Decoding Rewards in Competitive Games: Inverse Game Theory with Entropy Regularization

Junyi Liao [1] **Zihan Zhu** [2] **Ethan X. Fang** [1] **Zhuoran Yang** [3] **Vahid Tarokh** [1]

Abstract

Estimating the unknown reward functions driving agents' behavior is a central challenge in inverse games and reinforcement learning. This paper introduces a unified framework for reward function recovery in two-player zero-sum matrix games and Markov games with entropy regularization. Given observed player strategies and actions, we aim to reconstruct the underlying reward functions. This task is challenging due to the inherent ambiguity of inverse problems, the non-uniqueness of feasible rewards, and limited observational data coverage. To address these challenges, we establish reward function identifiability using the quantal response equilibrium (QRE) under linear assumptions. Building on this theoretical foundation, we propose an algorithm to learn reward from observed actions, designed to capture all plausible reward parameters by constructing confidence sets. Our algorithm works in both static and dynamic settings and is adaptable to incorporate other methods, such as Maximum Likelihood Estimation (MLE). We provide strong theoretical guarantees for the reliability and sample-efficiency of our algorithm. Empirical results demonstrate the framework's effectiveness in accurately recovering reward functions across various scenarios, offering new insights into decision-making in competitive environments.

1. Introduction

Understanding the underlying reward functions that drive agents' behavior is a central problem in inverse reinforcement learning (IRL) (Ng & Russell, 2000; Arora & Doshi, 2020). While traditional reinforcement learning (RL) (Szepesvári, 2010; Sutton & Barto, 2018) focuses on solving policies based on a known reward function, IRL inverts this process, aiming to infer the reward function from observed behavior. In competitive settings, such as two-player zero-sum games, this problem becomes even more complicated, as the agents' strategies depend not only on their own rewards but also on their opponents' strategies (Wang & Klabjan, 2018; Savas et al., 2019; Wei et al., 2021). These challenges motivate the study of inverse game theory (Lin et al., 2014; Yu et al., 2019), which seeks to recover reward functions from observed strategies in competitive games.

From a practical perspective, inferring the reward functions in competitive games has wide-ranging applications in economics, cyber security, robotics, and autonomous systems (Ng & Russell, 2000; Ziebart et al., 2008). Understanding the motivations behind players' actions in adversarial settings help optimize resource allocation in cyber security (Miehling et al., 2018), model strategic interactions in economic markets (Chow & Djavadian, 2015), or design better AI systems for competitive tasks (Huang et al., 2019).

Meanwhile, recovering reward functions in competitive games involves several key challenges: (i) Inverse problems are inherently ill-posed (Ahuja & Orlin, 2001; Yu et al., 2019), as multiple reward functions can lead to the same optimal strategy and equilibrium solutions. A well-designed algorithm should not merely recover a single reward function but instead identify the entire set of feasible reward functions (Lindner et al., 2022). (ii) In an offline setting (Jarboui & Perchet, 2021), insufficient dataset coverage is also a significant challenge. Observed strategies often fail to comprehensively cover the state-action space, making it difficult to ensure robust reward function recovery. These challenges are further amplified in Markov games (Littman, 1994), where agents' strategies evolve dynamically over time, introducing additional complexity in both reward identification and estimation.

1.1. Major Contributions

We propose a unified framework for inverse game theory that addresses the identification and estimation of reward

[1] Department of Electrical and Computer Engineering, Duke University, Durham NC, USA [2] Department of Statistics and Data Science, University of Pennsylvania, Philadelphia PA, USA [3] Department of Statistics and Data Science, Yale University, New Haven CT, USA. Correspondence to: Junyi Liao <junyi.liao@duke.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

functions in competitive games in both static and dynamic settings. Our contribution is four-fold:

- Identification of Reward Functions: We study the identification problem using the quantal response equilibrium (QRE) under a linear assumption. We formally define the conditions for reward parameter identifiability and characterize the feasible set when parameters are not uniquely identifiable.
- Algorithm for Reward Estimation: Building on the identification results, we propose an algorithm that estimates reward functions by constructing confidence sets to capture all feasible reward parameters.
- Extension to Markov Games: We extend our framework to entropy-regularized Markov games, combining reward recovery with transition kernel estimation to handle dynamic settings. This approach is designed to be sample-efficient and adaptable, incorporating methods like Maximum Likelihood Estimation (MLE).
- Theoretical and Empirical Validation: We provide rigorous theoretical guarantees to establish the reliability and efficiency of our algorithm. Additionally, numerical experiments demonstrate the effectiveness of our framework in accurately recovering reward functions across various competitive scenarios.

1.2. Related Work

Zero-sum Markov Games. The zero-sum Markov game (Shapley, 1953; Xie et al., 2020; Cen et al., 2023; Kalogiannis & Panageas, 2023) models the competitive interactions between two players in dynamic environments. The solution typically focuses on finding equilibrium strategies (Nash Jr, 1951; McKelvey & Palfrey, 1995; Xie et al., 2020) where neither player can unilaterally improve their outcome. With a primary focus on learning in a sample-efficient manner, learning algorithms are proposed, including policy-based methods (Cen et al., 2021; Wei et al., 2021; Zhao et al., 2022; Cen et al., 2023) and value-based methods (Xie et al., 2020; Chen et al., 2022; Kalogiannis & Panageas, 2023).

Inverse Optimization and Inverse Reinforcement Learning (IRL). Inverse optimization (Ahuja & Orlin, 2001; Chan et al., 2022; Ahmadi et al., 2023) reverses the traditional optimization process by taking observed decisions as input to infer an objective function (Ahuja & Orlin, 2001; Nourollahi & Ghate, 2018) and constraints (Chan & Kaw, 2019; Ghobadi & Mahmoudzadeh, 2021) that make these decisions approximately or exactly optimal. In practice, inverse optimization offers a powerful framework for understanding and modeling decision-making in complex systems across fields like marketing (Chow & Djavadian, 2015; Vatandoust et al., 2023), operations research (Brotcorne et al., 2005; Agarwal & Özlem Ergun, 2010; Yu et al., 2021), and machine learning (Konstantakopoulos et al., 2017; Dong et al., 2018; Tan et al., 2019).

Inverse reinforcement learning (Ng & Russell, 2000; Ziebart et al., 2008; Herman et al., 2016; Wulfmeier et al., 2016; Arora & Doshi, 2020) focuses on inferring the reward function based on the observed behavior or strategy of agents and experts, which is crucial for understanding various decision-making processes, from single-agent processes (Boularias et al., 2011; Herman et al., 2016; Fu et al., 2018) to competitive or cooperative games (Vorobeychik et al., 2007; Ling et al., 2018; Wang & Klabjan, 2018; Wu et al., 2024). A popular approach within the field of IRL is the Maximum Entropy IRL (Ziebart et al., 2008; Ziebart, 2018; Wulfmeier et al., 2016; Snoswell et al., 2020), which is based on the principle of maximum entropy and is provably efficient in handling uncertainty of agent behaviors (Snoswell et al., 2020; Gleave & Toyer, 2022) and high-dimensional observations (Wulfmeier et al., 2016; Snoswell et al., 2020; Song et al., 2022).

Entropy Regularization in RL and Games. We use the entropy regularization in our framework, which has become a widely used technique in reinforcement learning (Szepesvári, 2010; Ziebart, 2018) and game theory (Savas et al., 2019; Guan et al., 2021; Cen et al., 2023). Entropy regularization is provably effective in addressing challenges like exploration-exploitation tradeoff (Haarnoja et al., 2018; Wang et al., 2019; Ahmed et al., 2019; Neu et al., 2017), algorithm robustness (Zhao et al., 2020; Guo et al., 2021) and convergence acceleration (Cen et al., 2021; Cen et al., 2023; Zhan et al., 2023). Importantly, entropy regularization has also been shown to improve identifiability in inverse reinforcement learning (IRL) problems. Recent works in single-agent IRL, such as Cao et al. (2021) and Rolland et al. (2022), leverage entropy-regularized policies to transform ill-posed IRL problems into identifiable ones under mild assumptions. Our work builds on this insight by extending it to competitive multi-agent settings, where identifiability becomes even more subtle due to strategic interactions.

Paper Organization. In §2, we develop the framework of inverse game theory for entropy-regularized zero-sum games. In §3, we extend the framework introduced in §2 to a sequential decision-making setting, focusing on entropy-regularized zero-sum Markov games. We provide numerical experiments to validate the theoretical findings in §4, and conclude the paper in §5.

Notations. We introduce some useful notation before proceeding. Throughout this paper, we denote the set $1, 2, \cdots, n$ by $[n]$ for any positive integer n. For two positive sequences $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$, we write $a_n = \mathcal{O}(b_n)$ or $a_n \lesssim b_n$ if there exists a positive constant C such that $a_n \leq C \cdot b_n$. For any integer d, we denote the d-

dimensional Euclidean space by $\mathbb{R}^d$, with inner product $\langle x, y\rangle = x^\top y$ and the induced norm $\|x\| = \sqrt{\langle x, x\rangle}$. For any matrix $A = (a_{ij})$, the Frobenius norm of A is $\|A\|_{\mathrm{F}} = (\sum_{i,j} a_{ij}^2)^{1/2}$, and the operator norm (or spectral norm) of A is $\|A\|_{\mathrm{op}} = \sigma_1(A)$, where $\sigma_1(A)$ stands for the largest singular value of A. For any square matrix $A = (a_{ij})$, denote its trace by $\mathrm{tr}(A) = \sum_i a_{ii}$. For a nonempty set $\mathcal{X}$, we denote by $\Delta(\mathcal{X})$ the space of all probability distributions on $\mathcal{X}$.

2. Entropy-Regularized Zero-Sum Matrix Games

We derive the inverse game theory for entropy-regularized two-player zero-sum matrix games. We consider the identification problem of payoff matrices under the linear parametric assumption and derive a necessary and sufficient condition for strong identification. Furthermore, we propose methods to recover identified sets and payoff matrices.

2.1. Preliminary and Problem Formulation

We consider a two-player zero-sum matrix game, which is specified by a triple $(\mathcal{A}, \mathcal{B}, Q)$, where $\mathcal{A} = \{1, 2, \cdots, m\}$ and $\mathcal{B} = \{1, 2, \cdots, n\}$ are finite sets of actions that players $i \in \{1, 2\}$ can take, and $Q(\cdot, \cdot)$ is the payoff function. The zero-sum game can be formulated as the following min-max optimization problem

$$\max_{\mu} \min_{\nu} \mu^\top Q \nu,$$

where $\mu \in \Delta(\mathcal{A})$ and $\nu \in \Delta(\mathcal{B})$ are policies for each player, and $Q = (Q(a, b))_{a\in\mathcal{A}, b\in\mathcal{B}} \in \mathbb{R}^{m\times n}$ denotes the payoff matrix. The solution of this optimization problem is also known as the Nash equilibrium (Nash Jr, 1951), where both agents play the best response against the other agent.

Entropy-Regularized Two-Player Zero-Sum Matrix Game. We study the entropy-regularized matrix game. Formally, this amounts to solving the following matrix game with entropy regularization (Mertikopoulos & Sandholm, 2016):

$$\max_{\mu} \min_{\nu} \mu^\top Q \nu + \eta^{-1}\mathcal{H}(\mu) - \eta^{-1}\mathcal{H}(\nu),$$

where $\eta > 0$ is the regluarization parameter, and

$$\mathcal{H}(\pi) = -\sum_i \pi_i \log(\pi_i)$$

denotes the Shannon entropy (Shannon, 1948) of π. According to the von-Neumann minimax theorem (von Neumann, 1928), there exists a unique solution (μ^*, ν^*) to this min-max problem, denoted as the quantal response equilibrium (McKelvey & Palfrey, 1995), which satisfies the following fixed point equations:

$$\begin{cases} \mu^*(a) = \dfrac{e^{\eta Q(a,\cdot)\nu^*}}{\sum_{a\in\mathcal{A}} e^{\eta Q(a,\cdot)\nu^*}}, & \text{for all } a \in \mathcal{A}, \\ \nu^*(b) = \dfrac{e^{-\eta Q(\cdot,b)^\top \mu^*}}{\sum_{b\in\mathcal{B}} e^{-\eta Q(\cdot,b)^\top \mu^*}}, & \text{for all } b \in \mathcal{B}. \end{cases}$$

This non-linear system is equivalent to the following $m + n - 2$ linear constraints: for all $a \in \mathcal{A}$ and $b \in \mathcal{B}$,

$$\begin{cases} \big(Q(a,\cdot) - Q(1,\cdot)\big)\nu^* = \log(\mu^*(a)/\mu^*(1))/\eta, \\ \big(Q(\cdot,b) - Q(\cdot,1)\big)^\top \mu^* = -\log(\nu^*(b)/\nu^*(1))/\eta. \end{cases} \tag{1}$$

Goal. We study the inverse game theory for this entropy-regularized zero-sum game. To elaborate, we observe strategy pairs $(a^k, b^k) \overset{\text{iid}}{\sim} (\mu^*, \nu^*)$ follows the QRE, and we aim to recover all the feasible payoff functions $Q(\cdot, \cdot)$.

Identification of payoff matrices. To derive inverse game theory, it is important to study the identifiability of the payoff matrix, i.e. if there exists a unique payoff matrix that satisfies the QRE constraint. In this paper, we study the identification problem under the linear structure assumption (§2.2) and further generalize the analysis to the partial identification case (§2.3).

2.2. Strong Identification

Suppose (μ^*, ν^*) are the QRE for two players and we use the observed data to obtain an estimation denoted by $(\widehat{\mu}, \widehat{\nu})$. Next, we are going to estimate the payoff matrix from this estimated QRE. To ensure the game is identifiable, we leverage the following linear parametric assumption.

Assumption 2.1 (Linear payoff functions)**.** Suppose that there exists a vector-valued kernel $\phi : \mathcal{A} \times \mathcal{B} \to \mathbb{R}^d$ and a vector $\theta^* \in \mathbb{R}^d$ such that $\|\theta^*\| \le M$ for some $M > 0$, and

$$Q(a, b) = \langle \phi(a, b), \theta^* \rangle$$

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

To estimate the payoff matrix Q from the observed data, our essential goal is to estimate θ^*. Under Assumption 2.1, the linear system (1) can be rewritten as follows: for all $a \in \mathcal{A}$ and $b \in \mathcal{B}$,

$$\begin{cases} \big\langle (\phi(a,\cdot) - \phi(1,\cdot))\,\nu^*, \theta \big\rangle = \log(\mu^*(a)/\mu^*(1))/\eta, \\ \big\langle (\phi(\cdot,b) - \phi(\cdot,1))^\top \mu^*, \theta \big\rangle = -\log(\nu^*(b)/\nu^*(1))/\eta, \end{cases}$$

where $(\phi(a,\cdot) - \phi(1,\cdot))\,\nu^*, (\phi(\cdot,b) - \phi(\cdot,1))^\top \mu^* \in \mathbb{R}^d$. To simplify the notation, we define matrices

$$A(\nu) = ((\phi(a,\cdot) - \phi(1,\cdot))\,\nu)_{a\in\mathcal{A}/\{1\}} \in \mathbb{R}^{(m-1)\times d},$$
$$B(\mu) = ((\phi(\cdot,b) - \phi(\cdot,1))^\top \mu)_{b\in\mathcal{B}/\{1\}} \in \mathbb{R}^{(n-1)\times d},$$

and define vectors

$$c(\mu) = (\log(\mu(a)/\mu(1))/\eta)_{a\in\mathcal{A}/\{1\}} \in \mathbb{R}^{m-1},$$
$$d(\nu) = (-\log(\nu(b)/\nu(1))/\eta)_{b\in\mathcal{B}/\{1\}} \in \mathbb{R}^{n-1}$$

Then the linear constraints would be

$$\begin{bmatrix} A(\nu^*) \\ B(\mu^*) \end{bmatrix} \theta = \begin{bmatrix} c(\mu^*) \\ d(\nu^*) \end{bmatrix}. \tag{2}$$

Since the linear system has $m+n-2$ constraints and the dimension of θ is d. Intuitively, if $d \le m+n-2$ and the linear constraints are full rank, there is at most one solution of the above linear equations.

Proposition 2.2 (Necessary and sufficient condition for strong identification)**.** *Under Assumption 2.1, there is a unique $\theta \in \mathbb{R}^d$ such that $Q(a,b) = \langle \phi(a,b), \theta \rangle$ (i.e. $\theta = \theta^*$) for all $(a,b) \in \mathcal{A} \times \mathcal{B}$ if and only if the QRE satisfies the rank condition*

$$\textit{rank}\left(\begin{bmatrix} A(\nu^*) \\ B(\mu^*) \end{bmatrix}\right) = d. \tag{3}$$

Let the rank condition (3) hold, so that the game is strongly identifiable. In an offline setting, we propose a two-step method to estimate θ^*.

1. Estimate the QRE (μ^*, ν^*) from the observed data and obtain $(\widehat{\mu}, \widehat{\nu})$.
2. Leverge (2) to estimate θ. To be specific, we conduct the least-square estimation and obtain $\widehat{\theta}$:

$$\widehat{\theta} := \mathop{\arg\min}_{\theta\in\mathbb{R}^d} \left\| \begin{bmatrix} A(\widehat{\nu}) \\ B(\widehat{\mu}) \end{bmatrix} \theta - \begin{bmatrix} c(\widehat{\mu}) \\ d(\widehat{\nu}) \end{bmatrix} \right\|^2, \tag{4}$$

If the sample size is sufficiently large and $\mathrm{TV}(\widehat{\mu}, \mu^*)$ and $\mathrm{TV}(\widehat{\nu}, \nu^*)$ are close to zero, the coefficient matrix in (4) is of full column rank, and we can derive a closed form for $\widehat{\theta}$:

$$\widehat{\theta} = \left(\begin{bmatrix} A(\widehat{\nu}) \\ B(\widehat{\mu}) \end{bmatrix}^\top \begin{bmatrix} A(\widehat{\nu}) \\ B(\widehat{\mu}) \end{bmatrix} \right)^{-1} \begin{bmatrix} A(\widehat{\nu}) \\ B(\widehat{\mu}) \end{bmatrix}^\top \begin{bmatrix} c(\widehat{\mu}) \\ d(\widehat{\nu}) \end{bmatrix}. \tag{5}$$

Next, we derive the estimation error of the two-step method. Namely, given a finite sample bound for $\mathrm{TV}(\widehat{\mu}, \mu^*)$ and $\mathrm{TV}(\widehat{\nu}, \nu^*)$, we aim to derive $\|\widehat{\theta} - \theta^*\|$.

Theorem 2.3 (Parameter estimation error)**.** *Let ϵ_1 and ϵ_2 be two small numbers satisfying $\epsilon_1 < \min_{a\in[m]} \mu^*(a)$ and $\epsilon_2 < \min_{b\in[n]} \nu^*(b)$. Under Assumption 2.1 and the rank condition in (3), suppose $(\hat{\mu}, \hat{\nu})$ satisfies TV$(\hat{\mu}, \mu^*) \le \epsilon_1/2$ and TV$(\hat{\nu}, \nu^*) \le \epsilon_2/2$, then $\hat{\theta}$ constructed by (4) satisfies*

$$\|\hat{\theta} - \theta^*\|^2 \lesssim \epsilon_1^2 \cdot \big(1 + m \cdot (\epsilon_2^2 + 1)\big) + \epsilon_2^2 \cdot \big(1 + n \cdot (\epsilon_1^2 + 1)\big).$$

Proof. See Appendix A.1 for the complete proof. □

Now we present the finite sample result of the sample complexity. In the two-step method, given a dataset of agent actions following the true QRE, we first use a consistent estimator to approximate the true QRE and obtain $\widehat{\mu}, \widehat{\nu}$, then we use the estimated QRE to conduct the least square (5). Therefore, the sample complexity would be dependent on the convergence rate of the QRE estimator. A natural choice for QRE estimation is the frequency estimator.

Theorem 2.4 (Finite sample error bound)**.** *Given N samples $\{(a^k, b^k)\}_{k\in[N]}$ following the true QRE (μ^*, ν^*), we obtain $\widehat{\mu}, \widehat{\nu}$ by the frequency estimator. For any $\delta \in (0,1)$, the estimation error bound of the payoff matrix holds with probability at least $1-\delta$*

$$\|\widehat{Q} - Q\|_F^2 \lesssim \mathcal{O}\left(\frac{m^2 + n^2 + (m+n)\log(1/\delta)}{N} \right).$$

Proof. See Appendix A.2 for the complete proof. □

2.3. Partial Identification

If the rank condition (3) does not hold, there are infinitely many $\theta \in \mathbb{R}^d$ that satisfy the QRE constraint (2). Under Assumption 2.1, the identified set $\Theta \subset \mathbb{R}^d$ is

$$\Theta = \left\{ \theta : \begin{bmatrix} A(\nu^*) \\ B(\mu^*) \end{bmatrix} \theta = \begin{bmatrix} c(\mu^*) \\ d(\nu^*) \end{bmatrix}, \|\theta\| \le M \right\}.$$

Since the true parameter θ^* is partially identified, we construct a confidence set that contains the identified set with high probability. Given N strategy pairs following the true QRE, we first estimate the QRE from the observed data by frequency estimators $\widehat{\mu}$ and $\widehat{\nu}$. Next, we select a threshold $\kappa_N > 0$ and construct the confidence set as follows:

$$\widehat{\Theta}_N = \left\{ \theta : \left\| \begin{bmatrix} A(\widehat{\nu}) \\ B(\widehat{\mu}) \end{bmatrix} \theta - \begin{bmatrix} c(\widehat{\mu}) \\ d(\widehat{\nu}) \end{bmatrix} \right\|^2 \le \kappa_N, \|\theta\| \le M \right\}. \tag{6}$$

To recover the feasible payoff functions, we simply compute $\widehat{Q}(a,b) = \phi(a,b)^\top \widehat{\theta}$ for all $\widehat{\theta} \in \widehat{\Theta}$ according to the linear assumption. We summarize the procedure in Algorithm 1 (See Appendix A.3).

We demonstrate the effectiveness of Algorithm 1 by establishing its ability to construct accurate confidence sets. To be specific, we show that the confidence set $\widehat{\Theta}$ is close to the identified set Θ when the sample size N is large. The key to approximating feasible set Θ is to identify a suitable threshold κ_N that makes the confidence set $\widehat{\Theta}_N$ "similar" to Θ. The following theorem formalizes this intuition.

Theorem 2.5 (Convergence of confidence set)**.** *Let Assumption 2.1 hold. For each $N \in \mathbb{N}$, suppose we observe N samples $\{(a^k, b^k)\}_{k\in[N]}$ following the true QRE (μ^*, ν^*), and calculate $(\widehat{\mu}, \widehat{\nu})$ by the frequency estimator. Set the* *confidence set $\widehat{\Theta}_N$ as in (6), where $\kappa_N = \mathcal{O}(N^{-1})$. Then with probability at least $1-\delta$,*

$$d_H(\Theta, \widehat{\Theta}_N) \lesssim \frac{m+n+\sqrt{(m+n)\log(1/\delta)}}{\sqrt{N}}, \quad (7)$$

where d_H is the Hausdorff distance corresponding to the Euclidean distance in $\mathbb{R}^d$.

Proof. See Appendix A.3 for the complete proof. □

Theorem 2.5 establishes the asymptotic consistency of our confidence set $\widehat{\Theta}_N$ in the finite-sample setting, showing that it converges to the true feasible set Θ as the number of observed samples increases. The finite-sample bound (7) demonstrates that the estimation error decreases at the rate of $\mathcal{O}(N^{-1/2})$, which matches the standard concentration rate for empirical frequency estimators. The dependence on m and n highlights that larger action spaces require more samples for the same level of confidence. This result confirms that our method provides both statistical consistency and a well-characterized finite-sample error bound, making it a robust approach for inverse game-theoretic inference.

2.4. Selection in Confidence Sets

As discussed in §2.3, the true parameter θ^* is partially identifiable when the rank condition (3) does not hold, and there are infinitely many parameters that lead to the same QRE. To avoid unnecessary large coefficients that might overfit or lead to instability, we define the optimal solution θ^* as the vector that satisfies the QRE constraints and has the minimum Euclidean norm, i.e.,

$$\theta^* = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \|\theta\|, \quad \text{subject to} \begin{bmatrix} A(\nu^*) \\ B(\mu^*) \end{bmatrix} \theta = \begin{bmatrix} c(\mu^*) \\ d(\nu^*) \end{bmatrix}.$$

When the system is not full column rank, the minimum-norm solution of least square is uniquely determined by the Moore–Penrose inverse (Ben-Israel & Greville, 2006):

$$\theta^* = \begin{bmatrix} A(\nu^*) \\ B(\mu^*) \end{bmatrix}^{\dagger} \begin{bmatrix} c(\mu^*) \\ d(\nu^*) \end{bmatrix}.$$

Therefore, to estimate the optimal parameter θ^*, we propose the following plug-in estimator:

$$\widehat{\theta} = \begin{bmatrix} A(\widehat{\nu}) \\ B(\widehat{\mu}) \end{bmatrix}^{\dagger} \begin{bmatrix} c(\widehat{\mu}) \\ d(\widehat{\nu}) \end{bmatrix}.$$

Now we derive the estimation error bound $\|\widehat{\theta} - \theta^*\|$.

Theorem 2.6 (Convergence of the optimal QRE solution)**.** *Assume that the matrix*

$$X = \begin{bmatrix} A(\nu^*) \\ B(\mu^*) \end{bmatrix} \in \mathbb{R}^{(m+n-2)\times d}$$

is of full row rank, and its smallest singular value is bounded from below, that is, $\sigma_{m+n-2}(X) \geq \sigma_b$ for some $\sigma_b > 0$. Given N samples $\{(a^k, b^k)\}_{k\in[N]}$ following the true QRE (μ^*, ν^*), we obtain $(\widehat{\mu}, \widehat{\nu})$ by the frequency estimator. For any $\delta \in (0,1)$, when N is sufficiently large, the following estimation error bound of the optimal QRE solution holds with probability at least $1-\delta$:

$$\|\widehat{\theta} - \theta^*\| \lesssim \frac{m+n+\sqrt{(m+n)\log(1/\delta)}}{\sqrt{N}}.$$

Proof. See Appendix A.4 for the complete proof. We also discuss the assumption of this Theorem in Remark A.3. □

In practice, selecting the minimum-norm solution helps avoid overfitting and promotes stability (Hastie et al., 2009). The convergence rate $\mathcal{O}(N^{-1/2})$ matches standard results in statistical estimation, showing the reliability and efficiency of our method in practical settings.

3. Entropy-Regularized Zero-Sum Markov Games

In this section, we follow the same methodology in §2 and derive the inverse game theory for entropy-regularized two-player zero-sum Markov games.

3.1. Preliminary and Problem Formulation

We briefly review the setting of a two-player zero-sum Markov game (Littman, 1994), which is a framework that extends Markov decision processes (MDPs) to multi-agent settings, where two players with opposing objectives interact in a shared environment. A two-player zero-sum simultaneous-move episodic Markov game is defined by a sextuple $(\mathcal{S}, \mathcal{A}, \mathcal{B}, r, \mathbb{P}, H)$, where

- $\mathcal{S}$ is the state space, with $|\mathcal{S}| = S$,
- $\mathcal{A}$ and $\mathcal{B}$ are two finite sets of actions that players $i \in \{1, 2\}$ can take,
- $H \in \mathbb{N}$ is the number of time steps,
- $r = \{r_h\}_{h\in[H]}$ is a collection of reward functions, and
- $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$ is a collection of transition kernels.

At each time step $h \in [H]$, the players 1 and 2 simultaneously take actions $a \in \mathcal{A}$ and $b \in \mathcal{B}$ respectively upon observing the state $s \in \mathcal{S}$, and then player 1 receives the reward $r_h(s,a,b)$, while player 2 receives $-r_h(s,a,b)$. Namely, the gain of one player equals the loss of the other. The system then transitions to a new state $s' \sim \mathbb{P}_h(\cdot|s,a,b)$ according to the transition kernel $\mathbb{P}_h$.

Entropy-regularized two-player zero-sum Markov game. We study the two-player zero-sum Markov game with entropy regularization. We use (μ, ν) to denote the policy of two players, where $\mu = \{\mu_h\}_{h=1}^H$ and $\nu = \{\nu_h\}_{h=1}^H$. At step h, the entropy-regularized V-function is

$$V_h^{\mu,\nu}(s) = \mathbb{E}\biggl[\sum_{t=h}^{H} \gamma^{t-h}\bigl[r_t(s_t, a_t, b_t) - \eta^{-1}\log\mu_t(a_t|s_t) + \eta^{-1}\log\nu_t(b_t|s_t)\bigr]\bigg| s_h = s\biggr],$$

where $\gamma \in [0, 1]$ is the discount factor and $\eta > 0$ is the parameter of regularization. Meanwhile,, we define the entropy-regularized Q-function that

$$Q_h^{\mu,\nu}(s, a, b) = r_h(s, a, b) + \gamma\mathbb{E}_{\mathbb{P}_h(\cdot|s,a,b)}\left[V_{h+1}^{\mu,\nu}(\cdot)\right]. \quad (8)$$

For notation simplicity, we denote by $Q_h^{\mu,\nu}(s) \in \mathbb{R}^{m\times n}$ the collection of Q-functions at the state s, which is the matrix $[Q_h^{\mu,\nu}(s, a, b)]_{(a,b)\in\mathcal{A}\times\mathcal{B}}$. With this notation, we may write

$$\begin{aligned} V_h^{\mu,\nu}(s) &= \mu_h(s)^\top Q_h^{\mu,\nu}(s)\nu_h(s) \\ &\quad + \eta^{-1}\mathcal{H}(\mu_h(s)) - \eta^{-1}\mathcal{H}(\nu_h(s)). \end{aligned} \quad (9)$$

The equations (8) and (9) are also known as Bellman equations for Markov games. In a zero-sum game, one player seeks to maximize the value function while the other player wants to minimize it:

$$V_1^*(s) = \max_\mu \min_\nu V_1^{\mu,\nu}(s) = \min_\nu \max_\mu V_1^{\mu,\nu}(s).$$

Definition 3.1 (Quantal response equilibrium)**.** For each time step h, there is a unique pair of optimal policies (μ_h^*, ν_h^*) of the entropy-regularized Markov game, i.e. the quantal response equilibrium (QRE), characterized by the following minimax problem:

$$V_h^{\mu^*,\nu^*}(s) = \max_{\mu_h}\min_{\nu_h} V_h^{\mu,\nu}(s) = \min_{\nu_h}\max_{\mu_h} V_h^{\mu,\nu}(s).$$

which is equivalent to

$$\begin{aligned} V_h^{\mu^*,\nu^*}(s) &= \max_{\mu_h}\min_{\nu_h} \mu_h(s)^\top Q_h^{\mu,\nu}(s)\nu_h(s) \\ &\quad + \eta^{-1}\mathcal{H}(\mu_h(s)) - \eta^{-1}\mathcal{H}(\nu_h(s)), \end{aligned} \quad (10)$$

where $\mu_h : \mathcal{S} \to \Delta(\mathcal{A})$ is the policy followed by player 1 and $\nu_h : \mathcal{S} \to \Delta(\mathcal{B})$ is the policy followed by player 2, and $\mathcal{H}(\pi) := -\sum_i \pi_i \log(\pi_i)$ denotes the Shannon entropy of a distribution π. Also, it is known that the unique solution of this minimax problem (QRE) satisfies the following fixed point equations:

$$\begin{cases} \mu_h^*(a|s) = \dfrac{e^{\eta\langle Q_h^*(s,a,\cdot),\nu_h^*(\cdot|s)\rangle_{\mathcal{B}}}}{\sum_{a\in\mathcal{A}} e^{\eta\langle Q_h^*(s,a,\cdot),\nu_h^*(\cdot|s)\rangle_{\mathcal{B}}}}, \ \forall a \in \mathcal{A}, \\ \nu_h^*(b|s) = \dfrac{e^{-\eta\langle Q_h^*(s,\cdot,b),\mu_h^*(\cdot|s)\rangle_{\mathcal{A}}}}{\sum_{b\in\mathcal{B}} e^{-\eta\langle Q_h^*(s,\cdot,b),\mu_h^*(\cdot|s)\rangle_{\mathcal{A}}}}, \ \forall b \in \mathcal{B}. \end{cases} \quad (11)$$

Goal. We study the inverse game theory for this entropy-regularized two-player zero-sum Markov game, where both the rewards (r_h) and the transition kernels $(\mathbb{P}_h)$ are unknown. To elaborate, we observe i.i.d. trajectories

$$\{(s_1^t, a_1^t, b_1^t), \cdots, (s_H^t, a_H^t, b_H^t)\}_{t\in[T]}$$

following the QRE (μ^*, ν^*), and we aim to recover all the feasible reward functions r defined as follows.

