Corruption-robust Offline Multi-agent Reinforcement Learning From Human Feedback

Andi Nika
MPI-SWS

Debmalya Mandal
University of Warwick

Parameswaran Kamalaruban
Visa

Adish Singla
MPI-SWS

Goran Radanović
MPI-SWS

Abstract

We consider robustness against data corruption in offline multi-agent reinforcement learning from human feedback (MARLHF) under a strong-contamination model: given a dataset D of trajectory–preference tuples (each preference being an n-dimensional binary label vector representing each of the n agents' preferences), an ϵ-fraction of the samples may be arbitrarily corrupted. We model the problem using the framework of linear Markov games. First, under a *uniform coverage* assumption—where every policy of interest is sufficiently represented in the clean (prior to corruption) data—we introduce a robust estimator that guarantees an $O(\epsilon^{1-o(1)})$ bound on the Nash-equilibrium gap. Next, we move to the more challenging *unilateral coverage* setting, in which only a Nash equilibrium and its single-player deviations are covered: here our proposed algorithm achieves an $O(\sqrt{\epsilon})$ Nash-gap bound. Both of these procedures, however, suffer from intractable computation. To address this, we relax our solution concept to *coarse correlated equilibria* (CCE). Under the same unilateral-coverage regime, we then derive a quasi-polynomial-time algorithm whose CCE gap scales as $O(\sqrt{\epsilon})$. To the best of our knowledge, this is the first systematic treatment of adversarial data corruption in offline MARLHF.

1 INTRODUCTION

Reinforcement Learning from Human Feedback (RLHF) [Christiano et al., 2017] has surged in popularity as a straightforward, efficient means of fine-tuning large language models (LLMs) [Ouyang et al., 2022]. While most existing work focuses on single-agent settings, many real-world applications involve multiple interacting decision-makers, such as autonomous vehicles, distributed systems, or strategic marketplaces. Extending RLHF to such settings leads to multi-agent RLHF (MARLHF), where the goal is to learn aligned joint policies from preference data. Yet, despite its importance, research on MARLHF remains remarkably limited [Zhang et al., 2024].

At the same time, a critical shortcoming of all learning algorithms is their susceptibility to data-poisoning attacks—and RLHF is no exception [Wang et al., 2023a, Shi et al., 2023, Rando and Tramèr, 2023, Nika et al., 2025, Baumgärtner et al., 2024]. By injecting malicious feedback or subtly corrupting preference labels, adversaries can steer LLMs toward biased or harmful outputs—an especially serious risk as these models are increasingly deployed in safety-critical settings. Mandal et al. [2025] have proposed robust methods against data corruption for single-agent RLHF. However, no prior work has tackled the robustness of MARLHF systems against data-poisoning attacks. The added strategic complexity of MARLHF can amplify the impact of such attacks. It is therefore unclear whether single-agent methods extend directly to this setting.

Motivated by the above, we initiate the study of corruption-robust offline MARLHF. Our setting assumes access to preference data $D = \{(\tau, \tau', o)\}$ of size m, where τ, τ' are two sample trajectories and o is an n-dimensional vector of binary entries, each denoting a preference over the trajectory pair, corresponding to one of n agents. We assume that an ϵ-fraction of D is arbitrarily corrupted by an attacker. Using the standard Bradley-Terry (BT) [Bradley and Terry, 1952] preference model, we cast our problem as an instance of a linear Markov game. Previous work has already established that the optimal theoretical guarantees in corruption-robust offline RL and two-player zero-sum Markov games, under linear function approximation, exhibit a linear dependence on ϵ. However, such dependence is achieved only when the data covers all possible directions (i.e. *uniform cov-*

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

erage). When the data covers only a Nash policy and its unilateral deviations (i.e. *unilateral coverage*), prior work in RLHF (also RL and two-player zero-sum MGs) achieves $\sqrt{\epsilon}$ bounds. An immediate question is whether we can attain the same ϵ-dependent robustness rates in MARLHF as those of RLHF. Optimal dependence can be expected under uniform coverage. However, weaker coverage assumptions (e.g. unilateral coverage) imply a more challenging setting.

Challenges. The main challenge stems from uncertainty over multiple reward models: deriving worst-case guarantees means selecting a policy that performs well under *every* reward model in a confidence set obtained from a robust reward estimation, but computing that policy's worst-case performance is challenging—it requires optimizing over all candidate reward models, even though our data-coverage assumption only holds with respect to the true (unknown) reward. In the single-agent case, Mandal et al. [2025] resort to subgradient methods (yielding an $O(\epsilon^{1/4})$ rate) and then a primal-dual approach to recover $O(\sqrt{\epsilon})$. No analogous primal-dual theory exists for Markov games. Thus, we take a different approach.

Our approach. In multi-agent environments, the ultimate objective is to identify equilibrium policies, and minimizing the *Nash gap* provides a natural surrogate for that goal. Our key insight is that, for any reward model in the confidence set obtained by a robust reward estimation, the gap at a true Nash equilibrium policy $\boldsymbol{\pi}^*$ must lie within a small margin of the minimal gap over all policies, up to the reward-estimation error induced by the confidence set. Thus, if $\widetilde{\boldsymbol{\pi}}$ nearly minimizes the gap for a candidate reward model, then the gradient of $\boldsymbol{\pi}^*$ with respect to rewards can serve as a biased—but usable—proxy for the gradient at $\widetilde{\boldsymbol{\pi}}$. To estimate that proxy, we make use of our unilateral coverage (i.e. D sufficiently covers $\boldsymbol{\pi}^*$ and its unilateral deviations). In the linear Markov-game setting, the empirical feature differences between data-generating policies $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_{\text{ref}}$ then furnish a tractable approximation to the desired gradient. Plugging this into a first-order optimizer allows us to obtain the desired $O(\sqrt{\epsilon})$ guarantee on the Nash gap. In particular, we make the following contributions. A summary of our results is given in Table 1.

- First, assuming that D has uniform coverage, we design an algorithm that approximately computes a NE solely from corrupted preference data. Our algorithm first robustly estimates each reward function. Then, it runs a value-based backward-induction procedure to compute a policy that minimizes an estimate of the gap. We prove that our algorithm incurs $O(n\epsilon^{1-o(1)} + n/\sqrt{m})$ bounds on the Nash gap.
- Next, we relax our coverage assumption on our data and assume only unilateral coverage. We run projected gradient ascent (PGA) with a biased estimate of the gradient of the true gap for T_1 steps to compute the worst-case reward parameter. Using that parameter, we run the same procedure as in our previous method. We finally prove that our algorithm incurs $O(n\sqrt{\epsilon} + n/\sqrt{m} + n/\sqrt{T_1})$ bounds on the Nash gap.
- Our final contribution is on computational tractability. It is well-known that the NE computation is intractable for general-sum Markov games. We thus relax the NE notion into that of coarse correlated equilibrium (CCE). This allows us to frame each stage game as a saddle-point problem with convex-concave objective. We then utilize Optimistic Hedge to learn the CCE of each stage game. This yields and $O(n\sqrt{\epsilon} + n/\sqrt{m} + n/\sqrt{T_1} + n/T_2)$ bound on the CCE gap, where T_2 is the number of steps for which we run Optimistic Hedge.

2 PRELIMINARIES

This section contains the background technical material to be used throughout the paper.

2.1 Markov Games

A Markov game of finite horizon H between n agents is defined by the tuple $G = \big(S, \{A_i\}_{i=1}^n, \{P_h\}_{h=0}^{H-1}, \{\mathcal{R}_{i,h}\}_{h=0}^{H-1}, s_0\big)$, where S is the state space, A_i is the action set of agent i, $P_h : S \times A_1 \times \cdots \times A_n \to \Delta(S)$ is the state transition kernel at time-step h; the map $\mathcal{R}_{i,h} : S \times A_1 \times \cdots \times A_n \to \Delta(\mathbb{R})$ denotes the random reward of agent i at time-step h, with $R_{i,h}(s, \boldsymbol{a}) := \mathbb{E}[\mathcal{R}_{i,h}(s, \boldsymbol{a})|s, a]$; finally, $s_0 \in S$ denotes the initial state.

Policies and value functions. Given agent i, a Markov policy $\pi_i = (\pi_{i,0}, \ldots, \pi_{i,H-1})$ denotes the tuple containing the decision-making strategies of agent i, where, for each $h \in [H-1] := \{0, 1, \ldots, H-1\}$, $\pi_{i,h} : S \to \Delta(A_i)$ maps states to probability simplices over actions. A joint product policy is defined as the tuple $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ over all agents. We denote by $\Pi^{\text{PP}} = \Pi_1^{\text{PP}} \times \ldots \times \Pi_n^{\text{PP}}$ the overall product policy class and write $\boldsymbol{\pi} = (\pi_i, \boldsymbol{\pi}_{-i})$, where $\boldsymbol{\pi}_{-i} = (\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_n)$. Given joint policy $\boldsymbol{\pi}$ and state s at time-step h, the value function with respect to $\boldsymbol{\pi}$, s, and h is defined as

$$V_{i,h}^{\boldsymbol{\pi}}(s) = \mathbb{E}\left[\sum_{t=h}^{H-1} R_{i,t}(s_t, \boldsymbol{a}_t)|s_h = s, \boldsymbol{\pi}_t, P_t\right] ,$$

where $\boldsymbol{a} = (a_1, \ldots, a_n)$. Moreover, for given $\boldsymbol{a}$ the action value function is defined as

$$Q_{i,h}^{\boldsymbol{\pi}}(s, \boldsymbol{a}) = \mathbb{E}\left[\sum_{t=h}^{H-1} R_{i,t}(s_t, \boldsymbol{a}_t)|s_h = s, \boldsymbol{a}_h = \boldsymbol{a}, \boldsymbol{\pi}_t, P_t\right]$$

Nash equilibria. A product policy $\boldsymbol{\pi}^*$ is said to be an α-*Nash equilibrium* if there exists $\alpha \geq 0$, such that, for every agent i and state s, we have $V_{i,0}^{\boldsymbol{\pi}^*}(s) \geq V_{i,0}^{\pi_i', \boldsymbol{\pi}_{-i}^*}(s) - \alpha$, for every $\pi_i' \in \Pi_i^{\text{PP}}$. If $\alpha = 0$, then $\boldsymbol{\pi}^*$ is said to be a Nash equilibrium. We also define the notion of optimality of a

	Bounds on the NE (CCE) gap
NE & Unif. Cov.	$\widetilde{O}\left(\left(\frac{1}{\xi_R}+\frac{1}{\xi_P}\right) Hn\epsilon^{1-o(1)} + \frac{H^2 n\sqrt{\text{poly}(d)}}{\xi_P\sqrt{m}}\right)$
NE & Unil. Cov.	$\widetilde{O}\left(\left(\frac{1}{\sqrt{C_R}}+\frac{1}{\sqrt{C_P}}+\frac{1}{\sqrt{T_1}}\right)\left(H^{5/2}nd^{3/4}\sqrt{\epsilon}+\frac{H^2 n\sqrt{\text{poly}(d)}}{\sqrt{m}}\right)\right)$
CCE & Unil. Cov.	$\widetilde{O}\left(\left(\frac{1}{\sqrt{C_R}}+\frac{1}{\sqrt{C_P}}+\frac{1}{\sqrt{T_1}}\right)\left(H^{5/2}nd^{3/4}\sqrt{\epsilon}+\frac{H^2 n\sqrt{\text{poly}(d)}}{\sqrt{m}}\right)+\frac{Hn^2}{T_2}\right)$

Table 1: Summary of our bounds for: (i) NE gap minimization under uniform coverage, (ii) NE gap minimization under unilateral coverage, and (iii) CCE gap minimization under unilateral coverage. Here, $\widetilde{O}$ hides any poly-logarithmic factors, n denotes the number of agents, H denotes the horizon and d denotes the dimension. Moreover, ξ_R and ξ_P denote the uniform coverage coefficients, while C_R and C_P denote the unilateral coverage coefficients; ϵ denotes the corruption parameter, m is the data size, T_1 is the number of gradient steps while T_2 is the number of steps for which `OptimisticHedge` is run. It is worth mentioning that our bounds have optimal dependence on ϵ in the uniform coverage setting, while maintaining the same dependence as the single-agent (and two-player zero-sum Markov game) settings under non-uniform coverage. Moreover, note that the algorithm for CCE approximation is also computationally efficient.

given policy $\boldsymbol{\pi}$ profile. The *Nash gap* [Cui and Du, 2022] of $\boldsymbol{\pi}$ is defined as $\text{Gap}(\boldsymbol{\pi}) = \sum_{i\in[n]} V_{i,0}^{\dagger,\boldsymbol{\pi}_{-i}}(s_0) - V_{i,0}^{\boldsymbol{\pi}}(s_0)$, where $V_{i,0}^{\dagger,\boldsymbol{\pi}_{-i}}(s_0) = \max_{\pi_i'} V_{i,0}^{\pi_i',\boldsymbol{\pi}_{-i}}(s_0)$. Note that, by definition, any Nash equilibrium has 0 Nash gap, and any α-Nash equilibrium has at most α Nash gap.

Linear Markov games. In this paper, we consider linear Markov games [Zhong et al., 2022]. Formally, G is said to be a linear Markov game with feature map $\phi : S \times A \to \mathbb{R}^d$, for some $d \in \mathbb{N}$, if we have $P_h(s_{h+1}|s_h, \boldsymbol{a}_h) = \langle \phi(s_h, \boldsymbol{a}_h), \xi_h(s_{h+1})\rangle$, and $\mathcal{R}_{i,h}(s_h, \boldsymbol{a}_h) = \langle \phi(s_h, \boldsymbol{a}_h), \theta_{i,h}^*\rangle + \zeta_{i,h}, \forall (s_h, \boldsymbol{a}_h, i, h) \in S \times A \times [n] \times [H-1]$, where ξ_h and θ_i^* are unknown parameters and $\zeta_{i,h}$ zero-mean γ^2-subGaussian noise. Here, $\|\phi(s, \boldsymbol{a})\|_2 \leq 1$ for all state-action tuples $(s, \boldsymbol{a}) \in S \times A$, $\max\{\left\|\theta_{i,h}^*\right\|_2, \|\xi_h(s)\|_2\} \leq \sqrt{d}$, for all $i \in [n]$ and $h \in [H-1]$. Let Θ denote the set of all feasible θ as defined here.

Remark 1. *There are two main reasons why we consider linear Markov games to model our problem. First, the corruption-robust offline RL literature in the general function approximation [Ye et al., 2023] consider a corruption model which is defined in terms of Bellman residuals. Since we only assume access to preference data, this type of corruption model is not well-defined for our setting. Second, relaxing the linearity of rewards would then require corruption-robust maximum likelihood estimation procedures beyond generalized linear models, which, to the best of our knowledge, are not present in the current literature.*

2.2 Preference Data

Following the formulation in [Zhang et al., 2024], we denote by $\widetilde{D} = \{(\widetilde{\tau}_i, \widetilde{\tau}_i', \widetilde{o}_i)\}_{i=1}^m$ the clean preference dataset, where $\widetilde{\tau} = (\widetilde{s}_0, \widetilde{a}_{1,0}, \widetilde{a}_{2,0}, \ldots, \widetilde{a}_{n,0}, \widetilde{s}_1, \ldots, \widetilde{s}_{H-1})$ and $\widetilde{\tau}' = (\widetilde{s}_0', \widetilde{a}_{1,0}', \widetilde{a}_{2,0}', \ldots, \widetilde{a}_{n,0}', \widetilde{s}_1', \ldots, \widetilde{s}_{H-1}')$ denote sampled trajectories from behavior policies $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_{\text{ref}}$, respectively, and $\widetilde{o}_i = (\widetilde{o}_{i,1}, \ldots, \widetilde{o}_{i,n})$, with $\widetilde{o}_{i,j} \in \{-1, +1\}$, for all $j \in [n]$, gives information about individual preferences of agents for each pair of trajectories: $\widetilde{o}_{i,j} = 1$ implies that $\widetilde{\tau}_i$ is preferred to $\widetilde{\tau}_i'$ for agent j. We assume that the preferences are generated according to the Bradley-Terry (BT) model [Bradley and Terry, 1952]: for each agent j, we assume that we have

$$\mathbb{P}(\widetilde{o}_{i,j} = 1|\widetilde{\tau}_i, \widetilde{\tau}_i') = \sigma\left(\sum_{h=0}^{H-1} R_{i,h}(\widetilde{s}_h, \widetilde{\boldsymbol{a}}_h) - \sum_{h=0}^{H-1} R_{i,h}(\widetilde{s}_h', \widetilde{\boldsymbol{a}}_h')\right) ,$$

with $\sigma(x) = 1/(1+\exp(-x))$ being the sigmoid function.

2.3 Corruption Model

Following the ϵ-corruption model in offline RL(HF) [Zhang et al., 2022, Mandal et al., 2025], we assume that there exists an attacker that has full access to the dataset $\widetilde{D}$ and arbitrarily perturbs an ϵ-fraction of it. That is, given $\epsilon \in [0, 1/2)$, we assume that the attacker inspects $\widetilde{D}$ and modifies up to $\epsilon \cdot m$ samples in $\widetilde{D}$. We denote by D the poisoned dataset. In other words, there are at most $\epsilon \cdot m$ data samples in D such that $(\tau, \tau', o) \neq (\widetilde{\tau}, \widetilde{\tau}', \widetilde{o})$.[1]

2.4 Data Coverage

Offline learning problems necessitate access to a dataset that contains, at least to some extent, "good" samples, in the sense that they are traversed by policies we are trying to approximate. This condition is usually described by the notion of *data coverage*. In linear Markov games (MG), data coverage is captured by the feature covariance matrix. Formally, for every $h \in [H-1]$, we define

$$\Sigma_{\boldsymbol{\mu}}(h) = \mathbb{E}_{\boldsymbol{\mu}_h}\left[\phi(s_h, \boldsymbol{a}_h)\phi(s_h, \boldsymbol{a}_h)^\top\right]$$

[1] Following prior literature on corruption-robust RL [Zhang et al., 2022], we assume that, for each $h \in [H-1]$, the subset of D containing only the h steps of the samples, is also consistent with the ϵ-corruption model.

and

$$\Sigma^{-}_{\boldsymbol{\mu},\boldsymbol{\mu}_{\text{ref}}} = \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\mu}_{\text{ref}}}\left[\left(\phi(\tau)-\phi(\tau')\right)\left(\phi(\tau)-\phi(\tau')\right)^{\top}\right]$$

as the feature covariance matrices that will determine coverage of our given data. Here $\phi(\tau) = \sum_{h=0}^{H-1}\phi(s_h, \boldsymbol{a}_h)$. Note that $\Sigma_{\boldsymbol{\mu}}(h)$ is the standard covariance matrix used in the literature of offline RL, while $\Sigma^{-}_{\boldsymbol{\mu},\boldsymbol{\mu}_{\text{ref}}}$ is the difference covariance matrix which has been previously used in RLHF literature [Zhan et al., 2023].

3 ROBUST NE LEARNING UNDER UNIFORM COVERAGE

We start by considering the case when the preference dataset has uniform coverage, that is, all basis directions of the feature space are sufficiently covered. We state this assumption below.

Assumption 1 (Uniform Coverage)**.** *Let $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_{\text{ref}}$ be the behavior policies that were used to generate trajectories present in the data $\widetilde{D}$. We assume that we have $\Sigma^{-}_{\boldsymbol{\mu},\boldsymbol{\mu}_{\text{ref}}} \succeq \xi_R H \cdot I$ and $\Sigma_{\boldsymbol{\mu}}(h) \succeq \xi_P \cdot I$, for all $h \in [H-1]$, where ξ_R and ξ_P are strictly positive constants and $A \succeq B$ is equivalent to $x^{\top}(A-B)x \geq 0$, for all vectors $x \neq 0$.*

Remark 2. *Note that we require coverage with respect to two covariance matrices. The first one, $\Sigma^{-}_{\boldsymbol{\mu},\boldsymbol{\mu}_{\text{ref}}}$, captures coverage of the rewards, since rewards are estimated in terms of feature differences. The second one, $\Sigma_{\boldsymbol{\mu}}(h)$, captures coverage of transitions for each step h. In standard RL, the second condition is enough to provide guarantees. However, in preference-based RL, the first condition is necessary due to the lack of the reward signal in the given data [Zhan et al., 2023].*

On a high level, all our algorithms are based on the following pipeline. They first use the preference data to robustly estimate each agent's reward parameters. Then, using those rewards, they proceed to compute robust pessimistic and optimistic estimates of the individual Q-functions for policies of interest. Finally, they estimate the gap and output a joint policy that minimizes it. We will instantiate different versions of this pipeline under different coverage assumptions and notions of equilibria.

Robust Reward Estimation
$\Rightarrow$ Robust Q-function Estimation
$\Rightarrow$ Estimated Gap Minimization

3.1 Algorithm

The main idea of our proposed algorithm is as follows. First, note that our overall objective is to find a joint policy that minimizes the Nash gap with respect to the ground-truth reward functions. However, we have access neither to these functions, nor to the real environment. We are only given a finite preference dataset D, an ϵ-fraction of which is arbitrarily corrupted. Therefore, our first step is to compute robust estimates of the reward parameters of each agent. Note that, for linear rewards, maximum likelihood estimation becomes standard logistic regression. And for such an objective, it is known [Awasthi et al., 2022] that we can recover the true parameter of interest from ϵ-corrupted data with $O(\epsilon^{1-o(1)})$ accuracy via a robust method called `TrimmedMLE` (for a pseudocode, see Algorithm 6 in Appendix E). Thus, for each agent i, we let $\widetilde{\theta}_i = \texttt{TrimmedMLE}\left(D, \epsilon, \nu\right)$, where we denote by θ_i the Hd-dimensional result of the concatenation of $\theta_{i,h}$, for $h \in [H-1]$, and ν denotes a granularity hyperparameter.

Algorithm 1 Corruption-robust Equilibrium Learning from Human Feedback with Uniform Coverage

Require: Preference dataset D; confidence parameter δ.
Split D into equal D_1 and D_2.
▷ `Reward Estimation via Trimmed MLE using` D_1.
for $i \in [n]$ **do**
  Compute $\widetilde{\theta}_i = \texttt{TrimmedMLE}(D_1, \epsilon, \nu)$.
  Set the optimistic and pessimistic rewards $\overline{R}_{i,h}(\cdot,\cdot)$ and $\underline{R}_{i,h}(\cdot,\cdot)$ as in Equations (2) and (3), respectively.
  ▷ `Robust Value Function Estimation Phase using` D_2.
  for $\boldsymbol{\pi} \in \Pi^{\text{PP}}$ **do**
    Apply Algorithm 7 on D_2 using only the preferred trajectories generated by $\boldsymbol{\mu}$, with input $\boldsymbol{\pi}$, i, $\underline{R}_i$, and $\overline{R}_i$, and bonus function $\Gamma(\cdot,\cdot) = 0$, to obtain $\overline{V}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(\cdot)$ and $\underline{V}^{\boldsymbol{\pi}}_{i,h}(\cdot)$, for all $h \in [H-1]$.
  end for
end for
▷ `Nash Gap Estimation Phase`
for every policy $\boldsymbol{\pi} \in \Pi^{\text{PP}}$: **do**
  Compute the estimated gap $\widetilde{\text{Gap}}(\boldsymbol{\pi})$.
end for
return $\widetilde{\boldsymbol{\pi}} \in \arg\min_{\boldsymbol{\pi}\in\Pi^{\text{PP}}} \widetilde{\text{Gap}}(\boldsymbol{\pi})$.

