Inverse-Inverse Reinforcement Learning.

Masking Strategy from Inverse Reinforcement Learning

Kunal Pattanayak (Cornell University), Vikram Krishnamurthy (Advisor) and Christopher Berry (Lockheed Martin Collaborator)

Main Ideas.

- Utility Maximization (Microeconomics Theory, Machine Learning)
- (Adversarial IRL) Detecting Utility Maximization and Estimating Utility
- (Counter-Adversarial Move) Hiding Strategy from Adversarial IRL

Primer

Reinforcement Learning (RL):

$\rightarrow$ Markov Decision Process (MDP): Next state $x_{t+1} \sim f(x_t, a_t)$

$\rightarrow$ Maximize expected cumulative reward $R(x_t, a_t)$

$\rightarrow$ Examples: TD-Learning, SARSA, Q-learning (Robbins-Monro)

$\rightarrow$ Variance reduction, Off-Policy Evaluation

Inverse Reinforcement Learning (IRL) [1, 2]:

$\rightarrow$ Assumes RL algo. has converged to optimal policy $\pi^* : X \rightarrow A$

$\rightarrow$ *Reverse engineer MDP* - Find $\boldsymbol{R}$ s.t. π^* is optimal

$\rightarrow$ For inf-horizon MDP: Bellman optimality $\overset{LP}{\equiv}$ $A\boldsymbol{R}_{\text{est}} \leq \mathbf{0}$

$\rightarrow$ Ill-posed problem, but true reward $\boldsymbol{R}$ satisfies $A\boldsymbol{R} \leq 0$

Optimal policy for MDP $\rightarrow$ Bellman optimality,

IRL for MDP $\rightarrow$ Checking if Bellman optimality holds (LP)

Departing from MDPs to Constrained Utility Maximization

Utility Maximization: At time k, agent faces (possibly non-linear) resource constraint $g_k(\beta) \leq 0$, chooses **optimal** response β_k:

$$\beta_k = \operatorname{argmax}_{\beta \in \mathbb{R}_+^m} u(\beta),\ g_k(\beta) \leq 0,$$
$$\text{Active Constraint } g_k(\beta_k) = 0,\ k = 1, 2, \ldots, T\ (T < \infty)$$

Revealed Preference [3, 4]: Finds u_{est} that rationalizes analyst dataset $\mathbb{D} = \{g_k, \beta_k\}_{k=1}^T$:

(S1) There exists u_{est} if the following LP has a feasible solution:

$$\exists \{u_k, \lambda_k\} \in \mathbb{R}_+^{2T} \text{ s.t. } \mathrm{RP}(u, \mathbb{D}) \leq 0 \equiv\ u_s - u_k - \lambda_k g_k(\beta_s) \leq 0,\ \forall\ s, k$$

(S2) $u_{\text{est}}(\beta) = \min_k \{u_k + \lambda_k g_k(\beta)\}$ rationalizes $\mathbb{D}$

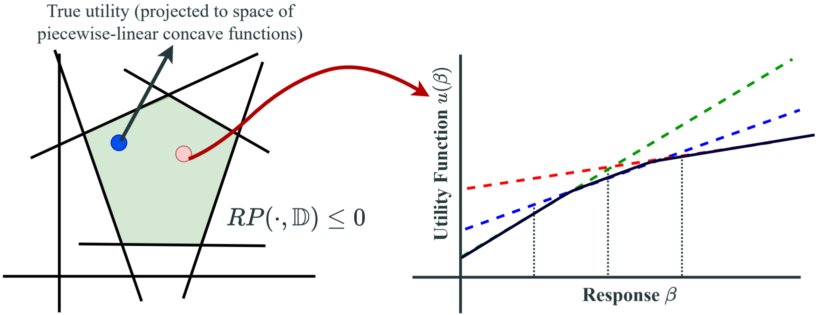
(Summary) Utility Maximization → KKT,

IRL (RP) for UM → Check for KKT, stitch *piece-wise utility.*

For quasi-convex g, reconstruction is piece-wise linear concave.