Definition 3.2 (Identified reward sets)**.** Given state and action space $\mathcal{S} \times \mathcal{A} \times \mathcal{B}$ and quantal response equilibrium (μ^*, ν^*), a reward function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}^H$ is identified if μ_h, ν_h is the solution of the minimax problem (10) induced by the reward function r_h for all $h \in [H]$.

3.2. Learning Reward Functions from Actions

In this section, we propose an algorithm to find all the feasible reward functions that lead to the QRE. We assume that both the reward function and transition kernel have a linear structure (Bradtke & Barto, 2004; Jin et al., 2020).

Assumption 3.3 (Linear MDP)**.** For the underlying MDP, we assume that for every reward function $r_h : \mathcal{S}\times\mathcal{A}\times\mathcal{B} \to [0, 1]$ and every transition kernel $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$, there exist $\omega_h \in \mathbb{R}^d$ and $\pi_h(\cdot) : \mathcal{S} \to \mathbb{R}^d$ such that

$$r_h(s, a, b) = \phi(s, a, b)^\top\omega_h,$$
$$\mathbb{P}_h(\cdot|s, a, b) = \phi(s, a, b)^\top\pi_h(\cdot)$$

for all $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$. In addition, the Q function is linear with respect to ϕ. Namely, for any QRE (μ, ν) and $h \in [H]$, there exists a vector $\theta_h \in \mathbb{R}^d$ such that

$$Q_h(s, a, b) = \phi(s, a, b)^\top\theta_h.$$

We assume $\|\phi(\cdot, \cdot, \cdot)\| \le 1$, $\|\theta_h\| \le R$, and $\|\pi_h(s)\| \le \sqrt{d}$ for all $h \in [H]$ and $s \in \mathcal{S}$.

Remark 3.4. In Assumption 3.3, since the reward functions r_h are normalized to the unit interval $[0, 1]$ and the number of time steps $[H]$ is finite, every Q-function Q_h must be bounded by some constant, and the constant $R \ge H(1 + \log m + \log n)$ exists. Since (ω_h) can be recovered by (θ_h), we prefer to make an assumption on (θ_h) instead of (ω_h) for the convenience of subsequent analysis.

We are going to find all the feasible ω_h for all $h \in [H]$ under Assumption 3.3. Analogous to matrix games, we first consider the identification problem of the Q-function. Namely, whether there is a unique θ_h corresponding to the QRE. Given the equilibrium constraint (11), we propose the following theorem for strong identification.

Proposition 3.5 (Strong identification of Q-function)**.** *Under Assumption 3.3, for each $h \in [H]$, the Q-function $Q_h(s, a, b) = \phi(s, a, b)^\top\theta_h$ is feasible for all $(s, a, b) \in$*

$\mathcal{S}\times\mathcal{A}\times\mathcal{B}$ if θ_h satisfies the following linear system:

$$\begin{bmatrix} A_h(s,\nu_h^*) \\ B_h(s,\mu_h^*) \end{bmatrix}\theta_h = \begin{bmatrix} c_h(s,\mu_h^*) \\ d_h(s,\nu_h^*) \end{bmatrix} \quad \textit{for all } s\in\mathcal{S}, \tag{12}$$

where

$$\begin{aligned} A_h(s,\nu_h) &= \left(\left(\phi(s,a,\cdot)-\phi(s,1,\cdot)\right)\nu_h(\cdot|s)\right)_{a\in\mathcal{A}\setminus\{1\}}, \\ B_h(s,\mu_h) &= \left(\left(\phi(s,\cdot,1)-\phi(s,\cdot,b)\right)\mu_h(\cdot|s)\right)_{b\in\mathcal{B}\setminus\{1\}} \end{aligned}$$

and

$$\begin{aligned} c_h(s,\mu_h) &= \left(\eta^{-1}\log\frac{\mu_h(a|s)}{\mu_h(1|s)}\right)_{a\in\mathcal{A}\setminus\{1\}} \in\mathbb{R}^{m-1}, \\ d_h(s,\nu_h) &= \left(-\eta^{-1}\log\frac{\nu_h(b|s)}{\nu_h(1|s)}\right)_{b\in\mathcal{B}\setminus\{1\}} \in\mathbb{R}^{n-1}. \end{aligned}$$

Moreover, there exists a unique $\theta_h\in\mathbb{R}^d$ if and only if the QRE satisfies the rank condition

$$\textit{rank}\left(\begin{bmatrix} A_h(\nu_h^*)^\top & B_h(\mu_h^*)^\top \end{bmatrix}\right) = d, \tag{13}$$

where

$$A_h(\nu_h) := \begin{bmatrix} A_h(1,\nu_h) \\ A_h(2,\nu_h) \\ \vdots \\ A_h(|\mathcal{S}|,\nu_h) \end{bmatrix}, \quad B_h(\mu_h) := \begin{bmatrix} B_h(1,\mu_h) \\ B_h(2,\mu_h) \\ \vdots \\ B_h(|\mathcal{S}|,\mu_h) \end{bmatrix}.$$

Proof. See Appendix B.1 for the complete proof. □

Following the Bellman equation (8), r_h is a feasible reward function iff there exists a feasible Q function Q_h and V function V_{h+1} such that

$$r_h(s,a,b) = Q_h(s,a,b) - \gamma\mathbb{E}_{\mathbb{P}_h(\cdot|s,a,b)}\left[V_{h+1}(\cdot)\right]. \tag{14}$$

Next, we propose an algorithm to recover the feasible reward functions. For all $h\in[H]$, the algorithm performs the following four steps:

- Recover the feasible set by solving the least square problem associated with the linear system (12):

$$\widehat{\Theta}_h = \left\{\|\theta\|\le R : \left\|\begin{bmatrix} A_h(\widehat{\nu}_h) \\ B_h(\widehat{\mu}_h)\end{bmatrix}\theta - \begin{bmatrix} c_h(\widehat{\mu}_h) \\ d_h(\widehat{\nu}_h)\end{bmatrix}\right\|^2 \le \kappa_h\right\}. \tag{15}$$

- Calculate the feasible Q and V functions (Q_h and V_h) for all $\widehat{\theta}_h\in\widehat{\Theta}_h$.

- Estimate the transition kernel $\mathbb{P}_h$ from the observed data. Since the transition kernel has a linear structure, we employ ridge regression for estimation:

$$\begin{aligned} \Lambda_h &= \sum_{t=1}^{T}\phi(s_h^t,a_h^t,b_h^t)\phi(s_h^t,a_h^t,b_h^t)^\top + \lambda\mathbf{I}_d, \\ \widehat{\mathbb{P}}_h\widehat{V}_{h+1}(s,a,b) &= \phi(s,a,b)^\top\Lambda_h^{-1} \\ &\quad\times\sum_{t=1}^{T}\phi(s_h^t,a_h^t,b_h^t)\widehat{V}_{h+1}(s_{h+1}^t); \end{aligned}$$

- Recover feasible set $\mathcal{R}_h$ by the Bellman equation (14).

We provide the pseudocode in Alg. 2 in Appendix §B.3.

3.3. Theoretical Guarantees

In this section, we present the theoretical results for Algorithm 2. To begin with, we define the base metric to measure the distance between rewards.

Definition 3.6 (Uniform metric for rewards)**.** We define the metric d between any pair of rewards r, r' as

$$D(r,r') = \sup_{(h,s,a,b)\in[H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{B}}\left|(r_h-r'_h)(s,a,b)\right|.$$

We aim to recover the feasible reward set defined below.

Definition 3.7 (Feasible reward set)**.** We say a reward function $r=(r_1,r_2,\cdots,r_H)$ is feasible with respect to a quantal response equilibrium μ and ν if the Q function $Q=(Q_1,Q_2,\cdots,Q_H)$ satisfies the identifability condition (11) and the norm constraint $\|\theta_h^*\|\le R$. We denote $\mathcal{R}$ as the feasible reward set corresponding to the quantal response equilibrium μ and ν, namely,

$$\mathcal{R} := \Bigg\{ r=(r_1,r_2,\cdots,r_H) : r \text{ is identified and } \left\|\omega_h + \gamma\sum_{s\in\mathcal{S}}\pi_h(s)V_{h+1}(s)\right\| \le R \text{ for all } h\in[H]\Bigg\}.$$

Also, we denote $\mathcal{Q}$ as the feasible Q function set:

$$\mathcal{Q} = \{(Q_h)_{h=1}^H : Q \text{ is identified and } \|\theta_h\|\le R,\ \forall h\in[H]\}.$$

Our formulation provides a principled way to handle partial identifiability in Markov games. Instead of forcing a single estimated reward function, we construct a structured set of feasible rewards, which offers a more robust approach to analyzing decision-making in complex multi-step strategic settings. Intuitively, the norm constraint $\|\theta_h\|\le R$ plays a key role in ensuring that the estimated reward functions remain well-conditioned, and do not include arbitrarily large coefficients. Additionally, by linking the feasible reward set to the recursive Bellman equations (8)-(9), our definition ensures that every element of $\widehat{\mathcal{R}}$ maintains temporal consistency. In other words, the inferred rewards lead to equilibrium strategies that are valid over multiple decision-making steps.

For the sake of clarity, we fix the initial state distribution in the Markov game $\rho_1 \in \Delta(\mathcal{S})$, and define the marginal state visitation distributions associated with policies μ, ν at each time step $h \in [H]$ as $d_h^{\mu,\nu}(s) = \mathbb{P}(s_h = s|\rho_1, \mu, \nu)$. Also, write the state-action visitation distributions as $d_h^{\mu,\nu}(s,a,b) = \mathbb{P}(s_h = s, a_h = a, b_h = b|\rho_1, \mu, \nu)$.

To control the uniform metric in Definition 3.6, we require an estimator of the QRE that performs uniformly well across all states $s \in \mathcal{S}$. When using frequency estimators to approximate the policies $\mu_h^*(\cdot|s)$ and $\nu_h^*(\cdot|s)$, the estimation at each state is conducted independently. As a result, it is essential that the dataset sufficiently covers all states in $\mathcal{S}$ to obtain reliable estimates. To ensure this, we impose the following assumption, which guarantees that every state is visited with a minimum frequency throughout the horizon.

Assumption 3.8 (C-well-posedness)**.** There exists a constant $C > 0$ such that

$$d_h^{\mu^*,\nu^*}(s) \geq C$$

for all $s \in \mathcal{S}$ and $h \in [H]$.

Now we are ready to present the theoretical results for the proposed algorithm.

Theorem 3.9 (Sample complexity of constructing feasible reward set)**.** *Under Assumptions 3.3 and 3.8, let $\rho_h = d_h^*$ be the stationary distribution associated with optimal policies μ^* and ν^*, where $h \in [H]$. We assume that the following $d \times d$ matrix*

$$\Psi_h = \mathbb{E}_{\rho_h}\left[\phi(s_h, a_h, b_h)\phi(s_h, a_h, b_h)^\top\right]$$

is nonsingular for all $h \in [H]$. Let $\mathcal{R}$ be the feasible reward set given in Definition 3.7. Given a dataset $\mathcal{D} = \{\mathcal{D}_h\}_{h\in[H]} = \{\{(s_h^t, a_h^t, b_h^t)\}_{t\in[T]}\}_{h\in[H]}$, we set $\lambda = \mathcal{O}(1)$, $\kappa_h = \mathcal{O}(T^{-1})$, and let $\widehat{\mathcal{R}}$ be the output of Algorithm 2. Let $\xi = \min_{h\in[H],s\in\mathcal{S},a\in\mathcal{A},b\in\mathcal{B}}\{\mu_h^*(a|s), \nu_h^*(b|s)\}$. For any $\delta \in (0,1)$, let $T > 0$ be sufficiently large, so

$$T \geq \max\Bigg\{\frac{1}{C^2}\log\frac{2HS}{\delta}, \frac{16(m\vee n)}{C\xi^2}\log\frac{4HS}{\delta}, \\ 512\|\Psi_h^{-1}\|_{\mathrm{op}}^2\log\frac{2Hd}{\delta}, 4\lambda\|\Psi_h^{-1}\|_{\mathrm{op}}\Bigg\}.$$

Then the following inequality holds with probability at least $1 - 3\delta$:

$$D(\mathcal{R}, \widehat{\mathcal{R}}) \lesssim \frac{1}{\sqrt{T}}\Bigg(\sqrt{S(m+n)\log\frac{HS}{\delta}}\log T \\ + S(m+n)\sqrt{\log\frac{HS}{\delta}} + \left(\sqrt{Sd} + \sqrt{d\log T}\right)\log(mn)\Bigg),$$

where D is the Hausdorff distance corresponding to the uniform metric in Definition 3.6.

Proof. See Appendix B.3 for the complete proof. □

Theorem 3.9 provides a strong guarantee on the accuracy of our reward recovery algorithm in Markov games. Our bound shows that the distance $D(\mathcal{R}, \widehat{\mathcal{R}})$ diminishes at the rate of $\mathcal{O}(T^{-1/2})$, which matches the optimal statistical rate for empirical risk minimization problems. This demonstrates that with sufficient data, the estimated reward functions remain close to the true feasible set, making our method statistically reliable and sample-efficient. The explicit dependence on problem parameters offers insights into how exploration, feature representations, and action space size affect the difficulty of inverse reward learning in Markov games.

We also note that the condition that Ψ_h is nonsingular ensures that the feature representation provides sufficient information for parameter recovery (Tu & Recht, 2017; Min et al., 2022). The norm $\|\Psi_h^{-1}\|_{\mathrm{op}}$ appears in the sample complexity bound, indicating that ill-conditioned feature matrices lead to larger estimation errors and require more samples to achieve the same level of accuracy.

In addition, instead of relying solely on frequency estimators for QRE estimation, we extend our framework to integrate Maximum Likelihood Estimation (MLE) into our method and establish a convergence result with the same $T^{-1/2}$ rate. We provide the details in Appendix §C.

4. Numerical Experiments

In this section, we implement our reward-learning algorithm and conduct numerical experiments in both entropy-regularized zero-sum matrix games and Markov games. All experiments are conducted in Google Colab. In this section we consider only two-player entropy-regularized entropy-regularized zero-sum Markov games. The experimental results of matrix games are presented in Appendix §E.

Setup. We define the kernel function $\phi : \mathcal{A} \times \mathcal{B} \to \mathbb{R}^d$ with dimension $d = 2$, and set the true parameter ω_h that specifying reward functions to be

$$\omega_h^* = (0.8, -0.6)^\top$$

for all steps $h \in [H]$. We set the sizes of action spaces to be $m = 5$ and $n = 5$, the size of state space $S = 4$, and the horizon $H = 6$. The entropy regularization term is $\eta = 0.5$.

We implement the algorithm proposed in §3.2. In each experiment, our algorithm outputs a parameter $\widehat{\theta}_h$ in the confidence set $\widehat{\Theta}_h$. We set the bound of feasible parameters θ_h to be $R = 10$, and set the threshold $\kappa_h = 10^3/N$, where

N is the sample size. The regularization term in ridge regression is $\lambda = 0.01$.

Metrics. We evaluate the performance of our algorithm using two metrics: (1) the error in the estimated reward function $(\widehat{r}_h)$, which measures how accurately the reconstructed payoff function matches the true reward function; and (2) the error in the estimated QRE, which quantifies the discrepancy between the QRE $(\widehat{\mu}, \widehat{\nu})$ derived from the estimated payoff function and the true QRE (μ^*, ν^*). We are particularly interested in the error in the estimated QRE, which validates whether the reconstructed reward functions interpret the observed strategy.

Results. As shown in Figures 1, 2 and Table 1, the overall error of our algorithm's output decreases as the sample size N increases from 10^4 to 10^5, demonstrating the improved accuracy of our approach with more data. While the estimation error of reward functions $(\widehat{r}_h)_{h=1}^6$ can be relatively large, the corresponding QRE $(\widehat{\mu}_h, \widehat{\nu}_h)$ remains well-aligned with the true QRE (μ_h^*, ν_h^*). Although some fluctuations are observed across time steps, the error remains small, especially for larger sample sizes. These results confirm that our method for reward estimation in Markov games is both statistically consistent and sample-efficient.

5. Conclusion

To conclude, we explore the challenge of recovering reward functions that explain agents' behavior in competitive games, with a focus on the entropy-regularized zero-sum setting. We propose a framework of inverse game theory concerning the underlying reward mechanisms driving observed behaviors, which applies to both the static setting (§2) and the dynamic setting (§3).

Under a linear assumption, we develop a novel approach for the identifiability of the parameter specifying the current-time payoff. To move forward, we develop an offline algorithm unifying QRE estimation, confidence set construction, transition kernel estimation, and reward recovery, and establish its convergence properties under regular conditions. Additionally, we adapt this algorithm to incorporate a MLE approach and provide theoretical guarantees for the adapted version. Our algorithms are reliable and effective in both static and dynamic settings, even in the presence of high-dimensional parameter spaces or rank deficiencies.

Future directions include exploring more complicated game settings, such as partially observable games and non-linear payoff functions, and extending the framework to online learning setting. Meanwhile, this research contributes to the broader effort to make competitive systems more interpretable, offering valuable insights at the intersection of game theory and reinforcement learning.

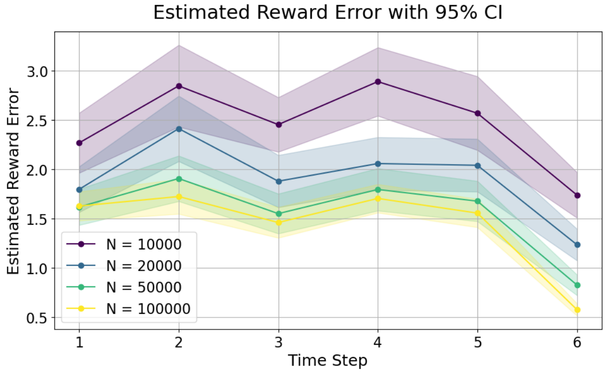


Figure 1. The reconstruction error of the reward functions $(\widehat{r}_h)_{h=1}^6$. The X-axis represents the time step h from 1 to 6, while the Y-axis represents the error $\|\widehat{r}_h - r_h^*\|_{\mathrm{F}}$ of the reward function $\widehat{r}$.

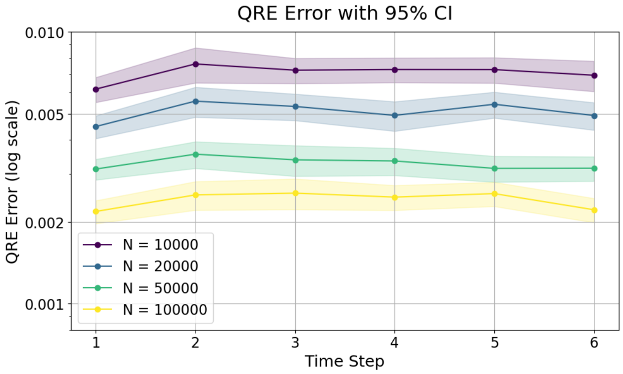


Figure 2. The discrepancy between the QRE $(\widehat{\mu}, \widehat{\nu})$ corresponding to the estimated reward functions $(\widehat{r}_h)_{h=1}^6$ and the true QRE (μ^*, ν^*). The X-axis represents the time step h from 1 to 6, while the Y-axis represents the errors $\mathrm{TV}(\widehat{\mu}_h, \mu_h^*) + \mathrm{TV}(\widehat{\nu}_h, \nu_h^*)$

Sample Size	**Reward Error**	
	Mean	95% CI
10,000	2.4611	± 0.1596
20,000	1.9031	± 0.1048
50,000	1.5609	± 0.0663
100,000	1.4398	± 0.0499

Sample Size	**QRE Error**	
	Mean	95% CI
10,000	7.08×10^{-3}	$\pm 4.61 \times 10^{-4}$
20,000	5.11×10^{-3}	$\pm 3.11 \times 10^{-4}$
50,000	3.28×10^{-3}	$\pm 1.70 \times 10^{-4}$
100,000	2.41×10^{-3}	$\pm 1.41 \times 10^{-4}$

Table 1. Mean error and 95% confidence intervals for reward and QRE estimation over 100 repetitions in the Markov game setting, across all time steps.

Impact Statement

This work advances the field of inverse reinforcement learning and game theory by introducing a unified framework for reward function identification and estimation in competitive multi-agent settings. Our findings contribute to a deeper understanding of decision-making in strategic environments, with potential applications in economics, automated negotiation, and multi-agent AI systems.

While our research provides theoretical and methodological advancements, we acknowledge potential ethical considerations. The ability to infer reward functions from observed behavior could be used both positively—to enhance transparency in AI decision-making and improve algorithmic fairness—and negatively, if applied to manipulate or exploit agents in competitive settings. Ensuring the responsible application of this work will require careful consideration of ethical safeguards and alignment with societal values.

Overall, this paper aims to advance Machine Learning and Game Theory research, and we do not foresee immediate societal risks. However, we encourage further discussion on the ethical implications of inverse game theory in real-world applications.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pp. 2312–2320, Red Hook, NY, USA, 2011. Curran Associates Inc. ISBN 9781618395993.

Agarwal, R. and Özlem Ergun. Network Design and Allocation Mechanisms for Carrier Alliances in Liner Shipping. *Operations Research*, 58(6):1726–1742, December 2010. doi: 10.1287/opre.1100.0848. URL https://ideas.repec.org/a/inm/oropre/v58y2010i6p1726-1742.html.

Ahmadi, F., Ganjkhanloo, F., and Ghobadi, K. Inverse learning: Solving partially known models using inverse optimization, 2023. URL https://arxiv.org/abs/2011.03038.

Ahmed, Z., Le Roux, N., Norouzi, M., and Schuurmans, D. Understanding the impact of entropy on policy optimization. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 151–160. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ahmed19a.html.

Ahuja, R. and Orlin, J. Inverse optimization. *Operations Research*, 49, 06 2001. doi: 10.1287/opre.49.5.771.10607.

Arora, S. and Doshi, P. A survey of inverse reinforcement learning: Challenges, methods and progress, 2020. URL https://arxiv.org/abs/1806.06877.

Ben-Israel, A. and Greville, T. *Generalized Inverses: Theory and Applications*. CMS Books in Mathematics. Springer New York, 2006. ISBN 9780387216348. URL https://books.google.com/books?id=abEPBwAAQBAJ.

Boularias, A., Kober, J., and Peters, J. Relative entropy inverse reinforcement learning. In Gordon, G., Dunson, D., and Dudík, M. (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 182–189, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/boularias11a.html.

Bradtke, S. J. and Barto, A. G. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22:33–57, 2004. URL https://api.semanticscholar.org/CorpusID:10316699.

Brotcorne, L., Marcotte, P., Savard, G., and WIART, M. Joint pricing and network capacity setting problem. 01 2005.

Cao, H., Cohen, S., and Szpruch, L. Identifiability in inverse reinforcement learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12362–12373. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/671f0311e2754fcdd37f70a8550379bc-Paper.pdf.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. Fast global convergence of natural policy gradient methods with entropy regularization, 2021. URL https://arxiv.org/abs/2007.06558.

Cen, S., Wei, Y., and Chi, Y. Fast policy extragradient methods for competitive games with entropy regularization, 2023. URL https://arxiv.org/abs/2105.15186.

Chan, T. C. Y. and Kaw, N. Inverse optimization for the recovery of constraint parameters, 2019. URL https://arxiv.org/abs/1811.00726.

Chan, T. C. Y., Mahmood, R., and Zhu, I. Y. Inverse optimization: Theory and applications, 2022. URL https://arxiv.org/abs/2109.03920.

Chen, S., Wang, M., and Yang, Z. Actions speak what you want: Provably sample-efficient reinforcement learning of the quantal stackelberg equilibrium from strategic feedbacks, 2023. URL https://arxiv.org/abs/2307.14085.

Chen, Z., Ma, S., and Zhou, Y. Sample efficient stochastic policy extragradient algorithm for zero-sum markov game. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=IvepFxYRDG.

Chow, J. Y. and Djavadian, S. Activity-based market equilibrium for capacitated multimodal transport systems. *Transportation Research Procedia*, 7:2–23, 2015. ISSN 2352-1465. doi: https://doi.org/10.1016/j.trpro.2015.06.001. URL https://www.sciencedirect.com/science/article/pii/S2352146515000691. 21st International Symposium on Transportation and Traffic Theory Kobe, Japan, 5-7 August, 2015.

Dong, C., Chen, Y., and Zeng, B. Generalized inverse optimization through online learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/28dd2c7955ce926456240b2ff0100bde-Paper.pdf.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making, 2023. URL https://arxiv.org/abs/2112.13487.

Fu, J., Luo, K., and Levine, S. Learning robust rewards with adversarial inverse reinforcement learning, 2018. URL https://arxiv.org/abs/1710.11248.

Ghobadi, K. and Mahmoudzadeh, H. Inferring linear feasible regions using inverse optimization. *European Journal of Operational Research*, 290(3):829–843, 2021. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2020.08.048. URL https://www.sciencedirect.com/science/article/pii/S037722172030761X.

Gleave, A. and Toyer, S. A primer on maximum causal entropy inverse reinforcement learning, 2022. URL https://arxiv.org/abs/2203.11409.

Guan, Y., Zhang, Q., and Tsiotras, P. Learning nash equilibria in zero-sum stochastic games via entropy-regularized policy approximation, 2021. URL https://arxiv.org/abs/2009.00162.

Guo, X., Xu, R., and Zariphopoulou, T. Entropy regularization for mean field games with learning, 2021. URL https://arxiv.org/abs/2010.00145.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018. URL https://arxiv.org/abs/1801.01290.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2009. ISBN 9780387848846. URL https://books.google.com/books?id=eBSgoAEACAAJ.

Herman, M., Gindele, T., Wagner, J., Schmitt, F., and Burgard, W. Inverse reinforcement learning with simultaneous estimation of rewards and dynamics. In Gretton, A. and Robert, C. C. (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 102–110, Cadiz, Spain, 09–11 May 2016. PMLR. URL https://proceedings.mlr.press/v51/herman16.html.

Huang, S., Held, D., Abbeel, P., and Dragan, A. Enabling robots to communicate their objectives. *Autonomous Robots*, 43, 02 2019. doi: 10.1007/s10514-018-9771-0.

Jarboui, F. and Perchet, V. Offline inverse reinforcement learning, 2021. URL https://arxiv.org/abs/2106.05068.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/jin20a.html.

Kalogiannis, F. and Panageas, I. Zero-sum polymatrix markov games: Equilibrium collapse and efficient computation of nash equilibria. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 59996–60020. Curran Associates, Inc., 2023.

Konstantakopoulos, I., Ratliff, L., Jin, M., Sastry, S., and Spanos, C. A robust utility learning framework via inverse optimization. *IEEE Transactions on Control Systems Technology*, PP, 04 2017. doi: 10.1109/TCST.2017.2699163.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lin, X., Beling, P., and Cogill, R. Multi-agent inverse reinforcement learning for zero-sum games. *IEEE Transactions on Computational Intelligence and AI in Games*, PP, 03 2014. doi: 10.1109/TCIAIG.2017.2679115.

Lindner, D., Krause, A., and Ramponi, G. Active exploration for inverse reinforcement learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=TPOJzwv2pc.

Ling, C. K., Fang, F., and Kolter, J. Z. What game are we playing? end-to-end learning in normal and extensive form games, 2018. URL https://arxiv.org/abs/1805.02777.

Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In Cohen, W. W. and Hirsh, H. (eds.), *Machine Learning Proceedings 1994*, pp. 157–163. Morgan Kaufmann, San Francisco (CA), 1994. ISBN 978-1-55860-335-6. doi: https://doi.org/10.1016/B978-1-55860-335-6.50027-1. URL https://www.sciencedirect.com/science/article/pii/B9781558603356500271.

McKelvey, R. D. and Palfrey, T. R. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995. ISSN 0899-8256. doi: https://doi.org/10.1006/game.1995.1023. URL https://www.sciencedirect.com/science/article/pii/S0899825685710238.

Mertikopoulos, P. and Sandholm, W. H. Learning in games via reinforcement and regularization. *Math. Oper. Res.*, 41(4):1297–1324, November 2016. ISSN 0364-765X.

Miehling, E., Rasouli, M., and Teneketzis, D. A pomdp approach to the dynamic defense of large-scale cyber networks. *IEEE Transactions on Information Forensics and Security*, 13(10):2490–2505, 2018. doi: 10.1109/TIFS.2018.2819967.

Min, Y., Wang, T., Zhou, D., and Gu, Q. Variance-aware off-policy evaluation with linear function approximation, 2022. URL https://arxiv.org/abs/2106.11960.

Nash Jr, J. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951. ISSN 0003486X, 19398980. URL http://www.jstor.org/stable/1969529.

Neu, G., Jonsson, A., and Gómez, V. A unified view of entropy-regularized markov decision processes, 2017. URL https://arxiv.org/abs/1705.07798.

Ng, A. Y. and Russell, S. J. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

Nourollahi, S. and Ghate, A. Inverse optimization in minimum cost flow problems on countably infinite networks. *Networks*, 73, 11 2018. doi: 10.1002/net.21862.

Rolland, P., Viano, L., Schürhoff, N., Nikolov, B., and Cevher, V. Identifiability and generalizability from multiple experts in inverse reinforcement learning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 550–564. Curran Associates, Inc., 2022.

Savas, Y., Ahmadi, M., Tanaka, T., and Topcu, U. Entropy-regularized stochastic games. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 5955–5962, 2019. doi: 10.1109/CDC40024.2019.9029555.

Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Shapley, L. S. Stochastic games*. *Proceedings of the National Academy of Sciences*, 39:1095 – 1100, 1953. URL https://api.semanticscholar.org/CorpusID:263414073.

