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 (Metelli et al., 2021; Lindner et al., 2022; Metelli et al., 2023). (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 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 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

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 $a_n = O(b_n)$ 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 $\|A\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$, 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 (A, B, Q), where $A = \{1, 2, \cdots, m\}$ and $B = \{1, 2, \cdots, n\}$ 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 $\mu \in \Delta(A)$ and $\nu \in \Delta(B)$ are policies for each player, and $Q = (Q(a, b))_{a\in A, b\in B} \in \mathbb{R}^{m\times n}$ 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):

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

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

$$H(\mu) = \sum_i \mu_i \log(\mu_i)$$

denotes the Shannon entropy (Shannon, 1948) of μ. 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:

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

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

$$\begin{pmatrix} Q(a,\cdot) - Q(1,\cdot) \\ Q(\cdot,b) - Q(\cdot,1) \end{pmatrix} \nu^* = \begin{pmatrix} \log(\mu^*(a)/\mu^*(1))/\varkappa, \\ \log(\nu(b)/\nu(1))/\varkappa. \end{pmatrix} \quad (1)$$

Goal. We study the inverse game theory for this entropy-regularized zero-sum game. To elaborate, we observe strategy pairs $(a^k, b^k) \overset{\text{iid}}{\sim} (\mu^*, \nu^*)$ where (μ^*, ν^*) 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 $(\hat{\mu}, \hat{\nu})$. 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 $\phi : A \times B \to \mathbb{R}^d$ 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 $(a, b) \in A \times B$.

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 $a \in A$ and $b \in B$,

$$\begin{pmatrix} \phi(a,\cdot) - \phi(1,\cdot) \\ \phi(\cdot,b) - \phi(\cdot,1) \end{pmatrix} \theta^* = \begin{pmatrix} \log(\hat{\mu}(a)/\hat{\mu}(1))/\varkappa, \\ \log(\hat{\nu}(b)/\hat{\nu}(1))/\varkappa, \end{pmatrix}$$

where $\begin{pmatrix} \phi(a,\cdot) - \phi(1,\cdot) \\ \phi(\cdot,b) - \phi(\cdot,1) \end{pmatrix} \in \mathbb{R}^d$. To simplify the notation, we define matrices

$$A(\hat{\nu}) = \big((\phi(a,\cdot) - \phi(1,\cdot))\hat{\nu}\big)_{a\in A\setminus\{1\}} \in \mathbb{R}^{(m-1)\times d},$$

$$B(\hat{\mu}) = \big((\phi(\cdot,b) - \phi(\cdot,1))^\top \hat{\mu}\big)_{b\in B\setminus\{1\}} \in \mathbb{R}^{(n-1)\times d},$$

and define vectors

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

Then the linear constraints would be

$$\begin{pmatrix} A(\beta) \\ B(\mu') \end{pmatrix} \theta = \begin{pmatrix} c(\mu') \\ d(\beta) \end{pmatrix}. \tag{2}$$

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 $Q(a,b) = h_\theta(a,b)$, θ (i.e. $\theta = \theta'$) for all $(a,b) \in A \times B$ if and only if the QRE satisfies the rank condition*

$$\operatorname{rank}\begin{pmatrix} A(\beta) \\ B(\mu') \end{pmatrix} = d. \tag{3}$$

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', β) from the observed data and obtain $(\hat{\mu}, \hat{\beta})$
2. Leverge (2) to estimate θ. To be specific, we conduct the least-square estimation and obtain $\hat{\theta}$:

$$\hat{\theta} := \arg\min_{\theta\in\mathbb{R}^d} \left\| \begin{pmatrix} A(\hat{\beta}) \\ B(\hat{\mu}) \end{pmatrix} \theta - \begin{pmatrix} c(\hat{\mu}) \\ d(\hat{\beta}) \end{pmatrix} \right\|^2, \tag{4}$$

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

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

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

Theorem 2.3 (Parameter estimation error)**.** *Let ε_1 and ε_2 be two small numbers satisfying $\varepsilon_1 < \min_{a\in[m]} \mu'(a)$ and $\varepsilon_2 < \min_{b\in[n]} \beta(b)$. Under Assumption 2.1 and the rank condition in (3), suppose $(\hat{\mu}, \hat{\beta})$ satisfies $\mathrm{TV}(\hat{\mu}, \mu') \le \varepsilon_1/2$ and $\mathrm{TV}(\hat{\beta}, \beta') \le \varepsilon_2/2$, then $\hat{\theta}$ constructed by (4) satisfies*

$$\|\hat{\theta} - \theta\|^2 \le \varepsilon_1^{-2} \cdot 1 + m \cdot (\varepsilon_2^{-2} + 1) + \varepsilon_2^{-2} \cdot 1 + n \cdot (\varepsilon_1^{-2} + 1) \cdot$$