Illustration. Revealed Preference Reconstruction

- Every feasible point in revealed preference LP corresponds to a rationalizing utility function
- Can have a smaller (precise) set by pinning down feasible variables $u_1, \lambda_1 = 1$ WLOG



Relating Revealed Preference and IRL

Variable	IRL	Revealed Preference
Probe	$\pi_0, P(\cdot\|x_k, a_k)$	$\{g_k(\cdot) \leq 0\}_{k=1}^T$
Response	π^*	$\{\beta_k\}_{k=1}^T$
Reward	$R(x, a)$	$u(\beta)$
IRL Rationale	Bellman Optimality	Rationalizability

- Revealed preference (RP) $\equiv$ **IRL for utility maximization**

- Equivalent RP variants [5] exist for sequential decision-making for cumulative utility maximization

For this talk: (i) Consider utility maximization framework,
(ii) View IRL as adversarial eavesdropper that extracts strategy

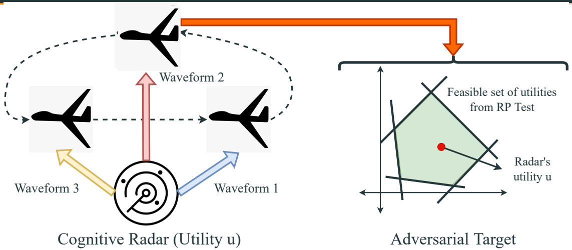
Let's Turn the Tables

"Can the decision maker spoof RP? If so, how?"

Some Comments on Inverse IRL

- IRL is System Identification (SI) [6, 7]. I-IRL aim is to ensure SI fails (not unidentifiable, but mis-specified utility estimate)
- Subject to budget constraints, make sub-optimal choices that:
  1. Ensure true utility function is **almost** infeasible for RP test
  2. Minimize utility loss due to sub-optimal response
- Inverse IRL focuses on ensuring utility (preferences) are not recoverable (revealed preference fails)
- Idea is gaining traction, for e.g. [8] that treats additive separable value and privacy term for maximization
- *Naive approach:* For all k, choose the same response β. This way, feasible set of utilities only contains the constant utility function and true utility lies outside the feasibility zone.

Running Example. Cognitive Radar Spoofing Adversary Target



Cognitive Radar: For adversary maneuvers $\{\alpha_k\}_{k=1}^K$, radar chooses waveforms (**response**) $\{\beta_k\}_{k=1}^K$ such that $\beta_k = \text{argmax}_\beta\, u(\beta),\ \alpha_k'\beta \leq 1$

Radar Bayesian tracker: α_k: state noise cov., β_k: inverse of observation noise cov., Radar SNR (Kalman precision) upper bound $\alpha_k'\beta_k \leq 1$

Adversary Target: Uses RP test to generate set-valued radar utility.
What if $\mathbb{D}$ is noisy? Test to detect feasibility [9] (*later*)

Radar $\rightarrow$ *"I need to safeguard my utility and spoof IRL (ensure poor utility reconstruction)"*

Introduction Summary

Testing for utility maximization $\equiv$ RP Test [10, 11] (LP Feasibility)

How to make checking linear feasibility difficult?

Ans. **Cognition (Strategy) Masking**

Intelligently perturbed actions successfully hide utility

We term this task as inverse IRL (I-IRL)

Key Ideas for I-IRL

- **Objective:** Ensure utility **almost fails** RP test
- **How?** Deliberately deviate from optimal response to *trick* IRL
- **Constraint:** Bounded Deviation from optimal response

"Performance-Obfuscation Trade-off"

Inspired from differential privacy [12], adversarial ML [13]

Deterministic I-IRL (Accurate Probe-Response Exchange)

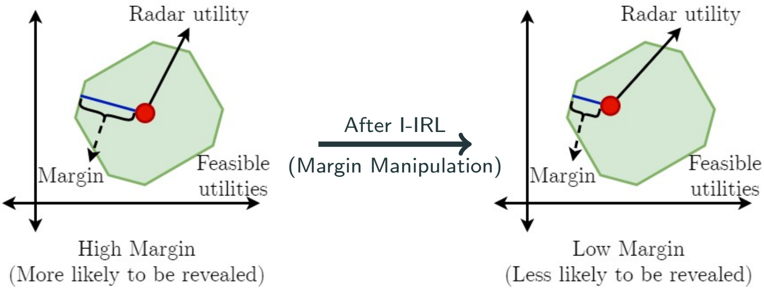
Adversarial target $\xrightarrow{\text{IRL}}$ RP Feasibility test (Reconstruct agent utility)

Key Question: How to rank utility functions in the feasible set?

