2022 IEEE 61st Conference on Decision and Control (CDC)
December 6-9, 2022. Cancún, Mexico

Inverse-Inverse Reinforcement Learning. How to Hide Strategy from an Adversarial Inverse Reinforcement Learner

Kunal Pattanayak, Vikram Krishnamurthy and Christopher Berry

2022 IEEE 61st Conference on Decision and Control (CDC) | 978-1-6654-6761-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/CDC51059.2022.9992959

Abstract*— Inverse reinforcement learning (IRL) deals with estimating an agent's utility function from its actions. In this paper, we consider how an agent can hide its strategy and mitigate an adversarial IRL attack; we call this inverse IRL (I-IRL).** ***How should the decision maker choose its response to ensure a poor reconstruction of its strategy by an adversary performing IRL to estimate the agent's strategy? **This paper comprises four results: First, we present an adversarial IRL algorithm that estimates the agent's strategy while controlling the agent's utility function. Then, we propose an I-IRL result that mitigates the IRL algorithm used by the adversary. Our I-IRL results are based on revealed preference theory in micro-economics. The key idea is for the agent to deliberately choose sub-optimal responses so that its true strategy is sufficiently masked. Third, we give a sample complexity result for our main I-IRL result when the agent has noisy estimates of the adversary-specified utility function. Finally, we illustrate our I-IRL scheme in a radar problem where a meta-cognitive radar is trying to mitigate an adversarial target.**

I. Introduction

This paper studies the interaction between two entities - a smart decision maker and an adversary that aims to estimate the plan of the decision maker; see Fig. 1 for a schematic representation. The adversary sends adversarial probes to the decision maker and controls the decision maker's utility function. In turn, the decision maker's response maximizes its utility function subject to the decision maker's budget constraint. The adversary's intent is to estimate the budget constraints of the decision maker. If the decision maker knows of the adversarial attack, how should the decision maker tweak its responses to mitigate the adversary?

We formulate this interaction between the decision maker and adversary as an *inverse-inverse reinforcement learning* problem. Reinforcement learning (RL) [1], [2] deals with learning the optimal decision strategy by observing the response to a control input. *Inverse* reinforcement learning (IRL) [3], [4], [5], [6] is the problem of reconstructing the utility function of a decision maker by observing its actions. Inverse IRL (I-IRL) is a natural extension of IRL: *If a decision maker knows that an adversary is using an IRL algorithm to reconstruct its strategy by observing its utility function, how should the decision maker deliberately tweak its response to mitigate the IRL algorithm?*

V. Krishnamurthy and K. Pattanayak are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, 14853 USA. e-mail: vikramk@cornell.edu, kp487@cornell.edu. C. Berry is with Lockheed Martin Advanced Technology Laboratories, Cherry Hill, NJ, 08002 USA. e-mail: christopher.m.berry@lmco.com. This research was supported in part by a research contract from Lockheed Martin and the Army Research Office grant W911NF-21-1-0093.

Outline and Main Results. This paper considers a revealed preference-based adversarial IRL scheme to estimate the decision maker's strategy.[1] Sec. II covers the key results from revealed preference theory in micro-economics. Revealed preference studies non-parametric detection of constrained utility maximization behavior. Theorem 1 in Sec. II presents a feasible test for identifying constrained utility maximization behavior, and generates a set-valued estimate of the decision maker's utility function. Before we address the problem of I-IRL for hiding strategy, we state Theorem 2, an IRL algorithm for estimating the strategy (budget constraint) of a decision maker when its utility function is known to the adversary. While Theorem 1 is well known in literature for estimating a utility function, Theorem 2 is new. Next, in Sec. III, we state our main result, Theorem 3. If the decision maker knows an adversary is using Theorem 2 to reconstruct, it deliberately chooses sub-optimal responses that minimally violate its strategic constraints using the I-IRL scheme of Theorem 3 to obfuscate the adversarial attack. Sec. III also presents a finite sample complexity result, Theorem 4 that upper bounds the probability that the I-IRL scheme of Theorem 3 fails when the decision maker has noisy measurements of the adversary specified utility functions. Finally, Sec. IV illustrate our I-IRL result for hiding strategy in a radar problem, wherein a cognitive radar is trying to mitigate an adversarial target.

Related Work. Our I-IRL result is based on adversarial obfuscation in machine learning. [10] provide a comprehensive list of adversarial attacks and robustness to adversarial attacks in machine learning. Our recent work [11] presents a cognition-masking scheme for a cognitive radar when the adversary has accurate measurements of the radar's response. This paper generalizes [11] in two major ways: First, we develop IRL results for estimating the decision maker's strategy followed by I-IRL result for masking strategy. Second, we analyze the performance of our I-IRL result in noisy settings via a finite sample complexity test.

This paper comprises a numerical example involving a cognitive radar trying to mitigate an adversarial target. A

[1] Revealed preference-based IRL [3], [7] is more fundamental than IRL in popular machine learning literature [5], [6], [8]. IRL in machine learning implicitly assumes the decision maker is optimal and then reconstructs its reward (utility). Revealed preference first identifies utility maximization behavior, and if so, generates a set-valued utility estimate. Indeed, one can impose additional constraints on the forward problem, and generate a more precise estimate of the decision maker's utility; one notable example being that of max-entropy IRL [8]. Another heuristic for a point-valued estimate is to extract the interior-most point from the set of feasible rewards using the concept of margins (for example, max-margin IRL [9]) which we also consider in this paper for inverse IRL.

