma + H)Hd\epsilon + H^2 d^{3/2} K^{-1/2}\right).$$

Note that one can always construct other examples in which the dependence of the gap on d is looser. However, here we are interested in showing that there is an explicit dependence in order to confirm the matching order.

It is worth mentioning that knowledge of ϵ is required by both SCRAM and RLS oracles. Moreover, agnostic learning without knowledge of ϵ is impossible for linear Markov games, without uniform data coverage, as the next result shows. The proof follows immediately from that of Theorem 3.4 of (Zhang et al., 2022), by restricting the second player's action space to a single action and observing that the LRU coverage reduces to their finite relative condition number assumption.

Corollary 1. *Under Assumption 3, for every algorithm L that achieves a diminishing suboptimality gap in a clean environment, there exist linear Markov games $\mathcal{G}_1$ and $\mathcal{G}_2$, an instance of corrupted data, and a data collecting distribution ρ, such that, for every $\epsilon \in (0, 1/2]$, L achieves $\mathit{SubOpt}(\widehat{\pi}, \widehat{\nu}, s) \geq 1/2$, for any $s \in \mathcal{S}$, with probability at least $1/4$, on at least one of the games.*

With this result, we end our discussion under corrupted data coverage assumptions. In order to give a formal characterization of the used coverage assumptions and other similar ones in the literature, we provide a formal discussion in Section 6. In the next section, we turn our attention to the more difficult setting where coverage guarantees are no longer present in the corrupted data and provide near-optimal bounds on the suboptimality of SCRAM-PMVI with novel bonus terms designed to account for this corrupted coverage.

5 RESULTS UNDER COVERAGE ON CLEAN DATA

In previous sections, we considered the problem of corruption in the linear setting, under the assumption that the corrupted data preserves the coverage that is necessary to solve the underlying Markov game. However, this might be too restrictive of an assumption since, in the worst case, the attacker would corrupt precisely those state-action tuples that belong to trajectories covered by LRU policies. Thus, the next natural question is whether we can approximately solve the problem given that we are not guaranteed that the available corrupted data satisfies the LRU coverage assumption.

Apart from not assuming any guarantees on coverage, we also assume fully arbitrary corruption, in the sense that the attacker can arbitrarily corrupt an ϵ-fraction of both covariates and observations. If we are to apply SCRAM in this setting, we would expect an additional error coming from the corruption of covariates, and another one coming from the corrupted coverage.

We find that by carefully designing a bonus term that also takes into account these two additional errors, we are able to approximately solve the corruption robust offline two-player zero-sum game, as the following result states.

Theorem 4. *Suppose that the condition of Assumption 3 is satisfied only on the clean dataset $\widetilde{D}$, for a given constant c_1, that is, assume that the clean sample covariance matrix $\sum_{\tau=1}^K \widetilde{\phi}_h^\tau (\widetilde{\phi}_h^\tau)^\top + I$, where $\widetilde{\phi}_h^\tau$ denotes the clean feature of the sample at time-step h of episode τ, satisfies Assumption 3. Furthermore, let $\delta > 0$, $\epsilon \in (0, 1/2)$, and $K \geq \log(\min\{K, d\})/\epsilon$. Then, under Assumption 1, with probability at least $1 - \delta$, SCRAM-PMVI with bonus defined as*

$$\Gamma_h(\cdot) = \left(\sqrt{(1-\epsilon)K}\mathcal{E} + (\sqrt{\epsilon K} + 2)H\sqrt{d}\right) \|\phi(\cdot)\|_{\Lambda_h^{-1}}$$

returns $(\widehat{\pi}, \widehat{\nu})$ that satisfy, for every $s \in \mathcal{S}$:

$$\mathit{SubOpt}(\widehat{\pi}, \widehat{\nu}, s) \leq O\left(\frac{1}{\sqrt{c_1}}(\gamma + H)Hd\sqrt{\epsilon} + \frac{H^2 d}{\sqrt{c_1 K}}\right) .$$

Note that we incur an additional $\sqrt{d}$ factor in the corruption error and that the order of ϵ is not optimal. This is due to both the error from corrupted covariates and the one from the corrupted coverage. Thus, the next natural question is whether these bounds can be improved, at least in terms of the dependence on ϵ. We answer this question in the affirmative. We show

that, under a mild assumption on the feature space, we can improve the order of ϵ at the cost of an extra $d^{3/2}$ factor, thus recovering the optimal dependency on the corruption level. First, we state the assumption.

Assumption 4. *Given a linear Markov game as in Definition 2, we assume that the features satisfy* $\min_{s,a,b} \|\phi(s,a,b)\|_2 \geq c_2$, *for some* $c_2 > 0$.

We emphasize that such an assumption is not restrictive and that c_2 is in the order of $1/\sqrt{d}$ under various feature constructions, such as random Fourier features (Rahimi and Recht, 2007). We find that Assumption 4 is enough to obtain order-optimal bounds in terms of ϵ, albeit having an additional $O(d)$ term when comparing with Theorem 4 coming from the constant c_2.

Theorem 5. *Suppose that the conditions of Theorem 4 and Assumption 4 hold. Then, with probability at least* $1-\delta$, *SCRAM-PMVI with bonus* $\Gamma_h(\cdot)$ *defined as*

$$\left(2(1-\epsilon)K\mathcal{E} + \epsilon KH\sqrt{d} + H\sqrt{Kd}\right)\left\|\phi(\cdot)^\top \Lambda_h^{-1}\right\|_2 ,$$

returns $(\widehat{\pi}, \widehat{\nu})$ *that satisfy, for every* $s \in \mathcal{S}$:

$$SubOpt(\widehat{\pi}, \widehat{\nu}, s) \leq \widetilde{O}\left(\frac{1}{c_1 c_2} H^2 d^{3/2}\epsilon + \frac{H^2 d^{3/2}}{c_1 c_2 \sqrt{K}}\right) .$$

This improvement comes as a result of a novel bonus term and a different style of analysis based on an application of the Woodbury matrix identity to account for the extra terms coming from corruption.

6 DISCUSSION ON MARL COVERAGE ASSUMPTIONS

In this section, we discuss the relationship between various coverage assumptions used in the literature. First, we state three additional assumptions, apart from Assumption 2 and 3.

Assumption 5 (Single-policy coverage)**.** *The NE strategy pair* (π^*, ν^*) *is covered by the dataset* D.

Assumption 6 (Unilateral coverage)**.** *For all strategies* $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ *and* $\nu : \mathcal{S} \to \Delta(\mathcal{B})$, *the strategy pairs* (π^*, ν) *and* (π, ν^*) *are covered by* D.

Assumption 7 (Uniform coverage)**.** *For all* $h \in [H]$ *and* $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, *the tuple* (s,a,b) *at time step* h *is covered by* D.

Note that Assumption 5 is the weakest assumption. Moreover, it is a direct extension of Assumption 3.2 of (Zhang et al., 2022), while Assumption 7 is an extension of the uniform policy coverage in the single-player setting (Cai et al., 2020). We show that Assumption 6 implies Assumption 3 with high probability, when the feature matrix has full rank, and that, for tabular Markov games,[6] these two assumptions are equivalent. All proofs of our results can be found in the Appendix.

Proposition 1. *Let* $\Phi \in \mathbb{R}^{SAB \times d}$ *denote the feature matrix. Assume* Φ *has full rank and let* $\delta \in (0,1)$. *Then, if Assumption 6 holds, there exists a positive constant that depends on* δ *for which Assumption 3 holds, with probability at least* $1-\delta$. *Moreover, in the tabular Markov game setting, these two assumptions are equivalent.*

Furthermore, it is obvious that uniform coverage is stronger than unilateral coverage. The relation between Assumption 2 and Assumption 7 is given as follows.

Proposition 2. *Assume that* Φ *has full rank. Then, if Assumption 2 holds, Assumption 7 holds. Moreover, in the tabular MG setting, these two assumptions are equivalent.*

As already shown in (Zhong et al., 2022), assuming that the collected data covers only the NE strategy pair $\pi^* = (\mu^*, \nu^*)$ is not enough to learn an approximate NE policy pair. Indeed, Assumption 3 is necessary for solving offline two-player zero-sum Markov games, even in clean environments.

Remark 3. *Note that, under the assumption that* Φ *is full rank (if not, orthogonalization can be applied), the given coverage assumptions are listed according to their strength. With high probability, we have*

$$A.2 \Rightarrow A.7 \Rightarrow A.6 \Rightarrow A.3 \Rightarrow A.5 .$$

Moreover, for the tabular setting, Assumption 2 is equivalent to 7 and Assumption 6 is equivalent to 3.

7 RELATED WORK

Offline RL. Our work is related to the offline RL literature, where there have been substantial developments in recent years, both on the empirical front (Jaques et al., 2019; Laroche et al., 2019; Fujimoto et al., 2019; Kumar et al., 2020; Agarwal et al., 2020; Kidambi et al., 2020) and the theoretical front (Jin et al., 2021; Xie et al., 2021; Rashidinejad et al., 2021; Uehara and Sun, 2021; Zanette et al., 2021b). As previously mentioned, coverage assumptions on the data are key in this setting, and there has been a variety of different assumptions in the single-player setting, starting from the all-policy coverage $\|d^\pi/d^\rho\|_\infty \leq B$, for all π (Munos and Szepesvári, 2008), to optimal policy coverage $\|d^*/d^\rho\|_\infty \leq B$ (Xie et al., 2021), and α-regularized optimal policy coverage $\|d^*_\alpha/d^\rho\|_\infty \leq B$ (Zhan et al., 2022). First, all of these assumptions are

[6] For tabular Markov games, we have $\phi(s,a,b) = e_{s,a,b}$, where $e_{s,a,b}$ is the SAB-dimensional zero vector with 1 in the (s,a,b) entry.

with respect to the original data-collecting distribution d^ρ, while our weakest assumption (LRU coverage) relies only upon the sample covariance matrix of the data. Second, note that the weakest of the assumptions above (α-regularized optimal policy coverage) requires coverage of only the optimal solution to the regularized LP problem, while coverage of the NE policy pair and its unilateral neighbors is necessary for finding NE policy pairs in our setting (Zhong et al., 2022). This is arguably due to the higher complexity of the problem compared to the single-player setting.

Adversarial attacks in RL. Our work also adds to the vast literature on adversarial attacks in ML (Szegedy et al., 2013; Biggio et al., 2013; Nguyen et al., 2015; Papernot et al., 2017; Biggio et al., 2012; Li et al., 2016; Xiao et al., 2012) and the existing body of work on adversarial attacks in RL and MARL (Huang et al., 2017; Lin et al., 2017; Wu et al., 2022; Gleave et al., 2020; Sun et al., 2020a,b; Ma et al., 2019; Rakhsha et al., 2021; Everitt et al., 2017; Huang and Zhu, 2019; Rangi et al., 2022; Mohammadi et al., 2023). Specifically, we consider the problem of robustness to data corruption, which is a type of training-time attack (Mei and Zhu, 2015; Xiao et al., 2015; Rakhsha et al., 2020). Popular types of defense against such attacks include randomized smoothing (Cohen et al., 2019; Wu et al., 2021), outlier detection (Diakonikolas et al., 2019) and robust estimator methods (Chen et al., 2022a; Diakonikolas et al., 2017; Banihashem et al., 2023). In this work, we use the latter methods, for both weight estimation in linear games and mean estimation in tabular ones. Arguably, the closest work to ours is (Zhang et al., 2022), which studies the corruption problem in single-agent RL. On the other hand, while our analysis of the lower bounds is inspired by their work, our analysis leads to tighter upper bounds in terms of ϵ and H for two-player zero-sum Markov games. Yang et al. (2022) also study the robustness problem in the offline RL setting. However, their attack model assumes observation perturbations of bounded radius. Our attack model is stronger since it allows for the arbitrary perturbation of the tuples. Recently, adversarial corruption in the online setting has been studied in linear contextual bandits (He et al., 2022) and more generally in MDPs with general function approximation (Ye et al., 2023). These works also broadly relate to corruption-robust approaches in distributed RL (Chen et al., 2022b; Fan et al., 2021), that focus on MDP settings.

Reward perturbation in MARL. Ma et al. (2022) study the problem of reward design in online systems with no-regret learners. Their goal is to modify the utility function iteratively so that the agents converge to a desired action profile, while maintaining low cost of perturbation. While we also consider multi-agent systems subject to a third-party intervention, our focus is in the offline setting with an adversarial corruption framework, with an emphasis on the defense front. On the other hand, Wu et al. (2023) is more closely related to ours. They also study an offline corruption model, where the attacker can perturb the reward signal in order to enforce a particular policy, while at the same time maintaining a low perturbation cost. While their focus in on designing cost-efficient attacks of that form, our work studies defenses against a broad class of data poisoning attacks, defined by the Huber contamination model.

Learning in Markov games. Finally, our work also relates to the research area of learning in Markov games (Vrancx et al., 2008; Littman, 1994, 2001; Tian et al., 2021; Wang and Sandholm, 2002; Sayin et al., 2021; Xie et al., 2020). In particular, we consider offline two-player zero-sum games, which have been recently considered in (Cui and Du, 2022) for tabular settings and (Zhong et al., 2022) for linear settings. While they solve the offline problem by assuming that the collected data is sampled from a benign behavioral policy, we consider the problem of robustness of their proposed methods to data corruption. More specifically, we assume that an ϵ fraction of collected data has been corrupted, and our aim is to learn NE strategy pairs under such an assumption. When $\epsilon = 0$, we recover the same bounds as theirs on the suboptimality gap.

8 CONCLUSION

We considered the problem of data corruption in offline two-player zero-sum Markov games. Our contribution was to provide an extensive characterization of the problem under various coverage assumptions on both the clean and corrupted data. To the best of our knowledge, we are the first to provide such a characterization for the problem of corruption in offline Markov games. For the hardest setting where minimal coverage is guaranteed only on the clean data, we are able to match the optimal order of ϵ under mild structural assumptions, thus providing a full picture of this setting. There are many interesting future directions to pursue: i) studying robustness under adversarial corruption in Markov games with general function approximation; ii) extending the two-player zero-sum Markov game setting to online data corruption, where, in each round the reward/transition data is corrupted with probability ϵ; iii) studying robustness to adversarial corruption in larger structured games (e.g. Markov potential games).

Acknowledgements

The authors thank anonymous reviewers for their valuable suggestions and comments. This research was, in

part, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 467367360.

References

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An Optimistic Perspective on Offline Reinforcement Learning. In *ICML*, 2020.