Next, since we are in the offline setting, the most reasonable approach is to apply pessimism with respect to the recovered parameters. To that end, we form confidence sets, for each agent i, based on the `TrimmedMLE` guarantees:

$$\Theta_{\text{Unif}}(\widetilde{\theta}_i) = \left\{\theta \in \Theta : \left\|\widetilde{\theta}_i - \theta\right\|_2 \leq O\left(\frac{\epsilon^{1-o(1)}}{\xi_R}\right)\right\}, \tag{1}$$

where $\delta > 0$ is a randomness parameter. Once we have access to the confidence set, we compute the boundary parameters

$$\overline{R}_{i,h}(s,\boldsymbol{a}) = \max_{\theta_i \in \Theta_{\text{Unif}}(\widetilde{\theta}_i)} \theta_{i,h}^{\top}\phi(s,\boldsymbol{a}), \tag{2}$$

$$\underline{R}_{i,h}(s,\boldsymbol{a}) = \min_{\theta_i \in \Theta_{\text{Unif}}(\widetilde{\theta}_i)} \theta_{i,h}^{\top}\phi(s,\boldsymbol{a}). \tag{3}$$

Note that the inner problems are convex programs that have closed-form solutions. Now that we have access to our estimate reward functions, we need to minimize the Nash gap with respect to these rewards. In order to do that, we apply backward induction. First, we initialize the value function estimates $\underline{V}_H^{\boldsymbol{\pi}}(\cdot) = \overline{V}_H^{\dagger,\boldsymbol{\pi}_{-i}}(\cdot) = 0$, for every joint policy $\boldsymbol{\pi}$. Then, for every step h down to 0, we apply a robust estimation algorithm `Rob-Q` (see Algorithm 2) for the Q-values. Essentially, the procedure first robustly estimates the parameters of the Bellman operator using a `RobEst` oracle which is guaranteed to return an $O(\epsilon)$-close value parameter under uniform coverage [Zhang et al., 2022]. Then, it computes the estimated Q-values by properly clipping the bonus-inflated (deflated) estimates so that they remain in $[-H\sqrt{d}, H\sqrt{d}]$. Once we have the Q-functions,

Algorithm 2 Robust Estimation of Q-Functions (`Rob-Q`)

Require: Dataset D; corruption level ϵ; policy $\boldsymbol{\pi}$; agent i; reward functions $\overline{R}_{i,h}$ and $\underline{R}_{i,h}$; step h; next-step value estimates $\underline{V}_{i,h+1}^{\boldsymbol{\pi}}(\cdot), \overline{V}_{i,h+1}^{\dagger,\boldsymbol{\pi}_{-i}}(\cdot)$; bonus $\Gamma(\cdot,\cdot)$.

1: Set $\underline{\omega}_{i,h}^{\boldsymbol{\pi}}$ as $\texttt{RobEst}\left(\phi(s_h,\boldsymbol{a}_h), \underline{R}_{i,h}(s_h,\boldsymbol{a}_h) + \underline{V}_{i,h+1}^{\boldsymbol{\pi}}(s_{h+1})\right)$.

2: Set $\overline{\omega}_{i,h}^{\dagger,\boldsymbol{\pi}_{-i}}$ as $\texttt{RobEst}\left(\phi(s_h,\boldsymbol{a}_h), \overline{R}_{i,h}(s_h,\boldsymbol{a}_h) + \overline{V}_{i,h+1}^{\dagger,\boldsymbol{\pi}_{-i}}(s_{h+1})\right)$.

3: Set $\underline{Q}_{i,h}^{\boldsymbol{\pi}}(\cdot,\cdot)$ as $\text{Clip}_{[-(H-h)\sqrt{d},(H-h)\sqrt{d}]}\left(\phi(\cdot,\cdot)^\top \underline{\omega}_{i,h}^{\boldsymbol{\pi}} - \Gamma(\cdot,\cdot)\right)$.

4: Set $\overline{Q}_{i,h}^{\dagger,\boldsymbol{\pi}_{-i}}(\cdot,\cdot)$ as $\text{Clip}_{[-(H-h)\sqrt{d},(H-h)\sqrt{d}]}\left(\phi(\cdot,\cdot)^\top \overline{\omega}_{i,h}^{\dagger,\boldsymbol{\pi}_{-i}} + \Gamma(\cdot,\cdot)\right)$.

5: **return** Q-functions $\underline{Q}_{i,h}^{\boldsymbol{\pi}}(\cdot,\cdot)$ and $\overline{Q}_{i,h}^{\dagger,\boldsymbol{\pi}_{-i}}(\cdot,\cdot)$.

the value function estimates $\underline{V}_{i,h}^{\boldsymbol{\pi}}(\cdot)$ (and $\overline{V}_{i,h}^{\dagger,\boldsymbol{\pi}_{-i}}(\cdot)$) for all steps, defined with respect to the estimated reward parameters, are then computed by taking expectations over (and taking $\max$ over actions for player i) the given policies.[2] Once we do this for every policy, we then return the policy $\widetilde{\boldsymbol{\pi}}$ that minimizes the estimated gap with respect to $(\overline{R}, \underline{R}) = (\overline{R}_1, \underline{R}_1, \ldots, \overline{R}_n, \underline{R}_n)$:

$$\arg\min_{\boldsymbol{\pi}} \widetilde{\text{Gap}}\left(\boldsymbol{\pi}, \overline{R}, \underline{R}\right) := \sum_{i\in[n]} \overline{V}_{i,0}^{\dagger,\boldsymbol{\pi}_{-i}}(s_0) - \underline{V}_{i,0}^{\boldsymbol{\pi}}(s_0) \ . \tag{4}$$

As shown in Appendix (see Lemma A.6), our optimistic and pessimistic value estimates are high-probability approximates of the true value function, which implies that the estimated gap is a high-probability upper bound on the actual Nash gap. Minimizing this surrogate gap therefore serves as a proxy for minimizing the true Nash gap—and, as the gap approaches zero, the resulting joint policy correspondingly approaches a Nash equilibrium. Algorithm 1 provides the pseudocode for the full procedure.

[2]Note that Algorithm 7 runs Algorithm 2 for H steps and finally returns the estimates of the value functions. We have used Algorithm 7 to present Algorithm 1 for ease of presentation. However, Algorithm 2 will be necessary in the following sections.

3.2 Theoretical guarantees

In this section, we state the theoretical guarantees on the convergence of the Algorithm 1. Proofs can be found in Appendix A.

Theorem 3.1. *Let $\epsilon \in [0, 1/2)$, $\delta > 0$ and $\Gamma(\cdot,\cdot) = 0$. Furthermore, assume that $m \geq \Omega((H^{3/2}/\epsilon^2)(d + \log(n/\delta)))$. Then, under Assumption 1 with $\xi_R \geq 5\epsilon$, for some positive constant c, there exist robust algorithms `TrimmedMLE` and `RobEst` such that, with probability at least $1-\delta$, the output $\widetilde{\boldsymbol{\pi}}$ of Algorithm 1 satisfies*

$$\text{Gap}(\widetilde{\boldsymbol{\pi}}) \leq O\left(Hn\left(\frac{\exp\left(H + \sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} + \frac{H\sqrt{d}+\gamma}{\xi_P}\right)\cdot\epsilon + Hn\sqrt{\frac{(H\sqrt{d}+\gamma)^2 \text{poly}(d)}{\xi_P^2 m}}\right)$$

Remark 3. *Note that the bounds of Theorem 3.1 have a quasi-linear dependence on ϵ, which is known to be optimal in the single agent setting [Zhang et al., 2022] and the two-agent zero-sum setting [Nika et al., 2024b]. This is due to the strong coverage assumption on the data. In practice, the data may cover only some directions of interest, in which case, a different approach is needed.*

4 ROBUST NE LEARNING UNDER UNILATERAL COVERAGE

In the previous section, we proposed an algorithm that returns an $O(n\epsilon + n/\sqrt{m})$-approximate Nash equilibrium under uniform coverage. However, such coverage is rarely possible in practice. The purpose of this section is to solve the Nash gap minimization problem under a more relaxed notion of coverage, namely *unilateral coverage*, which simply requires coverage of a Nash policy and all its unilateral deviations for each agent. We extend the notion of low relative uncertainty of [Zhong et al., 2022] to the MARLHF setting.

Assumption 2 (Unilateral Coverage)**.** *Given Nash equilibrium $\boldsymbol{\pi}^*$, we assume that there exist positive constants C_R and C_P, such that, for all $h \in [H-1]$ and $i \in \{1,\ldots,n\}$,*

$$\Sigma^-_{\boldsymbol{\rho},\boldsymbol{\rho}'} \succeq C_R \cdot \Sigma^-_{(\pi_i,\boldsymbol{\pi}^*_{-i}),\boldsymbol{\rho}'} \ , \quad \text{for} \ \ \boldsymbol{\rho},\boldsymbol{\rho}' \in \{\boldsymbol{\mu}, \boldsymbol{\mu}_{\text{ref}}\},$$
$$\text{and} \ \ \Sigma_{\boldsymbol{\mu}}(h) \succeq C_P \cdot \Sigma_{\pi_i,\boldsymbol{\pi}^*_{-i}}(h) \ .$$

The first condition simply says that behavior policies $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_{\text{ref}}$ sufficiently cover a Nash equilibrium and its unilateral deviations in the feature space. Different from single-agent

RL, where single policy concentrability is enough to provide theoretical guarantees, its extension to Markov games, where only Nash policies are covered does not allow for any guarantees. Unilateral coverage is in fact necessary and sufficient to provide any meaningful guarantees in zero-sum [Zhong et al., 2022] and general-sum Markov games [Zhang et al., 2023].

4.1 Algorithm

For the uniform coverage setting, we applied `TrimmedMLE` to obtain estimates for the ground-truth reward parameters. The benefit of such an approach is that, under such coverage, it comes with bounds on the ℓ^2-norm of the error, which then allows for defining our confidence set in terms of such error bounds. This, in turn, allows us to directly upper bound the difference between value functions and their estimates in terms of ℓ^2-difference of their respective reward parameters. When we do not have uniform coverage, the final estimate is not guaranteed to remain close to the true parameter in the ℓ^2 sense. In this case, as shown in Appendix (see Lemma B.1), the parameters are close in the log-sigmoid sense. Given the output $\widetilde{\theta}_i$ of `TrimmedMLE`, we define the confidence set for the unilateral coverage setting as

$$\Theta_{\text{Unil}}(\widetilde{\theta}_i) = \Bigg\{\theta \in \Theta : \frac{2}{m} \sum_{(\tau,\tau',o)\in D} \log \frac{\sigma\left(o \cdot \widetilde{\theta}_i^\top \left(\phi(\tau) - \phi(\tau')\right)\right)}{\sigma\left(o \cdot \theta^\top \left(\phi(\tau) - \phi(\tau')\right)\right)} \leq \kappa \Bigg\}, \quad (5)$$

where $\kappa = 6\epsilon H\sqrt{d} + (2d/m) \cdot \log(Hm/\delta)$ controls the 'radius' of the confidence set. We can provide theoretical guarantees that $\Theta_{\text{Unil}}(\widetilde{\theta}_i)$ contains the ground-truth parameter θ_i^* with high probability. Unfortunately though, the same analysis does not go through just by choosing reward estimates that maximize (minimize) over the confidence set, due to this more complicated notion of closeness. Hence, we follow a different approach in this section. Intuitively, if we can find θ in our confidence set that maximizes *the gap of our output policy*, and, similarly, find a policy $\boldsymbol{\pi}$ that minimizes *the gap with respect to this choice* of θ, we can finally bound the true gap on the ground-truth reward. This intuition is based on the observation that any Nash gaps of policies that used θ parameters in the confidence set should be close to each-other.

Algorithm 3 Reward Parameter Estimation (`RewardEst`)

Require: Dataset D; corruption level ϵ; confidence parameter δ; learning rate η; slackness parameter ν; number of steps T.

for $i \in [n]$ **do**
  Let $\widetilde{\theta}_i = \texttt{TrimmedMLE}(D, \epsilon, \nu)$.
  Initialize $\widehat{\theta}_i^{(0)}$ uniformly at random in $\Theta_{\text{Unil}}(\widetilde{\theta}_i)$ (defined in Equation (5)).
  for $t = 0, 1, \ldots, T-1$ **do**
    Take gradient step

$$\widehat{\theta}_i^{(t+1)} = \mathcal{P}_{\Theta_{\text{Unil}}(\widetilde{\theta}_i)}\left(\widehat{\theta}_i^{(t)} + \eta \widetilde{\nabla}_{\theta_i} \text{Gap}\left(\boldsymbol{\pi}^*, \widehat{\boldsymbol{\theta}}^{(t)}\right)\right).$$

  end for
  Set $\widehat{\theta}_i = (1/T)\sum_{t=1}^{T} \widehat{\theta}_i^{(t)}$.
end for
return $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)$.

Algorithm 4 Corruption-robust Nash Equilibrium Learning from Human Feedback

Require: Dataset D; corruption level ϵ; regularization parameter λ; confidence parameter δ; learning rate η_1; bonus functions $\Gamma(\cdot,\cdot)$; slackness parameter ν; number of gradient steps T_1.

Split D into equal D_1 and D_2.
Compute $\widehat{\boldsymbol{\theta}} = \texttt{RewardEst}\,(D_1, \epsilon, \delta, \eta_1, \nu, T_1)$.
for $i \in [n]$ **do**
  for $\boldsymbol{\pi} \in \Pi^{\text{PP}}$ **do**
    Initialize $\underline{V}_{i,H}^{\boldsymbol{\pi}}(\cdot) = \overline{V}_{i,H}^{\dagger,\boldsymbol{\pi}_{-i}}(\cdot) = 0$.
    for $h = H-1, \ldots, 0$ **do**
      Compute $\widehat{R}_{i,h}(\cdot,\cdot) = (\widehat{\theta}_{i,h})^\top \phi(\cdot,\cdot)$ to be the estimated reward.
      Compute $\left(\underline{Q}_{i,h}^{\boldsymbol{\pi}}(\cdot,\cdot), \overline{Q}_{i,h}^{\dagger,\boldsymbol{\pi}_{-i}}(\cdot,\cdot)\right) = \texttt{Rob-Q}\left(D_{2,h}, \boldsymbol{\pi}, \epsilon, \widehat{R}_{i,h}, \underline{V}_{i,h+1}^{\boldsymbol{\pi}}, \overline{V}_{i,h+1}^{\dagger,\boldsymbol{\pi}_{-i}}, \Gamma\right)$.
      Set $\underline{V}_{i,h}^{\boldsymbol{\pi}}(\cdot) = \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h}\left[\underline{Q}_{i,h}^{\boldsymbol{\pi}}(\cdot,\boldsymbol{a})\right]$ and $\overline{V}_{i,h}^{\dagger,\boldsymbol{\pi}_{-i}}(\cdot) = \max_{a_i} \mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_{-i,h}}\left[\overline{Q}_{i,h}^{\dagger,\boldsymbol{\pi}_{-i}}(\cdot,\boldsymbol{a})\right]$.
    end for
  end for
end for
return $\widetilde{\boldsymbol{\pi}} \in \arg\min_{\boldsymbol{\pi}\in\Pi^{\text{PP}}} \widetilde{\text{Gap}}(\boldsymbol{\pi}, \widehat{\boldsymbol{\theta}})$.

Based on the above discussion, we will utilize projected gradient ascent (PGA) to update our estimates of θ. However, we do not have access to the true gap given parameter θ, neither of its gradient with respect to θ. We thus resort to using biased estimates of it. As we show in Appendix (see Lemma B.6), for any $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_n) \in \Theta_{\text{Unil}}(\widetilde{\theta}_i) \times \ldots, \times\Theta_{\text{Unil}}(\widetilde{\theta}_n)$, any policy $\boldsymbol{\pi}$ that is the minimizer of the estimated gap computed via `Rob-Q` on $\boldsymbol{\theta}$, satisfies

$$|\text{Gap}(\boldsymbol{\pi}, \boldsymbol{\theta}) - \text{Gap}(\boldsymbol{\pi}^*, \boldsymbol{\theta})| \leq O(n\sqrt{\epsilon} + n/\sqrt{m}),$$

for Nash equilibrium policy $\boldsymbol{\pi}^*$ which is covered by D, where $\text{Gap}(\boldsymbol{\pi}, \boldsymbol{\theta})$ here denotes the true Nash gap of $\boldsymbol{\pi}$ under reward function parameterized by $\boldsymbol{\theta}$. Therefore, we optimize $\text{Gap}(\boldsymbol{\pi}^*, \boldsymbol{\theta})$ as a surrogate objective, and later transfer guarantees to $\text{Gap}(\boldsymbol{\pi}, \boldsymbol{\theta})$ using Lemma B.6.

However, we do not have access to $\nabla_{\boldsymbol{\theta}}\text{Gap}(\boldsymbol{\pi}^*, \boldsymbol{\theta})$. Here,

we use the following observation: in the linear setting, the (sub)gradient of the gap becomes the average feature differences over convex combinations of occupancy measures. Thus, we can use $\nabla_{\boldsymbol{\theta}} \sum_{i\in[n]}(V^{\boldsymbol{\mu}}_{i,0}(s_0,\theta_i) - V^{\boldsymbol{\mu}_{\text{ref}}}_{i,0}(s_0,\theta_i))$ as an estimate for $\nabla_{\boldsymbol{\theta}}\text{Gap}(\boldsymbol{\pi}^*,\boldsymbol{\theta})$, since $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_{\text{ref}}$ already cover $\boldsymbol{\pi}^*$ and its unilateral deviations. Here, $V^{\boldsymbol{\mu}}_{i,0}(s_0,\theta_i)$ denotes the value function of $\boldsymbol{\mu}$ with respect to reward function parametrized by θ_i. To estimate $\nabla_{\theta_i}(V^{\boldsymbol{\mu}}_{i,0}(s_0,\theta_i) - V^{\boldsymbol{\mu}_{\text{ref}}}_{i,0}(s_0,\theta_i))$, we use a robust mean oracle `RobMean` that takes as input corrupted feature differences and returns a $O(\sqrt{\epsilon})$-approximate estimate of their true mean, which in our case is just $\nabla_{\theta_i}(V^{\boldsymbol{\mu}}_{i,0}(s_0,\theta_i) - V^{\boldsymbol{\mu}_{\text{ref}}}_{i,0}(s_0,\theta_i))$. We thus define our gradient estimate $\widetilde{\nabla}_{\theta_i}\text{Gap}\,(\boldsymbol{\pi}^*,\boldsymbol{\theta})$ with respect to θ_i as

$$\sum_{h=0}^{H-1}\texttt{RobMean}\left(D^{\boldsymbol{\mu}}_{h,\phi}\right) - \sum_{h=0}^{H-1}\texttt{RobMean}\left(D^{\boldsymbol{\mu}_{\text{ref}}}_{h,\phi}\right)\ ,$$

where $D^{\boldsymbol{\mu}}_{h,\phi}$ and $D^{\boldsymbol{\mu}_{\text{ref}}}_{h,\phi}$ partition each h-level of D and store only the features of trajectories generated by $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_{\text{ref}}$, respectively. After running PGA for T_1 steps, we compute the empirical average of the iterates and use that to compute our estimated reward function. The reward estimation procedure `RewardEst` is described in Algorithm 3. Once we have access to this reward, we can run `Rob-Q` on it and obtain estimated gaps for each policy.

However, lack of uniform coverage implies weaker guarantees on `Rob-Q`. Thus, we need to properly define a bonus term that accounts for corruption and lack of coverage. First, let us define a scaled sample covariance matrix with respect to the preferred trajectories in the corrupted data, using regularization parameter $\lambda \geq 0$ (to be specified later) as

$$\Lambda_h = \frac{3}{5}\left(\frac{1}{m}\sum_{j=1}^{m}\phi(s^j_h,\boldsymbol{a}^j_h)\phi(s^j_h,\boldsymbol{a}^j_h)^\top + (\epsilon+\lambda)I\right).$$

Using Λ_h, we now define the bonus term to be used in this section as follows. For any $(s,\boldsymbol{a})$, let

$$\Gamma(s,\boldsymbol{a}) = E(d,m,\delta,\epsilon)\cdot\|\phi(s,\boldsymbol{a})\|_{\Lambda_h^{-1}}\ ,$$

where $E(d,m,\delta,\epsilon) = O(\sqrt{\epsilon} + 1/\sqrt{m})$ (see Appendix B for detailed definition). We run `Rob-Q` with bonus set as Γ and obtain estimated gaps for every joint policy. Finally, we return a joint policy that minimizes estimated gap. The full procedure is described in Algorithm 4.

4.2 Theoretical guarantees

In this section, we state the theoretical guarantee on the convergence of Algorithm 4.

Theorem 4.1. *Let $\epsilon \in [0,1/2)$, $\lambda \geq \Omega(dH\log(m/\delta)/m)$, and $\delta > 0$. Set $\Theta_{\text{Unil}}(\cdot)$ as in Equation (5) and $\Gamma(s,\boldsymbol{a}) = E(d,m,\delta,\epsilon)\cdot\|\phi(s,\boldsymbol{a})\|_{\Lambda_h^{-1}}$. Suppose Assumption 2 is satisfied and PGA is run for T_1 steps with learning rate $\eta = O(1/\sqrt{T_1})$. Then, there exist robust subroutines `RobEst`, `TrimmedMLE`, and `RobMean` such that, with probability at least $1-\delta$, the output $\widetilde{\boldsymbol{\pi}}$ of Algorithm 4 with subroutines `RobEst`, `TrimmedMLE`, `RobMean` and `RewardEst`, satisfies*

$$\text{Gap}(\widetilde{\boldsymbol{\pi}}) \leq \widetilde{O}\left(\left(\frac{1}{\sqrt{C_R}} + \frac{1}{\sqrt{C_P}} + \frac{1}{\sqrt{T_1}}\right)\cdot\left(H^{5/2}nd^{3/4}\sqrt{\epsilon} + H^2 n\sqrt{\frac{\text{poly}(d)}{m}}\right)\right).$$

Remark 4. *Note that the order of ϵ in the above bounds is $1/2$. This deterioration comes from the relaxation of uniform coverage. This dependence is identical to that in single-agent RL [Zhang et al., 2022], two-player zero-sum Markov games [Nika et al., 2024b], and single-agent RLHF [Mandal et al., 2025] under data corruption. The currently established linear lower bounds hold under uniform coverage. It remains an open question whether weaker coverage implies tighter lower bounds.*

5 ROBUST CCE LEARNING UNDER UNILATERAL COVERAGE

In the previous section, we provided an algorithm that was designed to compute an approximate NE using corrupted preference data under the minimal unilateral coverage assumption. However, a key bottleneck of Algorithm 4 is the intractability of the gap-minimization step. It is well-known that even normal-form general-sum games suffer from the curse of multi-agents—computational time scales exponentially with the number of agents (actions) [Foster et al., 2023]. Thus, to address this issue, previous work has considered more relaxed versions of the NE, such as *correlated equilibria* or *coarse correlated equilibria* (CCE) [Cui et al., 2023, Zhang et al., 2023, Ma et al., 2023, Song et al., 2021], the latter of which can be approximated using no-regret learning algorithms.

A general correlated policy is defined as a set of H maps $\boldsymbol{\pi} := \{\boldsymbol{\pi}_h : \Omega\times(S\times A)^{h-1}\times S \to \Delta(A)\}_{h\in[H-1]}$, where the first argument $w\in\Omega$ is sampled from some underlying distribution. A crucial difference from Markov policies is information about prior states that is given as input. We denote by Π^{GCP} the space of all general correlated policies. We denote by Π^{GCP}_i the set of general correlated policies for agent i. Then, policy $\boldsymbol{\pi}^*$ is said to be an α-CCE if there exists $\alpha\geq 0$, such that, for every agent i and state s, we have $V^{\boldsymbol{\pi}^*}_{i,0}(s) \geq V^{\pi'_i,\boldsymbol{\pi}^*_{-i}}_{i,0}(s) - \alpha$, for every $\pi'_i \in \Pi^{\text{GCP}}_i$. If $\alpha = 0$, then $\boldsymbol{\pi}^*$ is said to be a CCE. Note that the only difference between NEs and CCEs is that an NE is restricted to be a product policy, while a CCE can be any arbitrary combination of individual policies in the joint action space simplex. Hereafter, we overload notation and use $\text{Gap}(\boldsymbol{\pi})$ to denote the CCE gap of a joint policy $\boldsymbol{\pi}$.