Snoswell, A. J., Singh, S. P. N., and Ye, N. Revisiting maximum entropy inverse reinforcement learning: New perspectives and algorithms. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 241–249. IEEE, December 2020. doi: 10.1109/ssci47803.2020.9308391. URL http://dx.doi.org/10.1109/SSCI47803.2020.9308391.

Song, L., Li, D., Wang, X., and Xu, X. Adaboost maximum entropy deep inverse reinforcement learning with truncated gradient. *Information Sciences*, 602:328–350, 2022. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2022.04.017. URL https://www.sciencedirect.com/science/article/pii/S002002552200353X.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

Szepesvári, C. *Algorithms for Reinforcement Learning*, volume 4. 01 2010. doi: 10.2200/S00268ED1V01Y201005AIM009.

Tan, Y., Delong, A., and Terekhov, D. Deep inverse optimization. In Rousseau, L.-M. and Stergiou, K. (eds.), *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pp. 540–556, Cham, 2019. Springer International Publishing. ISBN 978-3-030-19212-9.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, August 2011. ISSN 1615-3383. doi: 10.1007/s10208-011-9099-z. URL http://dx.doi.org/10.1007/s10208-011-9099-z.

Tu, S. and Recht, B. Least-squares temporal difference learning for the linear quadratic regulator, 2017. URL https://arxiv.org/abs/1712.08642.

Vatandoust, B., Zad, B. B., Vallée, F., Toubeau, J.-F., and Bruninx, K. Integrated forecasting and scheduling of implicit demand response in balancing markets using inverse optimization. In *2023 19th International Conference on the European Energy Market (EEM)*, pp. 1–6, 2023. doi: 10.1109/EEM58374.2023.10161818.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

von Neumann, J. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928. URL https://api.semanticscholar.org/CorpusID:122961988.

Vorobeychik, Y., Wellman, M. P., and Singh, S. Learning payoff functions in infinite games. *Mach. Learn.*, 67(1–2):145–168, May 2007. ISSN 0885-6125. doi: 10.1007/s10994-007-0715-8. URL https://doi.org/10.1007/s10994-007-0715-8.

Wang, H., Zariphopoulou, T., and Zhou, X. Exploration versus exploitation in reinforcement learning: a stochastic control approach, 2019. URL https://arxiv.org/abs/1812.01552.

Wang, X. and Klabjan, D. Competitive multi-agent inverse reinforcement learning with sub-optimal demonstrations. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5143–5151. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/wang18d.html.

Wei, C.-Y., Lee, C.-W., Zhang, M., and Luo, H. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games, 2021. URL https://arxiv.org/abs/2102.04540.

Wu, J., Shen, W., Fang, F., and Xu, H. Inverse game theory for stackelberg games: the blessing of bounded rationality. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

Wulfmeier, M., Ondruska, P., and Posner, I. Maximum entropy deep inverse reinforcement learning, 2016. URL https://arxiv.org/abs/1507.04888.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3674–3682. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/xie20a.html.

Yu, L., Song, J., and Ermon, S. Multi-agent adversarial inverse reinforcement learning, 2019. URL https://arxiv.org/abs/1907.13220.

Yu, S., Wang, H., and Dong, C. Learning risk preferences from investment portfolios using inverse optimization, 2021. URL https://arxiv.org/abs/2010.01687.

Zhan, W., Cen, S., Huang, B., Chen, Y., Lee, J. D., and Chi, Y. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence, 2023. URL https://arxiv.org/abs/2105.11066.

Zhang, Z. Estimating mutual information via kolmogorov distance. *Information Theory, IEEE Transactions on*, 53: 3280 – 3282, 10 2007. doi: 10.1109/TIT.2007.903122.

Zhao, L., Liu, T., Peng, X., and Metaxas, D. Maximum-entropy adversarial data augmentation for improved generalization and robustness. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14435–14447. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/a5bfc9e07964f8dddeb95fc584cd965d-Paper.pdf.

Zhao, Y., Tian, Y., Lee, J., and Du, S. Provably efficient policy optimization for two-player zero-sum markov games. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 2736–2761. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/zhao22b.html.

Ziebart, B. D. Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy. 7 2018. doi: 10.1184/R1/6720692.v1.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2008. URL https://api.semanticscholar.org/CorpusID:336219.

A. Proof of Entropy-Regularized Matrix Games

A.1. Proof of Theorem 2.3

Proof. To begin with, we decompose $\|\widehat{\theta}-\theta^*\|^2$ as follows:

$$\|\widehat{\theta}-\theta^*\|^2 \lesssim \underbrace{\Big\|\big(A(\widehat{\nu})^\top A(\widehat{\nu})+B(\widehat{\mu})^\top B(\widehat{\mu})\big)^{-1}\big[A(\widehat{\nu})^\top c(\widehat{\mu})+B(\widehat{\mu})^\top d(\widehat{\nu})-A(\nu^*)^\top c(\mu^*)-B(\mu^*)^\top d(\nu^*)\big]\Big\|^2}_{\text{(I)}}$$

$$+\underbrace{\Big\|\Big[\big(A(\widehat{\nu})^\top A(\widehat{\nu})+B(\widehat{\mu})^\top B(\widehat{\mu})\big)^{-1}-\big(A(\nu^*)^\top A(\nu^*)+B(\mu^*)^\top B(\mu^*)\big)^{-1}\Big]\big[A(\nu^*)^\top c(\mu^*)+B(\mu^*)^\top d(\nu^*)\big]\Big\|^2}_{\text{(II)}}.$$

Bounding (I).

$$\begin{aligned}\text{(I)} \lesssim{}& \Big\|\big(A(\widehat{\nu})^\top A(\widehat{\nu})+B(\widehat{\mu})^\top B(\widehat{\mu})\big)^{-1}\Big\|_{\text{op}}^2\\ &\times\big[\|A(\widehat{\nu})^\top c(\widehat{\mu})-A(\nu^*)^\top c(\mu^*)\|^2+\|B(\widehat{\mu})^\top d(\widehat{\nu})-B(\mu^*)^\top d(\nu^*)\|^2\big]\end{aligned} \tag{16}$$

We will bound the three terms in the RHS of (16) in the following text. To begin with, we have

$$\|A(\widehat{\nu})^\top c(\widehat{\mu})-A(\nu^*)^\top c(\mu^*)\|^2 \lesssim \underbrace{\|(A(\widehat{\nu})-A(\nu^*))^\top c(\widehat{\mu})\|^2}_{1)}+\underbrace{\|A(\nu^*)^\top(c(\widehat{\mu})-c(\mu^*))\|^2}_{2)}.$$

Recall the definition of $c(\mu)=(\log(\mu_i/\mu_1)/\eta)_{i\in[m]/\{1\}}$, we have

$$\begin{aligned}2) &\le \|A(\nu^*)\|_{\text{op}}^2\cdot\sum_{i=2}^m(\log(\widehat{\mu}_i/\widehat{\mu}_1-\log(\mu_i^*/\mu_1^*)))^2/\eta^2\\ &\le 2\|A(\nu^*)\|_{\text{op}}^2\cdot\left[(m-2)(\log(\widehat{\mu}_1)-\log(\mu_1^*))^2+\sum_{i=1}^m(\log(\widehat{\mu}_i)-\log(\mu_i^*))^2\right]/\eta^2.\end{aligned} \tag{17}$$

Recall the definition of $A(\nu^*)=((\phi(i,\cdot)-\phi(1,\cdot))\,\nu^*)_{i\in[m]/\{1\}}$, which can be rewritten as follows:

$$A(\nu^*)^\top=\Phi_1\cdot(I_{m-1}\otimes\nu^*),$$

where $\Phi_1:=(\phi(i,\cdot)-\phi(1,\cdot))_{i\in[m]\setminus\{1\}}\in\mathbb{R}^{d\times(m-1)n}$ and $\otimes$ denotes the Kronecker product, and $I_{m-1}\otimes\nu^*\in\mathbb{R}^{(m-1)n\times(m-1)}$. Therefore, we have

$$\|A(\nu^*)\|_{\text{op}}^2\le\|\Phi_1\|_{\text{op}}^2\cdot\|\nu^*\|^2, \tag{18}$$

where we use the fact that $\|I_{m-1}\otimes\nu\|_{\text{op}}^2=\|\nu\|^2$. Besides, we notice that

$$\begin{aligned}\sum_{i=1}^m(\log(\widehat{\mu}_i)-\log(\mu_i^*))^2&=\sum_{i=1}^m\left(\log\left(1+\frac{\widehat{\mu}_i-\mu_i^*}{\mu_i^*}\right)\right)^2\\ &\le\sum_{i=1}^m\max\left\{\left(\frac{\widehat{\mu}_i-\mu_i^*}{\widehat{\mu}_i}\right)^2,\left(\frac{\widehat{\mu}_i-\mu_i^*}{\mu_i^*}\right)^2\right\},\end{aligned} \tag{19}$$

where we use the inequality $x/(1+x)\le\log(1+x)\le x$ for all $x>-1$. Moreover, we remark that $|\widehat{\mu}_i-\mu_i^*|\le\mathrm{TV}(\widehat{\mu},\mu^*)/2\le\epsilon_1/4$, we have

$$\sum_{i=1}^m(\log(\widehat{\mu}_i)-\log(\mu_i^*))^2\le\sum_{i=1}^m\frac{\epsilon_1^2/16}{\min\{\widehat{\mu}_i,\mu_i^*\}^2}\le\frac{m\cdot\epsilon_1^2}{16(\min_{i\in[m]}\mu_i^*-\epsilon_1)^2}.$$

Plugging these inequalities in (17), we obtain

$$2)\lesssim\frac{m\cdot\|\Phi_1\|_{\text{op}}^2\cdot\|\nu\|^2\cdot\epsilon_1^2}{\eta^2\cdot(\min_{i\in[m]}\mu_i-\epsilon_1)^2}. \tag{20}$$

To derive the bound for 1), we leverage the equivalent form of $A(\nu^*)$:

$$(A(\widehat{\nu}) - A(\nu^*))^\top = \Phi_1 \cdot (I_{m-1} \otimes (\widehat{\nu} - \nu^*)).$$

Plugging in this equality, we obtain

$$\begin{aligned} 1) &\leq \|\Phi_1\|_{\text{op}}^2 \cdot \|I_{m-1} \otimes (\widehat{\nu} - \nu^*)\|_{\text{op}}^2 \cdot \|c(\widehat{\mu})\|^2 \\ &\lesssim \|\Phi_1\|_{\text{op}}^2 \cdot \|\widehat{\nu} - \nu\|^2 \cdot (\|c(\mu^*)\|^2 + \|c(\widehat{\mu}) - c(\mu^*)\|^2) \\ &\lesssim \|\Phi_1\|_{\text{op}}^2 \cdot \epsilon_2^2 \cdot \left(\|c(\mu^*)\|^2 + \frac{m \cdot \epsilon_1^2}{\eta^2 \cdot (\min_{i \in [m]} \mu_i^* - \epsilon_1)^2} \right). \end{aligned} \tag{21}$$

Combining (20) and (21), we have

$$\begin{aligned} \|A(\widehat{\nu})^\top c(\widehat{\mu}) - A(\nu^*)^\top c(\mu^*)\|^2 &\lesssim \|\Phi_1\|_{\text{op}}^2 \cdot \epsilon_2^2 \cdot \left(\|c(\mu^*)\|^2 + \frac{m \cdot \epsilon_1^2}{\eta^2 \cdot (\min_{i \in [m]} \mu_i^* - \epsilon_1)^2} \right) \\ &\quad + \frac{m \cdot \|\Phi_1\|_{\text{op}}^2 \cdot \|\nu\|^2 \cdot \epsilon_1^2}{\eta^2 \cdot (\min_{i \in [m]} \mu_i^* - \epsilon_1)^2} \\ &\lesssim \epsilon_2^2 \cdot \|c(\mu^*)\|^2 + \frac{m \cdot \epsilon_1^2 \cdot (\epsilon_2^2 + 1)}{\eta^2 \cdot (\min_{i \in [m]} \mu_i^* - \epsilon_1)^2}. \end{aligned} \tag{22}$$

By the symmetric of $\|A(\widehat{\nu})^\top c(\widehat{\mu}) - A(\nu^*)^\top c(\mu^*)\|^2$ and $\|B(\widehat{\mu})^\top d(\widehat{\nu}) - B(\mu^*)^\top d(\nu^*)\|^2$, we have

$$\|B(\widehat{\mu})^\top d(\widehat{\nu}) - B(\mu^*)^\top d(\nu^*)\|^2 \quad \lesssim \epsilon_1^2 \cdot \|d(\nu^*)\|^2 + \frac{n \cdot \epsilon_2^2 \cdot (\epsilon_1^2 + 1)}{\eta^2 \cdot (\min_{i \in [n]} \nu_i^* - \epsilon_2)^2}. \tag{23}$$

Now we derive the bound for $\left\| \left(A(\widehat{\nu})^\top A(\widehat{\nu}) + B(\widehat{\mu})^\top B(\widehat{\mu}) \right)^{-1} \right\|_{\text{op}}^2$. To begin with, we have

$$\begin{aligned} \left\| \left(A(\widehat{\nu})^\top A(\widehat{\nu}) + B(\widehat{\mu})^\top B(\widehat{\mu}) \right)^{-1} \right\|_{\text{op}}^2 &\lesssim \left\| \left(A(\nu^*)^\top A(\nu^*) + B(\mu^*)^\top B(\mu^*) \right)^{-1} \right\|_{\text{op}}^2 \\ + \left\| \left(A(\widehat{\nu})^\top A(\widehat{\nu}) + B(\widehat{\mu})^\top B(\widehat{\mu}) \right)^{-1} \right. &- \left. \left(A(\nu^*)^\top A(\nu^*) + B(\mu^*)^\top B(\mu^*) \right)^{-1} \right\|_{\text{op}}^2. \end{aligned} \tag{24}$$

To simplify the notation, we define $U := A(\nu^*)^\top A(\nu^*) + B(\mu^*)^\top B(\mu^*)$ and $V(\widehat{\mu}, \widehat{\nu}) := (A(\widehat{\nu})^\top A(\widehat{\nu}) + B(\widehat{\mu})^\top B(\widehat{\mu})) - (A(\nu^*)^\top A(\nu^*) + B(\mu^*)^\top B(\mu^*))$, (24) can be rewritten as follow

$$\left\| \left(A(\widehat{\nu})^\top A(\widehat{\nu}) + B(\widehat{\mu})^\top B(\widehat{\mu}) \right)^{-1} \right\|_{\text{op}}^2 \lesssim \|U^{-1}\|_{\text{op}}^2 + \|(U + V(\widehat{\mu}, \widehat{\nu}))^{-1} - U^{-1}\|_{\text{op}}^2.$$

We leverage the Woodbury formula to derive the bound for the above inequality.

$$(U + V(\widehat{\mu}, \widehat{\nu}))^{-1} - U^{-1} = -U^{-1}(I + V(\widehat{\mu}, \widehat{\nu})U^{-1})^{-1} V(\widehat{\mu}, \widehat{\nu}) U^{-1},$$

which further implies that

$$\left\| \left(A(\widehat{\nu})^\top A(\widehat{\nu}) + B(\widehat{\mu})^\top B(\widehat{\mu}) \right)^{-1} \right\|_{\text{op}}^2 \lesssim \|U^{-1}\|_{\text{op}}^2 + \|U^{-1}\|_{\text{op}}^4 \cdot \|V(\widehat{\mu}, \widehat{\nu})\|_{\text{op}}^2 \cdot \|(I + V(\widehat{\mu}, \widehat{\nu})U^{-1})^{-1}\|_{\text{op}}^2.$$

We derive the bound for $\|U^{-1}\|_{\text{op}}^2$, $\|V(\widehat{\mu}, \widehat{\nu})\|_{\text{op}}^2$, $\|(I + V(\widehat{\mu}, \widehat{\nu})U^{-1})^{-1}\|_{\text{op}}^2$ in the following text. To begin with, by Weyl's inequality, we have $\sigma_{\min}(U) \geq \sigma_{\min}(A(\nu^*)^\top A(\nu^*)) + \sigma_{\min}(B(\mu^*)^\top B(\mu^*))$ and

$$\|U^{-1}\|_{\text{op}}^2 = \frac{1}{\sigma_{\min}(U)^2} \leq \frac{1}{(\sigma_{\min}(A(\nu^*))^2 + \sigma_{\min}(B(\mu^*))^2)^2}.$$

For $\|V(\widehat{\mu}, \widehat{\nu})\|_{\text{op}}^2 = \|(A(\widehat{\nu})^\top A(\widehat{\nu}) + B(\widehat{\mu})^\top B(\widehat{\mu})) - (A(\nu^*)^\top A(\nu^*) + B(\mu^*)^\top B(\mu^*))\|_{\text{op}}^2$, we have

$$\begin{aligned} \|V(\widehat{\mu}, \widehat{\nu})\|_{\text{op}}^2 &\lesssim \|A(\widehat{\nu})^\top A(\widehat{\nu}) - A(\nu^*)^\top A(\nu^*)\|_{\text{op}}^2 + \|B(\widehat{\mu})^\top B(\widehat{\mu}) - B(\mu^*)^\top B(\mu^*)\|_{\text{op}}^2 \\ &\lesssim \|(A(\widehat{\nu}) - A(\nu^*))^\top A(\widehat{\nu})\|_{\text{op}}^2 + \|A(\nu^*)^\top (A(\widehat{\nu}) - A(\nu^*))\|_{\text{op}}^2 \\ &+ \|(B(\widehat{\mu}) - B(\mu^*))^\top B(\widehat{\mu})\|_{\text{op}}^2 + \|B(\mu^*)^\top (B(\widehat{\mu}) - B(\mu^*))\|_{\text{op}}^2. \end{aligned}$$

Recalling (21), we have $\|A(\widehat{\nu}) - A(\nu^*)\|_{\text{op}}^2 \le \epsilon_2^2$ and $\|B(\widehat{\mu}) - B(\mu^*)\|_{\text{op}}^2 \le \epsilon_1^2$, which further implies that

$$\|V(\widehat{\mu}, \widehat{\nu})\|_{\text{op}}^2 \lesssim \epsilon_1^2 + \epsilon_2^2. \tag{25}$$

For the term $\|(I + V(\widehat{\mu}, \widehat{\nu})U^{-1})^{-1}\|_{\text{op}}^2$, we have $\|(I + V(\widehat{\mu}, \widehat{\nu})U^{-1})^{-1}\|_{\text{op}}^2 = 1/\sigma_{\min}(I + V(\widehat{\mu}, \widehat{\nu})U^{-1})^2$. By the property of the smallest singular value, we obtain

$$\begin{aligned}\sigma_{\min}(I + V(\widehat{\mu}, \widehat{\nu})U^{-1}) &= \inf_{\|x\|=1} \|(I + V(\widehat{\mu}, \widehat{\nu})U^{-1})x\| \ge 1 - \|V(\widehat{\mu}, \widehat{\nu})U^{-1}\|_{\text{op}} \\ &\ge 1 - \|V(\widehat{\mu}, \widehat{\nu})\|_{\text{op}} \|U^{-1}\|_{\text{op}}.\end{aligned} \tag{26}$$

Combining (16), (22), (23), and (24), we have

$$\text{(I)} \lesssim \epsilon_1^2 \cdot \left(\|d(\nu^*)\|^2 + \frac{m \cdot (\epsilon_2^2 + 1)}{\eta^2 \cdot (\min_{i \in [m]} \mu_i - \epsilon_1)^2} \right) + \epsilon_2^2 \cdot \left(\|c(\mu^*)\|^2 + \frac{n \cdot (\epsilon_1^2 + 1)}{\eta^2 \cdot (\min_{i \in [n]} \nu_i - \epsilon_2)^2} \right). \tag{27}$$

Bounding (II). Combining (24), (25), and (26), we obtain

$$\begin{aligned}\text{(II)} &\lesssim \left\| \big(A(\widehat{\nu})^\top A(\widehat{\nu}) + B(\widehat{\mu})^\top B(\widehat{\mu})\big)^{-1} - \big(A(\nu^*)^\top A(\nu^*) + B(\mu^*)^\top B(\mu^*)\big)^{-1} \right\|_{\text{op}}^2 \|A(\nu^*)^\top c(\mu^*) + B(\mu^*)^\top d(\nu^*)\|^2 \\ &\lesssim \|U^{-1}\|_{\text{op}}^4 \cdot \|V(\widehat{\mu}, \widehat{\nu})\|_{\text{op}}^2 \cdot \|(I + V(\widehat{\mu}, \widehat{\nu})U^{-1})^{-1}\|_{\text{op}}^2 \lesssim \epsilon_1^2 + \epsilon_2^2.\end{aligned} \tag{28}$$

Combining (27) and (28), we obtain that

$$\|\widehat{\theta} - \theta^*\|^2 \lesssim \epsilon_1^2 \cdot \big(1 + m \cdot (\epsilon_2^2 + 1)\big) + \epsilon_2^2 \cdot \big(1 + n \cdot (\epsilon_1^2 + 1)\big).$$

□

A.2. Proof of Theorem 2.4

Proof. Since we use the frequency estimator to estimate the QRE, we have

$$\begin{aligned}\mathbb{E}\left[\text{TV}(\widehat{\mu}, \mu^*)\right] &= \frac{1}{2} \sum_{a \in \mathcal{A}} \mathbb{E}\left[|\widehat{\mu}(a) - \mu^*(a)|\right] \le \frac{1}{2} \sum_{a \in \mathcal{A}} \sqrt{\mathbb{E}\left[(\widehat{\mu}(a) - \mu^*(a))^2\right]} \\ &= \frac{1}{2} \sum_{a \in \mathcal{A}} \sqrt{\frac{1}{N} \mu^*(a)(1 - \mu^*(a))} \le \frac{1}{2\sqrt{N}} \sum_{a \in \mathcal{A}} \sqrt{\mu^*(a)} \le \frac{1}{2} \sqrt{\frac{|\mathcal{A}|}{N}}.\end{aligned} \tag{29}$$

Let $A_1, \cdots, A_N \sim \text{Multinomial}(\mu^*)$ be the i.i.d. actions taken according to strategy μ^*. We then write the total variation as

$$\text{TV}(\widehat{\mu}, \mu) = f(A_1, \cdots, A_N) = \frac{1}{2} \sum_{a \in \mathcal{A}} \left| \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{A_i = a\}} - \mu^* \right|$$

Then the function $f : \mathcal{A}^n \to [0, 1]$ satisfy the bounded difference property for all $k \in [N]$:

$$\sup_{a_k, a_k' \in \mathcal{A}} |f(a_1, \cdots, a_{k-1}, a_k, a_{k+1} \cdots, a_N) - f(a_1, \cdots, a_{k-1}, a_k', a_{k+1} \cdots, a_N)| \le \frac{1}{N}.$$

By McDiarmid's inequality, for any $\epsilon > 0$, we have

$$\mathbb{P}\left(\text{TV}(\widehat{\mu}, \mu^*) - \mathbb{E}\left[\text{TV}(\widehat{\mu}, \mu^*)\right] \ge \epsilon\right) \le e^{-2N\epsilon^2}.$$

Combining this inequality with (29), we obtain the following tail bound:

$$\mathbb{P}\left(\text{TV}(\widehat{\mu}, \mu^*) \ge \frac{1}{2}\sqrt{\frac{m}{N}} + \sqrt{\frac{\log(2/\delta)}{2N}}\right) \le \frac{\delta}{2}. \tag{30}$$

Similarly, we can bound the total variation between $\widehat{\nu}$ and ν:

$$\mathbb{P}\left(\mathrm{TV}(\widehat{\nu}, \nu^*) \geq \frac{1}{2}\sqrt{\frac{n}{N}} + \sqrt{\frac{\log(2/\delta)}{2N}}\right) \leq \frac{\delta}{2}.$$

Using Theorem 2.3, $\mathrm{TV}(\widehat{\mu}, \mu^*) \leq \epsilon_1/2$ and $\mathrm{TV}(\widehat{\nu}, \nu^*) \leq \epsilon_2/2$ imply that

$$\|\widehat{\theta} - \theta^*\|^2 \lesssim \epsilon_1^2 \cdot \big(1 + m \cdot (\epsilon_2^2 + 1)\big) + \epsilon_2^2 \cdot \big(1 + n \cdot (\epsilon_1^2 + 1)\big) \simeq m\epsilon_1^2 + n\epsilon_2^2.$$

Since $Q(a, b) = \phi(a, b)^\top \theta$ for all $a, b \in \mathcal{A} \times \mathcal{B}$, we have

$$\begin{aligned}\|\widehat{Q} - Q\|_F^2 &= \sum_{a,b \in \mathcal{A}\times\mathcal{B}} (\widehat{Q}(a, b) - Q(a, b))^2 = \sum_{a,b \in \mathcal{A}\times\mathcal{B}} (\phi(a, b)^\top(\widehat{\theta} - \theta))^2 \\ &\leq \left(\sum_{a,b \in \mathcal{A}\times\mathcal{B}} \|\phi(a, b)\|^2\right) \|\widehat{\theta} - \theta\|^2 \lesssim m\epsilon_1^2 + n\epsilon_2^2.\end{aligned}$$

Therefore, for any $\delta \in (0, 1)$, we set

$$\epsilon_1 = \frac{\sqrt{m} + \sqrt{2\log(2/\delta)}}{\sqrt{N}}, \quad \epsilon_2 = \frac{\sqrt{n} + \sqrt{2\log(2/\delta)}}{\sqrt{N}}, \tag{31}$$

and obtain the following probability bound

$$\mathbb{P}\left(\|\widehat{Q} - Q\|_F^2 \lesssim \frac{m^2 + n^2 + (m + n)\log(1/\delta)}{N}\right) \geq 1 - \delta,$$

which is the desired result. □

A.3. Proof of Theorem 2.5

We first summarize the algorithm we propose in 2.3.

Algorithm 1 Learning payoff from actions

```
Require: Dataset $\mathcal{D} = \{(a^k, b^k)\}_{k\in[N]}$, kernel $\phi(\cdot,\cdot)$, entropy regularization parameter $\eta$, threshold parameter $\kappa$, ridge regularization term $\lambda$.
for $(a, b) \in \mathcal{A} \times \mathcal{B}$ do
    Compute the empirical QRE by
        $\widehat{\mu}(a) = \frac{1}{N}\sum_{k=1}^{N} \mathbb{1}_{\{a^k = a\}}, \quad \widehat{\nu}(b) = \frac{1}{N}\sum_{k=1}^{N} \mathbb{1}_{\{b^k = b\}}.$
    Construct the confidence set for $\theta$:
        $\widehat{\Theta} = \left\{\theta : \left\| \begin{bmatrix} A(\widehat{\nu}) \\ B(\widehat{\mu}) \end{bmatrix} \theta - \begin{bmatrix} c(\widehat{\mu}) \\ d(\widehat{\nu}) \end{bmatrix} \right\|^2 \leq \kappa, \|\theta\| \leq M \right\}.$
    Compute the feasible payoff matrices
        $\widehat{Q}(a, b) = \phi(a, b)^\top \widehat{\theta}, \quad \widehat{\theta} \in \widehat{\Theta}.$
end for
```

Mathematically, we use the Hausdorff distance $d_H(\cdot, \cdot)$ to quantify the difference between two sets in the parameter space.

Definition A.1 (Hausdorff distance)**.** Let $(\mathcal{M}, d)$ be a metric space. For each pair of non-empty subsets $X \subset \mathcal{M}$ and $Y \subset \mathcal{M}$, the Hausdorff distance between X and Y is defined as

$$d_H(X, Y) := \max\left\{\sup_{x\in X} d(x, Y), \sup_{y\in Y} d(X, y)\right\}.$$

Lemma A.2 (Construction error)**.** *Let ϵ_1 and ϵ_2 be two small numbers satisfying $\epsilon_1 < \min_{a\in[m]} \mu^*(a)$ and $\epsilon_2 < \min_{b\in[n]} \nu^*(b)$. We define normalized feature matrices Φ_1 and Φ_2 as*

$$\Phi_1 = ((\phi(i,\cdot) - \phi(1,\cdot)))_{i\in[m]\setminus\{1\}} \in \mathbb{R}^{d\times(m-1)n}, \quad \Phi_2 = ((\phi(\cdot,j) - \phi(\cdot,1)))_{j\in[n]\setminus\{1\}} \in \mathbb{R}^{d\times(n-1)m},$$

where we construct normalized features of the two players' actions by comparing to their first actions, respectively. Recall that the parameter M is specified in Assumption 2.1. Under this assumption, we define

$$\kappa = 2\left(\left(M\|\Phi_1\|_{op}^2 + \frac{n}{\eta^2\left(\min_{j\in[n]}\nu_j - \epsilon_2\right)^2}\right)\epsilon_2^2 + \left(M\|\Phi_2\|_{op}^2 + \frac{m}{\eta^2\left(\min_{i\in[m]}\mu_i - \epsilon_1\right)^2}\right)\epsilon_1^2\right).$$

Then, the confidence set $\hat{\Theta}$ constructed in (6) with parameters κ and M satisfies $\Theta \subseteq \hat{\Theta}$. That is, $\hat{\Theta}$ contains all feasible parameters. Moreover, when ϵ_1, ϵ_2 are sufficiently small, we have

$$d_H(\Theta, \hat{\Theta}) \lesssim \sqrt{\kappa}. \tag{32}$$

Proof. See Appendix D.1 for the complete proof. □

Under the estimate (31), the Theorem 2.5 is a direct corollary of Lemma A.2.