Ainesh Bakshi and Adarsh Prasad. Robust Linear regression: Optimal Rates in Polynomial time. In *ACM SIGACT*, 2021.

Kiarash Banihashem, Adish Singla, and Goran Radanovic. Defense Against Reward Poisoning Attacks in Reinforcement Learning. *Trans. Mach. Learn. Res.*, 2023.

Christopher Berner et al. Dota 2 with Large Scale Deep Reinforcement Learning. *CoRR*, abs/1912.06680, 2019.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning Attacks Against Support Vector Machines. *CoRR*, abs/1206.6389, 2012.

Battista Biggio et al. Evasion Attacks Against Machine Learning at Test Time. In *ECML PKDD*, 2013.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably Efficient Exploration in Policy Optimization. In *ICML*, 2020.

Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. Online and Distribution-free Robustness: Regression and Contextual Bandits with Huber Contamination. In *FOCS*, 2022a.

Yiding Chen, Xuezhou Zhang, Kaiqing Zhang, Mengdi Wang, and Xiaojin Zhu. Byzantine-robust Online and Offline Distributed Reinforcement Learning. *CoRR*, abs/2206.00165, 2022b.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing. In *ICML*, 2019.

Qiwen Cui and Simon S Du. When is Offline Two-player Zero-sum Markov Game Solvable? *CoRR*, abs/2201.03522, 2022.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being Robust (in High Dimensions) Can be Practical. In *ICML*, 2017.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A Robust Meta-algorithm for Stochastic Optimization. In *ICML*, 2019.

Tom Everitt, Victoria Krakovna, Laurent Orseau, Marcus Hutter, and Shane Legg. Reinforcement Learning with a Corrupted Reward Channel. *CoRR*, abs/1705.08417, 2017.

Xiaofeng Fan, Yining Ma, Zhongxiang Dai, Wei Jing, Cheston Tan, and Bryan Kian Hsiang Low. Fault-tolerant Federated Reinforcement Learning with Theoretical Guarantee. In *NeurIPS*, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy Deep Reinforcement Learning without Exploration. In *ICML*, 2019.

Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial Policies: Attacking Deep Reinforcement Learning. In *ICLR*, 2020.

Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly Optimal Algorithms for Linear Contextual Bandits with Adversarial Corruptions. In *NeurIPS*, 2022.

Harold V Henderson and Shayle R Searle. On Deriving the Inverse of a Sum of Matrices. *Siam Review*, 1981.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial Attacks on Neural Network Policies. *CoRR*, abs/1702.02284, 2017.

Yunhan Huang and Quanyan Zhu. Deceptive Reinforcement Learning under Adversarial Manipulations on Cost Signals. In *GameSec*, 2019.

Natasha Jaques et al. Way Off-policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog. *CoRR*, abs/1907.00456, 2019.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is Pessimism Provably Efficient for Offline RL? In *ICML*, 2021.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based Offline Rinforcement Learning. In *NeurIPS*, 2020.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for Offline Reinforcement Learning. In *NeurIPS*, 2020.

Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe Policy Improvement with Baseline Bootstrapping. In *ICML*, 2019.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *CoRR*, abs/2005.01643, 2020.

Bo Li, Yining Wang, Aarti Singh, and Yevgeniy Vorobeychik. Data Poisoning Attacks on Factorization-based Collaborative Filtering. In *NeurIPS*, 2016.

Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of Adversarial Attack on Deep Reinforcement Learning Agents. In *IJCAI*, 2017.

Michael L. Littman. Markov Games as a Framework for Multi-agent Reinforcement Learning. In *ICML*, 1994.

Michael L Littman. Value-function Reinforcement Learning in Markov Games. *Cognitive Systems Research*, 2:55–66, 2001.

Yuzhe Ma, Xuezhou Zhang, Wen Sun, and Jerry Zhu. Policy Poisoning in Batch Reinforcement Learning and Control. In *NeurIPS*, 2019.

Yuzhe Ma, Young Wu, and Xiaojin Zhu. Game Redesign in No-regret Game Playing. In *IJCAI*, 2022.

Shike Mei and Xiaojin Zhu. Using Machine Teaching to Identify Optimal Training-set Attacks on Machine Learners. In *AAAI*, 2015.

Mohammad Mohammadi, Jonathan Nöther, Debmalya Mandal, Adish Singla, and Goran Radanovic. Implicit poisoning attacks in two-agent reinforcement learning: Adversarial policies for training-time attacks. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 1835–1844, 2023.

Rémi Munos and Csaba Szepesvári. Finite-time Bounds for Fitted Value Iteration. *JMLR*, 2008.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In *CVPR*, 2015.

Yunpeng Pan et al. Agile Autonomous Driving Using End-to-end Deep Imitation Learning. *CoRR*, abs/1709.07174, 2017.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical Black-box Attacks Against Machine Learning. In *ACM*, 2017.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust Regression with Covariate Filtering: Heavy Tails and Adversarial Contamination. *CoRR*, abs/2009.12976, 2020.

Ali Rahimi and Benjamin Recht. Random Features for Large-scale Kernel Machines. In *NeurIPS*, 2007.

Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy Teaching via Environment Poisoning: Training-time Adversarial Attacks against Reinforcement Learning. In *ICML*, 2020.

Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy Teaching in Reinforcement Learning via Environment Poisoning Attacks. *JMLR*, 2021.

Anshuka Rangi, Haifeng Xu, Long Tran-Thanh, and Massimo Franceschetti. Understanding the Limits of Poisoning Attacks in Episodic Reinforcement Learning. *CoRR*, abs/2208.13663, 2022.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism. In *NeurIPS*, 2021.

Muhammed Sayin, Kaiqing Zhang, David Leslie, Tamer Basar, and Asuman Ozdaglar. Decentralized Q-learning in Zero-sum Markov Games. In *NeurIPS*, 2021.

Lloyd S Shapley. Stochastic Games. *PNAS*, 39, 1953.

David Silver et al. Mastering the Game of Go Without Human Knowledge. *Nature*, 550, 2017.

Jianwen Sun, Tianwei Zhang, Xiaofei Xie, Lei Ma, Yan Zheng, Kangjie Chen, and Yang Liu. Stealthy and Efficient Adversarial Attacks Against Deep Reinforcement Learning. In *AAAI*, 2020a.

Yanchao Sun, Da Huo, and Furong Huang. Vulnerability-aware Poisoning Mechanism for Online RL with Unknown Dynamics. In *ICLR*, 2020b.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing Properties of Neural Networks. *CoRR*, abs/1312.6199, 2013.

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Online Learning in Unknown Markov Games. In *ICML*, 2021.

Masatoshi Uehara and Wen Sun. Pessimistic Model-based Offline Reinforcement Learning Under Partial Coverage. *CoRR*, abs/2107.06226, 2021.

Peter Vrancx, Katja Verbeeck, and Ann Nowé. Decentralized Learning in Markov Games. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38, 2008.

Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. Supervised Reinforcement Learning with Recurrent Neural Network for Dynamic Treatment Recommendation. In *ACM SIGKDD*, 2018.

Xiaofeng Wang and Tuomas Sandholm. Reinforcement Learning to Play an Optimal Nash Equilibrium in Team Markov Games. In *NeurIPS*, 2002.

Fan Wu, Linyi Li, Zijian Huang, Yevgeniy Vorobeychik, Ding Zhao, and Bo Li. CROP: Certifying Robust Policies for Reinforcement Learning through Functional Smoothing. *CoRR*, abs/2106.09292, 2021.

Young Wu, Jermey McMahan, Xiaojin Zhu, and Qiaomin Xie. Reward Poisoning Attacks on Offline Multi-agent Reinforcement Learning. *CoRR*, abs/2206.01888, 2022.

Young Wu, Jeremy McMahan, Xiaojin Zhu, and Qiaomin Xie. Reward Poisoning Attacks on Offline Multi-agent Reinforcement Learning. In *AAAI*, 2023.

Han Xiao, Huang Xiao, and Claudia Eckert. Adversarial Label Flips Attack on Support Vector Machines. In *ECAI*, 2012.

Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is Feature Selection Secure Against Training Data Poisoning? In *ICML*, 2015.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning Zero-sum Simultaneous-move Markov Games Using Function Approximation and Correlated Equilibrium. In *COLT*, 2020.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent Pessimism for Offline Reinforcement Learning. In *NeurIPS*, 2021.

Rui Yang, Chenjia Bai, Xiaoteng Ma, Zhaoran Wang, Chongjie Zhang, and Lei Han. RORL: Robust Offline Reinforcement Learning via Conservative Smoothing. In *NeurIPS*, 2022.

Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Corruption-robust Algorithms with Uncertainty Weighting for Nonlinear Contextual Bandits and Markov Decision Processes. In *ICML*, 2023.

Andrea Zanette, Ching-An Cheng, and Alekh Agarwal. Cautiously Optimistic Policy Optimization and Exploration with Linear Function Approximation. In *COLT*, 2021a.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable Benefits of Actor-critic Methods for Offline Reinforcement Learning. In *NeurIPS*, 2021b.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline Reinforcement Learning with Realizability and Single-policy Concentrability. In *COLT*, 2022.

Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun. Corruption-robust Offline Reinforcement Learning. In *AISTATS*, 2022.

Han Zhong, Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang. Pessimistic Minimax Value Iteration: Provably Efficient Equilibrium Learning from Offline Datasets. In *ICML*, 2022.

Checklist

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

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

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

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

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

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

Appendix

Table of Contents

A Proofs of Section 4

In this section, we derive the proofs of the results in Section 4.

A.1 Proof of Theorem 1

We first restate the result.

Statement. *For every algorithm L, there exists a Markov game $\mathcal{G}$, an instance of the corrupted dataset, corruption level ϵ, and a data collecting distribution ρ, such that, with probability at least $1/4$, L will find a no-better than $\Omega(Hd\epsilon)$-approximate NE policy pair $(\widetilde{\pi}, \widetilde{\nu})$. That is, with a probability of at least $1/4$, we have, for every $s \in \mathcal{S}$:*

$$SubOpt(\widetilde{\pi}, \widetilde{\nu}, s) = \Omega(Hd\epsilon) \ .$$

Proof. We will construct an example to prove our statement. Consider the following Markov game $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, P, H, r, \gamma, s_0)$, with $|\mathcal{S}| = S$, $|\mathcal{A}| = A$, $|\mathcal{B}| = B$, deterministic transitions, episode length H and initial state s_0, with $S \leq (AB)^{H/2}$. Note that, in this case, we have $d = SAB$. Here r denotes the reward with respect to the max player. Assume that the transition dynamics follow a tree structure, that is, let $\mathcal{T}$ be a tree with nodes represented by states $s \in \mathcal{S}$ and edges represented by action tuples $(a, b) \in \mathcal{A} \times \mathcal{B}$, with root node s_0, such that, for every $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, node s is parent to node $\arg\max_{s'} P(s'|s, a, b)$. We denote by $p(s)$ the parent of node s. Moreover, assume that all states represented by the leaf nodes are self-absorbing states, i.e. the state does not change, no matter what action is taken. Let q denote the depth of $\mathcal{T}$. Note that we have

$$q = O(\lceil (\log_{AB}(S(AB-1)+1) - 1 \rceil) \ .$$

Now let us denote by $\mathcal{L}_i$ the subset of $\mathcal{S}$ containing the states represented by nodes in level i of $\mathcal{T}$, for all $i \in \{0\} \cup [q]$, and let s_i^j enumerate the states in level $\mathcal{L}_i$, for $j \in [(AB)^{i-1}]$. Let us define the reward function as follows. Fix a sequence of states $(s_0^*, s_1^*, \ldots, s_q^*) \in \mathcal{L}_0 \times \ldots \times \mathcal{L}_q$, where $s_0^* = s_0$ and s_i^* is a node in the ith level of $\mathcal{T}$ such that $P(s_i^* |\, s_{i-1}^*, a, b_1) = 1$, for all $i \in [q]$ and $a \in \mathcal{A}$. Furthermore, for all $s \neq s_{i-1}^*$ and $(a, b) \in \mathcal{A} \times \mathcal{B}$, we have $P(s_i^*|s, a, b) = 0$, for all $i \in [q]$. Let $\alpha \in (0, 1/3)$ and assume (s_q^*, a_1, b_1) is the least represented state according to data collecting distribution d^ρ.

Then, for all $i \in \{0, 1, \dots, q-1\}$, we define

$$r(s_i^*, a, b) = \begin{cases} \alpha & \text{if } a = a_1, b = b_1 \\ 2\alpha & \text{if } a = a_1, b \neq b_1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad r(s_q^*, a, b) = \begin{cases} \alpha & \text{if } a = a_1, b = b_1 \\ X & \text{if } a = a_1, b = b_2 \\ 3\alpha & \text{if } a = a_1, b \in \mathcal{B} \setminus \{b_1, b_2\} \\ 0 & \text{otherwise} \end{cases}$$

where X is a Bernoulli random variable with parameter 2α. On the other hand, for all $s \in \mathcal{S} \setminus \{s_0^*, \dots, s_q^*\}$, let

$$r(s, a, b) = \begin{cases} 1 & \text{if } a = a_1 \\ 0 & \text{otherwise} \end{cases}$$

Let us determine the value of the game in each state, using the method of backward induction. Set $V_{H+1}^*(s) = 0$, for all $s \in \mathcal{S}$. Then, for all $s \in \mathcal{L}_q \setminus \{s_q^*\}$, we have

$Q_H^*(s, \cdot, \cdot)$	b_1	b_2	$\dots$	b_B
a_1	1	1	$\dots$	1
a_2	0	0	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_A	0	0	$\dots$	0

and

$Q_H^*(s_q^*, \cdot, \cdot)$	b_1	b_2	b_3	$\dots$	b_B
a_1	α	2α	3α	$\dots$	3α
a_2	0	0	0	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_A	0	0	0	$\dots$	0

Thus, we obtain $V_H^*(s) = 1$, for all $s \in \mathcal{L}_d \setminus \{s_q^*\}$ and $V_H^*(s_q^*) = \alpha$. Moreover, note that, for all $s \in \mathcal{L}_{d-1} \setminus \{s_{q-1}^*\}$, we have

$Q_{H-1}^*(s, \cdot, \cdot)$	b_1	b_2	$\dots$	b_B
a_1	2	2	$\dots$	2
a_2	0	0	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_A	0	0	$\dots$	0

and

$Q_{H-1}^*(s_q^*, \cdot, \cdot)$	b_1	b_2	b_3	$\dots$	b_B
a_1	2α	3α	4α	$\dots$	4α
a_2	0	0	0	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_A	0	0	0	$\dots$	0

and thus $V_{H-1}^*(s_q^*) = 2\alpha$. Continuing in this fashion, we obtain

$$V_1^*(s_0) = H\alpha,$$

where the first q steps come from the trajectory $(s_0^*, \dots, s_q^*)$, and the rest of the $H - q$ steps come from staying in state s_q^*.