Soln.: **Margin** of RP test - **max. perturbation to fail RP test**

$$\boxed{\text{Margin}_{\mathbb{D}}(u) = \max_{\epsilon \geq 0} \epsilon, \quad \text{RP}(u, \mathbb{D}) + \epsilon \geq 0}$$

Resembles Afriat number [3], Houtman-Maks Index [14], Varian number [4] in economics for quantifying rationality



- Margin: Closeness to edge of feasible set (infeasibility of RP test)
- Center of feasible set: **max. margin**, edge of feasible set: **zero margin**
- $\downarrow$ Margin $\Leftrightarrow$ $\downarrow$ Goodness-of-fit to RP test (almost infeasible)
- But, $\downarrow$ Margin $\Leftrightarrow$ $\uparrow$ Deviation from optimal response
- **Deterministic** I-IRL: Deliberately perturb response to push utility **towards** edge of feasible set from RP test
- Focus on making u almost fail RP test, instead of ensuring no feasible set at all

Deterministic Inverse IRL for Masking Cognition

Suppose radar faces adversarial constraints $\{\alpha_k'\beta \leq 1\}_{k=1}^K$. The radar's *deterministic* I-IRL algorithm to hide its utility u is:

Step 1. Choose margin $\epsilon_{\text{thresh}} \in \mathbb{R}_+$
Step 2. Compute naive response β_k^*
Step 3. Compute optimal perturbation $\{\delta_k^*\}$ for I-IRL:

$$\{\delta_k^*\} = \underset{\{\delta_k\}\in\mathbb{R}^m}{\operatorname{argmin}} \underbrace{\sum_{k=1}^K \|\delta_k\|_2^2}_{\text{(Radar's degradation)}} \quad , \quad \underbrace{\text{Margin}_{\{\alpha_k,\beta_k^*+\delta_k\}}(u) \leq \epsilon_{\text{thresh}}}_{\text{(Mitigating adversarial RP Test)}} \qquad (1)$$

Step 4. Transmit engineered sub-optimal responses $\{\beta_k^* + \delta_k^*\}$.

Summary

Deterministic I-IRL: Small margin ϵ_{thresh}

$\iff$ Closer to failing RP test

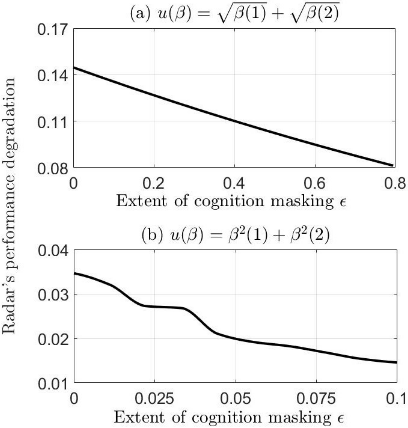
$\iff$ Larger deviation from radar's optimal strategy

- Margin Constraint is non-convex (bilinear).

Current research: *Formulate convex relaxations of bi-linear I-IRL constraints.*

Numerical Results: Deterministic Inverse IRL

- Simulation-based datasets to illustrate I-IRL for 2 utility functions
- Time horizon = 50, Response dimension = 2



Insights:

- **Small deviation** from *optimal strategy* masks *u* by a large extent.
- Performance degradation $\downarrow$ with ϵ (distance from edge of feasible set).
- Optimal deviation inversely proportional to utility's Lipschitz constant

Stochastic I-IRL. Noisy Response at Adversary IRL

(Adversary side): $\hat{\beta}_k = \beta_k + w_k,\ w_k \sim f_w$ (f_w known to radar) (2)