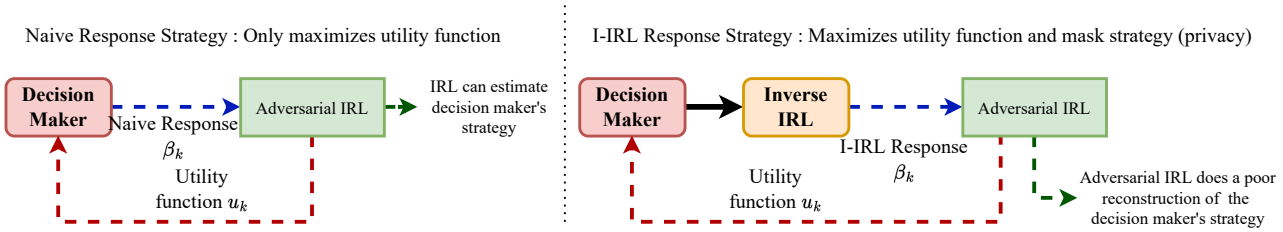

Fig. 1. Schematic of the I-IRL scheme for masking the strategy of a cognitive decision maker from adversarial IRL.
Naive response strategy (Left): The adversary sends a sequence of probe signals to the decision maker and records its responses. The probe signal parameterizes the decision maker's utility function. If the decision maker chooses the naive response that maximizes its utility function subject to its capability constraint, its capability can be estimated by the adversary using Theorem 2.
Adversarial inverse IRL strategy (Right): If the decision maker is aware that the adversary is trying to estimate its capability, the decision maker deliberately chooses sub-optimal responses via Theorem 3 to mitigate the adversary. The key idea is to ensure a poor reconstruction of the decision maker's constraint by the adversary by minimally perturbing its naive responses.

cognitive radar [12], [13], [14] uses the perception-action cycle of cognition to sense the environment and learn from it relevant information about the target and the environment. I-IRL for a cognitive radar can be viewed as a form of meta-cognition. Meta-cognition is a sophisticated form of electronic counter countermeasure (ECCM)[15], [16], [17], [18] to electronic countermeasures (ECM) in electronic warfare. However, meta-cognitive strategies involving deliberate violation of strategy to confuse the adversary's ECM have not been explored previously.

II. Background. Revealed Preference for Adversarial IRL

We start by briefly reviewing the key result in the area of revealed preference in microeconomics theory. Revealed preference studies non-parametric detection of utility maximization behavior. A utility maximizer is defined as:

Definition 1 ([19]): An agent is a *utility maximizer*[2] if for every constraint $g_k(\beta) \leq 0$, the response $\beta_k \in \mathbb{R}^m_+$ satisfies:

$$\beta_k \in \operatorname{argmax} u(\beta),\ g_k(\beta) \leq 0 \tag{1}$$

where $u(\beta)$ is a *monotone* utility function.

Definition 1 rationalizes consumer behavior in economics. The constraint $g_k(\beta) \leq 0$ in (1) is the budget faced by the consumer and β_k is the consumer's consumption vector. In the special case when $g_k(\beta)$ is linear, that is, $g_k(\beta) = \alpha_k'\beta - 1$, α_k can be interpreted as the price vector faced by the consumer; then $\alpha_k'\beta \leq 1$ is a natural budget constraint for a consumer with 1 dollar. Given a dataset of budget and consumption vectors, the aim in revealed preference is to determine if the consumer is a utility maximizer (rational) that satisfies (1). Indeed, the budget constraint $\alpha_k'\beta \leq 1$ is without loss of generality, and can be replaced by $\alpha_k'\beta \leq c$ for any positive scalar c.

[2]In machine learning literature for IRL, the decision maker typically maximizes its expected cumulative discounted reward in a Markov decision process (MDP) subject to an entropic constraint on its response (policy). Our radar-adversary interaction is a one-shot process - the adversary transmits a batch of probe signals, and then the radar responds with a batch of responses that masks its strategy. Hence, the forward optimization process for the decision maker is expressed as a utility maximization problem (1) subject to a resource constraint.

A. Adversarial IRL for Identifying Utility Function

The key result in revealed preference is Afriat's theorem [3], [7]. Afriat's theorem assumes a linear budget and specifies a set of linear inequalities that are both necessary and sufficient for a time series of constraints and responses to be consistent with utility maximization behavior (1). [19] propose a utility maximization test that generalizes Afriat's Theorem to non-linear budgets and is the key IRL algorithm used by the adversary in this paper:

Theorem 1 (Test for utility maximization [19]): Given a sequence of constraints and responses $\mathcal{D} = \{(g_k(\beta) \leq 0, \beta_k)\}_{k=1}^K$. Suppose the constraint is active at β_k ($g_k(\beta_k) = 0\ \forall k$). Then, the following statements are equivalent:

1) There exists a monotone, continuous utility function that satisfies (1).
2) There exist positive reals $\{u_t, \lambda_t\}_{t=1}^K$ such that the following inequalities are feasible:
$$u_s - u_t - \lambda_t g_t(\beta_s) \leq 0\ \forall t, s \in \{1, \ldots, K\}. \tag{2}$$
The IRL estimate of the decision maker's utility is:
$$u(\beta) = \min_{t \in \{1,2,\ldots,K\}} \{u_t + \lambda_t g_t(\beta)\} \tag{3}$$
constructed using feasible u_t and λ_t (2) rationalizes $\mathcal{D}$.
3) The data set $\mathcal{D}$ satisfies the Generalized Axiom of Revealed Preference (GARP), namely, for any $k \in \{1, 2, \ldots, K\}$, the following implication holds:
$$g_t(\beta_{t+1}) \leq g_t(\beta_t) \quad \forall t \leq k-1 \implies g_k(\beta_1) \geq g_k(\beta_k). \tag{4}$$