Now let $\mathcal{G}'$ be a Markov game that is identical to $\mathcal{G}$, except for one difference. Let r' denote the reward function of $\mathcal{G}'$. Then $r'(s_q^*, a_1, b_1) = r(s_q^*, a_1, b_1) + Ber(2\alpha)$, and $r'(s, a, b) = r(s, a, b)$, for all other state-action tuples. Then for $\mathcal{G}'$ we have

$\widetilde{Q}_H^*(s, \cdot, \cdot)$	b_1	$\dots$	b_B
a_1	1	$\dots$	1
a_2	0	$\dots$	0
$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_A	0	$\dots$	0

and

$\widetilde{Q}_H^*(s_q^*, \cdot, \cdot)$	b_1	b_2	b_3	$\dots$	b_B
a_1	3α	2α	2α	$\dots$	2α
a_2	0	0	0	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_A	0	0	0	$\dots$	0

where $\widetilde{Q}(\cdot, \cdot, \cdot)$ denotes the matrices of Q-values of NE policies for $\mathcal{G}'$. Note that the trajectory traversed by the NE policy pair in $\mathcal{G}'$ is still $(s_0^*, \dots, s_{q-1}^*, s_q^*)$. However, when at state s_q^*, the NE policy is (a_1, b_2), instead of (a_1, b_1). Thus, we have

$$\widetilde{V}_1^*(s_0) = (2H - d)\alpha,$$

since the system will stay in state s_q^* for $H-q$ steps, until the episode ends. Note that no policy pair can be simultaneously optimal in both games. In the worst case, a policy pair which is a NE in $\mathcal{G}'$ will incur a suboptimality gap of

$$(H-q)\alpha = (H - \lceil(\log_{AB}(S(AB-1)+1)-1\rceil)\alpha \geq \Omega(H\alpha),$$

in $\mathcal{G}$, where the second inequality follows from the fact that $S \leq (AB)^{H/2}$.

Now, let $\alpha = SAB\epsilon/2$. Since (s_q^*, a_1, b_1) is the least represented state with respect to d^ρ, by pigeon-hole principle, we must have $d^\rho(s_d^*, a_1, b_1) \leq 1/SAB$. Assume the adversary uses all its budget only to perturb the reward of this state-action tuple. Concretely, if the game is $\mathcal{G}$, then the adversary perturbs it into $\mathcal{G}'$ by adding $Ber(SAB\epsilon)$ to $r(s_q^*, a_1, b_1)$.

With probability at least $1/2$, the number of times (s_q^*, a_1, b_1) is counted in a dataset with KH tuples is no more than KH/SAB, since (s_q^*, a_1, b_1) is the least represented tuple. Conditioned on this, with probability at least $1/2$, the reward seen from (s_q^*, a_1, b_1) is $2SAB\epsilon$. Thus, perturbing the reward of (s_q^*, a_1, b_1) at least $KH\epsilon$ times is enough to make one of the games indistinguishable from the other one. Thus, the agent will inevitably incur

$$\mathrm{SubOpt}(\pi, \nu, s) \geq \Omega(HSAB\epsilon).$$

□

A.2 Proof of Theorem 2

The main result of this section relies on the RLS oracle guarantee stated below.

Theorem 6. *(Zhang et al., 2022) Given an ϵ-corrupted dataset $D = \{x_i, y_i\}_{i\in[n]}$, where the clean data is generated as $\widetilde{x}_i \sim \beta$, $\mathbb{P}(\|\widetilde{x}_i\| \leq 1) = 1$, $\widetilde{y}_i = \widetilde{x}_i^\top \omega^* + \xi_i$, where ξ_i is zero-mean σ^2-variance sub-Gaussian random noise, then a robust least square estimator returns an estimator ω such that, if $\mathbb{E}_\beta[xx^\top] \succeq \kappa I$, for some strictly positive constant κ, then with probability at least $1-\delta$, we have*

- *If $\mathbb{E}_\beta[xx^\top] \succeq \kappa I$, then with probability at least $1-\delta$, we have*

$$\|\omega^* - \omega\|_2 \leq c_1(\delta) \cdot \left(\sqrt{\frac{\sigma^2 poly(d)}{\kappa^2 n}} + \frac{\sigma}{\kappa}\epsilon\right);$$

- *With probability at least $1-\delta$, we have*

$$\mathbb{E}\left[\left\|\widetilde{x}^\top(\omega^* - \omega)\right\|_2^2\right] \leq c_2(\delta) \cdot \left(\frac{\sigma^2 poly(d)}{n} + \sigma^2\epsilon\right),$$

where c_1 and c_2 hide constants and polylog$(1/\delta)$ terms.

Using the above guarantee, we are now ready to prove Theorem 2. First, we recall its statement.

Statement. *Suppose that Assumption 2 holds on an ϵ-corrupted dataset D corresponding to a linear Markov game. Then, given $\delta > 0$, with probability at least $1-\delta$, RLS-PMVI with bonus term $\Gamma_h(s,a,b) = 0$ achieves suboptimality gap upper bounded by*

$$\widetilde{O}\left(\sqrt{\frac{H(H+\gamma)^2 poly(d)}{\kappa^2 K}} + \frac{H(H+\gamma)}{\kappa}\epsilon\right).$$

Proof. Let us define the error in Theorem 6 as

$$\mathcal{E}_1(\epsilon, K, D, H, \sigma) = c_1(\delta) \cdot \left(\sqrt{\frac{\sigma^2 poly(d)}{\kappa^2 K}} + \frac{\sigma}{\kappa}\epsilon\right). \tag{3}$$

First, we provide upper and lower bounds on the Bellman error.

Lemma 2. *Given the tuple (s, a, b), let us define the Bellman error as*

$$\underline{\iota}_h(s,a,b) = (\mathbb{B}_h \underline{V}_{h+1})(s,a,b) - \underline{Q}_h(s,a,b) \ ,$$
$$\overline{\iota}_h(s,a,b) = (\mathbb{B}_h \overline{V}_{h+1})(s,a,b) - \overline{Q}_h(s,a,b) \ ,$$

for all $h \in [H]$. Then, with probability at least $1-\delta$, we have

$$-\mathcal{E}_1(\epsilon, K, D, H, \sigma) \leq \underline{\iota}_h(s,a,b) \leq \mathcal{E}_1(\epsilon, K, D, H, \sigma) \ ,$$
$$-\mathcal{E}_1(\epsilon, K, D, H, \sigma) \leq -\overline{\iota}_h(s,a,b) \leq \mathcal{E}_1(\epsilon, K, D, H, \sigma) \ ,$$

for all $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ and $h \in [H]$.

Proof. We start by deriving upper bounds on the Bellman error. Note that, with probability at least $1-\delta$, Theorem 6 and Assumption 2 imply

$$|\phi^\top \underline{\omega}_h - (\mathbb{B}_h \underline{V}_{h+1})(s,a,b)| \leq \|\phi(s,a,b)\|_2 \, \|\underline{\omega}_h - \underline{\omega}_h^*\|_2 \leq \mathcal{E}_1(\epsilon, K, D, H, \sigma) \tag{4}$$

We show the first case. Note that we have

$$\underline{Q}_h(s,a,b) = \max\{0, \phi^\top \underline{\omega}_h\} \geq \phi^\top \underline{\omega}_h,$$

which, together with Equation (4), imply

$$\begin{aligned}
\underline{\iota}_h(s,a,b) &= (\mathbb{B}_h \underline{V}_{h+1})(s,a,b) - \underline{Q}_h(s,a,b) \\
&\leq (\mathbb{B}_h \underline{V}_{h+1})(s,a,b) - \phi^\top \underline{\omega}_h \\
&\leq \mathcal{E}_1(\epsilon, K, D, H, \sigma) \ .
\end{aligned}$$

For the lower bound, consider two cases. First, if $\phi^\top \underline{\omega}_h \leq 0$, then we have

$$\begin{aligned}
\underline{\iota}_h(s,a,b) &= (\mathbb{B}_h \underline{V}_{h+1})(s,a,b) - \underline{Q}_h(s,a,b) \\
&= (\mathbb{B}_h \underline{V}_{h+1})(s,a,b) - 0 \\
&\geq 0.
\end{aligned}$$

On the other hand, if $\phi^\top \underline{\omega}_h \geq 0$, then we have

$$\begin{aligned}
\underline{\iota}_h(s,a,b) &= (\mathbb{B}_h \underline{V}_{h+1})(s,a,b) - \underline{Q}_h(s,a,b) \\
&= (\mathbb{B}_h \underline{V}_{h+1})(s,a,b) - \phi^\top \underline{\omega}_h \\
&\geq -\mathcal{E}_1(\epsilon, K, D, H, \sigma) \ .
\end{aligned}$$

□

Next, we consider the relationship between the estimated value function and the true ones based on the best responses.

Lemma 3. *If the bounds on the Bellman error given above hold, then, for any $s \in \mathcal{S}$, we have*

$$\underline{V}_h(s) \leq V_h^{\hat{\pi},*}(s) + \mathcal{E}_1 \ \ \textit{and} \ \ V_h^{*,\hat{\nu}}(s) - \mathcal{E}_1 \leq \overline{V}_h(s) \ ,$$

where we omit the dependence of $\mathcal{E}_1$ on the relevant variables for brevity.

Proof. We prove the left inequality. The right follows similar arguments. We use backward induction. For $h = H+1$, we have $\underline{V}_h(s) = V_h^{\hat{\pi},*}(s) = 0$. We assume the inequality hold for $h+1$ and prove it for h. We have

$$\begin{aligned}
V_h^{\hat{\pi},*}(s) - \underline{V}_h(s) &= \mathbb{E}_{a\sim\hat{\pi}, b\sim *}\left[Q_h^{\hat{\pi},*}(s,a,b)\right] - \mathbb{E}_{a\sim\hat{\pi}, b\sim\hat{\nu}}\left[\underline{Q}_h(s,a,b)\right] \\
&= \mathbb{E}_{a\sim\hat{\pi}, b\sim *}\left[Q_h^{\hat{\pi},*}(s,a,b) - \underline{Q}_h(s,a,b)\right] + \left(\mathbb{E}_{a\sim\hat{\pi}, b\sim *}\left[\underline{Q}_h(s,a,b)\right] - \mathbb{E}_{a\sim\hat{\pi}, b\sim\hat{\nu}}\left[\underline{Q}_h(s,a,b)\right]\right) \ .
\end{aligned}$$

First, note that

$$Q_h^{\hat{\pi},*}(s,a,b) - \underline{Q}_h(s,a,b) = \mathbb{B}_h\left(V_{h+1}^{\hat{\pi},*}(s,a,b) - \underline{V}_{h+1}(s,a,b)\right) + \underline{\iota}_h(s,a,b) \geq -\mathcal{E}_1 \ ,$$

where the second inequality follows from Lemma 2 and the induction assumption. Furthermore, by the NE value property, we have

$$\mathbb{E}_{a\sim\hat{\pi},b\sim *}\left[\underline{Q}_h(s,a,b)\right] - \mathbb{E}_{a\sim\hat{\pi},b\sim\hat{\nu}}\left[\underline{Q}_h(s,a,b)\right] \geq 0 \ .$$

Thus, we obtain $\underline{V}_h(s) \leq V_h^{\hat{\pi},*}(s) + \mathcal{E}_1$. □

Next, we prove a result that gives us an upper bound on σ^2 in terms of the reward variance and H.

Lemma 4. *We have* $Var(\xi_h^\tau | s_h^\tau, a_h^\tau, b_h^\tau) = \sigma^2 \leq (H+\gamma)^2$.

Proof. We consider $\underline{V}_{h+1}$. The case for $\overline{V}_{h+1}$ is similar. We have

$$\begin{aligned}
Var(\xi_h^\tau | s_h^\tau, a_h^\tau, b_h^\tau) &= \mathbb{E}\left[(r_h^\tau + \underline{V}_{h+1}(s_{h+1}^\tau) - (\mathbb{B}_h\underline{V}_{h+1})(s_h^\tau, a_h^\tau, b_h^\tau))^2 | s_h^\tau, a_h^\tau, b_h^\tau\right] \\
&= \mathbb{E}\left[(r_h^\tau + \underline{V}_{h+1}(s_{h+1}^\tau) - \mathbb{E}[r_h^\tau + \underline{V}_{h+1}(s_{h+1}^\tau) | s_h^\tau, a_h^\tau, b_h^\tau])^2 | s_h^\tau, a_h^\tau, b_h^\tau\right] \\
&= \mathbb{E}\left[(r_h^\tau - \mathbb{E}[r_h^\tau | s_h^\tau, a_h^\tau, b_h^\tau])^2 | s_h^\tau, a_h^\tau, b_h^\tau\right] + \mathbb{E}\left[(\underline{V}_{h+1}(s_{h+1}^\tau) - \mathbb{E}[\underline{V}_{h+1}(s_{h+1}^\tau) | s_h^\tau, a_h^\tau, b_h^\tau])^2 | s_h^\tau, a_h^\tau, b_h^\tau\right] \\
&\quad + 2\mathbb{E}\left[(r_h^\tau - \mathbb{E}[r_h^\tau | s_h^\tau, a_h^\tau, b_h^\tau])(\underline{V}_{h+1}(s_{h+1}^\tau) - \mathbb{E}[\underline{V}_{h+1}(s_{h+1}^\tau) | s_h^\tau, a_h^\tau, b_h^\tau])\right] \\
&= Var(r_h^\tau | s_h^\tau, a_h^\tau, b_h^\tau) + Var(\underline{V}_{h+1}(s_{h+1}^\tau) | s_h^\tau, a_h^\tau, b_h^\tau) \\
&\quad + 2\sqrt{\mathbb{E}\left[(r_h^\tau - \mathbb{E}[r_h^\tau | s_h^\tau, a_h^\tau, b_h^\tau])^2\right]\mathbb{E}\left[(\underline{V}_{h+1}(s_{h+1}^\tau) - \mathbb{E}[\underline{V}_{h+1}(s_{h+1}^\tau) | s_h^\tau, a_h^\tau, b_h^\tau])^2\right]} \\
&\leq Var(r_h^\tau | s_h^\tau, a_h^\tau, b_h^\tau) + Var(\underline{V}_{h+1}(s_{h+1}^\tau) | s_h^\tau, a_h^\tau, b_h^\tau) \\
&\quad + 2\sqrt{Var(r_h^\tau | s_h^\tau, a_h^\tau, b_h^\tau) Var(\underline{V}_{h+1}(s_{h+1}^\tau) | s_h^\tau, a_h^\tau, b_h^\tau)} \\
&= \left(\sqrt{Var(r_h^\tau | s_h^\tau, a_h^\tau, b_h^\tau)} + \sqrt{Var(\underline{V}_{h+1}(s_{h+1}^\tau) | s_h^\tau, a_h^\tau, b_h^\tau)}\right)^2 \leq (\gamma + H)^2 \ ,
\end{aligned}$$

where the fourth equality uses Cauchy-Schwarz and the last one uses the fact that $0 \leq \underline{V}_{h+1}(s) \leq H$, for all $h \in [H]$, $s \in \mathcal{S}$. □

Next, we state the following well-known result which will help us express the suboptimality gap in terms of quantities that we can control. For a proof, see (Cai et al., 2020).