A.4. Proof of Theorem 2.6

Proof. We first control the error of Moore-Penrose inverse:

$$\begin{aligned}
\|\widehat{X}^\dagger - X^\dagger\|_{\text{op}} &= \left\|\widehat{X}^\top(\widehat{X}\widehat{X}^\top)^{-1} - X^\top(XX^\top)^{-1}\right\|_{\text{op}} \\
&\le \left\|\widehat{X}^\top(\widehat{X}\widehat{X}^\top)^{-1} - \widehat{X}^\top(XX^\top)^{-1}\right\|_{\text{op}} + \left\|(\widehat{X} - X)^\top(XX^\top)^{-1}\right\|_{\text{op}} \\
&\le \|\widehat{X}\|_{\text{op}}\left\|(\widehat{X}\widehat{X}^\top)^{-1}(XX^\top - \widehat{X}\widehat{X}^\top)(XX^\top)^{-1}\right\|_{\text{op}} + \|\widehat{X} - X\|_{\text{op}}\left\|(XX^\top)^{-1}\right\|_{\text{op}} \\
&\le \left(\|\widehat{X}\|_{\text{op}}\left\|(\widehat{X}\widehat{X}^\top)^{-1}\right\|_{\text{op}}\left\|XX^\top - \widehat{X}\widehat{X}^\top\right\|_{\text{op}} + \|\widehat{X} - X\|_{\text{op}}\right)\left\|(XX^\top)^{-1}\right\|_{\text{op}}.
\end{aligned} \tag{33}$$

Note that

$$\begin{aligned}
\|\widehat{X}\|_{\text{op}}^2 = \|\widehat{X}^\top\widehat{X}\|_{\text{op}} &= \|A(\widehat{\nu})^\top A(\widehat{\nu}) + B(\widehat{\mu})^\top B(\widehat{\mu})\|_{\text{op}} \\
&\le \|\Phi_1(I_{m-1}\otimes\widehat{\nu})\|_{\text{op}}^2 + \|\Phi_2(I_{n-1}\otimes\widehat{\mu})\|_{\text{op}}^2 \\
&\le \|\Phi_1\|_{\text{op}}^2 + \|\Phi_2\|_{\text{op}}^2,
\end{aligned} \tag{34}$$

and

$$\|XX^\top - \widehat{X}\widehat{X}^\top\|_{\text{op}} \le \|X(X - \widehat{X})^\top\|_{\text{op}} + \|\widehat{X}(X - \widehat{X})^\top\|_{\text{op}} \le 2\sqrt{\|\Phi_1\|_{\text{op}}^2 + \|\Phi_2\|_{\text{op}}^2}\|\widehat{X} - X\|_{\text{op}}. \tag{35}$$

Since the smallest singular value of X is bounded from below by $\sigma_b > 0$, we have

$$\left\|(XX^\top)^{-1}\right\|_{\text{op}} \le \frac{1}{\sigma_b^2}. \tag{36}$$

By Weyl's inequality, for sufficiently small $\epsilon_1, \epsilon_2 > 0$ with $\|\Phi_1\|_{\text{op}}^2\epsilon_2^2 + \|\Phi_2\|_{\text{op}}^2\epsilon_1^2 \le \sigma_b^2/4$, one have

$$\sigma_{\min}(\widehat{X}) \ge \sigma_{\min}(X) - \|\widehat{X} - X\|_{\text{op}} \ge \sigma_b - \sqrt{\|\Phi_1\|_{\text{op}}^2\epsilon_2^2 + \|\Phi_2\|_{\text{op}}^2\epsilon_1^2} \ge \frac{\sigma_b}{2}.$$

Hence

$$\left\|(\widehat{X}\widehat{X}^\top)^{-1}\right\|_{\mathrm{op}} \le \frac{4}{\sigma_b^2}. \tag{37}$$

Combining (33), (34), (35), (36) and (37), we have

$$\|\widehat{X}^\dagger - X^\dagger\|_{\mathrm{op}} \le \left(\frac{8(\|\Phi_1\|_{\mathrm{op}}^2 + \|\Phi_2\|_{\mathrm{op}}^2)}{\sigma_b^4} + \frac{1}{\sigma_b^2}\right)\sqrt{\|\Phi_1\|_{\mathrm{op}}^2\epsilon_2^2 + \|\Phi_2\|_{\mathrm{op}}^2\epsilon_1^2}. \tag{38}$$

Meanwhile, as in the proof of Theorem 2.3, we have

$$\|\widehat{y} - y\|^2 = \|c(\widehat{\nu}) - c(\nu^*)\|^2 + \|d(\widehat{\mu}) - d(\mu^*)\|^2 \le \frac{m\epsilon_1^2}{\eta^2(\min_{i\in[m]}\mu_i - \epsilon_1)^2} + \frac{n\epsilon_2^2}{\eta^2(\min_{j\in[n]}\nu_j - \epsilon_2)^2}. \tag{39}$$

Combining (38) and (39), we derive that

$$\begin{aligned}
\|\widehat{\theta} - \theta^*\| &\le \|X^\dagger - \widehat{X}^\dagger\|_{\mathrm{op}}\|y\| + \|\widehat{X}^\dagger\|_{\mathrm{op}}\|y - \widehat{y}\| \\
&\le \frac{(8\|\Phi_1\|_{\mathrm{op}}^2 + 8\|\Phi_2\|_{\mathrm{op}}^2 + \sigma_b^2)\|y\|}{\sigma_b^4}\sqrt{\|\Phi_1\|_{\mathrm{op}}^2\epsilon_2^2 + \|\Phi_2\|_{\mathrm{op}}^2\epsilon_1^2} \\
&\quad + \frac{2}{\sigma_b}\sqrt{\frac{m\epsilon_1^2}{\eta^2(\min_{i\in[m]}\mu_i - \epsilon_1)^2} + \frac{n\epsilon_2^2}{\eta^2(\min_{j\in[n]}\nu_j - \epsilon_2)^2}}.
\end{aligned}$$

According to our proof in Appendix A.2, with probability at least $1-\delta$, we have the bound (31) for the total variation error between optimal and empirical policies. Plugging in (31) to the last display, we have

$$\|\widehat{\theta} - \theta^*\| \lesssim \frac{m + n + \sqrt{(m+n)\log(1/\delta)}}{\sqrt{N}}.$$

This is the desired result. □

Remark A.3. The assumption of this Theorem is stronger than what we need in Theorem 2.5, because we require that $d > m + n - 2$ and X is of full row rank. This assumption is necessary for establishing the convergence of Moore-Penrose inverse:

$$\begin{bmatrix} A(\widehat{\nu}) \\ B(\widehat{\mu}) \end{bmatrix}^\dagger \to \begin{bmatrix} A(\nu^*) \\ B(\mu^*) \end{bmatrix}^\dagger.$$

This assumption is reasonable because, when the rank condition (4) is not satisfied, the kernel ϕ may capture redundant or irrelevant features in our system. Consequently, the dimensionality d becomes unnecessarily large.

B. Proof of Entropy-Regularized Markov Games

B.1. Proof of Proposition 3.5

Proof. For any $h \in [H]$, recall the QRE constraint

$$\mu_h^*(a|s) = \frac{e^{\eta\langle Q_h^*(s,a,\cdot),\nu_h^*(\cdot|s)\rangle_{\mathcal{B}}}}{\sum_{a\in\mathcal{A}} e^{\eta\langle Q_h^*(s,a,\cdot),\nu_h^*(\cdot|s)\rangle_{\mathcal{B}}}}, \quad \nu_h^*(b|s) = \frac{e^{-\eta\langle Q_h^*(s,\cdot,b),\mu_h^*(\cdot|s)\rangle_{\mathcal{A}}}}{\sum_{b\in\mathcal{B}} e^{-\eta\langle Q_h^*(s,\cdot,b),\mu_h^*(\cdot|s)\rangle_{\mathcal{A}}}}, \quad \forall (s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}.$$

This non-linear constraint is equivalent to the linear system:

$$\begin{cases} (Q_h(s,a,\cdot) - Q_h(s,1,\cdot))\,\nu_h^*(\cdot|s) = \log(\mu_h^*(a|s)/\mu_h^*(1|s))/\eta, & \text{for all } a \in \mathcal{A}, \\ (Q_h(s,\cdot,b) - Q_h(s,\cdot,1))\,\mu_h^*(\cdot|s) = -\log(\nu_h^*(b|s)/\nu_h^*(1|s))/\eta, & \text{for all } b \in \mathcal{B}. \end{cases}$$

Under Assumption 3.3, these linear equations can be rewritten as follows,

$$\begin{cases} \langle(\phi(s,a,\cdot) - \phi(s,1,\cdot))\,\nu_h^*(\cdot|s), \theta_h\rangle = \log(\mu_h^*(a|s)/\mu_h^*(1|s))/\eta, & \text{for all } a \in \mathcal{A}, \\ \langle(\phi(s,\cdot,b) - \phi(s,\cdot,1))\,\mu_h^*(\cdot|s), \theta_h\rangle = -\log(\nu_h^*(b|s)/\nu_h^*(1|s))/\eta, & \text{for all } b \in \mathcal{B}, \end{cases}$$

By the definition of A_h and B_h, the above linear system can be rewritten as

$$\begin{bmatrix} A_h(s,\nu_h^*) \\ B_h(s,\mu_h^*) \end{bmatrix} \theta_h = \begin{bmatrix} c_h(s,\mu_h^*) \\ d_h(s,\nu_h^*) \end{bmatrix},$$

where

$$c_h(s,\mu_h) \coloneqq (\log(\mu_h(a|s)/\mu_h(1|s))/\eta)_{a\in\mathcal{A}/\{1\}} \in \mathbb{R}^{m-1},$$
$$d_h(s,\nu_h) \coloneqq (-\log(\nu_h(b|s)/\nu_h(1|s))/\eta)_{b\in\mathcal{B}/\{1\}} \in \mathbb{R}^{n-1}.$$

Therefore, there exists a unique θ_h satisfies this linear system for all $s \in \mathcal{S}$ iff

$$\text{rank}\left(\begin{bmatrix} A_h(\nu_h^*)^\top & B_h(\mu_h^*)^\top \end{bmatrix}\right) = d.$$

Then we complete the proof. □

B.2. Analysis of Estimation Error of Q-functions

In this subsection, we analyze the estimation error of Q-functions in the Markov game. We first assume that the estimated QRE $(\widehat{\mu}, \widehat{\nu})$ is close to the true QRE (μ^*, ν^*), and analyze the accuracy of the confidence sets $\widehat{\Theta}_h$ for parameters θ_h.

Lemma B.1 (Estimation error of Q functions)**.** *Under Assumption 3.3, denote by $\widehat{\Theta}_h$ the confidence set (15) obtained by Algorithm 2 for each $h \in [H]$. Let ϵ_1 and ϵ_2 be two small positive numbers such that $\epsilon_1 < \min_{s\in\mathcal{S},a\in[m]} \mu_h^*(a|s)$ and $\epsilon_2 < \min_{s\in\mathcal{S},b\in[n]} \nu_h^*(b|s)$. Suppose that TV$(\hat{\mu}_h^*(\cdot|s), \mu_h^*(\cdot|s)) \le \epsilon_1/2$ and TV$(\hat{\nu}_h^*(\cdot|s), \nu_h^*(\cdot|s)) \le \epsilon_2/2$ for all $s \in \mathcal{S}$ and all $h \in [H]$. Let*

$$\kappa_h = \left(2R^2\|\Phi_1\|_{\text{op}}^2 + \frac{2Sm}{\eta^2\left(\min\limits_{s\in\mathcal{S},a\in[m]} \mu_h^*(a|s) - \epsilon_1\right)^2}\right)\epsilon_1^2 + \left(2R^2\|\Phi_2\|_{\text{op}}^2 + \frac{2Sn}{\eta^2\left(\min\limits_{s\in\mathcal{S},b\in[n]} \nu_h^*(b|s) - \epsilon_2\right)^2}\right)\epsilon_2^2,$$

where the normalized feature matrices $\Phi_1 \in \mathbb{R}^{d\times s(m-1)n}$ and $\Phi_2 \in \mathbb{R}^{d\times s(n-1)m}$ are defined as

$$\Phi_1 = (\phi(s,a,\cdot) - \phi(s,1,\cdot))_{s\in\mathcal{S},a\in[m]\setminus\{1\}}, \quad \Phi_2 = (\phi(s,\cdot,b) - \phi(s,\cdot,1))_{s\in\mathcal{S},b\in[n]\setminus\{1\}}.$$

Then, we have $\Theta_h \subseteq \hat{\Theta}_h$. Furthermore, for sufficiently small $\epsilon_1, \epsilon_2 > 0$, we have

$$D_H(\Theta_h, \hat{\Theta}_h) \lesssim \sqrt{\kappa_h}. \tag{40}$$

Proof. See Appendix D.2 for the complete proof. □

Next, we need to solve the sample complexity on the concentration of the QRE. Under the assumption 3.8, every state $s \in \mathcal{S}$ of our system is visited with sufficient frequency as the size of our dataset increases. We are then prepared to present the concentration result for the frequency-based estimator of the QRE.

Lemma B.2 (Concentration of QRE)**.** *Under Assumptions 3.3 and 3.8, let ϵ be a small positive number such that $\epsilon \le \min_{h\in[H],s\in\mathcal{S},a\in[m],b\in[n]}\{\mu_h^*(a|s), \nu_h^*(b|s)\}/2$. Let T, the number of sample episodes, be*

$$T = \frac{1}{2C^2}\cdot\log(2HS/\delta) + \frac{m\vee n}{2C\epsilon^2}\cdot\log(4HS/\delta).$$

where $\delta \in (0,1)$. Define the concentration event $\mathcal{E}_1$ as

$$\mathcal{E}_1 = \left\{\text{TV}(\widehat{\mu}_h(\cdot|s), \mu_h^*(\cdot|s)) \le \epsilon, \text{TV}(\widehat{\nu}_h(\cdot|s), \nu_h^*(\cdot|s)) \le \epsilon, \forall s\in\mathcal{S}, \forall h\in[H]\right\}. \tag{41}$$

Then the event $\mathcal{E}_1$ holds with probability at least $1-\delta$.

Proof. See Appendix D.3 for the complete proof. □

Combining Lemma B.2 and Lemma B.1, we obtain the following result concerning the sample complexity required to construct the confidence set $\widehat{\Theta}_h$.

Corollary B.3 (Sample complexity of constructing feasible Q function set)**.** *Under Assumptions 3.3 and 3.8, and given any $\delta \in (0,1)$, we set $T \in \mathbb{N}$ large and $\epsilon > 0$ small so that*

$$T \geq \frac{1}{C^2}\log\frac{2HS}{\delta}, \quad \textit{and} \quad \epsilon = \sqrt{\frac{m \vee n}{CT}\log\frac{4HS}{\delta}} \leq \frac{1}{4}\min_{s\in\mathcal{S},a\in[m],b\in[n]}\{\mu_h^*(a|s), \nu_h^*(b|s)\}.$$

Then the concentration event $\mathcal{E}_1$ (41) in Lemma B.2 holds with probability at least $1-\delta$. Moreover, we set the threshold parameter as

$$\kappa_h = 8\epsilon^2\left(R^2(\|\Phi_1\|_{\mathrm{op}}^2 + \|\Phi_2\|_{\mathrm{op}}^2) + \frac{4Sm}{\eta^2\min_{s\in\mathcal{S},a\in[m]}\mu_h^*(a|s)^2} + \frac{4Sn}{\eta^2\min_{s\in\mathcal{S},b\in[n]}\nu_h^*(b|s)^2}\right) = \mathcal{O}\left(\frac{1}{T}\right).$$

Then for each $h \in [H]$, the concentration event $\mathcal{E}_2$, defined by

$$\mathcal{E}_2 = \left\{\Theta_h \subseteq \hat{\Theta}_h \textit{ and } D_H(\Theta_h, \hat{\Theta}_h) \lesssim \sqrt{\kappa_h},\ \forall h \in [H]\right\},$$

holds with probability at least $1-\delta$.

B.3. Proof of Theorem 3.9

We first summarize the Algorithm we propose in §3.2 below.

Algorithm 2 Learning reward from actions with the frequency estimator of QRE

```
Require: Dataset $\mathcal{D} = \{(s_h^t, a_h^t, b_h^t)\}_{h\in[H],t\in[T]}$, kernel $\phi(\cdot,\cdot,\cdot)$, entropy regularization parameter $\eta$, discount factor $\gamma$, threshold parameter $(\kappa_h)$, ridge regularization term $\lambda$.
for $(h, s, a, b) \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ do
    Compute the empirical QRE by
        $$\widehat{\mu}_h(a|s) = \frac{1}{N_h(s) \vee 1} \sum_{(s_h,a_h)\in\mathcal{D}} \mathbb{1}_{(s_h,a_h)=(s,a)}, \quad \widehat{\nu}_h(b|s) = \frac{1}{N_h(s) \vee 1} \sum_{(s_h,b_h)\in\mathcal{D}} \mathbb{1}_{(s_h,b_h)=(s,b)},$$
    where $N_h(s) = \sum_{(s_h,a_h,b_h)\in\mathcal{D}_h} \mathbb{1}_{\{s_h=s\}}$;
end for
for $h = H, H-1, \cdots, 1$ do
    Construct the confidence set for $\theta_h$:
        $$\widehat{\Theta}_h = \left\{ \|\theta\| \le R : \left\| \begin{bmatrix} A_h(\widehat{\nu}_h) \\ B_h(\widehat{\mu}_h) \end{bmatrix} \theta - \begin{bmatrix} c_h(\widehat{\mu}_h) \\ d_h(\widehat{\nu}_h) \end{bmatrix} \right\|^2 \le \kappa_h \right\}$$
    Compute the feasible Q-functions and V-functions by containing all $\widehat{Q}_h$ and $\widehat{V}_h$ such that
        $$\widehat{Q}_h(s,a,b) = \phi(s,a,b)^\top \widehat{\theta}_h, \quad \text{where} \quad \widehat{\theta}_h \in \widehat{\Theta}_h,$$
        $$\widehat{V}_h(s) = \widehat{\mu}_h(s)^\top \widehat{Q}_h(s) \widehat{\nu}_h(s) + \eta^{-1}\mathcal{H}(\widehat{\mu}_h(s)) - \eta^{-1}\mathcal{H}(\widehat{\nu}_h(s));$$
    Compute the empirical transition kernel by
        $$\Lambda_h = \sum_{t=1}^T \phi(s_h^t, a_h^t, b_h^t)\phi(s_h^t, a_h^t, b_h^t)^\top + \lambda \mathbf{I}_d,$$
        $$\widehat{\mathbb{P}}_h \widehat{V}_{h+1}(s,a,b) = \phi(s,a,b)^\top \Lambda_h^{-1} \sum_{t=1}^T \phi(s_h^t, a_h^t, b_h^t) \widehat{V}_{h+1}(s_{h+1}^t);$$
    Compute the reward by
        $$\widehat{r}_h(s,a,b) = \widehat{Q}_h(s,a,b) - \gamma \widehat{\mathbb{P}}_h \widehat{V}_{h+1}(s,a,b).$$
end for
```

We require some technical results for the proof.

Ridge Regression Analysis. Indeed, given any $\lambda > 0$, Algorithm 2 estimates the true vectors $\pi_h(\cdot)$ specifying transition kernels $\mathbb{P}_h$ in Assumption 3.3 by solving the ridge regression problem:

$$\widehat{\Pi}_h = \operatorname*{argmin}_{\Pi_h \in \mathbb{R}^{S\times d}} \sum_{t=1}^T \left\| \Pi_h \phi(s_h^t, a_h^t, b_h^t) - \delta(s_{h+1}^t) \right\|^2 + \lambda \|\Pi_h\|_{\mathrm{F}}^2,$$

where $\Pi_h = (\pi_h(s))_{s\in\mathcal{S}}^\top \in \mathbb{R}^{S\times d}$. Then

$$\widehat{\Pi}_h = \sum_{t=1}^T \delta(s_{h+1}^t)\phi(s_h^t, a_h^t, b_h^t)^\top \Lambda_h^{-1}, \quad \text{where} \;\; \Lambda_h = \sum_{t=1}^T \phi(s_h^t, a_h^t, b_h^t)\phi(s_h^t, a_h^t, b_h^t)^\top + \lambda I_d.$$

Correspondingly, the estimate of the reward function is given by

$$\widehat{r}_h(s,a,b) = \widehat{Q}_h(s,a,b) - \gamma \widehat{\mathbb{P}}_h \widehat{V}_{h+1}(s,a,b) = \widehat{Q}_h(s,a,b) - \gamma \phi(s,a,b)^\top \widehat{\Pi}_h^\top \widehat{V}_{h+1}.$$

This is the approximation approach of Algorithm 2.

Property of estimated V-functions. According to Assumption 3.3 and row 3 of Algorithm 2, for all $h \in [H]$, we have the following bound for estimated V-functions:

$$-R - \eta^{-1}\log n \le \widehat{V}_h(s) \le R + \eta^{-1}\log m, \quad \forall s \in \mathcal{S}. \tag{42}$$

We denote by $\widehat{\mathcal{V}}_h$ the set of V-functions $\widehat{V}_h$ generated by $\widehat{\Theta}_h$. For any $\theta, \theta' \in \widehat{\Theta}_h$, we have

$$\Big|\widehat{V}_h(s;\theta) - \widehat{V}_h(s;\theta')\Big| = \Big|\widehat{\mu}_h(s)^\top(\widehat{Q}_h(s;\theta) - \widehat{Q}_h(s;\theta'))\widehat{\nu}_h(s)\Big| \le \|\theta - \theta'\|, \quad \forall s \in \mathcal{S}.$$

Consequently, the covering number of $\widehat{\mathcal{V}}_h$ is not greater by the covering number of $\widehat{\Theta}_h$, which is included in the ball $B(0,R) = \{\theta \in \mathbb{R}^d : \|\theta\| \le R\}$. By Lemma F.2, for any $\epsilon > 0$, we can find an ϵ-net $\mathcal{N}_\epsilon$ of $(\widehat{\mathcal{V}}_h, \|\cdot\|_\infty)$ such that

$$|\mathcal{N}_\epsilon| \le \left(1 + \frac{2R}{\epsilon}\right)^d. \tag{43}$$

Clearly, the same property also apply to the set of feasible V-functions $\mathcal{V}_h$ generated by Θ_h.

B.3.1. Bounding the Estimation Error of V-functions

In this subsection, we bound the estimation error of V-functions under the event $\mathcal{E}_2$ given by Corollary B.3. For any $\widehat{\theta}_h \in \widehat{\Theta}_h$, we can find a feasible $\theta_h \in \Theta_h$ such that $\|\widehat{\theta}_h - \theta_h\| \le C_h\sqrt{\kappa_h}$ for some constant C_h. Then for the V-functions generated by $\widehat{\theta}_h$ and θ_h, we have

$$\begin{aligned}
&|\widehat{V}_h(s) - V_h(s)| \\
&\le \left|\widehat{\mu}_h(s)^\top \widehat{Q}_h(s)\widehat{\nu}_h(s) + \frac{1}{\eta}\mathcal{H}(\widehat{\mu}_h(s)) - \frac{1}{\eta}\mathcal{H}(\widehat{\nu}_h(s)) - \mu_h^*(s)^\top Q_h(s)\nu_h^*(s) - \frac{1}{\eta}\mathcal{H}(\mu_h^*(s)) + \frac{1}{\eta}\mathcal{H}(\nu_h^*(s))\right| \\
&\le \underbrace{\left|\widehat{\mu}_h(s)^\top \widehat{Q}_h(s)\widehat{\nu}_h(s) - \mu_h^*(s)^\top Q_h(s)\nu_h^*(s)\right|}_{\text{(i)}} + \underbrace{\frac{1}{\eta}\left|\mathcal{H}(\widehat{\mu}_h(s)) - \mathcal{H}(\mu_h^*(s))\right|}_{\text{(ii)}} + \underbrace{\frac{1}{\eta}\left|\mathcal{H}(\widehat{\nu}_h(s)) - \mathcal{H}(\nu_h^*(s))\right|}_{\text{(iii)}}.
\end{aligned} \tag{44}$$

Now we successively bound the three terms above. Firstly,

$$\begin{aligned}
\text{(i)} &\le \left|\widehat{\mu}_h(s)^\top(\widehat{Q}_h - Q_h)(s)\widehat{\nu}_h(s)\right| + \left|(\widehat{\mu}_h - \mu_h^*)(s)^\top Q_h(s)\widehat{\nu}_h(s)\right| + \left|\mu_h^*(s)^\top Q_h(s)(\widehat{\nu}_h - \nu_h^*)(s)\right| \\
&\le \sup_{a\in\mathcal{A}, b\in\mathcal{B}} \left|\widehat{Q}_h - Q_h\right|(s,a,b) + \|(\widehat{\mu}_h - \mu_h^*)(s)\|_1 \|Q_h(s)\widehat{\nu}_h(s)\|_\infty + \|(\widehat{\nu}_h - \nu_h^*)(s)\|_1 \|Q_h(s)^\top \widehat{\mu}_h(s)\|_\infty \\
&\le \sup_{a\in\mathcal{A}, b\in\mathcal{B}} \left|\widehat{Q}_h - Q_h\right|(s,a,b) + 2\epsilon \sup_{a\in\mathcal{A}, b\in\mathcal{B}} |Q_h(s,a,b)| + 2\epsilon \sup_{a\in\mathcal{A}, b\in\mathcal{B}} |Q_h(s,a,b)| \\
&\le C_h\sqrt{\kappa_h} + 4R\epsilon,
\end{aligned} \tag{45}$$

where we use Hölder's inequality in the first inequality, and the third inequality follows from the following facts:

$$\begin{aligned}
\left|\widehat{Q}_h - Q_h\right|(s,a,b) &= \left|(\widehat{\theta}_h - \theta_h)^\top \phi(s,a,b)\right| \le \|\widehat{\theta}_h - \theta_h\|\,\|\phi(s,a,b)\| \le C_h\sqrt{\kappa_h}, \\
\left|Q_h(s,a,b)\right| &= |\theta_h^\top \phi(s,a,b)| \le \|\theta_h\|\,\|\phi(s,a,b)\| \le R, \quad \forall (s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}.
\end{aligned}$$

To bound terms (ii) and (iii), we use Lemma F.1:

$$\begin{aligned}
|\mathcal{H}(\mu_h^*(s)) - \mathcal{H}(\widehat{\mu}(s))| &\le -\epsilon\log\epsilon - (1-\epsilon)\log(1-\epsilon) + \epsilon\log(m-1) \le \epsilon\left(1 + \log\frac{m}{\epsilon}\right); \\
|\mathcal{H}(\nu_h^*(s)) - \mathcal{H}(\widehat{\nu}(s))| &\le -\epsilon\log\epsilon - (1-\epsilon)\log(1-\epsilon) + \epsilon\log(n-1) \le \epsilon\left(1 + \log\frac{n}{\epsilon}\right);
\end{aligned} \tag{46}$$

Combining (44), (45) and (46), for all $s \in \mathcal{S}$, we obtain

$$\begin{aligned}
|\widehat{V}_h(s) - V_h(s)| &\le C_h\sqrt{\kappa_h} + \epsilon\left(4R + 2 + \log\frac{mn}{\epsilon^2}\right) \\
&\lesssim \left(\sqrt{S(m+n)} + \log(Tmn)\right)\sqrt{\frac{m \vee n}{T}\log\frac{HS}{\delta}}.
\end{aligned}$$

Furthermore, for all $h \in [H]$ and all $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$, since $\|\phi(\cdot,\cdot,\cdot)\| \le 1$, we have

$$\begin{aligned}|\mathbb{P}_h \widehat{V}_{h+1} - \mathbb{P}_h V_{h+1}|(s,a,b) &\le \|\phi(s,a,b)\| \, \|\Pi_h\|_{\mathrm{op}} \, \|\widehat{V}_h - V_h\| \\ &\lesssim \frac{S(m+n) + \sqrt{S(m+n)}\log T}{\sqrt{T}} \sqrt{\log\frac{HS}{\delta}}.\end{aligned} \tag{47}$$

Note this bound is valid once the concentration event $\mathcal{E}_2$ holds.

B.3.2. Bounding the Estimation Error of Transition Kernels

In this section, we bound the difference between $\widehat{\mathbb{P}}_h \widehat{V}_{h+1}$ and $\mathbb{P}_h \widehat{V}_{h+1}$.

Error decomposition. We fix $\widehat{V}_{h+1} \in \widehat{\mathcal{V}}_{h+1}$, and let $\mathscr{H}_h^t$ be the σ-field generated by variables $(s_1, a_1, b_1, s_2, a_2, b_2, \cdots, s_h, a_h, b_h)$. Define random variables $\eta_h^t = \delta(s_{h+1}^t) - \Pi_h \phi(s_h^t, a_h^t, b_h^t)$. Then $\mathbb{E}[\eta_h^t | \mathscr{H}_h^t] = 0$. For all $s \in \mathcal{S}, a \in \mathcal{A}, b \in \mathcal{B}$,

$$\begin{aligned}\big(\widehat{\mathbb{P}}_h \widehat{V}_{h+1} - \mathbb{P}_h \widehat{V}_{h+1}\big)(s,a,b) &= \widehat{V}_{h+1}^\top \big(\widehat{\Pi}_h - \Pi_h\big)\phi(s,a,b) \\ &= \widehat{V}_{h+1}^\top \left(\sum_{t=1}^T \big(\eta_h^t + \Pi_h \phi(s_h^t, a_h^t, b_h^t)\big)\, \phi(s_h^t, a_h^t, b_h^t)^\top \Lambda_h^{-1}\phi(s,a,b) - \Pi_h\right)\phi(s,a,b) \\ &= \sum_{t=1}^T \left(\widehat{V}_{h+1}^\top \eta_h^t\right)\phi(s_h^t, a_h^t, b_h^t)^\top \Lambda_h^{-1}\phi(s,a,b) - \lambda \widehat{V}_{h+1}^\top \Pi_h \Lambda_h^{-1}\phi(s,a,b).\end{aligned}$$

By Cauchy's inequality, we obtain

$$\left|\widehat{\mathbb{P}}_h \widehat{V}_{h+1} - \mathbb{P}_h \widehat{V}_{h+1}\right|(s,a,b) \le \left(\left\|\sum_{t=1}^T \left(\widehat{V}_{h+1}^\top \eta_h^t\right)\phi(s_h^t, a_h^t, b_h^t)\right\|_{\Lambda_h^{-1}} + \lambda \left\|\Pi_h^\top \widehat{V}_{h+1}\right\|_{\Lambda_h^{-1}}\right) \|\phi(s,a,b)\|_{\Lambda_h^{-1}}, \tag{48}$$

where we write $\|x\|_M := \sqrt{x^\top M x}$ for any positive definite matrix M. Now we bound the three terms in (48).