Theorem 1 tests for economics-based rationality; its remarkable property is that it gives a *necessary and sufficient condition* for a agent to be a utility maximizer based on the agent's input-output response. The feasibility of the set of inequalities (2) can be checked using a linear programming solver; alternatively GARP can be checked using Warshall's algorithm with $O(K^3)$ computations [20], [21]. Theorem 1 can be viewed as set-valued system identification of an *argmax* system; set-valued since (3) yields a set of utility functions that rationalize the finite dataset $\mathcal{D}$.

Key Idea for I-IRL: Manipulating the Goodness-of-fit of revealed preference test (2). Theorem 1 also constructs a set-valued estimate (3) of the utility function u using the solution

of the set of feasibility inequalities (2). The estimated utility function (3) is ordinal since any positive monotone increasing transformation of (3) also satisfies Theorem 1. We make two observations here that are crucial for our I-IRL results in Sec. III:

1. Since the feasibility of (2) is necessary for utility maximization, the scalars $u(\beta_k), \lambda_k$ satisfy the revealed preference test of (2), where λ_k solves $\lambda_k \nabla g_k(\beta_k) = \nabla u(\beta_k)$. Due to the monotonicity of u, g_k and the assumption that the constraint is active ($g_k(\beta_k) = 0\ \forall k$), λ_k is well-defined.

2. The reconstructed utility function (3) is a point-wise minimum of monotone functions parameterized by positive reals $\{u_k, \lambda_k\}$ that satisfy (2). Hence, one can at best recover a lower envelope of the true utility function u that matches the function value and gradient value at the points $\beta_k, k = 1, 2, \ldots, K$ using Theorem 1. In other words, the closest approximation u_{best} to the decision maker's utility u via the reconstruction procedure of (3) is given by:

$$u_{\text{best}}(\beta) = \min_{k \in \{1,2,\ldots,K\}} \{u(\beta_k) + \lambda_k g_k(\beta)\}, \quad \text{where } \lambda_k \nabla g_k(\beta_k) = \nabla u(\beta_k). \tag{5}$$

Also, one can show that u_{best} (5) is the least squares estimate of u:

$$\{u(\beta_k), \lambda_k\} = \underset{\bar{\lambda}_k, u_k \geq 0}{\operatorname{argmin}} \int_{\mathcal{S}} \Big(u(\beta) - \min_t \{u_t + \bar{\lambda}_t g_t(\beta)\} \Big)^2 d\beta, \tag{6}$$

for any compact set $\mathcal{S} \subseteq \mathbb{R}^K_+$, where λ_k is defined in (5).

Our key idea for I-IRL is to perturb the response sequence $\{\beta_k\}$ so that the closest IRL estimate (5) of the decision maker's system parameters passes the revealed preference test of (2) by a low margin, where the margin is defined by:

$$\mathcal{M}_u(\{\beta_k, g_k\}) = \max_{j,k} u(\beta_j) - u(\beta_k) - \lambda_k g_k(\beta_j), \tag{7}$$

where $\lambda_k \nabla g_k(\beta_k) = \nabla u(\beta_k)$. The margin (7) is a measure of goodness-of-fit [22] of the revealed preference inequalities (2). Hence, a utility function that passes (2) with a large margin is a high-confidence point utility estimate for the adversary and vice versa.

Below, we present a revealed preference test, Theorem 2, that tests for feasible budget constraints estimating the decision maker's budget constraint when its utility function is known. The aim of our key I-IRL result of Theorem 3 in Sec. III is to ensure that the closest IRL estimate of the decision maker's constraint sequence $\{g_k(\cdot)\}$ passes the revealed preference test of Theorem 2 by a low margin (7).

B. Adversarial IRL for Identifying Strategy

Theorem 1 achieves IRL when an adversarial learner wants to estimate the decision maker's utility function and knows the decision maker's budget constraint sequence (strategy). We now consider the scenario where the adversary's probes parametrize the decision maker's utility, and the adversary's aim is to estimate the unknown budget constraint sequence $\{g_k(\beta) \leq 0\}$ (strategy) of the decision maker. Below, we present Theorem 2, a revealed preference test for the existence of feasible budget constraints when the utility function and decision maker's response is observed by the adversary.

Theorem 2 (IRL for Identifying Strategy): Given a time sequence of adversary controlled utility functions and decision maker's responses $\mathcal{D} = \{(u_k, \beta_k)\}_{k=1}^K$. Suppose the decision maker faces a budget constraint of the form $g(\beta) - \gamma_k \leq 0$ for every k. Then, the following statements are equivalent:

1) There exists a sequence of monotone continuous capability constraints $\{g_k(\beta) \leq 0\}$ that satisfy (1):

$$\beta_k = \operatorname{argmax}\ u_k(\beta),\ g_k(\beta) \leq 0 \tag{8}$$

2) There exist positive reals $\{\bar{g}_k, \lambda_k\}_{k=1}^K$ such that the following inequalities are feasible:

$$\bar{g}_s - \bar{g}_t - \lambda_t\ (u_t(\beta_s) - u_t(\beta_t)) \geq 0,\ \forall t, s. \tag{9}$$

The sequence of monotone constraints $\{g(\beta) - \bar{g}_k \leq 0\}$ rationalizes $\mathcal{D}$ (1), where budget g is given by:

$$g(\beta) = \max_{t \in \{1,2,\ldots,K\}} \{\bar{g}_t + \lambda_t\ (u_t(\beta) - u_t(\beta_t))\}. \tag{10}$$

3) The data set $\{u_t(\beta_t) - u_t(\cdot), \beta_t\}$ satisfies GARP (4).