There has been a lot of interest in efficiently computing approximate CCEs in Markov games using V-learning type algorithms [Jin et al., 2021, Wang et al., 2023b, Cui et al., 2023]. However, all these works consider the online setting, where the learner can explore the environment and increasingly gather more relevant data. We only have at our disposal offline corrupted preference data.

5.1 Algorithm

In this section, we propose an offline-learning algorithm for computing approximate CCE in linear Markov games. First, we again assume unilateral coverage on our data (Assumption 2). Given preference data D, we again run `RewardEst` procedure to obtain the reward estimates $\widehat{\boldsymbol{\theta}}$. At this point, different from the previous section, we take another approach at the estimated gap minimization problem. First, for a given joint policy $\boldsymbol{\pi}$, agent i, state s, step h, and actions a and $a^\dagger$, we define the loss $\mathcal{L}_i^s(a^\dagger, a')$ for this stage as

$$\mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_{-i,h}(\cdot|s)}\left[\overline{Q}_{i,h}^{\dagger,\boldsymbol{\pi}_{-i}}(s,a^\dagger,\boldsymbol{a}_{-i}) - \underline{Q}_{i,h}^{\boldsymbol{\pi}}(s,a',\boldsymbol{a}_{-i})\right] ,$$

where the optimistic and pessimistic Q-function estimates are computed using $\widehat{\boldsymbol{\theta}}$. Using this loss function, we can now express the estimated gap minimization problem at stage h as

$$\min_{\boldsymbol{\pi}'_h}\sum_{i\in[n]}\max_{\pi^\dagger_{i,h}}\mathbb{E}_{a_i^\dagger\sim\pi^\dagger_{i,h}(\cdot|s),a_i'\sim\pi'_{i,h}(\cdot|s)}\left[\mathcal{L}_i^s(a^\dagger,a')\right] .$$

Note that such an objective can be framed as a normal-form game at stage h. To solve the stage game, we utilize `OptimisticHedge` [Daskalakis et al., 2021], a no-regret learning algorithm which returns a $\widetilde{O}(1/T_2)$-approximate CCE of the game when run for T_2 iterations. Each player basically solves a $\max-\min$ problem at stage h and updates its policy using a multiplicative weights update style. We run the algorithm for T_2 iterations and return the average joint policy. The pseudo-code for Optimistic Hedge applied to our setting is given in Algorithm 8 (see Appendix E). Once we have computed our joint policy at stage h, we then compute the optimistic and pessimistic values by taking expectations of the Q-function estimates over the newly computed policies. We use these value estimates to run the next iteration $h-1$ of our algorithm. Finally, we return the joint policy $\widetilde{\boldsymbol{\pi}}$ which is a composition of the returned policies from `OptimisticHedge` at each stage h. The full procedure is given in Algorithm 5.

5.2 Theoretical guarantees

In this section, we provide upper bounds on the CCE gap for the output of Algorithm 5.

Theorem 5.1. *Let $\epsilon\in[0,1/2)$, $\lambda\geq\Omega(dH\log(m/\delta)/m)$, and $\delta>0$. Set $\Theta_{\text{Unil}}(\cdot)$ as in Equation (5) and $\Gamma(s,\boldsymbol{a}) = E(d,m,\delta,\epsilon)\cdot\|\phi(s,\boldsymbol{a})\|_{\Lambda_h^{-1}}$. Suppose Assumption 2 is satisfied, PGA is run for T_1 steps with learning rate $\eta_1 = O(1/\sqrt{T_1})$, and `OptimisticHedge` is run for T_2 steps with learning rate $\eta_2 = O(1/(n\log^4 T_2))$. Then, there exist robust subroutines `RobEst`, `TrimmedMLE`, and `RobMean` such that, with probability at least $1-\delta$, the output $\widetilde{\boldsymbol{\pi}}$ of Algorithm 5 with subroutines `RobEst`, `TrimmedMLE`, `RobMean` and `OptimisticHedge`, satisfies*

$$\text{Gap}(\widetilde{\boldsymbol{\pi}}) \leq \widetilde{O}\left(\left(\frac{1}{\sqrt{C_R}}+\frac{1}{\sqrt{C_P}}+\frac{1}{\sqrt{T_1}}\right)\cdot\left(H^{5/2}nd^{3/4}\sqrt{\epsilon}+H^2n\frac{\sqrt{\text{poly}(d)}}{\sqrt{m}}\right)+\frac{Hn^2}{T_2}\right) .$$

Remark 5. *Note that we only incur an additional $O(1/T_2)$ term on the CCE Gap, which comes from applying the no-regret sub-routine `OptimisticHedge`. The benefit of such procedure is that it can be run in quasi-polynomial time in dataset size and feature dimension. The computational complexity of Algorithm 5 is*

$$O\Big((nd)^{\log(\frac{1}{\epsilon})}+(T_1+nH)\cdot(\textit{poly}(m,d,\frac{1}{\epsilon}) + HT_2\max_i|A_i|)\Big).$$

Algorithm 5 Corruption-robust CCE Learning from Human Feedback

Require: Preference dataset D; regularization parameter λ; confidence parameter δ; learning rates η_1,η_2; bonus functions $\Gamma(\cdot,\cdot)$; slackness parameter ν; number of gradient steps T_1; number of optimization steps T_2.

1: Split D into equal D_1 and D_2.
2: Compute $\widehat{\boldsymbol{\theta}} = \texttt{RewardEst}\,(D_1,\epsilon,\delta,\eta_1,\nu,T_1)$.
3: Initialize $\widetilde{\boldsymbol{\pi}}$ uniformly at random and $\underline{V}_{i,H}^{\widetilde{\boldsymbol{\pi}}}(\cdot) = \overline{V}_{i,H}^{\dagger,\widetilde{\boldsymbol{\pi}}_{-i}}(\cdot) = 0$, for all $i\in\{1,\dots,n\}$.
4: **for** $h = H-1,\dots,0$ **do**
5: **for** $i = 1,\dots,n$ **do**
6: Compute $\widehat{R}_{i,h}(\cdot,\cdot) = (\widehat{\theta}_{i,h})^\top\phi(\cdot,\cdot)$ to be the estimated reward.
7: Compute $\left(\underline{Q}_{i,h}^{\widetilde{\boldsymbol{\pi}}}(\cdot,\cdot),\overline{Q}_{i,h}^{\dagger,\widetilde{\boldsymbol{\pi}}_{-i}}(\cdot,\cdot)\right) = \texttt{Rob-Q}\left(D_{2,h},\widetilde{\boldsymbol{\pi}},\epsilon,\widehat{R}_{i,h},\underline{V}_{i,h+1}^{\widetilde{\boldsymbol{\pi}}},\overline{V}_{i,h+1}^{\dagger,\widetilde{\boldsymbol{\pi}}_{-i}},\Gamma\right)$.
8: Compute loss $\mathcal{L}_i^s$, for states $s\in S$.
9: **end for**
10: Compute $\widetilde{\boldsymbol{\pi}}_h(\cdot|s) = \texttt{OptimisticHedge}\,(\mathcal{L}_1^s,\dots,\mathcal{L}_n^s,\eta_2,T_2)$.
11: Set $\underline{V}_{i,h}^{\widetilde{\boldsymbol{\pi}}}(\cdot) = \mathbb{E}_{\boldsymbol{a}\sim\widetilde{\boldsymbol{\pi}}_h}\left[\underline{Q}_{i,h}^{\widetilde{\boldsymbol{\pi}}}(\cdot,\boldsymbol{a})\right]$ and $\overline{V}_{i,h}^{\dagger,\widetilde{\boldsymbol{\pi}}_{-i}}(\cdot) = \max_{a_i}\mathbb{E}_{\boldsymbol{a}_{-i}\sim\widetilde{\boldsymbol{\pi}}_{-i,h}}\left[\overline{Q}_{i,h}^{\dagger,\widetilde{\boldsymbol{\pi}}_{-i}}(\cdot,\boldsymbol{a})\right]$, for $i\in\{1,\dots,n\}$.
12: **end for**
13: **return** $\widetilde{\boldsymbol{\pi}} = (\widetilde{\boldsymbol{\pi}}_0,\dots,\widetilde{\boldsymbol{\pi}}_{H-1})$.

6 RELATED WORK

Reinforcement Learning from Human Feedback (RLHF) RLHF has substantially grown in popularity in the recent years, largely due to LLMs [Ziegler et al., 2019, Nakano et al., 2021, Wu et al., 2021, Ouyang et al., 2022, Stiennon et al., 2020, Glaese et al., 2022, Ramamurthy et al., 2023, Menick et al., 2022, Ganguli et al., 2022, Bai et al., 2022, Gao et al., 2023]. Yet RLHF's utility extends far beyond LLMs, encompassing diverse applications—from game playing [Christiano et al., 2017, Warnell et al., 2018, Knox and Stone, 2008, MacGlashan et al., 2017] to robotic control [Shin et al., 2023, Brown et al., 2019]. Our work is related to recent theoretical studies on (MA)RLHF [Zhan et al., 2023, Zhu et al., 2023, Zhang et al., 2024, Li et al., 2023, Xiong et al., 2023, Nika et al., 2024a]. In particular, we consider data corruption on MARLHF. In the single player setting, Nika et al. [2025] propose a general data-poisoning framework in RLHF, while Mandal et al. [2025] propose robust algorithms trained on ϵ-corrupted data. The latter is the most closely related work to ours. While we share the preference-based model and the data corruption model, our setting is a generalization of the single-agent RLHF setting considered in [Mandal et al., 2025]. This introduces a new layer of complexity: instead of maximizing value functions over single policies, our goal is to minimize the Nash gap over joint policies. This involves a different style of analysis. Algorithmically, our methods diverge in two key ways. First, instead of relying on zeroth-order oracle calls to estimate gradients, we directly approximate the gradient of the Nash gap with respect to each agent's strategy via the biased gradient with respect to a Nash policy. This allows us to also maintain the $O(\sqrt{\epsilon})$ bounds on the gap. Second, we incorporate a quasi-polynomial-time subroutine that computes an approximate coarse correlated equilibrium (CCE) of the induced game.

Corruption-robust Offline Reinforcement Learning (RL) There has been a substantial body of research on adversarial attacks in (MA)RL [Huang et al., 2017, Lin et al., 2017, Wu et al., 2023, Rakhsha et al., 2021, Rangi et al., 2022, Nika et al., 2024b, Ma et al., 2023, Gleave et al., 2020]. Our research relates to a specific type of adversarial attack, namely, data corruption [Mei and Zhu, 2015, Xiao et al., 2015, Rakhsha et al., 2021]. Our focus is on designing robust algorithms trained on corrupted data generated via ϵ-corruption model (a.k.a. strong contamination model [Diakonikolas et al., 2019]). In this line of work, Zhang et al. [2022] first consider corruption-robust RL via linear Markov decision processes (MDP), which is later extended to linear zero-sum Markov games (MG) [Nika et al., 2024b]. Using a different contamination model, Ye et al. [2023] study the problem of corruption-robustness in RL with general function approximation. Our work diverges from the above in that we study strong data corruption in multi-agent reinforcement learning from human feedback, which, due to its dependence on preference data, introduces additional layers of complexity in providing robustness guarantees.

Offline Markov Games (MG) Our work also relates to the literature of learning in MGs [Tian et al., 2021, Vrancx et al., 2008, Littman, 1994, 2001]. We model our underlying environment from which the data is generated as a linear MG [Zhong et al., 2022]. We are interested in approximating notions of optimal joint policies from corrupted preference data. The primary notion of optimality in MGs is that of the Nash equilibrium (NE) [Nash Jr, 1950]. Due to its computational intractability in general-sum MGs, prior work has considered relaxed versions of it such as CCEs [Cui et al., 2023, Zhang et al., 2023, Ma et al., 2023, Song et al., 2021], and designed computationally efficient methods to compute them in the online setting [Jin et al., 2021, Wang et al., 2023b, Cui et al., 2023]. We depart from this line of work and consider the CCE computation problem in the offline setting, where we compute the CCE of each stage game via no-regret methods [Daskalakis et al., 2021].

7 DISCUSSION

In this paper, we studied the problem of data corruption in offline MARLHF. We proposed provable-robust algorithms, both under uniform and unilateral coverage assumptions. Finally, we proposed a computationally efficient algorithm that robustly approximates a coarse correlated equilibrium of the underlying Markov game.

A key technical contribution of our work is a new way to optimize the Nash gap without access to true reward functions or their gradients. Prior single-agent RLHF approaches rely on primal-dual methods or unbiased gradients, which do not extend to general-sum Markov games due to strategic coupling. We instead introduce *a biased but tractable gradient surrogate*: by leveraging the linear structure of the underlying Markov game, we approximate the gradient at a Nash equilibrium using feature expectations induced by behavior policies. Under unilateral coverage, these policies capture the occupancy measures of the equilibrium and its unilateral deviations, so their feature differences act as a proxy for the true gradient direction. Despite the bias, this estimate is accurate enough to guide projected gradient ascent over the reward confidence set, yielding $O(\sqrt{\epsilon})$ robustness guarantees. This idea—leveraging equilibrium structure to construct usable gradient surrogates from corrupted offline preference data—appears to be new and may be of independent interest.

Several interesting directions are worth pursuing. First, it is not clear how one can formulate the data corruption problem in MARLHF with general function approximation, and then how to design robust algorithms in that setting. Second, it would be interesting to address the open question of whether the $O(\sqrt{\epsilon})$ bound under non-uniform coverage is tight. Finally, implementing the proposed algorithms and experimentally testing them on MARL environments is another exciting future direction.

Acknowledgements

The work of Andi Nika and Goran Radanovic was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 467367360.

References

Pranjal Awasthi, Abhimanyu Das, Weihao Kong, and Rajat Sen. Trimmed Maximum Likelihood Estimation for Robust Generalized Linear Model. *NeurIPS*, 2022.

Yuntao Bai et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *CoRR*, abs/2204.05862, 2022.

Tim Baumgärtner, Yang Gao, Dana Alon, and Donald Metzler. Best-of-Venom: Attacking RLHF by Injecting Poisoned Preference Data. *CoRR*, abs/2404.05530, 2024.

Ralph Allan Bradley and Milton E Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4), 1952.

Daniel S. Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations. In *ICML*, 2019.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In *NeurIPS*, 2017.

Qiwen Cui and Simon S Du. Provably Efficient Offline Multi-agent Reinforcement Learning via Strategy-wise Bonus. *NeurIPS*, 2022.

Qiwen Cui, Kaiqing Zhang, and Simon Du. Breaking the Curse of Multiagents in a Large State Space: RL in Markov Games with Independent Linear Function Approximation. In *COLT*, 2023.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal No-regret Learning in General Games. In *NeurIPS*, 2021.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

Ilias Diakonikolas, Daniel M Kane, and Ankit Pensia. Outlier Robust Mean Estimation with Subgaussian Rates via Stability. In *NeurIPS*, 2020.

Ilias Diakonikolas, Samuel B Hopkins, Ankit Pensia, and Stefan Tiegel. Sos Certifiability of Subgaussian Distributions and its Algorithmic Applications. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 1689–1700, 2025.

Yihe Dong, Samuel Hopkins, and Jerry Li. Quantum Entropy Scoring for Fast Robust Mean Estimation and Improved Outlier Detection. *Advances in Neural Information Processing Systems*, 32, 2019.

Dylan J Foster, Noah Golowich, and Sham M Kakade. Hardness of Independent Learning and Sparse Equilibrium Computation in Markov Games. In *ICML*, 2023.

D Ganguli et al. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *CoRR*, abs/2209.07858, 2022.

Leo Gao, John Schulman, and Jacob Hilton. Scaling Laws for Reward Model Overoptimization. In *ICML*, 2023.

Amelia Glaese et al. Improving Alignment of Dialogue Agents via Targeted Human Judgements. *CoRR*, abs/2209.14375, 2022.

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

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

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning: A Simple, Efficient, Decentralized Algorithm for Multiagent RL. *arXiv preprint arXiv:2110.14555*, 2021.

W Bradley Knox and Peter Stone. Tamer: Training an Agent Manually via Evaluative Reinforcement. In *ICDL*, 2008.

Zihao Li, Zhuoran Yang, and Mengdi Wang. Reinforcement Learning with Human Feedback: Learning Dynamic Choices via Pessimism. *arXiv preprint arXiv:2305.18438*, 2023.

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*, 2001.

Shaocong Ma, Ziyi Chen, Shaofeng Zou, and Yi Zhou. Decentralized Robust V-Learning for Solving Markov Games with Model Uncertainty. *JMLR*, 2023.

James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. Interactive Learning from Policy-dependent Human Feedback. In *ICML*, 2017.

Debmalya Mandal, Andi Nika, Parameswaran Kamalaruban, Adish Singla, and Goran Radanović. Corruption Robust Offline Reinforcement Learning with Human Feedback. In *AISTATS*, 2025.

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

Jacob Menick et al. Teaching Language Models to Support Answers with Verified Quotes. *CoRR*, abs/2203.11147, 2022.

Reiichiro Nakano et al. Webgpt: Browser-assisted Question-answering with Human Feedback. *CoRR*, abs/2112.09332, 2021.

John F Nash Jr. Equilibrium Points in n-person Games. *Proceedings of the National Academy of Sciences*, 1950.

Andi Nika, Debmalya Mandal, Parameswaran Kamalaruban, Georgios Tzannetos, Goran Radanović, and Adish Singla. Reward Model Learning vs. Direct Policy Optimization: A Comparative Analysis of Learning from Human Preferences. In *ICML*, 2024a.

Andi Nika, Debmalya Mandal, Adish Singla, and Goran Radanovic. Corruption-Robust Offline Two-player Zero-sum Markov Games. In *AISTATS*, 2024b.

Andi Nika, Jonathan Nöther, Debmalya Mandal, Parameswaran Kamalaruban, Adish Singla, and Goran Radanović. Policy Teaching via Data poisoning in Learning from Human Preferences. In *AISTATS*, 2025.

Long Ouyang et al. Training Language Models to Follow Instructions with Human Feedback. In *NeurIPS*, 2022.

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

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. Is Reinforcement Learning (not) for Natural Language Processing: Benchmarks, Baselines, and Building Blocks for Natural Language Policy Optimization. In *ICLR*, 2023.

Javier Rando and Florian Tramèr. Universal Jailbreak Backdoors from Poisoned Human Feedback. In *ICLR*, 2023.

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.

Jiawen Shi, Yixin Liu, Pan Zhou, and Lichao Sun. Badgpt: Exploring Security Vulnerabilities of ChatGPT via Backdoor Attacks to InstructGPT. *CoRR*, abs/2304.12298, 2023.

Daniel Shin, Anca D. Dragan, and Daniel S. Brown. Benchmarks and Algorithms for Offline Preference-Based Reward Learning. *Transactions of Machine Learning Research*, 2023.

Ziang Song, Song Mei, and Yu Bai. When Can We Learn General-sum Markov Games with a Large Number of Players Sample-Efficiently? *arXiv preprint arXiv:2110.04184*, 2021.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to Summarize with Human Feedback. In *NeurIPS*, 2020.

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

Jiongxiao Wang, Junlin Wu, Muhao Chen, Yevgeniy Vorobeychik, and Chaowei Xiao. On the Exploitability of Reinforcement Learning with Human Feedback for Large Language Models. *CoRR*, abs/2311.09641, 2023a.

Yuanhao Wang, Qinghua Liu, Yu Bai, and Chi Jin. Breaking the Curse of Multiagency: Provably Efficient Decentralized Multi-agent RL with Function Approximation. In *COLT*, 2023b.

Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. Deep TAMER: Interactive Agent Shaping in High-dimensional State Spaces. In *AAAI*, 2018.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively Summarizing Books with Human Feedback. *CoRR*, abs/2109.10862, 2021.

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

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

Wei Xiong, Hanze Dong, Chenlu Ye, Han Zhong, Nan Jiang, and Tong Zhang. Gibbs Sampling from Human Feedback: A Provable KL-constrained Framework for RLHF. *CoRR*, 2023.

Chenlu Ye, Rui Yang, Quanquan Gu, and Tong Zhang. Corruption-robust Offline Reinforcement Learning with General Function Approximation. In *NeurIPS*, 2023.

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

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable Offline Reinforcement

Learning with Human Feedback. *CoRR*, abs/:2305.14816, 2023.

Natalia Zhang, Xinqi Wang, Qiwen Cui, Runlong Zhou, Sham M Kakade, and Simon S Du. Multi-Agent Reinforcement Learning from Human Feedback: Data Coverage and Algorithmic Techniques. *arXiv preprint arXiv:2409.00717*, 2024.

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

Yuheng Zhang, Yu Bai, and Nan Jiang. Offline Learning in Markov Games with General Function Approximation. In *ICML*, 2023.

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.

Banghua Zhu, Michael I. Jordan, and Jiantao Jiao. Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons. In *ICML*, 2023.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning Language Models from Human Preferences. *CoRR*, abs/1909.08593, 2019.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Corruption-robust Offline Multi-agent Reinforcement Learning from Human Feedback

Appendix

Table of Contents

A	Proof of Theorem 3.1	13
B	Proof of Theorem 4.1	18
C	Proof of Theorem 5.1	28
D	Technical Results	30
E	Additional Algorithm Pseudocodes	32

A Proof of Theorem 3.1

In this section, we provide the full proof for Theorem 3.1.

Lemma A.1. *[Lemma A.1 of Mandal et al. [2025]] Let Assumption 1 hold with $\xi_R \geq 5\epsilon$, for some positive constant c, and let*

$$m \geq \Omega\left(\frac{H^{3/2}}{\epsilon^2}\left(d + \log(n/\delta)\right)\right) .$$

Then, for every i, Algorithm 6 returns an estimator $\widetilde{\theta}_i$ such that, with probability at least $1-\delta/2$, satisfies

$$\left\|\widetilde{\theta}_i - \theta_i^*\right\|_2 \leq O\left(\frac{\epsilon}{\xi_R}\exp\left(H + \sqrt{\log(n/2\delta\epsilon)}\right)\right) ,$$

where $\widetilde{\theta}_i$ denotes the Hd-dimensional vector with sub-vectors $\widetilde{\theta}_{i,h}$ for every h.

Proof. This result is an immediate application of Lemma A.1 of [Mandal et al., 2025] to the multi-agent setting by applying the union bound over n agents. □

This upper bound provides us with a provable threshold function for our confidence sets. Next, we will make use of the following result.

Theorem A.1. *[Zhang et al., 2022] Given an ϵ-corrupted dataset $D = \{x_i, y_i\}_{i\in[m]}$, where the clean data is generated as $\widetilde{x}_i \sim \beta$, $\mathbb{P}(\|\widetilde{x}_i\| \leq 1) = 1$, $\widetilde{y}_i = \widetilde{x}_i^\top \omega^* + \zeta_i$, where ζ_i is zero-mean σ^2-variance sub-Gaussian random noise, then a robust least square estimator returns an estimator ω such that:*

- *If $\mathbb{E}_\beta[xx^\top] \succeq \xi I$, then with probability at least $1-\delta/2$, we have*

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

- *With probability at least $1-\delta/2$, we have*

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

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

Applying this to our setting means considering the corrupted Bellman operator samples from our data as signals generated from an unknown underlying distribution. We thus define, for every $i, h, s, \boldsymbol{a} \in [n] \times [H-1] \times S \times A$, the Bellman operator as

$$\mathbb{B}_{i,h} V_{i,h+1}(s, \boldsymbol{a}) = R_{i,h}(s, \boldsymbol{a}) + \sum_{s' \in S} P(s'|s, \boldsymbol{a}) V_{i,h+1}(s') \ .$$

We also define the Bellman operator with respect to the estimated rewards as

$$\underline{\mathbb{B}}_{i,h} \underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s, \boldsymbol{a}) = \underline{R}^{\boldsymbol{\pi}}_{i,h}(s, \boldsymbol{a}) + \sum_{s' \in S} P(s'|s, \boldsymbol{a}) \underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s') \ , \tag{6}$$

and

$$\overline{\mathbb{B}}_{i,h} \overline{V}^{\dagger, \boldsymbol{\pi}_{-i}}_{i,h+1}(s, \boldsymbol{a}) = \underline{R}^{\dagger, \boldsymbol{\pi}_{-i}}_{i,h}(s, \boldsymbol{a}) + \sum_{s' \in S} P(s'|s, \boldsymbol{a}) \overline{V}^{\dagger, \boldsymbol{\pi}_{-i}}_{i,h+1}(s') \ . \tag{7}$$

We then have the following result.