Step I: Analysis of the term $\|\Pi_h^\top \widehat{V}_{h+1}\|_{\Lambda_h^{-1}}$. Since $\Lambda_h = \lambda I_d + \sum_{t=1}^T \phi(s_h^t, a_h^t, b_h^t)\phi(s_h^t, a_h^t, b_h^t)^\top$, we have $\lambda_{\min}(\Lambda_h) \ge \lambda$. Hence

$$\|\Pi_h^\top \widehat{V}_{h+1}\|_{\Lambda_h^{-1}} \le \frac{1}{\sqrt{\lambda}}\|\Pi_h^\top \widehat{V}_{h+1}\| \le \frac{1}{\sqrt{\lambda}}\|\Pi_h\|_{\mathrm{op}}\|\widehat{V}_{h+1}\|.$$

Note that we assume $\|\pi_h(\cdot)\| \le \sqrt{d}$ in Assumption 3.3, we have $\|\Pi_h\|_{\mathrm{op}} \le \sqrt{Sd}$. Hence

$$\|\Pi_h^\top \widehat{V}_{h+1}\|_{\Lambda_h^{-1}} \le \sqrt{\frac{Sd}{\lambda}}(2R + \eta^{-1}\log m + \eta^{-1}\log n). \tag{49}$$

Step II: Analysis of the self-normalized process. By definition of η_h^t, we have $\|\eta_h^t\|_\infty \le 2$. Combining with the bound of $\widehat{V}_{h+1}$ given by (42), we obtain the following estimate:

$$-2(R + \eta^{-1}\log n) \le \widehat{V}_{h+1}^\top \eta_h^t \le 2(R + \eta^{-1}\log m).$$

By Hoeffding's inequality, $(\widehat{V}_{h+1}^\top \eta_h^t)_{t=1}^T$ are independent $(2R + \eta^{-1}\log m + \eta^{-1}\log n)$-sub-Gaussian random variables with zero mean. By Lemma F.4, the following inequality holds with probability at least $1-\delta$:

$$\left\|\sum_{t=1}^T \widehat{V}_{h+1}^\top \eta_h^t \phi(s_h^t, a_h^t, b_h^t)\right\|_{\Lambda_h^{-1}}^2 \le 2(2R + \eta^{-1}\log m + \eta^{-1}\log n)^2 \log\left(\frac{\det(\Lambda_h)^{1/2}\det(\lambda I_d)^{-1/2}}{\delta}\right).$$

Now we bound $\det(\Lambda_h)$. To this end, note that

$$\det(\Lambda_h) \le \|\Lambda_h\|_{\mathrm{op}}^d \le \left(\lambda + T\|\phi(s_h^t, a_h^t, b_h^t)\|^2\right)^d \le (\lambda + T)^d.$$

Consequently, we have

$$\left\|\sum_{t=1}^T \widehat{V}_{h+1}^\top \eta_h^t \phi(s_h^t, a_h^t, b_h^t)\right\|_{\Lambda_h^{-1}}^2 \le 2(2R + \eta^{-1}\log mn)^2 \left(\log\frac{1}{\delta} + \frac{d}{2}\log\left(1 + \frac{T}{\lambda}\right)\right). \tag{50}$$

Next, we derive a uniform bound for $\widehat{\mathcal{V}}_{h+1}$. We choose a ϵ-net of $(\widehat{\mathcal{V}}_{h+1}, \|\cdot\|_\infty)$ given in (43). Then for all $\widehat{V}_{h+1} \in \widehat{\mathcal{V}}_{h+1}$, we can find $V^* \in \mathcal{N}_\epsilon$ such that $\|\widehat{V}_{h+1} - V^*\|_\infty \le \epsilon$. Hence

$$\left\|\sum_{t=1}^T (\widehat{V}_{h+1} - V^*)^\top \eta_h^t \phi(s_h^t, a_h^t, b_h^t)\right\|_{\Lambda_h^{-1}} \le \frac{1}{\sqrt{\lambda}}\sum_{t=1}^T \|\widehat{V}_{h+1} - V^*\|_\infty \|\eta_h^t\|_1 \left\|\phi(s_h^t, a_h^t, b_h^t)\right\| \le \frac{2T\epsilon}{\sqrt{\lambda}}.$$

We apply a union bound version of (50) on $\mathcal{N}_\epsilon$. With probability at least $1-\delta$, the following inequality holds for all $\widehat{\mathcal{V}}_{h+1}$:

$$\begin{aligned}\left\|\sum_{t=1}^T \widehat{V}_{h+1}^\top \eta_h^t \phi(s_h^t, a_h^t, b_h^t)\right\|_{\Lambda_h^{-1}} &\le \left\|\sum_{t=1}^T V^{*\top} \eta_h^t \phi(s_h^t, a_h^t, b_h^t)\right\|_{\Lambda_h^{-1}} + \left\|\sum_{t=1}^T (\widehat{V}_{h+1} - V^*)^\top \eta_h^t \phi(s_h^t, a_h^t, b_h^t)\right\|_{\Lambda_h^{-1}} \\ &\le \sqrt{2}(2R + \eta^{-1}\log mn)\sqrt{\log\frac{|\mathcal{N}_\epsilon|}{\delta} + \frac{d}{2}\log\left(1 + \frac{T}{\lambda}\right)} + \frac{2T\epsilon}{\sqrt{\lambda}}.\end{aligned}$$

Take $\epsilon = 1/T$. Then we obtain

$$\left\|\sum_{t=1}^T \widehat{V}_{h+1}^\top \eta_h^t \phi(s_h^t, a_h^t, b_h^t)\right\|_{\Lambda_h^{-1}}^2 \le (2R + \eta^{-1}\log mn)\sqrt{2\log\frac{(1+2RT)^d}{\delta} + d\log\left(1 + \frac{T}{\lambda}\right)} + \frac{2}{\sqrt{\lambda}}. \tag{51}$$

Step III: Analysis of $\|\phi(s,a,b)\|_{\Lambda_h^{-1}}$. Let ρ_h be the visitation measure of (s_h, a_h, b_h) induced by QRE policies μ^* and ν^*. For simplicity, we use the following notations:

$$\overline{\Lambda}_h = \frac{1}{T}\Lambda_h = \frac{\lambda}{T}I_d + \frac{1}{T}\sum_{t=1}^T \phi(s_h^t, a_h^t, b_h^t)\phi(s_h^t, a_h^t, b_h^t)^\top, \quad \Psi_h = \mathbb{E}\left[\phi(s_h, a_h, b_h)\phi(s_h, a_h, b_h)^\top\right].$$

This step is somewhat tricky. We first present our main result below.

Lemma B.4 (Adapted from Min et al., 2022 Lemma H.5)**.** *Let $\{(s_h^t, a_h^t, b_h^t)\}_{t=1}^T$ be i.i.d. samples from the visitation distribution ρ_h. For any $\delta > 0$, if*

$$T \ge \max\left\{512\|\Psi_h^{-1}\|_{\mathrm{op}}^2 \log\frac{2d}{\delta}, 4\lambda\|\Psi_h^{-1}\|_{\mathrm{op}}\right\}, \tag{52}$$

then with probability at least $1-\delta$, it holds simultaneously for all $s \in \mathcal{S}, a \in \mathcal{A}, b \in \mathcal{B}$ that

$$\|\phi(s,a,b)\|_{\Lambda_h^{-1}} \le \frac{2}{\sqrt{T}}\|\phi(s,a,b)\|_{\Psi_h^{-1}}.$$

Proof. We first bound the difference between $\overline{\Lambda}_h$ and Ψ_h. We write $x_t = \phi(s_h^t, a_h^t, b_h^t)$, and define the matrix-valued function $\Sigma(x_1, \cdots, x_T) = \frac{\lambda}{T}I_d + \frac{1}{T}\sum_{t=1}^T x_t x_t^\top$. For any $t \in [T]$, if we replace x_t by some $\widetilde{x}_t$ with $\|\widetilde{x}_t\| \le 1$, we have

$$\begin{aligned}&\left(\Sigma(x_1, \cdots, x_{t-1}, x_t, x_{t+1}, \cdots, x_T) - \Sigma(x_1, \cdots, x_{t-1}, \widetilde{x}_t, x_{t+1}, \cdots, x_T)\right)^2 \\ &= \frac{1}{T^2}\left(x_t x_t^\top - \widetilde{x}_t\widetilde{x}_t^\top\right)^2 \preceq \frac{1}{T^2}\left(2x_t x_t^\top x_t x_t^\top + 2\widetilde{x}_t\widetilde{x}_t^\top\widetilde{x}_t\widetilde{x}_t^\top\right) \preceq \frac{4}{T^2}I_d =: A_t^2.\end{aligned}$$

Let $\sigma^2 = \left\|\sum_{t=1}^T A_t^2\right\| = \frac{4}{T}$. By Lemma F.3, with probability at least $1-\delta$, we have

$$\left\|\overline{\Lambda}_h - \mathbb{E}[\overline{\Lambda}_h]\right\|_{\mathrm{op}} \le \frac{4\sqrt{2}}{\sqrt{T}}\sqrt{\log\frac{2d}{\delta}} \quad \Rightarrow \quad \left\|\overline{\Lambda}_h - \Psi_h\right\|_{\mathrm{op}} \le \frac{4\sqrt{2}}{\sqrt{T}}\sqrt{\log\frac{2d}{\delta}} + \frac{\lambda}{T}. \tag{53}$$

Next, we bound the term $\|v\|_{\Lambda_h^{-1}}$ for any $v = \phi(s,a,b) \in \mathbb{R}^d$:

$$\begin{aligned}
\|v\|_{\Lambda_h^{-1}} &= \frac{1}{\sqrt{T}}\sqrt{v^\top \Psi_h^{-1} v + v^\top(\overline{\Lambda}_h^{-1} - \Psi_h^{-1})v} \\
&= \frac{1}{\sqrt{T}}\sqrt{v^\top \Psi_h^{-1} v + v^\top \Psi_h^{-1/2}(\Psi_h^{1/2}\overline{\Lambda}_h^{-1}\Psi_h^{1/2} - I_d)\Psi_h^{-1/2} v} \\
&\le \frac{1}{\sqrt{T}}\sqrt{\|v\|_{\Psi_h^{-1}}\left(1 + \left\|\Psi_h^{1/2}\overline{\Lambda}_h^{-1}\Psi_h^{1/2} - I_d\right\|_{\mathrm{op}}\right)\|v\|_{\Psi_h^{-1}}} \\
&\le \frac{1}{\sqrt{T}}\left(1 + \left\|\Psi_h^{1/2}\overline{\Lambda}_h^{-1}\Psi_h^{1/2} - I_d\right\|_{\mathrm{op}}^{1/2}\right)\|v\|_{\Psi_h^{-1}}.
\end{aligned} \tag{54}$$

To control the term $\left\|\Psi_h^{1/2}\overline{\Lambda}_h^{-1}\Psi_h^{1/2} - I_d\right\|_{\mathrm{op}}$, note that

$$\left\|\Psi_h^{1/2}\overline{\Lambda}_h^{-1}\Psi_h^{1/2} - I_d\right\|_{\mathrm{op}} \le \left\|\left(\Psi_h^{-1/2}\overline{\Lambda}_h\Psi_h^{-1/2}\right)^{-1}\right\|_{\mathrm{op}} \left\|I_d - \Psi_h^{-1/2}\overline{\Lambda}_h\Psi_h^{-1/2}\right\|_{\mathrm{op}}$$

Combining the hypothesis (52) and the bound (53), we have

$$\left\|I_d - \Psi_h^{-1/2}\overline{\Lambda}_h\Psi_h^{-1/2}\right\|_{\mathrm{op}} = \|\Psi_h^{-1}\|_{\mathrm{op}}\|\overline{\Lambda}_h - \Psi_h^{-1}\|_{\mathrm{op}} \le \frac{1}{2}.$$

Moreover, by Weyl's inequality, we have

$$\lambda_{\min}\left(\Psi_h^{-1/2}\overline{\Lambda}_h\Psi_h^{-1/2}\right) \ge 1 - \lambda_{\max}\left(I_d - \Psi_h^{-1/2}\overline{\Lambda}_h\Psi_h^{-1/2}\right) \ge \frac{1}{2}.$$

Consequently, we have

$$\left\|\Psi_h^{1/2}\overline{\Lambda}_h^{-1}\Psi_h^{1/2} - I_d\right\|_{\mathrm{op}} \le \frac{1}{\lambda_{\min}\left(\Psi_h^{-1/2}\overline{\Lambda}_h\Psi_h^{-1/2}\right)}\left\|I_d - \Psi_h^{-1/2}\overline{\Lambda}_h\Psi_h^{-1/2}\right\|_{\mathrm{op}} \le 2\cdot\frac{1}{2} = 1.$$

Plugging in the last display to (54) completes the proof. □

Final bound. We can obtain the final bound by combining (48), (49), (51) and Lemma B.4. For any $\delta > 0$, if

$$T \ge \max\left\{512\|\Psi_h^{-1}\|_{\mathrm{op}}^2 \log\frac{2Hd}{\delta}, 4\lambda\|\Psi_h^{-1}\|_{\mathrm{op}}\right\}, \tag{55}$$

then with probability at least $1-2\delta$, it holds simultaneously for all $h \in [H]$, all $\widehat{V}_{h+1} \in \widehat{\mathcal{V}}_{h+1}$, and all $s \in \mathcal{S}, a \in \mathcal{A}, b \in \mathcal{B}$ that

$$\begin{aligned}
&\left|\widehat{\mathbb{P}}_h\widehat{V}_{h+1} - \mathbb{P}_h\widehat{V}_{h+1}\right|(s,a,b) \\
&\le \frac{1}{\sqrt{T}}\left((2R + \eta^{-1}\log mn)\left(\sqrt{\frac{Sd}{\lambda}} + \sqrt{2\log\frac{H(1+2RT)^d}{\delta} + d\log\frac{\lambda+T}{\lambda}}\right) + \frac{2}{\sqrt{\lambda}}\right)\|\phi(s,a,b)\|_{\Psi_h^{-1}} \\
&\le \frac{(2R + \eta^{-1}\log mn)}{\sqrt{T}}\left(\sqrt{\frac{4Sd}{\lambda}} + \sqrt{2\log\frac{H}{\delta} + 2d\log(1+2RT) + d\log\frac{\lambda+T}{\lambda}}\right)\|\Psi_h^{-1}\|_{\mathrm{op}}^{1/2} \\
&\lesssim \frac{\sqrt{Sd} + \sqrt{d\log T} + \sqrt{\log(H/\delta)}}{\sqrt{T}}\log(mn)
\end{aligned} \tag{56}$$

Clearly, this upper bound converges to zero as the sample size $T \to \infty$.

B.3.3. Bounding the Estimation Error of Rewards

Now we go back to the proof of Theorem 3.9, and bound $D_H(\mathcal{R}, \widehat{\mathcal{R}})$.

Proof of Theorem 3.9. For any $h \in [H]$ and $\widehat{r} \in \widehat{\mathcal{R}}$, the estimated reward $\widehat{r}_h$ can be rewritten as

$$\widehat{r}_h(s,a,b) = \widehat{Q}_h(s,a,b) - \gamma \widehat{\mathbb{P}}_h \widehat{V}_{h+1}(s,a,b) \quad \text{for all } (s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B},$$

which is generated by some $\widehat{\theta}_h \in \widehat{\Theta}_h$ and $\widehat{V}_{h+1} \in \widehat{\mathcal{V}}_{h+1}$. When the concentration event $\mathcal{E}_2$ given in Corollary B.2 holds, we can find a feasible $\theta_h \in \Theta_h$ such that $\|\widehat{\theta}_h - \theta_h\| \leq C_h \sqrt{\kappa_h}$ for some constant $C_h > 0$. Moreover, according to our conclusion in Appendix B.3.1, we can also find a feasible $V_{h+1} \in \mathcal{V}_{h+1}$ such that (47) holds for all possible choice of h and (s,a,b). Now we consider the following feasible reward:

$$r_h(s,a,b) = \phi(s,a,b)^\top \theta_h - \gamma \mathbb{P}_h V_{h+1}(s,a,b), \quad r = (r_1, r_2, \cdots, r_H) \in \mathcal{R}.$$

We have the following decomposition:

$$\begin{aligned} |\widehat{r}_h - r_h|(s,a,b) &= \left|\phi(s,a,b)^\top (\widehat{\theta}_h - \theta_h) - \gamma(\widehat{\mathbb{P}}_h \widehat{V}_{h+1} - \mathbb{P}_h V_{h+1})(s,a,b)\right| \\ &\leq \underbrace{\|\phi(s,a,b)\| \, \|\widehat{\theta}_h - \theta_h\|}_{\leq C_h \sqrt{\kappa_h}} + \underbrace{\gamma \left|\widehat{\mathbb{P}}_h \widehat{V}_{h+1} - \mathbb{P}_h \widehat{V}_{h+1}\right|(s,a,b)}_{\text{bounded by (56)}} + \underbrace{\gamma \left|\mathbb{P}_h \widehat{V}_{h+1} - \mathbb{P}_h V_{h+1}\right|(s,a,b)}_{\text{bounded by (47)}} \end{aligned}$$

In this display, both the first and third bounds hold under the concentration event $\mathcal{E}_2$, and the second holds with probability $1 - 2\delta$ when T satisfies (55). Furthermore, all these bounds are uniform and do not depend on our choice of h, $\widehat{\theta}_h$, $\widehat{V}_{h+1}$ and (s,a,b). Consequently, with probability at least $1 - 3\delta$, we have

$$\sup_{\widehat{r} \in \widehat{\mathcal{R}}} d(\widehat{r}, \mathcal{R}) \lesssim \sqrt{\frac{S(m+n)}{T} \log \frac{HS}{\delta}} \left(\sqrt{S(m+n)} + \log T\right) + \frac{\sqrt{Sd} + \sqrt{d \log T}}{\sqrt{T}} \log(mn). \tag{57}$$

On the other hand, for any $r \in \mathcal{R}$ generated by $\theta_h \in \Theta_h \subseteq \widehat{\Theta}_h$ and $\theta_{h+1} \in \Theta_{h+1} \subseteq \widehat{\Theta}_{h+1}$, we pick the following estimated reward $\widehat{r} = (\widehat{r}_1, \widehat{r}_2, \cdots, \widehat{r}_H) \in \widehat{\mathcal{R}}$:

$$\begin{aligned} &\widetilde{V}_{h+1}(s) = \widehat{\mu}_{h+1}(s)^\top Q_{h+1}(s) \widehat{\nu}_{h+1}(s) + \eta^{-1} \mathcal{H}(\widehat{\nu}_{h+1}(s)) - \eta^{-1} \mathcal{H}(\widehat{\nu}_{h+1}(s)), \\ &\widehat{r}_h(s,a,b) = \phi(s,a,b)^\top \theta_h - \widehat{\mathbb{P}}_h \widetilde{V}_{h+1}(s,a,b), \quad h \in [H], s \in \mathcal{S}, a \in \mathcal{A}, b \in \mathcal{B}. \end{aligned}$$

Since $\Theta_{h+1} \subseteq \widehat{\Theta}_{h+1}$, the estimated V-function $\widetilde{V}_{h+1} \in \widehat{\mathcal{V}}_{h+1}$, and the uniform bound (56) still holds if we replacing $\widehat{V}_{h+1}$ by $\widetilde{V}_{h+1}$. Furthermore, similar to our estimation procedure (44)-(47), the estimation error of V-function has the following bound:

$$\begin{aligned} &|\widetilde{V}_{h+1}(s) - V_{h+1}(s)| \leq \left(4R + 2 + \log \frac{mn}{\epsilon^2}\right) \epsilon \lesssim \log(Tmn) \sqrt{\frac{m \vee n}{T} \log \frac{HS}{\delta}}, \\ &|\mathbb{P}_h \widetilde{V}_{h+1} - \mathbb{P}_h V_{h+1}|(s,a,b) \lesssim \frac{\sqrt{S(m+n)} \log(Tmn)}{\sqrt{T}} \sqrt{\log \frac{HS}{\delta}}, \quad \forall (s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}. \end{aligned} \tag{58}$$

Consequently, the following inequality simultaneously holds for all choice of h, r_h and (s,a,b):

$$\begin{aligned} |\widehat{r}_h - r_h|(s,a,b) &= |\widehat{\mathbb{P}}_h \widetilde{V}_{h+1} - \mathbb{P}_h V_{h+1}|(s,a,b) \\ &\leq \underbrace{|\widehat{\mathbb{P}}_h \widetilde{V}_{h+1} - \mathbb{P}_h \widetilde{V}_{h+1}|(s,a,b)}_{\text{bounded by (56)}} + \underbrace{|\mathbb{P}_h \widetilde{V}_{h+1} - \mathbb{P}_h V_{h+1}|(s,a,b)}_{\text{bounded by (58)}} \\ &\lesssim \sqrt{\frac{S(m+n)}{T} \log \frac{HS}{\delta}} \log(Tmn) + \left(\sqrt{\frac{Sd}{T}} + \sqrt{\frac{d \log T}{T}}\right) \log(mn). \end{aligned}$$

Therefore we have

$$\sup_{r\in\mathcal{R}} d(r,\widehat{\mathcal{R}}) \lesssim \sqrt{\frac{S(m+n)}{T}\log\frac{HS}{\delta}}\log(Tmn) + \left(\sqrt{\frac{Sd}{T}} + \sqrt{\frac{d\log T}{T}}\right)\log(mn). \tag{59}$$

Combining (57) and (59), we obtain the Hausdorff distance between $\mathcal{R}$ and $\widehat{\mathcal{R}}$:

$$\begin{aligned} D_H(\mathcal{R},\widehat{\mathcal{R}}) &= \max\left\{\sup_{\widehat{r}\in\widehat{\mathcal{R}}} d(\widehat{r},\mathcal{R}),\ \sup_{r\in\mathcal{R}} d(r,\widehat{\mathcal{R}})\right\} \\ &\lesssim \sqrt{\frac{S(m+n)}{T}\log\frac{HS}{\delta}}\left(\sqrt{S(m+n)}+\log T\right) + \frac{\sqrt{Sd}+\sqrt{d\log T}}{\sqrt{T}}\log(mn). \end{aligned}$$

This complete the proof. □

C. Incorporation of Maximum Likelihood Estimation (MLE)

C.1. Using MLE for QRE Estimation

To control the error of the estimated QRE $(\widehat{\mu},\widehat{\nu})$ in the tabular case, we assume that every state $s\in\mathcal{S}$ is visited with sufficiently high probability. This strong assumption can be restrictive in practical scenarios, as it requires an exhaustive exploration of the state space $\mathcal{S}$. To alleviate this limitation, we assume that the QRE (μ^*,ν^*) exhibits a sparse structure. To simplify our problem, we adopt a linear parameterization of the QRE. By adjusting the dimension of the parameters, we can capture the key features of the QRE.

Assumption C.1 (Linearly parameterized QRE)**.** Assume $d_a, d_b \in \mathbb{N}$. Let $\psi_a : \mathcal{S}\times\mathcal{A}\to\mathbb{R}^{d_a}$ and $\psi_b : \mathcal{S}\times\mathcal{B}\to\mathbb{R}^{d_b}$ be two bounded kernels such that

$$\|\psi_a(s,a)\|, \|\psi_b(s,b)\| \le K$$

for all $s\in\mathcal{S}, a\in\mathcal{A}$ and $b\in\mathcal{B}$. For each $h\in[H]$, there exists a vector $\vartheta_h^*\in\mathbb{R}^{d_a}$ such that $\|\vartheta_h^*\|\le 1$, and for all $s_h\in\mathcal{S}, a_h\in\mathcal{A}$,

$$\mu_h^*(a_h|s_h) = \frac{\exp\left(\vartheta_h^{*\top}\psi_a(s_h,a_h)\right)}{\sum_{a\in\mathcal{A}}\exp\left(\vartheta_h^{*\top}\psi_a(s_h,a)\right)};$$

Also, there exists a vector $\zeta_h^*\in\mathbb{R}^{d_b}$ such that $\|\zeta_h^*\|\le 1$, and for all $s_h\in\mathcal{S}, b_h\in\mathcal{B}$,

$$\nu_h^*(b_h|s_h) = \frac{\exp\left(\zeta_h^{*\top}\psi_b(s_h,b_h)\right)}{\sum_{b\in\mathcal{B}}\exp\left(\zeta_h^{*\top}\psi_b(s_h,b)\right)}.$$

Linear parametrization allows us to integrate the estimation of the QRE $(\mu_h^*(\cdot|s),\nu_h^*(\cdot|s))$ across different states $s\in\mathcal{S}$ by solving the parameters with maximum likelihood estimation (MLE). We can explicitly write the log-likelihood function of parameters ϑ_h and ζ_h given the dataset

$$\mathcal{D}_h = \left\{(s_h^1,a_h^1,b_h^1),(s_h^2,a_h^2,b_h^2),\cdots,(s_h^T,a_h^T,b_h^T)\right\}_{h=1}^H.$$

For the max player, the negative log-likelihood function is

$$\mathcal{L}_h(\vartheta_h) = -\sum_{t=1}^T \frac{\exp\left(\vartheta_h^\top\psi_a(s_h^t,a_h^t)\right)}{\sum_{a\in\mathcal{A}}\exp\left(\vartheta_h^\top\psi_a(s_h^t,a)\right)},$$

and for the min player, it is

$$\mathcal{L}_h(\zeta_h) = -\sum_{t=1}^T \frac{\exp\left(\zeta_h^\top\psi_b(s_h^t,b_h^t)\right)}{\sum_{b\in\mathcal{B}}\exp\left(\zeta_h^\top\psi_b(s_h^t,b)\right)}.$$

We estimate the parameters ϑ_h^* and ζ_h^* by minimizing their negative log-likelihood:

$$\widehat{\vartheta}_h = \operatorname*{argmin}_{\|\vartheta_h\|\leq 1} \mathcal{L}_h(\vartheta_h), \quad \text{and} \quad \widehat{\zeta}_h = \operatorname*{argmin}_{\|\zeta_h\|\leq 1} \mathcal{L}_h(\zeta_h). \tag{60}$$

As in the tabular case, it can be challenging to control the error of the estimated QRE across all states $s \in \mathcal{S}$. Nevertheless, we can focus on the average error and obtain the following L^2-convergence result for the MLEs.

Lemma C.2 (Convergence of MLE)**.** *Under Assumption C.1, we let $\widehat{\mu}_h$ and $\widehat{\nu}_h$ be the policies generated by $\widehat{\vartheta}_h$ and $\widehat{\zeta}_h$. Then there exists a constant $L_K > 0$ depending on m, n and K only, such that with probability at least $1-\delta$, it holds*

$$\mathbb{E}\left[H^2\left(\widehat{\mu}_h(\cdot|s_h), \mu_h^*(\cdot|s_h)\right)\right] \leq \frac{1}{T}\left(d_a \log(1+2TL_K) + \log\frac{1}{\delta} + \sqrt{2m}e^K + 2\right), \tag{61}$$

where $H(P,Q) = \|\sqrt{P} - \sqrt{Q}\|_2$ is the Hellinger distance between two probability distributions, and the expectation is taken with respect to the visitation distribution d_h^* of state s_h generated by the QRE (μ^*, ν^*). According to the relation $\mathrm{TV}(P,Q) \leq \sqrt{2}H(P,Q)$, the estimation (61) implies

$$\frac{1}{2}\mathbb{E}\left[\mathrm{TV}^2\left(\widehat{\mu}_h(\cdot|s_h), \mu_h^*(\cdot|s_h)\right)\right] \leq \frac{1}{T}\left(d_a \log(1+2TL_K) + \log\frac{1}{\delta} + \sqrt{2m}e^K + 2\right).$$

Similarly, with probability at least $1-\delta$, the following inequality holds:

$$\frac{1}{2}\mathbb{E}\left[\mathrm{TV}^2\left(\widehat{\nu}_h(\cdot|s_h), \nu_h^*(\cdot|s_h)\right)\right] \leq \frac{1}{T}\left(d_b \log(1+2TL_K) + \log\frac{1}{\delta} + \sqrt{2n}e^K + 2\right).$$

Proof. See Appendix C.2.2 for the complete proof. □

Since some states $s \in \mathcal{S}$ are less frequently visited, or worse, not contained in our dataset, the error of our estimator can be significantly higher in these states. Fortunately, such inaccuracies are less critical for states with low visitation probabilities, and it suffices to control the average error. We hence introduce a weaker metric to measure the distance between rewards.

Definition C.3 (L^1-metric for rewards)**.** Let ρ_h be the state-visitation measure corresponding to the true QRE (μ^*, ν^*). Define the L^1 metric D_1 between any pair of rewards $r, r' : [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ as

$$D_1(r, r') = \sup_{h\in[H], a\in\mathcal{A}, b\in\mathcal{B}} \mathbb{E}_{s\sim\rho_h} \left|(r_h - r_h')(s,a,b)\right|.$$

To align with the estimation of the QRE, it is necessary to adapt the estimation method of Q-functions. Since we have higher accuracies on frequently visited states, we may estimate the parameters (θ_h) by solving the least square problem weighted by visitation probabilities. Before we proceed, we introduce another type of identification of Q-functions.

Proposition C.4 (Strong identification of Q functions)**.** *Under Assumption 3.3, for any $h \in [H]$, $Q_h(s,a,b) = \phi(s,a,b)^\top \theta_h$ is feasible for all $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ if and only if θ_h satisfies the following linear system:*

$$\begin{bmatrix} A_h(\nu_h^*) \\ B_h(\mu_h^*) \end{bmatrix} \theta_h = \begin{bmatrix} c_h(\mu_h^*) \\ d_h(\nu_h^*) \end{bmatrix} \quad \textit{for all } s \in \mathcal{S}, \tag{62}$$

where

$$A_h(\nu_h^*) := \begin{bmatrix} \sqrt{\rho_h(1)} A_h(1, \nu_h^*) \\ \vdots \\ \sqrt{\rho_h(S)} A_h(S, \nu_h^*) \end{bmatrix}, \quad B_h(\mu_h^*) := \begin{bmatrix} \sqrt{\rho_h(1)} B_h(1, \mu_h^*) \\ \vdots \\ \sqrt{\rho_h(S)} B_h(S, \mu_h^*) \end{bmatrix},$$

and

$$c_h(\mu_h^*) := \begin{bmatrix} \sqrt{\rho_h(1)}\, c_h(1, \mu_h^*) \\ \vdots \\ \sqrt{\rho_h(S)}\, c_h(S, \mu_h^*) \end{bmatrix}, \quad d_h(\nu_h^*) := \begin{bmatrix} \sqrt{\rho_h(1)}\, d_h(1, \nu_h^*) \\ \vdots \\ \sqrt{\rho_h(S)}\, d_h(S, \nu_h^*) \end{bmatrix}.$$

There exists a unique feasible $\theta_h \in \mathbb{R}^d$ if and only if the QRE satisfies the rank condition

$$\textit{rank}\left(\begin{bmatrix} A_h(\nu_h^*)^\top & B_h(\mu_h^*)^\top \end{bmatrix}\right) = d. \tag{63}$$

Furthermore, the reward function r_h is uniquely identifiable if and only if the QRE satisfies the rank condition (63) for all indices $h, h+1, \cdots, H$.

Proof. The proof is exactly the same as Proposition 3.5. □

Estimation of Q-function. In accordance with the identification of Q, we need to change the construction of confidence set. Instead of simply concatenating the estimation equations over all $s \in \mathcal{S}$, we add a weight to each block to adjust the importance of each state:

$$A_h(\widehat{\nu}_h) := \begin{bmatrix} \sqrt{\rho_h(1)}A_h(1, \widehat{\nu}_h) \\ \vdots \\ \sqrt{\rho_h(S)}A_h(S, \widehat{\nu}_h) \end{bmatrix}, \quad B_h(\widehat{\mu}_h) := \begin{bmatrix} \sqrt{\rho_h(1)}B_h(1, \widehat{\mu}_h) \\ \vdots \\ \sqrt{\rho_h(S)}B_h(S, \widehat{\mu}_h) \end{bmatrix}.$$

Since the true distribution ρ_h of the state s_h is unknown, we replace it with the empirical estimator on the dataset $\mathcal{D}_h$:

$$\widehat{\rho}_h(s) = \mathbb{P}_{\mathcal{D}_h}(s_h = s) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\{s_h^t = s\}}, \quad s \in \mathcal{S}.$$

The matrices for confidence set construction are then given by

$$\widehat{A}_h(\widehat{\nu}_h) := \begin{bmatrix} \sqrt{\widehat{\rho}_h(1)}A_h(1, \widehat{\nu}_h) \\ \vdots \\ \sqrt{\widehat{\rho}_h(S)}A_h(S, \widehat{\nu}_h) \end{bmatrix}, \quad \widehat{B}_h(\widehat{\mu}_h) := \begin{bmatrix} \sqrt{\widehat{\rho}_h(1)}B_h(1, \widehat{\mu}_h) \\ \vdots \\ \sqrt{\widehat{\rho}_h(S)}B_h(S, \widehat{\mu}_h) \end{bmatrix}. \tag{64}$$

Similarly, we define

$$\widehat{c}_h(\widehat{\mu}_h) := \begin{bmatrix} \sqrt{\widehat{\rho}_h(1)}c_h(1, \widehat{\mu}_h) \\ \vdots \\ \sqrt{\widehat{\rho}_h(S)}c_h(S, \widehat{\mu}_h) \end{bmatrix}, \quad \widehat{d}_h(\widehat{\nu}_h) := \begin{bmatrix} \sqrt{\widehat{\rho}_h(1)}d_h(1, \widehat{\nu}_h) \\ \vdots \\ \sqrt{\widehat{\rho}_h(S)}d_h(S, \widehat{\nu}_h) \end{bmatrix}. \tag{65}$$

Given an appropriate threshold $\kappa_h > 0$, we construct the confidence set for parameter θ_h by solving the least square problem:

$$\widehat{\Theta}_h := \left\{ \theta : \left\| \begin{bmatrix} \widehat{A}_h(\widehat{\nu}_h) \\ \widehat{B}_h(\widehat{\mu}_h) \end{bmatrix} \theta - \begin{bmatrix} \widehat{c}_h(\widehat{\mu}_h) \\ \widehat{d}_h(\widehat{\nu}_h) \end{bmatrix} \right\|^2 \leq \kappa_h, \|\theta\| \leq R \right\},$$

This is equivalent to the weighted least square problem:

$$\sum_{s \in \mathcal{S}} \widehat{\rho}_h(s) \left\| \begin{bmatrix} A_h(s, \widehat{\nu}_h) \\ B_h(s, \widehat{\mu}_h) \end{bmatrix} \theta - \begin{bmatrix} c_h(s, \widehat{\mu}_h) \\ d_h(s, \widehat{\nu}_h) \end{bmatrix} \right\|^2 \leq \kappa_h.$$

We present the convergence result for the new confidence set below.

