INVERSE REINFORCEMENT LEARNING: A MICROECONOMICS-BASED APPROACH

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Kunal Pattanayak

December 2023

INVERSE REINFORCEMENT LEARNING:
A MICROECONOMICS-BASED APPROACH

Kunal Pattanayak, Ph.D.

Cornell University 2023

The general theme of this thesis is inverse reinforcement learning (IRL) for cognitive systems. By observing the end decisions generated from a cognitive system in multiple environments, the central idea revolves around how to identify if a system strategy is optimal and if so, estimate the system's utility function. Both traditional IRL in machine learning and revealed preference in microeconomics tackle the same question: given expert behavior, how to perform inverse optimization and reconstruct the expert's utility? This thesis advances and amalgamates the state-of-the-art in both machine learning and microeconomics theory in both a theoretical and applied sense.

First, we consider Bayesian stopping time problems and formulate necessary and sufficient conditions for an agent's behavior to be consistent with optimal stopping. The results advance the state-of-the-art in IRL in that an analyst can estimate system utility without requiring the agent's private observation likelihood. For practical applications, we also develop concentration inequalities for identifying strategy optimality when the analyst has empirical datasets. This class of IRL results may be used, for example, to formulate effective teaching strategies that cater to a particular student's attention span without intrusive probing.

Second, we exploit revealed preference-based IRL for adversarial identification and mitigation schemes for cognitive radars. The IRL techniques developed herein can be viewed as electronic countermeasures (ECM) for cognitive radars and facilitate non-parametric system identification of adversarial entities. For instance, by observing the emissions of an adversarial target, a cognitive radar can estimate the target's system

utility and can tune its sensing strategy to minimize the asymptotic covariance of the estimate of the target's coordinates.

Third, we unify two areas in microeconomics, namely, revealed preference and costly information acquisition. Tests for costly information acquisition identify if a decision maker expends attention optimally as a cognitive 'cost', where attention abstracts the decision maker's private subjective signals. Our result shows that the test for costly information acquisition is identical to a Bayesian (Blackwell order) analog of the test for quasi-linear utility maximization under an appropriate parameter map. The unification has several consequences: (i) we exploit the well-known equivalence between GARP and Afriat inequalities to reduce the computational complexity for testing costly information acquisition (combinatorial to quadratic); (ii) we can formulate robustness tests for costly information acquisition (translated from revealed preference) under noisy datasets. Finally, as an illustration, we perform a revealed preference-style analysis of user engagement metadata from a real-world YouTube dataset comprising 190k videos and show YouTube user engagement is consistent with costly information acquisition.

Finally, we take a step beyond the general philosophy of IRL and propose inverse-inverse reinforcement learning (I-IRL). The key idea is to presume the presence of an adversary performing IRL. If a cognitive system is aware of such an adversary, how can the system tweak its strategy to mask its utility from adversarial IRL? We specify how a cognitive system can deliberately choose 'optimal' sub-optimal responses that trade-off between maximizing the system utility and minimizing the probability of accurate utility reconstruction. From a cognitive radar's perspective, I-IRL can be viewed as an ECCM mechanism that minimizes strategy leakage subject to a bound on the radar's deviation from optimal sensing strategy. From a privacy-preserving perspective, I-IRL specifies the minimum magnitude of 'noise' required to be added to an offline dataset to minimize the recoverability of private attributes from the dataset.

BIOGRAPHICAL SKETCH

Kunal Pattanayak received the Bachelor and Master of Technology degree in Electronics and Electrical Communication Engineering from the Indian Institute of Technology Kharagpur, India in 2018. His doctoral research at Cornell University has been advised by Prof. Vikram Krishnamurthy. During his doctoral research, he collaborated with U.S. Air Force Research Laboratory and Lockheed Martin Advanced Technologies. He interned with RADAR, an RFID-based inventory tracking company, from May 2022 to August 2022. Post his graduate studies, he has joined Goldman Sachs as a Quantitative Strategist Associate in the Liquidity Risk division.

This thesis is dedicated to my parents, and to the entire Cornell and Ithaca community.

ACKNOWLEDGEMENTS

I am deeply grateful to my advisor, Dr. Vikram Krishnamurthy, for his invaluable guidance throughout my PhD journey. His uncanny knack of formulating interesting research problems, out-of-the-box thought process for problem-solving and sharp attention during research discussions have been instrumental and inspirational in shaping my research skills and this thesis. I would also like to extend my gratitude to my committee members, Dr. Siddhartha Banerjee and Dr. Jayadev Acharya, for their encouragement and feedback that helped strengthen my work.

During my graduate studies, I was extremely fortunate to collaborate with excellent research groups such as Lockheed Martin Advanced Technologies, Air Force Research Laboratory, and Army Research Laboratory. I am thankful for the enriching research exposure gained during our collaborations. In particular, I wish to acknowledge Dr. Muralidhar Rangaswamy, Dr. Erik Blasch, Dr. Alec Koppel, and Christopher Berry for the many fruitful discussions that ultimately led to this thesis.

I wish to express my sincere appreciation to my fellow lab members in the Cornell Statistical Signal Processing Lab, namely Dr. Sujay Bhatt, Dr. Buddhika Nettasinghe, Dr. Rui Luo, Omer Serbetci, Anurag Gupta, Shashwat Jain, Luke Snow and Adit Jain. Their constant support, encouragement, inspiration and camaraderie throughout the ups and downs of my PhD years have been invaluable. I am innately grateful to Buddhika for helping me assimilate into the group's style of conducting research. Our eternal brainstorming sessions will be forever etched in my memory. I wish the current group members all the very best to take the group forward.

I am also grateful to the wonderful Ithaca Latin dance community and Salsa Palante for helping me overcome the initial cultural shock during my time in Ithaca. More than the dance aspect, it was the sense of community that gave me an outlet to de-stress and feel at home in a foreign land. I would also like to acknowledge the support of my friends

at Cornell: Sripathi, Adrita, Rishabh, Sabyasachi, Abhradeep, Karan and Himani to name a few, who were ever so ready to help me brave the ups and downs of PhD life, and for whose company I can call Ithaca a second home. I am also thankful to the Cornell Club Squash team for helping me grow as an athlete and surrounding me with such smart and energetic minds, both on and off the court.

Most importantly, I am eternally grateful for the unconditional love and support of my parents and my fiancé, Trisala Mishra. They have been my pillars of strength through every challenge I have faced. I would not be where I am today without them by my side. I am thankful to my mother for helping me be prepared for the different stages of PhD in advance, my father for his unwavering and stoic faith in my capabilities when I was feeling the PhD pressure during the graduate mid-life crisis, and to Trisala for her constant support and for lending an ever-available ear throughout the last five years.

TABLE OF CONTENTS

Biographical Sketch . . . iii
Dedication . . . iv
Acknowledgements . . . v
Table of Contents . . . vii
List of Tables . . . xi
List of Figures . . . xii

1 Introduction 1

2 IRL for Bayesian Stopping Time Problems 5
2.1 Introduction . . . 5
2.1.1 Main results and context . . . 6
2.1.2 IRL for decisions made in multiple environments . . . 8
2.1.3 Context. Bayesian revealed preference for IRL . . . 10
2.2 Identifying optimal Bayesian stopping and reconstructing agent costs . . 12
2.2.1 Bayesian stopping agent . . . 12
2.2.2 Optimal Bayesian stopping agent in multiple environments . . . 14
2.2.3 IRL for inverse optimal stopping. Main result . . . 17
2.2.4 Discussion of Theorem 3 . . . 24
2.2.5 Discussion of (A1) and (A2) . . . 29
2.2.6 Outline of proof of Theorem 3 . . . 30
2.2.7 Summary . . . 32
2.3 Example 1. IRL for sequential hypothesis testing (SHT) . . . 33
2.3.1 Sequential hypothesis testing (SHT) Problem . . . 33
2.3.2 IRL for inverse SHT. Main assumptions . . . 34
2.3.3 IRL for inverse SHT. Main result . . . 36
2.3.4 Numerical example illustrating IRL for inverse SHT . . . 38
2.3.5 Numerical example. Regularized max-margin IRL for inverse SHT. 40
2.3.6 Performance Comparison. IRL for Inverse SHT and existing IRL methods for POMDPs . . . 43
2.3.7 Summary . . . 46
2.4 Example 2. IRL for inverse search . . . 47
2.4.1 Optimal Bayesian search agent in multiple search environments 48
2.4.2 IRL for inverse search. Main result . . . 51
2.4.3 Numerical example illustrating IRL for inverse search . . . 54
2.5 Inverse Optimal Stopping for Predicting YouTube Commenting Behavior 56
2.5.1 YouTube Dataset and Model Parameters . . . 58
2.5.2 YouTube Data Analysis Results . . . 60
2.6 Finite sample performance analysis of IRL decision test . . . 64
2.6.1 Finite sample statistical test for IRL . . . 65
2.6.2 Main result. Finite sample analysis for IRL . . . 67
2.6.3 Example 1. Finite sample effects for IRL in inverse SHT . . . 70

2.6.4 Example 2. Finite sample effects for IRL in inverse search . . . 74
2.7 Discussion and extensions . . . 77
2.8 Appendix . . . 80
2.8.1 Context and Perspective. IRL for Bayesian stopping problems . 80
2.8.2 Literature and Applications. IRL in Bayesian stopping problems 80
2.8.3 Related works in IRL . . . 82
2.9 Proof of Lemma 2 . . . 87
2.10 Proof of Theorem 3 . . . 88
2.10.1 Necessity of NIAS, NIAC inequalities . . . 89
2.10.2 Sufficiency of NIAS, NIAC inequalities for Bayes optimal stopping 91
2.10.3 Remarks . . . 95
2.11 Proof of Theorem 9 . . . 97
2.12 Proof of Theorem 13 . . . 97
2.13 Proof of Theorem 15 . . . 101
2.14 Context. IRL For Predicting YouTube Commenting Behavior . . . 103

3 IRL for Identification and Mitigation of Adversary Sensing Systems 106
3.1 Introduction . . . 106
3.2 Inverse Tracking and Estimating Adversary's Sensor . . . 111
3.2.1 Background and Preliminary Work . . . 111
3.2.2 Inverse Tracking Algorithms . . . 114
3.2.3 Estimating the Adversary's Sensor Gain . . . 117
3.2.4 Example. Estimating Adversary's Gain in Linear Gaussian case 118
3.3 Identifying Utility Maximization in a Cognitive Radar . . . 122
3.3.1 Background. Revealed Preferences and Afriat's Theorem . . . 124
3.3.2 Beam Allocation: Revealed Preference Test . . . 126
3.3.3 Waveform adaptation: Revealed Preference Test for Non-linear budgets . . . 128
3.4 Designing Smart Interference To Confuse Cognitive Radar . . . 132
3.4.1 Interference Signal Model . . . 133
3.4.2 Smart Interference to confuse the adversary radar . . . 135
3.4.3 Numerical example illustrating design of smart interference . . 138
3.5 Conclusion . . . 139

4 Unification of Economics-based IRL for Bayesian and non-Bayesian decision systems 142
4.1 Introduction . . . 142
4.2 Background . . . 150
4.2.1 Revealed preference (non-linear budget) . . . 150
4.2.2 Modifying the revealed preference test: Observed utility, unobserved budget constraint . . . 151
4.2.3 Revealed Rational Inattention . . . 155
4.2.4 Revealed rational inattention and observability of Bayesian agent's decisions by the analyst . . . 160

4.3 Main Result. Unification of Revealed preference and Revealed rational inattention . . . 163
4.3.1 Discussion of Theorem 27 . . . 165
4.3.2 Change of partial order from revealed preference to revealed rational inattention . . . 168
4.3.3 Generalizing Revealed Rational Inattention to Variable Attention Spans . . . 170
4.4 Related Works . . . 174
4.5 Extending Robustness Measures in Revealed Preference to Revealed Rational Inattention . . . 176
4.5.1 Afriat Efficiency Index for Rational Inattention (RI-AEI) . . . 178
4.5.2 Minimum Perturbation Test for Rational Inattention (RI-MPT) . 180
4.5.3 Money Pump Index for Rational Inattention (RI-MPI) . . . 182
4.6 Conclusion and Future Work . . . 185
4.7 Appendix . . . 186
4.7.1 Proof of Theorem 25 . . . 186
4.7.2 Proof of Theorem 27 . . . 188
4.7.3 Proof of Lemma 29 . . . 190
4.7.4 Proof of Corollary 30 . . . 192

5 IRL as a Principled Approach for Interpreting Deep Neural Networks 198
5.1 Introduction . . . 198
5.1.1 Summary of Results. . . . 199
5.1.2 Related Works . . . 203
5.2 Bayesian Revealed preference with Rational Inattention . . . 204
5.2.1 Utility Maximization with Rational Inattention (UMRI) . . . 205
5.2.2 Bayesian Revealed Preference (BRP) Test for Rationally Inattentive Utility Maximization . . . 207
5.2.3 Relating UMRI and BRP test to Interpretable Deep Image Classification . . . 209
5.3 Bayesian Revealed Preference explains CIFAR-10 Image Classification by Deep CNNs . . . 214
5.3.1 BRP and S-BRP Tests for deep CNN datasets. Results and Insights 216
5.3.2 Predicting deep CNN classification accuracy using our Interpretable Models . . . 221
5.4 Conclusions and Extensions . . . 225
5.5 Appendix . . . 227
5.5.1 Proof of Theorem 35 . . . 227
5.5.2 S-UMRI (Sparse UMRI) Model for Rationally Inattentive Bayesian Utility Maximization . . . 229
5.5.3 Construction of Deep CNN Dataset . . . 232

6 Inverse-Inverse Reinforcement Learning. Spoofing an Inverse Reinforcement Learner **234**
6.1 Introduction . . . 234
6.2 Background. IRL to Estimate Cognitive Radar . . . 238
6.2.1 Radar-Adversary Dynamics . . . 238
6.2.2 Radar Cognition as Constrained Utility Maximization . . . 240
6.2.3 Adversarial IRL for Identifying Strategy of Cognitive Radar . . . 246
6.2.4 Examples of IRL for Identifying Radar Cognition . . . 250
6.3 Inverse IRL. Masking Radar Utility and Constraints from Adversarial IRL 254
6.4 How to Mask Cognition from Detector? . . . 260
6.4.1 Noisy Adversarial IRL Detectors against Cognitive Radars . . . 261
6.4.2 Masking Cognition from IRL Detectors . . . 266
6.5 Numerical Results for I-IRL . . . 269
6.5.1 Cognition Masking via Theorem 44 for Noise-less Adversary Measurements . . . 270
6.5.2 Cognition Masking via Theorem 48 for Noisy Adversary Measurements . . . 273
6.6 Conclusion and Extensions . . . 276
6.7 Appendix . . . 278
6.7.1 Feasibility Test for Adversarial IRL . . . 278
6.7.2 IRL for Identifying Radar's Resource Constraint . . . 280
6.7.3 Example. Optimal Beam Allocation . . . 281
6.7.4 Proof of Theorem 46 . . . 284
6.7.5 Masking Radar's Utility Function for Multiple Constraints . . . 285
6.7.6 Cognition Masking Performance under Misspecified Radar Response Measurements . . . 288
6.7.7 Cognition Masking for Arbitrary IRL Algorithm . . . 291
6.7.8 Context. ECCM and Meta-Cognition . . . 292