Lemma A.2. *We have, for every tuple $(s_h, \boldsymbol{a}_h, s_{h+1})$ in D,*

$$Var\left(\underline{R}_{i,h}(s_h, \boldsymbol{a}_h) + \underline{V}_{i,h+1}(s_{h+1}) - \underline{\mathbb{B}}_{i,h} \underline{V}_{i,h+1}(s_h, \boldsymbol{a}_h) | s_h, \boldsymbol{a}_h\right) \leq (H\sqrt{d} + \gamma)^2 \ ,$$

and

$$Var\left(\overline{R}_{i,h}(s_h, \boldsymbol{a}_h) + \overline{V}_{i,h+1}(s_{h+1}) - \overline{\mathbb{B}}_{i,h} \overline{V}_{i,h+1}(s_h, \boldsymbol{a}_h) | s_h, \boldsymbol{a}_h\right) \leq (H\sqrt{d} + \gamma)^2 \ .$$

Proof. Note that we have

$$\begin{aligned}
&Var\left(\underline{R}_{i,h}(s_h, \boldsymbol{a}_h) + \underline{V}_{i,h+1}(s_{h+1}) - \underline{\mathbb{B}}_{i,h} \underline{V}_{i,h+1}(s_h, \boldsymbol{a}_h) | s_h, \boldsymbol{a}_h\right) \\
&\quad = \mathbb{E}\left[\left(\underline{R}_{i,h}(s_h, \boldsymbol{a}_h) + \underline{V}_{i,h+1}(s_{h+1}) - \mathbb{E}\left[\underline{R}_{i,h}(s_h, \boldsymbol{a}_h) + \underline{V}_{i,h+1}(s_{h+1})\right]\right)^2\right] \\
&\quad \leq Var(\underline{R}_{i,h}(s_h, \boldsymbol{a}_h)) + Var(\underline{V}_{i,h+1}(s_{h+1})) \\
&\quad \leq (H\sqrt{d} + \gamma)^2 \ ,
\end{aligned}$$

since both H and γ are nonnegative numbers. The proof of the second statement is similar. □

Using the above, we will define the error stated above using short-hand notation for ease of presentation:

$$E_1(d, m, \delta, \epsilon) = c_1(\delta) \cdot \left(\sqrt{\frac{(H\sqrt{d} + \gamma)^2 \text{poly}(d)}{\xi_P^2 m}} + \frac{H\sqrt{d} + \gamma}{\xi_P} \epsilon\right) \ . \tag{8}$$

Next, we prove upper bounds on the maximum and minimum values of estimated reward functions in terms of ground-truth rewards.

Lemma A.3. *With probability at least $1 - \delta/2$, we have*

$$\begin{aligned}
R_{i,h}(s, \boldsymbol{a}) - C_1 \frac{\epsilon \cdot \exp\left(H + \sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} &\leq \underline{R}_{i,h}(s, \boldsymbol{a}) \leq R_{i,h}(s, \boldsymbol{a}) \\
&\leq \overline{R}_{i,h}(s, \boldsymbol{a}) \leq R_{i,h}(s, \boldsymbol{a}) + C_1 \frac{\epsilon \cdot \exp\left(H + \sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} \ .
\end{aligned}$$

Proof. Let $\overline{\theta}_{i,h}$ be the parameter that corresponds to $\overline{R}_{i,h}$ and let $\underline{\theta}_{i,h}$ be defined similarly. Observe that, for every agent i and time-step h, we have

$$\left|\overline{R}_{i,h}(s, \boldsymbol{a}) - R_{i,h}(s, \boldsymbol{a})\right| = \left|\left\langle \phi(s, \boldsymbol{a}), \overline{\theta}_{i,h} \right\rangle - \left\langle \phi(s, \boldsymbol{a}), \theta^*_{i,h} \right\rangle\right|$$

$$
\begin{aligned}
&\leq \left|\left\langle \phi(s,\boldsymbol{a}), \overline{\theta}_{i,h}\right\rangle - \left\langle \phi(s,\boldsymbol{a}), \theta^*_{i,h}\right\rangle\right| \\
&= \left|\left\langle \phi(s,\boldsymbol{a}), \overline{\theta}_{i,h}\right\rangle - \left\langle \phi(s,\boldsymbol{a}), \widetilde{\theta}_{i,h}\right\rangle + \left\langle \phi(s,\boldsymbol{a}), \widetilde{\theta}_{i,h}\right\rangle - \left\langle \phi(s,\boldsymbol{a}), \theta^*_{i,h}\right\rangle\right| \\
&= \left|\left\langle \phi(s,\boldsymbol{a}), \widetilde{\theta}_{i,h} - \overline{\theta}_{i,h}\right\rangle + \left\langle \phi(s,\boldsymbol{a}), \overline{\theta}_{i,h} - \theta^*_{i,h}\right\rangle\right| \\
&\leq \|\phi(s,\boldsymbol{a})\|_2 \left\|\widetilde{\theta}_{i,h} - \overline{\theta}_{i,h}\right\|_2 + \|\phi(s,\boldsymbol{a})\|_2 \left\|\overline{\theta}_{i,h} - \theta^*_{i,h}\right\|_2 \\
&\leq \left\|\widetilde{\theta}_{i,h} - \overline{\theta}_{i,h}\right\|_2 + \left\|\overline{\theta}_{i,h} - \theta^*_{i,h}\right\|_2 \\
&\leq C_1 \frac{\epsilon \cdot \exp\left(H + \sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} ,
\end{aligned}
$$

where the first equality follows by definition, and the fact that the true expected rewards are already in $[-\sqrt{d}, \sqrt{d}]$; we have used Cauchy-Schwarz for the second inequality, the fact that $\|\phi(s,\boldsymbol{a})\|_2 \leq 1$ by assumption, and Lemma A.1 for the final inequality. Similarly, we have

$$
\left|R_{i,h}(s,\boldsymbol{a}) - \underline{R}_{i,h}(s,\boldsymbol{a})\right| \leq C_1 \frac{\epsilon \cdot \exp\left(H + \sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} .
$$

□

Next, we have the following result that provides bounds on the Bellman errors.

Lemma A.4. *With probability at least $1 - \delta/2$, we have, for every $i, h, s, \boldsymbol{a}$ and policy profile $\boldsymbol{\pi}$,*

$$
\begin{aligned}
-E_1(d,m,\delta,\epsilon) &\leq \underline{\mathbb{B}}_{i,h}\underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s,\boldsymbol{a}) - \underline{Q}^{\boldsymbol{\pi}}_{i,h}(s,\boldsymbol{a}) \leq E_1(d,m,\delta,\epsilon) , \\
-E_1(d,m,\delta,\epsilon) &\leq \overline{\mathbb{B}}_{i,h}\overline{V}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h+1}(s,\boldsymbol{a}) - \overline{Q}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(s,\boldsymbol{a}) \leq E_1(d,m,\delta,\epsilon) .
\end{aligned}
$$

Proof. First, as noted in [Zhong et al., 2022], in linear MDPs, the value functions are also linear in features. Thus, denoting by $\underline{\omega}^{\boldsymbol{\pi},*}_{i,h}$ the parameter of the Bellman transform of $\underline{V}^{\boldsymbol{\pi}}_{i,h}$ and defining $\overline{\omega}^{\dagger,\boldsymbol{\pi}_{-i},*}_{i,h}$ similarly, we have

$$
\begin{aligned}
\left|\phi(s,\boldsymbol{a})^\top \underline{\omega}^{\boldsymbol{\pi},*}_{i,h} - \underline{\mathbb{B}}_{i,h}\underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s,\boldsymbol{a})\right| &= \left\langle \phi(s,\boldsymbol{a}), \underline{\omega}^{\boldsymbol{\pi},*}_{i,h} - \underline{\omega}^{\boldsymbol{\pi}}_{i,h}\right\rangle \\
&\leq \|\phi(s,\boldsymbol{a})\|_2 \left\|\underline{\omega}^{\boldsymbol{\pi},*}_{i,h} - \underline{\omega}^{\boldsymbol{\pi}}_{i,h}\right\|_2 \\
&\leq E_1(d,m,\delta,\epsilon) ,
\end{aligned}
$$

where the penultimate step uses Cauchy-Schwarz and the final step uses the feature norm assumption and Theorem A.1. □

Next, we prove a similar result for the estimated value functions and best responses.

Lemma A.5. *Under the event of Lemma A.4, we have, for every agent i, state s, step h and policy $\boldsymbol{\pi}$:*

$$
\begin{aligned}
\underline{V}^{\boldsymbol{\pi}}_{i,h}(s) &\leq V^{\boldsymbol{\pi}}_{i,h}(s) + E_1(d,m,\delta,\epsilon) + C_1 \frac{\epsilon \cdot \exp\left(H + \sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R}, \quad \textit{and} \\
\overline{V}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(s) &\geq V^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(s) - E_1(d,m,\delta,\epsilon) - C_1 \frac{\epsilon \cdot \exp\left(H + \sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} .
\end{aligned}
$$

Proof. We prove the result by induction. Let $\underline{V}^{s}_{i,H}(\boldsymbol{\pi}) = 0$. The result holds for step H. Suppose the result holds for step $h+1$. Then, for step h we have

$$
\begin{aligned}
&\underline{V}^{\boldsymbol{\pi}}_{i,h}(s) = \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h}\left[\underline{Q}^{\boldsymbol{\pi}}_{i,h}(s,\boldsymbol{a})\right] && (9) \\
&\leq \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h}\left[\underline{\mathbb{B}}_{i,h}\underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s,\boldsymbol{a})\right] + E_1(d,m,\delta,\epsilon) && (10) \\
&\leq \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h}\left[\underline{\mathbb{B}}_{i,h}V^{\boldsymbol{\pi}}_{i,h+1}(s,\boldsymbol{a})\right] + E_1(d,m,\delta,\epsilon) && (11)
\end{aligned}
$$

$$= \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h}\left[\underline{R}_{i,h}(s,\boldsymbol{a}) + \sum_{s'} P(s'|s,\boldsymbol{a})V^{\boldsymbol{\pi}}_{i,h+1}(s')\right] + E_1(d,m,\delta,\epsilon)$$

$$\leq \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h}\left[R_{i,h}(s,\boldsymbol{a}) + \sum_{s'} P(s'|s,\boldsymbol{a})V^{\boldsymbol{\pi}}_{i,h+1}(s')\right] + C_1\frac{\epsilon\cdot\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} + E_1(d,m,\delta,\epsilon) \tag{12}$$

$$= \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h}\left[\mathbb{B}_{i,h}V^{\boldsymbol{\pi}}_{i,h+1}(s,\boldsymbol{a})\right] + C_1\frac{\epsilon\cdot\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} + E_1(d,m,\delta,\epsilon)$$

$$= V^{\boldsymbol{\pi}}_{i,h}(s) + C_1\frac{\epsilon\cdot\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} + E_1(d,m,\delta,\epsilon) \tag{13}$$

where Equation (9) follows by definition; Equation (10) follows from Lemma A.4; Equation (11) follows by the inductive assumption; Equation (12) follows from Lemma A.3; Equation (13) follows by definition. Similarly,

$$\overline{V}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(s) = \max_{a_i\in A_i}\mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_{-i,h}(\cdot|s)}\left[\overline{Q}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(s,\boldsymbol{a})\right] \tag{14}$$

$$\geq \max_{a_i\in A_i}\mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_{-i,h}(\cdot|s)}\left[\overline{\mathbb{B}}_{i,h}\overline{V}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h+1}(s,\boldsymbol{a})\right] - E_1(d,m,\delta,\epsilon) \tag{15}$$

$$\geq \max_{a_i\in A_i}\mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_{-i,h}(\cdot|s)}\left[\overline{\mathbb{B}}_{i,h}V^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h+1}(s,\boldsymbol{a})\right] - E_1(d,m,\delta,\epsilon) \tag{16}$$

$$\geq \max_{a_i\in A_i}\mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_{-i,h}(\cdot|s)}\left[\mathbb{B}_{i,h}V^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h+1}(s,\boldsymbol{a})\right] - C_1\frac{\epsilon\cdot\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} - E_1(d,m,\delta,\epsilon) \tag{17}$$

$$= V^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(s) - C_1\frac{\epsilon\cdot\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} - E_1(d,m,\delta,\epsilon)\ , \tag{18}$$

where again Equation (14) follows by definition; Equation (15) follows from Lemma A.4; Equation (16) follows from the linearity and monotonicity of $\overline{\mathbb{B}}_{i,h}$; Equation (17) follows by the inductive assumption and Lemma A.3, and (18) follows by definition. □

Next, we provide bounds on the difference between the estimated and true expected returns.

Lemma A.6. *Under the event of Lemma A.4 we have, for any $\boldsymbol{\pi}\in\Pi^{\text{PP}}$,*

$$V^{\boldsymbol{\pi}}_{i,0}(s_0) - \underline{V}^{\boldsymbol{\pi}}_{i,0}(s_0) \leq HC_1\frac{\epsilon\cdot\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} + HE_1(d,m,\delta,\epsilon)\ ,$$

$$V^{\boldsymbol{\pi}}_{i,0}(s_0) - \overline{V}^{\boldsymbol{\pi}}_{i,0}(s_0) \leq HC_1\frac{\epsilon\cdot\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} + HE_1(d,m,\delta,\epsilon)\ .$$

Proof. Note that we have

$$V^{\boldsymbol{\pi}}_{i,0}(s_0) - \underline{V}^{\boldsymbol{\pi}}_{i,0}(s_0) = \mathbb{E}_{\boldsymbol{a}_0\sim\boldsymbol{\pi}_0(\cdot|s_0)}\left[Q^{\boldsymbol{\pi}}_{i,0}(s_0,\boldsymbol{a}_0)\right] - \mathbb{E}_{\boldsymbol{a}_0\sim\boldsymbol{\pi}_0(\cdot|s_0)}\left[\underline{Q}^{\boldsymbol{\pi}}_{i,0}(s_0,\boldsymbol{a}_0)\right]$$

$$\leq \mathbb{E}_{\boldsymbol{a}_0\sim\boldsymbol{\pi}_0(\cdot|s_0)}\left[Q^{\boldsymbol{\pi}}_{i,0}(s_0,\boldsymbol{a}_0) - \underline{\mathbb{B}}_{i,0}\underline{V}^{\boldsymbol{\pi}}_{i,1}(s_0,\boldsymbol{a}_0) + E_1(d,m,\delta,\epsilon)\right] \tag{19}$$

$$= \mathbb{E}_{\boldsymbol{a}_0\sim\boldsymbol{\pi}_0(\cdot|s_0)}\left[\mathbb{B}_{i,0}V^{\boldsymbol{\pi}}_{i,1}(s_0,\boldsymbol{a}_0) - \underline{\mathbb{B}}_{i,0}\underline{V}^{\boldsymbol{\pi}}_{i,1}(s_0,\boldsymbol{a}_0) + E_1(d,m,\delta,\epsilon)\right] \tag{20}$$

$$= \mathbb{E}_{\boldsymbol{a}_0\sim\boldsymbol{\pi}_0(\cdot|s_0)}\left[R_{i,0}(s_0,\boldsymbol{a}_0) - \underline{R}_{i,0}(s_0,\boldsymbol{a}_0) + \mathbb{E}_{s_1\sim P(\cdot|s_0,\boldsymbol{a}_0)}\left[V_{i,1}(s_1) - \underline{V}_{i,1}(s_1)\right] + E_1(d,m,\delta,\epsilon)\right] \tag{21}$$

$$\leq \mathbb{E}_{\boldsymbol{a}_0\sim\boldsymbol{\pi}_0(\cdot|s_0)}\left[\mathbb{E}_{s_1\sim P(\cdot|s_0,\boldsymbol{a}_0)}\left[V_{i,1}(s_1) - \underline{V}_{i,1}(s_1)\right] + C_1\frac{\epsilon\cdot\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} + E_1(d,m,\delta,\epsilon)\right] \tag{22}$$

$$= C_1\frac{\epsilon\cdot\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} + E_1(d,m,\delta,\epsilon) + \mathbb{E}_{\boldsymbol{a}_0\sim\boldsymbol{\pi}_0(\cdot|s_0),s_1\sim P(\cdot|s_0,\boldsymbol{a}_0)}\left[V^{\boldsymbol{\pi}}_{i,1}(s_1) - \underline{V}^{\boldsymbol{\pi}}_{i,1}(s_1)\right]$$

$$= \dots$$

$$\leq H\left(C_1\frac{\epsilon\cdot\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R}+E_1(d,m,\delta,\epsilon)\right), \tag{23}$$

where Equation (19) follows from Lemma A.4; Equation (20) follows from the fact that the true action-value function has zero error with respect to the Bellman operator; Equation (21) follows from expanding the Bellman operator for both value functions; Equation (22) follows from Lemma A.3 and, finally, Equation (23) follows from applying the same bounds H steps.

Following similar arguments, for the best response value gap, we have:

$$\begin{aligned}
&V_{i,0}^{\dagger,\boldsymbol{\pi}_{-i}}(s_0)-\overline{V}_{i,0}^{\dagger,\boldsymbol{\pi}_{-i}}(s_0)=\max_{a_{i,0}\in A_i}\mathbb{E}_{\boldsymbol{a}_{-i,0}\sim\boldsymbol{\pi}_{-i,0}(\cdot|s_0)}\left[Q_{i,0}^{\dagger,\boldsymbol{\pi}_{-i}}(s_0,\boldsymbol{a}_0)-\overline{Q}_{i,0}^{\dagger,\boldsymbol{\pi}_{-i}}(s_0,\boldsymbol{a}_0)\right]\\
&\leq\max_{a_{i,0}\in A_i}\mathbb{E}_{\boldsymbol{a}_{-i,0}\sim\boldsymbol{\pi}_{-i,0}(\cdot|s_0)}\left[\mathbb{B}_{i,0}V_{i,1}^{\dagger,\boldsymbol{\pi}_{-i}}(s_0,\boldsymbol{a}_0)-\overline{\mathbb{B}}_{i,0}\overline{V}_{i,1}^{\dagger,\boldsymbol{\pi}_{-i}}(s_0,\boldsymbol{a}_0)+E_1(d,m,\delta,\epsilon)\right]\\
&\leq C_1\frac{\epsilon\cdot\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R}+E_1(d,m,\delta,\epsilon)+\max_{a_{i,0}\in A_i}\mathbb{E}_{\boldsymbol{a}_{-i,0}\sim\boldsymbol{\pi}_{-i,0}(\cdot|s_0),s_1\sim P(\cdot|s_0,\boldsymbol{a}_0)}\left[V_{i,1}^{\dagger,\boldsymbol{\pi}_{-i}}(s_1)-\overline{V}_{i,1}^{\dagger,\boldsymbol{\pi}_{-i}}(s_1)\right]\\
&\leq H\left(C_1\frac{\epsilon\cdot\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R}+E_1(d,m,\delta,\epsilon)\right).
\end{aligned}$$

□

Next, we state a result that provides an upper bound on the Nash gap in terms of the estimated value functions.

Lemma A.7. *Under the event of Lemma A.5, we have, for some $C_1>0$,*

$$\text{Gap}(\widetilde{\boldsymbol{\pi}})\leq\min_{\boldsymbol{\pi}}\sum_{i\in[n]}\overline{V}_{i,0}^{\dagger,\boldsymbol{\pi}_{-i}}(s_0)-\underline{V}_{i,0}^{\boldsymbol{\pi}}(s_0)+2nE_1(d,m,\delta,\epsilon)+2nC_1\frac{\epsilon\cdot\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R}.$$

Proof. Note that, by definition of the Nash gap and Lemma A.5, we have

$$\begin{aligned}
\text{Gap}(\widetilde{\boldsymbol{\pi}})&=\sum_{i\in[n]}V_{i,0}^{\dagger,\widetilde{\boldsymbol{\pi}}_{-i}}(s_0)-V_{i,0}^{\widetilde{\boldsymbol{\pi}}}(s_0)\\
&\leq\sum_{i\in[n]}\overline{V}_{i,0}^{\dagger,\widetilde{\boldsymbol{\pi}}_{-i}}(s_0)-\underline{V}_{i,0}^{\widetilde{\boldsymbol{\pi}}}(s_0)+2nE_1(d,m,\delta,\epsilon)+2nC_1\frac{\epsilon\cdot\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R}\\
&=\min_{\boldsymbol{\pi}}\sum_{i\in[n]}\overline{V}_{i,0}^{\dagger,\boldsymbol{\pi}_{-i}}(s_0)-\underline{V}_{i,0}^{\boldsymbol{\pi}}(s_0)+2nE_1(d,m,\delta,\epsilon)+2nC_1\frac{\epsilon\cdot\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R},
\end{aligned}$$

where the last step uses the fact that $\widetilde{\boldsymbol{\pi}}$ minimizes the quantity within the summation, as defined in Algorithm 1. □

Now we are ready to finalize the proof of the main theorem of Section 3. We restate it for convenience.

Theorem A.2. *Let $\epsilon\in[0,1/2)$, $\delta>0$ and $\Gamma(\cdot,\cdot)=0$. Furthermore, assume that $m\geq\Omega((H^{3/2}/\epsilon^2)(d+\log(n/\delta)))$. Then, under Assumption 1 with $\xi_R\geq 5\epsilon$, for some positive constant c, there exist robust algorithms* `TrimmedMLE` *and* `RobEst` *such that, with probability at least $1-\delta$, the output $\widetilde{\boldsymbol{\pi}}$ of Objective (4) satisfies*

$$\text{Gap}(\widetilde{\boldsymbol{\pi}})\leq O\left(Hn\left(\frac{\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R}+\frac{H\sqrt{d}+\gamma}{\xi_P}\right)\cdot\epsilon+Hn\sqrt{\frac{(H\sqrt{d}+\gamma)^2\text{poly}(d)}{\xi_P^2 m}}\right).$$

Proof. Let $\boldsymbol{\pi}^*$ be a Nash equilibrium. We have

$$\text{Gap}(\widetilde{\boldsymbol{\pi}})\leq\min_{\boldsymbol{\pi}}\sum_{i\in[n]}\overline{V}_{i,0}^{\dagger,\boldsymbol{\pi}_{-i}}(s_0)-\underline{V}_{i,0}^{\boldsymbol{\pi}}(s_0)+2nE_1(d,m,\delta,\epsilon)+2n\epsilon C_1\frac{\exp\left(H+\sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} \tag{24}$$

$$\leq \sum_{i\in[n]} \overline{V}_{i,0}^{\dagger,\boldsymbol{\pi}^{*-i}}(s_0) - V_{i,0}^{\dagger,\boldsymbol{\pi}^{*-i}}(s_0) + \sum_{i\in[n]} V_{i,0}^{\boldsymbol{\pi}^*}(s_0) - \underline{V}_{i,0}^{\boldsymbol{\pi}^*}(s_0) + \sum_{i\in[n]} V_{i,0}^{\dagger,\boldsymbol{\pi}^{*-i}}(s_0) - V_{i,0}^{\boldsymbol{\pi}^*}(s_0)$$

$$+ 2nE_1(d,m,\delta,\epsilon) + 2n\epsilon C_1 \frac{\exp\left(H + \sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} \tag{25}$$

$$\leq 2Hn \cdot \left(2\epsilon C_1 \frac{\exp\left(H + \sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} + 2E_1(d,m,\delta,\epsilon) \right) + \sum_{i\in[n]} V_{i,0}^{\dagger,\boldsymbol{\pi}^{*-i}}(s_0) - V_{i,0}^{\boldsymbol{\pi}^*}(s_0) \tag{26}$$

$$\leq 2Hn \cdot \left(2\epsilon C_1 \frac{\exp\left(H + \sqrt{\log(n/2\delta\epsilon)}\right)}{\xi_R} + 2E_1(d,m,\delta,\epsilon) \right) , \tag{27}$$

where Equation (24) follows from Lemma A.7; for Equation (25), we pick a Nash equilibrium $\boldsymbol{\pi}^*$ and use the fact that $\widetilde{\boldsymbol{\pi}}$ is the minimizer of the estimated gap; Equation (26) follows from Lemma A.6 and, finally, Equation (27) follows from the fact that $\boldsymbol{\pi}^*$ is a Nash equilibrium and thus any unilateral deviation yields a smaller value for agent i. □

B Proof of Theorem 4.1

In this section, we provide the full proof for Theorem 4.1. First, let us define state-action occupancy measures. Given policy $\boldsymbol{\pi}$, we define $d^{\boldsymbol{\pi}}(s) = (1/H)\sum_{h=0}^{H-1} \mathbb{P}(s_h = s|s_0, \boldsymbol{\pi})$ and $d^{\boldsymbol{\pi}}(s,a) = (1/H)\sum_{h=0}^{H-1} \mathbb{P}(s_h = s, a_h = a|s_0, \boldsymbol{\pi})$. We will use the notation $\tau \sim d^{\boldsymbol{\pi}}$ to imply that trajectory τ has been sampled according to $\boldsymbol{\pi}$ and the transition kernel of the underlying game. We begin by stating upper bound results that are used to specify the structure of our confidence set for this setting.