The proof of Theorem 2 is omitted for brevity; see [23] for a more elaborate discussion. At first sight, Theorem 2 appears to be a dual statement to the optimization in Theorem 1. Instead of testing for a rationalizing utility given a sequence of known budget constraints, Theorem 2 tests for a rationalizing sequence of budget constraints given the utility function and does not use duality in the proof.

In complete analogy to Theorem 1, the feasibility inequality of (9) is necessary and sufficient for the existence of a sequence of constraints that rationalizes the sequence of utility functions and responses. In complete analogy to (5), we now define g_{best}, the closest approximation (upper envelope) to the true budget g reconstructed via (9):

$$g_{\text{best}}(\beta) = \max_{k \in \{1,2,\ldots,K\}} \{\gamma_k + \lambda_k (u_k(\beta) - u_k(\beta_k))\}, \tag{11}$$

where $\lambda_k \nabla g_k(\beta_k) = \nabla u(\beta_k)$. Analogous to (7), we define the margin with which the true budget g passes the revealed preference test (9) of Theorem 2:

$$\mathcal{M}_g(\{\beta_k, u_k, \gamma_k\}) = \min_{j,k}\ g(\beta_j) - g(\beta_k) - \lambda_k\ (u_k(\beta_j) - u_k(\beta_k)),\ \text{where } \lambda_k \nabla u_k(\beta_k) = \nabla g(\beta_k). \tag{12}$$

In our I-IRL results in the next section, our key objective will be to minimally perturb the response sequence $\{\beta_k\}$ so that $\mathcal{M}_g(\cdot)$ lies below a pre-specified threshold.

Theorem 2 assumes the elements in the sequence of constraints $\{g(\beta) - \gamma_k\}$ differ only by a scalar shift. This assumption can indeed be relaxed to allow any sequence of budget constraints. But the *reconstructed* constraints (10) are restricted to the space of monotone piece-wise linear convex functions identical up to a constant. Hence, any constraint that lies outside this space is non-identifiable.

III. Inverse IRL (I-IRL) for Masking Decision Maker's Strategy

Sec. II presents IRL algorithms that an adversary uses to estimate the decision maker's strategy. If the decision maker is aware of the adversarial attack, how should it choose its responses to mask the strategy from the adversary? In Sec. III-A below, we present our main I-IRL result, Theorem 3. In Sec. III-B, we give a finite sample result for Theorem 3 that upper bounds the probability the I-IRL scheme of Theorem 3 fails when the decision maker's utility function is corrupted by additive noise.

A. Main Result. I-IRL for Adversarial IRL in Theorem 2

Theorem 3 (I-IRL for Masking Strategy): Let β_k^* denote the radar's naive response that maximizes adversary-specified utility u_k subject to constraint $g(\beta) \leq \gamma_k$ for time $k = 1, 2, \ldots, K$. Suppose the adversary uses Theorem 2 to reconstruct the decision maker's budget constraint $g(\cdot)$. Then, the I-IRL response sequence $\{\tilde{\beta}_k^*\}$ that masks $g(\cdot)$ from IRL (Theorem 2) is given by:

$$\tilde{\beta}_k^* = \text{argmax}_\beta \, u_k(\beta), \;\; g_k(\beta) \leq \gamma_k^*, \tag{13}$$

where the violated budget thresholds $\{\gamma_k^*\}$ solve the following optimization problem:

$$\{\gamma_k^*\} = \underset{\tilde{\gamma}_{1:K}}{\text{argmin}} \sum_{k=1}^{K} \|\tilde{\gamma}_k - \gamma_k\|_2^2, \tag{14}$$

$$\mathcal{M}_g(\{\tilde{\beta}_k, u_k, \tilde{\gamma}_k\}) \leq (1-\eta) \; \mathcal{M}_g(\{\beta_k^*, u_k, \gamma_k\}), \tag{15}$$

$$\tilde{\beta}_k = \text{argmax}_\beta \, u_k(\beta), \;\; g(\beta) \leq \tilde{\gamma}_k. \tag{16}$$

In (15), $\eta \in [0, 1]$ is a pre-defined scalar that parameterizes the extent of strategy masking for I-IRL.

Theorem 3 is the main I-IRL result of this paper. Simply put, the decision maker's response is the solution to the optimization problem (1) with purposefully distorted resource thresholds γ_k (1). Indeed, the decision maker's performance is degraded due to the violated constraints, but it is the price the decision maker pays for stealth - to mask its resource constraint g from adversarial IRL of Theorem 2.

Discussion.

• The I-IRL algorithm (13) computes the smallest perturbation needed in the decision maker's resource constraints that ensures a sufficiently poor resource constraint estimate (10) of the decision maker's budget constraint g (low margin of IRL feasibility test (9) parametrized by scalar η (15)). Hence, (14) computes the minimum violation that reduces the margin with the I-IRL response passes the feasibility test of (9) by a factor of $1/(1-\eta)$.