7 Concluding Remarks **295**

Bibliography **297**

LIST OF TABLES

2.1 IRL Identifiability of Optimal Bayesian stopping. 19

3.1 Comparison of sensitivity values (3.16) for log-likelihood of C wrt noise covariances $Q_k,\ R_k$ (3.4) - classical model vs inverse Kalman filter model. 121
3.2 Comparison of Cramér-Rao bounds for C - classical model vs inverse Kalman filter model. 122

5.1 Numerical experiments illustrating the goodness-of-fit of the constrained Bayesian utility maximization model to deep image classification . . . 216

6.1 Examples of Cognitive Radars in the literature abstracted as constrained utility maximizers 245
6.2 Numerical Experiments to illustrate how Inverse IRL (I-IRL) spoofs Afriat's Theorem based adversarial IRL 270

LIST OF FIGURES

2.1 Set diagrammatic view of strategy optimality for an inverse learner. Global optimality, relative optimality and ε-optimality 31
2.2 Key Idea behind IRL for optimal Bayesian stopping. Mapping the space of observation sequence $y_{1:\tau(\mu)}$ to a fictitious alphabet space 32
2.3 Numerical experiments illustrating feasibility tests for inverse SHT . . 39
2.4 Numerical Experiments. Inverse SHT with L_1-regularization on misclassification costs for 100 SHT environments 41
2.5 Inverse SHT Performance Comparison. Max-Margin NIAS-NIAC Test of Theorem 6 versus MMV and MMFE [1] 46
2.6 Numerical Experiments illustrating feasibility tests for inverse Bayesian search 56
2.7 YouTube Dataset Overview. Video Viewcount summed over all videos from 18 categories 58
2.8 Max-margin IRL and Entropy-regularized IRL prediction results on the YouTube user engagement dataset 64

3.1 Schematic Illustration of the Sense-Learn-Adapt (SLA) paradigm in context of the adversarial mitigation results in this chapter 108
3.2 Schematic illustration of the emission exchange between the cognitive radar and adversary entity in the SLA paradigm 111
3.3 Log-Likelihood plotted as a function of adversary's gain. The key takeaway is that the likelihood curve is flatter in the inverse filtering problem, and hence, computing the MLE is more difficult. 120
3.4 Micro-economic abstraction of interaction between cognitive radar and adversarial target 124
3.5 Schematic of transmit channel H_t, clutter channel H_c and interference signal $\boldsymbol{H}_p$ involving an adversarial cognitive radar and us. We observe the radar's waveform $\boldsymbol{W}$ in noise. The aim is to engineer the interference signal $\boldsymbol{H}_p$ to confuse the adversary cognitive radar. 132
3.6 Schematic of pulse-level smart interference design to confuse the cognitive radar. 136
3.7 Performance illustration of the proposed smart signal interference scheme 140

4.1 Schematic illustration of the relation between GARP in micro-economics and NIAC in information economics 146
4.2 Schematic of the deep auto-encoder architecture used to pre-process YouTube user engagement data 147

5.1 Schematic of the proposed rationally inattentive utility maximization model for interpreting deep convolutional neural networks 201
5.2 Illustration of the interpretable utility values computed for all image categories in the CIFAR-10 dataset 217

5.3 Illustration of the interpretable learning (rational inattention) costs computed for different neural networks architectures 221
5.4 Category-wise prediction of classification accuracy by the interpretable model for different neural network architectures 222

6.1 Schematic of adversarial IRL mitigating cognitive radars by identifying the radar utility 255
6.2 Schematic of the proposed inverse IRL (I-IRL) idea to safeguard system utility from adversarial IRL 258
6.3 Schematic illustration of I-IRL. Transmitting smart sub-optimal signals masks system utility from adversarial IRL 262
6.4 Numerical experiments illustrating how I-IRL successfully masks radar utility for optimal waveform adaptation from adversarial IRL 272
6.5 Numerical experiments illustrating how I-IRL successfully masks radar utility for optimal beam allocation from adversarial IRL 273
6.6 Numerical experiments illustrating how stochastic gradient-based I-IRL masks radar utility when the adversary uses an IRL detector for detecting utility maximization under noisy radar responses 276

CHAPTER 1

INTRODUCTION

The thesis is organized into the following chapters:

1. *IRL for Bayesian Stopping Time Problems.* This chapter presents an inverse reinforcement learning (IRL) framework for Bayesian stopping time problems. By observing the actions of a Bayesian decision maker, we provide a necessary and sufficient condition to identify if these actions are consistent with optimizing a cost function. In a Bayesian (partially observed) setting, the inverse learner can at best identify optimality wrt the observed strategies. Our IRL algorithm identifies optimality and then constructs set-valued estimates of the cost function. To achieve this IRL objective, we use novel ideas from Bayesian revealed preferences stemming from microeconomics. We illustrate the proposed IRL scheme using two important examples of stopping time problems, namely, sequential hypothesis testing and Bayesian search. Finally, for finite datasets, we propose an IRL detection algorithm and give finite sample bounds on its error probabilities.
2. *IRL for Identification and Mitigation of Adversary Sensing Systems.* This chapter considers three inter-related adversarial inference problems involving cognitive radars. We first discuss inverse tracking of the radar to estimate the adversary's estimate of us based on the radar's actions and calibrate the radar's sensing accuracy. Second, using revealed preference from microeconomics, we formulate a nonparametric test to identify if the cognitive radar is a constrained utility maximizer with signal processing constraints. We consider two radar functionalities, namely, beam allocation and waveform design, with respect to which the cognitive radar is assumed to maximize its utility and construct a set-valued estimator for the radar's utility function. Finally, we discuss how to engineer interference at the physical

layer level to confuse the radar which forces it to change its transmit waveform. The levels of abstraction range from smart interference design based on Wiener filters (at the pulse/waveform level), inverse Kalman filters at the tracking level and revealed preferences for identifying utility maximization at the systems level.

3. *Unification of Economics-based IRL for Bayesian and non-Bayesian agents.* This chapter unifies two key results from economic theory, namely, revealed rational inattention [2] and classical revealed preference [3,4]. Revealed rational inattention tests for rationality of information acquisition for Bayesian decision makers. On the other hand, classical revealed preference tests for utility maximization under known budget constraints. Our first result is an equivalence result - we unify revealed rational inattention [2] and revealed preference [3–5] through an equivalence map over decision parameters and partial order for payoff monotonicity over the decision space in both setups. Second, we exploit the unification result computationally to extend robustness measures for goodness-of-fit of revealed preference tests in the literature to revealed rational inattention. This extension facilitates quantifying how well a Bayesian decision maker's actions satisfy rational inattention.

4. *Rationalizing User Engagement Dynamics in Online Multimedia using IRL techniques.* Although not presented a separate chapter, all our theoretical results are illustrated on a real-world YouTube dataset comprising thumbnail, title and user engagement metadata from approximately 140,000 videos. All our numerical experiments are completely reproducible.

   (i) We compute the Bayesian analog of robustness measures from revealed preference literature on YouTube metadata features extracted from a deep auto-encoder, i.e., a deep neural network that learns low-dimensional features of the metadata. The computed robustness values show that YouTube user engagement fits the rational inattention model remarkably well.

(ii) We illustrate our IRL results for optimal Bayesian stopping on the YouTube dataset by predicting YouTube user engagement with high accuracy.

5. *IRL as a Principled Approach for Interpreting Deep Neural Networks.* Can deep convolutional neural networks (CNNs) for image classification be interpreted as utility maximizers with information costs? By performing set-valued system identification for Bayesian decision systems, this chapter demonstrate that deep CNNs behave equivalently (in terms of necessary and sufficient conditions) to rationally inattentive Bayesian utility maximizers, a generative model used extensively in economics for human decision-making. Our claim is based on approximately 500 numerical experiments on 5 widely used neural network architectures. The parameters of the resulting interpretable model are computed efficiently via convex feasibility algorithms. As a practical application, we also illustrate how the the reconstructed interpretable model can predict the classification performance of deep CNNs with high accuracy. The theoretical foundation of our approach lies in Bayesian revealed preference studied in micro-economics. All our results are on GitHub and completely reproducible.

6. *Inverse-Inverse Reinforcement Learning. Spoofing an Inverse Reinforcement Learner.* A cognitive radar is a constrained utility maximizer that adapts its sensing mode in response to a changing target environment. If an adversary can estimate the utility function of a cognitive radar, it can determine the radar's sensing strategy and mitigate the radar performance via electronic countermeasures (ECM). This chapter discusses how a cognitive radar can *hide* its strategy from an adversary that detects cognition. The radar does so by transmitting purposefully designed sub-optimal responses to spoof the adversary's Neyman-Pearson detector. We provide theoretical guarantees by ensuring the Type-I error probability of the adversary's detector exceeds a predefined level for a specified tolerance on the

radar's performance loss. We illustrate our cognition masking scheme via numerical examples involving waveform adaptation and beam allocation. We show that small purposeful deviations from the optimal strategy of the radar confuse the adversary by significant amounts, thereby masking the radar's cognition. Our approach uses ideas from revealed preference in microeconomics and adversarial inverse reinforcement learning. Our proposed algorithms provide a principled approach for system-level electronic counter-countermeasures (ECCM) to mask the radar's cognition, i.e., hide the radar's strategy from an adversary. We also provide performance bounds for our cognition masking scheme when the adversary has misspecified measurements of the radar's response.

CHAPTER 2

IRL FOR BAYESIAN STOPPING TIME PROBLEMS

2.1 Introduction

In a stopping time problem, a decision maker obtains noisy observations of a random variable (state of nature) x sequentially over time. Based on the observation history (sigma-algebra generated by the observations), the decision maker decides at each time whether to continue or stop. If the decision maker chooses the continue action, it pays a continuing cost and obtains the next observation. If the decision maker chooses the stop action at a specific time, then the problem terminates, and the decision maker pays a stopping cost. In a *Bayesian* stopping time problem, the decision maker knows the prior distribution of state of nature x and the observation likelihood (conditional distribution of the observations) $p(y|x)$ given the state x, and uses this information to update its belief and choose its continue and stop actions. Finally, in an *optimal* Bayesian stopping time problem, the decision maker chooses its continue and stop actions to minimize an expected cumulative cost function.

Traditional inverse reinforcement learning (IRL) aims to estimate the costs/rewards of a decision maker by observing its actions and was first studied by [6] and [7]. This chapter considers IRL for Bayesian stopping time problems. Suppose an inverse learner observes the actions of a decision maker performing Bayesian sequential stopping in *multiple environments*. The decision maker has a fixed observation likelihood and observation cost, and incurs a different stopping cost in each environment[1]. The inverse learner does not know the realizations of the observation sequence nor the observation likelihood

[1]We refer the reader to [8, Ch. 3.5] and [9] for motivating the need to for multiple environments for identifiability of Markov decision processes (MDPs).

of the decision maker; the inverse learner only knows the true state x and observes the stopping action a of the decision maker. The two main questions we address are:

1. How can the inverse learner check if the actions of a Bayesian decision maker are consistent with optimal stopping?
2. If the decision maker's actions are consistent with optimal stopping, how can the inverse learner estimate the stopping costs of the multiple environments?

2.1.1 Main results and context

The key results in this chapter are summarized as follows:

1. Inverse RL for Bayesian sequential stopping: Theorem 3 in Sec. 2.2 is our first key IRL result. Theorem 3 specifies a set of convex inequalities that are simultaneously necessary and sufficient for the actions of a Bayesian decision maker in multiple environments to be consistent with optimal stopping. If so, then Theorem 3 provides an algorithm for the inverse learner to generate set-valued estimates of the decision maker's costs in the multiple environments. Theorem 3 is especially useful in scenarios where the inverse learner has no knowledge of the decision maker's observation likelihood or observation sample paths, and yet can construct a set-valued estimate of the costs incurred by the decision maker.

2. Inverse RL for SHT and Search: Sec. 2.3 and Sec. 2.4 construct IRL algorithms for two specific examples of Bayesian stopping time problems, namely, Sequential Hypothesis Testing (SHT) and Search. The main results, Theorem 6 and Theorem 9 specify necessary and sufficient conditions for the decision maker's actions to be consistent with optimal SHT and optimal search, respectively. If the conditions hold, Theorems 6 and 9 provide algorithms to estimate the incurred misclassification costs (for SHT) and

search costs (for Bayesian search). In Sec. 2.3 for inverse SHT, we also propose an IRL algorithm to compute a point-estimate of the decision maker's costs. The point-estimate is computed by maximizing the regularized margin of the convex feasibility test for inverse SHT proposed in Theorem 6 and estimates the misclassification costs with high accuracy. Also, in Sec. 2.3.6, we compare numerically the performance of the IRL algorithm in Theorem 3 with two existing IRL algorithms [1] in the literature. This numerical comparison highlights how the IRL approach in this chapter complements the results of [1]. Theorem 6 achieves IRL when the inverse learner has partial information about the decision maker's costs.

3. Illustration of Inverse RL for Bayesian stopping on Real-World Dataset: One important use case of IRL is to extract preferences from expert human agents [10, 11]. In Sec. 2.5, we illustrate how our IRL algorithms extend to predicting human-level online multimedia user engagement using a massive YouTube dataset comprising video metadata from approximately 190000 videos.[2] From the set of costs that pass the convex feasibility test in Theorem 3 for optimal Bayesian stopping, we chose two point-valued IRL costs for IRL prediction, namely, max-margin IRL and entropy-regularized IRL. The main finding is that both point estimates accurately predict YouTube commenting behavior. Also, we observe that the max-margin estimate yields a more accurate prediction compared to the entropy-regularized estimate (in terms of the chi-square and total variation distance).

4. Sample Complexity for IRL: In Sec. 2.6, we propose IRL detection tests for optimal stopping, optimal SHT and optimal search under finite sample constraints. Theorems 11, 13 and 15 in Sec. 2.6 comprise our sample complexity results that characterize the robustness of the detection tests by specifying Type-I and Type-II error bounds for the IRL detection tests. To the best of our knowledge, our finite sample complexity results

[2]Although understanding YouTube commenting behavior was the main focus of our previous work [12], the inference methodology and numerical experiments in this chapter are new; see Sec. 2.14 and `https://github.com/KunalP117/YouTube-Commenting-Analysis` for details.

for the IRL detector, namely, the sample size required to achieve a Type-I or Type-II error probability below a specified value for IRL, are novel.

The proofs of all theorems are provided in the Appendix. For a practitioner's perspective, our key IRL algorithms are Theorems 3, 6 and 9, and finite sample complexity results for IRL error bounds are Theorems 11, 13 and 15.

2.1.2 IRL for decisions made in multiple environments

An important aspect of our IRL framework is that the inverse learner observes the decision maker in multiple environments.[3] The purpose of this section is to motivate this framework.

We consider a decision maker operating over M environments. In each environment, the decision maker solves a stopping time problem with a distinct stopping cost. The decision maker has a fixed observation likelihood (sensor accuracy) and sensing cost (operating cost), where both variables are invariant across multiple environments. Therefore, there are up to M distinct strategies exhibited by the decision maker, one for each environment. Let $J(\mu_m, s_m)$ denote the expected cost incurred by the decision maker when it chooses stopping strategy μ_m in environment m with stopping cost s_m.

Consider now the inverse learner that observes the decision maker. Assume that the inverse learner does not know the stopping costs s_m, but only observes[4] the set of strategies $\{\mu_m, m = 1, 2, \ldots, M\}$. To achieve IRL, the inverse learner must first establish

[3]The inverse learner in this chapter can be viewed as a *passive* analyst that does not control the environment variables, that is, the agent's stopping costs. An interesting extension of this chapter (for future work) is to consider an *active* inverse learner that purposefully adapts the environment variables to minimize IRL detection errors.

