Active Inverse Methods in Stackelberg Games with Bounded Rationality

Jianguo Chen [a] Jinlong Lei [b] Biqiang Mu [a] Yiguang Hong [b] Hongsheng Qi [a]

[a] *Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China; University of Chinese Academy of Sciences, Beijing, China*

[b] *Department of Control Science and Engineering, Tongji University, Shanghai, China; Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai, China; Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai, China; National Key Laboratory of Autonomous Intelligent Unmanned Systems, Shanghai, China; Frontiers Science Center for Intelligent Autonomous Systems, Ministry of Education, Shanghai, China*

Abstract

Inverse game theory is utilized to infer the cost functions of all players based on game outcomes. However, existing inverse game theory methods do not consider the learner as an active participant in the game, which could significantly enhance the learning process. In this paper, we extend inverse game theory to active inverse methods. For Stackelberg games with bounded rationality, the leader, acting as a learner, actively chooses actions to better understand the follower's cost functions. First, we develop a method of active learning by leveraging Fisher information to maximize information gain about the unknown parameters and prove the consistency and asymptotic normality. Additionally, when leaders consider its cost, we develop a method of active inverse game to balance exploration and exploitation, and prove the consistency and asymptotic Stackelberg equilibrium with quadratic cost functions. Finally, we verify the properties of these methods through simulations in the quadratic case and demonstrate that the active inverse game method can achieve Stackelberg equilibrium more quickly through active exploration.

Key words: Inverse game theory, active learning, exploration and exploitation, Stackelberg game, bounded rationality

1 Introduction

In recent years, the inverse problem has drawn increasing research attention in different areas such as control systems [44], robotics [31], and machine learning [26]. It aims to infer the system's objective function given the observed optimal behavior, and has been extended from single-agent to multi-agent scenarios.

In the single-agent case, inverse optimal control (IOC) and inverse reinforcement learning (IRL) are the two most common approaches. Specifically, IOC focuses on deriving an objective function from state/action samples under the assumption of a stable control system [1]. While IRL derives an objective function from expert demonstrations, assuming the expert's behavior is optimal [1]. Despite structural differences, both methods aim to reconstruct an objective function from observed data, as shown in Fig. 1 (a).

In the multi-agent case, inverse game theory is a framework that infers the cost functions of all players by observing the outcomes of the game when the cost functions are unknown [22], as shown in Fig. 1 (b). By now, various types of games have been addressed. For example, [30] aimed to infer parameters of player cost functions given that the state and control trajectories constitute an open-loop Nash equilibrium of a noncooperative differential game, and proposed a method by minimizing the violation of both the costate condition and the Hamiltonian condition. In two-person zero-sum stochastic games, [27] considered the problem of reconstructing rewards given a minimax bi-policy by solving an equivalent linear program. In addition, [45] inferred the cost function of the players in a matrix game such that a desired joint strategy is a Nash equilibrium based on semidefinite programs and bilevel optimization. However, the aforementioned literature only considers learning the objective function offline after ob-

Email addresses: chenjianguo@amss.ac.cn (Jianguo Chen), leijinlong@tongji.edu.cn (Jinlong Lei), bqmu@amss.ac.cn (Biqiang Mu), yghong@iss.ac.cn (Yiguang Hong), qihongsh@amss.ac.cn (Hongsheng Qi).

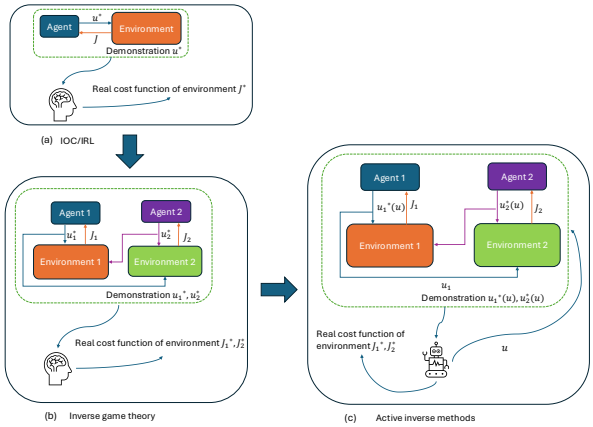

Fig. 1. The relationship develops from IOC/IRL to inverse game theory and active inverse methods.

serving all equilibrium data. In contrast, [7] considers the scenario where equilibria are acquired sequentially, and constructs this parameter identification problem as online optimization.

Nevertheless, these studies only focus on cases where the source of equilibrium data is given in advance. They do not consider scenarios where the learner, as a player in the game, can influence the outcome. In such cases, the learner can better understand other players' cost functions by actively choosing strategies, as shown in Fig. 1 (c). Simultaneously, participants in the game have their costs, so one should also consider their costs when actively understanding the opponent. As such, there is a balance between the exploration of active learning and the exploitation of cost reduction. We call the former *active learning* and the latter *active inverse game* to distinguish the two parts, although they are both included in our understanding of active inverse methods.

In this paper, we specifically consider a Stackelberg game with bounded rationality, where the leader, acting as a learner, can actively choose actions to better learn the true parameters in the follower's cost function. We design an algorithm for active inverse methods in Stackelberg games, focusing on learning and termed *active learning for Stackelberg games*, based on Fisher information. The leader will take the action that maximizes the Fisher information to get more information about the unknown parameters in the follower's cost function. Additionally, when the leader considers its own interaction costs while learning the follower's cost function, we propose a method of *active inverse game for Stackelberg games* (AIG for Stackelberg games) that achieves the best trade-off between exploration and exploitation like [9]. The main contributions of this paper are as follows:

(1) We propose a framework for active inverse games and use the Stackelberg game with bounded rationality as an example to illustrate it.
(2) We introduce an active learning approach based on optimal design, where the learner actively maximizes information gained using Fisher information. We prove its consistency and asymptotic normality, and verify its effectiveness through simulations.
(3) We further consider the scenario with the learner being a player in the game, which also needs to consider its cost that is influenced by other players' strategies. We then propose an active inverse game method for Stackelberg games that effectively balances exploration and exploitation, and prove its consistency as well.
(4) In particular, for a special case of Stackelberg game with quadratic cost functions, we further provide explicit expressions for all computational functions and prove that the leader's strategy achieves asymptotic Stackelberg equilibrium.

The remainder of the paper is organized as follows: Section 2 introduces the problem formulation of active inverse methods in Stackelberg games with bounded rationality. Section 3 presents the active learning algorithm for the problem based on Fisher information. Section 4 derives a method for AIG in Stackelberg games, balancing exploration and exploitation. Section 4.2 discusses the active inverse game when players' cost functions are quadratic. The proof of asymptotic consistency is detailed in Section 5. Section 6 presents simulations and compares the proposed framework with a uniform random strategy and inactive method, demonstrating superior performance. The paper concludes in Section 7.

The notation of the paper is shown as follows: $\mathbb{R}$ represents real number set; $\mathbb{R}^n$ represents the n-dimensional vector space; $\mathbb{R}^{n\times m}$ represents the space of $n \times m$-dimensional matrices; If the density function $p(X \mid \theta)$ of a random variable X is influenced by the parameter θ, we abbreviate $\mathbb{E}_{X\sim p(X|\theta)}[\cdot]$ and $\mathrm{Var}_{X\sim p(X|\theta)}[\cdot]$ to $\mathbb{E}_{X|\theta}[\cdot]$ and $\mathrm{Var}_{X|\theta}[\cdot]$, respectively; Similarly, for $p(X \mid y, \theta)$, we abbreviate $\mathbb{E}_{X\sim p(X|y,\theta)}[\cdot]$ and $\mathrm{Var}_{X\sim p(X|y,\theta)}[\cdot]$ to $\mathbb{E}_{X|y,\theta}[\cdot]$ and $\mathrm{Var}_{X|y,\theta}[\cdot]$, respectively; $S(\theta,\delta) \triangleq \{\theta' \mid \|\theta'-\theta\|_2 \le \delta\}$ denotes a closed neighborhood of radius δ around the point θ; $\xrightarrow{P}$ means convergence in probability; $\xrightarrow{a.s.}$ means convergence almost everywhere; $\xrightarrow{L}$ means convergence in distribution; Let $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

2 Problem formulation

In this section, we present the problem formulation for active inverse methods in Stackelberg games with bounded rationality.

2.1 Stackelberg games with bounded rationality

In a Stackelberg game, the strategic framework outlines two distinct roles: the leader and the follower [39]. Initially, the leader commits to an action, denoted by $\mathbf{u}^L \in \mathbb{R}^n$, selected from a set of feasible actions $\mathbf{U}^L \subseteq \mathbb{R}^n$. Subsequently, the follower observes the leader's chosen action and selects a response action $\mathbf{u}^F \in \mathbb{R}^h$ from its own set of feasible actions $\mathbf{U}^F \subseteq \mathbb{R}^h$. Given this sequential decision-making process, the leader and the follower incur costs denoted by $J^L(\mathbf{u}^L, \mathbf{u}^F)$ and $J^F(\mathbf{u}^F, \mathbf{u}^L)$, respectively. The cost functions, $J^L : \mathbf{U}^L \times \mathbf{U}^F \to \mathbb{R}$ for the leader and $J^F : \mathbf{U}^F \times \mathbf{U}^L \to \mathbb{R}$ for the follower, quantify the respective costs associated with each combination of actions chosen by the leader and the follower.

Given the leader's action $\mathbf{u}^L$, the follower can select its action $\mathbf{u}^F$ using various decision-making methods, such as best response or quantal response strategies. In the case of complete rationality, the follower chooses the best response strategy, expressed as:

$$\mathbf{u}^F \in \underset{\mathbf{u}^F \in \mathbf{U}^F}{\arg\min}\, J^F(\mathbf{u}^F, \mathbf{u}^L). \tag{1}$$

This involves selecting the action from its set $\mathbf{U}^F$ that minimizes its cost function $J^F(\mathbf{u}^F, \mathbf{u}^L)$, given the leader's action $\mathbf{u}^L$ [18]. However, in some cases, such as those involving human participants, the follower may not exhibit complete rationality [37,3]. This study addresses the scenario of bounded rationality. Consistent with prior works [29,11], we incorporate bounded rationality into the Stackelberg game framework by employing quantal response. Specifically, upon observing the leader's action $\mathbf{u}^L$, the follower determines its action $\mathbf{u}^F$ using a probability function:

$$p(\mathbf{u}^F \mid \mathbf{u}^L) \triangleq \frac{\exp\{-\lambda J^F(\mathbf{u}^F, \mathbf{u}^L)\}}{Z(\mathbf{u}^L)}, \mathbf{u}^F \in \mathbf{U}^F, \tag{2}$$

where $Z(\mathbf{u}^L) = \int_{\mathbf{U}^F} \exp\{-\lambda J^F(\mathbf{u}^F, \mathbf{u}^L)\}\, d\mathbf{u}^F$ and the rationality coefficient $\lambda \geq 0$ is used to modulate the follower's decision-making process [41]. Specifically, when $\lambda = 0$, the follower completely ignores the leader's influence. As λ approaches positive infinity, the follower's response converges to the best response strategy as (1).

With knowledge of the cost function and the follower's rationality coefficient, the leader can strategically determine its optimal deterministic strategy $\mathbf{u}^{L*}$, defined as the Stackelberg equilibrium strategy. This strategy is given by

$$\mathbf{u}^{L*} \in \underset{\mathbf{u}^L \in \mathbf{U}^L}{\arg\min}\, \mathbb{E}_{\mathbf{u}^F \mid \mathbf{u}^L} \left[J^L(\mathbf{u}^L, \mathbf{u}^F) \right], \tag{3}$$

where $\mathbb{E}_{\mathbf{u}^F \mid \mathbf{u}^L} \left[J^L(\mathbf{u}^L, \mathbf{u}^F) \right]$ denotes the expectation of the cost with respect to the probability density function $p(\mathbf{u}^F \mid \mathbf{u}^L)$.

2.2 The problem of active inverse methods in Stackelberg games

In a Stackelberg game, the leader may encounter a scenario where the follower's cost function is unknown [25,41]. Consequently, the leader lacks the necessary information to compute the Stackelberg equilibrium as defined in (3). To address the uncertainty surrounding the cost function, we adopt a parameterization approach wherein the leader knows the structure but lacks knowledge of the parameter $\theta \in \mathbb{R}^m$ within the follower's cost function $J^F(\mathbf{u}^F, \mathbf{u}^L; \theta)$, with the true parameters denoted as θ_0 [25,7]. The parameter θ is considered to belong to a set $\boldsymbol{\Theta} \subset \mathbb{R}^m$. Thus, if the leader assumes the unknown parameter to be θ, it anticipates the follower's action selection to be governed by the probability function:

$$p(\mathbf{u}^F \mid \mathbf{u}^L, \theta) \triangleq \frac{\exp\{-\lambda J^F(\mathbf{u}^F, \mathbf{u}^L; \theta)\}}{Z(\mathbf{u}^L, \theta)}, \mathbf{u}^F \in \mathbf{U}^F, \tag{4}$$

where $Z(\mathbf{u}^L, \theta) \triangleq \int_{\mathbf{U}^F} \exp\{-\lambda J^F(\mathbf{u}^F, \mathbf{u}^L; \theta)\}\, d\mathbf{u}^F$.