Lemma 5. *(Value difference lemma) Given an MG* $(\mathcal{S}, \mathcal{A}, \mathcal{B}, r, H)$*, let* $\hat{\pi} \otimes \hat{\nu} = \{\hat{\pi}_h \otimes \hat{\nu}_h : \mathcal{S} \to \Delta(\mathcal{A}) \times \Delta(\mathcal{B})\}_{h\in[H]}$ *be a product policy,* (π, ν) *be a policy pair, and* $(\widehat{Q}_h)_{h\in[H]}$ *be any estimated Q-functions. For any* $h \in [H]$ *we define the estimated value function* $\widehat{V}_h : \mathcal{S} \to \mathbb{R}$ *by setting* $\widehat{V}_h(s) = \langle \widehat{Q}_h(s,\cdot,\cdot), \hat{\pi}_h(\cdot|s) \otimes \hat{\nu}(\cdot|s)\rangle$*, for all* $s \in \mathcal{S}$*. Then, for all* $s \in \mathcal{S}$*, we have*

$$\begin{aligned}
\widehat{V}_1(s) - V_1^{\pi,\nu}(s) &= \sum_{h=1}^H \mathbb{E}_{\pi,\nu}\left[\langle \widehat{Q}_h(s,\cdot,\cdot), \hat{\pi}_h(\cdot|s) \otimes \hat{\nu}(\cdot|s) - \pi(\cdot|s) \otimes \nu(\cdot|s)\rangle | s_1 = s\right] \\
&\quad + \sum_{h=1}^H \mathbb{E}_{\pi,\nu}\left[\widehat{Q}_h(s_h,a_h,b_h) - (\mathbb{B}_h V_{h+1})(s_h,a_h,b_h) | s_1 = s\right] \ .
\end{aligned}$$

Now, the first term in the lemma above can be bounded by 0 (see Lemma A.3 of (Zhong et al., 2022)). Thus, we obtain

$$\widehat{V}_1(s) - V_1^{\pi,\nu}(s) \leq \sum_{h=1}^H \mathbb{E}_{\pi,\nu}\left[\widehat{Q}_h(s_h,a_h,b_h) - (\mathbb{B}_h V_{h+1})(s_h,a_h,b_h) | s_1 = s\right] \ . \tag{5}$$

Now we write the suboptimality gap as:

$$\text{SubOpt}(\widehat{\pi}, \widehat{\nu}, s) = V_1^{*,\widehat{\nu}}(s) - V_1^{\widehat{\pi},*}(s) = \left(V_1^{*,\widehat{\nu}}(s) - V_1^{*}(s)\right) + \left(V_1^{*}(s) - V_1^{\widehat{\pi},*}(s)\right) \ . \tag{6}$$

For the first term, we have

$$V_1^{*,\widehat{\nu}}(s) - V_1^{*}(s) \leq \overline{V}_1(s) - V_1^{*}(s) + \mathcal{E}_1 \leq \overline{V}_1(s) - V_1^{\pi',\nu^*}(s) + \mathcal{E}_1 \ ,$$

for some π', where the first inequality follows from Lemma 3 and the second inequality follows from the fact that (π^*, ν^*) is an NE policy. Similarly, we have that $V_1^{*}(s) - V_1^{\widehat{\pi},*}(s) \leq V_1^{\pi^*,\nu'}(s) - \underline{V}_1(s) + \mathcal{E}_1$, for some ν'. Then, Equation (5), Lemma 2 and Lemma 5 imply

$$\begin{aligned}
V_1^{*,\widehat{\nu}}(s) - V_1^{*}(s) &\leq \overline{V}_1(s) - V_1^{\pi',\widehat{\nu}}(s) + \mathcal{E}_1(\epsilon, K, H, d, \sigma) \\
&\leq \sum_{h=1}^{H} \mathbb{E}_{\pi',\nu^*}\left[-\overline{\iota}_h(s_h, a_h, b_h)\right] + H\mathcal{E}_1(\epsilon, K, H, d, \sigma) \\
&\leq 2H\mathbb{E}_{\pi',\nu^*}\left[\mathcal{E}_1(\epsilon, K, H, d, \sigma) | s_1 = s\right] \\
&\leq 2Hc_1(\delta) \cdot \left(\sqrt{\frac{\sigma^2 poly(d)}{\kappa^2 K}} + \frac{\sigma}{\kappa}\epsilon\right) \\
&\leq 2Hc_1(\delta) \cdot \left(\sqrt{\frac{(H+\gamma) poly(d)}{\kappa^2 K}} + \frac{H+\gamma}{\kappa}\epsilon\right) \ ,
\end{aligned}$$

where the last inequality follows from Lemma 4. Similarly, we have

$$V_1^{*}(s) - V_1^{\widehat{\pi},*}(s) \leq 2Hc_1(\delta) \cdot \left(\sqrt{\frac{(H+\gamma) poly(d)}{\kappa^2 K}} + \frac{H+\gamma}{\kappa}\epsilon\right) \ .$$

Finally, we obtain

$$\text{SubOpt}(\widehat{\pi}, \widehat{\nu}, s) \leq O\left(\sqrt{\frac{H(H+\gamma)^2 poly(d)}{\kappa^2 K}} + \frac{H(H+\gamma)}{\kappa}\epsilon\right) \ .$$

□

A.3 Proof of Theorem 3

Given clean dataset $\widetilde{D} = \{(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_K, \tilde{y}_K)\}$, consider the observation model

$$\tilde{y}_i = \langle \omega^*, \tilde{x}_i \rangle + \xi_i \ ,$$

where ξ_i are γ^2-sub-Gaussian independent noise variables and ω^* is the true regressor with $\|\omega^*\| \leq R$, for some $R < \infty$, and $\|\tilde{x}_i\|_2 \leq 1$, for $i \in [K]$. Furthermore, given $\epsilon < 1/2$, assume that ϵ-fraction of the outcomes in $\widetilde{D}$ are corrupted and let $D = \{(x_1, y_1), \ldots, (x_K, y_K)\}$ denote the corrupted dataset. Formally, we assume that the covariates remain clean, that is, $x_i = \tilde{x}_i$, for $i \in [K]$, and that, for any $i \in [K]$, a coin is flipped with success rate ϵ to determine whether $\tilde{y}_i$ is corrupted into y_i or not.

Furthermore, let

$$\Sigma = \frac{1}{K}\sum_{k=1}^{K} x_k x_k^T$$

denote the covariance matrix of D. The following result (Chen et al., 2022a) provides error norm bounds of the regressor estimated using the SCRAM method on the corrupted dataset D.

Theorem 7. *Let $0 < \epsilon < 0.5$ be an upper bound on the contamination level, and suppose K satisfies $K \geq O(\log(\min\{K,d\})/\epsilon)$. Then, given $\delta \in (0,1)$, there exists a poly(K,d) algorithm which takes as input the dataset D and, with probability at least $1-\delta$, outputs a vector ω that satisfies*

$$\|\omega^* - \omega\|_\Sigma \leq O\left(\epsilon\gamma\log(1/\epsilon) + \min\left\{\gamma\sqrt{\frac{d+\log(1/\delta)}{K}}, (R\gamma)^{1/2}\sqrt[4]{\frac{\log(1/\delta)}{K}}\right\}\right) .$$

For every $h \in [H]$, we define the datasets

$$\widetilde{D}_{min}(h) = \{\underbrace{\phi(\tilde{s}_h^\tau, \tilde{a}_h^\tau, \tilde{b}_h^\tau)}_{\text{covariates}}, \underbrace{\tilde{r}_h^\tau + \underline{V}_{h+1}(\tilde{s}_{h+1}^\tau)}_{\text{clean obs.}}\}_{\tau=1}^K$$

and

$$\widetilde{D}_{max}(h) = \{\phi(\tilde{s}_h^\tau, \tilde{a}_h^\tau, \tilde{b}_h^\tau), \tilde{r}_h^\tau + \overline{V}_{h+1}(\tilde{s}_{h+1}^\tau))\}_{\tau=1}^K .$$

Similarly, the partitions of the corrupted data are defined as

$$D_{min}(h) = \{\underbrace{\phi(s_h^\tau, a_h^\tau, b_h^\tau)}_{\text{covariates}}, \underbrace{r_h^\tau + \underline{V}_{h+1}(s_{h+1}^\tau)}_{\epsilon\text{-corrupted obs.}}\}_{\tau=1}^K$$

and

$$D_{max}(h) = \{\phi(s_h^\tau, a_h^\tau, b_h^\tau), r_h^\tau + \overline{V}_{h+1}(s_{h+1}^\tau))\}_{\tau=1}^K .$$

Note that we assume $\phi(\tilde{s}_h^\tau, \tilde{a}_h^\tau, \tilde{b}_h^\tau) = \phi(s_h^\tau, a_h^\tau, b_h^\tau)$, for all $\tau \in [K]$. Assumption 2 implies that there exist weight vectors $\underline{\omega}_h^*, \overline{\omega}_h^* \in \mathbb{R}^d$ such that we have

$$\phi(s_h^\tau, a_h^\tau, b_h^\tau)^\top \underline{\omega}_h^* + \xi_h^\tau = (\mathbb{B}_h \underline{V}_{h+1})(s_h^\tau, a_h^\tau, b_h^\tau) + \xi_h^\tau = \left(\tilde{r}_h^\tau + \underline{V}_{h+1}(\tilde{s}_{h+1}^\tau)\right) \tag{7}$$

and

$$\phi(s_h^\tau, a_h^\tau, b_h^\tau)^\top \overline{\omega}_h^* + \xi_h^\tau = (\mathbb{B}_h \overline{V}_{h+1})(s_h^\tau, a_h^\tau, b_h^\tau) + \xi_h^\tau = \left(\tilde{r}_h^\tau + \overline{V}_{h+1}(\tilde{s}_{h+1}^\tau)\right) , \tag{8}$$

where ξ_h^τ are zero-mean γ^2-subGaussian random variables.

Now, let us define the covariance matrices, for all $h \in [H]$, as:

$$\Sigma_h = \frac{1}{K}\sum_{\tau=1}^K \phi(s_h^\tau, a_h^\tau, b_h^\tau)\phi(s_h^\tau, a_h^\tau, b_h^\tau)^\top .$$

Note that Σ_h depend on the triples $(s_h^\tau, a_h^\tau, b_h^\tau)$, $\tau \in [K]$, which are not changed under corruption. Thus, the covariance matrices are clean. However, the observations on both $\widetilde{D}_{min}(h)$ and $\widetilde{D}_{max}(h)$, for all $h \in [H]$, are ϵ-corrupted. Then, Theorem 7 implies the following result.

Corollary 2. *Under the conditions of Theorem 7, there exist a poly(K,d) algorithm that returns a sequence of estimators $(\underline{\omega}_h, \overline{\omega}_h)_{h=1}^H$ such that, given $\delta > 0$, the following inequalities are satisfied, for all $h \in [H]$, with probability at least $1 - \delta/2$:*

$$\|\underline{\omega}_h^* - \underline{\omega}_h\|_{\Sigma_h} \leq O\left(\gamma\epsilon\log(1/\epsilon) + \min\left\{\gamma\sqrt{\frac{d+\log(8H/\delta)}{K}}, \sqrt{H}\gamma\sqrt[4]{\frac{d\log(8H/\delta)}{K}}\right\}\right) , \tag{9}$$

$$\|\overline{\omega}_h^* - \overline{\omega}_h\|_{\Sigma_h} \leq O\left(\gamma\epsilon\log(1/\epsilon) + \min\left\{\gamma\sqrt{\frac{d+\log(8H/\delta)}{K}}, \sqrt{H}\gamma\sqrt[4]{\frac{d\log(8H/\delta)}{K}}\right\}\right) . \tag{10}$$

We will denote the right-hand side bound on the errors of norms by $\mathcal{E}(\epsilon, K, H, d)$, as shorthand notation.

Proof. The result follows immediately from Theorem 7 by applying the union bound over $2H$ events, and also by noting that $\|\underline{\omega}_h^*\|, \|\overline{\omega}_h^*\| \leq H\sqrt{d}$, for all $h \in [H]$, by Lemma E.1 of (Zhong et al., 2022). □

Next, we prove the upper bounds on the Bellman error in terms of corruption level and bonus term. Given $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ and $h \in [H]$, recall that the bonus term is defined as

$$\Gamma_h(s,a,b) = \left(\sqrt{K}\mathcal{E}(\epsilon, K, H, d) + 2H\sqrt{d}\right) \|\phi(s,a,b)\|_{\Lambda_h^{-1}} \ ,$$

where

$$\Lambda_h = \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau, b_h^\tau)\phi(s_h^\tau, a_h^\tau, b_h^\tau)^\top + I \ .$$

Note that Λ_h is positive definite, and hence, invertible, since Σ_h is positive semi-definite.

From here on, let us use the following notation for ease of presentation. For a given (s,a,b) and $h \in [H]$, let

$$\phi := \phi(s,a,b), \ \text{ and } \ \phi_h := \phi(s_h, a_h, b_h) \ .$$

Lemma 6. *Given tuple (s,a,b), let $\underline{\iota}_h(s,a,b)$ and $\overline{\iota}_h(s,a,b)$ be defined as in Lemma 2. Then, with probability at least $1-\delta$, we have*

$$0 \leq \underline{\iota}_h(s,a,b) \leq 2\Gamma_h(s,a,b) \ , \tag{11}$$
$$0 \leq -\overline{\iota}_h(s,a,b) \leq 2\Gamma_h(s,a,b) \ , \tag{12}$$

for all $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$.