• *Computational Burden for I-IRL.* If $\eta = 0$ (no I-IRL), the decision maker simply solves (16) for its true resource thresholds γ_k. However, for $\eta \in (0, 1]$, the decision maker needs to solve a two-stage optimization problem - it first generates the set of all feasible resource thresholds for which the optimal response (16) passes the IRL feasibility test with sufficiently low margin (15) (parametrized by η), and then minimizes the deviation from the true resource thresholds over this feasible set.

• It is straightforward to show the minimum violation of constraints (14) is monotone in the parameter η. If $\eta = 0$, the I-IRL response $\{\tilde{\beta}_k^*\}$ is identical to the naive response $\{\beta_k^*\}$ and the minimum violation of budget is 0. On the other extreme, setting $\eta = 1$ requires maximal violation of the budget constraints $\{g(\beta) \leq \gamma_k\}$ since $\mathcal{M}_g(\{\tilde{\beta}_k, u_k, \tilde{\gamma}_k\}) \leq 0$ (15) implies the I-IRL response and decision maker's budget fail the revealed preference test of Theorem 2.

We illustrate the I-IRL result in the next section via a radar example; see Fig. 2 for the simulation result.

B. Finite Sample Complexity for I-IRL in Theorem 3

In the previous sections, we assumed both the adversary and the decision maker had accurate measurements of the response and the utility functions. In this section, we assume the decision maker's measurements of the utility function is noisy, and the noise is modeled as a random linear perturbation. The key question we address is:

Given a finite sequence of I-IRL responses to noisy utility functions $u_k(\beta) + \delta_k'\beta$, what is probability that the decision maker effectively masks its strategy from the adversary?

Let us now formalize the above question. Let $\mathcal{M}_g^{\text{true}} = \mathcal{M}_g(\{\beta_k^*, u_k, \gamma_k\})$ (12) denote the margin with which the naive response sequence $\{\beta_k^*\}$ (1) passes the revealed preference test of Theorem 2. We want to bound the following error probability for I-IRL in Theorem 3:

$$P_{\text{err}} = \mathbb{P}_{\delta_{1:K}} \left(\mathcal{M}_g(\{\tilde{\beta}_k^*, u_k(\cdot) + \delta_k'(\cdot), \gamma_k^*\}) \geq (1-\eta) \; \mathcal{M}_g^{\text{true}} \right) \tag{17}$$

Recall from Theorem 3 that our I-IRL aim is to ensure the margin of the revealed preference test (9) lies under a threshold. In (17), P_{err} is the probability with which the constraint (14) in Theorem 3 fails. In simple terms, P_{err} is the probability of the event that the margin with which the I-IRL response satisfies the inequalities (9) in Theorem 2 exceeds the margin threshold $(1-\eta)\mathcal{M}_g^{\text{true}}$.

We assume the following for Theorem 4:

(A1) The adversary controlled utility function u_k is monotone, concave and Lipschitz continuous with Lipschitz constant L.

(A2) The decision maker has a noisy estimate $\hat{u}_k = u_k(\beta) + \delta_k(\beta)$ of the adversary controlled utility function $u_k(\beta)$. The linear perturbation vector δ_k is a Gaussian zero mean random vector with covariance Σ.

(A3) Let $\Delta(g, \{\beta_k, u_k, \gamma_k\})$ denote the range with which $g, \{\beta_k, u_k, \gamma_k\}$ pass the revealed preference test of (9):

$$\Delta(g, \{\beta_k, u_k, \gamma_k\}) = \max_{j,k} \epsilon_{j,k} - \min_{j,k} \epsilon_{j,k}, \text{ where}$$
$$\epsilon_{j,k} = \gamma_j - \gamma_k - \lambda_k \; (u_k(\beta_j) - u_k(\beta_k)),$$
$$\lambda_k \nabla u_k(\beta_k) = \nabla g(\beta_k).$$

The random variable $\Delta(g, \{\hat{\beta}_k, \hat{u}_k, \hat{\gamma}_k\}) \leq \Delta_{\max}$ a.s., where $\hat{\beta}_k$ and $\hat{\gamma}_k$ are the decision maker's I-IRL response (13) and constraint threshold (14) due to noisy utility function $\hat{u}_k$ measured by the decision maker.

(A4) The random variable $\frac{\max_k\{||\nabla u_k(\hat{\beta}_k)||_2^2/||\nabla g(\hat{\beta}_k)||_2\}}{\min_{j,k}\{||\nabla u_k(\hat{\beta}_k)-\nabla u_k(\hat{\beta}_j)||_2^2\}}$ is upper bounded almost surely by $\kappa > 0$.

We are now ready our finite sample complexity result for I-IRL (Theorem 3); see the appendix for the proof.

Theorem 4 (Finite Sample Complexity for I-IRL): Consider the decision maker choosing I-IRL responses according to (13) in Theorem 3 in response to noisy utility functions controlled by the adversary. Let β_k^* denote the decision maker's naive response at time k that maximizes the noise-less utility u_k subject to budget constraint $g(\beta) \leq \gamma_k$. Suppose assumptions (A1)-(A4) hold. Then:

$$P_{\text{err}} \leq \boldsymbol{\phi}^K \left(\frac{2L\Delta_{\max}\kappa}{\sqrt{\text{Tr}(\Sigma)}} \right) \tag{18}$$

where P_{err} is the error probability for I-IRL (Theorem 3) (17) and $\boldsymbol{\phi}(\cdot)$ is the cdf of the standard normal distribution.