Although in reality, the follower selects its own action based on

$$p(\mathbf{u}^F \mid \mathbf{u}^L, \theta_0) = \frac{\exp\{-\lambda J^F(\mathbf{u}^F, \mathbf{u}^L; \theta_0)\}}{Z(\mathbf{u}^L, \theta_0)}, \mathbf{u}^F \in \mathbf{U}^F, \tag{5}$$

following the disclosure of the leader's action $\mathbf{u}^L$. Therefore, the Stackelberg equilibrium strategy in equation (3) is re-expressed by

$$\mathbf{u}^{L*} \in \underset{\mathbf{u}^L \in \mathbf{U}^L}{\arg\min}\, \mathbb{E}_{\mathbf{u}^F \mid \mathbf{u}^L, \theta_0} \left[J^L(\mathbf{u}^L, \mathbf{u}^F) \right]. \tag{6}$$

We examine a scenario in which the leader learns the unknown parameters within the follower's cost function through iterative interactions with the follower in a repeated Stackelberg game setting [41,7]. In the context of repeated Stackelberg games, the leader has the ability to actively employ effective strategies, thereby facilitating a more comprehensive acquisition of the unknown parameters in game-theoretic strategic interaction, with or without considering costs. This paper thus designates this problem as active inverse methods in Stackelberg games.

Specifically, during the T iterations of the repeated game, the leader actively selects strategies denoted as

$$\mathbf{D}^L(T) \triangleq \{\mathbf{u}^L(t)\}_{t=1}^T.$$

This generates a dataset

$$\mathbf{D}(T) \triangleq \{(\mathbf{u}^L(t), \mathbf{u}^F(t))\}_{t=1}^T,$$

where t represents the t-th interaction and $\mathbf{u}^F(t)$ is selected by the follower according to the probability (5) under $\mathbf{u}^L(t)$. Consequently, this iterative process enhances the leader's capability to acquire more accurate knowledge regarding the true parameter θ_0 in game-theoretic strategic interaction, with or without considering costs.

Then we present two commonly employed assumptions regarding the unknown parameter space and cost function [8,13,2,21].

Assumption 1 *The set $\mathbf{\Theta}$ is compact and convex and θ_0 is an interior point of $\mathbf{\Theta}$.*

Assumption 2 *For all $\mathbf{u}^F \in \mathbf{U}^F$, the function $p(\mathbf{u}^F \mid \mathbf{u}^L, \theta)$ is continuously differentiable with respect to both $\theta \in \mathbf{\Theta}$ and $\mathbf{u}^L \in \mathbf{U}^L$. Furthermore, it is twice continuously differentiable with respect to $\theta \in \mathbf{\Theta}$, and the Hessian matrix about θ is positive definite and continuous for all $\theta \in \mathbf{\Theta}$.*

For simplicity, when $\log p(\mathbf{u}^F \mid \mathbf{u}^L, \theta) \neq 0$, we define

$$\Phi(\mathbf{u}^F; \mathbf{u}^L, \theta) = \log p(\mathbf{u}^F \mid \mathbf{u}^L, \theta).$$

According to Assumption 2, for all $\mathbf{u}^F \in \mathbf{U}^F$, the functions Φ is continuously differentiable with respect to both $\theta \in \mathbf{\Theta}$ and $\mathbf{u}^L \in \mathbf{U}^L$, and let $\dot{\Phi}(\mathbf{u}^F; \mathbf{u}^L, \theta)$ denote the gradient of Φ with respect to (w.r.t.) θ. Furthermore, it is twice continuously differentiable w.r.t. $\theta \in \mathbf{\Theta}$, and the Hessian matrix w.r.t. θ of Φ denoted by $\ddot{\Phi}(\mathbf{u}^F; \mathbf{u}^L, \theta)$ is continuous for all $\theta \in \mathbf{\Theta}$.

The following assumption is the standard condition for the asymptotic normality of MLE [34,15,38].

Assumption 3 *Given $\mathbf{u}^L \in \mathbf{U}^L$ and $\theta \in \mathbf{\Theta}$, the following properties hold:*

$$\frac{\partial}{\partial \theta} \int_{\mathbf{U}^F} p(\mathbf{u}^F \mid \mathbf{u}^L, \theta) d\mathbf{u}^F = \int_{\mathbf{U}^F} \frac{\partial p(\mathbf{u}^F \mid \mathbf{u}^L, \theta)}{\partial \theta} d\mathbf{u}^F,$$

and

$$\frac{\partial^2 \int_{\mathbf{U}^F} p(\mathbf{u}^F \mid \mathbf{u}^L, \theta) d\mathbf{u}^F}{\partial \theta \partial \theta^\top} = \int_{\mathbf{U}^F} \frac{\partial^2 p(\mathbf{u}^F \mid \mathbf{u}^L, \theta)}{\partial \theta \partial \theta^\top} d\mathbf{u}^F.$$

Assumption 3 ensures that the following relationships hold (see [15, Theorem 2]): Given $\mathbf{u}^L \in \mathbf{U}^L$ and $\theta \in \mathbf{\Theta}$,

$$\begin{aligned} \mathbb{E}_{\mathbf{u}^F \mid \mathbf{u}^L, \theta} \left[\dot{\Phi}(\mathbf{u}^F; \mathbf{u}^L, \theta) \right] &= 0, \\ \mathbb{E}_{\mathbf{u}^F \mid \mathbf{u}^L, \theta} \left[\ddot{\Phi}(\mathbf{u}^F; \mathbf{u}^L, \theta) \right] &= -\mathbb{E}_{\mathbf{u}^F \mid \mathbf{u}^L, \theta} \left[\ddot{\Phi}(\mathbf{u}^F; \mathbf{u}^L, \theta) \right]. \end{aligned}$$

3 Active learning for Stackelberg games

In this section, we only focus on learning by designing an active learning algorithm for Stackelberg games based on Fisher information.

3.1 Query strategy based on Fisher information

The Fisher information quantifies the amount of information that an observable random variable X carries about an unknown parameter θ_0 on which the probability of X depends [35,28]. Let $p(X \mid \theta_0)$ denote the probability density function of X conditioned on the value of θ_0. Mathematically, the Fisher information $\mathcal{F}(\theta_0)$ is expressed as (see also[17, pp. 70–71])

$$\mathbb{E}_{X|\theta_0} \left[\left(\nabla_{\theta_0} \log p(X \mid \theta_0) \right) \left(\nabla_{\theta_0} \log p(X \mid \theta_0) \right)^\top \right].$$

In the Stackelberg game described in Section 2, once the leader obtains an estimate $\hat{\theta}$ for the unknown parameter θ, it aims to devise a query strategy that maximizes the acquisition of informative data. To achieve this goal, leveraging the Fisher information proves to be a promising approach. Accordingly, we give an observation information matrix for the leader in the following definition, which is computed using the estimate $\hat{\theta}$ instead of the true parameter θ_0 [19,47,46].

Definition 1 *Given the estimate $\hat{\theta} \in \mathbf{\Theta}$, the observation information matrix (OIM) $\mathcal{F}(\mathbf{u}^L \mid \hat{\theta}) \in \mathbb{R}^{m \times m}$ is defined as*

$$\mathcal{F}(\mathbf{u}^L \mid \hat{\theta}) \triangleq \mathbb{E}_{\mathbf{u}^F \mid \mathbf{u}^L, \hat{\theta}} \left[\dot{\Phi}(\mathbf{u}^F; \mathbf{u}^L, \hat{\theta}) \dot{\Phi}(\mathbf{u}^F; \mathbf{u}^L, \hat{\theta})^\top \right]. \tag{7}$$

Given the current estimate $\hat{\theta}$, we define a query function $\mathcal{H}(\mathbf{u}^L \mid \hat{\theta}) : \mathbb{R}^n \to \mathbb{R}$, where the leader seeks a strategy $\mathbf{u}^L \in \mathbf{U}^L$ that maximizes this function. In the relevant literature, the following representations of $\mathcal{H}$ are commonly encountered:

- $\mathcal{H}\left(\mathbf{u}^L \mid \hat{\theta}\right) = \operatorname{tr}\left[\mathcal{F}(\mathbf{u}^L \mid \hat{\theta})\right]$, which is known as A-optimality criterion. This criterion aims to maximize the trace of OIM, which is equivalent to minimizing the total parameter variance [23,14,32].
- $\mathcal{H}\left(\mathbf{u}^L \mid \hat{\theta}\right) = \left(\det\left[\mathcal{F}(\mathbf{u}^L \mid \hat{\theta})\right]\right)^{\frac{1}{m}}$, which is known as the D-optimality criterion in experimental design. This criterion seeks to maximize the determinant of OIM, thereby effectively reducing the volume of the joint confidence region for the estimated parameters [42,14,32].

- $\mathcal{H}\left(\mathbf{u}^L \mid \hat{\theta}\right) = \min \text{eig}\left[\mathcal{F}(\mathbf{u}^L \mid \hat{\theta})\right]$, which is known as the E-optimality criterion. The criterion aims to maximize the minimum eigenvalue of OIM, thereby reducing uncertainties in the most adverse direction within the parameter space [32,4].

Remark 1 *Under Assumption 2, the continuity in $\mathbf{u}^L$ of these query functions with A-optimality and D-optimality criteria is inherent, because they only perform basic arithmetic operations on elements of $\dot{\Phi}$. For the E-optimality criterion, continuity can be assured through Weyl's theorem on eigenvalue perturbation [40].*

3.2 Active learning based on Fisher information

After the t-th iteration of repeated games, the leader has gathered data $\mathbf{D}(T)$ and proceeds to refine its estimate $\hat{\theta}(t)$ by maximizing the log-likelihood function $l_t(\theta) : \mathbb{R}^m \to \mathbb{R}$ over the parameter space $\mathbf{\Theta}$ [33]. The log-likelihood function is defined as

$$l_t(\theta) \triangleq \frac{1}{t}\sum_{k=1}^{t} \Phi(\mathbf{u}^F(k); \mathbf{u}^L(k), \theta). \tag{8}$$

The parameter estimation method above is the maximum likelihood estimation (MLE), and its properties will be analyzed below. For convenience, we temporarily denote $p(\mathbf{u}^F \mid \mathbf{u}^L, \theta)$ as p, $Z(\mathbf{u}^L, \theta)$ as Z, and $J^F(\mathbf{u}^F, \mathbf{u}^L; \theta)$ as J^F, which are defined in (4). When J^F is continuously differentiable with respect to θ, define

$$\begin{aligned}&\text{Cov}\left(\frac{\partial J^F}{\partial \theta}\right)\\ =&\mathbb{E}\left[\left(\frac{\partial J^F}{\partial \theta} - \mathbb{E}\left[\frac{\partial J^F}{\partial \theta}\right]\right)\left(\frac{\partial J^F}{\partial \theta} - \mathbb{E}\left[\frac{\partial J^F}{\partial \theta}\right]\right)^T\right],\end{aligned}$$

which is positive definite.

The following assumption is satisfied when J^F possesses certain properties duo to the the Leibniz Rule in general forms [24, Theorem 3.2]. To simplify the assumptions, we only give the following assumption.

Assumption 4 *Given $\mathbf{u}^L \in \mathbf{U}^L$ and $\theta \in \mathbf{\Theta}$, the following properties hold:*

$$\frac{\partial}{\partial \theta}\int_{\mathbf{U}^F} \exp\{-\lambda J^F\}\, d\mathbf{u}^F = \int_{\mathbf{U}^F} \frac{\partial}{\partial \theta}\exp\{-\lambda J^F\}\, d\mathbf{u}^F,$$

$$\frac{\partial \int_{\mathbf{U}^F} \exp\{-\lambda J^F\}\frac{\partial J^F}{\partial \theta} d\mathbf{u}^F}{\partial \theta^\top} = \int_{\mathbf{U}^F} \frac{\partial \exp\{-\lambda J^F\}\frac{\partial J^F}{\partial \theta}}{\partial \theta^\top} d\mathbf{u}^F.$$

In certain scenarios, such as when J^F is linear with respect to θ, the log-likelihood function is concave, as shown in the following proposition.