[4]We are deliberately simplifying the IRL framework here for explanatory reasons. Our main result assumes the inverse learner only observes the actions of the decision maker, and not the strategy.

if the decision maker's strategy in each environment is consistent with minimizing an expected cost. Equivalently, the inverse learner must check if the expected cost incurred by the decision maker in environment m by choosing strategy μ_m is less than that incurred by all other (infinitely many) stopping strategies. However, the inverse learner does not observe infinitely many strategies, but only M strategies. Given the decision maker's strategies in M, each with a distinct stopping cost, the inverse learner's procedure to identify if the decision maker is optimal or not is defined below:

IRL identifiability of optimal stopping agent. *Consider a Bayesian stopping agent that chooses strategy μ_m in environment m, over multiple environments $m = 1, 2, \ldots, M$. Then, identifying an optimal Bayesian stopping time agent is equivalent to checking if the following inequalities have a feasible solution:*

$$\textit{There exists } s_1, s_2, \ldots, s_M \textit{ such that: } J(\mu_m, s_m) \leq J(\mu_n, s_m),\ \forall\, m, n. \tag{2.1}$$

Here, $J(\mu_m, s_n)$ is the decision maker's cumulative expected cost when the decision maker chooses strategy μ_m and incurs a stopping cost s_n.

The solution of the feasibility problem in (2.1) is the set-valued IRL estimate of the stopping costs incurred by the decision maker. The comparison in (2.1) between the performance of the decision maker's strategy in each environment to the strategies chosen in all other (finitely many) environments is formalized in Lemma 2 and is achieved by the inverse learner via the IRL procedure in Theorem 3. We also refer the reader to the seminal work of [8, Ch. 3.5] and [13] on identifiability of MDPs for further justification of multiple environments. The above framework of a Bayesian stopping time agent operating in multiple environments arises in several applications; see Sec. 2.8.2 for details.

2.1.3 Context. Bayesian revealed preference for IRL

The formalism used in this chapter to achieve IRL is *Bayesian revealed preferences* studied in microeconomics by [14, 15] and [2]; see Sec. 2.8.3 for more details. This Bayesian revealed preference-based approach *complements* existing IRL results for partially observed Markov decision processes (POMDP) including [1]. This chapter considers a subset of POMDPs, namely, Bayesian stopping time problems. Due to the problem structure, we show that our IRL algorithms *do not* require knowledge of the observation likelihood of the decision maker and also do not require solving a POMDP.

We now briefly discuss how the Bayesian revealed preference based IRL approach differs from classical IRL.

1. The classical IRL frameworks [6, 7] assume the observed agent is a reward maximizer (or equivalently, cost minimizer) and then seeks to estimate its cost function. The approach in this chapter is more fundamental. We first *identify* if the decisions of a single decision maker in multiple environments are consistent with optimality and if so, we then generate set-valued estimates of the costs that are consistent with the observed decisions.
2. Classical IRL assumes complete knowledge of the decision maker's observation likelihood. We assume the inverse learner only knows the state of nature and the action chosen when the decision maker stops, and does not know its observation likelihoods or the sequence of observation realizations. Two important scenarios where this situation arises are:

   (i) *Multimedia Datasets.* In online multimedia datasets such as the YouTube dataset analyzed in Sec. 2.5, it is impossible to know the attention span (observation likelihood) of the online user. All that is available are the online user's actions (interactions such as comments and comment ratings) and the underlying state of

nature (video metadata such as viewcount, thumbnail and video description); see also [12].

(ii) *Adversarial Signal Processing.* In adversarial signal processing and sensing applications, it is not realistic for the inverse learner to know the model dynamics of the agent. An important example is IRL for radars [16], where the radar is the adversary and so it is impossible to know its sensing modes (observation likelihood); however, the inverse learner records the electromagnetic waveforms (response) emitted by the radar.

Additional applications where only the agent decisions are available for IRL (and not the observation likelihood) include consumer insights and advertisement design research, interpretable ML in smart healthcare and electronic warfare. These are discussed in Sec. 2.8.1.

3. *Algorithmic Issues:* In classical IRL [7], the inverse learner solves the Bayesian stopping time problem iteratively for various choices of the cost. This can be computationally prohibitive since it involves stochastic dynamic programming over a belief space which is PSPACE hard [17]. The IRL procedure in this chapter does not require solving a POMDP and only requires testing for the feasibility of a set of convex inequalities.

For brevity, we discuss related IRL literature and applications of IRL for Bayesian stopping problems in Sec. 2.8.1.

2.2 Identifying optimal Bayesian stopping and reconstructing agent costs

Our IRL framework comprises a decision maker's actions in a stopping time problem over M environments and an *inverse learner* that observes these actions. This section defines the IRL problem that the inverse learner faces and then presents two results regarding the inverse learner:

1. *Identifying Optimal Stopping*. Theorem 3 below provides a necessary and sufficient condition for the inverse learner to identify if the Bayesian decision maker chooses its actions as the solution of an optimal stopping problem.
2. *IRL for Reconstructing Costs*. Theorem 3 is also constructive. It shows that the continue and stopping costs of the Bayesian decision maker can be reconstructed by solving a convex feasibility problem.

This section provides a complete IRL framework for Bayesian stopping time problems and sets the stage for subsequent sections where we formulate generalizations and examples.

2.2.1 Bayesian stopping agent

A Bayesian stopping time agent is parametrized by the tuple

$$\Xi = (\mathcal{X}, \pi_0, \mathcal{Y}, \mathcal{A}, B, \mu) \tag{2.2}$$

where

- $\mathcal{X} = \{1, 2, \dots X\}$ is a finite set of states.

- At time 0, the true state $x^o \in \mathcal{X}$ is sampled from prior distribution π_0. x^o is unknown to the agent.
- $\mathcal{Y} \subset \mathbb{R}$ is the observation space. Given state x^o, the observations $y \in \mathcal{Y}$ have conditional probability density $B(y, x^o) = p(y|x^o)$.
- $\mathcal{A} = \{1, 2, \dots A\}$ is the finite set of stopping actions.
- Finally, μ denotes the agent's stopping strategy. The stopping strategy operates sequentially on a sequence of observations $y_1, y_2, \dots$ as discussed below in Protocol 1.

Protocol 1 *Sequential Decision-making protocol: Assume the agent knows* Ξ.

1. *Generate* $x^o \sim \pi_0$*, at time* $t = 0$*. Here* x^o *is not known to the agent.*
2. *At time* $t > 0$*, agent records observation* $y_t \sim B(\cdot, x^o)$.
3. Belief Update: *Let* $\mathcal{F}_t$ *denote the sigma-algebra generated by observations* $\{y_1, y_2, \dots y_t\}$*. The agent updates its belief (posterior)* $\pi_t(x) = \mathbb{P}(x^o = x|\mathcal{F}_t), x \in \mathcal{X}$ *using Bayes formula as*
$$\pi_t = \frac{B(y_t)\pi_{t-1}}{\mathbf{1}'B(y_t)\pi_{t-1}}, \tag{2.3}$$
where $B(y) = \mathrm{diag}(\{B(y, x), x \in \mathcal{X}\})$*. The belief* π_t *is an* X*-dimensional probability vector in the* $X - 1$ *dimensional unit simplex*
$$\Delta(\mathcal{X}) \stackrel{def.}{=} \{\pi \in \mathbb{R}_+^X : \mathbf{1}'\pi = 1\}. \tag{2.4}$$
4. *Choose action* $a_t = \mu(\pi_t, t)$ *from the set* $\mathcal{A} \cup \{$*continue*$\}$*. If* $a_t \in \mathcal{A}$*, then stop, else if* $a_t =$ *continue*, *set* $t = t + 1$ *and go to Step 2.*

The stopping strategy μ is a (possibly randomized) time-dependent mapping from the agent's belief at time $t \in \mathbb{Z}^+$ to the set $\mathcal{A} \cup \{\text{continue}\}$ and belongs to $\boldsymbol{\mu}$, the set of

admissible stopping strategies:

$$\boldsymbol{\mu} = \{\mu : \Delta(\mathcal{X}) \times \mathbb{Z}^+ \to \mathcal{A} \cup \{\text{continue}\}\}. \tag{2.5}$$

We define the random variable τ as the time when the agent stops and takes a stop action from $\mathcal{A}$.

$$\tau = \inf\{t \geq 0 |\ \mu(\pi_t, t) \neq \{\text{continue}\}\}. \tag{2.6}$$

Clearly, the set $\{\tau = t\}$ is measurable wrt $\mathcal{F}_t$, the sigma-algebra generated by observations $\{y_1, y_2, \ldots y_t\}$. Hence, the random variable τ is adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$. In the following sub-section, we will introduce costs for the agent's stop and continue actions. We will use τ for expressing the expected cumulative cost of the agent.

To summarize, a Bayesian stopping agent is parameterized by Ξ and operates according to Protocol 1. Several decision problems such as SHT and sequential search fit this formulation.

2.2.2 Optimal Bayesian stopping agent in multiple environments

So far we have defined a Bayesian stopping agent. Our main IRL result is to identify if a Bayesian stopping agent's behavior in a set of environments $\mathcal{M}$ is *optimal*. The purpose of this section is to define optimal Bayesian stopping [18] in multiple environments. For identifiability reasons (see assumption (A2) below) we require at least two environments ($M \geq 2$).

An optimal Bayesian stopping agent in multiple environments is defined by the tuple

$$\Xi_{opt} = (\Xi, \mathcal{M}, \mathcal{C}, \boldsymbol{s}, \boldsymbol{\mu}^*). \tag{2.7}$$

In (2.7),

- $\mathcal{M}$ is the set of M environments.
- The parameters $\mathcal{X}, \mathcal{Y}, \mathcal{A}, \pi_0, p$ in Ξ (2.2) and continue cost $\mathcal{C}$ (defined below) are the same for all environments in $\mathcal{M}$.
- $\mathcal{C} = \{c_t\}_{t\geq 0}$, $c_t(x) \in \mathbb{R}^+$ is the continue cost incurred in any environment $m \in \mathcal{M}$ at time t given state $x^o = x$.
- $\boldsymbol{s} = \{s_m(x,a), x \in \mathcal{X}, a \in \mathcal{A}, m \in \mathcal{M}\}$, $s_m(x,a) < \infty$ is the cost for taking stop action a when the state $x^o = x$ in the m^{th} environment.
- $\boldsymbol{\mu}^* = \{\mu_m^*, m \in \mathcal{M}\}$ is the set of **optimal** stopping strategies of the Bayesian stopping agent over the set of environments $\mathcal{M}$, where the optimality is defined in Definition 1 below. In environment m, the Bayesian stopping agent employs its stopping strategy $\mu_m^*, m \in \mathcal{M}$ and operates according to Protocol 1.

Definition 1 (Optimal Stopping Strategy) *For each environment $m \in \mathcal{M}$, strategy μ_m^* is optimal for stopping cost $s_m(x,a)$ iff the following conditions hold:*

$$\mu_m^*(\pi, \tau) = \operatorname*{argmin}_{a \in \mathcal{A}} \pi' \bar{s}_{m,a}, \tag{2.8}$$

$$J(\mu_m^*, s_m) = \inf_{\mu \in \boldsymbol{\mu}} J(\mu, s_m), \tag{2.9}$$

Recall $\boldsymbol{\mu}$ (2.5) denotes the set of all stopping strategies. Also $J(\mu, s_m)$ is the expected cumulative cost defined as:

$$\begin{aligned} &J(\mu, s_m) = G(\mu, s_m) + C(\mu), \text{ where} \\ &G(\mu, s_m) = \mathbb{E}_\mu \left\{ \pi_\tau' \bar{s}_{m,\mu(\pi_\tau,\tau)} \right\}, \; C(\mu) = \mathbb{E}_\mu \left\{ \sum_{t=0}^{\tau-1} \pi_t' \bar{c}_t \right\}, \; \mu \in \boldsymbol{\mu}. \end{aligned} \tag{2.10}$$

$\mathbb{E}_\mu$ denotes expectation parametrized by μ wrt the probability measure induced by $y_{1:\tau}$. Also, $\bar{s}_a$, $\bar{c}_t$ are the stopping and continue[5] cost vectors, respectively, vectorized over states $x \in \mathcal{X}$.

[5]Since the continue cost is a positive real, the stopping time τ (2.6) is finite a.s.

Definition 1 is standard for the optimal strategy in a sequential stopping problem [19]. The optimal strategy naturally decomposes into two steps: choosing whether to continue or stop according to (2.9); and if the decision is to stop, then choose a specific stopping action from $\mathcal{A}$ according to (2.8). The optimal stopping strategies $\mu_m, m \in \mathcal{M}$ that satisfy the conditions (2.8), (2.9) can be obtained by solving a stochastic dynamic programming problem [19]. It is a well-known result [20] that the set of beliefs for which it is optimal to stop is convex.

Relation to Bayesian contextual bandits

For readers familiar with the multi-armed bandit problem, optimal Bayesian stopping can be viewed as an instance of the *partially-observed regularized contextual Bayesian bandit problem*; *contextual* [21] since the agent faces multiple ground truths x (context), *partially observed* [22] since the agent observes a sequence of noisy measurements of the underlying context x, *Bayesian* [23] since the agent minimizes its expected cumulative cost per context averaged over all contexts sampled from a prior distribution π_0, and *regularized* [24] since the agent minimizes the sum of expected stopping cost and a regularization term, namely, the expected continue cost. Loosely speaking, this chapter addresses the problem of IRL for partially-observed regularized contextual bandits. Although our IRL results are introduced in subsequent sections, we remark here that there is ample scope to extend the results in this chapter to typical RL decision frameworks that allow underlying state transitions. At a high level, this can be made possible by constructing feasibility tests in terms of the state-occupancy measure induced by the decision maker's policy in multiple environments.

2.2.3 IRL for inverse optimal stopping. Main result

We now discuss an inverse learner-centric view of the Bayesian stopping time problem and the main IRL result. Suppose the inverse learner observes the actions of a Bayesian stopping agent in M environments, where each environment is characterized by the stopping costs incurred by the agent. Suppose the agent performs several independent trials of Protocol 1 in all M environments. We make the following assumptions about the inverse learner performing IRL.

(A1) The inverse learner knows the dataset

$$\mathcal{D}_M = (\pi_0, \boldsymbol{p}), \text{ where } \boldsymbol{p} = \{p_m(a|x), x \in \mathcal{X}, a \in \mathcal{A}, m \in \mathcal{M}\}. \tag{2.11}$$

In (2.11), $p_m(a|x)$ is the Bayesian stopping agent's conditional probability of choosing stop action a at the stopping time given state $x^o = x$ in the m^{th} environment. We call $p_m(a|x)$ as the agent's *action selection policy*.

Note that:

(i) The inverse learner does not know the stopping times; it only has access to the conditional density of which stop action a was chosen given the true state x^o.

(ii) We assume the decision maker visits all states in the support of the prior pmf π_0 (2.11) infinitely often. In Sec. 2.6, we address the case where the decision maker visits the states finitely often and provide IRL performance guarantees via finite sample complexity.

(A2) Dataset $\mathcal{D}_M$ is generated by a Bayesian agent acting in at least $M \geq 2$ environments, where each environment has distinct stopping costs.