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 $\hat{\mu}, \hat{\beta}$, 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', β'), we obtain $\hat{\mu}, \hat{\beta}$ by the frequency estimator. For any $\delta \in (0,1)$, the estimation error bound of the payoff matrix holds with probability at least $1-\delta$*

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

Theorem 2.4 provides a probabilistic guarantee for the accuracy of the reconstructed payoff matrix $\hat{Q}$ in a finite-sample setting. The bound explicitly depends on the sample size N, the action space dimensions m, n, and the confidence parameter δ. The estimation error decreases at a rate of $O(1/N)$, which is consistent with the standard empirical result of the frequency estimator (van der Vaart, 1998). As the sample size N increases, the errors of $\hat{\mu}$ and $\hat{\beta}$ decrease, leading to a more accurate reconstruction of the reward Q'. On the other hand, the bound grows with the action space size in terms of $m^2 + n^2$, indicating that larger action spaces require more samples to achieve the same estimation accuracy.

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 feasible set $\Theta \subseteq \mathbb{R}^d$ is

$$\Theta = \left\{ \theta : \begin{pmatrix} A(\beta) \\ B(\mu') \end{pmatrix} \theta = \begin{pmatrix} c(\mu') \\ d(\beta) \end{pmatrix}, \|\theta\| \le M \right\}.$$

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 $\hat{\mu}$ and $\hat{\beta}$. Next, we select a threshold $\varepsilon_N > 0$ and construct the confidence set as follows:

$$\hat{\Theta}_N = \left\{ \theta : \left\| \begin{pmatrix} A(\hat{\beta}) \\ B(\hat{\mu}) \end{pmatrix} \theta - \begin{pmatrix} c(\hat{\mu}) \\ d(\hat{\beta}) \end{pmatrix} \right\|^2 \le \varepsilon_N, \|\theta\| \le M \right\}. \tag{6}$$

To recover the feasible payoff functions, we simply compute $\hat{Q}(a,b) = (a,b)^\top \hat{\theta}$ for all $\hat{\theta} \in \hat{\Theta}_N$ according to the linear assumption.

We demonstrate the effectiveness of our Algorithm by establishing its ability to construct accurate confidence sets. To be specific, we show that the confidence set $\hat{\Theta}_N$ 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 $\hat{\kappa}_N$ that makes the confidence set $\hat{\Theta}_N$ "similar" to Θ. The following theorem formalizes this intuition.

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

$$d_H(\Theta, \hat{\Theta}_N) \leq \frac{m+n+\sqrt{(m+n)\log(1/\beta)}}{N}, \quad (7)$$

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

Theorem 2.5 establishes the asymptotic consistency of our confidence set $\hat{\Theta}_N$ 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 $O(N^{-1/2})$, 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.,

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

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):

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

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

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

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

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

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

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

$$\|\hat{\theta} - \theta^*\| \leq \frac{m+n+\sqrt{(m+n)\log(1/\beta)}}{\sigma N}.$$

In practice, selecting the minimum-norm solution helps avoid overfitting and promotes stability (Hastie et al., 2009). The convergence rate $O(N^{-1/2})$ 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 (S, A, B, r, P, H), where

- S is the state space, with $|S| = S$,
- A and 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
- $P = \{P_h\}_{h\in[H]}$ is a collection of transition kernels.

At each time step $h \in [H]$, the players 1 and 2 simultaneously take actions $a \in A$ and $b \in B$ respectively upon observing the state $s \in S$, 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 $s' \sim P_h(\cdot|s, a, b)$ according to the transition kernel 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

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

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

$$Q_h^{\mu,\nu}(s, a, b) = r_h(s, a, b) + \mathbb{E}_{P_h(\cdot|s,a,b)}\, V_{h+1}^{\mu,\nu}(\cdot). \quad (8)$$

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 $[Q_h^{\mu,\nu}(s, a, b)]_{(a,b)\in A\times B}$. With this notation, we may write

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

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:

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

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:

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

which is equivalent to

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

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

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

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 (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 $S \times A \times B$ and quantal response equilibrium (μ^*, ν^*), a reward function $r : S \times A \times B \to \mathbb{R}^H$ is identified if $\{\mu_h, \nu_h\}$ 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 $r_h : S \times A \times B \to [0, 1]$ and every transition kernel $P_h : S \times A \times B \to \Delta(S)$, there exist $\theta_h \in \mathbb{R}^d$ and $\phi_h(\cdot) : S \to \mathbb{R}^d$ such that

$$r_h(s, a, b) = (s, a, b)^\top \theta_h,$$
$$P_h(\cdot|s, a, b) = (s, a, b)^\top \phi_h(\cdot)$$

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

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

We assume $\|(\cdot, \cdot, \cdot)\| \le 1$, $\|\omega_h\| \le R$, and $\|\phi_h(s)\|^p \le d$ for all $h \in [H]$ and $s \in S$.