Proposition 1 *Under Assumptions 2 and 4, the log-likelihood function (8) is concave if the term*

$$-\lambda\left(\frac{\partial^2 J^F}{\partial\theta\partial\theta^T} - \mathbb{E}\left[\frac{\partial^2 J^F}{\partial\theta\partial\theta^T}\right]\right) - \lambda^2 Cov\left(\frac{\partial J^F}{\partial\theta}\right)$$

is negative definite. In particular, this property holds true when J^F is linear with respect to θ.

PROOF. To prove the concavity of the log-likelihood function, we start by reviewing (4). We have

$$\log p = -\lambda J^F - \log(Z).$$

According to Assumption 2, $\log p$ is twice continuously differentiable with respect to θ. Next, we will compute the second derivative of $\log p$ with respect to θ.

First, we compute the first derivative of $\log p$ with respect to θ:

$$\frac{\partial \log p}{\partial \theta} = -\lambda\frac{\partial J^F}{\partial \theta} - \frac{1}{Z}\frac{\partial Z}{\partial \theta}.$$

Next, we compute the derivative of Z:

$$\begin{aligned}\frac{\partial Z}{\partial \theta} &= \int \frac{\partial}{\partial \theta}\exp\{-\lambda J^F\}\, d\mathbf{u}^F\\ &= -\lambda\int \exp\{-\lambda J^F\}\frac{\partial J^F}{\partial \theta}\, d\mathbf{u}^F,\end{aligned}$$

where the first equality holds by Assumption 4. Therefore, we have

$$\frac{\partial \log p}{\partial \theta} = -\lambda\frac{\partial J^F}{\partial \theta} + \lambda\frac{\int \exp\{-\lambda J^F\}\frac{\partial J^F}{\partial \theta}\, d\mathbf{u}^F}{Z}.$$

Then we compute the second derivative of $\log p$:

$$\begin{aligned}&\frac{\partial^2 \log p}{\partial\theta\partial\theta^T} = -\lambda\frac{\partial^2 J^F}{\partial\theta\partial\theta^T}\\ &+\lambda\frac{1}{Z^2}\left(Z\cdot\frac{\partial}{\partial\theta^T}\int_{\mathbf{U}^F}\exp\{-\lambda J^F\}\frac{\partial J^F}{\partial\theta}\, d\mathbf{u}^F\right.\\ &\left.-\frac{\partial Z}{\partial\theta^T}\cdot\int_{\mathbf{U}^F}\exp\{-\lambda J^F\}\frac{\partial J^F}{\partial\theta}\, d\mathbf{u}^F\right)\\ =&-\lambda\frac{\partial^2 J^F}{\partial\theta\partial\theta^T} + \lambda\mathbb{E}\left[\frac{\partial^2 J^F}{\partial\theta\partial\theta^T}\right] - \lambda^2\mathbb{E}\left[\frac{\partial J^F}{\partial\theta}\frac{\partial J^F}{\partial\theta^T}\right]\\ &+\lambda^2\mathbb{E}\left[\frac{\partial J^F}{\partial\theta}\right]\mathbb{E}\left[\frac{\partial J^F}{\partial\theta^T}\right]\\ =&-\lambda\left(\frac{\partial^2 J^F}{\partial\theta\partial\theta^T} - \mathbb{E}\left[\frac{\partial^2 J^F}{\partial\theta\partial\theta^T}\right]\right) - \lambda^2\text{Cov}\left(\frac{\partial J^F}{\partial\theta}\right).\end{aligned}$$

where the second equality holds by Assumption 4. If $\frac{\partial^2 \log p}{\partial\theta\partial\theta^T}$ is negative definite, the log-likelihood function (8) is concave.

When J^F is linear with respect to θ, we have $\frac{\partial^2 J^F}{\partial\theta\partial\theta^T} = \mathbf{0}$, then

$$\frac{\partial^2 \log p}{\partial\theta\partial\theta^T} = -\lambda^2 \mathrm{Cov}\left(\frac{\partial J^F}{\partial\theta}\right).$$

Thus, $\frac{\partial^2 \log p}{\partial\theta\partial\theta^T}$ is negative definite. Therefore, the log-likelihood function is concave. ■

Due to the concavity of the function in Proposition 1 and the convexity of the set $\mathbf{\Theta}$ in Assumption 1, MLE is a convex optimization.

Then we present a common assumption from the literature on parameter identification that ensures the identifiability of θ_0 [5,33].

Assumption 5 (Identifiability) *A necessary and sufficient condition for $\theta \neq \theta_0$ within the parameter space $\mathbf{\Theta}$ is that $p(\cdot|\mathbf{u}^L, \theta) \neq p(\cdot|\mathbf{u}^L, \theta_0)$ for all $\mathbf{u}^L \in \mathbf{U}^L$.*

From the definition of $p(\cdot \mid \mathbf{u}^L, \theta)$, identifiability in Assumption 5 means that if $\theta \neq \theta_0$, $J^F(\cdot, \mathbf{u}^L; \theta) \neq J^F(\cdot, \mathbf{u}^L; \theta_0)$ for all $\mathbf{u}^L \in \mathbf{U}^L$.

When the leader primarily focuses on learning the follower's parameters without incurring any cost, upon updating the estimate $\hat{\theta}$ using MLE, the leader can achieve this goal by choosing the action

$$\mathbf{u}^{L,\circ} \triangleq \arg\max_{\mathbf{u}^L \in \mathbf{U}^L} \mathcal{H}(\mathbf{u}^L \mid \hat{\theta}), \tag{9}$$

which maximizes the Fisher information criterion given in Section 3.1. This ensures the most informative data acquisition regarding the follower's parameters. Consequently, the active learning algorithm based on Fisher information is outlined in Algorithm 1.

3.3 Consistency and asymptotic normality of Alg. 1

We will give the consistency and asymptotic normality of Algorithm 1. First of all, some boundedness assumptions are given to ensure stability and tractability in models involving bounded rationality.

Assumption 6 *Given $\mathbf{u}^L \in \mathbf{U}^L$ and $\theta \in \mathbf{\Theta}$,*

(1) $\mathrm{Var}_{\mathbf{u}^F|\mathbf{u}^L,\theta}[J^F(\mathbf{u}^F, \mathbf{u}^L; \theta)]$ *is bounded.*

(2) $\mathbb{E}_{\mathbf{u}^F|\mathbf{u}^L,\theta}\|\dot{\Phi}(\mathbf{u}^F; \mathbf{u}^L, \theta)\|^3$ *is finite.*

(3) There exist $\epsilon > 0$ and random variables $B_{ij}(\mathbf{u}^F)$ such that $\sup\{|\ddot{\Phi}_{i,j}(\mathbf{u}^F; \mathbf{u}^L, \theta)| : ||(\theta^\top, \mathbf{u}^{L\top}) - (\theta_0^\top, \mathbf{u}_0^{L\top})|| \leq \epsilon\} \leq B_{ij}(\mathbf{u}^F)$ and $\mathbb{E}|B_{ij}(\mathbf{u}^F)|^{1+\delta}$ is bounded.

Algorithm 1 Active learning algorithm

Require: $T, \lambda, \hat{\theta}(0)$;

1: **for** $t = 1, 2, \ldots, T$ **do**
2: The leader chooses its action $\mathbf{u}^{L,\circ}(t)$ under estimate $\hat{\theta}(t-1)$ by using (9).
3: The follower observes the leader's action $\mathbf{u}^{L,\circ}(t)$, and updates its action $\mathbf{u}^F(t)$ according to (5).
4: The leader obverses the response of the follower $\mathbf{u}^F(t)$.
5: The leader gathers the data $\mathbf{D}(t) = \{(\mathbf{u}^{L,\circ}(k), \mathbf{u}^F(k))\}_{k=1}^t$.
6: The leader then gives an estimate $\hat{\theta}(t)$ of the follower's true parameter θ_0 by MLE, where $\hat{\theta}(t)$ maximizes l_t in (8) under data $\mathbf{D}(t)$.
7: **end for**

Ensure: $\hat{\theta}(T)$

(4) $\mathbb{E}_{\mathbf{u}^F|\mathbf{u}^L,\theta}\left[\max_{\theta\in\mathbf{\Theta}} |(\dot{\Phi}^\top \dot{\Phi})_{i,j}|\right]$ *is finite for all $\mathbf{u}^F \in \mathbf{U}^F$.*

These assumptions ensure stability and tractability by bounding variances, gradients, and higher-order terms. Bounded variance limits stochastic fluctuations of the cost function, while the bounded third moment of gradients controls extreme values. Assumptions 6. (3) and (4) further ensure the feasibility of parameterized systems. Relevant boundedness assumptions are widely used in the literature to ensure stability and tractability in models with bounded rationality [10,6].

The following theorem demonstrates that the estimates obtained by Algorithm 1 converge to the true parameters θ_0 in probability. Before giving the theorem, we review the notations $\mathbf{D}(T) = \{\mathbf{u}^{L,\circ}(t), \mathbf{u}^F(t))\}_{t=1}^T$, $\mathbf{D}^L(T) = \{\mathbf{u}^{L,\circ}(t)\}_{t=1}^T$, and define

$$\begin{aligned} l_T(\theta|\mathbf{D}(T)) &\triangleq \frac{1}{T}\sum_{t=1}^T \log p(\mathbf{u}^F(t) \mid \mathbf{u}^L(t), \theta), \\ L_T(\theta|\mathbf{D}^L(T)) &\triangleq \frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\mathbf{u}^F|\mathbf{u}^L(t),\theta_0}\left[\log p(\mathbf{u}^F \mid \mathbf{u}^L(t), \theta)\right]. \end{aligned} \tag{10}$$

Theorem 1 (Consistency of Alg. 1) *Under Assumptions 1, 5 and 6, the estimates $\hat{\theta}(T)$ obtained from Algorithm 1 exhibit consistency in probability. Namely, for any $\epsilon > 0$,*

$$\lim_{T\to\infty} \mathbb{P}\left\{\|\hat{\theta}(T) - \theta_0\| \geq \epsilon\right\} = 0. \tag{11}$$

which is denoted as $\hat{\theta}(T) \xrightarrow{P} \theta_0$.

Proof Outline. The proof of this theorem is detailed in Section 5, which consists of the following two steps:

Step 1: Using the law of large numbers to prove that

$$\lim_{T\to\infty} \mathbb{P}\left\{\left|l_T(\hat{\theta}(T)|\mathbf{D}(T)) - L_T(\theta_0|\mathbf{D}^L(T))\right| \geq \epsilon\right\} = 0.$$

Step 2: Using proof by contradiction to get the final conclusion. ■

In order to analyze the active action $\mathbf{u}^{L,\circ}$, we define

$$\mathbf{u}^{L\circ\circ} \in \arg\max_{\mathbf{u}^L\in\mathbf{U}^L} \mathcal{H}(\mathbf{u}^L \mid \theta_0). \tag{12}$$

The following theorem gives the convergence of the active action obtained by Algorithm 1.

Theorem 2 *Under Assumptions 1, 2, 5 and 6, if the maximum $\mathbf{u}^{L\circ\circ}$ is unique, we have*

$$\mathbf{u}^{L\circ}(T) \xrightarrow{P} \mathbf{u}^{L\circ\circ}, \quad as \quad T\to\infty.$$

PROOF. Firstly, we have

$$\begin{aligned}
0 \leq& \mathcal{H}\left(\mathbf{u}^{L\circ\circ} \mid \theta_{\mathbf{0}}\right) - \mathcal{H}\left(\mathbf{u}^{L\circ}(T) \mid \theta_{\mathbf{0}}\right)\\
=& \mathcal{H}\left(\mathbf{u}^{L\circ\circ} \mid \theta_{\mathbf{0}}\right) - \mathcal{H}\left(\mathbf{u}^{L\circ\circ} \mid \hat{\theta}(T)\right)\\
&+ \mathcal{H}\left(\mathbf{u}^{L\circ\circ} \mid \hat{\theta}(T)\right) - \mathcal{H}\left(\mathbf{u}^{L\circ}(T) \mid \hat{\theta}(T)\right)\\
&+ \mathcal{H}\left(\mathbf{u}^{L\circ}(T) \mid \hat{\theta}(T)\right) - \mathcal{H}\left(\mathbf{u}^{L\circ}(T) \mid \theta_{\mathbf{0}}\right)\\
\leq& \mathcal{H}(\mathbf{u}^{L\circ\circ} \mid \theta_{\mathbf{0}}) - \mathcal{H}\left(\mathbf{u}^{L\circ\circ} \mid \hat{\theta}(T)\right)\\
&+ \mathcal{H}\left(\mathbf{u}^{L\circ}(T) \mid \hat{\theta}(T)\right) - \mathcal{H}\left(\mathbf{u}^{L\circ}(T) \mid \theta_{\mathbf{0}}\right),
\end{aligned}$$

where the last inequality holds because

$$\mathcal{H}\left(\mathbf{u}^{L\circ\circ} \mid \hat{\theta}(T)\right) - \mathcal{H}\left(\mathbf{u}^{L\circ}(T) \mid \hat{\theta}(T)\right) \leq 0,$$

according to equation (9) of Algorithm 1. As a consequence, we derive

$$\begin{aligned}
&\left|\mathcal{H}\left(\mathbf{u}^{L\circ\circ} \mid \theta_{\mathbf{0}}\right) - \mathcal{H}\left(\mathbf{u}^{L\circ}(T) \mid \theta_{\mathbf{0}}\right)\right|\\
\leq& \left|\mathcal{H}(\mathbf{u}^{L\circ\circ} \mid \theta_{\mathbf{0}}) - \mathcal{H}\left(\mathbf{u}^{L\circ\circ} \mid \hat{\theta}(T)\right)\right|\\
&+ \left|\mathcal{H}\left(\mathbf{u}^{L\circ}(T) \mid \hat{\theta}(T)\right) - \mathcal{H}\left(\mathbf{u}^{L\circ}(T) \mid \theta_{\mathbf{0}}\right)\right|.
\end{aligned} \tag{13}$$