Both assumptions are discussed below after the main theorem, but let us make some preliminary remarks at this stage. (A1) implies the inverse learner observes the stopping

actions chosen by a Bayesian stopping agent in a finite number (M) of environments, where the agent performs an infinite number of independent trials of Protocol 1 in each environment; see discussion in Sec. 2.2.5 for asymptotic interpretation. In Sec. 2.6 we will consider finite sample effects where the inverse learner observes the agent performing a finite number of independent trials of Protocol 1. Assumption (A2) is necessary for the inverse optimal stopping problem to be well-posed.

Let μ_m denote the policy chosen by the agent in the m^{th} environment, and $\boldsymbol{\mu}_{\mathcal{M}} = \{\mu_m, m \in \mathcal{M}\}$ denote the set of chosen strategies.[6] The finite assumption on $|\mathcal{M}|$ in (A1) imposes a restriction on our IRL task of identifying optimality of a Bayesian stopping agent formalized below:

Lemma 2 (IRL identifiability of optimal Bayesian stopping agent.) *Given the dataset* $\mathcal{D}_M$ (2.11)*, the inverse learner can identify an optimal Bayesian stopping agent* (2.7) *acting in* M *environments if and only if* (2.8) *and the following relaxation of* (2.9) *holds:*

$$G_{m,m} + C_m \leq G_{n,m} + C_n, \ \forall m, \ n \in \mathcal{M}, \ m \neq n. \tag{2.12}$$

In (2.12)*,* $G_{n,m} = G(\mu_n, s_m)$ *is the expected stopping cost and* $C_m = C(\mu_m)$ *is the expected cumulative continue cost for the policy* μ_m *chosen in environment* $m, \ m \in \mathcal{M}$*.*

The proof of Lemma 2 is in Sec. 2.9. Lemma 2 formalizes the IRL identification procedure of the inverse learner in (2.1). Since the inverse learner only observes the agent's actions from M strategies chosen by the stopping agent, the best the inverse learner can do is check if μ_m is optimal for environment m out of the finite strategies in $\boldsymbol{\mu}_{\mathcal{M}}$. Indeed, the expected stopping cost $G_{n,m}$ is a function of the policy μ_n. However, in Sec. 2.10, we

[6] Recall that μ is a generic variable of a stopping policy, $\boldsymbol{\mu}$ is the space of admissible policies, μ_m^* is the optimal policy in environment m and μ_m is a realization of the agent's policy.

	C unknown	$C \in \mathcal{C}$ convex in $p(a\|x)$	$C \in \mathcal{C}$ non-convex in $p(a\|x)$
Identifiability	Absolute Optimality	Absolute Optimality	Relative Optimality
Conditions	(2.8), (2.9) in Def. 1	(2.8), (2.9) in Def. 1	(2.8), (2.12) in Lemma 2
IRL Example	$--$	Inverse Optimal Stopping with Entropic Running Cost	Inverse SHT (Sec. 2.3)
Reconstruction	Convex reconstruction (2.94)	Convex reconstruction (2.94)	Reconstructed cost for a finite set of strategies/

Table 2.1: IRL Identifiability of Optimal Bayesian stopping.

show how the expected stopping cost can be expressed only in terms of the observed variables in $\mathcal{D}_M$, namely, the action selection policies $\{p_m(a|x)\}_{m=1}^M$ of the agent induced by the stopping strategies $\{\mu_m\}_{m=1}^M$. This is precisely what Theorem 3 below achieves when the inverse learner has access to the agent's action selection policies.

Remarks:

(1) If the analyst does not know *a priori* the structure of the expected continue cost in (2.10), then the IRL identifiability can be generalized from testing for relative optimality (2.8), (2.12) to testing for absolute optimality (2.8), (2.9) in Definition 1. Specifically, we show a certain reconstruction of the expected continue cost (see (2.94) in Sec. 2.10) ensures if relative optimality (2.12) holds, then absolute optimality (2.9) holds.

(2) In contrast to remark (1) above, if the analyst does know a functional form of the expected continue cost, IRL identifiability cannot be improved from testing for relative optimality. One example is IRL for inverse SHT discussed in Sec. 2.3 below where the expected continue cost is known to be the expected stopping time of the agent. On a deeper and more subtle level, knowledge of the structure of the expected continue cost imposes an implicit constraint on the reconstructed cost. Ensuring the reconstructed

expected continue cost (2.94) in Sec. 2.10 satisfies this implicit constraint is non-trivial and beyond the scope of this chapter.

We now present our first main IRL result. The result specifies a set of inequalities that, given the inverse learner's specifications in assumptions (A1) and (A2), are simultaneously *necessary* and *sufficient* for the inverse learner to identify a Bayesian stopping agent's actions to be optimal in the sense of Lemma 2. For readability, we provide the exact expressions for the feasibility inequalities introduced below after the main theorem.

Theorem 3 (IRL for inverse Bayesian optimal stopping [2]) *Consider the inverse learner with dataset $\mathcal{D}_M$ (2.11) obtained from a Bayesian stopping agent's actions over M environments. Assume (A1) and (A2) hold. Then:*

1. *Identifiability: The inverse learner can identify if the dataset $\mathcal{D}_M$ is generated by an optimal Bayesian stopping agent, i.e.,* (2.8) *and* (2.9)*; see Lemma 2.*

2. *Existence: There exists an optimal stopping agent parameterized by tuple Ξ_{opt}* (2.7) *, if and only if there exists a feasible solution to the following convex (in stopping costs) inequalities:*

$$\textit{Find } s_m(x,a) \in \mathbb{R}_+ \ \forall m \in \mathcal{M} \textit{ s.t.}$$

$$\mathrm{NIAS}(\mathcal{D}_M, \{s_m(x,a), x \in \mathcal{X}, a \in \mathcal{A}, m \in \mathcal{M}\}) \leq 0, \tag{2.13}$$

$$\mathrm{NIAC}(\mathcal{D}_M, \{s_m(x,a), x \in \mathcal{X}, a \in \mathcal{A}, m \in \mathcal{M}\}) \leq 0. \tag{2.14}$$

The NIAS (No Improving Action Switches) and NIAC (No Improving Action Cycles) inequalities are defined in (2.16), (2.17) below, and are convex in the stopping cost $s_m(x,a), m \in \mathcal{M}$.

3. *Reconstruction of costs:*

(a) *If the inverse learner knows the agent's expected continue cost C_m for all environments m, the set-valued IRL estimate of the agent's stopping costs is the set of all feasible*

costs $\{s_m(x,a), m \in \mathcal{M}\}$ *that satisfy the NIAS* (2.13)*, NIAC* (2.14) *and SUMCOST inequalities below:*

$$\text{SUMCOST}(\mathcal{D}_M, \{s_m(x,a), C_m, m \in \mathcal{M}\}) \leq 0, \tag{2.15}$$

where SUMCOST *is defined in (2.18), and* C_m *is the expected cost of the Bayesian stopping agent in environment* m*.*

(b) *Suppose the inverse learner knows the agent's stopping costs, and the NIAS* (2.13) *and NIAC* (2.14) *inequalities are feasible. Then, the set-valued IRL estimate of the agent's expected continue cost is given by the set of all feasible costs* C_m *that satisfy the SUMCOST inequality* (2.15)*. Also, if the inverse learner knows the agent's expected continue cost is convex, then the SUMCOST inequality structure permits a convex reconstruction of the cost outlined in Definition 4.* ■

Theorem 3 is proved in Sec. 2.10. It says that identifying if a set $\mathcal{M}$ comprising stopping actions of a Bayesian stopping agent in multiple environments is optimal and then reconstructing the costs incurred in the environments is equivalent to solving a convex feasibility problem. Theorem 3 provides a constructive procedure for the inverse learner to generate set valued estimates of the stopping cost $s_m(x,a)$ and expected cumulative continue cost C_m for all environments $m \in \mathcal{M}$. Algorithms for convex feasibility such as interior points methods [25] can be used to check feasibility of (2.13) and (2.14) (defined in (2.16) and (2.17) below) and construct a feasible solution.

The inequalities NIAS, NIAC and SUMCOST denoted abstractly in Theorem 3 are defined below:

Definition 4 (NIAS, NIAC and SUMCOST inequalities) *Given dataset* $\mathcal{D}_M$*, stopping costs*

$\{s_m(x,a), m \in \mathcal{M}\}$ *and expected continue costs* $\{C_m, m \in \mathcal{M}\}$*:*

$$\text{NIAS} : \sum_{x \in \mathcal{X}} p_m(x|a)(s_m(x,a) - s_m(x,b)) \leq 0, \forall a, m. \tag{2.16}$$

$$\text{NIAC} : \sum_{m \in \widehat{\mathcal{M}}} \mathbb{E}_{a \sim \sum_x \pi_0(x) p_m(\cdot|x)} \left\{ \min_{a' \in \mathcal{A}} \mathbb{E}_{x \sim p_m(\cdot|a)} \{ s_m(x,a) - s_{m+1}(x,a') \} \right\} \leq 0,$$

for any subset of indices $\widehat{\mathcal{M}} \subseteq \mathcal{M}$*, where* $m_k + 1 = m_{k+1}$ *if* $k < l$ *and* $m_l + 1 = m_1$*.*

(2.17)

$$\text{SUMCOST} : \ \mathbb{E}_{x \sim \pi_0, a \sim p_m(\cdot|x)} \{ s_m(x,a) \} + C_m \leq \mathbb{E}_{a \sim p_n(a)} \{ \min_{a' \in \mathcal{A}} \mathbb{E}_{x \sim p_n(\cdot|a)} \{ s_m(x,a') \} \} + C_n,$$

$$\forall\, m, n \in \mathcal{M}. \tag{2.18}$$

Reconstruction of expected cumulative continue cost. *If NIAS, NIAC and SUMCOST inequalities defined above have a feasible solution, the following convex reconstruction of the agent's expected continue cost is consistent with optimal Bayesian stopping* (2.8)*,* (2.9)*, a stronger condition compared to relative optimality* (2.8)*,* (2.12)*:*

$$\widehat{C}(\mu) = \max_{m=1,2,\ldots,M} \ \left\{ C_m + G_{m,m} - \tilde{G}(\mu, s_m) \right\}, \ \textit{where} \tag{2.19}$$

$$\tilde{G}(\mu, s_m) = \sum_{a \in \mathcal{A}} \left(\sum_{x \in \mathcal{X}} p_\mu(a|x) \pi_0(x) \right) \min_{b \in \mathcal{A}} \sum_{x \in \mathcal{X}} p_\mu(x|a) s_m(x,b), \ \textit{and} \tag{2.20}$$

$$G_{m,m} = \sum_{x \in \mathcal{X}, a \in \mathcal{A}} \pi_0(x) p_m(a|x) s_m(x,a) \tag{2.21}$$

The above reconstruction assumes the agent's mapping from the sequence of observations $y_{1:\tau(\mu)}$ *to the space of actions is one-to-one, and is valid if and only if the agent's expected cumulative continue cost is convex.*

Let us now provide an intuitive explanation for the abstract inequalities of Theorem 3. NIAS (2.13): NIAS applies to each of the M environments in $\mathcal{M}$. NIAS checks if, for every environment, the agent chooses the optimal stop action given its stopping belief and stopping strategy.

NIAC (2.14): NIAC checks for optimality of the agent's stopping strategies in M environments. Since the stopping agent chooses its strategies in a finite number (M) environments, NIAC checks if the agent's strategy in the m^{th} environment performs at least as well as the strategies of the agent in all other environments given the environment's stopping cost $s_m(x, a)$, for all $m \in \mathcal{M}$. If so, it constructs a feasible set of stopping costs in the M environments so that the chosen strategies are consistent with an optimal stopping agent.

SUMCOST (2.15): If the Bayesian agent is an optimal stopping agent (NIAS and NIAC have a feasible solution), SUMCOST constructs a set of feasible expected continue costs incurred by the Bayesian agent in the multiple environments. The feasibility of NIAS and NIAC ensures that the SUMCOST inequalities have a feasible solution. In (2.18), the RHS term is the expected cumulative cost of the agent in environment n given the stopping costs in environment m. The feasibility inequality (2.18) checks for feasible expected cumulative continue costs so that the agent's stopping strategies in $\mathcal{M}$ are identified as optimal by the inverse learner, i.e. , (2.12) is satisfied. The reconstructed cost $\widehat{C}$ (2.19) is a convex interpolation of expected stopping costs and feasible scalars C_m (2.18) such that conditions (2.8) and (2.9) for optimal Bayesian stopping hold; see Sec. 2.10 for a detailed discussion. We remark that the reconstruction in (2.19) is only valid when (a) the inverse learner has no information about the agent's observation likelihood, and (b) the inverse learner does not know the agent's expected continue cost. In Table 2.1, we highlight the subtle issues underpinning IRL identifiability for optimal Bayesian stopping in more detail. In Sec. 2.3 below, we discuss IRL for optimal Bayesian stopping when the inverse learner knows the agent's expected continue cost; hence, the reconstruction (2.19) is no more required for achieving IRL.

2.2.4 Discussion of Theorem 3

We now discuss the implications of Theorem 3 and contextualize the NIAS and NIAC feasibility inequalities (2.13), (2.14) of Theorem 3.

(i) Necessity and Sufficiency.

The NIAS and NIAC conditions (2.13), (2.14) are necessary and sufficient for the inverse learner to identify an optimal stopping agent. This makes Theorem 3 a remarkable result. If no feasible solution exists, then the dataset $\mathcal{D}_M$ cannot be rationalized by an optimal Bayesian stopping agent. Also, if there exists a feasible solution, then the dataset $\mathcal{D}_M$ must be generated by an optimal stopping agent in multiple environments (Lemma 2).

(ii) Set valued estimate vs point estimate.

An important consequence of Theorem 3 is that the reconstructed utilities are set-valued estimates rather than point valued estimates even though the dataset $\mathcal{D}_M$ has $K \to \infty$ samples. Estimating the costs from the solution of a cost minimization problem is an ill-posed problem. Put differently, all points in the feasible set of rationalizing costs explain the dataset $\mathcal{D}_M$ equally well.

(iii) Consistency of Set-Valued Estimate.

The NIAS and NIAC inequalities are both necessary and sufficient for optimal Bayesian stopping. The necessity implies that the true stopping costs and expected continue costs incurred by the agent are feasible wrt the convex NIAS and NIAC inequalities. Hence, the IRL procedure is consistent in that the set-valued estimator contains the true generating model.

(iv) Context: NIAS and NIAC.