IV. Example. I-IRL for Meta-cognitive radar

Theorem 3 specified the procedure for a decision maker to effectively mask its cognition from an adversary. Here, we apply our I-IRL result to the problem of a cognitive radar optimizing waveform based on the SINR of the adversarial target measurement [24]:

$$\beta_k \in \text{argmax}_\beta \, \text{SINR}(\alpha_k, \beta), \;\; p'\beta \leq p_k. \tag{19}$$

In (19), $p(i)\beta(i)$ is the cost of transmitting signal power $\beta(i)$ on the i^{th} waveform. The radar's SINR as a function of the adversary's probe and the radar's response is defined as:

$$\text{SINR}(\alpha, \beta) = \frac{\beta' Q \beta}{\beta' P(\alpha)\beta + \zeta}, \tag{20}$$

where ζ denotes the noise power. In (20), the radar's signal power and interference power are assumed to be quadratic forms of positive definite matrices $Q, P(\alpha)$ respectively. Clearly, the above setup falls under the non-linear utility maximization setup in Definition 1. For appropriately chosen matrices (see [24] for a detailed discussion), the utility in (19) can be shown to be monotonically increasing in β.

Suppose an adversary's aim is to learn the radar's resource constraint p (19).[3] The radar knows of the adversary's motives and wants to mask its plan p. Thus, the radar modifies its strategy (19) as per the I-IRL scheme of Theorem 3 to mask its non-linear budget (19) from the adversary. We illustrate the I-IRL performance via a simple numerical example with the following parameters:

- Time horizon $K = 100$, Response dimension $m = 6$.
- Budget vector $p = [p(1) \dots p(m)]$, $p(i) \sim \text{Unif}(1, 4)$.
- Extent of strategy masking η: Varied from 0.05 to 0.95.
- Matrix $Q = [Q_{i,j}]$, where $Q_{i,i} = 5$, $Q_{i,j} = 0$ if $j \neq i$, and $P(\alpha_k) = [P_{i,j}]$, where $P_{i,i} \sim \text{Unif}(1, 3)$ and $P_{i,j} = -0.05$ if $j \neq i$, Noise power $\zeta = 1$.

Our numerical results are shown in Fig. 2. Recall from Theorem 3 that the scalar η parametrizes the extent of masking of the radar's resource constraint from adversarial IRL of Theorem 2. The key observation from Fig. 2 is that the radar's deliberate performance degradation increases with extent of cognition masking η. Also, we see that a small constraint violation by the radar suffices to confuse the adversary to a large extent, hence successfully masking the radar's strategy.

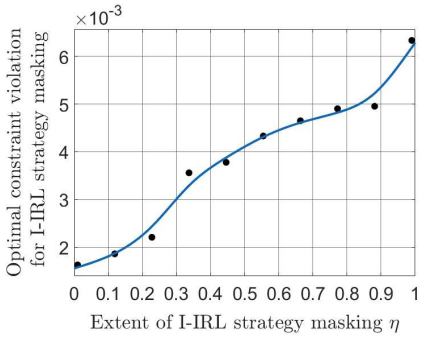

Fig. 2. I-IRL for masking the strategy of a cognitive radar: Small deliberate constraint violation of the radar (vertical axis) results in large performance loss (extent of strategy masking η) of the adversarial IRL algorithm (horizontal axis). The I-IRL constraint violation by the decision maker increases with η. $\eta = 0$ corresponds to zero strategy masking, and $\eta = 1$ corresponds to complete strategy masking by the decision maker.

V. Conclusion and Extensions

This paper focuses on masking a decision maker's strategy when probed by an adversarial inverse reinforcement learner. We term this problem inverse-inverse reinforcement learning (I-IRL). If the decision maker knows an adversary is trying to reconstruct its strategy, how should it tweak its responses to hide its strategy? Our main I-IRL result is Theorem 3. The key idea is for the decision maker to deliberately choose sub-optimal responses that violates its strategic resource constraints while ensuring the adversary does a poor reconstruction of the decision maker's strategy. Our finite sample result, Theorem 4, upper bounds the probability that our I-IRL result is ineffective in noisy settings; when the decision maker has noisy estimates of the adversary-specified utility functions.

Finally, a useful extension of this paper would be to study more general game-theoretic settings where even the adversary knows the radar is trying to mask its cognition.

VI. Appendix

A. Proof of Theorem 4

We start by computing the margin with which the I-IRL response of the decision maker passes the feasibility inequalities (9) of Theorem 3. Let $\hat{u}_k(\beta) = u_k(\beta) + \delta_k'\beta$ denote the noisy utility function estimate available to the decision maker. Let $\{\hat{\beta}_k\}$ and $\{\hat{\gamma}_k\}$ denote the I-IRL responses and

[3] Since the adversary knows and also controls the radar's utility function, it can benefit from knowing the radar's budget constraint. The adversary can, via carefully chosen probes, dupe the radar by forcing the radar to transmit low power signals (low tracking precision) on some time instants when the target performs malicious maneuvers. [25] shows how a seller can maximize its profit by effectively learning a consumer's utility from the consumer's responses. In the radar context, [26] computes the optimal probe sequence for an adversary that minimizes its IRL algorithm's Type-II error probability (incorrectly detecting utility maximization behavior).