Due to Assumption 2 and Assumption 6.(4) , by using Lebesgue's dominated convergence theorem [36, Theorem 1.34] we can establish that the Fisher information $\mathcal{F}(\mathbf{u}^L \mid \theta)$ defined by (7) is continuous with respect to θ for any $\mathbf{u}^L \in \mathbf{U}^L$. Additionally, $\mathcal{H}(\mathbf{u}^L \mid \theta)$ exhibits continuity at θ_0 for all $\mathbf{u}^L \in \mathbf{U}^L$ from Remark 1. Therefore, for any $\epsilon > 0$, there exists $\delta > 0$ such that for all $\theta \in S(\theta_0,\delta)\cap\mathbf{\Theta}$, it holds that $|\mathcal{H}(\mathbf{u}^L \mid \theta_{\mathbf{0}}) - \mathcal{H}(\mathbf{u}^L \mid \theta)| < \epsilon/2$ for all $\mathbf{u}^L \in \mathbf{U}^L$. Particularly, this implies

$$\begin{aligned}
&|\mathcal{H}(\mathbf{u}^{L\circ\circ} \mid \theta_{\mathbf{0}}) - \mathcal{H}(\mathbf{u}^{L\circ\circ} \mid \hat{\theta}(T))| < \tfrac{\epsilon}{2},\\
&|\mathcal{H}(\mathbf{u}^{L\circ}(T), \hat{\theta}(T)) - \mathcal{H}(\mathbf{u}^{L\circ}(T), \theta_{\mathbf{0}})| < \tfrac{\epsilon}{2},
\end{aligned}$$

for all $\hat{\theta}(T)) \in S(\theta_0,\delta)\cap\mathbf{\Theta}$.

Moreover, in conjunction with (13), it can be demonstrated that

$$\begin{aligned}
&\mathbb{P}\left\{\hat{\theta}(T) \in S(\theta_0,\delta)\cap\mathbf{\Theta}\right\}\\
\leq& \mathbb{P}\left\{\left|\mathcal{H}\left(\mathbf{u}^{L\circ\circ} \mid \theta_{\mathbf{0}}\right) - \mathcal{H}\left(\mathbf{u}^{L\circ}(T) \mid \theta_{\mathbf{0}}\right)\right| < \epsilon\right\}\\
\leq& 1.
\end{aligned}$$

Because $\hat{\theta}(T) \xrightarrow{P} \theta_0$ as $T\to\infty$ from Theorem 1, we obtain

$$\lim_{T\to\infty} \mathbb{P}\left\{\left|\mathcal{H}\left(\mathbf{u}^{L\circ\circ} \mid \theta_{\mathbf{0}}\right) - \mathcal{H}\left(\mathbf{u}^{L\circ}(T) \mid \theta_{\mathbf{0}}\right)\right| < \epsilon\right\} = 1.$$

This implies that

$$\mathcal{H}\left(\mathbf{u}^{L\circ}(T) \mid \theta_{\mathbf{0}}\right) \xrightarrow{P} \mathcal{H}\left(\mathbf{u}^{L\circ\circ} \mid \theta_{\mathbf{0}}\right) \quad \text{as} \quad T\to\infty.$$

Consequently, we deduce

$$\mathbf{u}^{L\circ}(T) \xrightarrow{P} \mathbf{u}^{L\circ\circ}, \quad \text{as} \quad T\to\infty,$$

from Remark 1 and the uniqueness of $\mathbf{u}^{L\circ\circ}$. The proof is completed. ■

To elucidate additional properties of Algorithm 1, we introduce the following notation:

$$\mathcal{F}_T(\theta) \triangleq \tfrac{1}{T}\textstyle\sum_{t=1}^{T} \mathcal{F}(\mathbf{u}^{L\circ}(t) \mid \theta). \tag{14}$$

Now, we present the following corollary concerning $\mathcal{F}_T$, which is crucial for proving Theorem 3.

Corollary 1 *Under Assumptions 1, 2, 5 and 6, if the maximum $\mathbf{u}^{L\circ\circ}$ is unique, we have that for any $\theta \in \Theta$*

$$\mathcal{F}_T(\theta) \xrightarrow{P} \mathcal{F}(\mathbf{u}^{L\circ\circ} \mid \theta). \tag{15}$$

PROOF. Because $\mathcal{F}$ is continuous w.r.t. $\mathbf{u}^F$ from Remark 1 under Assumption 2 and

$$\mathbf{u}^{L\circ}(t) \xrightarrow{P} \mathbf{u}^{L\circ\circ}, \quad \text{as} \quad t\to\infty,$$

from Theorem 2, we obtain

$$\mathcal{F}(\mathbf{u}^{L\circ}(t) \mid \theta) \xrightarrow{P} \mathcal{F}(\mathbf{u}^{L\circ\circ} \mid \theta).$$

Furthermore, it follows that $\mathcal{F}_T(\theta) \xrightarrow{P} \mathcal{F}(\mathbf{u}^{L\circ\circ} \mid \theta)$ as $T \to \infty$. Thus, the conclusion follows. ■

$\mathcal{F}(\mathbf{u}^{L\circ\circ} \mid \theta)$ is related to the asymptotic variance of the estimate, and the details will be given in the subsequent theorem. But before that, let us first introduce the following lemma.

Lemma 1 (Theorem 2 in [15]) *Let $Y_1, Y_2, \ldots$ be a sequence of independent random variables. Assume that Y_t has density $p_t(y_t \mid \theta_0)$. And the MLE of the true parameter value θ_0 is denoted by $\hat{\theta}_n \in \mathbf{\Theta}$, which maximizing the log-likelihood*

$$l_T(\theta) = \frac{1}{T}\sum_{t=1}^{n} \Phi_t(y_t, \theta),$$

like equation (8), where $\Phi_t(y_t, \theta) = \log p_t(y_t \mid \theta)$. If the following conditions 1) to 9) are satisfied, then

$$\sqrt{T}(\hat{\theta}(T) - \theta_0) \xrightarrow{L} N(0, \Gamma^{-1}(\theta_0)).$$

1). θ_0 is an interior point of $\mathbf{\Theta}$.
2). $\hat{\theta}(T) \xrightarrow{P} \theta_0$.
3). $\dot{\Phi}_t(y_t, \theta) = \frac{\partial \log p_t(y_t|\theta)}{\partial \theta}$ and $\ddot{\Phi}_t(y_t, \theta) = \frac{\partial^2 \log p_t(y_t|\theta)}{\partial \theta \partial \theta^\top}$ exist.
4). For all t, $\ddot{\Phi}_t(y_t, \theta)$ is a continuous function of θ and is a measurable function of Y_t.
5). $\mathbb{E}[\dot{\Phi}_t(y_t, \theta) \mid \theta] = 0 \quad for \quad t = 1, 2, \cdots$.
6). $\Gamma_t(\theta) \triangleq \mathbb{E}[\dot{\Phi}_t(y_t, \theta)\dot{\Phi}_t(y_t, \theta)^\top|\theta] = -\mathbb{E}[\ddot{\Phi}_t(y_t, \theta)|\theta]$, for $t = 1, 2, \cdots$.
7). For some $\delta > 0$ and all $\theta \in S(\theta_0, \delta) \cap \mathbf{\Theta}$, if $T^{-1}\sum_k \left[-\ddot{\Phi}_{t,ij}(y_t, \theta)\right] - T^{-1}\sum_k \Gamma_{t,ij}(\theta) \xrightarrow{P} 0$, then $n^{-1}\sum_t \left[-\ddot{\Phi}_{t,ij}(y_t, \theta)\right] \xrightarrow{P} \Gamma_{ij}(\theta)$ holds, where $\Gamma(\theta)$ is positive definite.
8). $\mathbb{E}\left\|\frac{\partial \log p_t(y|\theta)}{\partial \theta}\right\|^3$ is bounded.
9). There exist $\epsilon > 0$ and random variables $B_{t,ij}(Y_t)$ such that $\sup\{|\frac{\partial^2 \log p_t(y|\theta)}{\partial \theta_i \partial \theta_j^\top}| : ||\theta - \theta_0|| \le \epsilon\} \le B_{t,ij}(Y_t)$ and $\mathbb{E}|B_{t,ij}(Y_t)|^{1+\delta}$ is bounded.

PROOF. Among these conditions, only *1)* and *7)* are different from those in the literature [15]. However, these do not affect the proof process because of the following two points.

(1) The original condition *1)* in [15] states that 'Θ is an open subset of $\mathbb{R}^m$', implying the existence of $\eta > 0$ such that $S(\theta_0, \eta) \subset \mathbf{\Theta}$. Thus, instead of requiring 'θ_0 is an interior point of $\mathbf{\Theta}$', this condition suffices.

(2) The original condition *7)* in [15] is only used to ensure that for some $\delta > 0$ and all $\theta \in S(\theta_0, \delta) \cap \mathbf{\Theta}$, if $T^{-1}\sum_t \left[-\ddot{\Phi}_{t,ij}(y_t, \theta)\right] - T^{-1}\sum_t \Gamma_{t,ij}(\theta) \xrightarrow{P} 0$, then $T^{-1}\sum_t \left[-\ddot{\Phi}_{t,ij}(y_t, \theta)\right] \xrightarrow{P} \Gamma_{ij}(\theta)$ holds, where $\Gamma(\theta)$ is positive definite. Therefore, we directly use the conclusion to replace the original condition *7)*.

In addition to these, the proofs in [15] remain applicable with this substitution. ■

By employing Corollary 1 and Lemma 1, we establish the following theorem, affirming that the algorithm establishes the asymptotic normality of the estimate and achieves optimal estimate under a certain metric.

Theorem 3 (Asymptotic normality of Alg. 1) *Under Assumptions 1, 2, 3, 5 and 6, if the maximum $\mathbf{u}^{L\circ\circ}$ is unique, and $\mathcal{F}(\mathbf{u}^{L\circ\circ} \mid \theta)$ is positive definite, then the estimates $\hat{\theta}(T)$ obtained from Algorithm 1 converge to a normal distribution as $T \to \infty$:*

$$\sqrt{T}\left(\hat{\theta}(T) - \theta_0\right) \xrightarrow{L} N\left(0, \mathcal{F}(\mathbf{u}^{L\circ\circ}, \theta_0)^{-1}\right),$$

where $\mathbf{u}^{L\circ\circ}$ is defined in (12).

PROOF. Our proof process primarily relies on the conclusions from Lemma 1, and the conditions are verified as follows.

(A). Condition *1)* in Lemma 1 is satisfied by Assumption 1.

(B). Condition *2)* in Lemma 1 is satisfied by Theorem 1.

(C). Under Assumption 2, conditions *3)* and *4)* in Lemma 1 are satisfied.

(D). Condition *5)* and Condition *6)* in Lemma 1 are satisfied by Assumption 3.

(E). For some $\delta > 0$ and all $\theta \in S(\theta_0, \delta) \cap \mathbf{\Theta}$, assume

$$T^{-1}\sum_t \left[-\ddot{\Phi}_{t,ij}(\mathbf{u}^F(t), \theta)\right] - \mathcal{F}_T(\theta) \xrightarrow{P} 0, \tag{16}$$

where $\mathcal{F}_T(\theta)$ is defined in (14). From Corollary 1,

$$\mathcal{F}_T(\theta) \xrightarrow{P} \mathcal{F}(\mathbf{u}^{L\circ\circ} \mid \theta), \tag{17}$$

where $\mathcal{F}(\mathbf{u}^{L\circ\circ} \mid \theta)$ is positive definite. Therefore, combining equations 16 and 17, we have

$$T^{-1}\sum_t \left[-\ddot{\Phi}_{t,ij}(\mathbf{u}^F(t), \theta)\right] \xrightarrow{P} \mathcal{F}(\mathbf{u}^{L\circ\circ} \mid \theta).$$

Thus, condition *7)* in Lemma 1 is satisfied.