The inequalities (2.8), (2.12) for the inverse learner to identify an optimal stopping agent can be written in abstract notation as (2.22), (2.23), respectively, in terms of the variables

$\{s_m, C_m\}_{m=1}^{\mathcal{M}}$:

$$\text{NIAS}(\{\{p(y_{1:\tau(\mu_m)}|x), x \in \mathcal{X}\}, s_m, m \in \mathcal{M}\}, \pi_0) \leq 0, \tag{2.22}$$

$$\text{NIAC}^*(\{\{p(y_{1:\tau(\mu_m)}|x), x \in \mathcal{X}\}, s_m, C_m, m \in \mathcal{M}\}, \pi_0) \leq 0. \tag{2.23}$$

The inverse learner in our setup does not know the agent's observation sequences $\{y_{1:\tau}, m \in \mathcal{M}\}$, observation likelihood B or the continue cost $\mathcal{C}$. Hence, as shown in Sec. 2.10, the best the inverse learner can do is check for the feasibility of the NIAC (2.17) that does not depend on C_m. Otherwise, the IRL task is equivalent to using optimality equations (2.8), (2.12) expressed abstractly as NIAS and NIAC* above to reconstruct the costs. Eq. 2.13 and 2.14 in Theorem 3 specialize to (2.22) and (2.23) by replacing the action selection policy $p_m(a|x)$ with the unknown likelihood $p(y_{1:\tau(\mu_m)}|x)$. Put differently, (2.13) and (2.14) defined in (2.16), (2.17) can be viewed as surrogates of the feasibility conditions (2.22) and (2.23), respectively. However, as discussed in the proof in Sec. 2.10, the action selection policy $p_m(a|x)$ suffices for both necessity and sufficiency of Bayes optimality (2.22), (2.23) in spite of being a Blackwell noisy measurement of $p(y_{1:\tau(\mu_m)}|x)$. Also, observe the NIAC inequality (2.17) is independent of C_m and expressed only in terms of stopping costs s_m. However, as shown in Sec. 2.10, the feasibility of both inequalities (2.23) and (2.17) are equivalent. Finally, in some examples of stopping time problems such as SHT discussed in Sec. 2.3, the inverse learner knows the agent's expected cumulative continue cost and hence, can use the NIAC* inequality as is to identify optimality and achieve IRL.

NIAS and NIAC with ε-feasibility. One trivial solution that satisfies both NIAS and NIAC inequalities in Theorem 3 is the degenerate cost of all zeros. Such degeneracy is common in IRL literature due to the fundamental ill-posedness of the inverse optimization problem. In practice, one can ensure only non-trivial solutions pass the NIAS and NIAC feasibility

inequalities by introducing a margin constraint:

$$\text{NIAS}(\cdot) \leq -\varepsilon, \;\; \text{NIAC}(\cdot) \leq -\varepsilon, \; \varepsilon > 0. \tag{2.24}$$

Margin constraints for ensuring non-degenerate solutions to feasibility tests are common practices in IRL [26]. In complete analogy, using the ϵ restriction of (2.24), we can ensure only non-trivial informative costs pass the NIAS and NIAC feasibility test of Theorem 3.

(v) Private and Public Beliefs.

The stopping belief π_τ in (2.9) can be interpreted as the *private belief* evaluated by the agent after measuring $y_{1:\tau}$ in the sense of Bayesian social learning [19, 27]. Since π_τ is unavailable to the inverse learner, it uses the *public belief* $p(x|a)$ as a result of the agent's stop action to estimate its incurred costs.

(vi) IRL for stopping agent whose observation likelihood changes with the environment. For notational convenience, we assume the Bayesian agent's observation likelihood is fixed across different environments. However, in Sec. 2.10, we discuss under what conditions the inverse learner can achieve IRL when the Bayesian agent's observation likelihoods change with the environment. We provide a specific example of the agent continue cost, namely, the entropic continue cost that facilitates the inverse learner to achieve IRL for different agent observation likelihoods in different environments. The agent's stopping cost in this case is a logistic function in terms of its action selection policy; the logistic function also arises in Max-Entropy IRL [28]. This resemblance is not surprising; the agent in [28] maximizes its cumulative expected reward subject to a bound on the mutual information between the prior and the distribution of beliefs induced by its policy. The objective function in (2.9) where C is the mutual information between the prior and the stopping belief is simply the Lagrangian form of the objective the agent aims to optimize in [28]. The IRL problem for agents that Maximize their expected

terminal rewards with a mutual information penalty has also been studied in the Bayesian revealed preference literature by [29].

(vii) IRL for boundedly-rational forward learner.

For general POMDPs, it is difficult[7] for a Bayesian sequential decision maker to compute the optimal policy μ^* in (2.8), (2.9). We say that a strategy $\hat{\mu}$ is ϵ-optimal if the following condition holds:

$$\epsilon\text{-optimal Bayesian stopping: } J(\hat{\mu}) - J(\mu^*) \leq \epsilon, \text{ for some } \epsilon \geq 0. \tag{2.25}$$

Eq. 2.25 arises when the forward learner uses sub-optimal procedures for solving the POMDP such as approximate value iteration, open loop feedback and finite state controllers. When both the stopping cost and the expected continue cost are free variables like in Theorem 3, detecting ϵ-optimality is non-identifiable and a difficult task. However, if either the stopping cost or the expected continue cost, (such as in the case of SHT discussed in Sec. 2.3) is known to the inverse learner, one can identify ϵ-optimality based on the feasibility of the IRL inequalities. We briefly discuss identification of ϵ-optimality after Theorem 6; a general framework is beyond the scope of this chapter and the subject of future work. Indeed, more precise knowledge of the agent's sub-optimality allows the inverse learner to achieve IRL; see [30] for a discussion on how to achieve IRL when the inverse learner has access to a ranked set of forward learner's decision trajectories, ranked according to the extent of sub-optimality in each trajectory.

(viii) **No knowledge of observation likelihood by the inverse learner.** This chapter assumes the inverse learner has no knowledge of the agent's observation likelihood. The sufficiency proof of Theorem 3 exploits this zero-knowledge assumption and posits that the inverse learner can thus assume a one-to-one mapping from the space of observation

[7] [17] show that solving partially observed Markov decision processes are in general PSPACE hard. The SHT and Search problems discussed in this chapter are special cases where the optimal stopping strategy is stationary due to the problem structure and characterized as a threshold policy in the belief space.

sequences $y_{1:\tau(\mu)}$ to the space of stopping actions. Indeed, one can show that if the instantaneous continue cost has an entropic form, for example, the Shannon-Gibbs entropy, Rényi entropy or Tsallis entropy, the optimal mapping from observation sequences to stopping actions is one-to-one due to the strongly concave nature of these costs; see [29] for a discussion of IRL for entropic costs.

(ix) **Partial knowledge of agent costs.** If the Bayesian agent's instantaneous continue cost is zero, then it is optimal to never stop sensing, i.e., the agent observe infinitely many samples and the posterior belief approaches the Dirac delta function centered at the state x[8]. Hence, the optimal $p_m(a|x)$ has non-zero weights if and only if $a \in \operatorname{argmin}_{a' \in \mathcal{A}} \; s_m(x, a')$. Then checking for optimal Bayesian stopping with zero running cost is equivalent to identifying feasible stopping costs that satisfy the following condition:

$$p_m(a|x) \neq 0 \iff a \in \operatorname*{argmin}_{a' \in \mathcal{A}} \; s_m(x, a'). \tag{2.26}$$

Sec. 2.3 considers the case where the instantaneous continue cost is a constant, hence the cumulative expected continue cost is proportional to the expected stopping time of the agent. If the inverse learner knows the expected continue cost, IRL is achieved by checking for the existence of feasible stopping costs that satisfy the NIAS (2.16) and SUMCOST (2.18) inequalities with C_m set to the agent's expected continue cost in environment m.

(x) **IRL with ε-feasibility.** If neither the stopping costs nor the expected continue costs are known to the inverse learner, the NIAS, NIAC and SUMCOST inequalities are trivially feasible by choosing the degenerate solution of constant costs. In this case it makes sense to construct the inverse learner's non-trivial IRL cost estimate as the set

[8] It follows from Bernstein-von Mises theorem [31] that, under mild smoothness conditions, the agent's posterior belief converges asymptotically to a normal distribution centered around the maximum likelihood estimate with covariance $\lim_{t\to\infty}(t\, I(x))^{-1}$, where I denotes the Fisher information matrix.

of feasible costs $\{s_m(x,a), C_m, m \in \mathcal{M}\}$ that are ϵ-feasible wrt the NIAC, NIAC and SUMCOST inequalities:

- Choose feasibility margins $\epsilon_{NIAS},\ \epsilon_{NIAC},\ \epsilon_{SUMCOST} \geq 0,$ not all zero.
- Construct the set-valued IRL estimate as the set of all tuples $\{s_m(x,a), C_m, m \in \mathcal{M}\}$ that satisfy

$$\text{NIAS}(\cdot) \leq \epsilon_{NIAS},\ \text{NIAC}(\cdot) \leq \epsilon_{NIAC} \text{ and } \text{SUMCOST}(\cdot) \leq \epsilon_{SUMCOST}. \tag{2.27}$$

2.2.5 Discussion of (A1) and (A2)

(A1): To motivate (A1), suppose for each environment $m \in \mathcal{M}$, the inverse learner records the Bayesian stopping agent's true state $x^o_{k,m}$, stopping action $a_{k,m}$ and stopping time $\tau_k(\mu_m)$ over $k = 1, 2, \ldots, K$ independent trials. Then the pmf $p_m(a|x)$ in (2.11) is the limit pmf of the empirical pmf $\hat{p}_m(a|x)$ as the number of trials $K \to \infty$ defined as:

$$\hat{p}_m(a|x) = \frac{\sum_{k=1}^K \mathbb{1}\{x^o_{k,m} = x, a_{k,m} = a\}}{\sum_{k=1}^K \mathbb{1}\{x^o_{k,m} = x\}}. \tag{2.28}$$

Specifically, since for each $m \in \mathcal{M}$ the sequence $\{x^o_{k,m}, a_{k,m}\}$ is i.i.d for $k = 1, 2, \ldots K$, by Kolmogorov's strong law of large numbers, as the number of trials $K \to \infty$, $\hat{p}_m(a|x)$ converges with probability 1 to the pmf $p_m(a|x)$. In the remainder of the chapter (apart from Sec. 2.6), we will work with the asymptotic dataset $\mathcal{D}_M$ for IRL. In Sec. 2.6 we analyze the effect of finite sample size K on the inverse learner using concentration inequalities.

(A2): (A2) is necessary for the identification of an optimal stopping agent (Lemma 2) to be well-posed. Suppose (A2) does not hold. Then, for $M = 2$ and true stopping costs $s_1 = s_2$, we have $p_1(a|x) = p_2(a|x)$ in $\mathcal{D}_M$. This implies the set of feasible solutions

(C_1, C_2) for the feasibility inequality (2.18) is the set $\{(C_1, C_2) : C_1 = C_2, C_1, C_2 \in \mathbb{R}_+\}$ and is hence, unidentifiable. [9]

2.2.6 Outline of proof of Theorem 3

The proof of Theorem 3 in Sec. 2.10 involves two main ideas. The first key idea is to specify a fictitious likelihood $\mathbb{P}_\mu(\tilde{y}_\pi|x)$ parametrized by the stopping strategy so that given strategy μ, observation likelihood B and prior π_0, the observation trajectory $y_{1:\tau}$ of the stopping time problem yields an identical stopping belief π_μ, i.e.,

$$\mathbb{P}(\tilde{y}_\pi|x, \mu) = \mathbb{P}\left(\{y_{1:\tau}\} : \pi_\tau = \pi|x\right).$$

A more precise statement is given in (2.75). In other words, a one-step Bayesian update using the likelihood $\mathbb{P}(\tilde{y}_\pi|x, \mu)$ is equivalent to the multi-step Bayesian update (2.3) of the state till the stopping time. This idea is shown in Fig. 2.2. Recall that the cumulative expected cost of the agent comprises two components, the stopping cost and cumulative continue cost. A useful property of this fictitious likelihood is that it is a sufficient statistic for the expected stopping cost $G(\cdot)$.

The second main idea is to formulate the agent's expected cumulative cost using the observed action selection policy $p(a|x)$ of the agent instead of the unobserved fictitious likelihood $p(y_{1:\tau(\mu_m)}|x)$ that determines the expected stopping cost. $p_m(a|x)$ (2.11) is a stochastically garbled (noisy) version of $p(y_{1:\tau(\mu_m)}|x)$. We use this concept to formulate the NIAS and NIAC inequalities whose feasibility given $\mathcal{D}_M$ is necessary and sufficient for identifying an optimal stopping by a Bayesian stopping agent in multiple environments.

[9] The condition $M = 1$ (or equivalently, $M = 2$ with equal stopping costs) is analogous to probing an agent with the same probe vector in classical revealed preferences [3, 32]. The obtained dataset of probes and responses can be rationalized by any concave, locally non-satiated, monotone utility function thus leading to loss of identifiability of the agent's utilities.

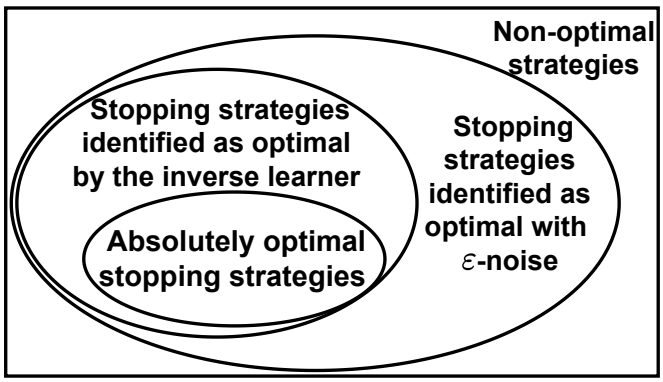

Figure 2.1: Given tuple $(\Xi, \mathcal{C}, \{s_m(x,a), m \in \mathcal{M}\})$, the set of stopping strategies (2.12) of a stopping agent identified as an optimal stopping agent by the inverse learner (Lemma 2) contains the stopping strategies of an absolutely optimal agent defined by (2.8), (2.9). Such strategies can be obtained by small perturbations of the absolute optimal strategies such that the Bayesian stopping agent's strategy in each environment still performs better than that chosen by the agent in any other environment in $\mathcal{M}$. Like Sec. 2.2, Sec. 2.3 and 2.4 deal with identifying such optimal strategies for the SHT and search problems. In Sec. 2.6, we will detect if the agent's strategies corrupted by noise (due to finite sample constraints) belong to the set of strategies identified as optimal strategies by the inverse learner.

Showing that feasibility of the NIAS and NIAC inequalities (2.13), (2.14) is a necessary condition for the stopping strategies chosen by the Bayesian stopping agent to be optimal, (2.8), (2.12) is straightforward. The key idea in the sufficiency proof is to note that the elements of the garbling matrix that maps the fictitious observation likelihood to the action selection policy is unknown to the inverse learner. Hence, the inverse learner can arbitrarily assume $p_m(a|x)$ to be an accurate measurement of $p(y_{1:\tau(\mu_m)}|x)$. We then show that for a feasible set of viable stopping costs $\{s_m(x,a), C_m, m \in \mathcal{M}\}$ that satisfy the NIAS and NIAC inequalities, there exist a set of positive reals $\{C_m, m \in \mathcal{M}\}$ that satisfy (2.8), (2.9) with the expected cumulative continue cost incurred by the agent in the m^{th} environment set to C_m.

The NIAS and NIAC inequalities are convex in the stopping costs $s_m, m \in \mathcal{M}$. The inverse learner can solve for these convex feasibility constraints to obtain a feasible solution. Thus, we have a constructive IRL procedure for reconstructing the stopping

and expected cumulative continue costs for the inverse optimal stopping time problem.