Lemma C.5. *Under Assumptions 3.3 and C.1, let $\widehat{\Theta}_h$ be the confidence set obtained. Set*

$$\epsilon^2 = \mathcal{O}\left(\frac{(d_a + d_b)\log T + \log(H/\delta) + \sqrt{m} + \sqrt{n}}{T}\right),$$

and

$$\kappa_h = \mathcal{O}\bigg(\frac{1}{T}\Big(m^{7/2} + n^{7/2} + S(\log mn)^2\Big) + \frac{m^3 + n^3}{T}\left((d_a + d_b)\log T + \log\frac{H}{\delta}\right)\bigg). \tag{66}$$

Then with probability at least $1-3\delta$, we have

$$\mathbb{E}_{s\sim\rho_h}[\mathrm{TV}^2(\widehat{\mu}_h(\cdot|s),\mu_h^*(\cdot|s))]\le\epsilon^2,\quad \mathbb{E}_{s\sim\rho_h}[\mathrm{TV}^2(\widehat{\nu}_h(\cdot|s),\nu_h^*(\cdot|s))]\le\epsilon^2,$$

and $\Theta_h\subseteq\widehat{\Theta}_h$ for all $h\in[H]$. Furthermore, for each $h\in[H]$, the Hausdorff distance between the feasible set and the confidence set satisfies the estimate

$$d_H(\Theta_h,\widehat{\Theta}_h)\lesssim\sqrt{\kappa_h}.$$

Proof. See Appendix D.4 for the complete proof. □

We summarize the algorithm for learning reward from actions in Algorithm 3.

Algorithm 3 Learning reward from actions (MLE case)

```
Require: Dataset $\mathcal{D}=\{(s_h^t,a_h^t,b_h^t)\}_{h\in[H],t\in[T]}$, kernels $\phi(\cdot,\cdot,\cdot)$, $\psi_a(\cdot,\cdot)$, $\psi_b(\cdot,\cdot)$, entropy regularization term $\eta$, discount factor $\gamma$, threshold parameter $(\kappa_h)$, ridge regularization term $\lambda$.
for $(h,s,a,b)\in[H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{B}$ do
    Compute the empirical QRE by maximal likelihood estimation, according to (60)
    Construct the confidence set for $\theta_h$:
        $$\widehat{\Theta}_h=\left\{\|\theta\|\le R:\left\|\begin{bmatrix}\widehat{A}_h(\widehat{\nu}_h)\\\widehat{B}_h(\widehat{\mu}_h)\end{bmatrix}\theta-\begin{bmatrix}\widehat{c}_h(\widehat{\mu}_h)\\\widehat{d}_h(\widehat{\nu}_h)\end{bmatrix}\right\|^2\le\kappa_h\right\},$$
    where $\widehat{A}_h(\widehat{\nu}_h),\widehat{B}_h(\widehat{\mu}_h),\widehat{c}_h(\widehat{\mu}_h)$ and $\widehat{d}_h(\widehat{\nu}_h)$ are given in (64) and (65);
    Compute the feasible Q-functions and V-functions by containing all $\widehat{Q}_h$ and $\widehat{V}_h$ such that
        $$\widehat{Q}_h(s,a,b)=\phi(s,a,b)^\top\widehat{\theta}_h\quad\text{where}\quad\widehat{\theta}_h\in\widehat{\Theta}_h,$$
        $$\widehat{V}_h(s)=\widehat{\mu}_h(s)^\top\widehat{Q}_h(s)\widehat{\nu}_h(s)+\eta^{-1}\mathcal{H}(\widehat{\mu}_h(s))-\eta^{-1}\mathcal{H}(\widehat{\nu}_h(s));$$
    Compute the empirical transition kernel by
        $$\Lambda_h=\sum_{t=1}^T\phi(s_h^t,a_h^t,b_h^t)\phi(s_h^t,a_h^t,b_h^t)^\top+\lambda\mathbf{I}_d,$$
        $$\widehat{\mathbb{P}}_h\widehat{V}_{h+1}(s,a,b)=\phi(s,a,b)^\top\Lambda_h^{-1}\sum_{t=1}^T\phi(s_h^t,a_h^t,b_h^t)\widehat{V}_{h+1}(s_{h+1}^t);$$
    Compute the reward by
        $$\widehat{r}_h(s,a,b)=\widehat{Q}_h(s,a,b)-\gamma\widehat{\mathbb{P}}_h\widehat{V}_{h+1}(s,a,b).$$
end for
```

To demonstrate the effectiveness of Algorithm 3, we provide the theoretical results for the feasible sets constructed by Algorithm 3 in the following theorem.

Theorem C.6. *Under Assumptions 3.3 and C.1, we let $\rho_h=d_h^*$ be the stationary distribution associated with the optimal policies μ^* and ν^*, where $h\in[H]$. We also assume that the following $d\times d$ matrix*

$$\Psi_h=\mathbb{E}_{\rho_h}\left[\phi(s_h,a_h,b_h)\phi(s_h,a_h,b_h)^\top\right]$$

is nonsingular for all $h\in[H]$. Let $\mathcal{R}$ be the feasible reward set in Definition 3.7. Given a dataset

$$\mathcal{D}=\{\mathcal{D}_h\}_{h\in[H]}=\{\{(s_h^t,a_h^t,b_h^t)\}_{t\in[T]}\}_{h\in[H]},$$

we set $\lambda = \mathcal{O}(1)$*, and as in (66), set* $\kappa_h = \mathcal{O}(T^{-1})$*, and denote by* $\hat{\mathcal{R}}$ *the output of Algorithm 3. For any* $\delta \in (0,1)$*, we have that, with probability at least* $1-4\delta$*,*

$$D_1(\mathcal{R}, \widehat{\mathcal{R}}) \lesssim \frac{1}{\sqrt{T}}\bigg(m^{7/4} + n^{7/4} + (m^{3/2} + n^{3/2} + \log T)\sqrt{(d_a + d_b + d)\log T + \log(H/\delta)} + \sqrt{S}d\log(mn)\bigg).$$

Proof. See Appendix C.3 for the complete proof. □

This theorem provides a finite-sample bound on the Hausdorff distance between the identifiable set of feasible rewards $\mathcal{R}$ and the estimated reward set $\widehat{\mathcal{R}}$ obtained via our algorithm with MLE-based QRE estimation. This result quantifies the sample complexity required for reliable and efficient reward inference in entropy-regularized Markov games. We point out several key insights emerge from this theorem:

- *Convergence Rate.* The bound reveals a convergence rate of roughly $\mathcal{O}(T^{-1/2})$, consistent with standard statistical results. This ensures that as the number of observed samples T grows, our estimated feasible reward set converges toward the true identifiable set.

- *Dependence on Problem Complexity.* Larger state or action spaces require more data to achieve the same accuracy level. In particular, the fractional exponents on action spaces m and n in this bound originate from the estimate for Hellinger distance in (61) and the application of Cauchy-Schwartz inequality. Since the exponent $7/4$ is still less than 2, the effect is milder than in worst-case scenarios where complexity scales quadratically or cubically.

- *Role of Feature Representations.* The terms involving the dimensions d, d_a and d_b, where d is the dimension of kernel ϕ for modeling reward and transitions, and d_a, d_n are the dimensions of policy kernels ψ_a, ψ_b for players $1, 2$, respectively. This involvement emphasizes the impact of the complexity of the feature representation on estimation accuracy. A richer representation (with larger dimensions) increases expressive power but also requires more data to ensure robust recovery.

- *Implications for Practical Applications.* This result reassures practitioners that our approach, which leverages maximum likelihood estimation for QRE recovery, provides reliable inference, even when the true reward parameter is partially identified. It establishes clear guidelines on data requirements for accurate reward reconstruction.

C.2. Convergence of MLE in Policy Estimation

C.2.1. Linear Parameterized Strategies

We prove the case for the max player, whose action space is $\mathcal{A}$. We assume that we have a feature map $\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_a}$ such that $\|\psi(s,a)\| \le K$ for all $s \in \mathcal{S}, a \in \mathcal{A}$. The strategy μ of the max player is then parameterized by a group of vectors $\vartheta_h \in \mathbb{R}^{d_a}$:

$$\mu_h(a_h|s_h) = \mu_{\vartheta_h}(a_h, s_h) := \frac{\exp\left(\vartheta_h^\top \psi(s_h, a_h)\right)}{\sum_{a\in\mathcal{A}} \exp\left(\vartheta_h^\top \psi(s_h, a)\right)}, \quad s_h \in \mathcal{S}, a_n \in \mathcal{A}, h = 1, 2, \cdots, H.$$

We assume that the true parameters $\vartheta_h^* \in \Gamma$, where $\Gamma = B(0,1)$ is the unit closed ball in $\mathbb{R}^{d_a}$ centered at 0. We define a pseudometric ρ on Γ:

$$\begin{aligned}\rho(\vartheta, \widetilde{\vartheta}) &= \max_{s\in\mathcal{S}} H\left(\mu_\vartheta(\cdot|s), \mu_{\widetilde{\vartheta}}(\cdot|s)\right) = \max_{s\in\mathcal{S}} \sqrt{H^2\left(\mu_\vartheta(\cdot|s), \mu_{\widetilde{\vartheta}}(\cdot|s)\right)} \\ &= \max_{s\in\mathcal{S}} \sqrt{\frac{1}{2}\sum_{a\in\mathcal{A}}\left(\sqrt{\mu_\vartheta(a|s)} - \sqrt{\mu_{\widetilde{\vartheta}}(a|s)}\right)^2},\end{aligned}$$

where $H^2(\cdot,\cdot)$ is the squared Hellinger distance between two probability distributions.

Lower bound for actions. Since the parameters $\vartheta_h \in \Gamma$, and the feature map $\psi : \mathcal{S} \times \mathcal{B} \to \mathbb{R}^{d_a}$ ranges in $B(0, K)$, we have

$$|\vartheta_h^\top \psi(s_h, a_h)| \leq \|\vartheta_h\| \, \|\psi(s_h, a_h)\| \leq K, \quad \forall s_h \in \mathcal{S}, a_h \in \mathcal{A}.$$

Consequently,

$$\mu_h(a_h|s_h) = \frac{\exp\left(\vartheta_h^\top \psi(s_h, a_h)\right)}{\sum_{a \in \mathcal{A}} \exp\left(\vartheta_h^\top \psi(s_h, a)\right)} \geq \frac{e^{-K}}{\sum_{a \in \mathcal{A}} e^K} = \frac{e^{-2K}}{m}.$$

Therefore, for every $h \in [H]$, $a_h \in \mathcal{A}$, $s_h \in \mathcal{S}$ and $\vartheta_h \in \Gamma$, the following bound holds:

$$\frac{e^{-2K}}{m} \leq \mu_h(a_h|s_h) \leq 1. \tag{67}$$

This bound is very useful when we analyze the variation of parameters ϑ_h.

Lemma C.7. *There exists a constant $L_K > 0$ such that for any $\vartheta, \widetilde{\vartheta} \in \Gamma$,*

$$\rho(\vartheta, \widetilde{\vartheta}) \leq L_K \|\vartheta - \widetilde{\vartheta}\|. \tag{68}$$

Proof. Define the softmax function $p = (p_1, \cdots, p_m) : \mathbb{R}^m \to [0,1]^m$ by

$$p(x_1, \cdots, x_m) = \left(\frac{e^{x_1}}{\sum_{i=1}^m e^{x_i}}, \cdots, \frac{e^{x_m}}{\sum_{i=1}^m e^{x_i}} \right).$$

The Jacobian matrix is then given by

$$J_p = \begin{pmatrix} p_1(1-p_1) & -p_1 p_2 & \cdots & -p_1 p_m \\ -p_2 p_1 & p_2(1-p_2) & \cdots & -p_2 p_m \\ \vdots & \vdots & \ddots & \vdots \\ -p_m p_1 & -p_m p_2 & \cdots & p_m(1-p_m). \end{pmatrix}$$

Note that $p_1 + p_2 + \cdots + p_m = 1$. By triangle inequality,

$$\|J_p\|_{\mathrm{op}} = \|\mathrm{diag}(p) - pp^\top\|_{\mathrm{op}} \leq \|\mathrm{diag}(p)\|_{\mathrm{op}} + \|pp^\top\|_{\mathrm{op}} \leq 2.$$

Hence p is a Lipschitz function. Now for every $s \in \mathcal{S}$, define

$$x^s = \left(\vartheta^\top \psi(s,1), \cdots, \vartheta^\top \psi(s,m)\right)^\top, \quad \widetilde{x}^s = \left(\widetilde{\vartheta}^\top \psi(s,1), \cdots, \widetilde{\vartheta}^\top \psi(s,m)\right)^\top.$$

Then

$$\|x^s - \widetilde{x}^s\| \leq \sqrt{\|\vartheta - \widetilde{\vartheta}\|^2 \sum_{a \in \mathcal{A}} \|\psi(x,a)\|^2} \leq \sqrt{m} K \|\vartheta - \widetilde{\vartheta}\|. \tag{69}$$

Furthermore, we can bound the Hellinger distance between μ_ϑ and $\mu_{\widetilde{\vartheta}}$ as follows:

$$\begin{aligned} \sqrt{H^2\left(\mu_\vartheta(\cdot|s), \mu_{\widetilde{\vartheta}}(\cdot|s)\right)} &\leq \sqrt{\frac{1}{2} \sum_{a \in \mathcal{A}} \frac{me^{2K}}{2} \left|\mu_\vartheta(a|s) - \mu_{\widetilde{\vartheta}}(a|s)\right|^2} = \sqrt{\frac{m}{4}} e^K \|p(x^s) - p(\widetilde{x}^s)\| \\ &\leq \sqrt{m} e^K \|x^s - \widetilde{x}\| \overset{(69)}{\leq} mKe^K \|\vartheta - \widetilde{\vartheta}\|, \end{aligned} \tag{70}$$

where in the first inequality, we use the fact that

$$\begin{aligned} \left|\sqrt{\mu_\vartheta(a|s)} - \sqrt{\mu_{\widetilde{\vartheta}}(a|s)}\right| &\leq \frac{1}{2\sqrt{\min\{\mu_\vartheta(a|s), \mu_{\widetilde{\vartheta}}(a|s)\}}} \left|\mu_\vartheta(a|s) - \mu_{\widetilde{\vartheta}}(a|s)\right| \\ &\overset{(67)}{\leq} \sqrt{\frac{me^{2K}}{2}} \left|\mu_\vartheta(a|s) - \mu_{\widetilde{\vartheta}}(a|s)\right|. \end{aligned}$$

The result (68) follows by taking the uniform bound (70) for all $s \in \mathcal{S}$. □

C.2.2. Proof of Lemma C.2

Proof. This conclusion is inspired by Chen et al., 2023. To recover the strategy taken by the max player from an offline dataset

$$\{(s_1^t, a_1^t), (s_2^t, a_2^t), \cdots, (s_H^t, a_H^t)\}_{t=1}^T,$$

we employ maximum likelihood estimation on parameters ϑ_h. The estimator $\widehat{\vartheta}_h$ is solved by minimzing the following negative log-likelihood:

$$\mathcal{L}_h(\vartheta_h) = -\sum_{t=1}^T \log \mu_{\vartheta_h}(a_h^t | s_h^t).$$

Fix $\epsilon > 0$. We take a minimal ϵ-net $\widetilde{\Gamma}_\epsilon$ of Γ with respect to the pseudometric ρ, which satisfies $|\widetilde{\Gamma}_\epsilon| = \mathcal{N}(\epsilon, \Gamma, \rho)$. Then, by taking a union bound of Lemma F.5, with probability at least $1 - \delta$, the following inequality holds for all $\widetilde{\vartheta}_h \in \widetilde{\Gamma}_\epsilon$:

$$\begin{aligned}
\frac{1}{2}\left(\mathcal{L}_h(\vartheta_h^*) - \mathcal{L}_h(\widetilde{\vartheta}_h)\right) &\leq \sum_{t=1}^T \log \mathbb{E}\left[\sqrt{\frac{\mu_{\widetilde{\vartheta}_h}(a_h^t|s_h^t)}{\mu_h^*(a_h^t|s_h^t)}}\right] + \log\frac{\mathcal{N}(\epsilon,\Gamma,\rho)}{\delta} \\
&\leq \sum_{t=1}^T \left(\mathbb{E}\left[\sqrt{\frac{\mu_{\widetilde{\vartheta}_h}(a_h^t|s_h^t)}{\mu_h^*(a_h^t|s_h^t)}}\right] - 1\right) + \log\frac{\mathcal{N}(\epsilon,\Gamma,\rho)}{\delta} \\
&= T\mathbb{E}_{s_h\sim\rho_h}\left[\mathbb{E}_{a_h\sim\mu_h^*(\cdot|s_h)}\left[\sqrt{\frac{\mu_{\widetilde{\vartheta}_h}(a_h|s_h)}{\mu_h^*(a_h|s_h)}}\right] - 1\right] + \log\frac{\mathcal{N}(\epsilon,\Gamma,\rho)}{\delta} \\
&= -T\mathbb{E}\left[H^2\left(\mu_{\widetilde{\vartheta}_h}(\cdot|s_h), \mu_h^*(\cdot|s_h)\right)\right] + \log\frac{\mathcal{N}(\epsilon,\Gamma,\rho)}{\delta}.
\end{aligned}$$

By (67), for all $\vartheta_h \in \Gamma$ and all $\widetilde{\vartheta}_h \in \widetilde{\Gamma}_\epsilon$, we have

$$\begin{aligned}
\frac{1}{2}\left|\mathcal{L}_h(\vartheta_h) - \mathcal{L}_h(\widetilde{\vartheta}_h)\right| &= \sum_{t=1}^T \left|\log\sqrt{\mu_{\widetilde{\vartheta}_h}(a_h^t|s_h^t)} - \log\sqrt{\mu_{\vartheta_h}(a_h^t|s_h^t)}\right| \\
&\leq \sum_{t=1}^T \sqrt{m}e^K \left|\sqrt{\mu_{\widetilde{\vartheta}_h}(a_h^t|s_h^t)} - \sqrt{\mu_{\vartheta_h}(a_h^t|s_h^t)}\right| \\
&\leq \sqrt{2m}Te^K\rho(\vartheta_h, \widetilde{\vartheta}_h).
\end{aligned}$$

Therefore, with probability at least $1 - \delta$, it holds for all $\vartheta_h \in \Gamma$ and all $\widetilde{\vartheta}_h \in \widetilde{\Gamma}_\epsilon$ that

$$\mathbb{E}\left[H^2\left(\mu_{\widetilde{\vartheta}_h}(\cdot|s_h), \mu_h^*(\cdot|s_h)\right)\right] \leq \frac{\mathcal{L}_h(\vartheta_h) - \mathcal{L}_h(\vartheta_h^*)}{2T} + \frac{1}{T}\log\frac{\mathcal{N}(\epsilon,\Gamma,\rho)}{\delta} + \sqrt{2m}e^K\rho(\vartheta_h, \widetilde{\vartheta}_h). \tag{71}$$

On the other hand, for all $s_h \in \mathcal{S}$,

$$\begin{aligned}
&\left|H^2\left(\mu_{\widetilde{\vartheta_h}}(\cdot|s_h), \mu_h^*(\cdot|s_h)\right) - H^2\left(\mu_{\vartheta_h}(\cdot|s_h), \mu_h^*(\cdot|s_h)\right)\right| \\
&= \left(H\left(\mu_{\widetilde{\vartheta_h}}(\cdot|s_h), \mu_h^*(\cdot|s_h)\right) + H\left(\mu_{\vartheta_h}(\cdot|s_h), \mu_h^*(\cdot|s_h)\right)\right) \\
&\quad\times\left|H\left(\mu_{\widetilde{\vartheta_h}}(\cdot|s_h), \mu_h^*(\cdot|s_h)\right) - H\left(\mu_{\vartheta_h}(\cdot|s_h), \mu_h^*(\cdot|s_h)\right)\right| \\
&\leq 2H\left(\mu_{\widetilde{\vartheta_h}}(\cdot|s_h), \mu_{\vartheta_h}(\cdot|s_h)\right) \leq 2\rho(\widetilde{\vartheta}_h, \vartheta_h).
\end{aligned} \tag{72}$$

Combining (71) and (72), and take $\widetilde{\vartheta}_h \in \widetilde{\Gamma}_\epsilon$ for each $\vartheta_h \in \Gamma$ such that $\rho(\vartheta_h, \widetilde{\vartheta}_h) \leq \epsilon$, we have

$$\mathbb{E}\left[H^2\left(\mu_{\vartheta_h}(\cdot|s_h), \mu_h^*(\cdot|s_h)\right)\right] \leq \frac{\mathcal{L}_h(\vartheta_h) - \mathcal{L}_h(\vartheta_h^*)}{2T} + \frac{1}{T}\log\frac{\mathcal{N}(\epsilon,\Gamma,\rho)}{\delta} + (\sqrt{2m}e^K + 2)\epsilon, \tag{73}$$

which holds with probability at least $1 - \delta$ for all $\vartheta_h \in \Gamma$.

Bound the Covering Number. Now we are going to bound the covering number $\mathcal{N}(\epsilon, \Gamma, \rho)$. By Lemma 68, we have the following inclusion of balls under different metrics:

$$B_{\|\cdot\|}\left(\vartheta, \frac{\epsilon}{L_K}\right) \subseteq B_\rho(\vartheta, \epsilon).$$

Combining this result with Lemma 43, we have

$$\mathcal{N}\left(\epsilon, \Gamma, \rho\right) \leq \mathcal{N}\left(\frac{\epsilon}{L_K}, \Gamma, \|\cdot\|\right) \leq \left(1 + \frac{2L_K}{\epsilon}\right)^{d_a}. \tag{74}$$

Final bound. We combine (73) and (74), and take $\epsilon = 1/T$. Then with probability at least $1-\delta$, it holds for all $\vartheta_h \in \Gamma$ that

$$\mathbb{E}\left[H^2\left(\mu_{\vartheta_h}(\cdot|s_h), \mu_h^*(\cdot|s_h)\right)\right] \leq \frac{\mathcal{L}_h(\vartheta_h) - \mathcal{L}_h(\vartheta_h^*)}{2T} + \frac{d_a \log(1+2TL_K) + \log \delta^{-1} + \sqrt{2m}e^K + 2}{T}.$$

We take the maximum likelihood estimator:

$$\widehat{\vartheta}_h = \operatorname*{argmin}_{\vartheta_h \in \Gamma} \mathcal{L}_j(\vartheta_h).$$

Since $\mathcal{L}_h(\widehat{\vartheta}_h) \leq \mathcal{L}_h(\vartheta_h^*)$, we have

$$\mathbb{E}\left[H^2\left(\widehat{\mu}_h(\cdot|s_h), \mu_h^*(\cdot|s_h)\right)\right] \leq \frac{d_a \log(1+2TL_K) + \log \delta^{-1} + \sqrt{2m}e^K + 2}{T},$$

where $\widehat{\mu}_h$ is the strategy associated with the estimator $\widehat{\vartheta}_h$. Note that

$$\mathrm{TV}(P, Q) \leq \sqrt{2}H(P, Q)$$

holds for all distributions P and Q, we have

$$\mathbb{E}\left[\mathrm{TV}^2\left(\widehat{\mu}_h(\cdot|s_h), \mu_h^*(\cdot|s_h)\right)\right] \leq \frac{2d_a \log(1+2TL_K) + 2\log \delta^{-1} + 2\sqrt{2m}e^K + 4}{T}. \tag{75}$$

□

C.3. Proof of Theorem C.6

Proof. The proof has three steps.

Bound the estimation error of V-functions. Similar to §B.3.1, we need to bound the error between $\widehat{V}_h$ and V_h in the sense of expectation:

$$\begin{aligned}\mathbb{E}_{s\sim\rho_h}\left[\left|\widehat{V}_h(s) - V_h(s)\right|\right] \leq{}& \underbrace{\mathbb{E}_{s\sim\rho_h}\left[\left|\widehat{\mu}_h(s)^\top \widehat{Q}_h(s)\widehat{\nu}_h(s) - \mu_h^*(s)^\top Q_h(s)\nu_h^*(s)\right|\right]}_{\text{(i)}} \\ &+ \underbrace{\mathbb{E}_{s\sim\rho_h}\left[\frac{1}{\eta}\left|\mathcal{H}(\widehat{\mu}_h(s)) - \mathcal{H}(\mu_h^*(s))\right|\right]}_{\text{(ii)}} + \underbrace{\mathbb{E}_{s\sim\rho_h}\left[\frac{1}{\eta}\left|\mathcal{H}(\widehat{\nu}_h(s)) - \mathcal{H}(\nu_h^*(s))\right|\right]}_{\text{(iii)}}.\end{aligned} \tag{76}$$

Let us bound the three terms above. Note that,

$$\begin{aligned}&\left|\widehat{\mu}_h(s)^\top \widehat{Q}_h(s)\widehat{\nu}_h(s) - \mu_h^*(s)^\top Q_h(s)\nu_h^*(s)\right| \\ &\leq \left|\widehat{\mu}_h(s)^\top(\widehat{Q}_h - Q_h)(s)\widehat{\nu}_h(s)\right| + \left|(\widehat{\mu}_h - \mu_h^*)(s)^\top Q_h(s)\widehat{\nu}_h(s)\right| + \left|\mu_h(s)^\top Q_h(s)(\widehat{\nu}_h - \nu_h^*)(s)\right| \\ &\leq \sup_{a\in\mathcal{A}, b\in\mathcal{B}}\left|\widehat{Q}_h - Q_h\right|(s,a,b) + \|(\widehat{\mu}_h - \mu_h^*)(s)\|_1 \|Q_h(s)\widehat{\nu}_h(s)\|_\infty + \|(\widehat{\nu}_h - \nu_h^*)(s)\|_1 \|Q_h(s)^\top \widehat{\mu}_h(s)\|_\infty \\ &\leq \sup_{a\in\mathcal{A}, b\in\mathcal{B}}\left|\widehat{Q}_h - Q_h\right|(s,a,b) + 2\left(\mathrm{TV}(\mu_h^*(\cdot|s_h), \widehat{\mu}_h(\cdot|s_h)) + \mathrm{TV}(\nu_h^*(\cdot|s_h), \widehat{\nu}_h(\cdot|s_h))\right) \sup_{a\in\mathcal{A}, b\in\mathcal{B}} |Q_h(s,a,b)| \\ &\leq C_h\sqrt{\kappa_h} + 2R \cdot \mathrm{TV}(\mu_h^*(\cdot|s_h), \widehat{\mu}_h(\cdot|s_h)) + 2R \cdot \mathrm{TV}(\nu_h^*(\cdot|s_h), \widehat{\nu}_h(\cdot|s_h)),\end{aligned}$$

where the second and third inequalities follow from Hölder's inequality, and the fourth from boundedness of parameter θ_h. By taking expectation on both sides of the last display, we have

$$\text{(i)} \le C_h\sqrt{\kappa_h} + 4R\epsilon. \tag{77}$$

Again, we use Lemma F.1 to bound the term (ii). Written $\tau = \mathrm{TV}(\mu_h^*(\cdot|s_h), \widehat{\mu}_h(\cdot|s_h))$, we have

$$\begin{aligned}|\mathcal{H}(\mu_h^*(s)) - \mathcal{H}(\widehat{\mu}_h(s))| &\le -\tau\log\tau - (1-\tau)\log(1-\tau) + \tau\log(m-1)\\ &\le \tau\left(1 + \log\frac{m}{\tau}\right).\end{aligned}$$

Similarly, by taking expectation on both sides and using Jensen's inequality,

$$\text{(ii)} \le \epsilon\left(1 + \log\frac{m}{\epsilon}\right).$$

A similar bound holds for term (iii) by replacing m by n. Combining (76), (77) and the entropy bounds, we get

$$\mathbb{E}_{s\sim\rho_h}\Big[|\widehat{V}_h(s) - V_h(s)|\Big] \le C_h\sqrt{\kappa_h} + \epsilon\left(4R + 2 + \log\frac{mn}{\epsilon^2}\right).$$

Furthermore, for all $h \in [H]$ and all $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$, since $\|\phi(\cdot,\cdot,\cdot)\| \le 1$, we have

$$\begin{aligned}&\mathbb{E}_{s\sim\rho_h}\Big[|\mathbb{P}_h\widehat{V}_{h+1} - \mathbb{P}_h V_{h+1}|(s,a,b)\Big] \le \|\phi(s,a,b)\|\,\|\Pi_h\|_{\mathrm{op}}\,\mathbb{E}_{s\sim\rho_h}\Big[\|\widehat{V}_h - V_h\|\Big]\\ &\quad\le \|\Pi_h\|_{\mathrm{op}}\left(C_h\sqrt{\kappa_h} + \epsilon\left(4R + 2 + \log\frac{mn}{\epsilon^2}\right)\right)\\ &\quad\lesssim \frac{1}{\sqrt{T}}\left(m^{7/4} + n^{7/4} + (m^{3/2} + n^{3/2} + \log T)\sqrt{(d_a + d_b)\log T + \log(H/\delta)} + \sqrt{S}\log(mn)\right).\end{aligned} \tag{78}$$

Bound the transition error. In §B.3.2, we bound the error between $\widehat{\mathbb{P}}_h\widehat{V}_{h+1}$ and $\mathbb{P}_h\widehat{V}_{h+1}$ over all $V_{h+1} \in \widehat{\mathcal{V}}_{h+1}$. This bound does not depend on the total variation distance between the true and the estimated policies. Therefore, the bound (56) still holds in this situation for all $s \in \mathcal{S}, a \in \mathcal{A}$ and $b \in \mathcal{B}$. Therefore, under a certain concentration event $\mathcal{E}_t$ with probability at least $1-\delta$,

$$\mathbb{E}_{s\sim\rho_h}\Big[\big|\widehat{\mathbb{P}}_h\widehat{V}_{h+1} - \mathbb{P}_h\widehat{V}_{h+1}\big|(s,a,b)\Big] \lesssim \frac{\sqrt{Sd} + \sqrt{d\log T} + \sqrt{\log(H/\delta)}}{\sqrt{T}}\log(mn). \tag{79}$$