(F). It follows straight forwardly that conditions *8)* and *9)* hold by Assumption 6.

Therefore, from Lemma 1, the estimates $\hat{\theta}(T)$ obtained from Alg. 1 converge to a normal distribution as $T \to \infty$:

$$\sqrt{T}\left(\hat{\theta}(T) - \theta_0\right) \xrightarrow{L} N\left(0, \mathcal{F}(\mathbf{u}^{L\circ\circ}, \theta_0)^{-1}\right),$$

where $\mathbf{u}^{L\circ\circ}$ is defined in (12). Thus, the conclusion follows. ■

From the definition of $\mathbf{u}^{L\circ\circ}$ in (12), it can be concluded that the algorithm achieves a certain optimal optimality to ensure that the asymptotic variance is as small as possible in the sense of MLE.

4 Active inverse game for Stackelberg games

In this section, we focus on the scenario where the leader also considers its cost while actively learning the unknown parameter in the follower's cost function.

4.1 Balance between exploitation and exploration

In Section 3, we have investigated a scenario where the leader merely focuses on learning the follower's cost function while disregarding its own costs. However, the follower also acts as a player within the game, necessitating consideration of its own costs influenced by the strategies of other players. This interdependence between strategies complicates both exploration (attempting to discover more information about the unknown parameters by selecting actions to increase the amount of Fisher information) and exploitation (leveraging existing knowledge to achieve the locally optimal solutions) tasks. Therefore, we now explore strategies for leader to effectively balance exploration and exploitation to achieve optimal trade-offs [9].

Remark 2 *The difference from reinforcement learning (RL): The exploration-exploitation trade-off also exists in RL [20,43]. However, RL typically involves a single agent learning from an environment without explicit interaction with opponents. Different from RL agents that explore and exploit based on environmental feedback, active inverse methods incorporate game-theoretic feedback into the learning process, as the strategies of players influence each other. Moreover, active inverse methods aim to achieve equilibrium efficiently, unlike RL, which focuses solely on optimizing rewards without considering strategic equilibrium between multiple players.*

If the leader's estimate of the unknown parameter is $\hat{\theta}$, then the estimated expected cost of the action $\mathbf{u}^L$ is

$$\mathcal{J}(\mathbf{u}^L \mid \hat{\theta}) \triangleq \mathbb{E}_{\mathbf{u}^F \mid \mathbf{u}^L, \hat{\theta}}\left[J^L(\mathbf{u}^L, \mathbf{u}^F; \hat{\theta})\right], \tag{18}$$

which is based entirely on current information.

Next, we propose a query strategy that effectively balances exploitation and exploration. At the t-th iteration of repeated games, the leader selects action $\mathbf{u}^{L\diamond}(t)$ by solving the following optimization problem:

$$\min_{\mathbf{u}^L \in \mathbf{U}^L} \underbrace{\mathcal{J}(\mathbf{u}^L \mid \hat{\theta}(t))}_{\text{exploitation}} - \underbrace{\rho_t \mathcal{H}(\mathbf{u}^L \mid \hat{\theta}(t))}_{\text{exploration}}, \tag{19}$$

where ρ_t is a carefully chosen coefficient that balances exploitation and exploration, as detailed later. In (19), the *exploitation* term represents the expected cost of the leader's action based on current knowledge, aiming to achieve locally optimal results known so far; the *exploration* term utilizes the Fisher information associated with the leader's action, as discussed in Section 3, to discover more new information regarding the follower. Because $\mathcal{H}$ is always greater than 0, we set $\rho_t \to 0$ as $t \to \infty$ to reduce the focus on learning the follower's information after a certain level of learning.

Except for the query strategy, the process is the same as Algorithm 1 in Section 4. The algorithm of active inverse game for Stackelberg games, which balances cost and learning, is summarized in Algorithm 2.

Algorithm 2 Active inverse game algorithm

Require: $T, \lambda, \hat{\theta}(0)$;
1: **for** $t = 1, 2, \ldots, T$ **do**
2: The leader computes the coefficient ρ_t.
3: The leader chooses its action $\mathbf{u}^{L\diamond}(t)$ under estimate $\hat{\theta}(t-1)$ by using (19).
4: The leader broadcasts $\mathbf{u}^{L\diamond}(t)$ to the follower.
5: The follower obverses the response of the follower $\mathbf{u}^F(t)$.
6: The leader gathers the data $\mathbf{D}(t) = \{(\mathbf{u}^{L\diamond}(k), \mathbf{u}^F(k))\}_{k=1}^t$.
7: The leader then finds the $\hat{\theta}(t)$ by MLE, where $\hat{\theta}(t)$ maximizes l_t in (8) under data $\mathbf{D}(t)$.
8: **end for**
Ensure: $\hat{\theta}(T)$

Remark 3 *Specifically, we give an example of the coefficient design which has good experimental results. Set*

$$\rho_t \triangleq \mu_t \sigma\left(\alpha\left(\|\hat{\theta}(t) - \hat{\theta}(t-1)\|_2 - \eta\right)\right),$$

where μ_t is a non-negative decreasing sequence such that $\mu_t \to 0$ as $t \to \infty$, and $\sigma(x) = \frac{1}{1+e^{-x}}$ denotes the stan-

dard sigmoid function. Here, $\alpha > 0$ is a large enough constant, and $\eta > 0$ is a small threshold. Thus,

(1) If $\|\hat{\theta}(t) - \hat{\theta}(t-1)\|_2 > \eta$,

$$\sigma\left(\alpha\left(\|\hat{\theta}(t) - \hat{\theta}(t-1)\|_2 - \eta\right)\right) \approx 1$$

duo to the large enough constant α. Then $\rho_t \approx \mu_t$;
(2) Similarly, if $\|\hat{\theta}(t) - \hat{\theta}(t-1)\|_2 < \eta$, then $\rho_t \approx 0$;
(3) As t becomes sufficiently large, $\rho_t \approx 0$.

This formulation is guided by the following rationale:

(1) In the early stages of interaction, when information about the parameters is limited, the estimation of expected costs may be inaccurate, necessitating an emphasis on exploration efficiency.
(2) The convergence rate of the variance in MLE parameters is known to be $O(\frac{1}{t})$. As the number of interactions increases, parameter learning variance improves. When the difference between parameter estimates is small, i.e., $\|\hat{\theta}(t) - \hat{\theta}(t-1)\|_2 < \eta$, the strategy should focus more on exploitation to minimize estimated costs effectively.

The following theorem states that the estimates obtained by Algorithm 2 converge in probability to the true parameters θ_0.

Theorem 4 (Consistency of Alg. 2) *Under Assumptions 1, 5 and 6, the estimates $\hat{\theta}(T)$ obtained from Algorithm 2 exhibit consistency in probability, denoted as $\hat{\theta}(T) \xrightarrow{P} \theta_0$.*

PROOF. The proof process is similar to Theorem 1, and the detailed process will be given in Section 5. ■

4.2 Convergence to Stackelberg equilibrium

In this part, we will introduce the consistency of actions under the quadratic cost function. The cost functions of the leader and follower are specified as:

$$\begin{aligned} J^L(\mathbf{u}^L, \mathbf{u}^F) =& \frac{1}{2}\mathbf{u}^{L\top}\mathbf{Q}^L\mathbf{u}^L + \mathbf{u}^{F\top}\mathbf{R}_1^L\mathbf{u}^L \\ &+ \frac{1}{2}\mathbf{u}^{F\top}\mathbf{R}_2^L\mathbf{u}^F, \\ J^F(\mathbf{u}^F, \mathbf{u}^L, \theta_0) =& \frac{1}{2}\mathbf{u}^{F\top}\mathbf{Q}(\theta_0)\mathbf{u}^F + \mathbf{u}^{L\top}\mathbf{R}_1^F\mathbf{u}^F \\ &+ \frac{1}{2}\mathbf{u}^{L\top}\mathbf{R}_2^F\mathbf{u}^L, \end{aligned} \tag{20}$$

where $\mathbf{Q}^L$, $\mathbf{Q}(\theta_0)$ are symmetric positive definite matrices, and $\mathbf{R}_2^L$, $\mathbf{R}_2^F$ are symmetric matrices. This game framework has many potential applications, such as modeling public goods games in [12]. The matrix $\mathbf{Q}(\theta_0)$ depends on the unknown parameter θ, with θ_0 representing its true value.

Therefore, for the quadratic game shown in (20), the probability function (4) can be refined to

$$\begin{aligned} p(\mathbf{u}^F \mid \mathbf{u}^L, \theta) =& \frac{1}{Z(\mathbf{u}^L, \theta)} \exp\{-\lambda J^F(\mathbf{u}^F, \mathbf{u}^L, \theta)\} \\ =& (2\pi)^{-h/2} \det(\mathbf{\Sigma}(\theta))^{-1/2} \\ & \cdot \exp\left\{-\frac{1}{2}(\mathbf{u}^F - \mu(\theta))^{\mathrm{T}}\mathbf{\Sigma}(\theta)^{-1}(\mathbf{u}^F - \mu(\theta))\right\} \end{aligned} \tag{21}$$

where $\mu(\theta) \triangleq -\mathbf{Q}^{-1}(\theta)\mathbf{R}_1^F\mathbf{u}^L$ and $\mathbf{\Sigma}(\theta) \triangleq \frac{1}{\lambda}\mathbf{Q}(\theta)^{-1}$. Thus, $\mathbf{u}^F$ follows a multivariate normal distribution with mean $\mu(\theta)$ and covariance $\mathbf{\Sigma}(\theta)$ in the quadratic games shown in (20). Furthermore, Assumption 5 is simplified to: *The sufficient and necessary condition for $\theta \neq \theta_0$ within the parameter space $\mathbf{\Theta}$ is that $Q(\theta) \neq Q(\theta_0)$ for all $\mathbf{u}^L \in \mathbf{U}^L$.*

First of all, in order to analyze the properties of the function $\mathcal{J}$ in (18), we define

$$\begin{aligned} \mathbf{C}(\theta) \triangleq& \mathbf{Q}^L + 2\mathbf{R}_1^{F\top}\mathbf{Q}(\theta)^{-1}\mathbf{R}_2^L\mathbf{Q}(\theta)^{-1}\mathbf{R}_1^F \\ & - \mathbf{R}_1^{F\top}\mathbf{Q}(\theta)^{-1}\mathbf{R}_1^L - \mathbf{R}_1^{L\top}\mathbf{Q}(\theta)^{-1}\mathbf{R}_1^F. \end{aligned} \tag{22}$$

Next, we give an assumption about the continuity of function $\mathbf{Q}(\theta)$ and positive definiteness of $\mathbf{C}(\theta_0)$.

Assumption 7 *$\mathbf{Q}(\theta)$ is a continuous function of θ, and $\mathbf{C}(\theta_0)$ is positive definite.*

Assumption 7 can ensure that the expected cost with respect to $\mathbf{u}^L$ is strongly convex, which will be rigorously given in the following lemma.

Lemma 2 *Under Assumption 7, there exists $\eta > 0$ such that for all $\hat{\theta} \in S(\theta_0, \eta)$, $\mathcal{J}(\mathbf{u}^L \mid \hat{\theta})$ is strongly convex with respect to $\mathbf{u}^F \in \mathbf{U}^F$.*

PROOF. From Assumption 7, the matrix $\mathbf{Q}(\theta)$ is continuous. Then the matrix expression $\mathbf{C}(\theta)$, which involves $\mathbf{Q}(\theta)$ and its inverse, is also continuous with respect to θ, since matrix operations like addition, multiplication, and inversion preserve continuity. Therefore, there exists $\eta > 0$ such that for all $\hat{\theta} \in S(\theta_0, \eta)$, the matrix $\mathbf{C}(\hat{\theta})$ is positive definite.