perturbed constraint thresholds computed via (13) and (14), respectively, in response to noisy utility functions $\{\hat{u}_k\}$. The margin $\mathcal{M}_g(\hat{\beta}_k, u_k, \hat{\gamma}_k)$ is defined as:

$$\mathcal{M}_g(\hat{\beta}_k^*, u_k, \hat{\gamma}_k) = \min_{j,k} \underbrace{\hat{\gamma}_j - \hat{\gamma}_k - \lambda_k(u_k(\hat{\beta}_j) - u_k(\hat{\beta}_j))}_{=\epsilon_{j,k}}, \tag{21}$$

where $\lambda_k \nabla u_k(\hat{\beta}_k) = \nabla g(\hat{\beta}_k)$. If $\hat{u}_k$ were the true utility function at time k generated by the adversary, the margin definition in (21) changes to:

$$\mathcal{M}_g(\hat{\beta}_k, \hat{u}_k, \hat{\gamma}_k) = \min_{j,k} \underbrace{\hat{\gamma}_j - \hat{\gamma}_k - \hat{\lambda}_k(\hat{u}_k(\hat{\beta}_j) - \hat{u}_k(\hat{\beta}_j))}_{=\hat{\epsilon}_{j,k}}, \tag{22}$$

where $\hat{\lambda}_k \hat{u}_k(\hat{\beta}_k) = \nabla g(\hat{\beta}_k)$. Observe that by definition (14), $\mathcal{M}_g(\hat{\beta}_k, \hat{u}_k, \hat{\gamma}_k) = (1-\eta)\mathcal{M}_g^{\text{true}}$. Also, we observe that the margin definitions in (21) and (22) differ only in the term involving the utility functions. Our aim is to find necessary conditions for which the event $\{\mathcal{M}_g(\hat{\beta}_k, u_k, \hat{\gamma}_k) \geq (1-\eta)\mathcal{M}_g^{\text{true}}\}$ holds, or equivalently, the event $\{\mathcal{M}_g(\hat{\beta}_k, \hat{u}_k, \hat{\gamma}_k) \leq \mathcal{M}_g(\hat{\beta}_k, u_k, \hat{\gamma}_k)\}$ holds.

Due to assumption (A3), a necessary condition for the event $\{\mathcal{M}_g(\hat{\beta}_k, u_k, \hat{\gamma}_k) \geq (1-\eta)\mathcal{M}_g^{\text{true}}\}$ to hold is $\{\epsilon_{j,k} \geq \hat{\epsilon}_{j,k} - \Delta_{\max},\ \forall j,k\}$. We wish to bound the term $(\hat{\epsilon}_{j,k} - \epsilon_{j,k})$:

$$\begin{aligned}
&\hat{\epsilon}_{j,k} - \epsilon_{j,k} = \lambda_k(u_k(\hat{\beta}_j) - u_k(\hat{\beta}_j)) - \hat{\lambda}_k(\hat{u}_k(\hat{\beta}_j) - \hat{u}_k(\hat{\beta}_j)) \\
=&\lambda_k(u_k(\hat{\beta}_j) - u_k(\hat{\beta}_j)) - (\lambda_k + (\hat{\lambda}_k - \lambda_k))(u_k(\hat{\beta}_j) \\
&\qquad - u_k(\hat{\beta}_j) + \delta_k'(\hat{\beta}_j - \hat{\beta}_k)) \\
=&-(\hat{\lambda}_k - \lambda_k)(u_k(\hat{\beta}_j) - u_k(\hat{\beta}_k) - \nabla u_k(\beta_k)'(\hat{\beta}_j - \hat{\beta}_k)) \\
&\quad (\text{since } \lambda_k \nabla u_k(\hat{\beta}_k) = \hat{\lambda}_k \nabla \hat{u}_k(\hat{\beta}_k)) \\
=&(\hat{\lambda}_k - \lambda_k)(u_k(\hat{\beta}_k) + \nabla u_k(\beta_k)'(\hat{\beta}_j - \hat{\beta}_k) - u_k(\hat{\beta}_j)) \\
\geq&(\hat{\lambda}_k - \lambda_k)\frac{1}{2L}||\nabla u_k(\hat{\beta}_k) - \nabla u_k(\hat{\beta}_j)||_2^2 \ (\text{Asmp. (A1)})
\end{aligned} \tag{23}$$

From (21), (22), we rewrite $\hat{\lambda}_k - \lambda_k$ in (23) as:

$$\hat{\lambda}_k - \lambda_k = \lambda_k \frac{\delta_k' \nabla u_k(\hat{\beta}_k)}{||\nabla u_k(\hat{\beta}_k)||_2^2} = \frac{\delta_k' \nabla g(\hat{\beta}_k)}{||\nabla u_k(\hat{\beta}_k)||_2^2} \tag{24}$$

Combining (23) and (24), the following inequality results:

$$\begin{aligned}
&\hat{\epsilon}_{j,k} - \epsilon_{j,k} \leq \Delta_{\max} \\
\Rightarrow &\delta_k' \nabla g(\hat{\beta}_k) \leq \frac{2L\Delta_{\max}||\nabla u_k(\hat{\beta}_k)||_2^2}{\min_{j,k}\{||\nabla u_k(\hat{\beta}_k) - \nabla u_k(\hat{\beta}_j)||_2^2\}}
\end{aligned}$$