Bound the L^1-Hausdorff distance between sets. According to Lemma C.5, for any $\widehat{\theta}_h \in \widehat{\Theta}_h$, we can find a feasible $\theta_h \in \Theta_h$ such that $\|\widehat{\theta}_h - \theta_h\| \le C_h\sqrt{\kappa_h}$ for some constant C_h. Moreover, for any $\widehat{V}_{h+1} \in \widehat{\mathcal{V}}_{h+1}$, we can also find a feasible V-function $V_{h+1} \in \mathcal{V}_{h+1}$ such that (78) holds for all possible choice of (a,b). Consider the reward functions $\widehat{r}_h$ and r_h generated by $(\widehat{\theta}_h, \widehat{V}_{h+1})$ and (θ_h, V_{h+1}), respectively:

$$\begin{aligned}&\mathbb{E}_{s\sim\rho_h}\left[|\widehat{r}_h - r_h|(s,a,b)\right] = \mathbb{E}_{s\sim\rho_h}\Big[\Big|\phi(s,a,b)^\top(\widehat{\theta}_h - \theta_h) - \gamma(\widehat{\mathbb{P}}_h\widehat{V}_{h+1} - \mathbb{P}_h V_{h+1})(s,a,b)\Big|\Big]\\ &\quad\le \mathbb{E}_{s\sim\rho_h}\Big[\underbrace{\|\phi(s,a,b)\|\,\|\widehat{\theta}_h - \theta_h\|}_{\le R\sqrt{\kappa_h}} + \gamma\underbrace{\big|\widehat{\mathbb{P}}_h\widehat{V}_{h+1} - \mathbb{P}_h\widehat{V}_{h+1}\big|(s,a,b)}_{\text{bounded by (78)}} + \gamma\underbrace{\big|\mathbb{P}_h\widehat{V}_{h+1} - \mathbb{P}_h V_{h+1}\big|(s,a,b)}_{\text{bounded by (79)}}\Big]\\ &\quad\lesssim \frac{1}{\sqrt{T}}\left(m^{7/4} + n^{7/4} + (m^{3/2} + n^{3/2} + \log T)\sqrt{(d_a + d_b + d)\log T + \log(H/\delta)} + \sqrt{Sd}\log(mn)\right).\end{aligned}$$

Since this bound is valid for any choice of $(\widehat{\theta}_h)$, we have

$$\begin{aligned}\sup_{\widehat{r}\in\widehat{\mathcal{R}}} D_1(\widehat{r}, \mathcal{R}) \lesssim \frac{1}{\sqrt{T}}\Bigg(&m^{7/4} + n^{7/4}\\ &+ (m^{3/2} + n^{3/2} + \log T)\sqrt{(d_a + d_b + d)\log T + \log(H/\delta)} + \sqrt{Sd}\log(mn)\Bigg).\end{aligned} \tag{80}$$

On the other hand, for any $r \in \mathcal{R}$ generated by $\theta_h \in \Theta_h \subseteq \widehat{\Theta}_h$ and $\theta_{h+1} \in \Theta_{h+1} \subseteq \widehat{\Theta}_{h+1}$, we pick the following estimated reward $\widehat{r} = (\widehat{r}_1, \widehat{r}_2, \cdots, \widehat{r}_H) \in \widehat{\mathcal{R}}$:

$$\widetilde{V}_{h+1}(s) = \widehat{\mu}_{h+1}(s)^\top Q_{h+1}(s)\widehat{\nu}_{h+1}(s) + \eta^{-1}\mathcal{H}(\widehat{\nu}_{h+1}(s)) - \eta^{-1}\mathcal{H}(\widehat{\nu}_{h+1}(s)),$$
$$\widehat{r}_h(s,a,b) = \phi(s,a,b)^\top \theta_h - \widehat{\mathbb{P}}_h \widetilde{V}_{h+1}(s,a,b), \quad h \in [H], s \in \mathcal{S}, a \in \mathcal{A}, b \in \mathcal{B}.$$

Since $\Theta_{h+1} \subseteq \widehat{\Theta}_{h+1}$, the estimated V-function $\widetilde{V}_{h+1} \in \widehat{\mathcal{V}}_{h+1}$, and the bound (79) still holds if we replacing $\widehat{V}_{h+1}$ by $\widetilde{V}_{h+1}$. Furthermore, similar to our estimation procedure (76)-(78), the estimation error of V-function has the following bound:

$$\begin{aligned}\mathbb{E}_{s\sim\rho_h}\left[|\mathbb{P}_h\widetilde{V}_{h+1} - \mathbb{P}_h V_{h+1}|(s,a,b)\right] &\le \|\Pi_h\|_{\text{op}}\left(C_h\sqrt{\kappa_h} + \epsilon\left(4R + 2 + \log\frac{mn}{\epsilon^2}\right)\right), \\ &\lesssim \frac{1}{\sqrt{T}}\left(m^{7/4} + n^{7/4} + (m^{3/2} + n^{3/2} + \log T)\sqrt{(d_a + d_b)\log T + \log(H/\delta)} + \sqrt{S}\log(mn)\right). \end{aligned} \tag{81}$$

Consequently, the following inequality simultaneously holds for all choice of h, r_h and (a,b):

$$\begin{aligned}\mathbb{E}_{s\sim\rho_h}\left[|\widehat{r}_h - r_h|(s,a,b)\right] &= \mathbb{E}_{s\sim\rho_h}\left[|\widehat{\mathbb{P}}_h\widetilde{V}_{h+1} - \mathbb{P}_h V_{h+1}|(s,a,b)\right] \\ &\le \underbrace{\mathbb{E}_{s\sim\rho_h}\left[|\widehat{\mathbb{P}}_h\widetilde{V}_{h+1} - \mathbb{P}_h \widetilde{V}_{h+1}|(s,a,b)\right]}_{\text{bounded by (79)}} + \underbrace{\mathbb{E}_{s\sim\rho_h}\left[|\mathbb{P}_h\widetilde{V}_{h+1} - \mathbb{P}_h V_{h+1}|(s,a,b)\right]}_{\text{bounded by (81)}} \\ &\lesssim \frac{1}{\sqrt{T}}\left(m^{7/4} + n^{7/4} + (m^{3/2} + n^{3/2} + \log T)\sqrt{(d_a + d_b + d)\log T + \log(H/\delta)} + \sqrt{Sd}\log(mn)\right).\end{aligned}$$

Therefore we have

$$\begin{aligned}\sup_{r\in\mathcal{R}} D_1(r,\widehat{\mathcal{R}}) \lesssim \frac{1}{\sqrt{T}}\Bigg(&m^{7/4} + n^{7/4} \\ &+ (m^{3/2} + n^{3/2} + \log T)\sqrt{(d_a + d_b + d)\log T + \log(H/\delta)} + \sqrt{Sd}\log(mn)\Bigg).\end{aligned} \tag{82}$$

Combining (80) and (82), we obtain the Hausdorff distance between $\mathcal{R}$ and $\widehat{\mathcal{R}}$:

$$\begin{aligned} D_1(\mathcal{R},\widehat{\mathcal{R}}) &= \max\left\{\sup_{\widehat{r}\in\widehat{\mathcal{R}}} D_1(\widehat{r},\mathcal{R}),\ \sup_{r\in\mathcal{R}} D_1(r,\widehat{\mathcal{R}})\right\} \\ &\lesssim \frac{1}{\sqrt{T}}\left(m^{7/4} + n^{7/4} + (m^{3/2} + n^{3/2} + \log T)\sqrt{(d_a + d_b + d)\log T + \log(H/\delta)} + \sqrt{Sd}\log(mn)\right).\end{aligned}$$

□

D. Proof of Auxiliary Lemmas

D.1. Proof of Lemma A.2

Proof. To begin with, we prove that $\Theta \subseteq \widehat{\Theta}$. We will use the following notation for convenience:

$$X := \begin{bmatrix} A(\nu^*) \\ B(\mu^*) \end{bmatrix}, y := \begin{bmatrix} c(\mu^*) \\ d(\nu^*) \end{bmatrix} \quad \text{and} \quad \widehat{X} := \begin{bmatrix} A(\widehat{\nu}) \\ B(\widehat{\mu}) \end{bmatrix}, \widehat{y} := \begin{bmatrix} c(\widehat{\mu}) \\ d(\widehat{\nu}) \end{bmatrix}.$$

For any $\theta \in \Theta$, we have $X\theta = y$ by the definition of Θ. Plugging in this equation, we have

$$\|\widehat{X}\theta - \widehat{y}\|^2 = \|(\widehat{X} - X)\theta - (\widehat{y} - y)\|^2 \le 2(\|\widehat{X} - X\|_{\text{op}}^2\|\theta\|^2 + \|\widehat{y} - y\|^2). \tag{83}$$

Next, we bound the operator norm $\|\widehat{X} - X\|_{\text{op}}^2$:

$$\begin{aligned}\|\widehat{X} - X\|_{\text{op}}^2 &= \|(\widehat{X} - X)^\top(\widehat{X} - X)\|_{\text{op}} \\ &= \|(A(\widehat{\nu}) - A(\nu^*))^\top(A(\widehat{\nu}) - A(\nu^*)) + (B(\widehat{\mu}) - B(\mu^*))^\top(B(\widehat{\mu}) - B(\mu^*))\|_{\text{op}} \\ &\le \|(A(\widehat{\nu}) - A(\nu^*))^\top(A(\widehat{\nu}) - A(\nu^*))\|_{\text{op}} + \|(B(\widehat{\mu}) - B(\mu^*))^\top(B(\widehat{\mu}) - B(\mu^*))\|_{\text{op}}.\end{aligned}$$

Using (18), we have

$$\|\widehat{X} - X\|_{\text{op}}^2 \le \|A(\widehat{\nu}) - A(\nu^*)\|_{\text{op}}^2 + \|B(\widehat{\mu}) - B(\mu^*)\|_{\text{op}}^2 \le \|\Phi_1\|_{\text{op}}^2 \cdot \epsilon_2^2 + \|\Phi_2\|_{\text{op}}^2 \cdot \epsilon_1^2, \tag{84}$$

and

$$\|\widehat{y} - y\|^2 = \|c(\widehat{\nu}) - c(\nu^*)\|^2 + \|d(\widehat{\mu}) - d(\mu^*)\|^2 \le \frac{m\epsilon_1^2}{\eta^2(\min_{i\in[m]} \mu_i - \epsilon_1)^2} + \frac{n\epsilon_2^2}{\eta^2(\min_{j\in[n]} \nu_j - \epsilon_2)^2}. \tag{85}$$

Plugging in (84) and (85) to (83), we obtain that

$$\|\widehat{X}\theta - \widehat{y}\|^2 \le \kappa, \tag{86}$$

where

$$\kappa := 2\left(M\|\Phi_1\|_{\text{op}}^2 + \frac{n}{\eta^2(\min_{j\in[n]} \nu_j - \epsilon_2)^2}\right)\epsilon_2^2 + 2\left(M\|\Phi_2\|_{\text{op}}^2 + \frac{m}{\eta^2(\min_{i\in[m]} \mu_i - \epsilon_1)^2}\right)\epsilon_1^2.$$

Consequently, we have $\Theta \subseteq \widehat{\Theta}$, and $\inf_{\widehat{\theta}\in\widehat{\Theta}} \|\widehat{\theta} - \theta\|^2 = 0$ for any $\theta \in \Theta$. Therefore, the Hausdorff distance $d_H(\Theta, \widehat{\Theta})$ would be

$$d_H(\Theta, \widehat{\Theta}) = \sup_{\widehat{\theta}\in\widehat{\Theta}} \inf_{\theta\in\Theta} \|\widehat{\theta} - \theta\|.$$

Intuitively, $\inf_{\theta\in\Theta} \|\widehat{\theta} - \theta\|^2$ is the distance of $\widehat{\theta}$ to its projection onto the candidate set Θ. For any $\widehat{\theta} \in \widehat{\Theta}$, the projection of $\widehat{\theta}$ onto the affine subspace $S_{X,y} = \{\theta \in \mathbb{R}^d : X\theta = y\}$ is given by

$$\widetilde{\theta} = X^\dagger y + (I - X^\dagger X)\widehat{\theta}.$$

Since $\Theta = \{\theta : X\theta = y, \|\theta\|^2 \le M\}$ is the ball of radius $\sqrt{M - \|X^\dagger y\|^2}$ in $S_{X,y}$ centered at $X^\dagger y$, we decompose the distance from $\widehat{\theta} \in \widehat{\Theta}$ to its projection θ^* onto Θ as follows:

$$\begin{aligned}
\|\widehat{\theta} - \theta^*\|^2 &= \|\widehat{\theta} - \widetilde{\theta}\|^2 + \|\widetilde{\theta} - \theta^*\|^2 && (87)\\
&= \underbrace{\|X^\dagger(X\widehat{\theta} - y)\|^2}_{\text{(i)}} + \underbrace{\left[\|(I - X^\dagger X)\widehat{\theta}\| - \sqrt{M - \|X^\dagger y\|^2}\right]_+^2}_{\text{(ii)}}, && (88)
\end{aligned}$$

where $[x]_+ = \max\{x, 0\}$. By the triangle inequality,

$$\begin{aligned}
\|X^\dagger(X\widehat{\theta} - y)\| &\le \|X^\dagger\|_{\text{op}} \|X\widehat{\theta} - y\| \\
&\le \|X^\dagger\|_{\text{op}} \left(\|\widehat{X}\widehat{\theta} - \widehat{y}\| + \|X - \widehat{X}\|_{\text{op}}\|\widehat{\theta}\| + \|\widehat{y} - y\|\right) \\
&\le \|X^\dagger\|_{\text{op}} \|\widehat{X}\widehat{\theta} - \widehat{y}\| + \|X^\dagger\|_{\text{op}} \sqrt{2\|X - \widehat{X}\|_{\text{op}}^2 \|\widehat{\theta}\|^2 + 2\|\widehat{y} - y\|^2}.
\end{aligned} \tag{89}$$

Since $\widehat{\theta} \in \widehat{\Theta}$, we have $\|\widehat{X}\widehat{\theta} - \widehat{y}\| \le \sqrt{\kappa}$ and $\|\widehat{\theta}\|^2 \le M$. Plugging in the estimates (84) and (85) to (89), we obtain

$$\|X^\dagger(X\widehat{\theta} - y)\| \le \|X^\dagger\|_{\text{op}}\sqrt{\kappa} + \|X^\dagger\|_{\text{op}}\sqrt{2M\|X - \widehat{X}\|_{\text{op}}^2 + 2\|\widehat{y} - y\|^2} \le 2\sqrt{\kappa} \cdot \|X^\dagger\|_{\text{op}}. \tag{90}$$

Hence

$$\text{(i)} = \|X^\dagger(X\widehat{\theta} - y)\|^2 \le 4\kappa \cdot \|X^\dagger\|_{\text{op}}. \tag{91}$$

Applying the triangle inequality to (90), we also have

$$\|X^\dagger X\widehat{\theta}\| \ge \|X^\dagger y\| - 2\sqrt{\kappa} \cdot \|X^\dagger\|_{\text{op}}$$

Using the orthogonal decomposition $\widehat{\theta} = X^\dagger X \widehat{\theta} + (I - X^\dagger X)\widehat{\theta}$, we have

$$\|(I - X^\dagger X)\widehat{\theta}\|^2 = \|\widehat{\theta}\|^2 - \|X^\dagger X \widehat{\theta}\|^2 \leq M - \left(\|X^\dagger y\| - 2\sqrt{\kappa} \cdot \|X^\dagger\|_{\mathrm{op}}\right)^2 .$$

Therefore,

$$\begin{aligned}
\text{(ii)} &\leq \left[\sqrt{M - \left(\|X^\dagger y\| - 2\sqrt{\kappa} \cdot \|X^\dagger\|_{\mathrm{op}}\right)^2} - \sqrt{M - \|X^\dagger y\|^2}\right]^2 \\
&\leq \left[\sqrt{M - \|X^\dagger y\|^2 + 4\sqrt{\kappa} \cdot \|X^\dagger y\| \|X^\dagger\|_{\mathrm{op}}} - \sqrt{M - \|X^\dagger y\|^2}\right]^2 \\
&\leq \frac{8\kappa \cdot \|X^\dagger y\|^2 \|X^\dagger\|_{\mathrm{op}}^2}{\sqrt{M - \|X^\dagger y\|}},
\end{aligned} \tag{92}$$

where we use the inequality $\sqrt{a+b} - \sqrt{a} \leq \frac{b}{2\sqrt{a}}$ in the last inequality. Combining (88), (91) and (92), we have

$$\|\widehat{\theta} - \theta^*\|^2 \leq 4\kappa \cdot \left(\|X^\dagger\|_{\mathrm{op}} + \frac{2\|X^\dagger y\|^2 \|X^\dagger\|_{\mathrm{op}}^2}{\sqrt{M - \|X^\dagger y\|}}\right).$$

Therefore we obtain the upper bound of the Hausdoff distance:

$$d_H(\Theta, \widehat{\Theta}) = \sup_{\widehat{\theta} \in \widehat{\Theta}} \inf_{\theta \in \Theta} \|\widehat{\theta} - \theta\| \leq \mathcal{O}(\sqrt{\kappa}),$$

where the notation $\mathcal{O}(\cdot)$ absorbs a constant depending only on Φ_1, Φ_2, X, y and M. Thus we conclude the whole proof. □

D.2. Proof of Lemma B.1

Proof. Akin to the proof of Lemma A.2, we use the notation

$$X_h := \begin{bmatrix} A_h(\nu_h^*) \\ B_h(\mu_h^*) \end{bmatrix}, y_h := \begin{bmatrix} c_h(\mu_h^*) \\ d_h(\nu_h^*) \end{bmatrix} \quad \text{and} \quad \widehat{X}_h := \begin{bmatrix} A_h(\widehat{\nu}_h) \\ B_h(\widehat{\mu}_h) \end{bmatrix}, \widehat{y}_h := \begin{bmatrix} c_h(\widehat{\mu}_h) \\ d_h(\widehat{\nu}_h) \end{bmatrix}.$$

Then we have

$$\begin{aligned}
\|\widehat{X}_h - X_h\|_{\mathrm{op}}^2 &= \|(\widehat{X}_h - X_h)^\top (\widehat{X}_h - X_h)\|_{\mathrm{op}} \\
&= \Bigg\| \sum_{s \in \mathcal{S}} (A_h(s, \widehat{\nu}_h) - A_h(s, \nu_h^*))^\top (A_h(s, \widehat{\nu}_h) - A_h(s, \nu_h^*)) \\
&\qquad + \sum_{s \in \mathcal{S}} (B_h(s, \widehat{\mu}_h) - B_h(s, \mu_h^*))^\top (B_h(s, \widehat{\mu}_h^*) - B_h(s, \mu_h^*)) \Bigg\|_{\mathrm{op}} \\
&\leq \sum_{s \in \mathcal{S}} \left\|(A_h(s, \widehat{\nu}_h) - A_h(s, \nu_h^*))^\top (A_h(s, \widehat{\nu}_h) - A_h(s, \nu_h^*))\right\|_{\mathrm{op}} \\
&\qquad + \sum_{s \in \mathcal{S}} \left\|(B_h(s, \widehat{\mu}_h) - B_h(s, \mu_h^*))^\top (B_h(s, \widehat{\mu}_h) - B_h(s, \mu_h^*))\right\|_{\mathrm{op}} \\
&\leq \sum_{s \in \mathcal{S}} \|A_h(s, \widehat{\nu}_h) - A_h(s, \nu_h^*)\|_{\mathrm{op}}^2 + \sum_{s \in \mathcal{S}} \|B_h(s, \widehat{\mu}_h) - B_h(s, \mu_h^*)\|_{\mathrm{op}}^2 \\
&\leq \sum_{s \in \mathcal{S}} \|\Phi_{1,s}\|_{\mathrm{op}}^2 \epsilon_1^2 + \sum_{s \in \mathcal{S}} \|\Phi_{2,s}\|_{\mathrm{op}}^2 \epsilon_2^2 \\
&= \|\Phi_1\|_{\mathrm{op}}^2 \epsilon_1^2 + \|\Phi_2\|_{\mathrm{op}}^2 \epsilon_2^2,
\end{aligned}$$

where $\Phi_{1,s} \in \mathbb{R}^{d \times (m-1)n}$ and $\Phi_{2,s} \in \mathbb{R}^{d \times (n-1)m}$ are defined as

$$\Phi_{1,s} := \left(\phi(s, a, \cdot) - \phi(s, 1, \cdot)\right)_{a \in [m] \backslash \{1\}}, \quad \Phi_{2,s} := \left(\phi(s, \cdot, b) - \phi(s, \cdot, 1)\right)_{b \in [n] \backslash \{1\}},$$

and we also define

$$\Phi_1 = \begin{bmatrix}\Phi_{1,1} & \Phi_{1,2} & \cdots & \Phi_{1,S}\end{bmatrix}, \quad \Phi_2 = \begin{bmatrix}\Phi_{2,1} & \Phi_{2,2} & \cdots & \Phi_{2,S}\end{bmatrix},$$

Furthermore,

$$\begin{aligned}\|\widehat{y}_h - y_h\|^2 &= \sum_{s\in\mathcal{S}} \|c_h(s,\widehat{\nu}_h) - c_h(s,\nu_h^*)\|^2 + \sum_{s\in\mathcal{S}} \|d_h(s,\widehat{\mu}_h) - d_h(s,\mu_h^*)\|^2 \\ &\le \frac{Sm\epsilon_1^2}{\eta^2\left(\min_{s\in\mathcal{S},a\in[m]}\mu_h^*(a|s)-\epsilon_1\right)^2} + \frac{Sn\epsilon_2^2}{\eta^2\left(\min_{s\in\mathcal{S},b\in[n]}\nu_h^*(b|s)-\epsilon_2\right)^2}.\end{aligned}$$

Consequently, for any $\theta_h \in \Theta_h$, we have

$$\begin{aligned}\|\widehat{X}_h\theta_h - \widehat{y}_h\|^2 &= \|(\widehat{X}_h - X_h)\theta_h - (\widehat{y}_h - y_h)\|^2 \\ &\le 2\left(\|\widehat{X}_h - X_h\|_{\text{op}}^2\|\theta_h\|^2 + \|\widehat{y}_h - y_h\|^2\right) \\ &\le \kappa_h,\end{aligned}$$

where

$$\begin{aligned}\kappa_h = &\,2\left(R^2\|\Phi_1\|_{\text{op}}^2 + \frac{Sm}{\eta^2\left(\min_{s\in\mathcal{S},a\in[m]}\mu_h^*(a|s)-\epsilon_1\right)^2}\right)\epsilon_1^2 \\ &+ 2\left(R^2\|\Phi_2\|_{\text{op}}^2 + \frac{Sn}{\eta^2\left(\min_{s\in\mathcal{S},b\in[n]}\nu_h^*(b|s)-\epsilon_2\right)^2}\right)\epsilon_2^2.\end{aligned}$$

Therefore $\Theta_h \subseteq \widehat{\Theta}_h$, and the Hausdorff distance is

$$d_H(\Theta_h, \widehat{\Theta}_h) = \sup_{\widehat{\theta}_h\in\widehat{\Theta}_h}\inf_{\theta_h\in\Theta_h}\|\widehat{\theta}_h - \theta_h\|.$$

Analogous to the proof of Lemma A.2, for each $h \in [H]$, there exists a constant C_h depending on Φ_1, Φ_2, X_h, y_h and d such that

$$d_H(\Theta_h, \widehat{\Theta}_h) \le C_h\sqrt{\kappa_h}.$$

□

D.3. Proof of Lemma B.2

Proof. Similar to (30), given any $h \in [H]$ and $s \in \mathcal{S}$, we have the following McDiarmid's bound for all $\epsilon > 0$:

$$\begin{aligned}\mathbb{P}\left(\text{TV}(\widehat{\mu}_h(\cdot|s),\mu_h^*(\cdot|s)) \le \frac{1}{2}\sqrt{\frac{m}{N_h(s)}} + \epsilon \,\middle|\, N_h(s)\right) \ge 1 - e^{-2N_h(s)\epsilon^2}, \\ \mathbb{P}\left(\text{TV}(\widehat{\nu}_h(\cdot|s),\nu_h^*(\cdot|s)) \le \frac{1}{2}\sqrt{\frac{n}{N_h(s)}} + \epsilon \,\middle|\, N_h(s)\right) \ge 1 - e^{-2N_h(s)\epsilon^2}.\end{aligned}$$

Let $\lambda > 0$, and define $\mathcal{H}_\lambda$ to be the event that $N_h(s) \ge \lambda$ for all $h \in [H]$ and $s \in \mathcal{S}$. Then under event $\mathcal{H}_\lambda$, one can derive the union bound

$$\begin{aligned}&\mathbb{P}\left(\text{TV}(\widehat{\mu}_h(\cdot|s),\mu_h^*(\cdot|s)) \le \sqrt{\frac{m}{4\lambda}} + \epsilon,\ \text{TV}(\widehat{\nu}_h(\cdot|s),\nu_h^*(\cdot|s)) \le \sqrt{\frac{n}{4\lambda}} + \epsilon,\ \forall s\in\mathcal{S},\forall h\in[H] \,\middle|\, \mathcal{H}_\lambda\right) \\ &\ge 1 - 2HSe^{-2\lambda\epsilon^2}.\end{aligned} \tag{93}$$

We let $\lambda(\epsilon) = \frac{1}{2\epsilon^2}\log\frac{4HS}{\delta} \ge \frac{1}{2\epsilon^2}$. Since $m, n \ge 2$, (93) becomes

$$\begin{aligned}&\mathbb{P}\left(\text{TV}(\widehat{\mu}_h(\cdot|s),\mu_h^*(\cdot|s)) \le \sqrt{2m}\epsilon,\ \text{TV}(\widehat{\nu}_h(\cdot|s),\nu_h^*(\cdot|s)) \le \sqrt{2n}\epsilon,\ \forall s\in\mathcal{S},\forall h\in[H] \,\middle|\, \mathcal{H}_{\lambda(\epsilon)}\right) \\ &\ge 1 - \frac{\delta}{2}.\end{aligned} \tag{94}$$

For sufficiently large sample size $T > 0$, the Hoeffding's inequality implies

$$\begin{aligned}\mathbb{P}\left(N_h(s) \leq \lambda\right) &= \mathbb{P}\left(\frac{1}{T}N_h(s) - d_h^{\mu^*,\nu^*}(s) \leq \frac{\lambda}{T} - d_h^{\mu^*,\nu^*}(s)\right) \\ &\leq \mathbb{P}\left(\frac{1}{T}\left(N_h(s) - \mathbb{E}[N_h(s)]\right) \leq \frac{\lambda}{T} - C\right) \leq \exp\left(-\frac{2}{T}\left(CT - \lambda\right)^2\right).\end{aligned}$$

And again one can derive the union bound

$$\mathbb{P}(\mathcal{H}_\lambda) \geq 1 - HS\exp\left(-\frac{2}{T}\left(CT - \lambda\right)^2\right) \geq 1 - HS\exp\left(-2C^2T + C\lambda\right).$$

We replace ϵ by $\frac{\epsilon}{\sqrt{2(m\vee n)}}$ in (94), and fix $\lambda = \frac{m\vee n}{\epsilon^2}\log\frac{4HS}{\delta}$. We set

$$T = \frac{1}{2C^2}\log\frac{2HS}{\delta} + \frac{\lambda}{2C} = \frac{1}{2C^2}\log\frac{2HS}{\delta} + \frac{m\vee n}{2C\epsilon^2}\log\frac{4HS}{\delta}. \tag{95}$$

Consequently, we have $\mathbb{P}(\mathcal{H}_\lambda) \geq 1 - \delta/2$, and

$$\mathbb{P}\left(\mathcal{E}_1\right) = \mathbb{P}\left(\mathcal{E}_1 \,|\, \mathcal{H}_\lambda\right)\mathbb{P}(\mathcal{H}_\lambda) \geq 1 - \delta.$$

Therefore (95) concludes the proof. □

D.4. Proof of Lemma C.5

Notations. We use the following notations for simplicity:

$$A_h(\widehat{\nu}_h) = \begin{bmatrix}\sqrt{\rho_h(1)}A_h(1,\widehat{\nu}_h) \\ \vdots \\ \sqrt{\rho_h(S)}A_h(S,\widehat{\nu}_h)\end{bmatrix}, \quad B_h(\widehat{\nu}_h) = \begin{bmatrix}\sqrt{\rho_h(1)}B_h(1,\widehat{\mu}_h) \\ \vdots \\ \sqrt{\rho_h(S)}B_h(S,\widehat{\mu}_h)\end{bmatrix},$$
$$c_h(\widehat{\mu}_h) = \begin{bmatrix}\sqrt{\rho_h(1)}\,c_h(1,\widehat{\nu}_h) \\ \vdots \\ \sqrt{\rho_h(S)}\,c_h(S,\widehat{\nu}_h)\end{bmatrix}, \quad d_h(\widehat{\nu}_h) = \begin{bmatrix}\sqrt{\rho_h(1)}\,d_h(1,\widehat{\mu}_h) \\ \vdots \\ \sqrt{\rho_h(S)}\,d_h(S,\widehat{\mu}_h)\end{bmatrix},$$

and

$$X_h = \begin{bmatrix}A_h(\nu_h^*) \\ B_h(\mu_h^*)\end{bmatrix}, \quad \tilde{X}_h = \begin{bmatrix}A_h(\widehat{\nu}_h) \\ B_h(\widehat{\mu}_h)\end{bmatrix}, \quad \widehat{X}_h = \begin{bmatrix}\widehat{A}_h(\widehat{\nu}_h) \\ \widehat{B}_h(\widehat{\mu}_h)\end{bmatrix},$$
$$y_h = \begin{bmatrix}c_h(\mu_h^*) \\ d_h(\nu_h^*)\end{bmatrix}, \quad \tilde{y}_h = \begin{bmatrix}c_h(\widehat{\mu}_h) \\ d_h(\widehat{\nu}_h)\end{bmatrix}, \quad \widehat{y}_h = \begin{bmatrix}\widehat{c}_h(\widehat{\mu}_h) \\ \widehat{d}_h(\widehat{\nu}_h)\end{bmatrix}.$$

where ρ_h is the visitation measure of the agents at step h using the QRE policies μ^* and ν^*, and $\mathcal{D}_h$ is the empirical distribution at step h associated with the dataset. The confidence set is

$$\widehat{\Theta}_h := \left\{\theta : \left\|\widehat{X}_h\theta - \widehat{y}_h\right\|^2 \leq \kappa_h, \|\theta\| \leq R\right\}.$$

We focus on finding an appropriate threshold $\kappa_h > 0$.

Proof. The formal proof has three main steps.