Note that to compute the expectation of the quadratic form $\mathbb{E}[\mathbf{x}^T\mathbf{Q}\mathbf{x}]$ for a multivariate normal random vector $\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$, we can use the formula:

$$\mathbb{E}[\mathbf{x}^T\mathbf{Q}\mathbf{x}] = \mathrm{tr}(\mathbf{Q}\mathbf{\Sigma}) + \boldsymbol{\mu}^T\mathbf{Q}\boldsymbol{\mu},$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix, which is the sum of its diagonal elements. Since $\mathbf{u}^F$ follows a normal distribution with mean $\mu(\theta) \triangleq -\mathbf{Q}^{-1}(\theta)\mathbf{R}_1^F\mathbf{u}^L$ and covariance matrix $\mathbf{\Sigma}(\theta) \triangleq \frac{1}{\lambda}\mathbf{Q}(\theta)^{-1}$, , we can conclude from (20) that

$$\begin{aligned}
&\mathcal{J}(\mathbf{u}^L \mid \hat{\theta}) = \mathbb{E}_{\mathbf{u}^F|\mathbf{u}^L,\hat{\theta}}\left[J^L\left(\mathbf{u}^F, \mathbf{u}^L, \hat{\theta}\right)\right] \\
=&\frac{1}{2}\mathbf{u}^{L\top}\mathbf{Q}^L\mathbf{u}^L + \mu\left(\hat{\theta}\right)^\top\mathbf{R}_1^L\mathbf{u}^L + \mathrm{tr}\left[\mathbf{R}_2^L\mathbf{\Sigma}\left(\hat{\theta}\right)\right] \\
&+ \mu\left(\hat{\theta}\right)^T\mathbf{R}_2^L\mu\left(\hat{\theta}\right) \\
=&\frac{1}{2}\mathbf{u}^{L\top}\left(\mathbf{Q}^L + 2\mathbf{R}_1^{F\intercal}\mathbf{Q}^{-1}\mathbf{R}_2^L\mathbf{Q}^{-1}\mathbf{R}_1^F\right)\mathbf{u}^L \\
&- \mathbf{u}^{L\top}\mathbf{R}_1^{F\intercal}\mathbf{Q}^{-1}\mathbf{R}_1^L\mathbf{u}^L + \mathrm{tr}\left[\mathbf{R}_2^L\mathbf{\Sigma}\left(\hat{\theta}\right)\right] \\
=&\frac{1}{2}\mathbf{u}^{L\top}\mathbf{C}(\hat{\theta})\mathbf{u}^L + \mathrm{tr}\left[\mathbf{R}_2^L\mathbf{\Sigma}\left(\hat{\theta}\right)\right].
\end{aligned} \tag{23}$$

Thus, we conclude that $\mathcal{J}(\mathbf{u}^L \mid \hat{\theta})$ is convex with respect to $\mathbf{u}^F$. ■

With the property of function $\mathcal{J}$ given in Lemma 2, we can obtain the asymptotic property about Stackelberg equilibrium in Algorithm 2.

Theorem 5 *Under Assumptions 1, 5, 6 and 7, the actions $\mathbf{u}^{L\diamond}(t)$ generated by Algorithm 2 converge to the Stackelberg equilibrium. Namely, as $T \to \infty$,*

$$\mathbf{u}^{L\diamond}(T) \xrightarrow{P} \mathbf{u}^{L*},$$

where $\mathbf{u}^{L*}$ represents the Stackelberg equilibrium of the leader as defined in (6).

PROOF. From Theorem 4, it follows that $\hat{\theta}(T) \xrightarrow{P} \theta_0$ as $T \to \infty$. Therefore, there exists a sufficiently large T_0 such that for $T > T_0$, ρ_T is sufficiently small, and with probability $1 - \varepsilon_T$ ($\varepsilon_T \to 0$), $\hat{\theta}(T) \in S(\theta_0, \eta)$. Furthermore, according to Lemma 2,

$$\mathcal{J}(\mathbf{u}^L \mid \hat{\theta}(T)) - \rho_T\mathcal{H}(\mathbf{u}^L \mid \hat{\theta}(T)),$$

is strongly convex with respect to $\mathbf{u}^L$, ensuring a unique minimum $\mathbf{u}^{L\diamond}(T)$. Therefore,

$$\nabla_{\mathbf{u}^L}\mathcal{J}(\mathbf{u}^{L\diamond}(T) \mid \hat{\theta}(T)) - \rho_T\nabla_{\mathbf{u}^L}\mathcal{H}(\mathbf{u}^{L\diamond}(T) \mid \hat{\theta}(T)) = 0.$$

Given $\rho_T \to 0$, it follows that

$$\lim_{T\to\infty}\nabla_{\mathbf{u}^L}\mathcal{J}(\mathbf{u}^{L\diamond}(T) \mid \hat{\theta}(T)) = 0,$$

implying $\mathbf{u}^{L\diamond}(T) \to \mathbf{u}^{L*}$. This convergence demonstrates that $\mathbf{u}^{L\diamond}(T)$ asymptotically approaches the Stackelberg equilibrium action $\mathbf{u}^{L*}$ of the leader. ■

Theorem 5 establishes the asymptotic properties of Algorithm 2 in infinite time. Additionally, Section 4.1 has discussed its properties in finite time, which will be further verified in the simulation part. Note that while the empirical results suggest that the algorithm performs well in finite time, a theoretical analysis of its finite-time behavior is not yet provided and remains an open question for future investigation.

5 Proof of consistency

In this section, we present the proof of consistency for Algorithm 1 and 2. The proofs of Theorems 1 and 4 are similar, and thus, we demonstrate them uniformly. First, we review the notations $\mathbf{D}(T) = \{(\mathbf{u}^L(t), \mathbf{u}^F(t))\}_{t=1}^T$ and $\mathbf{D}^L(T) = \{\mathbf{u}^L(t)\}_{t=1}^T$ which are obtained by either Algorithm 1 or Algorithm 2.

To establish consistency, we first introduce the following lemma.

Lemma 3 *Under the identifiability of θ_0 in Assumption 5, for all $\theta \neq \theta_0$ in $\mathbf{\Theta}$, there exists $\epsilon > 0$ such that*

$$L_T(\theta_0 \mid \mathbf{D}^L(T)) > L_T(\theta \mid \mathbf{D}^L(T)) + \epsilon,$$

where L_T is defined in (10).

PROOF. For all $\mathbf{u}^L \in \mathbf{U}^L$, consider the difference

$$\begin{aligned}
&\mathbb{E}_{\mathbf{u}^F|\mathbf{u}^L,\theta_0}\left[\log\frac{p(\mathbf{u}^F \mid \mathbf{u}^L, \theta)}{p(\mathbf{u}^F \mid \mathbf{u}^L, \theta_0)}\right] \\
=&\mathbb{E}_{\mathbf{u}^F|\mathbf{u}^L,\theta_0}\left[\log p(\mathbf{u}^F \mid \mathbf{u}^L, \theta)\right] \\
&- \mathbb{E}_{\mathbf{u}^F|\mathbf{u}^L,\theta_0}\left[\log p(\mathbf{u}^F \mid \mathbf{u}^L, \boldsymbol{\theta_0}\right]).
\end{aligned} \tag{24}$$

Under Assumption 5, we have $\frac{p(\mathbf{u}^F|\mathbf{u}^L,\theta)}{p(\mathbf{u}^F|\mathbf{u}^L,\theta_0)} \neq 1$ for all $\theta \neq \theta_0$ in $\mathbf{\Theta}$. Since $\log t < t - 1$ for any $t \neq 1$, we deduce that for all $\mathbf{u}^L \in \mathbf{U}^L$,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{u}^F|\mathbf{u}^L,\theta_0}\left[\log\frac{p(\mathbf{u}^F|\mathbf{u}^L, \theta)}{p(\mathbf{u}^F|\mathbf{u}^L, \theta_0)}\right] \\
<&\mathbb{E}_{\mathbf{u}^F|\mathbf{u}^L,\theta_0}\left[\frac{p(\mathbf{u}^F|\mathbf{u}^L, \theta)}{p(\mathbf{u}^F|\mathbf{u}^L, \theta_0)} - 1\right] \\
=&\int_{\mathbf{U}^F}\left(\frac{p(\mathbf{u}^F|\mathbf{u}^L, \theta)}{p(\mathbf{u}^F|\mathbf{u}^L, \theta_0)} - 1\right)p(\mathbf{u}^F|\mathbf{u}^L, \theta_0)d\mathbf{u}^F \\
=&\int_{\mathbf{U}^F}p(\mathbf{u}^F|\mathbf{u}^L, \theta)d\mathbf{u}^F - \int_{\mathbf{U}^F}p(\mathbf{u}^F|\mathbf{u}^L, \theta_0)d\mathbf{u}^F \\
=&1 - 1 = 0.
\end{aligned}$$

Let $\epsilon = -\inf_{\theta\in\mathbf{\Theta}} \min_{\mathbf{u^L}\in\mathbf{U}^L} \mathbb{E}_{\mathbf{u}^F|\mathbf{u}^L,\theta_0}\left[\log\frac{p(\mathbf{u}^F|\mathbf{u}^L,\theta)}{p(\mathbf{u}^F|\mathbf{u}^L,\theta_0)}\right]$. Therefore, we can derive $\epsilon > 0$ from Assumption 1. This together with (24) demonstrates that

$$\begin{aligned}
&L_T(\theta|\mathbf{D}^L(T)) - L_T(\theta_0|\mathbf{D}^L(T)) \\
=&\frac{1}{T}\sum_{t=1}^{T}\Big[\mathbb{E}_{\mathbf{u}^F|\mathbf{u}^L,\theta_0}[\log p(\mathbf{u}^F|\mathbf{u}^L(t),\theta)] \\
&- \mathbb{E}_{\mathbf{u}^F|\mathbf{u}^L,\theta_0}[\log p(\mathbf{u}^F \mid \mathbf{u}^L(t),\theta_0)]\Big] \\
\leq& -\epsilon.
\end{aligned}$$

The proof is completed. ■

The following lemma gives an intermediate result, which is one of the sufficient conditions for the Kolmogorov strong law of large numbers to hold [16].

Lemma 4 *Under Assumption 6, we have that for all* $\theta \in \mathbf{\Theta}$*,*

$$\sum_{t=1}^{\infty}\frac{1}{t^2}\mathrm{Var}_{\mathbf{u}^F|\mathbf{u}^L,\theta_0}\left[\log p(\mathbf{u}^F|\mathbf{u}^L(t),\theta)\right] < \infty.$$

PROOF. By reviewing the definition of $p(\mathbf{u}^F \mid \mathbf{u}^L(t),\theta)$ in (4), we have

$$\begin{aligned}
&\sum_{t=1}^{\infty}\frac{1}{t^2}\mathrm{Var}_{\mathbf{u}^F|\mathbf{u}^L,\theta_0}\left[\log p(\mathbf{u}^F \mid \mathbf{u}^L(t),\theta)\right] \\
=&\sum_{t=1}^{\infty}\frac{1}{t^2}\mathrm{Var}_{\mathbf{u}^F|\mathbf{u}^L,\theta_0}\left[-\lambda J^F(\mathbf{u}^F \mid \mathbf{u}^L(t),\theta)) - \log(Z(\mathbf{u}^L,\theta))\right] \\
=&\sum_{t=1}^{\infty}\frac{\lambda^2}{t^2}\mathrm{Var}_{\mathbf{u}^F|\mathbf{u}^L,\theta_0}\left[J^F(\mathbf{u}^F \mid \mathbf{u}^L(t),\theta))\right].
\end{aligned}$$

Because $\mathrm{Var}_{\mathbf{u}^F|\mathbf{u}^L,\theta_0}\left[J^F(\mathbf{u}^F\mathbf{u}^L(t),\theta))\right]$ is bounded from Assumpton 6.(1). Thus, we get

$$\sum_{t=1}^{\infty}\frac{1}{t^2}\mathrm{Var}_{\mathbf{u}^F|\mathbf{u}^L,\theta_0}\left[\log p(\mathbf{u}^F|\mathbf{u}^L(t),\theta)\right] < \infty.$$

The proof is completed. ■

Proof of Theorems 1 and 4:

Since the estimate $\hat{\theta}(T)$ is the maximizer of $l_T(\theta \mid \mathbf{D}(T))$ obtained through maximum likelihood estimation,

$$\begin{aligned}
&l_T\left(\theta_0|\mathbf{D}(T)\right) - L_T(\theta_0|\mathbf{D}^L(T)) \\
\leq& l_T(\hat{\theta}(T)|\mathbf{D}(T)) - L_T(\theta_0|\mathbf{D}^L(T)).
\end{aligned} \tag{25}$$

Additionally, from Lemma 3, we know that θ_0 is the maximizer of $L_T(\theta \mid \mathbf{D}^L(T))$. Therefore,

$$\begin{aligned}
&l_T(\hat{\theta}(T)|\mathbf{D}(T)) - L_T(\theta_0|\mathbf{D}^L(T)) \\
\leq& l_T(\hat{\theta}(T)|\mathbf{D}(T)) - L_T(\hat{\theta}(T)|\mathbf{D}^L(T)).
\end{aligned} \tag{26}$$

Taking into account both (25) and (26), we can deduce

$$\begin{aligned}
&\left|l_T(\hat{\theta}(T)|\mathbf{D}(T)) - L_T(\theta_0|\mathbf{D}^L(T))\right| \\
\leq& \max\Bigg\{\left|l_T(\hat{\theta}(T)|\mathbf{D}(T)) - L_T(\hat{\theta}(T)|\mathbf{D}^L(T))\right|, \\
&\left|l_T(\theta_0|\mathbf{D}(T)) - L_n(\theta_0|\mathbf{D}^L(T))\right|\Bigg\}.
\end{aligned}$$