Proof. We will prove the first inequality, and the second will follow by symmetry of argument. Let $\underline{\omega}_h^*$ be defined as in Equation 7. First, note that Corollary 2 implies

$$\begin{aligned}
\|\underline{\omega}_h - \underline{\omega}_h^*\|_{\Lambda_h}^2 &= (\underline{\omega}_h - \underline{\omega}_h^*)^\top \Lambda_h (\underline{\omega}_h - \underline{\omega}_h^*) \\
&= (\underline{\omega}_h - \underline{\omega}_h^*)^\top (K\Sigma_h + I)(\underline{\omega}_h - \underline{\omega}_h^*) \\
&= K(\underline{\omega}_h - \underline{\omega}_h^*)^\top \Sigma_h (\underline{\omega}_h - \underline{\omega}_h^*) + (\underline{\omega}_h - \underline{\omega}_h^*)^\top I (\underline{\omega}_h - \underline{\omega}_h^*) \\
&\leq K \|\underline{\omega}_h - \underline{\omega}_h^*\|_{\Sigma_h}^2 + 4H^2 d \\
&\leq K\mathcal{E}(\epsilon, K, H, d)^2 + 4H^2 d
\end{aligned}$$

where the first inequality comes from the fact that $\|\underline{\omega}_h^*\|_2 \leq H\sqrt{d}$, from Lemma E.1 of (Zhong et al., 2022) and also $\|\underline{\omega}_h\|_2 \leq H\sqrt{d}$, by design of SCRAM (Chen et al., 2022a). Thus, taking the square roots of both sides, we obtain

$$\|\underline{\omega}_h - \underline{\omega}_h^*\|_{\Lambda_h} \leq \sqrt{K\mathcal{E}(\epsilon, K, H, d)^2 + 4H^2 d} \leq \sqrt{K}\mathcal{E}(\epsilon, K, H, d) + 2H\sqrt{d} \ . \tag{13}$$

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

$$|\phi^\top \underline{\omega}_h - (\mathbb{B}_h \underline{V}_{h+1})(s,a,b)| = |\phi^\top (\underline{\omega}_h - \underline{\omega}_h^*)| \leq \|\underline{\omega}_h - \underline{\omega}_h^*\|_{\Lambda_h} \|\phi\|_{\Lambda_h^{-1}} \leq \left(\sqrt{K}\mathcal{E}(\epsilon, K, H, d) + 2H\sqrt{d}\right) \|\phi\|_{\Lambda_h^{-1}} \ ,$$

where the first inequality follows from extended Cauchy-Schwarz.

Note that each of the bounds on the right-hand side of the last inequality holds with probability at least $1-\delta/2$, thus, by union bound, the inequality above holds with probability at least $1-\delta$. Therefore, we obtain

$$\phi^\top \underline{\omega}_h - \Gamma_h(s,a,b) \leq \mathbb{B}_h \underline{V}_{h+1} \leq H - h + 1 \ ,$$

where in the last inequality we have used the fact that $|r_h| \leq 1$ and $|\underline{V}_{h+1}(s)| \leq H - h$. Furthermore,

$$\underline{Q}_h(s,a,b) = \max\{0, \phi^\top \underline{\omega}_h - \Gamma_h(s,a,b)\} \geq \phi^\top \underline{\omega}_h - \Gamma_h(s,a,b) \ .$$

Thus, we have

$$\begin{aligned}\underline{\iota}_h(s,a,b) &= (\mathbb{B}_h\underline{V}_{h+1})(s,a,b) - \underline{Q}_h(s,a,b)\\ &\le (\mathbb{B}_h\underline{V}_{h+1})(s,a,b) - \phi^\top\underline{\omega}_h + \Gamma_h(s,a,b)\\ &\le 2\Gamma_h(s,a,b)\ .\end{aligned}$$

To prove the non-negativity of $\underline{\iota}_h(s,a,b)$, we consider two cases. First, if $\phi^\top\underline{\omega}_h - \Gamma_h(s,a,b) \ge 0$, then we have

$$\begin{aligned}\underline{\iota}_h(s,a,b) &= (\mathbb{B}_h\underline{V}_{h+1})(s,a,b) - \underline{Q}_h(s,a,b)\\ &= (\mathbb{B}_h\underline{V}_{h+1})(s,a,b) - 0 \ge 0\ ,\end{aligned}$$

since $\underline{V}_{h+1}(s) \in [0,H]$ and $r(s,a,b) \in [0,1]$. On the other hand, if $\phi^\top\underline{\omega}_h - \Gamma_h(s,a,b) \le 0$, we have

$$\begin{aligned}\underline{\iota}_h(s,a,b) &= (\mathbb{B}_h\underline{V}_{h+1})(s,a,b) - \underline{Q}_h(s,a,b)\\ &= (\mathbb{B}_h\underline{V}_{h+1})(s,a,b) - \phi^\top\underline{\omega}_h + \Gamma_h(s,a,b) \ge 0\ .\end{aligned}$$

□

Next, we give bounds on the best response values in terms of the estimated value functions.

Lemma 7. *If Equations (11) and (12) in Lemma 6 hold, then, for any $s \in \mathcal{S}$, we have*

$$\underline{V}_h(s) \le V_h^{\hat{\pi},*}(s), \quad \text{and} \quad V_h^{*,\hat{\nu}}(s) \le \overline{V}_h(s)\ .$$

Proof. This is an immediate implication of Lemma A.2 of (Zhong et al., 2022). □

The next result provides upper bounds on the expected values of the feature norms, for all the trajectories followed by the policy pairs in the unilateral set of the NE policy pair. This result is based on the Low Relative Uncertainty assumption.

Lemma 8. *Assume that Assumption 3 holds, that is, assume that there exists a constant $c_1 > 0$ such that, for all $h \in [H]$, we have*

$$\Lambda_h \succeq I + c_1 K \max\left\{\sup_\nu\ \mathbb{E}_{\pi^*,\nu}[\phi_h\phi_h^\top|s_1 = x],\ \sup_\pi\ \mathbb{E}_{\pi,\nu^*}[\phi_h\phi_h^\top|s_1 = x]\right\}\ .$$

Then, for all $h \in [H]$, with probability at least $1-\delta$, we have

$$\mathbb{E}_{\pi^*,\nu'}\left[\|\phi(s,a,b)\|_{\Lambda_h^{-1}}\right] + \mathbb{E}_{\pi',\nu^*}\left[\|\phi(s,a,b)\|_{\Lambda_h^{-1}}\right] \le 2\sqrt{\frac{d}{c_1 K}}\ .$$

Proof. We derive an upper bound for the first expectation. The argument is identical to the second one. Let

$$\overline{\Sigma}_h(x) = \mathbb{E}_{\pi^*,\nu'}\left[\phi(s_h,a_h,b_h)^\top\phi(s_h,a_h,b_h)|s_1 = x\right]\ . \tag{14}$$

First, note that, given two matrices $A, B \in \mathbb{R}^{n\times n}$ such that $A \preceq B$, then, for any $x \in \mathbb{R}^n_+$, we have $\|x\|_A \le \|x\|_B$. This follows immediately by observing that, since $A - B \succeq 0$, which means that $A - B$ is positive semi-definite, we obtain $x^T(A-B)x \ge 0$, which implies the desired result. Now, for any $h \in [H]$, we have

$$\begin{aligned}\mathbb{E}_{\pi^*,\nu'}\left[\|\phi(s_h,a_h,b_h)\|_{\Lambda_h^{-1}}\right] &= \mathbb{E}_{\pi^*,\nu'}\left[\sqrt{\phi_h^\top\Lambda_h^{-1}\phi_h}\right]\\ &\le \mathbb{E}_{\pi^*,\nu'}\left[\sqrt{\phi_h^\top\left(c_1K\overline{\Sigma}_h(x) + I\right)^{-1}\phi_h}\Big|s_1 = x\right]\\ &\le \sqrt{\mathbb{E}_{\pi^*,\nu'}\left[Tr\left(\left(c_1K\overline{\Sigma}_h(x) + I\right)^{-1}\phi_h\phi_h^\top\right)\Big|s_1 = x\right]}\\ &= \sqrt{Tr\left(\left(c_1K\overline{\Sigma}_h(x) + I\right)^{-1}\mathbb{E}_{\pi^*,\nu'}\left[\phi_h\phi_h^\top|s_1 = x\right]\right)}\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{Tr\left(\left(c_1 K \overline{\Sigma}_h(x) + I\right)^{-1} \overline{\Sigma}_h(x)\right)} \\
&= \sqrt{\frac{1}{c_1 K} Tr\left(\left(c_1 K \overline{\Sigma}_h(x) + I\right)^{-1} \left(\left(c_1 K \overline{\Sigma}_h(x) + I\right) - I\right)\right)} \\
&= \sqrt{\frac{1}{c_1 K}} \sqrt{Tr\left(I - \left(c_1 K \overline{\Sigma}_h(x) + I\right)^{-1}\right)} \\
&\leq \sqrt{\frac{d}{c_1 K}} \ ,
\end{aligned}
$$

where the first inequality follows by assumption; the second inequality follows from Jensen; the second equality follows from the fact that $Tr(\mathbb{E}X) = \mathbb{E}[Tr(X)]$, for a given random matrix X; for the fourth equality, we have used the observation that $Tr(I - A) \leq \dim(A)$, for a positive definite matrix A. □

Now we are ready to prove Theorem 3. We restate it below for convenience.

Statement. *Suppose that Assumption 3 holds with given constant c_1. Let $\delta > 0$, $\epsilon < 1/2$ and let D be the ϵ-corrupted version of the dataset $\widetilde{D}$ comprised of K trajectories of length H, where $K \geq \log(\min(K, d))/\epsilon$ and $(\widetilde{s}_h^\tau, \widetilde{a}_h^\tau, \widetilde{b}_h^\tau) = (s_h^\tau, a_h^\tau, b_h^\tau)$, for all $\tau \in [K]$, $h \in [H]$. Then, with probability at least $1 - \delta$, SCRAM-PMVI outputs $(\widehat{\pi}, \widehat{\nu})$ that satisfy, for every $s \in \mathcal{S}$:*

$$
SubOpt(\widehat{\pi}, \widehat{\nu}, s) \leq \widetilde{O}\left(\frac{1}{\sqrt{c_1}}(\gamma + H)H\sqrt{d}\epsilon + \frac{H^2 d}{\sqrt{c_1 K}}\right) \ .
$$

We decompose the suboptimality gap as in Equation (6):

$$
\text{SubOpt}(\widehat{\pi}, \widehat{\nu}, s) = V_1^{*,\widehat{\nu}}(s) - V_1^{\widehat{\pi},*}(s) = \left(V_1^{*,\widehat{\nu}}(s) - V_1^*(s)\right) + \left(V_1^*(s) - V_1^{\widehat{\pi},*}(s)\right) \ .
$$

First, we bound the left difference.

$$
V_1^{*,\widehat{\nu}}(s) - V_1^*(s) \leq \overline{V}_1(s) - V_1^*(s) \leq \overline{V}_1(s) - V_1^{\pi',\nu^*}(s) \ ,
$$

for some π', where the first inequality follows from Lemma 7 and the second inequality follows from the fact that (π^*, ν^*) is an NE policy. Similarly, we have that $V_1^*(s) - V_1^{\widehat{\pi},*}(s) \leq V_1^{\pi^*,\nu'}(s) - \underline{V}_1(s)$, for some ν'. Then, similar to the proof of Theorem 2, we have

$$
\begin{aligned}
V_1^{*,\widehat{\nu}}(s) - V_1^*(s) &\leq \overline{V}_1(s) - V_1^{\pi',\widehat{\nu}}(s) \\
&\leq \sum_{h=1}^{H} \mathbb{E}_{\pi',\nu^*}\left[-\bar{\iota}_h(s_h, a_h, b_h)\right] \\
&\leq 2\sum_{h=1}^{H} \mathbb{E}_{\pi',\nu^*}\left[\Gamma_h(s_h, a_h, b_h) | s_1 = s\right] \\
&\leq 2\left(\sqrt{K}\mathcal{E}(\epsilon, K, H, d) + 2H\sqrt{d}\right) \sum_{h=1}^{H} \mathbb{E}_{\pi',\nu^*}\left[\|\phi(s, a, b)\|_{\Lambda_h^{-1}}\right]
\end{aligned}
$$

Similarly, we have

$$
V_1^*(s) - V_1^{\widehat{\pi},*}(s) \leq 2\left(\sqrt{K}\mathcal{E}(\epsilon, K, H, d) + 2H\sqrt{d}\right) \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\nu'}\left[\|\phi(s, a, b)\|_{\Lambda_h^{-1}}\right] \ .
$$

Finally, applying Lemma 8, we obtain

$$
\text{SubOpt}(\widehat{\pi}, \widehat{\nu}, s) \leq 2\left(\sqrt{K}\mathcal{E}(\epsilon, K, H, d) + 2H\sqrt{d}\right) \sum_{h=1}^{H} \left(\mathbb{E}_{\pi',\nu^*}\left[\|\phi(s, a, b)\|_{\Lambda_h^{-1}}\right] + \mathbb{E}_{\pi^*,\nu'}\left[\|\phi(s, a, b)\|_{\Lambda_h^{-1}}\right]\right)
$$

$$
\begin{aligned}
&\leq 4\left(\sqrt{K}\mathcal{E}(\epsilon,K,H,d)+2H\sqrt{d}\right)\sum_{h=1}^{H}\sqrt{\frac{d}{c_1K}}\\
&=O\left(\frac{\gamma}{c_1}H\sqrt{d}\epsilon+H^2K^{-1/2}d\right)\\
&\leq O\left(\frac{1}{c_1}(\gamma+H)H\sqrt{d}\epsilon+\frac{H^2d}{c_1\sqrt{K}}\right)\ ,
\end{aligned}
$$

where the last inequality follows from Lemma 4.

B Proofs of Section 5

In this section, we derive the proofs of the results in Section 5.

B.1 Proof of Theorem 4

For this section, we prove a similar result which gives us general bounds when no coverage is guaranteed on the corrupted dataset, but the underlying clean data has LRU coverage.