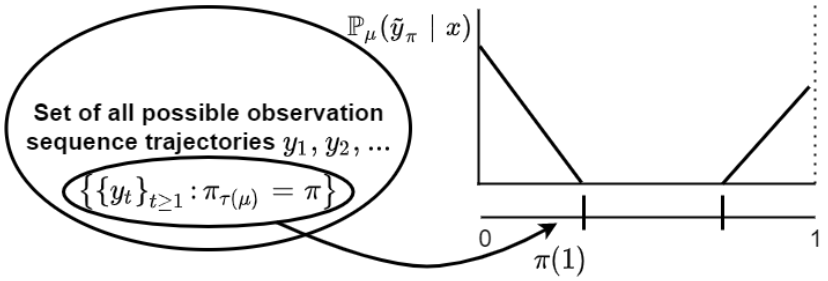

Figure 2.2: Schematic illustration of first main idea of proof of Theorem 3 for the case when $X = 2$. The key idea is to construct a fictitious observation likelihood $\mathbb{P}_\mu(\tilde{y}_\pi|x)$ for compact representation of the agent's expected stopping cost. The probability of generating the fictitious observation $\tilde{y}_\pi$ is equal to the probability of a sequence of observations yielding a stopping belief π for a given stopping time μ.

2.2.7 Summary

This section has laid the groundwork for IRL of a Bayesian stopping time agent. Specifically, we discussed the dynamics of the Bayesian stopping time agent in a single environment (2.2) and multiple environments (2.7). We then described the IRL problem that the inverse learner aims to solve. Theorem 3 gave a necessary and sufficient condition for a Bayesian stopping time agent to be identified as an optimal stopping agent when its decisions in multiple environments are observed by the inverse learner. The agent's stopping cost in each environment can be estimated by solving a convex feasibility problem. Theorem 3 forms the basis of the IRL framework in this chapter. Next, we develop IRL results for 2 examples of stopping time problems, namely, sequential hypothesis testing and Bayesian search.

2.3 Example 1. IRL for sequential hypothesis testing (SHT)

We now discuss our first example of IRL for an optimal Bayesian stopping time problem, namely, *inverse* Sequential Hypothesis Testing (SHT). Our main result below (Theorem 6) specifies a necessary and sufficient condition for IRL in SHT. The SHT problem is a special case of the optimal Bayesian stopping problem discussed in Sec. 2.2.2 since the continue cost c_t (2.2) is a constant for all time t in the SHT problem. For our IRL task, the continue cost can be chosen as 1 WLOG.

2.3.1 Sequential hypothesis testing (SHT) Problem

Let $y_1, y_2, \ldots$ be a sequence of i.i.d observations. Suppose the Bayesian agent knows that the pdf of y_i is either $p(y|x=1)$ or $p(y|x=2)$. The aim of classical SHT is to decide sequentially on whether $x=1$ or $x=2$ by minimizing a combination of the continue (measurement) cost and misclassification cost. In analogy to Sec. 2.2.2, we now define a set of SHT environments in which a Bayesian stopping agent operates.

Definition 5 (Optimal SHT in multiple environments) *The set $\mathcal{M}$ of optimal SHT in multiple environments is a special case of optimal stopping in multiple environments Ξ_{opt} (2.7) with:*

- $\mathcal{X} = \{1, 2\}$, $\mathcal{Y} \subset \mathbb{R}$, $\mathcal{A} = \mathcal{X}$.
- $\mathcal{C} = \{c_t\}_{t\geq 0},\ c_t(x) = c \in \mathbb{R}^+,\ \forall x \in \mathcal{X}$ *is the constant continue cost.*
- $\{\mu_m, m \in \mathcal{M}\}$ *are the SHT stopping strategies chosen by the Bayesian agent over M SHT environments defined below.*
- $s_m(x, a)$ *is the stopping cost incurred by the agent in the m^{th} SHT environment*

parametrized by misclassification costs $(\bar{L}_{m,1}, \bar{L}_{m,2})$.

$$s_m(x,a) = \begin{cases} \bar{L}_{m,1}, & \textit{if } x = 1, a = 2, \\ \bar{L}_{m,2}, & \textit{if } x = 2, a = 1, \\ 0, & \textit{if } x = a \in \{1,2\}. \end{cases}$$

The SHT stopping strategies in the above definition satisfy the optimality conditions in Definition 1 and can be computed using stochastic dynamic programming [19]. The solution for μ_m for the m^{th} SHT environment is well-known [20] to be a stationary policy with the following threshold rule parameterized by scalars $\alpha_m, \beta_m \in (0,1)$:

$$\mu_m(\pi) = \begin{cases} \text{choose action } 2, & \text{if } 0 \leq \pi(x=2) \leq \beta_m \\ \text{continue}, & \text{if } \beta_m < \pi(x=2) \leq \alpha_m \\ \text{choose action } 1, & \text{if } \alpha_m < \pi(x=2) \leq 1. \end{cases} \tag{2.29}$$

Remark: Since the SHT dynamics can be parameterized by $c, \bar{L}_1, \bar{L}_2$, we can set $c = 1$ without loss of generality since the optimal policy is unaffected. Also, the expected cumulative continue cost of the agent is simply the expected stopping time of the agent.

2.3.2 IRL for inverse SHT. Main assumptions

Suppose the inverse learner observes the actions of a Bayesian stopping agent in M SHT environments. In addition to assumptions (A2), we assume the following about the inverse learner performing IRL for identifying an SHT agent:

(A3) The inverse learner has the dataset

$$\mathcal{D}_M(SHT) = (\mathcal{D}_M, \{C_m, m \in \mathcal{M}\}), \tag{2.30}$$

where $\mathcal{D}_M$ is defined in (2.11), $C_m = \mathbb{E}_{\mu_m}\{\tau\}$ is the expected continue cost incurred by the Bayesian agent in the m^{th} environment.

(A4) The stopping strategies $\{\mu_m, m \in \mathcal{M}\}$ are stationary strategies characterized by the threshold structure in (2.29).

(A5) There exist reals $\delta_1,\ \delta_2 \in (0,1)$ such that the following conditions are satisfied:

$$(i)\ \beta_m \leq \delta_1 \leq \delta_2 \leq \alpha_m,\ \forall m \in \mathcal{M},\ (ii)\quad \delta_1/(1-\delta_1) \leq \bar{L}_{m,1}/\bar{L}_{m,2} \leq \delta_2/(1-\delta_2),$$

where α_m, β_m are the threshold values of the stationary strategy μ_m chosen by the Bayesian agent in environment m.

Remarks: (i) Assumption (A3) specifies additional information the inverse learner has for performing IRL for SHT by recording the agent decisions over $K \to \infty$ independent trials. Since the continue cost is 1, the expected cumulative continue cost is simply the expected stopping time of the agent. The inverse learner obtains an a.s. consistent estimate of the expected stopping time by computing the sample average of the K stopping times. Since the expected continue cost is simply the expected stopping time of the agent and known to the inverse learner, it is no more a feasible variable in the feasibility equations (2.17). This yields a smaller feasibility set for the stopping costs.

(ii) Assumption (A4) comprises partial information the inverse learner has about the stopping strategies chosen by the agent and its observation likelihood. Since the optimal stopping strategy is well-known to have a threshold structure [20], the inverse learner only needs to compare the expected cost incurred from threshold policies to check for optimality and achieving IRL.

(iii) Assumption (A5) ensures the expected stopping cost of the SHT agent $G(\mu_m, s)$ (2.9) that depends on the unobserved strategy μ_m can be expressed in terms of the induced action selection policy $p_m(a|x)$ for any stopping cost s, i.e., $G(\mu_m, s_n) = \mathbb{E}_{p_m(a)}\{\mathbb{E}_{x\sim p_m(\cdot|a)}\{s(x,a)\}\}$.

2.3.3 IRL for inverse SHT. Main result

Our main result below specifies a set of linear inequalities that are necessary and sufficient for the Bayesian agent's actions observed by the inverse learner to be identified as that of an optimal SHT agent (Lemma 2). Any feasible solution constitutes a viable SHT misclassification cost for the M SHT environments in which the Bayesian agent operates.

Theorem 6 (IRL for inverse SHT) *Consider the inverse learner with dataset $\mathcal{D}_M(\mathrm{SHT})$ (2.30) obtained from a Bayesian agent taking actions in M SHT environments. Assume (A2) holds. Then:*

1. *Identifiability: The inverse learner can identify if the dataset $\mathcal{D}_M(\mathrm{SHT})$ is generated by an optimal SHT agent (Lemma 2).*

2. *Existence: There exists an optimal SHT agent parameterized by tuple Ξ_{opt} (2.7), if and only if there exists a feasible solution to the following convex (in stopping costs) inequalities:*

$$\begin{aligned}
&\textit{Find } s_m(x,a) > 0, s_m(x,x) = 0,\ \forall x,a \in \mathcal{X},\ m \in \mathcal{M} \textit{ s.t.}\\
&\mathrm{NIAS}: \sum_{x\in\mathcal{X}} p_m(x|a)(s_m(x,a) - s_m(x,b)) \le 0, \forall a,b,m.\\
&\mathrm{NIAC}^*: \left(\sum_{x,a} \pi_0(x) p_m(a|x) s_m(x,a) + C_m\right) - \left(\sum_a p_n(a) \min_b \sum_x p_n(x|a) s_m(x,b) + C_n\right) \le 0,\\
&\qquad \forall m,n \in \mathcal{M}, m \ne n. \qquad (2.31)
\end{aligned}$$

(Recall that $C_m = \mathbb{E}_{\mu_m}\{\tau\}$ is known to the inverse learner, and hence is not a free variable).

3. *Reconstruction: The set-valued IRL estimates of the SHT misclassification costs $\{\bar{L}_m, m \in \mathcal{M}\}$ are defined below where $\bar{L}_m = (\bar{L}_{1,m}, \bar{L}_{2,m})$:*

$$\bar{L}_{1,m} = s_m(1,2),\ \bar{L}_{2,m} = s_m(2,1)\ \forall m \in \mathcal{M},$$

where $\{s_m(x,a), m \in \mathcal{M}\}$ is any feasible solution to the NIAS and NIAC* inequalities.■

Theorem 6 is a special instance of Theorem 3 for identifying an optimal stopping agent operating in multiple environments. The NIAC* resembles SUMCOST (2.18) with the only difference that C_m is the expected stopping time of the agent in environment m instead of being a feasible variable like in (2.18). We note that since the expected stopping time is non-convex in the agent's action selection policy $p_m(a|x)$, the inverse learner cannot use the convex reconstruction procedure of (2.19) to estimate the expected stopping time for any other policy.

Remarks:

1. Inverse SHT is an IRL task with partially specified costs: out of the continue and stopping costs, the continue cost incurred by the Bayesian agent is already known to the inverse learner. As a consequence, the feasibility test for identifying an optimal SHT agent imposes tighter restrictions (fewer feasible variables) compared to identifying optimal stopping in Theorem 3 and avoids degenerate feasible solutions that trivially satisfy the inequalities (2.31) of Theorem 6.

2. *IRL for Multi-state SHT.* Theorem 6 is independent of the number of states X. When $X > 2$, IRL for inverse SHT comprises estimating the misclassification costs $\{\bar{L}_{m,x,a},\ x \neq a, x, a \in X\}$, and is achieved by solving the feasibility inequalities (2.16) and (2.31) of Theorem 6.[10]

3. *Inverse SHT for boundedly-rational forward learner.* In Sec. 2.2.4, we discussed the concept of ϵ-optimality for a forward learner. Below, we briefly discuss how the NIAS and NIAC* feasibility inequalities of Theorem 6 can identify if an agent performs ϵ-optimal SHT when the inverse learner knows the agent's expected continue cost.

If NIAS and NIAC* (2.31) are feasible, then one cannot say if the dataset $\mathcal{D}_M$ (2.30) is generated from an absolutely optimal Bayesian agent (Definition 1) or an ϵ-optimal Bayesian agent (2.25). However, if $\mathcal{D}_M$ fails the feasibility test (2.31) of Theorem 6,

[10]Since the state and environment index suffice to denote the misclassification cost when $X = 2$, the subscript 'a' is dropped from the misclassification cost notation in Lemma 2 for notational clarity.

then it is clear $\mathcal{D}_M$ results from an ϵ-optimal Bayesian agent, where a bound on ϵ can be obtained by finding the minimum relaxation needed for passing the feasibility test (2.31):

$$\min_{\epsilon_{\text{relax}} \geq 0} \epsilon_{\text{relax}}, \text{ such that } \text{NIAS}(\mathcal{D}_M, \{s_m(x,a)\}) \leq \epsilon_{\text{relax}}, \text{ NIAS}(\mathcal{D}_M, \{s_m(x,a)\}) \leq \epsilon_{\text{relax}}. \tag{2.32}$$

The ϵ-relaxation in (2.32) arises frequently in microeconomic theory in robustness tests to measure how far an economic agent is from satisfying economics-based rationality. Some examples of widely used robustness measures in economics literature include the Houtman index (HM-Index) [33], Afriat measure [34] and Varian measure [35].

2.3.4 Numerical example illustrating IRL for inverse SHT

We now present a toy numerical example for inverse SHT with 3 SHT environments and 3 states. The aim of this example is to illustrate the consistency property of Theorem 6. That is, that the true misclassification costs lie in the set of feasible costs computed by the inverse learner by solving the convex feasibility test of Theorem 6.

SHT environments. We consider $M = 3$ SHT environments with:

- *Prior* $\pi_0 = [0.5\ 0.5]'$.
- *Observation likelihood*: $p(y|x=1) = \mathcal{N}(1,2)$, $p(y|x=2) = \mathcal{N}(-1,2)$, where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2.
- *Misclassification costs*:
  Environment 1: $(\bar{L}_{1,1}, \bar{L}_{1,2}) = (2, 2.5)$, Environment 2: $(\bar{L}_{2,1}, \bar{L}_{2,2}) = (4, 3)$, Environment 3: $(\bar{L}_{3,1}, \bar{L}_{3,2}) = (6, 6)$.

Inverse Learner specification. Next we consider the inverse learner. We generate $K = 10^5$ samples for the 3 SHT environments using the above parameters. Recall from

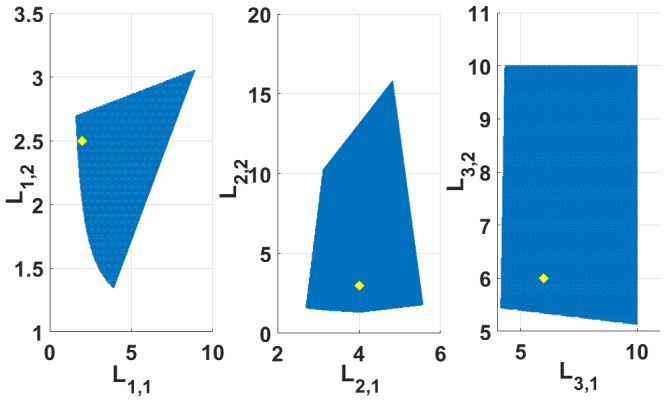

Figure 2.3: Inverse SHT numerical example with parameters specified in Sec. 2.3.4. The key observation is that the true misclassification costs (yellow points) lie in the feasible set (blue region) of costs computed via Theorem 6. This follows from the necessity proof of Theorem 6 which says if the Bayesian agent is an optimal SHT agent, then the true costs lie in the feasible set of costs that satisfy the NIAS and NIAC* inequalities. Highlighting the advantage of the set-valued estimate of our IRL algorithm, we note that all points in the blue feasible region rationalize the observed stop actions of the Bayesian agent equally well. Indeed, the feasible region shrinks with the number of environments M.

Theorem 6 that the inverse learner uses the dataset $\mathcal{D}_M(\mathrm{SHT})$ to perform IRL for inverse SHT, where $\mathcal{D}_M(\mathrm{SHT})$ is defined as:

$$\mathcal{D}_M(\mathrm{SHT}) = (\pi_0, (\hat{p}_m(a|x), \sum_{k=1}^{K} \tau_k(\mu_m)/K), m \in \{1,2,3\}), \tag{2.33}$$

where $K = 10^5$, the second and third terms are the empirically calculated action selection policy and expected stopping time for SHT environments m from the 10^5 generated samples. We denote the action selection policy in (2.33) as $\hat{p}_m(a|x)$ and not $p_m(a|x)$ since the numerical example uses an empirical estimate.