Adversarial target $\overset{\text{IRL}}{\to}$ Feasibility Detector (see also [10] for details)

H_0 : RP Test has a feasible solution for $\{\alpha_k, \beta_k\}$

H_1 : RP Test has NO feasible solution for $\{\alpha_k, \beta_k\}$

IRL Detector : $\boxed{\phi^*(\widehat{\mathbb{D}}) \gtrless_{H_0}^{H_1} F_L^{-1}(1-\eta)}$ $(\widehat{\mathbb{D}} = \{\alpha_k, \hat{\beta}_k\})$

Test Statistic $\phi^*(\widehat{\mathbb{D}})$: Min. perturbation to pass RP test,

Reference r.v. $L := \max_{j,k} \alpha_j'(w_j - w_k)$,

Variable η : Adversary chosen bound for $\mathbb{P}(H_1|H_0)$

"Radar labeled non-cognitive if margin $\leq$ threshold"

Differences compared to Deterministic I-IRL

- Radar **can no more** manipulate margin of RP test
- Can *at best* manipulate $\mathbb{P}(H_1|\{\alpha_k, \beta_k\}, u)$, the Conditional Type-I error probability, conditioned on u
- **Stochastic** I-IRL: Deliberately perturb radar's response to mitigate IRL detector (**increase** conditional Type-I error probability)
- Computing optimal I-IRL requires non-deterministic constraints (threshold on $\mathbb{P}(H_1|\{\alpha_k, \beta_k\}, u)$)
- Stochastic approximation (finite perturbation methods) methods to achieve local optimal

Stochastic Inverse IRL for Masking Cognition

Adversary's sensor is noisy; everything else the same as deterministic case. Radar's *stochastic* I-IRL algorithm is:

Step 1. Choose sensitivity parameter $\lambda > 0$

Step 2. Compute naive response β_k^*

Step 3. Compute optimal perturbation $\{\delta_k^*\}$ for I-IRL:

$$\{\delta_k^*\} = \underset{\{\delta_k\}\in\mathbb{R}^m}{\operatorname{argmin}} \sum_{k=1}^{K} \Big(\underbrace{u(\beta_k^*) - u(\beta_k^* + \delta_k))}_{\text{(Radar's deliberate performance loss)}} - \lambda \underbrace{\mathbb{P}(H_1 | \{\alpha_k, \beta_k^* + \delta_k\}, u)}_{\text{(Mitigating adversarial IRL detector)}}$$

Step 4. Transmit engineered sub-optimal responses $\{\beta_k^* + \delta_k^*\}$

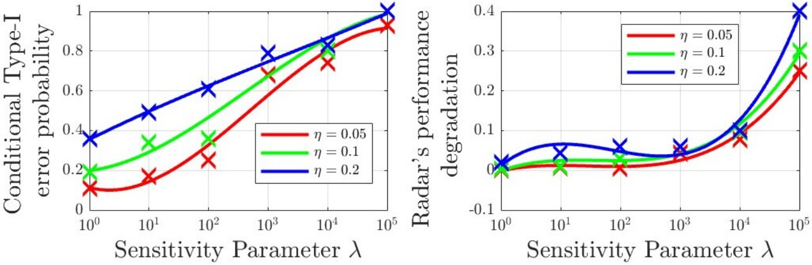
Objective: Ensuring low margin of RP Test with high probability

Summary

- **Stochastic I-IRL**: Trade-off between $\uparrow$ *QoS* and $\uparrow$ *adversarial obfuscation*.
- Radar can only <u>estimate</u> $\mathbb{P}(H_1 | H_0, u)$ via Monte-Carlo methods.
- Stochastic approximation based algorithms like **SPSA [15]** can be used.
- SPSA $\rightarrow$ Fewer (only 2) computations/update wrt finite diff. methods.