First, since we assume corrupted covariates, we rewrite the overall covariance as

$$
\Lambda_h=\widetilde{\Lambda}_h+\widehat{\Lambda}_h=\sum_{\text{clean }\tau}\widetilde{\phi}_h^\tau\left(\widetilde{\phi}_h^\tau\right)^\top+\sum_{\text{corrupted }\tau}\widehat{\phi}_h^\tau\left(\widehat{\phi}_h^\tau\right)^\top+I=(1-\epsilon)\left(K\widetilde{\Sigma}_h+I\right)+\epsilon\left(K\widehat{\Sigma}_h+I\right) \tag{15}
$$

We start by deriving upper bounds on the Bellman error.

Lemma 9. *With probability at least $1-\delta$, we have*

$$
\begin{aligned}
0\leq \underline{\iota}_h(s,a,b)\leq 2\widehat{\Gamma}_h(s,a,b)\ ,\\
0\leq -\overline{\iota}_h(s,a,b)\leq 2\widehat{\Gamma}_h(s,a,b)\ ,
\end{aligned}
$$

where

$$
\widehat{\Gamma}_h(s,a,b)=\left(\sqrt{(1-\epsilon)K}\mathcal{E}+(\sqrt{\epsilon K}+2)H\sqrt{d}\right)\|\phi(s,a,b)\|_{\Lambda_h^{-1}}
$$

Proof. Let us consider the first part. The second part will follow by a similar argument. Similar to the argument of Lemma 6, we have

$$
\begin{aligned}
\|\underline{\omega}_h-\underline{\omega}_h^*\|_{\Lambda_h}^2&=(\underline{\omega}_h-\underline{\omega}_h^*)^\top\Lambda_h(\underline{\omega}_h-\underline{\omega}_h^*)\\
&=(\underline{\omega}_h-\underline{\omega}_h^*)^\top\left((1-\epsilon)\left(K\widetilde{\Sigma}_h+I\right)+\epsilon\left(K\widehat{\Sigma}_h+I\right)\right)(\underline{\omega}_h-\underline{\omega}_h^*)\\
&=(1-\epsilon)K\|\underline{\omega}_h-\underline{\omega}_h^*\|_{\widetilde{\Sigma}_h}^2+\epsilon K\|\underline{\omega}_h-\underline{\omega}_h^*\|_{\widehat{\Sigma}_h}^2+H^2d\\
&\leq(1-\epsilon)K\mathcal{E}^2+\epsilon K\left\|\widehat{\Sigma}_h\right\|_2H^2d+4H^2d\\
&\leq(1-\epsilon)K\mathcal{E}^2+\epsilon KH^2d+4H^2d\ .
\end{aligned}
$$

Thus, we obtain

$$
\|\underline{\omega}_h-\underline{\omega}_h^*\|_{\Lambda_h}\leq\sqrt{(1-\epsilon)K}\mathcal{E}+(\sqrt{\epsilon K}+2)H\sqrt{d}\ .
$$

Then, similar to Lemma 6, with probability at least $1-\delta$, we have

$$
|\phi^\top\underline{\omega}_h-(\mathbb{B}_h\underline{V}_{h+1})(s,a,b)|=|\phi^\top(\underline{\omega}_h-\underline{\omega}_h^*)|\leq\|\underline{\omega}_h-\underline{\omega}_h^*\|_{\Lambda_h}\|\phi\|_{\Lambda_h^{-1}}\leq\left(\sqrt{(1-\epsilon)K}\mathcal{E}+(\sqrt{\epsilon K}+2)H\sqrt{d}\right)\|\phi\|_{\Lambda_h^{-1}}\ .
$$

The rest of the proof follows similar arguments to Lemma 6. □

Next, we derive upper bounds on the additional error term coming from the damage on coverage.

Lemma 10. *Assume that the condition of Lemma 8 holds only for the underlying clean data, but not for the corrupted dataset. Then, we have*

$$\mathbb{E}_{\pi^*,\nu'}\left[\|\phi(s,a,b)\|_{\Lambda_h^{-1}}\right] + \mathbb{E}_{\pi',\nu^*}\left[\|\phi(s,a,b)\|_{\Lambda_h^{-1}}\right] \leq 2\left(\sqrt{\frac{d}{c_1 K}} + \sqrt{\frac{d\epsilon}{c_1 K(1-\epsilon)}}\right) .$$

Proof. First, using the definition of Λ_h, we can equivalently write $\Lambda_h = S_h - \Delta_h$, where

$$S_h = \sum_{\tau=1}^{K} \widetilde{\phi}_h^\tau \left(\widetilde{\phi}_h^\tau\right)^\top + I \tag{16}$$

is the clean sample covariance matrix coming from $\widetilde{D}$, and Δ_h is defined as

$$\Delta_h = \sum_{\text{corrupted } \tau} \tilde{\phi}_h^\tau \tilde{\phi}_h^{\tau\top} - \phi_h^\tau \phi_h^{\tau\top} . \tag{17}$$

Since Δ_h is a symmetric matrix, we can write $\Delta_h = U_h E_h U_h^\top$, where E_h is a diagonal matrix composed of the eigenvalues of Δ_h. Note that the absolute value of each entry of E_h is at most ϵK. We will use the following formula for the inverse of a sum of two matrices (Henderson and Searle, 1981):

$$(A+UBV)^{-1} = A^{-1} - A^{-1}U(I+BVA^{-1}U)^{-1}BVA^{-1} ,$$

provided that $I + BVA^{-1}U$ is non-singular. In contrast to the Woodbury matrix identity, the matrix B here can be non-invertible. Using $A = S_h$ and $UBV = -U_h E_h U_h^\top$ we get the following identity.

$$\Lambda_h^{-1} = (S_h - \Delta_h)^{-1} = \left(S_h - U_h E_h U_h^\top\right)^{-1} = S_h^{-1} + S_h^{-1} \underbrace{U_h \left(I - E_h U_h^\top S_h^{-1} U_h\right)^{-1} E_h U_h^\top}_{:=M_h} S_h^{-1} . \tag{18}$$

This identity gives us the following bound on the bonus term with respect to π^* and ν'.

$$\mathbb{E}_{\pi^*,\nu'}\left[\|\phi(s_h,a_h,b_h)\|_{\Lambda_h^{-1}}\right] = \mathbb{E}_{\pi^*,\nu'}\left[\sqrt{\phi_h^\top \Lambda_h^{-1}\phi_h}\right] \leq \mathbb{E}_{\pi^*,\nu'}\left[\sqrt{\phi_h^\top S_h^{-1}\phi_h}\right] + \mathbb{E}_{\pi^*,\nu'}\left[\sqrt{|\phi_h^\top S_h^{-1} M_h S_h^{-1}\phi_h|}\right] \tag{19}$$

The first term can be bounded by $\sqrt{\frac{d}{c_1 K}}$ by Lemma 8, since it depends on the clean covariance matrix. For the second term, we first upper bound the maximum eigenvalue of the inverse of the middle expression in M_h. For that, we make use of the following Lemma.

Lemma 11. *Given $\Sigma_h(x)$ as defined in Equation (14), S_h as defined in Equation (16), $\Delta_h = U_h S_h U_h^\top$ as defined in Equation (17), and M_h as in Equation (18), we have*

$$Tr\left(S_h^{-1} M_h S_h^{-1} \Sigma_h(x)\right) \leq \frac{d\epsilon}{c_1 K(1-\epsilon)} .$$

Proof. First, note that, since U_h is an orthonormal matrix, we can write

$$E_h U_h^\top S_h^{-1} U_h = E_h U_h^{-1} S_h^{-1} U_h = E_h \left(S_h U_h\right)^{-1} U_h .$$

Next, observe that

$$\begin{aligned}
\|S_h U_h\|_2 &= \max_{x\neq 0} \frac{\left\|U_h x + \sum_{\tau=1}^K \widetilde{\phi}_h^\tau (\widetilde{\phi}_h^\tau)^\top U_h x\right\|_2}{\|x\|_2} \\
&\leq \max_{x\neq 0} \frac{\|U_h x\|_2 + \left\|\sum_{\tau=1}^K \widetilde{\phi}_h^\tau (\widetilde{\phi}_h^\tau)^\top\right\|_2 \|U_h x\|_2}{\|x\|_2}
\end{aligned}$$

$$\leq \max_{x \neq 0} \frac{\|x\|_2 + K\|x\|_2}{\|x\|_2}$$
$$= 1 + K \ ,$$

where the first inequality follows by the triangle inequality and matrix norm properties, and for the second inequality we have used the fact that the maximum eigenvalue of the clean covariance matrix does not exceed $1 + K$. This implies that

$$(S_h U_h)^{-1} \succcurlyeq \frac{1}{1+K} I_d$$

and, consequently,

$$I - E_h U_h^\top S_h^{-1} U_h \succcurlyeq \left(1 - \frac{\epsilon K}{1+K}\right) I_d \ , \tag{20}$$

since the minimum eigenvalue of $-E_h$ is at least $-\epsilon K$. Thus, we obtain

$$\left(I - E_h U_h^\top S_h^{-1} U_h\right)^{-1} \preceq \frac{1}{1 - \frac{\epsilon K}{1+K}} I_d \ \preceq \frac{1}{1-\epsilon} I_d \ .$$

Using the argument above, we have

$$\begin{aligned}
Tr\left(S_h^{-1} M_h S_h^{-1} \Sigma_h(x)\right) &= Tr\left(S_h^{-1} U_h \left(I - E_h U_h^\top S_h^{-1} U_h\right)^{-1} E_h U_h^\top S_h^{-1} U_h U_h^\top \Sigma_h(x)\right) \\
&= Tr\left(S_h^{-1} U_h \left(\underbrace{\left(I - E_h U_h^\top S_h^{-1} U_h\right)^{-1}}_{A} - I\right) U_h^\top \Sigma_h(x)\right) \\
&\leq \frac{1}{c_1 K} Tr\left(U_h \left(A - I\right) U_h^\top \left(c_1 K \Sigma_h(x) + I - I\right) \underbrace{\left(c_1 K \Sigma_h(x) + I\right)^{-1}}_{B}\right) \\
&= \frac{1}{c_1 K} Tr\left(U_h \left(A - I\right) U_h^\top \left(I - B\right)\right) \\
&\leq \frac{d}{c_1 K} \left\|U_h \left(A - I\right) U_h^\top \left(I - B\right)\right\|_2 \\
&\leq \frac{d}{c_1 K} \|A - I\|_2 \|I - B\|_2 \\
&\leq \frac{d}{c_1 K} \left(\frac{1}{1-\epsilon} - 1\right) \\
&= \frac{d\epsilon}{c_1 K (1-\epsilon)} \ ,
\end{aligned}$$

where the first inequality uses Assumption 3 and the fact that S_h^{-1} is symmetric; the second inequality uses the fact that $Tr(A) \leq d\|A\|_2$; the third inequality uses the property $\|AB\|_2 \leq \|A\|_2 \|B\|_2$, and the fact that, for orthonormal matrices, we have $\|U_h\|_2 \leq 1$. Thus, by putting everything together, we obtain the desired bounds. □

Now, note that

$$\mathbb{E}_{\pi^*, \nu'}\left[\sqrt{\phi^\top S_h^{-1} M_h S_h^{-1} \phi_h}\right] \leq \sqrt{Tr\left(S_h^{-1} M_h S_h^{-1} \Sigma_h(x)\right)} \leq \sqrt{\frac{d\epsilon}{c_1 K (1-\epsilon)}} \ ,$$

by Lemma 11. Thus, by putting everything together, we obtain

$$\mathbb{E}_{\pi^*, \nu'}\left[\|\phi(s_h, a_h, b_h)\|_{\Lambda_h^{-1}}\right] \leq \sqrt{\frac{d}{c_1 K}} + \sqrt{\frac{d\epsilon}{c_1 K (1-\epsilon)}} \ .$$

□

Now we are ready to prove the main result of this section. We restate the result for convenience.

Statement. *Suppose that the condition of Assumption 3 is satisfied only on the clean dataset $\widetilde{D}$, for a given constant c_1, and that an ϵ-fraction of tuples in D is arbitrarily corrupted. Furthermore, let $\delta > 0$ and $\epsilon \in (0, 1/2)$, and $K \geq \log(\min\{K, d\})/\epsilon$. Then, with probability at least $1-\delta$, $S-PMVI$ returns $(\widehat{\pi}, \widehat{\nu})$ that satisfy, for any $s \in \mathcal{S}$*

$$SubOpt(\widehat{\pi}, \widehat{\nu}, s) \leq O\left(\underbrace{\frac{1}{\sqrt{c_1 K}} H^2 d}_{\substack{\text{clean signal} \\ \text{\& clean covariates} \\ \text{\& LRU coverage}}} + \underbrace{\frac{1}{\sqrt{c_1}} H^2 \sqrt{d\epsilon}}_{\substack{\text{corrupted signal} \\ \text{\& clean covariates} \\ \text{\& LRU coverage}}} + \underbrace{\frac{1}{\sqrt{c_1}} H^2 d\sqrt{\epsilon}}_{\substack{\text{corrupted signal} \\ \text{\& corrupted covariates} \\ \text{\& LRU coverage}}} + \underbrace{\frac{1}{\sqrt{c_1(1-\epsilon)}} H^2 d^{3/2}\epsilon}_{\substack{\text{corrupted signal} \\ \text{\& corrupted covariates} \\ \text{\& corrupted coverage}}} \right)$$