IRL Result. The inverse learner performs IRL by using the dataset $\mathcal{D}_M(\mathrm{SHT})$ (2.33) to solve the linear feasibility problem in Theorem 6. The result of the feasibility test is shown in Fig. 2.3. The blue region is the set of feasible misclassification costs for each SHT environment. The feasible set of costs is $\{(\bar{L}_{m,1}, \bar{L}_{m,2}), m \in \{1,2,3\}\} \subseteq \mathbb{R}^6_+$. Fig. 2.3 displays the feasible misclassification costs for a single environment keeping

the costs for the other two environments fixed at their true values. The need to fix costs for the other two environments for plotting the set of feasible costs is only for visualization purposes. It is not possible to plot a 6 dimensional point (vector of estimated misclassification costs for 3 SHT environments) on the 2-d plane.

The true misclassification costs for each SHT environment are highlighted by a yellow point. The key observation is that these true costs belong to the set of feasible costs (blue region) computed via Theorem 6. Thus, Theorem 6 successfully performs IRL for the SHT problem and the set of feasible misclassification costs can be reconstructed as the solution to a linear feasibility problem. Also, all points in the set of misclassification costs explain the SHT dataset equally well.

2.3.5 Numerical example. Regularized max-margin IRL for inverse SHT.

We now present a numerical example for inverse SHT involving $M = 100$ environments where we compute a point-valued IRL estimate of the SHT misclassification costs. This inference task is in contrast to the set-valued IRL flavor considered thus far in the chapter. Given dataset $\mathcal{D}_M(\text{SHT})$ (2.33), we compute a point estimate $\bar{L}^*$ of misclassification costs that maximizes the $\mathcal{L}_2$-regularized margin of the NIAC* feasibility inequalities of Theorem 6. The point estimate $\bar{L}^*$ is inspired by max-margin IRL methods in the literature [7, 26] and defined as:

$$\bar{L}^* = \underset{\bar{L}}{\text{argmin}} \sum_{m,n=1,m\neq n}^{M} \text{Margin}_{\mathcal{D}_M(\text{SHT})}(m, n, \bar{L}) - \lambda \|\bar{L}\|_2^2, \tag{2.34}$$

$$\text{Margin}_{\mathcal{D}_M(\text{SHT})}(m, n, \bar{L}) = \left(G(\hat{p}_n, \bar{L}_m) + \hat{C}_n\right) - \left(G(\hat{p}_m, \bar{L}_m) + \hat{C}_m\right), \tag{2.35}$$

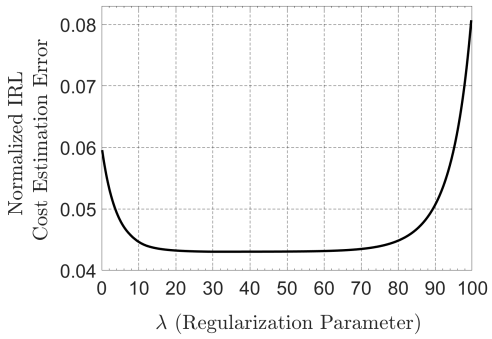

Figure 2.4: Inverse SHT numerical example for 100 SHT environments, with parameters specified in Sec. 2.3.5. The main takeaway is that regularized max-margin IRL for inverse SHT (2.34) can estimate the misclassification costs incurred by the stopping agent in the 100 SHT environments with up to 95% accuracy by varying the regularization parameter λ in (2.34).

where $G(\hat{p}, \bar{L}_m)$ is the expected misclassification cost for SHT with action selection policy $\hat{p}$ and misclassification costs $\bar{L}_m$, and $\hat{C}_m = \sum_{k=1}^{K} \tau_k(\mu_m)/K)$ is the agent's expected continue cost in environment m computed empirically from K independent trials. In simple terms, (2.35) is the difference in expected cumulative cost between action policies $\hat{p}_m$ and $\hat{p}_n$ for a fixed misclassification cost $\bar{L}_m$. The objective function in (2.34) is the $\mathcal{L}_2$-norm regularized margin with which the *candidate* SHT misclassification costs pass the NIAC* convex feasibility test of (2.31). In (2.34), $\lambda > 0$ is a tunable regularization parameter and $G(\cdot)$ is the expected misclassification cost defined in (2.12). Setting λ to 0 yields the max-margin IRL estimate of the stopping agent's misclassification costs and lies within the feasible set of costs generated by Theorem 6. The other extreme is setting λ to ∞ which results in $\bar{L}^* = 0$.

The following numerical example illustrates regularized IRL (2.34) for inverse SHT.

SHT environments. We consider $M = 100$ SHT environments with:

- *Prior* $\pi_0 = [1/4 \;\; 1/4 \;\; 1/4 \;\; 1/4]'$ (The state space is now $\mathcal{X} = \{1, 2, 3, 4\}$).

- *Observation likelihood*: $p(y|x=1) = \mathcal{N}(-2, 8),\ p(y|x=2) = \mathcal{N}(0, 8)$, , $p(y|x=3) = \mathcal{N}(2, 8)$ and $p(y|x=4) = \mathcal{N}(4, 8)$.
- *Misclassification costs*: The misclassification costs $\bar{L} = \{\bar{L}_{m,x,a}\}$ in the M environments is uniformly sampled from the interval $[4, 10]^{M \times X \times (X-1)}$.

Inverse Learner Specification: The inverse learner aggregates the dataset $\mathcal{D}_M(\mathrm{SHT})$ according to the procedure described in (2.33) by generating $K = 10^7$ independent trials for the SHT agent in all $M = 100$ environments. Then, the inverse learner computes the regularized max-margin IRL estimate $\bar{L}^*$ by solving the optimization problem (2.34).

IRL Results: The inverse learner performs IRL by using the dataset $\mathcal{D}_M(\mathrm{SHT})$ to solve the optimization problem (2.34). Recall the dataset $\mathcal{D}_M(\mathrm{SHT})$ is generated by observing the actions of an SHT agent in multiple environments with misclassification costs $\bar{L}$. Figure 4 shows the estimation error $\|\bar{L}^* - \bar{L}\|_2 / \|\bar{L}\|\|_2$ of the inverse learner's IRL estimate $\bar{L}^*$ (2.34) computed by the inverse learner as the regularization parameter λ in (2.34) is varied. The error is normalized wrt the $\mathcal{L}_2$-norm of the true misclassification costs in multiple environments incurred by the Bayesian agent whose actions comprise $\mathcal{D}_M(\mathrm{SHT})$.

The least estimation error obtained by varying λ over the interval $[0, 100]$ was observed to be 0.042. In other words, the point IRL estimate obtained by solving the optimization problem (2.34) can estimate the true misclassification costs of the SHT environments with up to 95% accuracy. Indeed, the estimation accuracy increases with the number of environments at the cost of greater computation resources. Second, we observed that the error starts increasing sharply from $\lambda \sim 75$. This is expected since the regularization term in (2.34) dominates the margin term at large values of λ.

2.3.6 Performance Comparison. IRL for Inverse SHT and existing IRL methods for POMDPs

In this section, we compare the IRL performance of Theorem 6 for inverse SHT against two well-known algorithms for IRL of POMDPs, namely, Max-Margin between Values (MMV) [1, Alg. 4] and Max-Margin between Feature Expectations (MMFE) [1, Alg. 5]. We compare the performance of MMV and MMFE algorithms against max-margin inverse SHT (2.34) with regularization parameter λ set to 0.

Recall from (2.30) that our inverse SHT result of Theorem 6 requires state-terminal action pairs of the SHT agent over several independent trials and the expected stopping time of the SHT agent. In comparison, MMV and MMFE do not require the expected stopping time, but instead require complete knowledge of: (a) the observation likelihood of the Bayesian agent, and (b) the beliefs of the SHT agent at every time step. Moreover, MMV and MMFE require a POMDP solver for IRL.

To compare the performance of our IRL scheme (2.34) against MMV and MMFE, we perform two sets of numerical experiments with different specifications of the agent's observation likelihood:

Case 1: Perfect Knowledge of SHT Model Dynamics. MMV and MMFE have perfect knowledge of the SHT agent's observation likelihood.

Case 2: Misspecified SHT Model Dynamics. MMV and MMFE have misspecified knowledge of the SHT agent's observation likelihood. For environment m the observation likelihood $p_m(y|x)$ is misspecified to be the agent's action policy $p_m(a|x)$.

Experimental Setup

For our numerical experiments, we consider $M = 4$ SHT environments with:

- *Prior* $\pi_0 = [1/2 \;\; 1/2]'$ (The state space is $\mathcal{X} = \{1, 2\}$).
- *Observation likelihood*: $p(y|x = 1) = \mathcal{N}(+2, 4),\; p(y|x = 2) = \mathcal{N}(-2, 4)$,
- *Misclassification costs*: The misclassification costs $\bar{L} = \{\bar{L}_m, m \in \mathcal{M}\}$ in the M environments are uniformly sampled from the interval $[5, 25]$ for all states and actions in $\mathcal{X}$. Recall that we assume the continue cost is set to 1 WLOG.

For every environment $m = 1, 2, 3, 4$, we computed $\bar{L}_{m,\mathrm{MMV}},\; \bar{L}_{m,\mathrm{MMFE}}$ and $\bar{L}_{m,\mathrm{Margin}}$, the point-valued IRL estimate of the agent's misclassification cost from MMV, MMFE and max-margin inverse SHT (defined in (2.34) with regularization parameter $\lambda = 0$), respectively. For estimated misclassification cost $\bar{L}_{m,\mathrm{est}} \in \{\bar{L}_{m,\mathrm{MMV}},\; \bar{L}_{m,\mathrm{MMFE}},\; \bar{L}_{m,\mathrm{Margin}}\}$ with true cost $\bar{L}_m$ and chosen stopping strategy μ_m (Lemma 2), the normalized IRL estimation error is defined as:

$$\text{IRL Estimation Error} = \frac{|J(\mu_m, \bar{L}_m) - J(\mu_m, \bar{L}_{m,\mathrm{est}})|}{J(\mu_m, \bar{L}_m)}, \tag{2.36}$$

where $J(\cdot)$ is the expected cumulative cost defined in (2.9).

Our experimental results are displayed in Fig. 2.5. Our results show that our proposed IRL algorithm yields a lower IRL estimation error (2.36) than MMV and MMFE algorithms when model dynamics are misspecified. We observe that, on average, our max-margin IRL algorithm yields 60% lower estimation error compared to MMV and MMFE algorithms with misspecified model dynamics, and yields 27% higher estimation error compared to MMV and MMFE algorithms with accurate model dynamics.

Key Findings

Our key findings from the numerical experiments[11] can be summarized as:

- For the case of perfect knowledge of model dynamics (case 1), we observed that the MMV and MMFE algorithms of [1] perform better than max-margin IRL (2.34), and yield approximately $\mathbf{27}\%$ lower IRL estimation error compared to max-margin IRL. This is expected since both MMV and MMFE have access to private information the forward learner uses for decision-making and hence generates a more accurate IRL estimate.
- When the model dynamics are misspecified (case 2), our max-margin IRL algorithm outperforms both MMV and MMFE algorithms and yields approximately $\mathbf{60}\%$ lower IRL estimation error compared to MMV and MMFE.

Indeed, when no assumptions are placed on the underlying POMDP structure like in [1], achieving IRL requires perfect knowledge of the model dynamics. Hence, MMV and MMFE fail when model dynamics are misspecified.

Perspective

Cases 1 and 2 highlight the fact that our approach is complementary to that of [1]. [1] achieve IRL where the model dynamics are perfectly specified (case 1). In comparison, our IRL methods yield necessary and sufficient methods for optimal Bayesian stopping when no knowledge of model dynamics is provided to the inverse learner.

[11] All our numerical results are completely reproducible and can be accessed from the GitHub repository `https://github.com/KunalP117/YouTube-Commenting-Analysis`

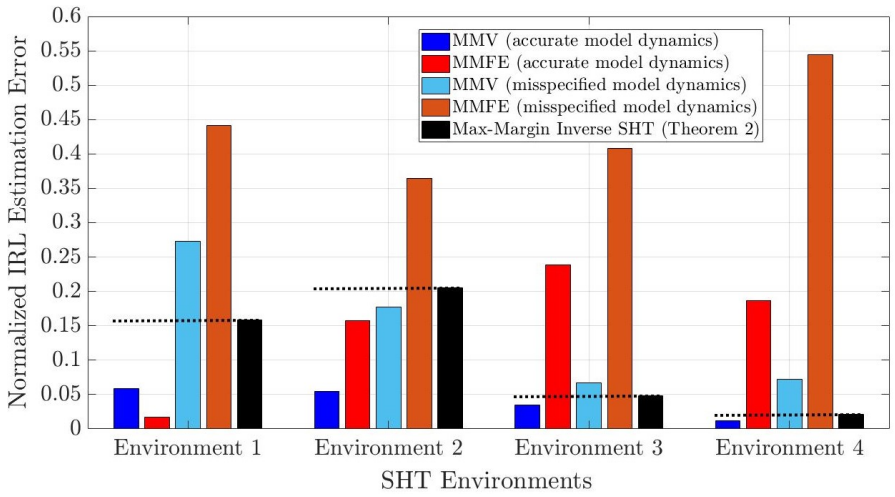

Figure 2.5: Inverse SHT Performance Comparison. Max-Margin NIAS-NIAC Test of Theorem 6 versus MMV and MMFE [1]

2.3.7 Summary

Theorem 6 specified necessary and sufficient conditions for identifying an optimal SHT agent acting in multiple environments. These conditions constitute a linear feasibility program that the inverse learner can solve to estimate SHT misclassification costs of the environments. The IRL task of solving the inverse SHT problem is more structured than the inverse optimal stopping problem in Sec. 2.2, since the agent's costs are partially known (expected continue cost is known) to the inverse learner. Hence, the feasible set of costs generated using Theorem 6 is smaller than that generated by Theorem 3 for the inverse SHT problem. We also proposed an IRL algorithm for point-valued estimation of the environments' misclassification costs and illustrated its performance in Sec. 2.3.5. Our key finding is that this point-valued IRL algorithm reconstructs the misclassification costs with up to 95% accuracy. Recall from Sec. 2.8.2 that an online user in multimedia platforms can be viewed as a Bayesian agent performing SHT. In the context of online multimedia platforms, the continue and stopping cost of the SHT agent can be viewed as the online user's sensing cost (attention to visual cues) and preference for viewing the online content, respectively. Hence, the numerical example in Sec. 2.3.5

can be viewed as an IRL methodology to reconstruct an online user's preferences for advertisements/movie thumbnails by observing his/her actions in multiple environments (webpages). We illustrate this claim in Sec. 2.5 with an IRL analysis on a real-world dataset. Finally, in Sec. 2.3.6 we compared our inverse SHT algorithm to two existing algorithms in the literature for IRL for POMDPs, namely, MMV and MMFE [1]. Our key observation was that our inverse SHT algorithm outperforms MMV and MMFE in scenarios where the inverse learner has limited information about the forward learner, i.e. , the learner's model dynamics are misspecified.