$\{\delta_k' \nabla g(\hat{\beta}_k)\}$ is a sequence of independent zero mean Gaussian random variables with variance $\{\text{Tr}(\Sigma)||\nabla g(\hat{\beta}_k)||_2^2\}$ Also, notice how the LHS does not depend on the index j. Thus, we express our error probability P_{err} as:

$$\begin{aligned}
P_{\text{err}} =&\mathbb{P}(\hat{\epsilon}_{j,k} - \epsilon_{j,k} \leq \Delta_{\max},\ \forall j,k) \leq \prod_{k=1}^{K} \mathbb{P}\left(\delta_k' \nabla g(\hat{\beta}_k) \leq \pi_k\right) \\
=&\prod_{k=1}^{K} \phi\left(\frac{2L\Delta_{\max}||\nabla u_k(\hat{\beta}_k)||_2^2/||\nabla g(\hat{\beta}_k)||_2}{\sqrt{\text{Tr}(\Sigma)}\min_{j,k}\{||\nabla u_k(\hat{\beta}_k) - \nabla u_k(\hat{\beta}_j)||_2^2\}}\right) \\
\leq&\phi^K\left(\frac{2L\Delta_{\max}\max_k\{||\nabla u_k(\hat{\beta}_k)||_2^2/||\nabla g(\hat{\beta}_k)||_2\}}{\sqrt{\text{Tr}(\Sigma)}\min_{j,k}\{||\nabla u_k(\hat{\beta}_k) - \nabla u_k(\hat{\beta}_j)||_2^2\}}\right) \\
=&\phi^K\left(\frac{2L\Delta_{\max}\kappa}{\sqrt{\text{Tr}(\Sigma)}}\right) \ (\text{from (A4)}) \qquad \blacksquare
\end{aligned}$$

REFERENCES

[1] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[2] L. Kang, J. Bo, L. Hongwei, and L. Siyuan. Reinforcement learning based anti-jamming frequency hopping strategies design for cognitive radar. In *2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, pages 1–5. IEEE, 2018.

[3] S. Afriat. The construction of utility functions from expenditure data. *International economic review*, 8(1):67–77, 1967.

[4] H. Varian. Non-parametric tests of consumer behaviour. *The Review of Economic Studies*, 50(1):99–110, 1983.

[5] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[6] A. Y. Ng, S. J. Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

[7] H. Varian. Revealed preference and its applications. *The Economic Journal*, 122(560):332–338, 2012.

[8] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

[9] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich. Maximum margin planning. In *Proceedings of the 23rd international conference on Machine learning*, pages 729–736, 2006.

[10] S. H. Silva and P. Najafirad. Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*, 2020.

[11] K. Pattanayak, V. Krishnamurthy, and C. Berry. Meta-cognition. an inverse-inverse reinforcement learning approach for cognitive radars. *arXiv preprint arXiv:2205.01794*, 2022.

[12] S. Haykin. Cognitive radar. *IEEE Signal Processing Magazine*, pages 30–40, Jan. 2006.

[13] S. Haykin. Cognitive dynamic systems: Radar, control, and radio [point of view]. *Proceedings of the IEEE*, 100(7):2095–2103, 2012.

[14] K. Bell, C. Baker, G. Smith, J. Johnson, and M. Rangaswamy. Cognitive radar framework for target detection and tracking. *IEEE Journal of Selected Topics in Signal Processing*, 9(8):1427–1439, 2015.

[15] L. Neng-Jing and Z. Yi-Ting. A survey of radar ecm and eccm. *IEEE Transactions on Aerospace and Electronic Systems*, 31(3):1110–1120, 1995.

[16] C. Shi, F. Wang, M. Sellathurai, and J. Zhou. Low probability of intercept-based distributed mimo radar waveform design against barrage jamming in signal-dependent clutter and coloured noise. *IET Signal Processing*, 13(4):415–423, 2019.

[17] W.-Q. Wang. Moving-target tracking by cognitive rf stealth radar using frequency diverse array antenna. *IEEE Transactions on Geoscience and Remote Sensing*, 54(7):3764–3773, 2016.

[18] W.-Q. Wang. Adaptive rf stealth beamforming for frequency diverse array radar. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1158–1161. IEEE, 2015.

[19] F. Forges and E. Minelli. Afriat's theorem for general budget sets. *Journal of Economic Theory*, 144(1):135–145, 2009.

[20] H. Varian. Revealed preference. *Samuelsonian economics and the twenty-first century*, pages 99–115, 2006.

[21] H. Varian. The nonparametric approach to demand analysis. *Econometrica*, 50(1):945–973, 1982.

[22] H. R. Varian et al. *Goodness-of-fit for revealed preference tests*. Citeseer, 1991.

[23] K. Pattanayak and V. Krishnamurthy. Unifying classical and bayesian revealed preference. *arXiv preprint arXiv:2106.14486*, 2021.

[24] V. Krishnamurthy, K. Pattanayak, S. Gogineni, B. Kang, and M. Rangaswamy. Adversarial radar inference: Inverse tracking, identifying cognition, and designing smart interference. *IEEE Transactions on Aerospace and Electronic Systems*, 57(4):2067–2081, 2021.

[25] K. Amin, R. Cummings, L. Dworkin, M. Kearns, and A. Roth. Online learning and profit maximization from revealed preferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[26] V. Krishnamurthy and W. Hoiles. Afriat's test for detecting malicious agents. *IEEE Signal Processing Letters*, 19(12):801–804, 2012.