Lemma B.1. *Let $\mathbb{P}(o|\tau,\tau',\theta) = 1/(1+\exp(-o\cdot\theta^\top(\phi(\tau)-\phi(\tau'))$. Then, given $\delta > 0$, with probability at least $1-\delta$, we have for any agent i:*

$$\frac{2}{m}\sum_{(\tau,\tau',o)\in D} \log\left(\frac{\mathbb{P}(o|\tau,\tau',\widetilde{\theta}_i)}{\mathbb{P}(o|\tau,\tau',\theta_i^*)}\right) \leq 6H\sqrt{d}\epsilon + c\frac{d}{m}\log\left(\frac{Hmn}{\delta}\right) ,$$

where $\widetilde{\theta}_i$ is the output of Algorithm 6.

Proof. This is an immediate application of Lemma 4.2 of [Mandal et al., 2025] by applying the union bound for all agents. □

Note that the above bounds characterize the confidence set that we used throughout Section 4. As a robust estimation technique, we again make use of `RobEst` based on the second part of Theorem A.1, which does not require uniform coverage. Lemma A.2 and Theorem A.1 give us the following guarantee for the output of the robust estimate $\widetilde{\omega}$, where, without loss of generality, we use the behavior policy $\boldsymbol{\mu}$:

$$\mathbb{E}_{\boldsymbol{\mu}}\left[\left\|\phi(s,\boldsymbol{a})^\top(\omega^* - \widetilde{\omega})\right\|_2\right] \leq c_2(\delta)\cdot\sqrt{\frac{(H\sqrt{d}+\gamma)^2\text{poly}(d)}{m} + (H\sqrt{d}+\gamma)^2\epsilon} . \tag{28}$$

Note that, using the above, we can equivalently write

$$\|\omega^* - \widetilde{\omega}\|^2_{\Sigma_{\boldsymbol{\mu}}(h)} \leq c_2(\delta)\left(\frac{(H\sqrt{d}+\gamma)^2\text{poly}(d)}{m} + (H\sqrt{d}+\gamma)^2\epsilon\right) ,$$

which implies that

$$\|\omega^* - \widetilde{\omega}\|^2_{\Sigma_{\boldsymbol{\mu}}(h)+(2\epsilon+\lambda)I} \leq c_2(\delta)\left(\frac{(H\sqrt{d}+\gamma)^2\text{poly}(d)}{m} + (H\sqrt{d}+\gamma)^2\epsilon + (2\epsilon+\lambda)H\sqrt{d}\right) ,$$

since $\|\omega^*\| \leq H\sqrt{d}$ (Lemma A.1 of [Zhang et al., 2022]). Let us define

$$E(d,m,\delta,\epsilon) := \sqrt{c_2(\delta)\left(\frac{(H\sqrt{d}+\gamma)^2\text{poly}(d)}{m} + (H\sqrt{d}+\gamma)^2\epsilon + (2\epsilon+\lambda)H\sqrt{d}\right)} . \tag{29}$$

This term will be useful in defining our bonus for this section. Recall that, for each step h, we have defined the scaled sample covariance matrix with respect to the corrupted data as follows:

$$\Lambda_h = \frac{3}{5}\left(\frac{1}{m}\sum_{i=1}^{m}\left(\phi(s_h,\boldsymbol{a}_h)\phi(s_h,\boldsymbol{a}_h)^\top\right) + (\epsilon+\lambda)I\right) , \tag{30}$$

while the bonus has been defined as

$$\Gamma(s,\boldsymbol{a}) = E(d,m,\delta,\epsilon)\cdot\|\phi(s,\boldsymbol{a})\|_{\Lambda_h^{-1}} . \tag{31}$$

In the absence of bounds on the norm of the parameter, we cannot bound the difference in rewards directly, as we did in Lemma A.3. Thus, we will need to follow another approach. First, similar to the previous section, we have the following result.

Lemma B.2. *Let $\lambda \geq \Omega(dH\log(m/\delta))$ and Γ be defined as in Equation (31). Then, with probability at least $1-\delta/2$ we have, for every $i,h,s,\boldsymbol{a}$, and policy $\boldsymbol{\pi}$,*

$$\begin{aligned}
0 &\leq \underline{\mathbb{B}}_{i,h}\underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s,\boldsymbol{a}) - \underline{Q}^{\boldsymbol{\pi}}_{i,h}(s,\boldsymbol{a}) \leq 2\Gamma(s,\boldsymbol{a}) ,\\
0 &\geq \overline{\mathbb{B}}_{i,h}\overline{V}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h+1}(s,\boldsymbol{a}) - \overline{Q}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(s,\boldsymbol{a}) \geq -2\Gamma(s,\boldsymbol{a}) .
\end{aligned}$$

Proof. Following a similar approach as the proof of Lemma A.4, we have

$$\begin{aligned}
\left|\phi(s,\boldsymbol{a})\top\underline{\omega}^*_{i,h} - \underline{\mathbb{B}}_{i,h}\underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s,\boldsymbol{a})\right| &= \left|\left\langle\phi(s,\boldsymbol{a}),\underline{\omega}^*_{i,h} - \underline{\omega}^{\boldsymbol{\pi}}_{i,h}\right\rangle\right|\\
&\leq \left\|\underline{\omega}^*_{i,h} - \underline{\omega}^{\boldsymbol{\pi}}_{i,h}\right\|_{\Sigma_{\boldsymbol{\mu}}(h)+(2\epsilon+\lambda)I}\|\phi(s,\boldsymbol{a})\|_{(\Sigma_{\boldsymbol{\mu}}(h)+(2\epsilon+\lambda)I)^{-1}}\\
&\leq E(d,m,\delta,\epsilon)\cdot\|\phi(s,\boldsymbol{a})\|_{(\Sigma_{\boldsymbol{\mu}}(h)+(2\epsilon+\lambda)I)^{-1}}\\
&\leq E(d,m,\delta,\epsilon)\cdot\|\phi(s,\boldsymbol{a})\|_{(\Sigma_{\boldsymbol{\mu}}(h)+(2\epsilon+\lambda)I)^{-1}}\\
&\leq E(d,m,\delta,\epsilon)\cdot\|\phi(s,\boldsymbol{a})\|_{\Lambda_h^{-1}}\\
&= \Gamma(s,\boldsymbol{a}) ,
\end{aligned}$$

where the penultimate inequality uses the fact that $\left\|\underline{\omega}^*_{i,h}\right\|_2 \leq H\sqrt{d}$ and the final inequality follows from the following observation:

$$\begin{aligned}
\Lambda_h &= \frac{3}{5}\left(\frac{1}{m}\sum_{i=1}^{m}\phi(s_h,\boldsymbol{a}_h)\phi(s_h,\boldsymbol{a}_h)^\top + (\epsilon+\lambda)I\right)\\
&\preceq \frac{3}{5}\left(\frac{1}{m}\sum_{i=1}^{m}\phi(\widetilde{s}_h,\widetilde{\boldsymbol{a}}_h)\phi(\widetilde{s}_h,\widetilde{\boldsymbol{a}}_h)^\top + (2\epsilon+\lambda)I\right)\\
&\preceq \Sigma_{\boldsymbol{\mu}}(h) + (2\epsilon+\lambda)I ,
\end{aligned}$$

where the second step uses the fact that $\|\phi(s,\boldsymbol{a})\|_2 \leq 1$ and that only $\epsilon\cdot m$ samples are corrupted, while the last step uses Lemma D.1 and the fact that $m(2\epsilon+\lambda) \geq \Omega(d\log(m/\delta))$, due to our choice of λ and the fact that $\epsilon \geq 0$.

Thus, we obtained that

$$\underline{\mathbb{B}}_{i,h}\underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s,\boldsymbol{a}) - \Gamma(s,\boldsymbol{a}) \leq \phi(s,\boldsymbol{a})\top\underline{\omega}^*_{i,h} \leq \underline{\mathbb{B}}_{i,h}\underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s,\boldsymbol{a}) + \Gamma(s,\boldsymbol{a}) ,$$

which, by subtracting $\Gamma(s,\boldsymbol{a})$ from all sides, further implies that

$$\underline{\mathbb{B}}_{i,h}\underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s,\boldsymbol{a}) - 2\Gamma(s,\boldsymbol{a}) \leq \phi(s,\boldsymbol{a})\top\underline{\omega}^*_{i,h} - \Gamma(s,\boldsymbol{a}) \leq \underline{\mathbb{B}}_{i,h}\underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s,\boldsymbol{a}) .$$

Now, since $\underline{\mathbb{B}}_{i,h}\underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s,\boldsymbol{a}) \in [-(H-h)\sqrt{d},(H-h)\sqrt{d}]$, and since the clipping operator is monotone, we have

$$\begin{aligned}
\underline{\mathbb{B}}_{i,h}\underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s,\boldsymbol{a}) - 2\Gamma(s,\boldsymbol{a}) &\leq \mathrm{Clip}_{[-(H-h)\sqrt{d},(H-h)\sqrt{d}]}\left(\underline{\mathbb{B}}_{i,h}\underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s,\boldsymbol{a}) - 2\Gamma(s,\boldsymbol{a})\right)\\
&\leq \mathrm{Clip}_{[-(H-h)\sqrt{d},(H-h)\sqrt{d}]}\left(\phi(s,\boldsymbol{a})\top\underline{\omega}^*_{i,h} - \Gamma(s,\boldsymbol{a})\right)
\end{aligned}$$

$$= \underline{Q}^{\boldsymbol{\pi}}_{i,h}(s, \boldsymbol{a})$$
$$\leq \underline{\mathbb{B}}_{i,h} \underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s, \boldsymbol{a}) \ .$$

This finally implies that

$$0 \leq \underline{\mathbb{B}}_{i,h} \underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s, \boldsymbol{a}) - \underline{Q}^{\boldsymbol{\pi}}_{i,h}(s, \boldsymbol{a}) \leq 2\Gamma(s, \boldsymbol{a}) \ .$$

For the optimistic estimates, we argue in a symmetrical fashion, thus, we omit the proof. □

Next, we will state a result which is the analogue of Lemma A.5 for this section. We define $\underline{V}^{\boldsymbol{\pi}}_{i,h}(s, \widehat{\theta}_i)$ and $\overline{V}^{\dagger, \boldsymbol{\pi}_{-i}}_{i,h}(s, \widehat{\theta}_i)$ to be the lower and upper estimates of the value functions of given policy $\boldsymbol{\pi}$ with respect to parameter $\widehat{\theta}_i$. We use similar notation for the Q-function estimates.

Lemma B.3. *Let $\widehat{\theta}_i \in \Theta_{\text{Unil}}(\widetilde{\theta}_i)$ be a parameter used by the robust subroutine. Then, under the event of Lemma B.2, we have, for every agent i, state s, step h and policy $\boldsymbol{\pi}$:*

$$\underline{V}^{\boldsymbol{\pi}}_{i,h}(s, \widehat{\theta}_i) \leq V^{\boldsymbol{\pi}}_{i,h}(s, \widehat{\theta}_i) \ , \quad \textit{and} \quad \overline{V}^{\dagger, \boldsymbol{\pi}_{-i}}_{i,h}(s, \widehat{\theta}_i) \geq V^{\dagger, \boldsymbol{\pi}_{-i}}_{i,h}(s, \widehat{\theta}_i) \ .$$

Proof. Similar to Lemma A.5, we again apply induction. Note that the result holds for step H where all value estimates are 0, since the bound term is non-negative. Suppose the statement holds for step $h+1$. Then, for step h, we have

$$\begin{aligned}
\underline{V}^{\boldsymbol{\pi}}_{i,h}(s, \widehat{\theta}_i) &= \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_h} \left[\underline{Q}^{\boldsymbol{\pi}}_{i,h}(s, \boldsymbol{a}, \widehat{\theta}_i) \right] \\
&\leq \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_h} \left[\underline{\mathbb{B}}_{i,h} \underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s, \boldsymbol{a}, \widehat{\theta}_i) \right] \\
&\leq \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_h} \left[\underline{\mathbb{B}}_{i,h} V^{\boldsymbol{\pi}}_{i,h+1}(s, \boldsymbol{a}, \widehat{\theta}_i) \right] \\
&= V^{\boldsymbol{\pi}}_{i,h}(s, \widehat{\theta}_i) \ .
\end{aligned}$$

For $\overline{V}^{\dagger, \boldsymbol{\pi}_{-i}}_{i,h}(s, \widehat{\theta}_i)$ we have

$$\begin{aligned}
\overline{V}^{\dagger, \boldsymbol{\pi}_{-i}}_{i,h}(s, \widehat{\theta}_i) &= \max_{a_i \in A_i} \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}_{-i,h}(\cdot|s)} \left[\overline{Q}^{\dagger, \boldsymbol{\pi}_{-i}}_{i,h}(s, \boldsymbol{a}, \widehat{\theta}_i) \right] \\
&\geq \max_{a_i \in A_i} \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}_{-i,h}(\cdot|s)} \left[\overline{\mathbb{B}}_{i,h} \overline{V}^{\dagger, \boldsymbol{\pi}_{-i}}_{i,h+1}(s, \boldsymbol{a}, \widehat{\theta}_i) \right] \\
&\geq \max_{a_i \in A_i} \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}_{-i,h}(\cdot|s)} \left[\overline{\mathbb{B}}_{i,h} V^{\dagger, \boldsymbol{\pi}_{-i}}_{i,h+1}(s, \boldsymbol{a}, \widehat{\theta}_i) \right] \\
&= V^{\dagger, \boldsymbol{\pi}_{-i}}_{i,h}(s, \widehat{\theta}_i) \ .
\end{aligned}$$

□

Next, we prove an upper bound on the expected sum of bonuses.

Lemma B.4. *Let $\boldsymbol{\pi}^*$ be a Nash equilibrium which is covered by D. Then, for every agent i, we have*

$$\mathbb{E}_{\boldsymbol{\pi}^*} \left[\sum_{h=0}^{H-1} \Gamma(s_h, \boldsymbol{a}_h) \right] \leq H \cdot E(d, m, \delta, \epsilon) \cdot \sqrt{\frac{5d}{C_P}} \ .$$

Proof. Using the definition of the bonus in Equation (31), we have

$$\mathbb{E}_{\boldsymbol{\pi}^*} \left[\sum_{h=0}^{H-1} \Gamma(s_h, \boldsymbol{a}_h) \right] = E(d, m, \delta, \epsilon) \cdot \mathbb{E}_{\boldsymbol{\pi}^*} \left[\sum_{h=0}^{H-1} \|\phi(s_h, \boldsymbol{a}_h)\|_{\Lambda_h^{-1}} \right] \ .$$

We bound the last factor on the right-hand side of the equation above. Using the definition of Λ_h in Equation (30), we have, for every $h \in [H-1]$:

$$\mathbb{E}_{\boldsymbol{\pi}^*} \left[\|\phi(s_h, \boldsymbol{a}_h)\|_{\Lambda_h^{-1}} \right] \leq \mathbb{E}_{\boldsymbol{\pi}^*} \left[\|\phi(s_h, \boldsymbol{a}_h)\|_{((\Sigma_{\boldsymbol{\mu}}(h) + \lambda I))^{-1}} \right] \tag{32}$$

$$
\begin{aligned}
&\leq \mathbb{E}_{\boldsymbol{\pi}^*}\left[\sqrt{\phi(s_h,\boldsymbol{a}_h)^\top\left((\Sigma_{\boldsymbol{\mu}}(h)+\lambda I)\right)^{-1}\phi(s_h,\boldsymbol{a}_h)}\right] \\
&\leq \sqrt{\mathbb{E}_{\boldsymbol{\pi}^*}\left[\phi(s_h,\boldsymbol{a}_h)^\top\left((\Sigma_{\boldsymbol{\mu}}(h)+\lambda I)\right)^{-1}\phi(s_h,\boldsymbol{a}_h)\right]} && (33)\\
&= \sqrt{Tr\left(\mathbb{E}_{\boldsymbol{\pi}^*}\left[\phi(s_h,\boldsymbol{a}_h)\phi(s_h,\boldsymbol{a}_h)^\top\right]\left((\Sigma_{\boldsymbol{\mu}}(h)+\lambda I)\right)^{-1}\right)} && (34)\\
&\leq \sqrt{\frac{1}{C_P}}\sqrt{Tr\left(\mathbb{E}_{\boldsymbol{\mu}_h}\left[\phi(s_h,\boldsymbol{a}_h)\phi(s_h,\boldsymbol{a}_h)^\top\right]\left((\Sigma_{\boldsymbol{\mu}}(h)+\lambda I)\right)^{-1}\right)} && (35)\\
&= \sqrt{\frac{1}{C_P}}\sqrt{Tr\left(\Sigma_{\boldsymbol{\mu}}(h)\left((\Sigma_{\boldsymbol{\mu}}(h)+\lambda I)\right)^{-1}\right)} \\
&\leq \sqrt{\frac{1}{C_P}}\sqrt{\sum_{j=1}^{d}\frac{\sigma_j}{\sigma_j+\lambda}} && (36)\\
&\leq \sqrt{\frac{d}{C_P}}\,, && (37)
\end{aligned}
$$

where $Tr(M)$ denotes the trace of matrix M and σ_j denote the eigenvalues of covariance matrix $\Sigma_{\boldsymbol{\mu}}(h)$. Above, Equation (32) uses the observation

$$
\begin{aligned}
\Lambda_h^{-1} &= \left(\frac{3}{5}\left(\frac{1}{m}\sum_{i=1}^{m}\phi(s_h,\boldsymbol{a}_h)\phi(s_h,\boldsymbol{a}_h)^\top+(\epsilon+\lambda)I\right)\right)^{-1} \\
&\preceq \left(\frac{3}{5}\left(\frac{1}{m}\sum_{i=1}^{m}\phi(\widetilde{s}_h,\widetilde{\boldsymbol{a}}_h)\phi(\widetilde{s}_h,\widetilde{\boldsymbol{a}}_h)^\top+\lambda I\right)\right)^{-1} \\
&\preceq \left((\Sigma_{\boldsymbol{\mu}}(h)+\lambda I)\right)^{-1}\,,
\end{aligned}
$$

which follows from Lemma D.1 and similar arguments as in the proof of Lemma B.2; Equation (33) uses Jensen's inequality; Equation (34) uses the commutativity of the trace operator: $Tr(x^\top Mx)=Tr(Mxx^\top)$; Equation (35) uses the transition coverage part of Assumption 2; finally, Equation (36) uses the fact that the eigenvalues of $\Sigma_{\boldsymbol{\mu}}(h)$ and λ are nonnegative real numbers. □

Next, we will provide an upper bound on the difference in preference functions between the ground-truth and estimated parameters.

Lemma B.5. *For any agent i and $\theta_i \in \Theta_{\text{Unil}}(\widetilde{\theta}_i)$, with probability at least $1-\delta/2$, we have*

$$
\mathbb{E}_{\substack{\tau\sim d^{\boldsymbol{\mu}}\\ \tau'\sim d^{\boldsymbol{\mu}_{\text{ref}}}}}\left[\left\|\mathbb{P}\left(\cdot|\tau,\tau',\theta_i^*\right)-\mathbb{P}\left(\cdot|\tau,\tau',\theta_i\right)\right\|_1^2\right] \leq 8H\sqrt{d}\epsilon + c\cdot\frac{d}{m}\log\left(\frac{2Hnm\sqrt{d}}{\delta}\right)\,,
$$

where c is an absolute constant.

Proof. Lemma D.2 gives us the following bound:

$$
\mathbb{E}_{\substack{\tau\sim d^{\boldsymbol{\mu}}\\ \tau'\sim d^{\boldsymbol{\mu}_{\text{ref}}}}}\left[\left\|\mathbb{P}\left(\cdot|\tau,\tau',\theta_i^*\right)-\mathbb{P}\left(\cdot|\tau,\tau',\theta\right)\right\|_1^2\right] \leq \frac{c_1}{m}\sum_{j=1}^{m}\log\left(\frac{\mathbb{P}\left(\widetilde{o}_j|\widetilde{\tau}_j,\widetilde{\tau}'_j,\theta\right)}{\mathbb{P}\left(\widetilde{o}_j|\widetilde{\tau}_j,\widetilde{\tau}'_j,\theta_i^*\right)}\right)+\log\left(d\log\left(2n/\delta\right)\right)\,.
$$

Let us deal with the first term of the bound below. Note that the bound depends on clean samples. Let $\widehat{D}$ be the given dataset and S denote the set of corrupted trajectories in $\widehat{D}$. We can write

$$
\begin{aligned}
\sum_{j=1}^{m}\log\left(\frac{\mathbb{P}\left(\widetilde{o}_j|\widetilde{\tau}_j,\widetilde{\tau}'_j,\theta\right)}{\mathbb{P}\left(\widetilde{o}_j|\widetilde{\tau}_j,\widetilde{\tau}'_j,\theta_i^*\right)}\right) &= \sum_{(\tau,\tau',o)\in S}\log\left(\frac{\mathbb{P}\left(o_j|\tau_j,\tau'_j,\theta\right)}{\mathbb{P}\left(o_j|\tau_j,\tau'_j,\theta_i^*\right)}\right)+\sum_{(\tau,\tau',o)\notin S}\log\left(\frac{\mathbb{P}\left(o_j|\tau_j,\tau'_j,\theta\right)}{\mathbb{P}\left(o_j|\tau_j,\tau'_j,\theta_i^*\right)}\right) \\
&\leq \sum_{(\tau,\tau',o)\in\widehat{D}}\log\left(\frac{\mathbb{P}\left(o_j|\tau_j,\tau'_j,\theta\right)}{\mathbb{P}\left(o_j|\tau_j,\tau'_j,\theta_i^*\right)}\right)+\epsilon\cdot\log\left(\frac{1+\exp\left(H\sqrt{d}\right)}{1+\exp\left(-H\sqrt{d}\right)}\right)
\end{aligned}
$$

$$\leq \sum_{(\tau,\tau',o)\in\widehat{D}} \log\left(\frac{\mathbb{P}\left(o_j|\tau_j,\tau'_j,\widetilde{\theta}_i\right)}{\mathbb{P}\left(o_j|\tau_j,\tau'_j,\theta^*_i\right)}\right) + 2\epsilon\cdot H\sqrt{d}$$
$$\leq 8H\sqrt{d}\epsilon + c_2\cdot\frac{d}{m}\log\left(\frac{2Hmn}{\delta}\right) ,$$

where the first inequality uses the fact that the corrupted subset comprises an ϵ-fraction of the whole dataset and the fact that $\phi^\top\theta \leq H\sqrt{d}$, by assumption of linear MDPs; the second inequality uses the fact that $\widetilde{\theta}_i$ maximizes the log-likelihood with respect to the corrupted data, and that H and d are natural numbers so we can bound the $\log$ expression directly in terms of the bounds on the rewards; finally, for the last inequality we have used Lemma B.1 with some constant c_2. Putting things together, we obtain the stated bound. □

Next, we prove bounds on the gaps with respect to any chosen reward parameters from the confidence set.