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 $R \le H(1 + \log m + \log n)$ exists. Since (θ_h) can be recovered by (ω_h), we prefer to make an assumption on (ω_h) instead of (θ_h) for the convenience of subsequent analysis.

We are going to find all the feasible θ_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 ω_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 $Q_h(s, a, b) = (s, a, b)^\top \omega_h$ is feasible for all $(s, a, b) \in$*

$S \to A \to B$ *if* ψ_h *satisfies the following linear system:*

$$\begin{pmatrix} A_h(s, \pi) \\ B_h(s, \mu_h) \end{pmatrix} \psi_h = \begin{pmatrix} c_h(s, \mu_h) \\ d_h(s, \pi) \end{pmatrix} \quad \text{for all } s \in S, \tag{12}$$

where

$$A_h(s, \pi) = ((\quad (s, a, \cdot) \qquad (s, 1, \cdot)) \; \pi(\cdot|s))_{a \in A\setminus\{1\}},$$
$$B_h(s, \mu_h) = ((\quad (s, \cdot, 1) \qquad (s, \cdot, b)) \; \mu_h(\cdot|s))_{b \in B\setminus\{1\}}$$

and

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

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

$$\text{rank} \begin{pmatrix} A_h(\pi) & \mathbf{0} \\ \mathbf{0} & B_h(\mu_h)^\top \end{pmatrix} = d, \tag{13}$$

where

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

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

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

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):

$$\Psi_h = \{\psi_h \in \mathbb{R}^d : \left\| \begin{pmatrix} A_h(\pi) \\ B_h(\mu_h) \end{pmatrix} \psi_h - \begin{pmatrix} c_h(\mu_h) \\ d_h(\pi) \end{pmatrix} \right\|_2^2 \le \varepsilon_h\}. \tag{15}$$

- Calculate the feasible Q and V functions (Q_h and V_h) for all $\psi_h \in \Psi_h$.
- Estimate the transition kernel P_h from the observed data. Since the transition kernel has a linear structure, we employ ridge regression for estimation:

$$\Lambda_h = \sum_{t=1}^{T} (s_h^t, a_h^t, b_h^t)\,(s_h^t, a_h^t, b_h^t)^\top + \mathbf{I}_d,$$
$$\hat{P}_h V_{h+1}(s, a, b) = (s, a, b)^\top \Lambda_h^{-1} \sum_{t=1}^{T} (s_h^t, a_h^t, b_h^t) V_{h+1}(s_{h+1}^t);$$

- Recover feasible set R_h by the Bellman equation (14).

3.3. Theoretical Guarantees

In this section, we present the theoretical results for our Algorithm. 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

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

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 π if the Q function $Q = (Q_1, Q_2, \cdots, Q_H)$ satisfies the identifability condition (11) and the norm constraint $\|\psi_h\| \in R$. We denote R as the feasible reward set corresponding to the quantal response equilibrium μ and π, namely,

$$R := \Big(r = (r_1, r_2, \cdots, r_H) : r \text{ is identified and } \psi_h + \sum_{s \in S} P_h(s) V_{h+1}(s) \in R \text{ for all } h \in [H] \Big).$$

Also, we denote Q as the feasible Q function set:

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

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 $\|\psi_h\| \in R$ 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 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 $\nu_1 \in \Delta(S)$, and define the marginal state visitation distributions associated with policies μ, π at each time step $h \in [H]$ as $d_h^{\mu,\pi}(s) = \mathbb{P}(s_h = s | \nu_1, \mu, \pi)$ Also, write the state-action visitation distributions as $d_h^{\mu,\pi}(s, a, b) = \mathbb{P}(s_h = s, a_h = a, b_h = b | \nu_1, \mu, \pi)$

To control the uniform metric in Definition 3.6, we require an estimator of the QRE that performs uniformly well across

all states $s \in S$. When using frequency estimators to approximate the policies $\mu_h^*(\cdot|s)$ and $\hat{\pi}_h(\cdot|s)$, the estimation at each state is conducted independently. As a result, it is essential that the dataset sufficiently covers all states in 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

$$d_h^{\mu^*, \hat{\pi}}(s) \quad C$$

for all $s \in S$ 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 $\nu_h = d_h^{\mu^*, \hat{\pi}}$ be the stationary distribution associated with optimal policies μ^* and $\hat{\pi}$, where $h \in [H]$. We assume that the following $d \times d$ matrix*

$$_h = \mathrm{E}_{\nu_h} \quad (s_h, a_h, b_h) \quad (s_h, a_h, b_h)^\top$$

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

$$T \quad \max \quad \frac{1}{C^2}\log\frac{2HS}{}, \frac{16(m+n)}{C^2}\log\frac{4HS}{},$$