Numerical Results: Stochastic Inverse IRL

- Utility function $u(\beta) = \sqrt{\beta_1} + \sqrt{\beta_2}$, Time horizon $K = 50$



Key Insights:

- Small *performance loss* sufficiently confuses IRL detector (large cond. Type-I error).
- **Both** adversarial confusion and performance loss $\uparrow$ with λ.
- Interestingly, performance degradation $\downarrow$ with η (error bound).
- On right figure, notice the elbow point at $\lambda \approx 10^3$

Finite Sample Effects for Inverse IRL

Suppose:
- Radar has noisy (additive Gaussian) measurements of the adversary's probes α_k.
- Radar oblivious to sensor noise and uses deterministic I-IRL.

Want to Study: Effects of noisy constraint on utility spoofing

Recall: Deterministic I-IRL $\rightarrow$ RP test margin $\leq \epsilon_{\text{thresh}}$
Want to bound: Probability that utility is **NOT** within ϵ_{thresh} margin for RP test:

$$\boxed{\mathbb{P}(\text{Margin}_{\{\alpha_k + w_k, \tilde{\beta}_k^*\}}(u) \not\leq \epsilon_{\text{thresh}})}$$

$w_k \rightarrow$: Radar sensor's measurement noise,
$\tilde{\beta}_k^* \rightarrow$: I-IRL response.
Assume i.i.d $w_k \sim \mathcal{N}(\mathbf{0}, \Sigma)$.

Finite Sample Result for I-IRL

Finite Sample Complexity for Deterministic I-IRL

For **deterministic** I-IRL responses, observes adversary signals in noise. Then, under mild conditions, the I-IRL error probability is bounded as:

$$\mathbb{P}(\text{Margin}_{\{\alpha_k + w_k, \tilde{\beta}_k^*\}}(u) > \epsilon_{\text{thresh}}) \leq 1 - \frac{T\ e^{-\psi^2/2}}{\psi\sqrt{2\pi}}$$

- $\psi(\cdot)$: proportional to range of allowable probes,
inversely proportional to Lipschitz constant of utility, noise power

Takeaway: Error probability $\downarrow$ with horizon T, utility's Lipschitz constant and $\uparrow$ with noise power.

Remark. Above error bound is loose, currently investigating tighter convergence rates.

Summary

- Considered the task of inverse IRL - *how to spoof a strategy extracting system*.
- **Main Idea**: *Deliberately perturb optimal response to sufficiently reduce margin of RP test for utility maximization and 'hide' utility.*
- Sub-optimality in response trades-off between **Privacy** and **Performance**
- Discussed both noise-less and noisy exchange scenarios: both cases are challenging (*non-convex, stochastic approximation*)
- Finite sample complexity for I-IRL error - *How robust is I-IRL to noise in adversary signal measurement?*
- Methodology inspired from adversarial obfuscation [13] in deep learning and differential privacy [12]

Extensions

1. *Online IRL.* Current strategy hiding idea is offline (since IRL via Afriat's Theorem is intrinsically offline). Bandit approach for approximating IRL detector?
2. *Semi-parametric I-IRL.* Jointly optimize over response perturbations and variance of additive Laplacian noise for robust I-IRL.
3. **Counter**-(counter-)^{n}measure: What if adversary knows radar's spoofing strategy? *Game theoretic approach*

If you have any ideas (even if vaguely related), let's chat! Eager to know your thoughts.

Thank You!

Miscellaneous

FAQs

- **How justified is the constrained utility maximization abstraction for radar operation?**

Quite prevalent in literature:

(i) Multi-UAV network [16]: Utility $\rightarrow$ Fairness and downlink data rate, Constraint $\rightarrow$ Transmission power, Cramer-Rao bound on localization accuracy
(ii) Q-RAM (Resource Allocation) [17]: Utility $\rightarrow$ QoS for tracking and search, Constraint $\rightarrow$ Bandwidth, Short-term and Long-term constraints
(iii) Radar Tracking with ECM [18]: Utility $\rightarrow$ Neg. of weighted mean of radar energy and dwell time, Constraint $\rightarrow$ 4% Cap on lost tracks due to ECM

FAQs

- **Is conditional Type-I probability the only I-IRL metric for adversarial obfuscation in stochastic I-IRL?**

No fixed formula, does need more work. Some intuitive alternatives: (a) Use deterministic I-IRL as is. Formulate concentration inequalities for margin of the noisy dataset.