Bound the X-matrix error. We are going to bound the error of the matrix $\widehat{X}_h$ in the least square problem. First note that

$$\begin{aligned}
&\|\widetilde{X}_h - X_h\|_{\text{op}}^2 = \left\|(\widetilde{X}_h - X_h)^\top(\widetilde{X}_h - X_h)\right\|_{\text{op}} \\
&\le \left\|(A_h(\widehat{\nu}_h) - A_h(\nu_h))^\top(A_h(\widehat{\nu}_h) - A_h(\nu_h))\right\|_{\text{op}} + \left\|(B_h(\widehat{\mu}_h) - B_h(\mu_h))^\top(B_h(\widehat{\mu}_h) - B_h(\mu_h))\right\|_{\text{op}} \\
&= \sum_{s\in\mathcal{S}} \rho_h(s_h)\,\|A_h(s_h,\widehat{\nu}_h) - A_h(s_h,\nu_h)\|_{\text{op}}^2 + \sum_{s\in\mathcal{S}} \rho_h(s_h)\,\|B_h(s_h,\widehat{\mu}_h) - B_h(s_h,\mu_h)\|_{\text{op}}^2 \\
&= \mathbb{E}_{s_h\sim\rho_h}\left[\|A_h(s_h,\widehat{\nu}_h) - A_h(s_h,\nu_h)\|_{\text{op}}^2\right] + \mathbb{E}_{s_h\sim\rho_h}\left[\|B_h(s_h,\widehat{\mu}_h) - B_h(s_h,\mu_h)\|_{\text{op}}^2\right].
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}_{s_h\sim\rho_h}\left[\|A_h(s_h,\widehat{\nu}_h) - A_h(s_h,\nu_h)\|_{\text{op}}^2\right] &= \mathbb{E}_{s_h\sim\rho_h}\left[\|\Phi_{1,s_h}\left(I_{m-1}\otimes\widehat{\nu}_h(\cdot|s_h) - I_{m-1}\otimes\nu_h(\cdot|s_h)\right)\|_{\text{op}}^2\right] \\
&\le \max_{s\in\mathcal{S}}\|\Phi_{1,s}\|_{\text{op}}^2\,\mathbb{E}_{s_h\sim\rho_h}\left[\|\nu_h(\cdot|s_h) - \widehat{\nu}_h(\cdot|s_h)\|^2\right] \\
&\le 4\max_{s\in\mathcal{S}}\|\Phi_{1,s}\|_{\text{op}}^2\,\mathbb{E}_{s_h\sim\rho_h}\left[\text{TV}^2(\nu_h(\cdot|s_h),\widehat{\nu}_h(\cdot|s_h))\right],
\end{aligned}$$

and similarly,

$$\mathbb{E}_{s_h\sim\rho_h}\left[\|B_h(s_h,\widehat{\mu}_h) - B_h(s_h,\mu_h)\|_{\text{op}}\right] \le 4\max_{s\in\mathcal{S}}\|\Phi_{2,s}\|_{\text{op}}^2\,\mathbb{E}_{s_h\sim\rho_h}\left[\text{TV}^2(\mu_h(\cdot|s_h),\widehat{\mu}_h(\cdot|s_h))\right].$$

Since $\mathbb{E}_{s_h\sim\rho_h}\left[\text{TV}^2(\nu_h(\cdot|s_h),\widehat{\nu}_h(\cdot|s_h))\right] \le \epsilon^2$ and $\mathbb{E}_{s_h\sim\rho_h}\left[\text{TV}^2(\mu_h(\cdot|s_h),\widehat{\mu}_h(\cdot|s_h))\right] \le \epsilon^2$, we have

$$\|X_h - \widetilde{X}_h\|_{\text{op}} \le 2\epsilon\sqrt{\max_{s\in\mathcal{S}}\|\Phi_{1,s}\|_{\text{op}}^2 + \max_{s\in\mathcal{S}}\|\Phi_{2,s}\|_{\text{op}}^2}. \tag{96}$$

Now we analyze the error from approximating the visitation measure by the frequency estimator. Note that

$$\begin{aligned}
&\|\widehat{X}_h - \widetilde{X}_h\|_{\text{op}}^2 = \left\|(\widehat{X}_h - \widetilde{X}_h)^\top(\widehat{X}_h - \widetilde{X}_h)\right\|_{\text{op}} \\
&\le \left\|(\widehat{A}_h(\widehat{\nu}_h) - A_h(\widehat{\nu}_h))^\top(\widehat{A}_h(\widehat{\nu}_h) - A_h(\widehat{\nu}_h))\right\|_{\text{op}} + \left\|(\widehat{B}_h(\widehat{\mu}_h) - B_h(\widehat{\mu}_h))^\top(\widehat{B}_h(\widehat{\mu}_h) - B_h(\widehat{\mu}_h))\right\|_{\text{op}} \\
&= \sum_{s_h\in\mathcal{S}}\left|\sqrt{\widehat{\rho}_h(s_h)} - \sqrt{\rho_h(s_h)}\right|^2\|A_h(s_h,\widehat{\nu}_h)\|_{\text{op}}^2 + \sum_{s_h\in\mathcal{S}}\left|\sqrt{\widehat{\rho}_h(s_h)} - \sqrt{\rho_h(s_h)}\right|^2\|B_h(s_h,\widehat{\mu}_h)\|_{\text{op}}^2 \\
&\le \sum_{s_h\in\mathcal{S}}|\widehat{\rho}_h(s_h) - \rho_h(s_h)|\,\|A_h(s_h,\widehat{\nu}_h)\|_{\text{op}}^2 + \sum_{s_h\in\mathcal{S}}|\widehat{\rho}_h(s_h) - \rho_h(s_h)|\,\|B_h(s_h,\widehat{\mu}_h)\|_{\text{op}}^2,
\end{aligned}$$

where $\widehat{\rho}_h(s)$ is the frequency of the visitation to a state $s\in\mathcal{S}$ at the step h. Then

$$\begin{aligned}
\sum_{s_h\in\mathcal{S}}|\widehat{\rho}_h(s_h) - \rho_h(s_h)|\,\|A_h(s_h,\widehat{\nu}_h)\|_{\text{op}}^2 &= \sum_{s\in\mathcal{S}}|\widehat{\rho}_h(s) - \rho_h(s)|\,\|\Phi_{1,s}\|_{\text{op}}^2\|\nu_h(\cdot|s)\|^2 \\
&\le 2\max_{s\in\mathcal{S}}\|\Phi_{1,s}\|_{\text{op}}^2\text{TV}(\widehat{\rho}_h,\rho_h).
\end{aligned}$$

Applying the same procedure to $\sum_{s_h\in\mathcal{S}}|\widehat{\rho}_h(s_h) - \rho_h(s_h)|\,\|B_h(s_h,\widehat{\mu}_h)\|_{\text{op}}^2$, we have

$$\|\widehat{X}_h - \widetilde{X}_h\|_{\text{op}} \le 2\sqrt{\max_{s\in\mathcal{S}}\|\Phi_{1,s}\|_{\text{op}}^2 + \max_{s\in\mathcal{S}}\|\Phi_{2,s}\|_{\text{op}}^2}\text{TV}(\widehat{\rho}_h,\rho_h).$$

Using the triangle inequality,

$$\begin{aligned}
\|\widehat{X}_h - X_h\|_{\text{op}} &\le \|\widehat{X}_h - \widetilde{X}_h\|_{\text{op}} + \|\widetilde{X}_h - X_h\|_{\text{op}} \\
&\le 2\sqrt{\max_{s\in\mathcal{S}}\|\Phi_{1,s}\|_{\text{op}}^2 + \max_{s\in\mathcal{S}}\|\Phi_{2,s}\|_{\text{op}}^2}\left(2\epsilon + \text{TV}(\widehat{\rho}_h,\rho_h)\right).
\end{aligned} \tag{97}$$

Bound the y-vector error. We are going to bound the error of the vector $\widehat{y}_h$ in the least square problem. By definition and Jensen's inequality,

$$
\begin{aligned}
\|\widetilde{y}_h - y_h\|^2 &= \sum_{s_h \in \mathcal{S}} \rho_h(s_h) \left\| c_h(s_h, \widehat{\mu}_h) - c_h(s_h, \mu_h) \right\|^2 + \sum_{s_h \in \mathcal{S}} \rho_h(s_h) \left\| d_h(s_h, \widehat{\nu}_h) - d_h(s_h, \nu_h) \right\|^2 \\
&= \mathbb{E}_{s_h \sim \rho_h} \left[\left\| c_h(s_h, \widehat{\mu}_h) - c_h(s_h, \mu_h) \right\|^2 \right] + \mathbb{E}_{s_h \sim \rho_h} \left[\left\| d_h(s_h, \widehat{\nu}_h) - d_h(s_h, \nu_h) \right\|^2 \right].
\end{aligned}
$$

By definition of $c_h(s_h, \mu_h)$ and $c_h(s_h, \widehat{\mu}_h)$,

$$
\begin{aligned}
\mathbb{E}_{s_h \sim \rho_h} \left[\left\| c_h(s_h, \widehat{\mu}_h) - c_h(s_h, \mu_h) \right\|^2 \right] &= \frac{1}{\eta^2} \mathbb{E}_{s_h \sim \rho_h} \left[\sum_{a=2}^{m} \left(\log \frac{\widehat{\mu}_h(a|s_h)}{\widehat{\mu}_h(1|s_h)} - \log \frac{\mu_h(a|s_h)}{\mu_h(1|s_h)} \right)^2 \right] \\
&\le \frac{2}{\eta^2} \mathbb{E}_{s_h \sim \rho_h} \left[(m-1) \left(\log \frac{\widehat{\mu}_h(1|s_h)}{\mu_h(1|s_h)} \right)^2 + \sum_{a=2}^{m} \left(\log \frac{\widehat{\mu}_h(a|s_h)}{\mu_h(a|s_h)} \right)^2 \right].
\end{aligned}
$$

According to the estimate (19) and the bound (67), for all $a \in \mathcal{A}$ and $s_h \in \mathcal{S}$, we have

$$
\left(\log \frac{\widehat{\mu}_h(a|s_h)}{\mu_h(a|s_h)} \right)^2 \le \left(\frac{\widehat{\mu}_h(a|s_h) - \mu_h(a|s_h)}{e^{-2K}/m} \right)^2 \le m^2 e^{4K} \mathrm{TV}^2(\widehat{\mu}_h(\cdot|s_h), \mu_h(\cdot|s_h)).
$$

Hence

$$
\mathbb{E}_{s_h \sim \rho_h} \left[\left\| c_h(s_h, \widehat{\mu}_h) - c_h(s_h, \mu_h) \right\|^2 \right] \le \frac{4m^3 e^{4K}}{\eta^2} \mathbb{E}_{s_h \sim \rho_h} \left[\mathrm{TV}^2(\widehat{\mu}_h(\cdot|s_h), \mu_h(\cdot|s_h)) \right].
$$

Applying the same procedure to the other term in the decomposition of $\|\widetilde{y}_h - y_h\|_{\mathrm{op}}^2$, and using the fact that $\mathbb{E}_{s_h \sim \rho_h} \left[\mathrm{TV}^2(\nu_h(\cdot|s_h), \widehat{\nu}_h(\cdot|s_h)) \right] \le \epsilon^2$ and $\mathbb{E}_{s_h \sim \rho_h} \left[\mathrm{TV}^2(\mu_h(\cdot|s_h), \widehat{\mu}_h(\cdot|s_h)) \right] \le \epsilon^2$, we have

$$
\|\widetilde{y}_h - y_h\|^2 \le \frac{4e^{4K}(m^3 + n^3)}{\eta^2} \epsilon^2. \tag{98}
$$

Now it remains to bound the empirical error $\|\widehat{y}_h - \widetilde{y}_h\|$.

$$
\begin{aligned}
\|\widehat{y}_h - \widetilde{y}_h\|^2 &= \sum_{s_h \in \mathcal{S}} \left| \sqrt{\widehat{\rho}_h(s_h)} - \sqrt{\rho_h(s_h)} \right|^2 \left\| c_h(s_h, \widehat{\mu}_h) \right\|^2 + \sum_{s_h \in \mathcal{S}} \left| \sqrt{\widehat{\rho}_h(s_h)} - \sqrt{\rho_h(s_h)} \right|^2 \left\| d_h(s_h, \widehat{\nu}_h) \right\|^2 \\
&\le \sum_{s_h \in \mathcal{S}} \left| \widehat{\rho}_h(s_h) - \rho_h(s_h) \right| \left\| c_h(s_h, \widehat{\mu}_h) \right\| + \sum_{s_h \in \mathcal{S}} \left| \widehat{\rho}_h(s_h) - \rho_h(s_h) \right| \left\| d_h(s_h, \widehat{\nu}_h) \right\|^2.
\end{aligned}
$$

By definition of $c_h(s_h, \widehat{\mu_h})$,

$$
\sum_{s \in \mathcal{S}} \left| \widehat{\rho}_h(s) - \rho_h(s) \right| \left\| c_h(s_h, \widehat{\mu}_h) \right\| \le \sum_{s \in \mathcal{S}} \left| \widehat{\rho}_h(s) - \rho_h(s) \right| \frac{2K \log m}{\eta} \le \frac{4K \log m}{\eta} \mathrm{TV}(\widehat{\rho}_h, \rho_h).
$$

Similarly,

$$
\sum_{s_h \in \mathcal{S}} \left| \widehat{\rho}_h(s_h) - \rho_h(s_h) \right| \left\| d_h(s_h, \widehat{\nu}_h) \right\|^2 \le \frac{4K \log n}{\eta} \mathrm{TV}(\widehat{\rho}_h, \rho_h).
$$

Hence

$$
\|\widehat{y}_h - \widetilde{y}_h\| \le \frac{4K \sqrt{(\log m)^2 + (\log n)^2}}{\eta} \mathrm{TV}(\widehat{\rho}_h, \rho_h). \tag{99}
$$

Combining (98) and (99) by the triangle inequality, we have

$$
\|\widehat{y}_h - y_h\| \le \frac{2e^{2K}\sqrt{m^3 + n^3}}{\eta} \epsilon + \frac{4K \sqrt{(\log m)^2 + (\log n)^2}}{\eta} \mathrm{TV}(\widehat{\rho}_h, \rho_h). \tag{100}
$$

Determine the threshold. Combining the estimates (97) and (100), we have

$$\begin{aligned}\|\widehat{X}_h\theta_h - \widehat{y}_h\|^2 &= \|(\widehat{X}_h - X_h)\theta_h - (\widehat{y}_h - y_h)\|^2 \\ &\leq 2\left(\|\widehat{X}_h - X_h\|_{\mathrm{op}}^2\|\theta_h\|^2 + \|\widehat{y}_h - y_h\|^2\right) \\ &\lesssim R^2\left(\max_{s\in\mathcal{S}}\|\Phi_{1,s}\|_{\mathrm{op}}^2 + \max_{s\in\mathcal{S}}\|\Phi_{2,s}\|_{\mathrm{op}}^2 + \frac{e^{4K}(m^3+n^3)}{\eta^2}\right)\epsilon^2 \\ &\quad + \left(\max_{s\in\mathcal{S}}\|\Phi_{1,s}\|_{\mathrm{op}}^2 + \max_{s\in\mathcal{S}}\|\Phi_{2,s}\|_{\mathrm{op}}^2 + \frac{4K(\log mn)^2}{\eta^2}\right)\mathrm{TV}^2(\widehat{\rho}_h, \rho_h).\end{aligned} \tag{101}$$

We consider the following three concentration events:

(i) *The estimated policy of the max player has small error.* By the bound (75), with probability at least $1-\delta$, the following event holds:

$$\mathcal{E}_a = \left\{\mathbb{E}_{s_h\sim\rho_h}\left[\mathrm{TV}^2(\widehat{\mu}_h(\cdot|s_h), \mu_h^*(\cdot|s_h))\right] \lesssim \frac{d_a\log T + \log(H/\delta) + \sqrt{m}}{T}, \quad \forall h\in[H]\right\}.$$

(ii) *The estimated policy of the min player has small error.* Similar to the bound (75), with probability at least $1-\delta$, the following event holds:

$$\mathcal{E}_b = \left\{\mathbb{E}_{s_h\sim\rho_h}\left[\mathrm{TV}^2(\widehat{\nu}_h(\cdot|s_h), \nu_h^*(\cdot|s_h))\right] \lesssim \frac{d_b\log T + \log(H/\delta) + \sqrt{n}}{T}, \quad \forall h\in[H]\right\}.$$

(iii) *The frequency estimator of the visitation measure has small error.* Similar to the tail bound (30), with probability at least $1-\delta$, the following event holds:

$$\mathcal{E}_s = \left\{\mathrm{TV}(\widehat{\rho}_h, \rho_h) \leq \frac{1}{2}\sqrt{\frac{S}{T}} + \sqrt{\frac{\log(H/\delta)}{2T}}, \quad \forall h\in[H]\right\}.$$

The event $\mathcal{E}_a\cap\mathcal{E}_b\cap\mathcal{E}_s$ holds with probability at least $1-3\delta$. Corresponding to $\mathcal{E}_a$ and $\mathcal{E}_b$, we take

$$\epsilon^2 = \mathcal{O}\left(\frac{(d_a+d_b)\log T + \log(H/\delta) + \sqrt{m} + \sqrt{n}}{T}\right).$$

According to (101), we set

$$\kappa_h = \mathcal{O}\left(\frac{m^{7/2} + n^{7/2} + (m^3+n^3)((d_a+d_b)\log T + \log(H/\delta)) + S(\log mn)^2}{T}\right).$$

Then we have $\|\widehat{X}_h\theta_h - \widehat{y}_h\|^2 \leq \kappa_h$, and the confidence set with respect to this threshold contains the true parameter θ_h with probability at least $1-3\delta$.

Since both $\|\widehat{X}_h - X_h\|_{\mathrm{op}}$ and $\|\widehat{y}_h - y_h\|$ are bounded by $\mathcal{O}(T^{-1})$ under the event $\mathcal{E}_a\cap\mathcal{E}_b\cap\mathcal{E}_s$, by applying the same procedure as in Lemma A.2, we conclude that, for each $h\in[H]$, there exists a constant C_h independent of T such that

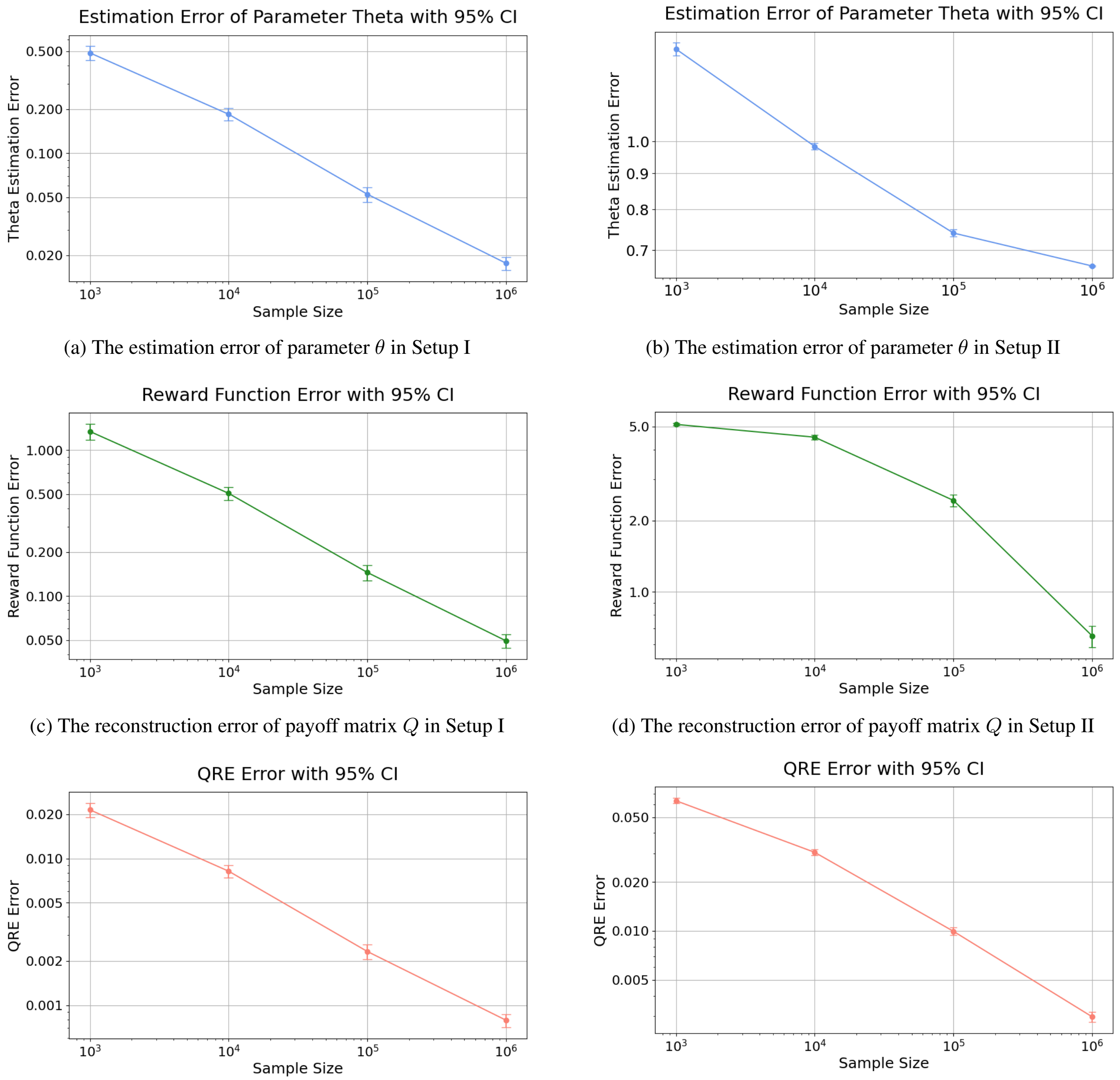
$$d_{\mathrm{H}}(\Theta_h, \widehat{\Theta}_h) \leq C_h\sqrt{\kappa_h}.$$

Thus we complete the proof. □

E. Additional Results of Numerical Experiments

In this section, we present the results of our numerical experiments on two-player entropy-regularized zero-sum matrix games. We discuss both the strong identification case and the partial identification case.

We evaluate the performance of our algorithm by three key metrics: (1) the error in parameter estimation, which measures the difference between the estimated reward parameter $\widehat{\theta}$ and the true parameter θ^*; (2) the error in the estimated payoff function $\widehat{Q}$, which evaluates how accurately the reconstructed payoff function matches the true payoff function; and (3) the error in the estimated QRE, which quantifies the discrepancy between the QRE $(\widehat{\mu}, \widehat{\nu})$ derived from the estimated payoff function and the true QRE (μ^*, ν^*). Among these metrics, we are particularly interested in the error in the estimated QRE, which validates whether the algorithm aligns with the observed data and behavior.



(a) The estimation error of parameter θ in Setup I

(b) The estimation error of parameter θ in Setup II

(c) The reconstruction error of payoff matrix Q in Setup I

(d) The reconstruction error of payoff matrix Q in Setup II

(e) The discrepancy between the QRE $(\widehat{\mu}, \widehat{\nu})$ derived from the estimated payoff $\widehat{Q}$ and the true QRE (μ^*, ν^*) in Setup I

(f) The discrepancy between the QRE $(\widehat{\mu}, \widehat{\nu})$ derived from the estimated payoff $\widehat{Q}$ and the true QRE (μ^*, ν^*) in Setup II

Figure 3. The results of numerical simulation on zero-sum matrix games. Both X and Y axes are log-scaled. The X-axis represents the sample size from 10^3 to 10^6. The Y-axis represents (a,b) the error $\|\widehat{\theta} - \theta^*\|$ of the estimate $\widehat{\theta}$; (c,d) The Y-axis represents the error $\|\widehat{Q} - Q^*\|_{\mathrm{F}}$ of the reward function $\widehat{Q}$; (e,f) The Y-axis represents the error $\mathrm{TV}(\widehat{\mu}, \mu^*) + \mathrm{TV}(\widehat{\nu}, \nu^*)$. We repeat 100 experiments for each sample size and plot 95% confidence interval for the error.

Setup I. We define the kernel function $\phi : \mathcal{A} \times \mathcal{B} \to \mathbb{R}^d$ with dimension $d = 2$, and set the true parameter to be

$$\theta^* = (0.8, -0.6)^\top.$$

We set the sizes of action spaces to be $m = 4,\ n = 6$. The entropy regularization term is $\eta = 0.5$.

To generate the dataset, we sample i.i.d. pairs $\{(a_i, b_i)\}_{i=1}^N$ from the QRE (μ^*, ν^*) corresponding to the true payoff function $Q^*(a, b) = \phi(a, b)^\top \theta^*$. We conduct experiments for varying sample sizes $N \in \{10^3, 10^4, 10^5, 10^6\}$, and repeat 100 times for each N.

We implement the algorithm proposed in §2.2.

Setup II. We define the kernel function $\phi : \mathcal{A} \times \mathcal{B} \to \mathbb{R}^d$ with dimension $d = 6$, and set the true parameter to be

$$\theta^* = (0.8, -0.6, 0.75, 0.2, 0.5, -0.5)^\top.$$

We set the size of action spaces to be $m = 6,\ n = 6$. The entropy regularization term is $\eta = 0.5$.

To generate the dataset, we sample i.i.d. pairs $\{(a_i, b_i)\}_{i=1}^N$ from the QRE (μ^*, ν^*) corresponding to the true payoff function $Q^*(a, b) = \phi(a, b)^\top \theta^*$. We conduct experiments for varying sample sizes $N \in \{10^3, 10^4, 10^5, 10^6\}$, and repeat 100 times for each N.

We implement the algorithm proposed in §2.3. In each experiment, our algorithm outputs a parameter $\widehat{\theta}$ in the confidence set $\widehat{\Theta}$. We set the bound of feasible parameters θ to be $M = 4$, and set the threshold $\kappa = 10^3/N$, where N is the sample size.

Results. We conduct experiments for both strong identification and partial identification cases.

In Setup I, the true parameter θ^* is uniquely identifiable. The results are presented in Figures 3a, 3c and 3e, where both X and Y axis take the log scale. Figures 3a and 3c demonstrate that the estimated parameter $\widehat{\theta}$ and the reconstructed payoff matrix $\widehat{Q}$ converge to their true values, with the estimation error following an order close to $\mathcal{O}(N^{-1/2})$, which is consistent with our theoretical results. Figure 3e shows that the QRE corresponding to our estimated payoff matrix aligns with the true QRE.

In Setup II, the true parameter θ^* is partially identifiable, meaning that multiple feasible parameters can explain the observed strategies. The results are shown in Figures 3b, 3d and 3f, where both X and Y axis take the log scale. As expected, Figures 3b and 3d illustrate that the estimated parameter $\widehat{\theta}$ and the payoff matrix $\widehat{Q}$ do not necessarily converge to the true values. Nevertheless, Figure 3f shows that the QRE derived from the estimated payoff still converges to the true QRE. This indicates that even when the reward function is not uniquely identifiable, our estimated payoff structure remains a valid explanation of the observed agents' behavior.

F. Auxiliary Lemmas

Lemma F.1 (Zhang, 2007)**.** *Let μ and μ^* be two discrete probability distributions on $\{1, 2, \cdots, n\}$ such that $\mathrm{TV}(\mu, \mu^*) \le \epsilon \le 1/2$. Then*

$$|\mathcal{H}(\mu^*) - \mathcal{H}(\mu^*)| \le -\epsilon \log \epsilon - (1 - \epsilon) \log(1 - \epsilon) + \epsilon \log(n - 1).$$

Lemma F.2 (Vershynin, 2018, Proposition 5.2.2)**.** *Let $R > 0$, and let $B(0, R) = \{x \in \mathbb{R}^d : \|x\| \le R\}$ be the R-ball centered at 0 in the Euclidean space $(\mathbb{R}^d, \|\cdot\|)$. Then for any $\epsilon > 0$, the covering number $N(\epsilon, B(0, R), \|\cdot\|)$ admits the following bound:*

$$N(\epsilon, B(0, R), \|\cdot\|) \le \left(1 + \frac{2R}{\epsilon}\right)^d.$$

Lemma F.3 (Tropp, 2011, Corollary 7.5)**.** *Let $Z_1, \cdots, Z_n$ be a family of independent random variables, and let H be a function that maps n variables to a self-adjoint matrix of dimension d. Consider a sequence $A_1, \cdots, A_n$ of fixed self-adjoint matrices that satisfy*

$$\left(H(z_1, \cdots, z_{k-1}, z_k, z_{k+1}, \cdots, z_n) - H(z_1, \cdots, z_{k-1}, z_k', z_{k+1}, \cdots, z_n)\right)^2 \preceq A_k^2,$$

where z_i and z_i' range over all possible values of Z_i for each index i. Let

$$\sigma^2 = \left\| \sum_{k=1}^{n} A_k^2 \right\|.$$

Then for all $t > 0$,

$$\mathbb{P}\left(\lambda_{\max}\left(H(Z_1, \cdots, Z_n) - \mathbb{E}\left[H(Z_1, \cdots, Z_n)\right]\right) \geq t\right) \leq 1 - d \cdot e^{-t^2/8\sigma^2}.$$

Lemma F.4 (Abbasi-Yadkori et al., 2011, Lemma 9)**.** *Let $(\mathscr{F}_t)_{t=1}^{\infty}$ be a filtration, and let $(\eta_t)_{t=1}^{\infty}$ be an adapted and conditional σ-sub-Gaussian process, i.e. η_t is $\mathscr{F}_t$-measurable, and*

$$\mathbb{E}[\eta_t | \mathscr{F}_{t-1}] = 0, \qquad \mathbb{E}\left[e^{\lambda \eta_t} \middle| \mathscr{F}_{t-1}\right] \leq e^{\lambda^2 \sigma^2/2}, \quad \forall \lambda \in \mathbb{R}.$$

Let $(X_t)_{t=1}^{\infty}$ be a predictable $\mathbb{R}^d$-valued process with respect to $(\mathscr{F}_t)_{t=1}^{\infty}$, i.e. X_t is $\mathscr{F}_{t-1}$ measurable. Assume that $V_0 \in \mathbb{R}^{d \times d}$ is a positive definite matrix, and for any $t \geq 0$, let $V_t = V_0 + \sum_{s=1}^{t} X_s X_s^\top$. Let τ be a stopping time with respect to $(\mathscr{F}_t)_{t=1}^{\infty}$. Then, for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$:

$$\left\| \sum_{t=1}^{\tau} X_t \eta_t \right\|_{V_\tau^{-1}}^2 \leq 2\sigma^2 \log\left(\frac{\det(V_\tau)^{1/2} \det(V_0)^{-1/2}}{\delta} \right).$$

Lemma F.5 (Foster et al., 2023, Lemma A.4)**.** *Let $(\mathscr{F}_t)_{t=1}^{T}$ be a filtration, and let $(X_t)_{t=1}^{T}$ be an adapted process. Then for any $\delta > 0$, it holds with probability at least $1 - \delta$ that for all $t \leq T$,*

$$\sum_{i=1}^{t} X_i \leq \sum_{i=1}^{t} \log \mathbb{E}\left[e^{X_i} | \mathscr{F}_{i-1}\right] + \log \frac{1}{\delta}.$$