Moreover, for any $\epsilon > 0$, it can be derived that the event

$$\left\{\left|l_T(\hat{\theta}(T)|\mathbf{D}(T)) - L_T(\theta_0|\mathbf{D}^L(T))\right| \geq \epsilon\right\}$$

is a subset of the event

$$\begin{aligned}
&\left\{\left|l_T(\hat{\theta}(T)|\mathbf{D}(T)) - L_T(\hat{\theta}(T)|\mathbf{D}^L(T))\right| \geq \epsilon\right\} \\
\cdot&\bigcup\left\{\left|l_T(\theta_0|\mathbf{D}(T)) - L_T(\theta_0|\mathbf{D}^L(T))\right| \geq \epsilon\right\}.
\end{aligned}$$

Therefore, we derive

$$\begin{aligned}
&\mathbb{P}\left\{\left|l_T(\hat{\theta}(T)|\mathbf{D}(T)) - L_T(\theta_0|\mathbf{D}^L(T))\right| \geq \epsilon\right\} \\
\leq&\mathbb{P}\left\{\left|l_T(\hat{\theta}(T)|\mathbf{D}(T)) - L_T(\hat{\theta}(T)|\mathbf{D}^L(T))\right| \geq \epsilon\right\} \\
&+\mathbb{P}\left\{\left|l_T(\theta_0|\mathbf{D}(T)) - L_T(\theta_0|\mathbf{D}^L(T))\right| \geq \epsilon\right\}.
\end{aligned} \tag{27}$$

Let's review the Kolmogorov strong law of large numbers [16, Theorem 2.3.10]: Let $X_i, i > 1$, be independent random variables such that $\mathbb{E}X_i = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2$ exist for every $i \geq 1$, and let $\bar{\mu}_n = n^{-1}\sum_{i=1}^{n}\mu_i$. Then if $\sum_{k\geq 1} k^{-2}\sigma_i^2 < \infty$, then $\bar{X}_n - \bar{\mu}_n \xrightarrow{P} 0$. Therefore, by reviewing the definitions of $l_T(\theta|\mathbf{D}(T))$ and $L_T(\theta|\mathbf{D}^L(T))$ in (10) and the result of Lemma 4, we can apply Kolmogorov strong law of large numbers to directly get that for any $\theta \in \mathbf{\Theta}$, as $T \to \infty$,

$$l_T(\theta|\mathbf{D}(T)) - L_T(\theta|\mathbf{D}^L(T)) \xrightarrow{P} 0. \tag{28}$$

Thus,

$$\begin{aligned}
&\lim_{T\to\infty}\mathbb{P}\left\{\left|l_T(\hat{\theta}(T)|\mathbf{D}(T)) - L_T(\hat{\theta}(T)|\mathbf{D}^L(T))\right| \geq \epsilon\right\} = 0, \\
&\lim_{T\to\infty}\mathbb{P}\left\{\left|l_T(\theta_0|\mathbf{D}(T)) - L_T(\theta_0|\mathbf{D}^L(T))\right| \geq \epsilon\right\} = 0.
\end{aligned} \tag{29}$$

Combining (27) and (29), we conclude that

$$\lim_{T\to\infty} \mathbb{P}\left\{\left|l_T(\hat{\theta}(T)|\mathbf{D}(T)) - L_T(\theta_0|\mathbf{D}^L(T))\right| \geq \epsilon\right\} = 0,$$

i.e., as $T \to \infty$, $l_T(\hat{\theta}(T)|\mathbf{D}(T)) - L_T(\theta_0|\mathbf{D}^L(T)) \xrightarrow{P} 0$. According to (28) and the convergence property of line combination of two random variables by probability, we infer that as $T \to \infty$,

$$L_T(\hat{\theta}(T)|\mathbf{D}^L(T)) - L_T(\theta_0|\mathbf{D}^L(T)) \xrightarrow{P} 0. \tag{30}$$

Then we use proof by contradiction to establish that $\hat{\theta}(T) \xrightarrow{P} \theta_0$ as $T \to \infty$. Assume that $\hat{\theta}(T) \xrightarrow{P} \theta_0$ does not hold. In other words, there exist $\epsilon_0 > 0$ and $\delta_0 > 0$ such that, for any $T > 0$, there exists $t_0 > T$ with $\mathbb{P}\left\{\|\hat{\theta}(t_0) - \theta_0\| \geq \epsilon_0\right\} > \delta_0$. From Lemma 3, we know that for any $\theta \notin S(\theta_0, \epsilon_0)$, there exists $\epsilon > 0$ such that $L_{t_0}(\theta|\mathbf{D}^L(t_0)) - L_{t_0}(\theta_0|\mathbf{D}^L(t_0)) < -\epsilon$. Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{\left|L_{t_0}(\hat{\theta}(t_0)|\mathbf{D}^L(t_0)) - L_{t_0}(\theta_0|\mathbf{D}^L(t_0))\right| \geq \epsilon\right\} \\
\geq&\mathbb{P}\left\{\hat{\theta}(t_0) \notin S(\theta_0, \epsilon_0)\right\} \\
=&\mathbb{P}\left\{\|\hat{\theta}(t_0) - \theta_0\| \geq \epsilon_0\right\} \\
>&\delta_0.
\end{aligned}$$

This contradicts with (30). Thus, the proof is completed. ■

6 Numerical Simulations

Consider a Stackelberg game involving a leader who also acts as a learner and a follower, where their cost functions are quadratic as described in (20).

We set

$$\mathbf{Q}^L = \begin{bmatrix} 41 & 2 \\ 2 & 8 \end{bmatrix}, \mathbf{R}_1^L = \begin{bmatrix} 12 & 42 \\ 13 & 1 \end{bmatrix}, \mathbf{R}_2^L = \begin{bmatrix} 400 & 34 \\ 34 & 4 \end{bmatrix},$$
$$\mathbf{Q}(\theta_0) = \begin{bmatrix} \theta_{10} & \theta_{20} \\ \theta_{20} & \theta_{30} \end{bmatrix} \text{ and } \mathbf{R}_1^F = \begin{bmatrix} 16 & 8 \\ 9 & 31 \end{bmatrix},$$

where $\mathbf{Q}(\theta_0)$ is a symmetric positive definite matrix and $\theta_0 = [\theta_{10}, \theta_{20}, \theta_{30}]$ is the unknown parameter to be estimated. The term $\mathbf{R}_2^F$ in (20) is disregarded as it does not influence the follower's decision-making.

We set the truth value of the parameter $\theta = [\theta_1, \theta_2, \theta_3]$ to $\theta_0 = [20, 10, 30]$ and its value space to $\mathbf{\Theta} = \{\theta|\theta_1 \geq \kappa, \theta_1\theta_3 - \theta_2^2 \geq \kappa\}$, where κ represents a small positive number to ensure that $\mathbf{Q}(\theta)$ is a positive definite matrix and $\mathbf{\Theta}$ is compact. Upon calculation, the eigenvalue vector of $\mathbf{C}(\theta_0)$ defined in (22) is $[261.46728326, 5.66471674]$, confirming that Assumption 7 holds.

6.1 Simulation of Alg. 1

Settings: The parameter settings are presented in Table 1. To validate the active learning for the Stackelberg game detailed in Section 3, we run 50 independent sample paths using Algorithm 1. Each experiment is configured with $T = 20$ and $\lambda = 1$. As a benchmark, we employ a uniform random strategy, where the leader selects uniformly from $\mathbf{U}^L$, defined by the box constraints in Table 1. Algorithm 1 is executed three times, optimizing for D-optimality, A-optimality, and E-optimality, respectively.

Results: The results of active learning for Stackelberg games are depicted in Fig. 2, where the horizontal axis represents the regularized estimated bias $\sqrt{T}(\hat{\theta}(T)-\theta_0)$. Each subplot displays the empirical probability density of each element of the estimate derived from samples. It is evident that the estimates obtained through Algorithm 1 exhibit a trend towards a normal distribution and consistently converge towards the true parameters. Moreover, the variance of the estimates, whether optimized for D-optimality, E-optimality, or A-optimality, is lower compared to those derived using the uniform random strategy, consistent with theoretical results.

Table 1
Parameters settings

Operational Parameters	Value
T of Alg. 1	20
Num. of sample paths for Alg. 1	50
T of Alg. 2	100
Num. of sample paths for Alg. 2	300
$\mathbf{U}^L$	$[10, 100] \times [10, 100]$
κ	1e-3
λ	1
$\mu_t = \mu_0/t$	4e7/t
η	2
α	1e3

6.2 Simulation of Alg. 2

Settings: To evaluate the active inverse game for the Stackelberg game, focusing on the exploration-exploitation trade-off, we run 300 independent sample paths for Algorithm 2. Each experiment is conducted with $T = 100$ and $\lambda = 1$, where the other parameters are listed in Table 1. Specifically, Algorithm 2 is implemented exclusively for E-optimality. As a comparison, we test a method without active exploration, which

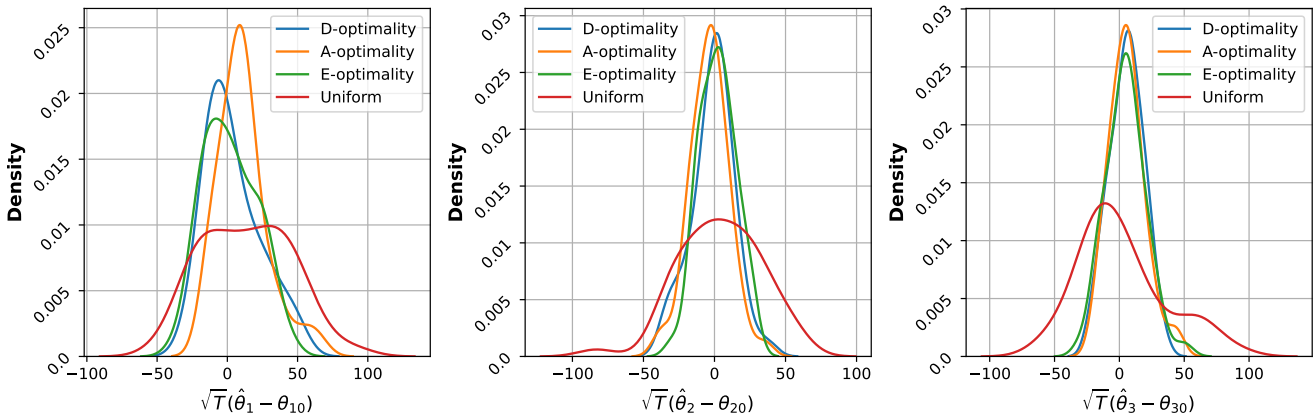

Fig. 2. Active learning for Stackelberg games with D-optimality, A-optimality and E-optimality vs. uniform random strategy.

only considers the *exploitation* term in Eq. (19). After obtaining the estimate $\hat{\theta}(t)$, the leader solves

$$\min_{\mathbf{u}^L \in \mathbf{U}^L} \mathbb{E}_{\mathbf{u}^F|\mathbf{u}^L,\hat{\theta}(t)} \left[J^L(\mathbf{u}^F, \mathbf{u}^L, \hat{\theta}(t)) \right]$$

to obtain query action.

Results: The empirical results are presented in Fig. 3 and Fig. 4. Fig. 3 illustrates the relative error $\frac{\|\mathbf{u}^L(T) - \mathbf{u}^{L*}\|}{\|\mathbf{u}^{L*}\|}$, where $\mathbf{u}^L(T)$ represents the leader's action obtained through different methods and $\mathbf{u}^{L*}$ denotes the true Stackelberg equilibrium defined in (6). It is observed that early-time performance of metrics such as the best, median, and interquartile range favor active inverse game methods over the approach without exploration. However, over time, both methods converge with random oscillations.

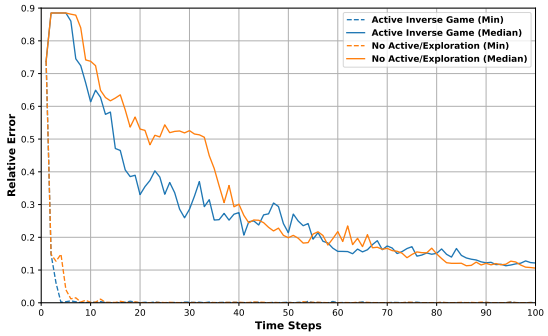

Fig. 3. The relative error $\frac{\|\mathbf{u}^L(T) - \mathbf{u}^{L*}\|}{\|\mathbf{u}^{L*}\|}$ of the active inverse game for Stackelberg games (Algorithm 2) vs. strategy without exploration. The dotted lines represent the best results (Min), and the solid line represents the median of the results (Median) of 300 experiments by methods, respectively.