Proof. As in the proof of Theorem 3, we have, for some policies π' and ν':

$$\begin{aligned}
\text{SubOpt}(\widehat{\pi}, \widehat{\nu}, s) &= V_1^{*,\widehat{\nu}}(s) - V_1^{\widehat{\pi},*}(s) = \left(V_1^{*,\widehat{\nu}}(s) - V_1^{*}(s)\right) + \left(V_1^{*}(s) - V_1^{\widehat{\pi},*}(s)\right) \\
&\leq \left(\overline{V}_1(s) - V_1^{\pi',\nu^*}(s)\right) + \left(V_1^{\pi^*,\nu'}(s) - \underline{V}_1(s)\right) \qquad (21) \\
&\leq \sum_{h=1}^{H} \left(\mathbb{E}_{\pi',\nu^*}\left[-\overline{\iota}_h(s_h, a_h, b_h)\right] + \mathbb{E}_{\pi^*,\nu'}\left[\underline{\iota}_h(s_h, a_h, b_h)\right] \right) \qquad (22) \\
&\leq 2\sum_{h=1}^{H} \left(\mathbb{E}_{\pi',\nu^*}\left[\widehat{\Gamma}_h(s,a,b)|s_1 = s\right] + \mathbb{E}_{\pi^*,\nu'}\left[\widehat{\Gamma}_h(s,a,b)|s_1 = s\right] \right) \qquad (23) \\
&\leq 2\left(\sqrt{(1-\epsilon)K\mathcal{E}} + (\sqrt{\epsilon K} + 2)H\sqrt{d}\right) \sum_{h=1}^{H} \left(\mathbb{E}_{\pi',\nu^*}\left[\|\phi(s,a,b)\|_{\Lambda_h^{-1}}\right] + \mathbb{E}_{\pi^*,\nu'}\left[\|\phi(s,a,b)\|_{\Lambda_h^{-1}}\right] \right) \\
&\leq 4H\left(\sqrt{(1-\epsilon)K\mathcal{E}} + \sqrt{\epsilon K d}H + 2H\sqrt{d}\right) \left(\sqrt{\frac{d}{c_1 K}} + \sqrt{\frac{d\epsilon}{c_1 K(1-\epsilon)}}\right) \qquad (24) \\
&\leq O\left(\underbrace{\frac{1}{\sqrt{c_1 K}} H^2 d}_{\substack{\text{clean signal} \\ \text{\& clean covariates} \\ \text{\& LRU coverage}}} + \underbrace{\frac{1}{\sqrt{c_1}} H^2 \sqrt{d\epsilon}}_{\substack{\text{corrupted signal} \\ \text{\& clean covariates} \\ \text{\& LRU coverage}}} + \underbrace{\frac{1}{\sqrt{c_1}} H^2 d\sqrt{\epsilon}}_{\substack{\text{corrupted signal} \\ \text{\& corrupted covariates} \\ \text{\& LRU coverage}}} + \underbrace{\frac{1}{\sqrt{c_1(1-\epsilon)}} H^2 d\epsilon}_{\substack{\text{corrupted signal} \\ \text{\& corrupted covariates} \\ \text{\& corrupted coverage}}} \right) \\
&\leq O\left(H^2 d^{3/2} K^{-1/2} + H^2 d^{3/2}\sqrt{\epsilon}\right) , \qquad (25)
\end{aligned}$$

where (21) follows from Lemma 3; (21) follows from Equation (5); (23) follows from Lemma 6; (24) follows from Lemma 10 and (25) follows from the fact that, for $\epsilon \in (0, 1/2)$, we have that $\epsilon/\sqrt{1-\epsilon} \leq \sqrt{\epsilon}$. □

B.2 Proof of Theorem 5

In this section, we improve the order of ϵ in Theorem 4 under an additional assumption on the feature space (Assumption 4. We start by proving upper bounds on the Bellman error in terms of the bonus term.

Lemma 12. *With probability at least $1-\delta$, we have*

$$\begin{aligned}
0 \leq \underline{\iota}_h(s,a,b) &\leq 2\widehat{\Gamma}_h(s,a,b) , \\
0 \leq -\overline{\iota}_h(s,a,b) &\leq 2\widehat{\Gamma}_h(s,a,b) ,
\end{aligned}$$

for all $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ and $h \in [H]$, where

$$\widehat{\Gamma}_h(s,a,b) = \left(2(1-\epsilon)K\mathcal{E} + \epsilon K H\sqrt{d} + H\sqrt{Kd}\right) \left\|\phi(s,a,b)^\top \Lambda_h^{-1}\right\|_2 .$$

Proof. First, we express the Bellman error of any given (s, a, b) tuple at time-step h as

$$\left|\overline{Q}_h(s,a,b) - (\mathbb{B}_h \overline{V}_{h+1})(s,a,b)\right| = \left|\phi_h^\top \left(\underline{\omega}_h - \underline{\omega}_h^*\right)\right| = \left|\phi_h^\top \Lambda_h^{-1} \Lambda_h \left(\underline{\omega}_h - \underline{\omega}_h^*\right)\right| \leq \left\|\phi_h^\top \Lambda_h^{-1}\right\|_2 \left\|\Lambda_h \left(\underline{\omega}_h - \underline{\omega}_h^*\right)\right\|_2 \ .$$

The error coming from the second player is similarly bounded. Note that we have

$$\begin{aligned}
&\left\|\Lambda_h \left(\underline{\omega}_h - \underline{\omega}_h^*\right)\right\|_2^2 = \left|\left(\underline{\omega}_h - \underline{\omega}_h^*\right)^\top \left(\widetilde{\Lambda}_h + \widehat{\Lambda}_h\right)^2 \left(\underline{\omega}_h - \underline{\omega}_h^*\right)\right| \\
&= \left|\left(\underline{\omega}_h - \underline{\omega}_h^*\right)^\top \left(\widetilde{\Lambda}_h^2 + \widetilde{\Lambda}_h \widehat{\Lambda}_h + \widehat{\Lambda}_h \widetilde{\Lambda}_h + \widehat{\Lambda}_h^2\right) \left(\underline{\omega}_h - \underline{\omega}_h^*\right)\right| \\
&\leq \underbrace{\left|\left(\underline{\omega}_h - \underline{\omega}_h^*\right)^\top \widetilde{\Lambda}_h^2 \left(\underline{\omega}_h - \underline{\omega}_h^*\right)\right|}_{P_1} + \underbrace{\left|\left(\underline{\omega}_h - \underline{\omega}_h^*\right)^\top \left(\widetilde{\Lambda}_h \widehat{\Lambda}_h + \widehat{\Lambda}_h \widetilde{\Lambda}_h\right) \left(\underline{\omega}_h - \underline{\omega}_h^*\right)\right|}_{P_2} + \underbrace{\left|\left(\underline{\omega}_h - \underline{\omega}_h^*\right)^\top \widehat{\Lambda}_h^2 \left(\underline{\omega}_h - \underline{\omega}_h^*\right)\right|}_{P_3}
\end{aligned}$$

We derive upper bounds for the three terms above separately. First, we have

$$\begin{aligned}
P_1 &= (1-\epsilon) \left|\left(\underline{\omega}_h - \underline{\omega}_h^*\right)^\top \left(K\widetilde{\Sigma}_h + I\right) \widetilde{\Lambda}_h \left(\underline{\omega}_h - \underline{\omega}_h^*\right)\right| \\
&\leq (1-\epsilon) \left\|\widetilde{\Lambda}_h\right\|_2 \left(K \left\|\underline{\omega}_h - \underline{\omega}_h^*\right\|_{\widetilde{\Sigma}_h}^2 + H^2 d\right) \\
&\leq (1-\epsilon)^2 K^2 \mathcal{E}^2 + (1-\epsilon)^2 K H^2 d \ ,
\end{aligned}$$

where the first equality follows from Equation (15), the first inequality follows from the fact that $\langle x, Ax\rangle \leq \|A\|_2 \|x\|_2^2$ and the last inequality follows from the fact that the maximum eigenvalue of $\widetilde{\Lambda}_h$ is at most $(1-\epsilon)K$, the fact that $\|\underline{\omega}_h^*\|_2 \leq H\sqrt{d}$ (Lemma E.1 of Zhong et al. (2022)) and Corollary 2, where we deliberately omit the dependencies of $\mathcal{E}$ for brevity.

Next, we have

$$\begin{aligned}
P_2 &\leq \left|\left(\underline{\omega}_h - \underline{\omega}_h^*\right)^\top \widetilde{\Lambda}_h \widehat{\Lambda}_h \left(\underline{\omega}_h - \underline{\omega}_h^*\right)\right| + \left|\left(\underline{\omega}_h - \underline{\omega}_h^*\right)^\top \widehat{\Lambda}_h \widetilde{\Lambda}_h \left(\underline{\omega}_h - \underline{\omega}_h^*\right)^\top\right| \\
&\leq 2 \left\|\widehat{\Lambda}_h\right\|_2 \left(\underline{\omega}_h - \underline{\omega}_h^*\right)^\top \widetilde{\Lambda}_h \left(\underline{\omega}_h - \underline{\omega}_h^*\right) \\
&\leq 2\epsilon K \left(K \left\|\underline{\omega}_h - \underline{\omega}_h^*\right\|_{\widetilde{\Sigma}_h}^2 + H^2 d\right) \\
&= 2\epsilon(1-\epsilon) K^2 \mathcal{E}^2 + 2\epsilon(1-\epsilon) K H^2 d \ ,
\end{aligned}$$

where we have used the fact that the maximum eigenvalue of $\widehat{\Lambda}_h$ is at most ϵK and applied similar arguments as for P_1. For the last term, we have

$$\begin{aligned}
P_3 &= \epsilon \left|\left(\underline{\omega}_h - \underline{\omega}_h^*\right)^\top \left(K\widehat{\Sigma}_h + I\right) \widehat{\Lambda}_h \left(\underline{\omega}_h - \underline{\omega}_h^*\right)\right| \\
&\leq \epsilon \left\|\widehat{\Lambda}_h\right\|_2 \left(K \left\|\underline{\omega}_h - \underline{\omega}_h^*\right\|_2^2 \left\|\widehat{\Sigma}_h\right\|_2 + H^2 d\right) \\
&\leq \epsilon^2 K \left(K H^2 d + H^2 d\right) \\
&= \epsilon^2 K^2 H^2 d + \epsilon^2 K H^2 d \ ,
\end{aligned}$$

by applying the same arguments as above. Putting everything together, we obtain

$$\begin{aligned}
\left\|\Lambda_h \left(\underline{\omega}_h - \underline{\omega}_h^*\right)\right\|_2^2 &\leq P_1 + P_2 + P_3 \\
&= (1-\epsilon)^2 K^2 \mathcal{E}^2 + (1-\epsilon)^2 K H^2 d + 2\epsilon(1-\epsilon) K^2 \mathcal{E}^2 + 2\epsilon(1-\epsilon) K H^2 d + \epsilon^2 K^2 H^2 d + \epsilon^2 K H^2 d \\
&= (1-\epsilon)^2 K^2 \mathcal{E}^2 + 2\epsilon(1-\epsilon) K^2 \mathcal{E}^2 + \epsilon^2 K^2 H^2 d + \left((1-\epsilon) H\sqrt{Kd} + \epsilon H \sqrt{Kd}\right)^2 \\
&\leq 3(1-\epsilon)^2 K^2 \mathcal{E}^2 + \epsilon^2 K^2 H^2 d + K H^2 d \ ,
\end{aligned}$$

where the last inequality follows from the fact that $\epsilon < 1/2$. Taking the square root of both sides and using the triangle inequality, we finally obtain

$$\left\|\Lambda_h \left(\underline{\omega}_h - \underline{\omega}_h^*\right)\right\|_2 \leq 2(1-\epsilon) K \mathcal{E} + \epsilon K H \sqrt{d} + H\sqrt{Kd} \ .$$

□

Next, we derive an upper bound on the expected value of the feature norms with respect to policy pairs lying in the LRU set.

Lemma 13. *Assume that Assumption 3 holds on the clean dataset $\widetilde{D}$ only, with given constant $c_1 > 0$. Then, for all $h \in [H]$, with probability at least $1-\delta$, we have*

$$\mathbb{E}_{\pi^*,\nu'}\left[\left\|\phi(s_h,a_h,b_h)^\top \Lambda_h^{-1}\right\|_2\right] + \mathbb{E}_{\pi',\nu^*}\left[\left\|\phi(s_h,a_h,b_h)^\top \Lambda_h^{-1}\right\|_2\right] \leq 2\left(\frac{d}{c_1c_2K} + \frac{d\epsilon}{c_1c_2K(1-\epsilon)}\right) ,$$

where $c_2 \leq \min_{\phi\in\Phi}\|\phi\|_2$.

Proof. Using the short-hand notation ϕ_h instead of $\phi(s_h,a_h,b_h)$, we have

$$\begin{aligned}
\mathbb{E}_{\pi^*,\nu'}\left[\left\|\phi_h^\top \Lambda_h^{-1}\right\|_2\right] &\leq \frac{1}{c_2}\mathbb{E}_{\pi^*,\nu'}\left[\sqrt{\phi_h^\top\left(S_h^{-1}+S_h^{-1}M_hS_h^{-1}\right)^2\phi_h\phi_h^\top\phi_h}\right] \\
&= \frac{1}{c_2}\mathbb{E}_{\pi^*,\nu'}\left[\sqrt{Tr\left(S_h^{-2}\left(I+M_hS_h^{-1}\right)^2\left(\phi_h\phi_h^\top\right)^2\right)}\right] \\
&\leq \frac{1}{c_2}\mathbb{E}_{\pi^*,\nu'}\left[Tr\left(S_h^{-1}\left(I+M_hS_h^{-1}\right)\phi_h\phi_h^\top\right)\right] \\
&= \frac{1}{c_2}Tr\left(S_h^{-1}\left(I+M_hS_h^{-1}\right)\left(\Sigma_h(x)\right)\right) \\
&\leq \frac{1}{c_2}\left(Tr\left(S_h^{-1}\Sigma_h(x)\right)+Tr\left(S_h^{-1}M_hS_h^{-1}\Sigma_h(x)\right)\right) .
\end{aligned}$$

The first factor is bounded as

$$Tr\left(S_h^{-1}\Sigma_h(x)\right) \leq \frac{d}{c_1K}$$

by Lemma 8, while the second term is bounded as

$$Tr\left(S_h^{-1}M_hS_h^{-1}\Sigma_h(x)\right) \leq \frac{d\epsilon}{c_1K(1-\epsilon)} ,$$

by Lemma 11. The result follows. □

Statement. *Suppose that the conditions of Theorem 4 and Assumption 4 hold. Then, with probability at least $1-\delta$, $S-PMVI$ returns $(\widehat{\pi},\widehat{\nu})$ that satisfy, for any $s \in \mathcal{S}$:*

$$SubOpt(\widehat{\pi},\widehat{\nu},s) \leq O\left(\frac{1}{c_1c_2}H^2d^{3/2}K^{-1/2} + \frac{1}{c_1c_2}H^2d^{3/2}\epsilon\right) .$$

Proof. Again, as in Theorem 3, we have

$$\begin{aligned}
&\text{SubOpt}(\widehat{\pi},\widehat{\nu},s) = V_1^{*,\widehat{\nu}}(s) - V_1^{\widehat{\pi},*}(s) = \left(V_1^{*,\widehat{\nu}}(s) - V_1^*(s)\right) + \left(V_1^*(s) - V_1^{\widehat{\pi},*}(s)\right) \\
&\leq \sum_{h=1}^H\left(\mathbb{E}_{\pi',\nu^*}\left[-\bar{\iota}_h(s_h,a_h,b_h)\right] + \mathbb{E}_{\pi^*,\nu'}\left[\underline{\iota}_h(s_h,a_h,b_h)\right]\right) \\
&\leq 2\sum_{h=1}^H\left(\mathbb{E}_{\pi',\nu^*}\left[\widehat{\Gamma}_h(s,a,b)|s_1=s\right] + \mathbb{E}_{\pi^*,\nu'}\left[\widehat{\Gamma}_h(s,a,b)|s_1=s\right]\right) \\
&\leq 2\left(2(1-\epsilon)K\mathcal{E} + \epsilon KH\sqrt{d} + H\sqrt{Kd}\right)\sum_{h=1}^H\left(\mathbb{E}_{\pi',\nu^*}\left[\left\|\phi(s,a,b)^\top\Lambda_h^{-1}\right\|_2\right] + \mathbb{E}_{\pi^*,\nu'}\left[\left\|\phi(s,a,b)^\top\Lambda_h^{-1}\right\|_2\right]\right) \qquad (26) \\
&\leq 4H\left(2(1-\epsilon)K\mathcal{E} + \epsilon KH\sqrt{d} + H\sqrt{Kd}\right)\sum_{h=1}^H\left(\frac{d}{c_1c_2K} + \frac{d\epsilon}{c_1c_2K(1-\epsilon)}\right) \qquad (27) \\
&\leq O\left(\frac{1}{c_1c_2}H^2d^{3/2}K^{-1/2} + \frac{1}{c_1c_2}H^2d^{3/2}\epsilon\right) ,
\end{aligned}$$

where (27) follows from Lemma 12, and (27) follows from Lemma 13 . □

C Proofs of Section 6

In this section, we provide the proofs of results related to coverage assumptions in Section 6.