$$512k_{\mathrm{op}}^2\log\frac{2Hd}{}, 4k_{\mathrm{op}}.$$

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

$$D(R, \hat{R}) \le \frac{1}{T}\sqrt{S(m+n)\log\frac{HS}{}\log T}$$

$$+ S(m+n)\log\frac{HS}{} + \sqrt{\frac{}{Sd}} + \sqrt{d\log T}\log(mn),$$

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

Theorem 3.9 provides a strong guarantee on the accuracy of our reward recovery algorithm in Markov games. Our bound shows that the distance $D(R, \hat{R})$ diminishes at the rate of $O(T^{-1/2})$, 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 $_h$ is nonsingular ensures that the feature representation provides sufficient information for parameter recovery (Tu & Recht, 2017; Min et al., 2022). The norm $k \quad {}_h{}^{-1}k_{\mathrm{op}}$ 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 can extend our framework to integrate Maximum Likelihood Estimation (MLE) into our method and establish a convergence result with the same $T^{-1/2}$ rate.

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.

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

$$\theta_h^* = (0.8, \quad 0.6)^\top$$

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 $\tau = 0.5$

We implement the algorithm proposed in §3.2. In each experiment, our algorithm outputs a parameter $\hat{\theta}_h$ in the confidence set Θ_h. We set the bound of feasible parameters θ_h to be $R = 10$, and set the threshold $\epsilon_h = 10^{-3}/N$, where N is the sample size. The regularization term in ridge regression is $= 0.01$.

Metrics. We evaluate the performance of our algorithm using two metrics: (1) the error in the estimated reward function ($\hat{\theta}_h$), 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 ($\hat{\mu}, \hat{\pi}$) derived from the estimated payoff function and the true QRE ($\mu^*, \hat{\pi}$). 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 $(\hat{b}_h)_{h=1}^6$ can be relatively large, the corresponding QRE $(\hat{\mu}_h, \hat{\pi})$ remains well-aligned with the true QRE (μ_h^*, π^*). 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.

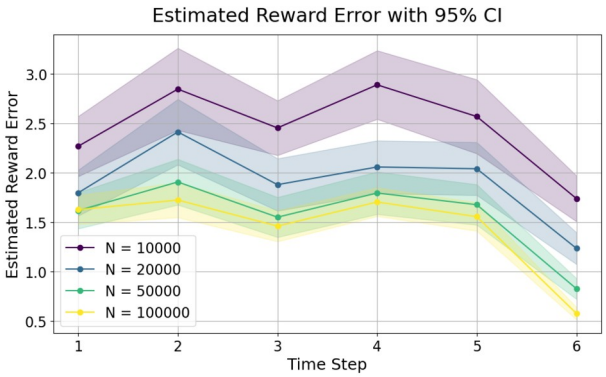

Figure 1. The reconstruction error of the reward functions $(\hat{b}_h)_{h=1}^6$. The X-axis represents the time step h from 1 to 6, while the Y-axis represents the error $\|\hat{b}_h - r_h^*\|_F$ of the reward function b.

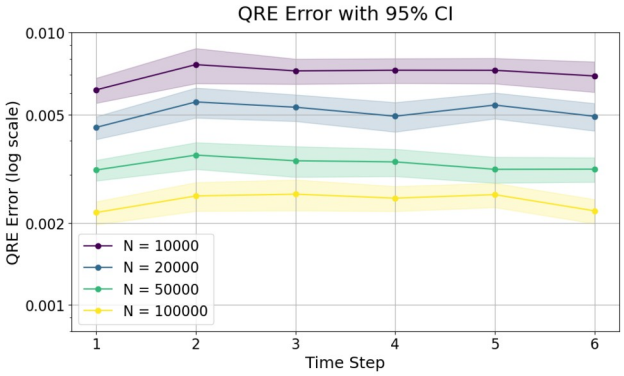

Figure 2. The discrepancy between the QRE $(\hat{\mu}, \hat{\pi})$ corresponding to the estimated reward functions $(\hat{b}_h)_{h=1}^6$ and the true QRE (μ^*, π^*). The X-axis represents the time step h from 1 to 6, while the Y-axis represents the errors $\mathrm{TV}(\hat{\mu}_h, \mu_h^*) + \mathrm{TV}(\hat{\pi},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

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.

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

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.

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

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, 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.

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.

Metelli, A. M., Ramponi, G., Concetti, A., and Restelli, M. Provably efficient learning of transferable rewards. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7665–7676. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/metelli21a.html.

Metelli, A. M., Lazzati, F., and Restelli, M. Towards theoretical understanding of inverse reinforcement learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24555–24591. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/metelli23a.html.

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.

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

van der Vaart, A. W. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

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.

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.

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.