Lemma B.6. *Let $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1,\ldots,\widehat{\theta}_n)$, where $\widehat{\theta}_i \in \Theta_{\text{Unil}}(\widetilde{\theta}_i)$, for every $i\in[n]$, and let $\boldsymbol{\pi}^*$ be a Nash equilibrium covered by D. Then, if*

$$\widetilde{\boldsymbol{\pi}} \in \arg\min_{\boldsymbol{\pi}} \widetilde{\text{Gap}}\left(\boldsymbol{\pi},\widehat{\boldsymbol{\theta}}\right) ,$$

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

$$0 \leq \text{Gap}\left(\widetilde{\boldsymbol{\pi}},\widehat{\boldsymbol{\theta}}\right) \leq \widetilde{O}\left(n\cdot\left(\frac{1}{\sqrt{C_R}}+\frac{1}{\sqrt{C_P}}\right)\cdot\left(H^{5/2}d^{3/4}\sqrt{\epsilon}+H^2\text{poly}(d)\frac{1}{\sqrt{m}}\right)\right) ,$$
$$0 \leq \text{Gap}\left(\boldsymbol{\pi}^*,\widehat{\boldsymbol{\theta}}\right) \leq \widetilde{O}\left(n\cdot\left(\frac{1}{\sqrt{C_R}}+\frac{1}{\sqrt{C_P}}\right)\cdot\left(H^{5/2}d^{3/4}\sqrt{\epsilon}+H^2\text{poly}(d)\frac{1}{\sqrt{m}}\right)\right) ,$$

Proof. First, note that both the true gap and estimated gap are, by definition, non-negative. Now, given $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\pi}}$, as specified in the statement, we have

$$\text{Gap}\left(\widetilde{\boldsymbol{\pi}},\widehat{\boldsymbol{\theta}}\right) = \sum_{i\in[n]} V_{i,0}^{\dagger,\widetilde{\boldsymbol{\pi}}_{-i}}\left(s_0,\widehat{\theta}_i\right) - V_{i,0}^{\widetilde{\boldsymbol{\pi}}}\left(s_0,\widehat{\theta}_i\right) \tag{38}$$
$$\leq \sum_{i\in[n]} \overline{V}_{i,0}^{\dagger,\widetilde{\boldsymbol{\pi}}_{-i}}\left(s_0,\widehat{\theta}_i\right) - \underline{V}_{i,0}^{\widetilde{\boldsymbol{\pi}}}\left(s_0,\widehat{\theta}_i\right) \tag{39}$$
$$\leq \min_{\boldsymbol{\pi}} \sum_{i\in[n]} \overline{V}_{i,0}^{\dagger,\boldsymbol{\pi}_{-i}}\left(s_0,\widehat{\theta}_i\right) - \underline{V}_{i,0}^{\boldsymbol{\pi}}\left(s_0,\widehat{\theta}_i\right) \tag{40}$$
$$\leq \sum_{i\in[n]} \overline{V}_{i,0}^{\dagger,\boldsymbol{\pi}^*_{-i}}\left(s_0,\widehat{\theta}_i\right) - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right) - \left(\underline{V}_{i,0}^{\boldsymbol{\pi}^*}\left(s_0,\widehat{\theta}_i\right) - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right)\right) \tag{41}$$
$$= \sum_{i\in[n]} \overline{V}_{i,0}^{\dagger,\boldsymbol{\pi}^*_{-i}}\left(s_0,\widehat{\theta}_i\right) + V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\theta^*\right) - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right)$$
$$- \left(\underline{V}_{i,0}^{\boldsymbol{\pi}^*}\left(s_0,\widehat{\theta}_i\right) + V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\theta^*\right) - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right)\right) \tag{42}$$
$$= \sum_{i\in[n]} \overline{V}_{i,0}^{\dagger,\boldsymbol{\pi}^*_{-i}}\left(s_0,\widehat{\theta}_i\right) + V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\theta^*\right) - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right) - V_{i,0}^{\dagger,\boldsymbol{\pi}^*_{-i}}\left(s_0,\theta^*\right)$$
$$+ V_{i,0}^{\boldsymbol{\pi}^*}\left(s_0,\theta^*\right) - \left(\underline{V}_{i,0}^{\boldsymbol{\pi}^*}\left(s_0,\widehat{\theta}_i\right) + V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\theta^*\right) - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right)\right)$$
$$+ \underbrace{\left(V_{i,0}^{\dagger,\boldsymbol{\pi}^*_{-i}}\left(s_0,\theta^*\right) - V_{i,0}^{\boldsymbol{\pi}^*}\left(s_0,\theta^*\right)\right)}_{=0} \tag{43}$$
$$= \sum_{i\in[n]} \underbrace{\overline{V}_{i,0}^{\dagger,\boldsymbol{\pi}^*_{-i}}\left(s_0,\widehat{\theta}_i\right) - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right) + V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\theta^*\right) - V_{i,0}^{\dagger,\boldsymbol{\pi}^*_{-i}}\left(s_0,\theta^*\right)}_{:=Z_1}$$
$$+ \underbrace{V_{i,0}^{\boldsymbol{\pi}^*}\left(s_0,\theta^*\right) - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\theta^*\right) - \left(\underline{V}_{i,0}^{\boldsymbol{\pi}^*}\left(s_0,\widehat{\theta}_i\right) - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right)\right)}_{:=Z_2} \tag{44}$$

where Equation (38) follows by definition; Equation (39) follows from Lemma B.3; Equation (40) follows by design of Algorithm 4; in Equation (41) we just substitute $\boldsymbol{\pi}^*$ and add and subtract the same term; in Equation (42) and Equation (43) we again add and subtract identical terms; Equation (44) follows from the fact that $\boldsymbol{\pi}^*$ is a NE under θ^*.

Now, we will deal with the two terms above, Z_1 and Z_2, separately. First, let us consider the term Z_2. For every i, we have

$$
\begin{aligned}
&V_{i,0}^{\boldsymbol{\pi}^*}(s_0,\theta_i^*) - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}(s_0,\theta_i^*) - \left(\underline{V}_{i,0}^{\boldsymbol{\pi}^*}\left(s_0,\widehat{\theta}_i\right) - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right)\right)\\
&= \mathbb{E}_{\boldsymbol{a}_0\sim\boldsymbol{\pi}_0^*(\cdot|s_0)}\left[Q_{i,0}^{\boldsymbol{\pi}^*}(s_0,\boldsymbol{a}_0) - \underline{Q}_{i,0}^{\boldsymbol{\pi}^*}(s_0,\boldsymbol{a}_0,\widehat{\theta}_i)\right] - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}(s_0,\theta_i^*) + V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right)\\
&\le \mathbb{E}_{\boldsymbol{a}_0\sim\boldsymbol{\pi}_0^*(\cdot|s_0)}\left[\mathbb{B}_{i,0}V_{i,1}^{\boldsymbol{\pi}^*}(s_0,\boldsymbol{a}_0) - \underline{\mathbb{B}}_{i,0}\underline{V}_{i,0}^{\boldsymbol{\pi}^*}(s_0,\boldsymbol{a}_0,\widehat{\theta}_i) + 2\Gamma(s_0,\boldsymbol{a}_0)\right] - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}(s_0,\theta_i^*) + V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right) \qquad (45)\\
&= \mathbb{E}_{\boldsymbol{a}_0\sim\boldsymbol{\pi}_0^*(\cdot|s_0)}\left[R_{i,0}(s_0,\boldsymbol{a}_0) - \underline{R}_{i,0}(s_0,\boldsymbol{a}_0) + \mathbb{E}_{s_1\sim P_1(\cdot|s_0,\boldsymbol{a}_0)}\left[V_{i,1}^{\boldsymbol{\pi}^*}(s_1) - \underline{V}_{i,1}^{\boldsymbol{\pi}^*}(s_1,\widehat{\theta}_i)\right] + 2\Gamma(s_0,\boldsymbol{a}_0)\right]\\
&\qquad - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}(s_0,\theta_i^*) + V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right) \qquad (46)\\
&\le \mathbb{E}_{\boldsymbol{a}_0\sim\boldsymbol{\pi}_0^*(\cdot|s_0),s_1\sim P_1(\cdot|s_0,\boldsymbol{a}_0),\boldsymbol{a}_1\sim\boldsymbol{\pi}_1^*(\cdot|s_1)}\left[\sum_{h=0}^{1}\left(R_{i,h}(s_h,\boldsymbol{a}_h) - \underline{R}_{i,h}(s_h,\boldsymbol{a}_h) + 2\Gamma(s_h,\boldsymbol{a}_h)\right)\right]\\
&\qquad + \mathbb{E}_{\boldsymbol{a}_0\sim\boldsymbol{\pi}_0^*(\cdot|s_0),s_1\sim P_1(\cdot|s_0,\boldsymbol{a}_0),\boldsymbol{a}_1\sim\boldsymbol{\pi}_1^*(\cdot|s_1),s_2\sim P_2(\cdot|s_1,\boldsymbol{a}_1)}\left[V_{i,2}^{\boldsymbol{\pi}^*}(s_2) - \underline{V}_{i,2}^{\boldsymbol{\pi}^*}(s_2,\widehat{\theta}_i)\right]\\
&\qquad - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}(s_0,\theta_i^*) + V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right)\\
&\le \ldots\ldots\\
&\le \mathbb{E}_{\boldsymbol{\pi}^*}\left[\sum_{h=0}^{H-1}\left(R_{i,h}(s_h,\boldsymbol{a}_h) - \underline{R}_{i,h}(s_h,\boldsymbol{a}_h) + 2\Gamma(s_h,\boldsymbol{a}_h)\right)\right] - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}(s_0,\theta_i^*) + V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right)\\
&= \mathbb{E}_{\tau\sim d^{\boldsymbol{\pi}^*}}\left[\phi(\tau)^\top\theta_i^* - \phi(\tau)^\top\widehat{\theta}_i\right] - \mathbb{E}_{\tau\sim d^{\boldsymbol{\mu}_{\text{ref}}}}\left[\phi(\tau)^\top\theta_i^* - \phi(\tau)^\top\widehat{\theta}_i\right] + 2\mathbb{E}_{\boldsymbol{\pi}*}\left[\sum_{h=0}^{H-1}\Gamma(s_h,\boldsymbol{a}_h)\right] \qquad (47)\\
&= \mathbb{E}_{\tau\sim d^{\boldsymbol{\pi}^*},\tau'\sim d^{\boldsymbol{\mu}_{\text{ref}}}}\left[\left(\phi(\tau)-\phi(\tau')\right)^\top\left(\theta_i^* - \widehat{\theta}_i\right)\right] + 2\mathbb{E}_{\boldsymbol{\pi}*}\left[\sum_{h=0}^{H-1}\Gamma(s_h,\boldsymbol{a}_h)\right]\\
&\le \sqrt{\mathbb{E}_{\tau\sim d^{\boldsymbol{\pi}^*},\tau'\sim d^{\boldsymbol{\mu}_{\text{ref}}}}\left[\left(\left(\phi(\tau)-\phi(\tau')\right)^\top\left(\theta_i^* - \widehat{\theta}_i\right)\right)^2\right]} + 2\mathbb{E}_{\boldsymbol{\pi}*}\left[\sum_{h=0}^{H-1}\Gamma(s_h,\boldsymbol{a}_h)\right] \qquad (48)\\
&= \sqrt{\left(\theta_i^* - \widehat{\theta}_i\right)^\top\mathbb{E}_{\tau\sim d^{\boldsymbol{\pi}^*},\tau'\sim d^{\boldsymbol{\mu}_{\text{ref}}}}\left[\left(\phi(\tau)-\phi(\tau')\right)\left(\phi(\tau)-\phi(\tau')\right)^\top\right]\left(\theta_i^* - \widehat{\theta}_i\right)}\\
&\qquad + 2\mathbb{E}_{\boldsymbol{\pi}*}\left[\sum_{h=0}^{H-1}\Gamma(s_h,\boldsymbol{a}_h)\right]\\
&= \sqrt{\left(\theta_i^* - \widehat{\theta}_i\right)^\top\Sigma^-_{\boldsymbol{\pi}^*,\boldsymbol{\mu}_{\text{ref}}}\left(\theta_i^* - \widehat{\theta}_i\right)} + 2\mathbb{E}_{\boldsymbol{\pi}*}\left[\sum_{h=0}^{H-1}\Gamma(s_h,\boldsymbol{a}_h)\right] \qquad (49)\\
&\le \sqrt{\frac{1}{C_R}}\sqrt{\left(\theta_i^* - \widehat{\theta}_i\right)^\top\Sigma^-_{\boldsymbol{\mu},\boldsymbol{\mu}_{\text{ref}}}\left(\theta_i^* - \widehat{\theta}_i\right)} + 2\mathbb{E}_{\boldsymbol{\pi}*}\left[\sum_{h=0}^{H-1}\Gamma(s_h,\boldsymbol{a}_h)\right] \qquad (50)\\
&= \sqrt{\frac{1}{C_R}}\sqrt{\mathbb{E}_{\tau\sim d^{\boldsymbol{\mu}},\tau'\sim d^{\boldsymbol{\mu}_{\text{ref}}}}\left[\left(\left(\phi(\tau)-\phi(\tau')\right)^\top\theta_i^* - \left(\phi(\tau)-\phi(\tau')\right)^\top\widehat{\theta}_i\right)^2\right]} + 2\mathbb{E}_{\boldsymbol{\pi}*}\left[\sum_{h=0}^{H-1}\Gamma(s_h,\boldsymbol{a}_h)\right]\\
&= \sqrt{\frac{1}{C_R}}\sqrt{\mathbb{E}_{\tau\sim d^{\boldsymbol{\mu}},\tau'\sim d^{\boldsymbol{\mu}_{\text{ref}}}}\left[\left|\sigma^{-1}\left(\mathbb{P}\left(o=1|\tau,\tau',\theta_i^*\right)\right) - \sigma^{-1}\left(\mathbb{P}\left(o=1|\tau,\tau',\widehat{\theta}_i\right)\right)\right|^2\right]}\\
&\qquad + 2\mathbb{E}_{\boldsymbol{\pi}*}\left[\sum_{h=0}^{H-1}\Gamma(s_h,\boldsymbol{a}_h)\right] \qquad (51)
\end{aligned}
$$

$$\leq \sqrt{\frac{\iota^2}{C_R}}\sqrt{\mathbb{E}_{\tau\sim d^{\boldsymbol{\mu}},\tau'\sim d^{\boldsymbol{\mu}_{\text{ref}}}}\left[\left|\mathbb{P}\left(o=1|\tau,\tau',\theta_i^*\right)-\mathbb{P}\left(o=1|\tau,\tau',\widehat{\theta}_i\right)\right|^2\right]}+2\mathbb{E}_{\boldsymbol{\pi}*}\left[\sum_{h=0}^{H-1}\Gamma(s_h,\boldsymbol{a}_h)\right] \tag{52}$$

$$\leq \sqrt{\frac{\iota^2}{2C_R}}\sqrt{\mathbb{E}_{\tau\sim d^{\boldsymbol{\mu}},\tau'\sim d^{\boldsymbol{\mu}_{\text{ref}}}}\left[\left\|\mathbb{P}\left(\cdot|\tau,\tau',\theta_i^*\right)-\mathbb{P}\left(\cdot|\tau,\tau',\widehat{\theta}_i\right)\right\|_1^2\right]}+2H\cdot E(d,m,\delta,\epsilon)\cdot\sqrt{\frac{5d}{C_P}} \tag{53}$$

$$\leq \sqrt{\frac{\iota^2}{2C_R}}\sqrt{8\epsilon+c\cdot\frac{d}{m}\log\left(\frac{nm}{\delta}\right)}+2H\cdot E(d,m,\delta,\epsilon)\cdot\sqrt{\frac{5d}{C_P}} \tag{54}$$

$$\tag{55}$$

where Equation (45) follows by definition of Q-functions and the Bellman operator, and Lemma B.2; Equation (46) follows from the definition of the Bellman operator; Equation (47) uses the trajectory-based definition of the return; Equation (48) uses Jensen's inequality; Equation (49) uses the definition of the difference covariance matrix with respect to $\boldsymbol{\pi}^*$ and $\boldsymbol{\mu}_{\text{ref}}$; Equation (50) uses the first part of Assumption 2; Equation (51) uses the definition of the link function and the preference data generation assumption; Equation (52) uses uses Lemma D.3; finally, for Equation (53) we have used Lemma B.4.

Now, denote $\pi_i^\dagger \in \arg\max_{\pi'} V_{i,0}^{\pi',\boldsymbol{\pi}^*_{-i}}(s_0,\widehat{\theta}_i)$. For the final term, Z_1, we similarly have

$$\begin{aligned}
&\overline{V}_{i,0}^{\dagger,\boldsymbol{\pi}^*_{-i}}\left(s_0,\widehat{\theta}_i\right)-V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right)-\left(V_{i,0}^{\dagger,\boldsymbol{\pi}^*_{-i}}\left(s_0,\theta_i^*\right)-V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\theta_i^*\right)\right)\\
&\quad\leq \mathbb{E}_{\tau\sim d^{\pi_i^\dagger},\tau'\sim d^{\boldsymbol{\pi}^*_{-i}}}\left[\left(\phi(\tau)-\phi(\tau')\right)^\top\left(\widehat{\theta}_i-\theta_i^*\right)\right]+2\mathbb{E}_{\pi_i^\dagger,\boldsymbol{\pi}^*_{-i}}\left[\sum_{h=0}^{H-1}\Gamma(s_h,\boldsymbol{a}_h)\right]\\
&\quad\leq \sqrt{\frac{1}{C_R}}\sqrt{\left(\widehat{\theta}_i-\theta_i^*\right)^\top\Sigma^-_{\boldsymbol{\mu},\boldsymbol{\mu}_{\text{ref}}}\left(\widehat{\theta}_i-\theta_i^*\right)}+2H\cdot E(d,m,\delta,\epsilon)\cdot\sqrt{\frac{5d}{C_P}}\\
&\quad\leq \sqrt{\frac{\iota^2}{2C_R}}\sqrt{\mathbb{E}_{\tau\sim d^{\boldsymbol{\mu}},\tau'\sim d^{\boldsymbol{\mu}_{\text{ref}}}}\left[\left\|\mathbb{P}\left(\cdot|\tau,\tau',\theta_i^*\right)-\mathbb{P}\left(\cdot|\tau,\tau',\underline{\theta}_i\right)\right\|_1^2\right]}+2H\cdot E(d,m,\delta,\epsilon)\cdot\sqrt{\frac{5d}{C_P}}\\
&\quad\leq \sqrt{\frac{\iota^2}{2C_R}}\sqrt{8\epsilon+c\cdot\frac{d}{m}\log\left(\frac{nm}{\delta}\right)}+2H\cdot E(d,m,\delta,\epsilon)\cdot\sqrt{\frac{5d}{C_P}}\,,
\end{aligned}$$

where we have used similar arguments as above (note that this is the part where low relative uncertainty is needed, as opposed to coverage of only a Nash equilibrium).

Putting everything together, and using the definition of $E(d,m,\delta,\epsilon)$ from Equation (28) we obtain

$$\begin{aligned}
\text{Gap}(\widetilde{\boldsymbol{\pi}},\widehat{\theta}_1,\ldots,\widehat{\theta}_n)&\leq 2n\cdot\sqrt{\frac{\iota^2}{2C_R}}\sqrt{8\epsilon+c\cdot\frac{d}{m}\log\left(\frac{nm}{\delta}\right)}\\
&\quad+4nH\cdot\sqrt{c_2(\delta)\left(\frac{(H\sqrt{d}+\gamma)^2\text{poly}(d)}{m}+(H\sqrt{d}+\gamma)^2\epsilon+(2\epsilon+\lambda)H\sqrt{d}\right)}\cdot\sqrt{\frac{5d}{C_P}}\\
&\leq \widetilde{O}\left(n\cdot\left(\frac{1}{\sqrt{C_R}}+\frac{1}{\sqrt{C_P}}\right)\cdot\left(H^{5/2}d^{3/4}\sqrt{\epsilon}+H^2\text{poly}(d)\frac{1}{\sqrt{m}}\right)\right)\,,
\end{aligned}$$

where we have also used our choice of λ. Finally, for the third statement, note that

$$\begin{aligned}
\text{Gap}\left(\boldsymbol{\pi}^*,\widehat{\theta}_1,\ldots,\widehat{\theta}_n\right)&=\sum_{i\in[n]}V_{i,0}^{\dagger,\boldsymbol{\pi}^*_{-i}}\left(s_0,\widehat{\theta}_i\right)-V_{i,0}^{\boldsymbol{\pi}^*}\left(s_0,\widehat{\theta}_i\right)\\
&\leq\sum_{i\in[n]}\overline{V}_{i,0}^{\dagger,\boldsymbol{\pi}^*_{-i}}\left(s_0,\widehat{\theta}_i\right)-V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right)-\left(\underline{V}_{i,0}^{\boldsymbol{\pi}^*}\left(s_0,\widehat{\theta}_i\right)-V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0,\widehat{\theta}_i\right)\right)\\
&\leq\widetilde{O}\left(n\cdot\left(\frac{1}{\sqrt{C_R}}+\frac{1}{\sqrt{C_P}}\right)\cdot\left(H^{5/2}d^{3/4}\sqrt{\epsilon}+H^2\text{poly}(d)\frac{1}{\sqrt{m}}\right)\right)\,,
\end{aligned}$$

where the first inequality uses Lemma B.3 as if the robust subroutine were applied on $\boldsymbol{\pi}^*$, while the second inequality follows from noting that we already have a bound on the previous quantity from Equation (41). □

Before we proceed, we need to provide guarantees on the output of the PGA methods used in Algorithm 4.