(b) Manipulate the average margin instead of margin. BUT, might be underplaying robustness of IRL detector.

(c) [**Speculative**] Use a neural network to learn IRL method on the fly and disrupt ECM.

Remark: I-IRL hinges delicately on IRL methodology.

Other heuristic ideas to hide utility?

FAQs

- **What's next after IRL, and inverse IRL? I2-IRL?**

Game-theoretic formulation.

Key challenge: Formulate a utility function in terms of both adversary probes and radar response.

Anticipated outcome: Inverse game theory - Detecting play from the Nash equilibrium of a game between adversary and radar.

References

[1] Pieter Abbeel and Andrew Y Ng. "Apprenticeship learning via inverse reinforcement learning". In: *Proceedings of the twenty-first international conference on Machine learning*. 2004, p. 1.

[2] A. Y. Ng, S. J. Russell, et al. "Algorithms for inverse reinforcement learning.". In: *Icml*. Vol. 1. 2000, p. 2.

[3] S. N. Afriat. "The construction of utility functions from expenditure data". In: *International economic review* 8.1 (1967), pp. 67–77.

[4] H. R. Varian. "Revealed preference and its applications". In: *The Economic Journal* 122.560 (2012), pp. 332–338.

[5] Victor H Aguiar and Nail Kashaev. "Stochastic revealed preferences with measurement error". In: *The Review of Economic Studies* 88.4 (2021), pp. 2042–2093.

[6] Romain Pasquier and Ian FC Smith. "Robust system identification and model predictions in the presence of systematic uncertainty". In: *Advanced Engineering Informatics* 29.4 (2015), pp. 1096–1109.

[7] Lennart Ljung. "System identification". In: *Signal analysis and prediction*. Springer, 1998, pp. 163–173.

[8] Tesary Lin. "Valuing intrinsic and instrumental preferences for privacy". In: *Marketing Science* (2022).

[9] Vikram Krishnamurthy and William Hoiles. "Afriat's test for detecting malicious agents". In: *IEEE Signal Processing Letters* 19.12 (2012), pp. 801–804.

[10] Vikram Krishnamurthy et al. "Identifying cognitive radars-inverse reinforcement learning using revealed preferences". In: *IEEE Transactions on Signal Processing* 68 (2020), pp. 4529–4542.

[11] Vikram Krishnamurthy et al. "Adversarial Radar Inference: Inverse Tracking, Identifying Cognition, and Designing Smart Interference". In: *IEEE Transactions on Aerospace and Electronic Systems* 57.4 (2021), pp. 2067–2081. DOI: 10.1109/TAES.2021.3090901.

[12] Rathindra Sarathy and Krishnamurty Muralidhar. "Evaluating Laplace noise addition to satisfy differential privacy for numeric data.". In: *Trans. Data Priv.* 4.1 (2011), pp. 1–17.

[13] Varun Chandrasekaran et al. "Face-off: Adversarial face obfuscation". In: *arXiv preprint arXiv:2003.08861* (2020).

[14] Martijn Houtman and J Maks. “Determining all maximal data subsets consistent with revealed preference”. In: *Kwantitatieve methoden* 19.1 (1985), pp. 89–104.

[15] James C Spall. “An overview of the simultaneous perturbation method for efficient optimization”. In: *Johns Hopkins apl technical digest* 19.4 (1998), pp. 482–492.

[16] Xinyi Wang et al. “Constrained utility maximization in dual-functional radar-communication multi-UAV networks”. In: *IEEE Transactions on Communications* 69.4 (2020), pp. 2660–2672.

[17] Jeffery Hansen et al. “Resource management for radar tracking”. In: *2006 IEEE Conference on Radar*. IEEE. 2006, 8–pp.

[18] WD Blair et al. “Benchmark for radar allocation and tracking in ECM”. In: *IEEE Transactions on Aerospace and Electronic Systems* 34.4 (1998), pp. 1097–1114.