C.1 Proof of Proposition 1

Before proving Proposition 1, we state a result that provides bounds on the concentration of covariance matrices. For proof see (Zanette et al., 2021a).

Lemma 14. *Let $\{\phi_i\}_{i\in[K]} \subset \mathbb{R}^d$ be i.i.d. samples from an underlying bounded distribution ν, with $\|\phi_i\|_2 \leq 1$, for $i \in [K]$ and covariance Σ. Define*

$$\Lambda = \sum_{k=1}^{K} \phi_i \phi_i^\top + \lambda I \ ,$$

where $\lambda \geq \Omega(d \log(K/\delta))$. Then, with probability at least $1-\delta$, we have

$$\frac{1}{3}(K\Sigma + \lambda I) \preceq \Lambda \preceq \frac{5}{3}(K\Sigma + \lambda I) \ .$$

We restate Proposition 1.

Statement. *Assume Φ has full rank and let $\delta \in (0,1)$. Then, if Assumption 6 holds, there exists a positive constant that depends on δ for which Assumption 3 holds, with probability at least $1-\delta$. Moreover, in the tabular MG setting, these two assumptions are equivalent.*

Proof. Fix $x \in \mathcal{S}$. Given policy pair (π, ν) and $h \in [H]$, let the matrix $D_h^{\pi,\nu} \in \mathbb{R}^{SAB \times SAB}$ denote the diagonal matrix composed of $d_h^{\pi,\nu}(s,a,b)$ diagonal entries, where we set $d_1^{\pi,\nu}(x,a,b) = 1$. Note that

$$\mathbb{E}_{\pi,\nu}\left[\phi(s_h,a_h,b_h)\phi(s_h,a_h,b_h)^\top | s_1 = x\right] = \Phi^\top D_h^{\pi,\nu} \Phi \ . \tag{28}$$

Further, let us denote by ρ the behavioral policy from which the clean tuples (s,a,b) from the given offline data are sampled and let us define the diagonal matrix $D_{\tau,h}^{\rho} \in \mathbb{R}^{SAB \times SAB}$, for every $\tau \in [K], h \in [H]$, with diagonal entries

$$\widetilde{D}_{\tau,h}^{\rho}[(s,a,b)] := \widetilde{d}_{\tau,h}^{\rho}(s,a,b) = \begin{cases} 1 & \text{if } (s_h^\tau, a_h^\tau, b_h^\tau) = (s,a,b) \\ 0 & \text{otherwise} \end{cases}$$

and let

$$\widehat{D}_h^{\rho} = \frac{1}{K}\sum_{\tau=1}^{K} \widetilde{D}_{\tau,h}^{\rho}$$

denote the sample covariance matrix. By assumption, there exists a finite positive constant c such that

$$\frac{d_h^{\pi,\nu}(s,a,b)}{d_h^{\rho}(s,a,b)} < c < \infty, \forall h \in [H], (\pi,\nu) \in \mathcal{U}(\pi^*,\nu^*), (s,a,b) \in \{(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B} : d_h^{\pi,\nu}(s,a,b) > 0\} \ ,$$

where $\mathcal{U}(\pi^*,\nu^*) = \{(\pi^*,\nu), (\pi,\nu^*), \forall\pi, \forall\nu\}$. This implies that

$$\min_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}} \left(d_h^{\rho}(s,a,b) - \frac{1}{c} d_h^{\pi,\nu}(s,a,b) \right) \geq 0 \ , \tag{29}$$

which implies that

$$D_h^{\rho} - \frac{1}{c} D_h^{\pi,\nu} \succeq 0 \ . \tag{30}$$

Now, Lemma 14 implies that there exist constants $\lambda_h \geq \Omega(d\log(KH/\delta))$, for $h \in [H]$, such that, with probability at least $1-\delta$, we have

$$\frac{1}{3}(KD_h^{\rho} + \lambda_h I) \preceq \sum_{k=1}^{K} \widetilde{D}_{k,h}^{\rho} + \lambda_h I \preceq \frac{5}{3}(KD_h^{\rho} + \lambda_h I) \ ,$$

for all $h \in [H]$. This implies that

$$\frac{1}{3}D_h^\rho - \frac{2\lambda_h}{3K}I \preceq \widehat{D}_h^\rho \preceq \frac{5}{3}D_h^\rho + \frac{2\lambda_h}{3K}I \ . \tag{31}$$

Equations (30) and (31) imply that, with probability at least $1-\delta$, we have

$$\widehat{D}_h^\rho - \left(\frac{1}{c}D_h^{\pi,\nu} - \frac{2\lambda_h}{3K}I\right) \succeq 0 \ .$$

Since the left-hand side is a diagonal matrix, the above is equivalent to

$$\min_{(s,a,b)} \left(\widehat{d}_h^\rho(s,a,b) - \left(\frac{1}{c} - \frac{2\lambda_h}{3Kd_h^{\pi,\nu}(s,a,b)}\right) d_h^{\pi,\nu}(s,a,b)\right) \geq 0 \ .$$

Now let us define constant c_1 as

$$c_1 = \max\{\frac{1}{c} - \frac{2d\log(KH/\delta)}{3Kd_h^{\pi,\nu}(s,a,b)} : d_h^{\pi,\nu}(s,a,b) > 0\} \ .$$

Then we obtain

$$\min_{(s,a,b)} \left(\widehat{d}_h^\rho(s,a,b) - c_1 d_h^{\pi,\nu}(s,a,b)\right) \geq 0,$$

which means that we have

$$\widehat{D}_h^\rho - c_1 D_h^{\pi,\nu} \succeq 0 \ .$$

Now, since Φ has full rank, its null space is 0. Thus, given non-zero $x \in \mathbb{R}^d$, we have $\Phi x \neq 0$ and thus

$$(\Phi x)^\top \left(\widehat{D}_h^\rho - c_1 D_h^{\pi,\nu}\right)(\Phi x) \geq 0 \ ,$$

since $\widehat{D}_h^\rho - c_1 D_h^{\pi,\nu} \succeq 0$. Therefore,

$$\Phi^\top \left(\widehat{D}_h^\rho - c_1 D_h^{\pi,\nu}\right)\Phi \succeq 0 \ ,$$

which implies

$$\sum_{\tau=1}^{K} \Phi^\top \widetilde{D}_{\tau,h}^\rho \Phi + I \succeq I + c_1 K \Phi^\top D_h^{\pi,\nu}\Phi \ .$$

Note that the above can be written as

$$\Lambda_h \succeq I + c_1 K \max\left\{\sup_\nu \mathbb{E}_{\pi^*,\nu}\left[\phi_h\phi_h^\top | s_1 = x\right], \sup_\nu \mathbb{E}_{\pi,\nu^*}\left[\phi_h\phi_h^\top | s_1 = x\right]\right\} \ .$$

Now assume the MG is tabular. Then Φ is just the identity matrix. Thus, if Assumption 3 holds, then there exists a positive constant c_1 such that, for any $h \in [H]$, $(\pi,\nu) \in \mathcal{U}(\pi^*,\nu^*)$, we have

$$\begin{aligned}
\Lambda_h \succeq I + c_1 K \Phi^\top D_h^{\pi,\nu}\Phi &\Rightarrow \Phi^\top \left(\widehat{D}_h^\rho - c_1 D_h^{\pi,nu}\right)\Phi \succeq 0 \\
&\Rightarrow \widehat{D}_h^\rho - c_1 D_h^{\pi,\nu} \succeq 0 \\
&\Rightarrow \frac{5}{3}D_h^\rho + \frac{2\lambda_h}{3K}I - c_1 D_h^{\pi,\nu} \succeq 0 \\
&\Rightarrow D_h^\rho + \frac{2\lambda_h}{5K}I - \frac{3}{5}c_1 D_h^{\pi,\nu} \succeq 0 \\
&\Rightarrow \min_{(s,a,b)} \left(d_h^\rho(s,a,b) - \left(\frac{3}{5}c_1 - \frac{2\lambda_h}{3Kd_h^{\pi,\nu}(s,a,b)}\right) d_h^{\pi,\nu}(s,a,b)\right) \geq 0 \ .
\end{aligned}$$

Now let us define

$$c' = \max\left\{\frac{3}{5}c_1 - \frac{2\lambda_h}{5Kd_h^{\pi,\nu}(s,a,b)} : d_h^{\pi,\nu}(s,a,b) > 0\right\} \ .$$

We have

$$\min_{(s,a,b)} \left(d_h^\rho(s,a,b) - c' d_h^{\pi,\nu}(s,a,b)\right) \geq 0 \ .$$

Thus, for $c = 1/c'$, we obtain

$$\frac{d_h^{\pi,\nu}(s,a,b)}{d_h^\rho(s,a,b)} < c < \infty, \forall h \in [H], (\pi,\nu) \in \mathcal{U}(\pi^*,\nu^*), (s,a,b) \in \{(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B} : d_h^{\pi,\nu}(s,a,b) > 0\} \ .$$

□

C.2 Proof of Proposition 2

First, we show the equivalence of both assumptions in the tabular Markov game setting. Note that, in the tabular setting, the features are SAB dimensional and $\phi(s,a,b)$ is the unit vector $e_{s,a,b}$ with coordinate (s,a,b) set to 1. First, suppose Assumption 2 is true. Then we have,

$$\mathbb{E}_{d_h^\rho}\left[e_{s,a,b}e_{s,a,b}^\top\right] = \text{diag}\left[\{d_h^\rho(s,a,b)\}_{s,a,b}\right] \succeq \xi \mathrm{I} \tag{32}$$

This implies that $d_h^\rho(s,a,b) \geq \xi$ for any h and s,a,b and Assumption 7 is satisfied.

Now, suppose Assumption 7 is true. Then $d_h^\rho(s,a,b) > 0$ for any h and tuple (s,a,b). Since the number of states is finite, there exists a constant C such that $d_h^\rho(s,a,b) \geq C$ for any h and any tuple (s,a,b). Therefore, Equation (32) is satisfied with $\xi = C$, and thus, Assumption 2 is satisfied.

Next, we show that Assumption 2 is actually a stronger assumption than Assumption 7 for the more general linear MDP model. We will write $\Phi \in \mathbb{R}^{SAB\times d}$ to denote the feature matrix where the row (s,a,b) corresponds to the d-dimensional feature $\phi(s,a,b)$.[7]

Lemma 15. *Suppose rank$(\Phi) = d$. Then Assumption 2 implies Assumption 7.*

Proof. We will assume $\mathcal{S}$ is finite but possibly very large. The proof can be easily generalized for infinite $\mathcal{S}$. Since Φ has rank d let us write $\Phi = U\Lambda V^\top$ where $U \in \mathbb{R}^{SAB\times d}$, Λ is a d-dimensional diagonal matrix, and $V \in \mathbb{R}^{d\times d}$. Moreover, we can take V to be orthonormal i.e. $V^\top V = \mathrm{I}$. Let $D_h^\rho \in \mathbb{R}^{SAB\times SAB}$ be a diagonal matrix with $D^\rho(s,a,b) = d_h^\rho(s,a,b)$.

$$\begin{aligned}&\mathbb{E}_{d_h^\rho}\left[\phi(s,a,b)\phi(s,a,b)^\top\right] = \Phi^\top D_h^\rho \Phi \\ &= V\Lambda U^\top D_h^\rho U \Lambda V^\top \succeq \xi \mathrm{I}_{d\times d}\end{aligned}$$

Since Φ has rank d, both Λ and V are invertible. This gives us

$$\begin{aligned}&U^\top D_h^\rho U \succeq \xi (V\Lambda)^{-1}(\Lambda V^\top)^{-1} \\ &= \xi \Lambda^{-1}(V^\top V)^{-1}\Lambda^{-1} \succeq \xi \Lambda^{-2}\mathrm{I}_{d\times d}\end{aligned}$$

Since the matrix $U^\top D_h^\rho U$ is already in diagonalized representation, the minimum eigenvalue of $U^\top D_h^\rho U$ is the smallest diagonal entry of D_h^ρ. Therefore,

$$\lambda_{\min}\left(U^\top D_h^\rho U\right) = \min_{s,a,b} d_h^\rho(s,a,b) \geq \frac{\xi}{\max_{j\in[d]}\Lambda(j)^2}$$

Since Φ has rank d at least one entry of the diagonal matrix Λ is non-zero. Therefore, for any h, the tuple (s,a,b) is covered with probability at least p where $p = \xi/\max_j \Lambda(j)^2$.

On the other hand, note that, for general rank-d feature matrix Φ, assumption 7 need not imply assumption 2. However, if we put additional restrictions on the features e.g. diversity, these two assumptions could be equivalent. □

[7] For infinite $\mathcal{S}$, Φ is interpreted as a function.