Proposition B.1. *Let $\eta_1 = 1/\sqrt{T_1}$ and, for every agent i, let*

$$\widehat{\theta}_i^* \in \arg\max_{\theta_i \in \Theta_{\text{Unil}}(\widetilde{\theta}_i)} \text{Gap}_i(\boldsymbol{\pi}^*, \theta_i),$$

where

$$\widehat{\theta}_i := \frac{1}{T_1} \sum_{t=1}^{T_1} \widehat{\theta}_i^{(t)},$$

and the iterates $\widehat{\theta}_i^{(t)}$ are generated by

$$\widehat{\theta}_i^{(t+1)} = \mathcal{P}_{\Theta_{\text{Unil}}(\widetilde{\theta}_i)}\left(\widehat{\theta}_i^{(t)} + \eta_1 \widetilde{\nabla}_\theta \text{Gap}_i(\boldsymbol{\pi}^*, \widehat{\theta}_i^{(t)})\right)$$

for T_1 steps. Then,

$$\text{Gap}_i(\boldsymbol{\pi}^*, \widehat{\theta}_i^*) - \text{Gap}_i(\boldsymbol{\pi}^*, \widehat{\theta}_i) \leq \widetilde{\mathcal{O}}\left(\left(\frac{1}{\sqrt{C_R}} + \frac{1}{\sqrt{C_P}}\right)\left(H^{5/2} d^{3/4}\sqrt{\epsilon} + H^2 \frac{\text{poly}(d)}{\sqrt{m}}\right) + \frac{H^2\sqrt{\text{poly}(d)}}{\sqrt{T_1}}\left(\sqrt{\epsilon} + \frac{1}{\sqrt{m}}\right)\right).$$

Proof. First, note that, for any agent i and parameter θ_i, we have

$$\begin{aligned}\text{Gap}_i\left(\boldsymbol{\pi}^*, \theta_i\right) &= V_{i,0}^{\dagger, \boldsymbol{\pi}^*_{-i}}(s_0, \theta_i) - V_{i,0}^{\boldsymbol{\pi}^*}(s_0, \theta_i) \\ &= \max_{\pi_i'} V_{i,0}^{\pi_i', \boldsymbol{\pi}^*_{-i}}(s_0, \theta_i) - V_{i,0}^{\boldsymbol{\pi}^*}(s_0, \theta_i)\ .\end{aligned}$$

Now, let us define

$$\Pi_i^{\text{PP},\dagger}(\theta_i) = \left\{\pi_i^\dagger \in \Pi_i^{\text{PP}} : V_{i,0}^{\pi_i^\dagger, \boldsymbol{\pi}^*_{-i}}(s_0, \theta_i) = \max_{\pi_i'} V_{i,0}^{\pi_i', \boldsymbol{\pi}^*_{-i}}(s_0, \theta_i)\right\}\ ,$$

as the set of unilateral maximizer policies for player i at θ_i. Let $\pi_i^\dagger \in \Pi_i^{\text{PP},\dagger}(\widehat{\theta}_i^*)$ be any unilateral maximizer at $\widehat{\theta}_i^*$, and define

$$g_i^* := \nabla_{\theta_i}\left(V_{i,0}^{\pi_i^\dagger, \boldsymbol{\pi}^*_{-i}}(s_0, \theta_i) - V_{i,0}^{\boldsymbol{\pi}^*}(s_0, \theta_i)\right)\ .$$

Since the value function is linear in θ_i, g_i^* does not depend on θ_i. Moreover, Danskin's subdifferential formula implies that

$$g_i^* \in \partial_{\theta_i} \text{Gap}_i\left(\boldsymbol{\pi}^*, \widehat{\theta}_i^*\right)\ .$$

By linearity in θ_i, for any θ_i' we have

$$(\theta_i')^\top g_i^* = V_{i,0}^{\pi_i^\dagger, \boldsymbol{\pi}^*_{-i}}(s_0, \theta_i') - V_{i,0}^{\boldsymbol{\pi}^*}(s_0, \theta_i')\ . \tag{56}$$

In particular, since $\pi_i^\dagger$ is active at $\widehat{\theta}_i^*$,

$$\text{Gap}_i\left(\boldsymbol{\pi}^*, \widehat{\theta}_i^*\right) = \left(\widehat{\theta}_i^*\right)^\top g_i^*\ ,$$

while for any θ_i,

$$\text{Gap}_i\left(\boldsymbol{\pi}^*, \theta_i\right) \geq \theta_i^\top g_i^*\ .$$

Thus, we can write:

$$\text{Gap}_i\left(\boldsymbol{\pi}^*, \widehat{\theta}_i^*\right) - \text{Gap}_i\left(\boldsymbol{\pi}^*, \theta_i\right) \leq \left\langle \widehat{\theta}_i^* - \theta_i, g_i^*\right\rangle\ . \tag{57}$$

Now, since $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_{\text{ref}}$ cover $\boldsymbol{\pi}^*$ and its unilateral deviations, we can estimate the gradient of the value function at $\boldsymbol{\pi}^*$, or its unilateral deviations, from the gradient of the value function at $\boldsymbol{\mu}$ or $\boldsymbol{\mu}_{\text{ref}}$. Let $\widehat{\theta}_i = (1/T_1)\sum_{t=1}^{T_1}\widehat{\theta}_i^{(t)}$. Note that, since $\widehat{\theta}_i \in \Theta_{\text{Unil}}(\widetilde{\theta}_i)$ and $\widehat{\theta}_i^* \in \Theta_{\text{Unil}}(\widetilde{\theta}_i)$, we have

$$\begin{aligned}
&\left|\left\langle \widehat{\theta}_i^* - \widehat{\theta}_i, g_i^* - \nabla_\theta V_{i,0}^{\boldsymbol{\mu}}\left(s_0, \widehat{\theta}_i\right) + \nabla_\theta V_{i,0}^{\boldsymbol{\pi}^*}\left(s_0, \widehat{\theta}_i\right)\right\rangle\right| \\
&= \left|\left(V_{i,0}^{\pi_i^\dagger, \boldsymbol{\pi}_{-i}^*}\left(s_0, \widehat{\theta}_i^*\right) - V_{i,0}^{\boldsymbol{\mu}}\left(s_0, \widehat{\theta}_i^*\right)\right) - \left(V_{i,0}^{\pi_i^\dagger, \boldsymbol{\pi}_{-i}^*}\left(s_0, \widehat{\theta}_i\right) - V_{i,0}^{\boldsymbol{\mu}}\left(s_0, \widehat{\theta}_i\right)\right)\right| \\
&= \left|\mathbb{E}_{\tau\sim d^{\pi_i^\dagger, \boldsymbol{\pi}_{-i}^*}, \tau'\sim d^{\boldsymbol{\mu}}}\left[\left(\phi(\tau) - \phi(\tau')\right)^\top\left(\widehat{\theta}_i - \widehat{\theta}_i^*\right)\right]\right| \\
&\leq \left|\mathbb{E}_{\tau\sim d^{\pi_i^\dagger, \boldsymbol{\pi}_{-i}^*}, \tau'\sim d^{\boldsymbol{\mu}}}\left[\left(\phi(\tau) - \phi(\tau')\right)^\top\left(\widehat{\theta}_i - \theta_i^*\right)\right]\right| + \left|\mathbb{E}_{\tau\sim d^{\pi_i^\dagger, \boldsymbol{\pi}_{-i}^*}, \tau'\sim d^{\boldsymbol{\mu}}}\left[\left(\phi(\tau) - \phi(\tau')\right)^\top\left(\theta_i^* - \widehat{\theta}_i^*\right)\right]\right| \\
&\leq 2\sqrt{\frac{\iota^2}{2C_R}}\sqrt{8\epsilon + c\cdot\frac{d}{m}\log\left(\frac{nm}{\delta}\right)}\,,
\end{aligned} \tag{58}$$

where the first equality follows from Equation (56) and linearity of the value function in θ; the rest follows the same argument as the proof of Lemma B.5. Similarly, we have

$$\left|\left\langle \widehat{\theta}_i^* - \widehat{\theta}_i, \nabla_{\theta_i} V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0, \widehat{\theta}_i\right) - \nabla_{\theta_i} V_{i,0}^{\boldsymbol{\pi}^*}\left(s_0, \widehat{\theta}_i\right)\right\rangle\right| \leq 2\sqrt{\frac{\iota^2}{2C_R}}\sqrt{8\epsilon + c\cdot\frac{d}{m}\log\left(\frac{nm}{\delta}\right)}\,.$$

Combining, we get

$$\left|\left\langle \widehat{\theta}_i^* - \widehat{\theta}_i, g_i^* - \nabla_{\theta_i}\mathcal{R}(\widehat{\theta}_i)\right\rangle\right| \leq 4\sqrt{\frac{\iota^2}{2C_R}}\sqrt{8\epsilon + c\cdot\frac{d}{m}\log\left(\frac{nm}{\delta}\right)}\,, \tag{59}$$

where

$$\mathcal{R}(\widehat{\theta}_i) = V_{i,0}^{\boldsymbol{\mu}}\left(s_0, \widehat{\theta}_i\right) - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0, \widehat{\theta}_i\right)\,.$$

So, along the direction of $\widehat{\theta}_i^* - \widehat{\theta}_i$, the gradient of the difference of values at $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_{\text{ref}}$ is a good approximation of the active linear branch of the gap at $\widehat{\theta}_i^*$. Now, in order to approximate the gradient at $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_{\text{ref}}$, we use the fact that

$$\nabla_{\theta_i} V_{i,0}^{\boldsymbol{\mu}}(s_0, \widehat{\theta}_i) = \sum_{h=0}^{H-1}(d_h^{\boldsymbol{\mu}})^\top\Phi\,,$$

together with the fact that we already have access to ϵ-corrupted features from $d^{\boldsymbol{\mu}}$. Thus, we can define a robust estimate of the above as

$$\widetilde{\nabla}_{\theta_i} V_{i,0}^{\boldsymbol{\mu}}\left(s_0, \widehat{\theta}_i\right) = \sum_{h=0}^{H-1}\texttt{RobMean}\left(D_{h,\phi}^{\boldsymbol{\mu}}\right)\,,$$

and

$$\widetilde{\nabla}_{\theta_i} V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0, \widehat{\theta}_i\right) = \sum_{h=0}^{H-1}\texttt{RobMean}\left(D_{h,\phi}^{\boldsymbol{\mu}_{\text{ref}}}\right)\,.$$

Corollary 2 gives us bounds $f(\epsilon)$ on the L2-error:

$$\left\|\widetilde{\nabla}_\theta V_{i,0}^{\boldsymbol{\mu}}(s_0, \theta) - \nabla_\theta V_{i,0}^{\boldsymbol{\mu}}(s_0, \theta)\right\| \leq O(Hf(\epsilon))\,,$$

and

$$\left\|\widetilde{\nabla}_\theta V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}(s_0, \theta) - \nabla_\theta V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}(s_0, \theta)\right\| \leq O(Hf(\epsilon))\,,$$

where

$$f(\epsilon) = \sqrt{\frac{d\log(\text{poly}(d))}{m}} + \sqrt{d\epsilon} + \sqrt{\frac{d\log(1/\delta)}{m}} \ .$$

Let us now define

$$\widetilde{\nabla}_{\theta_i}\text{Gap}_i\left(\boldsymbol{\pi}^*, \widehat{\theta}_i\right) = \widetilde{\nabla}_{\theta_i} V_{i,0}^{\boldsymbol{\mu}}\left(s_0, \widehat{\theta}_i\right) - \widetilde{\nabla}_{\theta_i} V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0, \widehat{\theta}_i\right) \ .$$

Note that we have

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_{\theta}\text{Gap}_i\left(\boldsymbol{\pi}^*, \widehat{\theta}_i\right)\right\|\right] \leq 4H + O(Hf(\epsilon)) \ ,$$

due to the guarantees of `RobMean` and the feature norm bounds. Therefore, we are in the conditions of Lemma D.4. Recall that $\widehat{\theta}_i^*$ is the true maximizer of $\text{Gap}_i(\boldsymbol{\pi}^*, \theta_i)$ over $\Theta_{\text{Unil}}(\widetilde{\theta}_i)$. Then, using the active branch g_i^* selected above, we have

$$\text{Gap}_i\left(\boldsymbol{\pi}^*, \widehat{\theta}_i^*\right) - \frac{1}{T_1}\sum_{t=1}^{T_1}\text{Gap}_i\left(\boldsymbol{\pi}^*, \widehat{\theta}_i^{(t)}\right) \leq \frac{1}{T_1}\sum_{t=1}^{T_1}\left\langle \widehat{\theta}_i^* - \widehat{\theta}_i^{(t)}, g_i^*\right\rangle \tag{60}$$

$$= \frac{1}{T_1}\sum_{t=1}^{T_1}\left\langle \widehat{\theta}_i^* - \widehat{\theta}_i^{(t)}, \widetilde{\nabla}_{\theta}\text{Gap}_i\left(\boldsymbol{\pi}^*, \widehat{\theta}_i^{(t)}\right)\right\rangle + \frac{1}{T_1}\sum_{t=1}^{T_1}\left\langle \widehat{\theta}_i^* - \widehat{\theta}_i^{(t)}, g_i^* - \widetilde{\nabla}_{\theta}\text{Gap}_i\left(\boldsymbol{\pi}^*, \widehat{\theta}_i^{(t)}\right)\right\rangle \tag{61}$$

$$\begin{aligned}
&\leq \frac{1}{T_1}\left(\frac{\left\|\widehat{\theta}_i^* - \widehat{\theta}_i^{(1)}\right\|^2}{2\eta} + \frac{\eta T_1(4H + O(Hf(\epsilon)))^2}{2}\right) \\
&\quad + \frac{1}{T_1}\sum_{t=1}^{T_1}\left\langle \widehat{\theta}_i^* - \widehat{\theta}_i^{(t)}, g_i^* - \nabla_{\theta} V_{i,0}^{\boldsymbol{\mu}}\left(s_0, \widehat{\theta}_i^{(t)}\right) + \nabla_{\theta} V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0, \widehat{\theta}_i^{(t)}\right)\right\rangle \\
&\quad + \frac{1}{T_1}\sum_{t=1}^{T_1}\left\langle \widehat{\theta}_i^* - \widehat{\theta}_i^{(t)}, \nabla_{\theta} V_{i,0}^{\boldsymbol{\mu}}\left(s_0, \widehat{\theta}_i^{(t)}\right) - \nabla_{\theta} V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0, \widehat{\theta}_i^{(t)}\right) - \widetilde{\nabla}_{\theta}\text{Gap}_i\left(\boldsymbol{\pi}^*, \widehat{\theta}_i^{(t)}\right)\right\rangle
\end{aligned} \tag{62}$$

$$\begin{aligned}
&\leq O\left(\frac{H^2 + f(\epsilon)}{\sqrt{T_1}}\right) + 4\sqrt{\frac{\iota^2}{2C_R}}\sqrt{8\epsilon + c\cdot\frac{d}{m}\log\left(\frac{nm}{\delta}\right)} \\
&\quad + \frac{1}{T_1}\sum_{t=1}^{T_1}\left\|\widehat{\theta}_i^* - \widehat{\theta}_i^{(t)}\right\|\left\|\nabla_{\theta} V_{i,0}^{\boldsymbol{\mu}}\left(s_0, \widehat{\theta}_i^{(t)}\right) - \nabla_{\theta} V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0, \widehat{\theta}_i^{(t)}\right) - \widetilde{\nabla}_{\theta}\text{Gap}_i\left(\boldsymbol{\pi}^*, \widehat{\theta}_i^{(t)}\right)\right\|
\end{aligned} \tag{63}$$

$$\begin{aligned}
&\leq O\left(\frac{H^2 + f(\epsilon)}{\sqrt{T_1}}\right) + 4\sqrt{\frac{\iota^2}{2C_R}}\sqrt{8\epsilon + c\cdot\frac{d}{m}\log\left(\frac{nm}{\delta}\right)} + 2H^2\sqrt{d}f(\epsilon) \\
&\leq \widetilde{O}\left(\left(\frac{1}{\sqrt{C_R}} + \frac{1}{\sqrt{C_P}}\right)\cdot\left(H^{5/2}d^{3/4}\sqrt{\epsilon} + H^2\text{poly}(d)\frac{1}{\sqrt{m}}\right) + \frac{H^2\sqrt{\text{poly}(d)}}{\sqrt{T_1}}\left(\sqrt{\epsilon} + \frac{1}{\sqrt{m}}\right)\right) \ ,
\end{aligned}$$

where Equation (60) follows from Equation (57); Equation (61) adds and subtracts the same term; Equation (62) follows from Lemma D.4; Equation (63) follows from Cauchy-Schwarz and Equation (59); and the last inequality follows from Corollary 2. □

Theorem B.1. *Let $\epsilon \in [0, 1/2)$, $\lambda \geq \Omega(dH\log(m/\delta)/m)$, and $\delta > 0$. Set $\Theta_{\text{Unil}}(\cdot)$ as in Equation (5) and $\Gamma(s, \boldsymbol{a}) = E(d, m, \delta, \epsilon)\cdot\|\phi(s, \boldsymbol{a})\|_{\Lambda_h^{-1}}$. Suppose Assumption 2 is satisfied and PGA is run for T_1 steps with learning rate $\eta = O(1/\sqrt{T_1})$. Then, there exist robust subroutines `RobEst`, `TrimmedMLE`, and `RobMean` such that, with probability at least $1-\delta$, the output $\widetilde{\boldsymbol{\pi}}$ of Algorithm 4 with subroutines `RobEst`, `TrimmedMLE`, `RobMean` and `RewardEst`, satisfies*

$$\text{Gap}(\widetilde{\boldsymbol{\pi}}) \leq \widetilde{O}\left(\left(\frac{1}{\sqrt{C_R}} + \frac{1}{\sqrt{C_P}} + \frac{1}{\sqrt{T_1}}\right)\cdot\left(H^{5/2}nd^{3/4}\sqrt{\epsilon} + H^2 n\sqrt{\frac{\text{poly}(d)}{m}}\right)\right) \ .$$

Proof. Define $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_n^*)$ and let

$$f(\epsilon) = \widetilde{O}\left(\left(\frac{1}{\sqrt{C_R}} + \frac{1}{\sqrt{C_P}}\right) \cdot \left(H^{5/2} d^{3/4} \sqrt{\epsilon} + H^2 \text{poly}(d) \frac{1}{\sqrt{m}}\right) + \frac{H^2 \sqrt{\text{poly}(d)}}{\sqrt{T_1}} \left(\sqrt{\epsilon} + \frac{1}{\sqrt{m}}\right)\right) .$$

Also define

$$\Delta := \widetilde{O}\left(n \cdot \left(\frac{1}{\sqrt{C_R}} + \frac{1}{\sqrt{C_P}}\right) \cdot \left(H^{5/2} d^{3/4} \sqrt{\epsilon} + H^2 \text{poly}(d) \frac{1}{\sqrt{m}}\right)\right) .$$

Moreover, let $\widehat{\boldsymbol{\theta}}^*$ denote the maximizer of the gap with respect to $\boldsymbol{\pi}^*$ over the unilateral confidence set, i.e.,

$$\widehat{\boldsymbol{\theta}}^* \in \arg \max_{\boldsymbol{\theta} \in \Theta_{\text{Unil}}(\widetilde{\theta}_1) \times \ldots \times \Theta_{\text{Unil}}(\widetilde{\theta}_n)} \text{Gap}\left(\boldsymbol{\pi}^*, \boldsymbol{\theta}\right) .$$

We have

$$\begin{aligned}
&\text{Gap}\left(\widetilde{\boldsymbol{\pi}}, \boldsymbol{\theta}^*\right) \leq \text{Gap}\left(\boldsymbol{\pi}^*, \boldsymbol{\theta}^*\right) + \Delta && (64)\\
&\leq \text{Gap}\left(\boldsymbol{\pi}^*, \widehat{\boldsymbol{\theta}}^*\right) + \Delta && (65)\\
&\leq \text{Gap}\left(\boldsymbol{\pi}^*, \widehat{\boldsymbol{\theta}}\right) + n \cdot f(\epsilon) + \Delta && (66)\\
&\leq \text{Gap}\left(\widetilde{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}}\right) + n \cdot f(\epsilon) + 2\Delta && (67)\\
&\leq \widetilde{\text{Gap}}\left(\widetilde{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}}\right) + n \cdot f(\epsilon) + 2\Delta && (68)\\
&\leq \widetilde{\text{Gap}}\left(\boldsymbol{\pi}^*, \widehat{\boldsymbol{\theta}}\right) + n \cdot f(\epsilon) + 2\Delta && (69)\\
&= \sum_{i \in [n]} \overline{V}_{i,0}^{\dagger, \boldsymbol{\pi}^*_{-i}}\left(s_0, \widehat{\theta}_i\right) - \underline{V}_{i,0}^{\boldsymbol{\pi}^*}\left(s_0, \widehat{\theta}_i\right) + n \cdot f(\epsilon) + 2\Delta && (70)\\
&\leq \sum_{i \in [n]} \overline{V}_{i,0}^{\dagger, \boldsymbol{\pi}^*_{-i}}\left(s_0, \widehat{\theta}_i\right) - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0, \widehat{\theta}_i\right)\\
&\qquad - \left(\underline{V}_{i,0}^{\boldsymbol{\pi}^*}\left(s_0, \widehat{\theta}_i\right) - V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}\left(s_0, \widehat{\theta}_i\right)\right) + n \cdot f(\epsilon) + 2\Delta && (71)\\
&\leq \widetilde{O}\left(n \cdot \left(\frac{1}{\sqrt{C_R}} + \frac{1}{\sqrt{C_P}}\right) \cdot \left(H^{5/2} d^{3/4} \sqrt{\epsilon} + H^2 \text{poly}(d) \frac{1}{\sqrt{m}}\right)\right) + n \cdot f(\epsilon) + 2\Delta && (72)\\
&\leq \widetilde{O}\left(n \cdot \left(\frac{1}{\sqrt{C_R}} + \frac{1}{\sqrt{C_P}}\right) \cdot \left(H^{5/2} d^{3/4} \sqrt{\epsilon} + H^2 \text{poly}(d) \frac{1}{\sqrt{m}}\right) + \frac{H^2 n \sqrt{\text{poly}(d)}}{\sqrt{T_1}} \left(\sqrt{\epsilon} + \frac{1}{\sqrt{m}}\right)\right) .
\end{aligned}$$

Equation (64) follows from Lemma B.6. Equation (65) follows from the definition of $\widehat{\boldsymbol{\theta}}^*$. Equation (66) follows from Proposition B.1, applied coordinate-wise at $\boldsymbol{\pi}^*$. Equation (67) follows again from Lemma B.6. Equation (68) follows from Lemma B.3. Equation (69) follows by design of Algorithm 4, since $\widetilde{\boldsymbol{\pi}}$ minimizes the estimated gap at $\widehat{\boldsymbol{\theta}}$. Equation (70) follows by definition of estimated gap. In Equation (71) we add and subtract the same term. Equation (72) follows from Lemma B.6. The final inequality follows from the definitions of $f(\epsilon)$ and Δ. □

C Proof of Theorem 5.1

This section includes full proof of Theorem 5.1. We begin by establishing regret guarantees on the Optimistic Hedge algorithm applied to our setting.