Fig. 4 displays the empirical probability density plots of estimates at different time steps. At time step 20, it is evident that the variance of estimates obtained through the active method is significantly smaller than that of the inactive method, also providing an explanation for the early-stage superiority of the active method. As time progresses, when the variance diminishes sufficiently (indicating $\rho \approx 0$, where ρ is the coefficient of exploration), the exploration component in the query function diminishes, leading to comparable variances for both methods, as depicted in Fig. 4. This observation further elucidates why the relative errors of both methods converge in later time steps, as illustrated in Fig. 3.

7 Conclusion

In this paper, we have developed a framework of active inverse games. We introduced active inverse methods framework specifically in Stackelberg games with bounded rationality, where the leader, acting as a learner, strategically selects actions to better understand the follower's cost functions. We developed an active learning algorithm leveraging Fisher information to maximize information gain about the unknown parameters, ensuring consistency and asymptotic normality with minimal variance in the estimates. Additionally, when considering the leader's cost, we designed an algorithm of active inverse game for Stackelberg games that balances exploration and exploitation, ensuring consistency and asymptotic Stackelberg equilibrium in quadratic games. Finally, we experimentally validated our theoretical findings and advantages of our method through an example of quadratic game.

References

[1] Nematollah Ab Azar, Aref Shahmansoorian, and Mohsen Davoudi. From inverse optimal control to inverse reinforcement learning: A historical review. *Annual Reviews in Control*, 50:119–138, 2020.

[2] Michele Aghassi and Dimitris Bertsimas. Robust game theory. *Mathematical Programming*, 107(1):231–273, 2006.

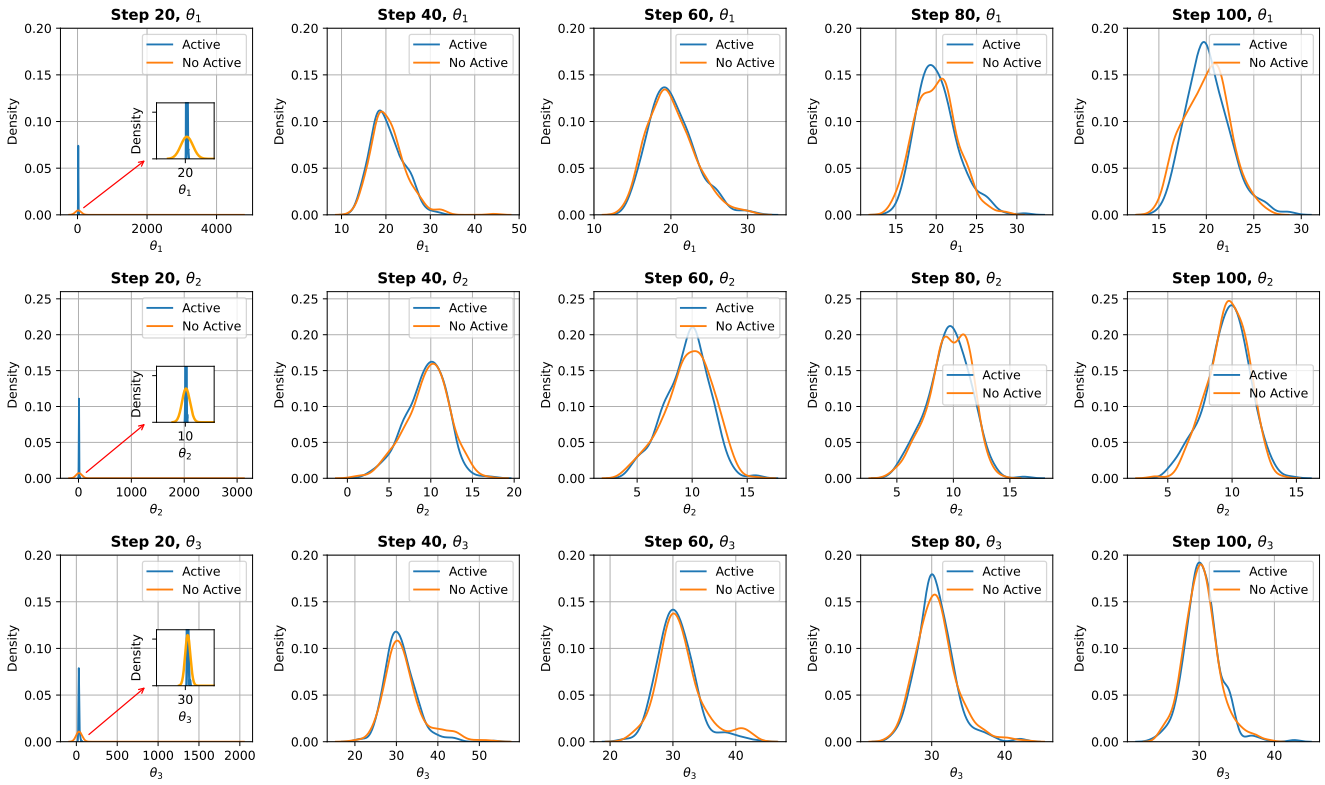

Fig. 4. Comparison of parameter estimate by active inverse methods in Stackelberg games and strategy without exploration by minimizing estimated expected cost over time at steps 20, 40, 60, 80 and 100.

[3] Robert J Aumann. Rationality and bounded rationality. *Games and Economic Behavior*, 21(1-2):2–14, 1997.

[4] Harvey Thomas Banks, Kathleen Holm, and Franz Kappel. Comparison of optimal design methods in inverse problems. *Inverse Problems*, 27(7):075002, 2011.

[5] Torsten Bohlin. On the maximum likelihood method of identification. *IBM Journal of Research and Development*, 14(1):41–51, 1970.

[6] Colin F Camerer and Ernst Fehr. When does "economic man" dominate social behavior? *science*, 311(5757):47–52, 2006.

[7] Jianguo Chen, Jinlong Lei, Hongsheng Qi, and Yiguang Hong. Online parameter identification of generalized non-cooperative game. *arXiv preprint arXiv:2310.09511*, 2023.

[8] Zhaoyang Cheng, Guanpu Chen, and Yiguang Hong. Single-leader-multiple-followers Stackelberg security game with hypergame framework. *IEEE Transactions on Information Forensics and Security*, 17:954–969, 2022.

[9] Melanie Coggan. Exploration and exploitation in reinforcement learning. *Research Supervised by Prof. Doina Precup, CRA-W DMP Project at McGill University*, 2004.

[10] Drew Fudenberg and David K Levine. *The Theory of Learning in Games*, volume 2. MIT press, 1998.

[11] Jacob K Goeree, Charles A Holt, and Thomas R Palfrey. Regular quantal response equilibrium. *Experimental Economics*, 8:347–367, 2005.

[12] Victor Gorelik and Tatiana Zolotova. Stackelberg and Nash equilibria in games with linear-quadratic payoff functions as models of public goods. In *Optimization and Applications: 12th International Conference, OPTIMA 2021, Petrovac, Montenegro, September 27–October 1, 2021, Proceedings 12*, pages 275–287. Springer, 2021.

[13] Joseph Greenberg. Avoiding tax avoidance: A (repeated) game-theoretic approach. *Journal of Economic Theory*, 32(1):1–13, 1984.

[14] Ting He, Chang Liu, Ananthram Swami, Don Towsley, Theodoros Salonidis, Andrei Iu Bejan, and Paul Yu. Fisher information-based experiment design for network tomography. *ACM SIGMETRICS Performance Evaluation Review*, 43(1):389–402, 2015.

[15] Bruce Hoadley. Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *The Annals of Mathematical Statistics*, pages 1977–1991, 1971.

[16] Jiming Jiang et al. *Large Sample Techniques for Statistics*. Springer, 2010.

[17] Don H Johnson. Statistical signal processing. *Houston: Rice University*, 2013.

[18] Ludovic A Julien. Stackelberg games. In *Handbook of Game Theory and Industrial Organization, Volume I*, pages 261–311. Edward Elgar Publishing, 2018.

[19] Yongsu Jung and Ikjin Lee. Optimal design of experiments for optimization-based model calibration using Fisher information matrix. *Reliability Engineering & System Safety*, 216:107968, 2021.

[20] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[21] Michael Kearns, Michael L Littman, and Satinder Singh. Graphical models for game theory. *arXiv preprint arXiv:1301.2281*, 2013.

[22] Volodymyr Kuleshov and Okke Schrijvers. Inverse game theory: Learning utilities in succinct games. In *Web and Internet Economics: 11th International Conference, WINE 2015, Amsterdam, The Netherlands, December 9-12, 2015, Proceedings 11*, pages 413–427. Springer, 2015.

[23] Adam Lane. Adaptive designs for optimal observed Fisher information. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):1029–1058, 2020.

[24] Serge Lang. *Undergraduate analysis*. Springer Science & Business Media, 2013.

[25] Niklas Lauffer, Mahsa Ghasemi, Abolfazl Hashemi, Yagiz Savas, and Ufuk Topcu. No-regret learning in dynamic Stackelberg games. *IEEE Transactions on Automatic Control*, 2023.

[26] Bosen Lian, Wenqian Xue, Frank L Lewis, and Tianyou Chai. Inverse reinforcement learning for multi-player noncooperative apprentice games. *Automatica*, 145:110524, 2022.

[27] Xiaomin Lin, Peter A Beling, and Randy Cogill. Multiagent inverse reinforcement learning for two-person zero-sum games. *IEEE Transactions on Games*, 10(1):56–68, 2017.

[28] Alexander Ly, Maarten Marsman, Josine Verhagen, Raoul PPP Grasman, and Eric-Jan Wagenmakers. A tutorial on Fisher information. *Journal of Mathematical Psychology*, 80:40–55, 2017.

[29] Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

[30] Timothy L Molloy, Jairo Inga, Michael Flad, Jason J Ford, Tristan Perez, and Sören Hohmann. Inverse open-loop noncooperative differential games and inverse optimal control. *IEEE Transactions on Automatic Control*, 65(2):897–904, 2019.

[31] Katja Mombaur, Anh Truong, and Jean-Paul Laumond. From human to humanoid locomotion—an inverse optimal control approach. *Autonomous Robots*, 28:369–383, 2010.

[32] Vladimir Palmin. *Fisher Information-driven Optimal Experiment Design*. PhD thesis, National Research University, 2021.

[33] Jian-Xin Pan, Kai-Tai Fang, Jian-Xin Pan, and Kai-Tai Fang. Maximum likelihood estimation. *Growth Curve Models and Statistical Diagnostics*, pages 77–158, 2002.

[34] Calyampudi Radhakrishna Rao. *Linear statistical inference and its applications*, volume 2. Wiley New York, 1973.

[35] Vijay K Rohatgi and AK Md Ehsanes Saleh. *An introduction to probability and statistics*. John Wiley & Sons, 2015.

[36] Walter Rudin. *Real and complex analysis*. McGraw-Hill, Inc., 1987.

[37] Reinhard Selten. Bounded rationality. *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für Die Gesamte Staatswissenschaft*, 146(4):649–658, 1990.

[38] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

[39] Heinrich Von Stackelberg. *Market structure and equilibrium*. Springer Science & Business Media, 2010.

[40] Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

[41] Jibang Wu, Weiran Shen, Fei Fang, and Haifeng Xu. Inverse game theory for Stackelberg games: the blessing of bounded rationality. *Advances in Neural Information Processing Systems*, 35:32186–32198, 2022.

[42] Themistoklis C Xygkis, George N Korres, and Nikolaos M Manousakis. Fisher information-based meter placement in distribution grids via the d-optimal experimental design. *IEEE Transactions on Smart Grid*, 9(2):1452–1461, 2016.

[43] Mohan Yogeswaran and SG Ponnambalam. Reinforcement learning: Exploration–exploitation dilemma in multi-agent foraging task. *Opsearch*, 49:223–236, 2012.

[44] Chengpu Yu, Yao Li, Hao Fang, and Jie Chen. System identification approach for inverse optimal control of finite-horizon linear quadratic regulators. *Automatica*, 129:109636, 2021.

[45] Yue Yu, Jonathan Salfity, David Fridovich-Keil, and Ufuk Topcu. Inverse matrix games with unique quantal response equilibrium. *IEEE Control Systems Letters*, 7:643–648, 2022.

[46] Guangyang Zeng, Biqiang Mu, Jiming Chen, Zhiguo Shi, and Junfeng Wu. Global and asymptotically efficient localization from range measurements. *IEEE Transactions on Signal Processing*, 70:5041–5057, 2022.

[47] Guangyang Zeng, Biqiang Mu, Jieqiang Wei, Wing Shing Wong, and Junfeng Wu. Localizability with range-difference measurements: Numerical computation and error bound analysis. *IEEE/ACM Transactions on Networking*, 30(5):2117–2130, 2022.