Lemma C.1. *Denote by $\widetilde{\boldsymbol{\pi}}$ the joint policy returned by Optimistic Hedge run for T_2 rounds. For each $i \in [n]$, let*

$$\begin{aligned}
\overline{\text{Reg}}_{i,h,T_2} &= \max_{\pi^\dagger_{i,h}} \mathbb{E}_{a_i^\dagger \sim \pi^\dagger_{i,h}(\cdot|s), a_i' \sim \widetilde{\pi}_{i,h}(\cdot|s)} \left[\mathbb{E}_{\boldsymbol{a}_{-i} \sim \widetilde{\boldsymbol{\pi}}_{-i,h}(\cdot|s)} \left[\overline{Q}_{i,h}^{\pi_i^\dagger, \widetilde{\boldsymbol{\pi}}_{-i,h}}(s, a_i^\dagger, \boldsymbol{a}_{-i}) - \underline{Q}^{\widetilde{\boldsymbol{\pi}}}(s, a_i', \boldsymbol{a}_{-i})\right]\right]\\
&- \min_{\pi'_{i,h}} \max_{\pi^\dagger_{i,h}} \mathbb{E}_{a_i^\dagger \sim \pi^\dagger_{i,h}(\cdot|s), a_i' \sim \pi'_{i,h}(\cdot|s)} \left[\mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}_{-i,h}(\cdot|s)} \left[\overline{Q}_{i,h}^{\pi_i^\dagger, \widetilde{\boldsymbol{\pi}}_{-i}}(s, a_i^\dagger, \boldsymbol{a}_{-i}) - \underline{Q}^{\pi'_i, \widetilde{\boldsymbol{\pi}}_{-i}}(s, a_i', \boldsymbol{a}_{-i})\right]\right] .
\end{aligned}$$

For every $i \in [n]$ and $h \in [H-1]$, we have

$$\sum_{i\in[n]} \overline{\text{Reg}}_{i,h,T_2} \leq O\left(n^2 H \cdot \log|A| \cdot \log^4 T_2\right)$$

Proof. Recall that the loss we use for Optimistic Hedge is defined as

$$\mathcal{L}_i^s(a^\dagger, a') = \mathbb{E}_{\boldsymbol{a}_{-i}\sim\widetilde{\boldsymbol{\pi}}_{-i,h}(\cdot|s)}\left[\overline{Q}_{i,h}^{\dagger,\widetilde{\boldsymbol{\pi}}_{-i}}(s, a_i^\dagger, \boldsymbol{a}_{-i}) - \underline{Q}_{i,h}^{\widetilde{\boldsymbol{\pi}}}(s, a', \boldsymbol{a}_{-i})\right] .$$

Now, let $\pi_i^{(\dagger,(t)}$ denote the iterations of player i with respect to the $\max$ problem. Note that, for every agent i and state s, after T_2 steps, we have

$$\begin{aligned}
&\max_{\pi_i^\dagger} \sum_{t=1}^{T_2} \mathbb{E}_{a^\dagger\sim\pi_i^\dagger, a'\sim\widetilde{\pi}_i^{(t)}}\left[\mathcal{L}_i^s(a^\dagger, a')\right] - \min_{\pi_i}\max_{\pi_i^\dagger} \sum_{t=1}^{T_2} \mathbb{E}_{a^\dagger\sim\pi_i^\dagger, a'\sim\widetilde{\pi}_i^{(t)}}\left[\mathcal{L}_i^s(a^\dagger, a')\right] \\
&= \underbrace{\max_{\pi_i^\dagger} \sum_{t=1}^{T_2} \mathbb{E}_{a^\dagger\sim\pi_i^\dagger, a'\sim\widetilde{\pi}_i^{(t)}}\left[\mathcal{L}_i^s(a^\dagger, a')\right] - \sum_{t=1}^{T_2} \mathbb{E}_{a^\dagger\sim\pi_i^{\dagger,(t)}, a'\sim\widetilde{\pi}_i^{(t)}}\left[\mathcal{L}_i^s(a^\dagger, a')\right]}_{\text{Reg}_{i,h,T_2} \text{ for } \pi_i^{\dagger,(t)}} \\
&\quad + \underbrace{\sum_{t=1}^{T_2} \mathbb{E}_{a^\dagger\sim\pi_i^{\dagger,(t)}, a'\sim\widetilde{\pi}_i^{(t)}}\left[\mathcal{L}_i^s(a^\dagger, a')\right] - \min_{\pi_i} \sum_{t=1}^{T_2} \mathbb{E}_{a^\dagger\sim\pi_i^{\dagger,(t)}, a'\sim\widetilde{\pi}_i^{(t)}}\left[\mathcal{L}_i^s(a^\dagger, a')\right]}_{\text{Reg}_{i,h,T_2} \text{ for } \pi_i^{(t)}} \\
&\quad + \underbrace{\min_{\pi_i} \sum_{t=1}^{T_2} \mathbb{E}_{a^\dagger\sim\pi_i^{\dagger,(t)}, a'\sim\widetilde{\pi}_i^{(t)}}\left[\mathcal{L}_i^s(a^\dagger, a')\right] - \min_{\pi_i}\max_{\pi_i^\dagger} \sum_{t=1}^{T_2} \mathbb{E}_{a^\dagger\sim\pi_i^\dagger, a'\sim\widetilde{\pi}_i^{(t)}}\left[\mathcal{L}_i^s(a^\dagger, a')\right]}_{\leq 0} \\
&\leq O\left(nH \log|A_i| \log^4 T_2\right) ,
\end{aligned}$$

where for the inequality we have used Theorem D.2 applied on the regrets with respect to the $\max$ and $\min$ players, and the fact that $\min_x f(x,y) \leq \min_x \max_y f(x,y)$, due to the monotonicity of the $\min$ operator. This implies that, the empirical distribution of the sequence of policies up to time step T_2, which equals the returned policy $\widetilde{\boldsymbol{\pi}}$, satisfies, for any player i and state s,

$$\max_{\pi_i^\dagger} \mathbb{E}_{a^\dagger\sim\pi_i^\dagger, a'\sim\widetilde{\pi}_i}\left[\mathcal{L}_i^s(a^\dagger, a')\right] - \min_{\pi_i}\max_{\pi_i^\dagger} \mathbb{E}_{a^\dagger\sim\pi_i^\dagger, a'\sim\pi_i}\left[\mathcal{L}_i^s(a^\dagger, a')\right] \leq O\left(\frac{n \cdot \log|A| \cdot \log^4 T_2}{T_2}\right) .$$

In particular, we have

$$\begin{aligned}
&\sum_{i\in[n]} \max_{\pi_{i,h}^\dagger} \mathbb{E}_{a_i^\dagger\sim\pi_{i,h}^\dagger(\cdot|s), a_i'\sim\widetilde{\pi}_{i,h}(\cdot|s)}\left[\mathbb{E}_{\boldsymbol{a}_{-i}\sim\widetilde{\boldsymbol{\pi}}_{-i,h}(\cdot|s)}\left[\overline{Q}_{i,h}^{\dagger,\widetilde{\boldsymbol{\pi}}_{-i,h}}(s, a_i^\dagger, \boldsymbol{a}_{-i}) - \underline{Q}_{i,h}^{\widetilde{\boldsymbol{\pi}}}(s, a_i', \boldsymbol{a}_{-i})\right]\right] \\
&- \min_{\boldsymbol{\pi}_h'} \sum_{i\in[n]} \max_{\pi_{i,h}^\dagger} \mathbb{E}_{a_i^\dagger\sim\pi_{i,h}^\dagger(\cdot|s), a_i'\sim\pi_{i,h}'(\cdot|s)}\left[\mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_{-i,h}(\cdot|s)}\left[\overline{Q}_{i,h}^{\dagger,\widetilde{\boldsymbol{\pi}}_{-i}}(s, a_i^\dagger, \boldsymbol{a}_{-i}) - \underline{Q}_{i,h}^{\widetilde{\boldsymbol{\pi}}}(s, a_i', \boldsymbol{a}_{-i})\right]\right] \\
&\leq O\left(\frac{n^2 H \cdot \log|A| \cdot \log^4 T_2}{T_2}\right) ,
\end{aligned}$$

where we have used the fact that the $\min$ of the sum is larger than the sum of individual $\min$s. □

Finally, we are ready to state and prove Theorem 5.1. We restate it here for convenience.

Theorem C.1. *Let $\epsilon \in [0, 1/2), \lambda \geq \Omega(dH \log(m/\delta)/m)$, and $\delta > 0$. Set $\Theta_{\text{Unil}}(\cdot)$ as in Equation (5) and $\Gamma(s, \boldsymbol{a}) = E(d, m, \delta, \epsilon) \cdot \|\phi(s, \boldsymbol{a})\|_{\Lambda_h^{-1}}$. Suppose Assumption 2 is satisfied, PGA is run for T_1 steps with learning rate $\eta_1 = O(1/\sqrt{T_1})$, and* `OptimisticHedge` *is run for T_2 steps with learning rate $\eta_2 = O(1/(n \log^4 T_2))$. Then, there exist robust subroutines*

`RobEst`, `TrimmedMLE`, and `RobMean` such that, with probability at least $1-\delta$, the output $\widetilde{\boldsymbol{\pi}}$ of Algorithm 5 with subroutines `RobEst`, `TrimmedMLE`, `RobMean` and `OptimisticHedge`, satisfies

$$\text{Gap}(\widetilde{\boldsymbol{\pi}}) \leq \widetilde{O}\left(\left(\frac{1}{\sqrt{C_R}}+\frac{1}{\sqrt{C_P}}+\frac{1}{\sqrt{T_1}}\right)\cdot\left(H^{5/2}nd^{3/4}\sqrt{\epsilon}+H^2n\frac{\sqrt{\text{poly}(d)}}{\sqrt{m}}\right)+\frac{Hn^2}{T_2}\right) .$$

Proof. Similar to the proof of Theorem 4.1, we can write

$$\begin{aligned}
\text{Gap}\left(\widetilde{\boldsymbol{\pi}},\boldsymbol{\theta}^*\right) &\leq \text{Gap}\left(\widetilde{\boldsymbol{\pi}},\widehat{\boldsymbol{\theta}}^*\right)\\
&\leq \text{Gap}\left(\widetilde{\boldsymbol{\pi}},\widehat{\boldsymbol{\theta}}\right)+n\cdot f(\epsilon)\\
&\leq \widetilde{\text{Gap}}\left(\widetilde{\boldsymbol{\pi}},\widehat{\boldsymbol{\theta}}\right)+n\cdot f(\epsilon)\\
&\leq \min_{\boldsymbol{\pi}}\widetilde{\text{Gap}}\left(\boldsymbol{\pi},\widehat{\boldsymbol{\theta}}\right)+O\left(\frac{n^2H\cdot\log|A|\cdot\log^4 T_2}{T_2}\right)+n\cdot f(\epsilon) \qquad (73)\\
&\leq \widetilde{\text{Gap}}\left(\boldsymbol{\pi}^*,\widehat{\boldsymbol{\theta}}\right)+O\left(\frac{n^2H\cdot\log|A|\cdot\log^4 T_2}{T_2}\right)+n\cdot f(\epsilon)\\
&= \sum_{i\in[n]}\overline{V}_{i,0}^{\dagger,\boldsymbol{\pi}^*_{-i}}\left(s_0,\widehat{\theta}_i\right)-\underline{V}_{i,0}^{\boldsymbol{\pi}^*}\left(s_0,\widehat{\theta}_i\right)+O\left(\frac{n^2H\cdot\log|A|\cdot\log^4 T_2}{T_2}\right)+n\cdot f(\epsilon)\\
&\leq \widetilde{O}\left(n\cdot\left(\frac{1}{\sqrt{C_R}}+\frac{1}{\sqrt{C_P}}+\frac{1}{\sqrt{T_1}}\right)\cdot\left(H^{5/2}d^{3/4}\sqrt{\epsilon}+H^2\text{poly}(d)\frac{1}{\sqrt{m}}\right)+\frac{n^2H}{T_2}\right) , \qquad (74)
\end{aligned}$$

where Equation (73) follows from Lemma C.1 and Equation (74) follows from Theorem 4.1. □

D Technical Results

This section includes various miscellaneous technical results that are used throughout the proofs in the paper.

Lemma D.1 (Zanette et al. [2021])**.** *Let $\{\phi_i\}_{i=1}^m \subset \mathbb{R}^d$ be i.i.d. samples from an underlying bounded distribution $\boldsymbol{\mu}_{\text{ref}}$, with $\|\phi_i\|_2 \leq 1$ and covariance $\Sigma_{\boldsymbol{\mu}_{\text{ref}}}$. Define*

$$\Lambda = \sum_{i=1}^m \phi_i\phi_i^\top + \lambda I ,$$

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

$$\frac{1}{3}\left(m\Sigma_{\boldsymbol{\mu}_{\text{ref}}}+\lambda I\right) \preceq \Lambda \preceq \frac{5}{3}\left(m\Sigma_{\boldsymbol{\mu}_{\text{ref}}}+\lambda I\right) .$$

Next, we state a result that bounds the difference in log probabilities of parameters in a given space.

Lemma D.2 (Lemma 2 of [Zhan et al., 2023] for the linear setting)**.** *With probability at least $1-\delta$, we have, for every agent i and $\theta \in \Theta_{\text{Unil}}(\widetilde{\theta}_i)$:*

$$\mathbb{E}_{\substack{\tau\sim d^{\boldsymbol{\mu}}\\ \tau'\sim d^{\boldsymbol{\mu}_{\text{ref}}}}}\left[\left\|\mathbb{P}\left(\cdot|\tau,\tau',\theta_i^*\right)-\mathbb{P}\left(\cdot|\tau,\tau',\theta\right)\right\|_1^2\right] \leq \frac{c}{m}\sum_{j=1}^m\log\left(\frac{\mathbb{P}\left(\widetilde{o}_j|\widetilde{\tau}_j,\widetilde{\tau}'_j,\theta_i^*\right)}{\mathbb{P}\left(\widetilde{o}_j|\widetilde{\tau}_j,\widetilde{\tau}'_j,\theta\right)}\right)+\log\left(d\log\left(n/\delta\right)\right) ,$$

where $c>0$ is an absolute constant.

Proof. This is just an application of Proposition 1 of [Zhan et al., 2023] to the linear setting, which then induces the result above by applying the union bound over all agents. □

Next, we show that the inverse of the sigmoid link function is Lipschitz on a bounded domain.

Lemma D.3. *Let $-\infty < a, b < \infty$ be two real numbers and let $\sigma(x) = 1/(1+\exp(-x))$ be defined on the domain $x \in [a, b]$. Then, there exists a positive number $\iota < \infty$, such that the inverse σ^{-1} of σ is Lipschitz with constant ι, that is,*

$$\sup_{p(x)\in(0,1):x\in[a,b]} \left| \frac{\partial \sigma^{-1}(p)}{\partial p} \right| \leq \iota \ .$$

As a consequence, if the rewards are bounded, then the inverse of the preference link function is Lipschitz continuous for some $\iota > 0$.

Proof. First, note that the derivative of the inverse of the sigmoid can be written as

$$\frac{\partial \sigma^{-1}(p)}{\partial p} = \frac{\partial}{\partial p} \log \frac{p}{1-p} = \frac{1}{p(1-p)} \ .$$

Now, since $x \in [a, b]$ and σ is continuous in $\mathbb{R}$, then, there exist $-\infty < a', b' < \infty$, such that $p = \sigma(x) \in [a', b']$, for every $x \in [a, b]$. Moreover, since the function $p(1-p)$ is also continuous in $\mathbb{R}$, then, there exist $-\infty < a'', b'' < \infty$, such that $p(1-p) \in [a'', b'']$, for all $p \in [a', b']$. Thus, there exists a positive constant ι, such that

$$\frac{1}{p(1-p)} \leq \iota \ ,$$

for all $x \in [a, b]$. This implies that the function σ^{-1} is Lipschitz on the domain of σ.

For the final statement of the result, note that the preference function uses differences in expected rewards that are individually bounded in $[-\sqrt{d}, \sqrt{d}]$. Thus, we have

$$\sum_{h=0}^{H-1} R(s_h, a_h) - R'(s_h, a_h) \in [-2H\sqrt{d}, 2H\sqrt{d}] \ .$$

Thus, the domain of the sigmoid link function, used for our preference model, has a bounded domain, which implies that its inverse is Lipschitz continuous. □

Corollary 1. *The function $f(\theta) = \mathbb{E}_{\tau\sim d^{\boldsymbol{\mu}_{\text{ref}}}} \left[\phi(\tau)^\top \theta\right]$ is H-Lipschitz and convex. Moreover, the set $\Theta_{\text{Unil}}(\theta')$ is a convex set, for any θ' and $\lambda > 0$.*

Proof. Note that we have

$$\|\nabla_\theta f(\theta)\| = \left\|(d^{\boldsymbol{\mu}_{\text{ref}}})^\top \Phi\right\| \leq \|d^{\boldsymbol{\mu}_{\text{ref}}}\|_1 \|\Phi\|_\infty \leq \max_{(s,a)} \|\phi(s,a)\|_1 \leq H \max_{(s,a)} \|\phi(s,a)\|_2 \leq H \ .$$

Convexity follows from the direct observation that $\nabla_\theta (d^{\boldsymbol{\mu}_{\text{ref}}})^\top \Phi = 0$. The convexity of $\Theta_{\text{Unil}}(\theta')$ is observed in [Mandal et al., 2025]. □

Next, we provide an upper bound on the error of the estimate returned by the `RobMean` algorithm.

Theorem D.1 (Proposition 1.5 of [Diakonikolas et al., 2020])**.** *Let T be an ϵ-corrupted set of m samples from a distribution in $\mathbb{R}^d$ with mean ρ and covariance Σ. Let ϵ' be in the order of $(\log(1/\delta)/m + \epsilon) \leq c$, for a constant $c > 0$. Then any stability-based algorithm on input T and ϵ', efficiently computes $\widetilde{\rho}$ such that with probability at least $1-\delta$, we have*

$$\|\rho - \widetilde{\rho}\| = O\left(\sqrt{\frac{Tr(\Sigma)\log(Tr(\Sigma)/\|\Sigma)\|}{m}} + \sqrt{\|\Sigma\|\,\epsilon} + \sqrt{\frac{\|\Sigma\|\log(1/\delta)}{m}}\right) \ .$$

Corollary 2. *The `RobMean` algorithm returns a gradient estimate that satisfies*

$$\left\|\widetilde{\nabla}_\theta V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}(s_0) - \nabla_\theta V_{i,0}^{\boldsymbol{\mu}_{\text{ref}}}(s_0)\right\| \leq O\left(\sqrt{\frac{d\log(\text{poly}(d))}{m}} + \sqrt{d\epsilon} + \sqrt{\frac{d\log(1/\delta)}{m}}\right) \ .$$

Proof. Note that, since $\|\phi(s, \boldsymbol{a})\| \leq 1$, for all state action tuples, then $\|\Phi\| \leq d$ and $Tr(\Phi) \leq d$. □

Next, we state an upper bound on the individual regret of each agent when playing Optimistic Hedge.

Theorem D.2 (Theorem 1.1 of [Daskalakis et al., 2021]). *There are constants $C, C' > 1$ so that the following holds. Suppose a time horizon $T \in \mathbb{N}$ and a game G with n players and $|A_i|$ actions for each player $i \in [n]$ is given. Suppose all players play according to Optimistic Hedge with any positive step size $\eta = 1/(Cn \log^4 T)$. Then, for any player $i \in [n]$, the regret of player i satisfies*

$$\mathrm{Reg}_{i,T} \leq O\left(n \cdot \log |A| \cdot \log^4 T\right) .$$

Lemma D.4 (Lemma E.6 of [Mandal et al., 2025]). *Let $y_1 \in W$, and $\eta > 0$. Define the sequence $y_2, \ldots, y_{n+1}$ and $h_1, \ldots, h_n$ such that, for $k = 1, \ldots, n$*

$$y_{k+1} = \mathcal{P}_W\left(y_k - \eta \widehat{h}_k\right) ,$$

and $\widehat{h}_k$ satisfies

$$\mathbb{E}\left[\widehat{h}_k | \mathcal{F}_{k-1}\right] = h_k, \text{ and } \mathbb{E}\left[\left\|\widehat{h}_k\right\|^2 | \mathcal{F}_{k-1}\right] \leq G^2 ,$$

where $\mathcal{F}_k$ are the σ-algebras on which the variables up to k are defined. Then, for any $y^* \in W$, we have

$$\mathbb{E}\left[\sum_{k=1}^{n} \langle y^* - y_k, h_k \rangle\right] \leq \frac{\|y_1 - y^*\|^2}{2\eta} + \frac{\eta n G^2}{2} .$$

E Additional Algorithm Pseudocodes

Algorithm 6 Alternating Minimization (`Trimmed MLE`) for full sample corruption.

Require: Corrupted data D; corruption parameter ϵ; slackness parameter ν.

Split D into equally-sized D_1 and D_2, uniformly at random.
Use D_1 to build a robust estimate $\widehat{\Sigma}$ of the $\Sigma^-_{\boldsymbol{\mu},\boldsymbol{\mu}_{\mathrm{ref}}}$ [Diakonikolas et al., 2025].
Whiten covariates using $\widehat{\Sigma}$, i.e. form $\widetilde{D} = \{\widehat{\Sigma}^{-1/2}(\phi(\tau) - \phi(\tau')) | (\tau, \tau') \in D_2\}$.
Let $\widehat{D} \leftarrow \texttt{Filtering}(\widetilde{D}, \epsilon)$ (Algorithm 4 of [Dong et al., 2019]).
Define $L_\theta(\tau, \tau', o) = \log \sigma\left(o \cdot \theta^\top \widehat{\Sigma}^{1/2}\left(\phi(\tau) - \phi(\tau')\right)\right)$, for $\tau \in \widehat{D}$.
Set $\widetilde{\theta}_0 = 0$.
for $t = 1, 2, \ldots$ **do**
  $\widetilde{S}_t = \arg\max_{\substack{S \subset \widehat{D}: \\ |S| = (1-\epsilon)m}} \sum_{(\tau,\tau',o) \in S} L_{\widetilde{\theta}_t}(\tau, \tau', o)$.
  $\widetilde{\theta}_{t+1} = \arg\max_{\theta : \|\theta\| \leq \sqrt{Hd}} \sum_{(\tau,\tau',o) \in \widetilde{S}_t} L_\theta(\tau, \tau', o)$.
  if $\sum_{(\tau,\tau',o) \in \widetilde{S}_t} L_{\widetilde{\theta}_{t+1}}(\tau, \tau', o) \leq \sum_{(\tau,\tau',o) \in \widetilde{S}_t} L_{\widetilde{\theta}_t}(\tau, \tau', o) + \nu$ **then**
    Return $\widetilde{\theta}_{t+1}$.
  end if
end for

Algorithm 7 Robust Estimation of Value Functions

Require: Dataset D, policy $\boldsymbol{\pi}$, agent i, reward functions $\overline{R}_i$ and $\underline{R}_i$, bonus function $\Gamma(\cdot,\cdot)$.

Initialize $\underline{V}^{\boldsymbol{\pi}}_{i,H}(\cdot) = \overline{V}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,H}(\cdot) = 0$, for all agents $i \in [n]$.
for $h = H-1,\ldots,0$: **do**
  $\underline{\omega}^{\boldsymbol{\pi}}_{i,h} = \texttt{RobEst}\left(\phi(s_h,\boldsymbol{a}_h), \underline{R}_{i,h}(s_h,\boldsymbol{a}_h) + \underline{V}^{\boldsymbol{\pi}}_{i,h+1}(s)\right)$.
  $\overline{\omega}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h} = \texttt{RobEst}\left(\phi(s_h,\boldsymbol{a}_h), \overline{R}_{i,h}(s_h,\boldsymbol{a}_h) + \overline{V}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h+1}(s)\right)$.
  $\underline{Q}^{\boldsymbol{\pi}}_{i,h}(\cdot,\cdot) = \text{Clip}_{\left[-(H-h)\sqrt{d},(H-h)\sqrt{d}\right]}\left(\phi(\cdot,\cdot)^\top \underline{\omega}^{\boldsymbol{\pi}}_{i,h} - \Gamma(\cdot,\cdot)\right)$.
  $\overline{Q}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(\cdot,\cdot) = \text{Clip}_{\left[-(H-h)\sqrt{d},(H-h)\sqrt{d}\right]}\left(\phi(\cdot,\cdot)^\top \overline{\omega}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h} + \Gamma(\cdot,\cdot)\right)$.
  $\underline{V}^{\boldsymbol{\pi}}_{i,h}(s) = \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h}\left[\underline{Q}^{\boldsymbol{\pi}}_{i,h}(s,\boldsymbol{a})\right]$.
  $\overline{V}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(s) = \max_{a_i\in A_i} \mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_{-i}}\left[\overline{Q}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(s,\boldsymbol{a})\right]$.
end for
return Value functions $\overline{V}^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(\cdot)$ and $\underline{V}^{\boldsymbol{\pi}}_{i,h}(\cdot)$, for all $h \in [H-1]$.

Algorithm 8 Optimistic Hedge for n min $-$ max Games (`OptimisticHedge`)

Require: Loss functions $\mathcal{L}_1(\cdot),\ldots,\mathcal{L}_n(\cdot)$; step size ν; steps T.

Initialize policies $\pi_i^{(0)}, \pi_i^{\dagger,(0)}$ uniformly at random, for every $i \in \{1,\ldots,n\}$
for $t = 0,\ldots,T-1$ **do**
  for $i \in \{1,\ldots,n\}$ **do**
    Let $u_i^{(t)}(a^\dagger) = \mathbb{E}_{a'\sim\pi_i^{(t)}}\left[\mathcal{L}_i(a^\dagger,a')\right]$, for every individual action a of player i.
    Let $\ell_i^{(t)}(a') = \mathbb{E}_{a^\dagger\sim\pi_i^{\dagger,(t)}}\left[\mathcal{L}_i(a^\dagger,a')\right]$, for every individual action a of player i.
    for $a_i \in A_i$ **do**
      Update for the $\max$ player:

$$\pi_i^{\dagger,(t+1)}(a^\dagger) = \frac{\pi_i^{\dagger,(t)}(a^\dagger)\cdot\exp\left(\nu\cdot u_i^{(t)}(a^\dagger)\right)}{\sum_{a^\ddagger}\pi_i^{\dagger,(t)}(a^\ddagger)\cdot\exp\left(\nu\cdot u_i^{(t)}(a^\ddagger)\right)}\,.$$

      Update for the $\min$ player:

$$\pi_i^{(t+1)}(a') = \frac{\pi_i^{(t)}(a')\cdot\exp\left(-\nu\cdot \ell_i^{(t)}(a')\right)}{\sum_{a''}\pi_i^{(t)}(a'')\cdot\exp\left(-\nu\cdot \ell_i^{(t)}(a'')\right)}\,.$$

    end for
  end for
end for
return Average policy profile over T rounds $\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\pi}^{(t)}$.