2.4 Example 2. IRL for inverse search

In this section, we present a second example of IRL for an optimal Bayesian stopping time problem, namely, *inverse* Bayesian Search. In the search problem, a Bayesian agent sequentially searches over a set of target locations until a static (non-moving) target is found. The optimal search problem is a special case of a Bayesian multi-armed bandit problem, and also of the optimal Bayesian stopping problem discussed in Sec. 2.2.2 since the continue cost (2.2) is the cost of searching a location and the stopping cost is 0 in the Bayesian search problem. Our IRL task in this section will be to estimate the search costs.

The optimal search problem is a modification of the sequential stopping problem in Sec. 2.2 with the following changes:

- There is only 1 stop action but multiple continue actions, namely, which of the X locations to search at each time. We will call the continue actions as search actions, or simply, actions.
- The observation likelihood B depends both on the true state x^o and the continue

action a.

Suppose an inverse learner observes the decisions of a Bayesian search agent over M search environments. The aim of the inverse search problem is to identify if the search actions of the agent are optimal and if so, estimate their search costs. Our IRL result for Bayesian search (Theorem 9 below) gives a necessary and sufficient condition for identifying an optimal search agent (formalized in Lemma 8 below) as equivalent to the existence of a feasible solution to a set of linear inequalities.

2.4.1 Optimal Bayesian search agent in multiple search environments

Suppose an agent searches for a target location $x \in \mathcal{X}$. When the agent chooses action $a \in \mathcal{X}$ to search location a, it obtains an observation y. Assume the agent knows the set of conditional pmfs of y, namely, $\{p(y|x^o = x), x \in \{1, 2, \dots X\}\}$. The aim of optimal search is to decide sequentially which location to search at each time to minimize the cumulative search cost until the target is found.

We define an optimal Bayesian search agent in M search environments as

$$\Xi_{opt} = (\mathcal{X}, \pi_0, \mathcal{Y}, \mathcal{A}, \boldsymbol{\alpha}, \{l_m, \mu_m, m \in \mathcal{M}\}) \tag{2.37}$$

where

- $\mathcal{X} = \{1, 2, \dots X\}$ is a finite set of states (target locations).
- At time 0, the true state $x^o \in \mathcal{X}$ is sampled from prior pmf π_0. This location x is not known to the agent but is known to the inverse learner (performing IRL).
- $\mathcal{Y} = \{0, 1\}$, where $y = 1$ (found) and $y = 0$ (not found) after searching a location.

- The set of actions $\mathcal{A} = \mathcal{X}$, $a \in \mathcal{A}$ is the location searched by the agent.
- The Bayesian agent in search environments m incurs instantaneous cost $l_m(a) > 0$ for searching location a.
- $\boldsymbol{\alpha} = \{\alpha(a), a \in \mathcal{A}\}$, $\alpha(a)$ is the reveal probability for location a, i.e. , the probability that the target is found when the agent searches the target location ($x = a$) in search environment $m \in \mathcal{M}$. $\boldsymbol{\alpha}$ characterizes the action dependent observation likelihood $B(y, x, a)$.

$$B(y, x, a) = p(y|x, a) = \begin{cases} \alpha(a), & y = 1, x = a \\ 1 - \alpha(a), & y = 0, x = a \\ 1, & y = 0, x \neq a. \end{cases} \tag{2.38}$$

  For IRL identifiability, we assume that the reveal probabilities are the same for all search environments in $\mathcal{M}$.
- $\{\mu_m, m \in \mathcal{M}\}$ are the optimal search strategies of the Bayesian agent over all environments in $\mathcal{M}$, when the agent operates sequentially on a sequence of observations $y_1, y_2, \ldots$ as discussed below in Protocol 2.

Protocol 2 *Sequential Decision-making protocol for Search:*

1. *Generate $x^o \sim \pi_0$ at time $t = 0$.*
2. *At time $t \geq 1$, agent records observation $y_t \sim B(\cdot, a_{t-1}, x^o)$.*
3. *If $y_t = 1$, then stop. Otherwise, if $y_t = 0$:*

   (i) Update belief $\pi_{t-1} \to \pi_t$ (described below).

   (ii) For search policy μ, agent takes action $a_t = \mu(\pi_t)$. (Note the first action is taken at time $t = 0$, while the first observation is at $t = 1$).

   (iii) Set $t = t + 1$ and go to Step 2.

Belief Update: *Let $\mathcal{F}_t$ denote the sigma-algebra generated by the action and observation sequence $\{a_1, y_1, \ldots a_t, y_t\}$. The agent updates its belief $\pi_t = \mathbb{P}(x^o = x|\mathcal{F}_t),\ x \in \mathcal{X}$ using Bayes formula as*

$$\pi_t = \frac{B(y_t, a_{t-1})\pi_{t-1}}{\mathbf{1}' B(y_t, a_{t-1})\pi_{t-1}}, \tag{2.39}$$

where $B(y, a) = \text{diag}(\{B(y, x, a), x \in \mathcal{X}\})$. The belief π_t is an $X-$dimensional probability vector belonging to the $(X-1)$ dimensional unit simplex (2.4).

Remark: The search agent's stopping region is simply the set of distinct vertices of the $X-1$ dimensional unit simplex.

We define the random variable τ as the time when the agent stops (target is found).

$$\tau = \inf\ \{t > 0|\ y_t = 1\} \tag{2.40}$$

Clearly, the set $\{\tau = t\}$ is measurable wrt $\mathcal{F}_t$, hence, the random variable τ is adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$. Below, we define the optimal search strategies $\{\mu_m, m \in \mathcal{M}\}$.

Definition 7 (Search strategy optimality) *The optimal search strategy μ_m of the Bayesian agent operating according to Protocol 2 in environment $m \in \mathcal{M}$ that minimizes the agent's cumulative expected search cost is well known [19] to be a stationary policy as defined below:*

$$J(\mu_m, l_m) = \min_{\mu} J(\mu, l_m) = \mathbb{E}_\mu \left\{ \sum_{t=0}^{\tau-1} l_m(\mu(\pi_t)) \right\}, \tag{2.41}$$

$$\mu_m(\pi) = \underset{a\in\mathcal{A}}{\text{argmax}} \left(\frac{\pi(a)\alpha}{l_m(a)} \right). \tag{2.42}$$

Here, $\mathbb{E}_\mu\{\cdot\}$ denotes expectation parametrized by μ induced by the probability measure $\{a_t, y_{t+1}\}_{t=1}^{\tau-1}$, $J(\cdot)$ denotes the expected search cost and μ belongs to the class of stationary search strategies.

Remarks. (1) Note that the minimization in (2.41) is over stationary search strategies. It is well known that the optimal search strategy has a threshold structure [19]. Since the set of all threshold strategies forms a compact set, we can replace the '$\inf$' in (2.9) for generic optimal stopping problems by '$\min$' in (2.41).

(2) Since the expected cumulative cost of an agent depends only on the search costs (for constant reveal probabilities), we can set $l_m(1) = 1,\ \forall m \in \mathcal{M}$ WLOG.

2.4.2 IRL for inverse search. Main result

In this subsection, we provide an inverse learner-centric view of the Bayesian stopping time problem and the main IRL result for inverse search. Suppose the inverse learner observes a search agent taking actions over M search environments where the agent performs several independent trials of Protocol 2 for Bayesian sequential search in each environment. We make the following assumptions about the inverse learner performing IRL to identify if $\mathcal{M}$ comprises an optimal search agent.

(A6) The inverse learner knows the dataset

$$\mathcal{D}_M(\text{Search}) = (\pi_0, \{g_m(a,x), m \in \mathcal{M}\}). \tag{2.43}$$

Here, $g_m(a,x)$ is the average number of times the agent searches location a when the target is in x in environment m:

$$g_m(a,x) = \mathbb{E}_{\mu_m}\left\{\sum_{t=1}^{\tau} \mathbb{1}\{\mu_m(\pi_t) = a\}|x\right\}. \tag{2.44}$$

We call $g_m(a,x)$ as the agent's *search action policy* in search environment m.

(A7) In dataset $\mathcal{D}_M(\text{Search})$, there are at least $M \geq 2$ environments with distinct search costs.

Assumption (A6) is discussed after the main result. In complete analogy to (A2), assumption (A7) is needed for identifiability of the search costs. We emphasize that the inverse learner only has the average number of times the agent searches a particular location in any environment. The inverse learner does not know the stopping time or the order in which the agent search the locations. In completely analogy to Lemma 2, Lemma 8 below specifies the inverse learner's identifiability of an optimal search agent under assumptions (A6) and (A7):

Lemma 8 (IRL identifiability of optimal Bayesian search agent) *The inverse learner identifies the tuple* Ξ_{opt} (2.37) *as an optimal Bayesian search agent iff* (2.45) *holds.*

$$J(\mu_m, l_m(a)) \leq J(\mu_n, l_m(a)),\ \forall\ m, n \in \mathcal{M},\ m \neq n. \tag{2.45}$$

In complete analogy with (2.12) in Lemma 2 for identifying an optimal stopping time agent, $J(\cdot)$ *in the above equation is the expected cumulative search cost of the agent.*

We omit the proof of Lemma 8 since it is identical to that of Lemma 2. Eq. 2.45 in Lemma 8 is analogous to (2.12) in Lemma 2. The inverse learner simply checks if the expected cumulative search cost for environment m is the smallest possible given the finite strategies $\{\mu_m, m \in \mathcal{M}\}$. We are now ready to present our main IRL result for the inverse search problem. The result specifies a set of linear inequalities that are simultaneously necessary and sufficient for a search agent's actions in multiple environments $\mathcal{M}$ to be identified as that of an optimal search agent (2.45).

Theorem 9 (IRL for inverse Bayesian search) *Consider the inverse learner with dataset* $\mathcal{D}_M(\text{Search})$ *(2.43) obtained from a search agent acting in multiple environments* $\mathcal{M}$*. Assume (A6) holds. Then:*

1. *Identifiability: The inverse learner can identify if the dataset* $\mathcal{D}_M(\text{Search})$ *is generated*

by an optimal search agent (Definition 8).

2. *Existence: There exists an optimal search agent parameterized by tuple* Ξ_{opt} (2.37) *if and only if there exists a feasible solution to the following linear (in search costs) inequalities:*

$$\textit{Find } l_m(a) \in \mathbb{R}_+, l_m(1) = 1 \quad \textit{s.t.} \quad \text{NIAC}^\dagger(\mathcal{D}_M(\text{Search})) \leq 0, \textit{ where}$$

$$\text{NIAC}^\dagger: \sum_{x \in \mathcal{X}} \pi_0(x)(g_m(a,x) - g_n(a,x))\, l_m(a) < 0 \; \forall m, n \in \mathcal{M},\; m \neq n. \tag{2.46}$$

3. *Reconstruction: The set-valued IRL estimate of the agent's search costs in environments* $\mathcal{M}$ *is the set of all feasible solutions to the NIAC*† *inequalities.* ■

The proof of Theorem 9 is in Sec. 2.11. Theorem 9 provides a set of linear inequalities whose feasibility is equivalent to identifying the optimality of a Bayesian search agent in multiple environments with different search costs. Note that Theorem 9 uses the search action policies $\{p_m(a|x), m \in \mathcal{M}\}$ to construct the expected cumulative search costs of the agent in multiple environments and verify if the inequality for identifying optimality (2.45) for Bayesian search holds. The key idea for the IRL result is to express the expected cost of the search agent in environment m in terms of its chosen search action policy $g_m(a,x)$ (2.43). Algorithms for linear feasibility such as the simplex method [25] can be used to check feasibility of (2.46) in Theorem 9 and construct a feasible set of search costs for the optimal search agent.

Discussion of assumption (A6). To motivate (A6), suppose for each environment $m \in \mathcal{M}$, the inverse learner records the state $x_{k,m}$ and agent actions $\{a_{1:\tau_{k,m},k,m}\}$ over $k = 1, 2, \ldots K$ independent trials. Then, the variable $g_m(a,x)$ in $\mathcal{D}_M(\text{Search})$ (2.43) is the limit pmf of the empirical pmf $\hat{g}_m(a,x)$ as the number of trials $K \to \infty$.

$$\hat{g}_m(a,x) = \frac{\sum_{k=1}^K \sum_{t=1}^{\tau_{k,m}} \mathbb{1}\{x_{k,m} = x, a_{t,k,m} = a\}}{\sum_{k=1}^K \mathbb{1}\{x_{k,m} = x\}}. \tag{2.47}$$

In complete analogy to Sec. 2.2.5, almost sure convergence holds by Kolmogorov's strong law of large numbers. $g_m(a, x)$ is the average number of times the agent searches location a when the target is in location x in environment m. More formally, for a fixed state x, $g_m(a, x)$ is the number of times the posterior belief of the agent visits the region in the unit simplex of pmfs where it is optimal to choose action a. In Sec. 2.11, we discuss how the search action policy $g_m(a, x)$ can be used to express the agent's cumulative expected search cost (2.41) in the m^{th} environment.

Remark: Analogous to the action selection policy (2.28) for stopping problems with multiple stopping actions, the inverse learner uses the search action policy to identify Bayes optimality in stopping problems with multiple continue actions (and single stop action).

2.4.3 Numerical example illustrating IRL for inverse search

We now present a numerical example for inverse search with 3 search environments and 3 search locations. The aim of this example is to illustrate the consistency property of Theorem 9. That is, that the true search costs lie in the set of feasible costs computed by the inverse learner by solving the feasibility test of Theorem 9.

Search environments. We consider $M = 3$ search environments with:

- *Prior* $\pi_0 = [1/3\ 1/3\ 1/3]'$.
- *Search locations:* $X = A = 3$.
- *Reveal probability*: $\alpha(1) = 0.7, \alpha(2) = 0.68, \alpha(3) = 0.6$.
- *Search costs*:

  Environment 1: $l_1(1) = 1,\ l_1(2) = 3,\ l_1(3) = 4$,

Environment 2: $l_2(1) = 1,\ l_2(2) = 1,\ l_2(3) = 2,$

Environment 3: $l_3(1) = 1,\ l_3(2) = 0.5,\ l_3(3) = 3.$

(Recall that WLOG the search cost $l_m(1)$ can be set to 1 for all $m \in \{1, 2, 3\}$.)

Inverse Learner specification. Next we consider the inverse learner. We generate $K = 10^6$ samples for the search agent in all 3 environments using the above parameters. Recall from Theorem 9 that the inverse learner uses the dataset $\mathcal{D}_M(\text{Search})$ to perform IRL for search. Here

$$\mathcal{D}_M(\text{Search}) = (\pi_0, (\hat{g}_m(a, x), m \in \{1, 2, 3\}), \tag{2.48}$$

where $K = 10^6$, the second term in the dataset is the empirically calculated search action policy (2.47) of the agent in environment m from the 10^6 generated samples.

IRL Result. The inverse learner performs IRL by using the dataset $\mathcal{D}_M(\text{Search})$ (2.48) to solve the linear feasibility problem in Theorem 9. The result of the feasibility test is shown in Fig. 2.6. The blue region is the set of feasible search costs for each environment. The feasible set of costs is $\{(l_m(2), l_m(3), m \in \{1, 2, 3\}\} \subseteq \mathbb{R}^6_+$. For visualization purposes, Fig. 2.6 displays the feasible search costs for each environment in a different sub-figure. In complete analogy to Fig. 2.3, the feasible search costs for each environment are shown in each sub-figure by keeping the search costs of the other 2 environments fixed at their true values. The true search cost for every environment is highlighted by a yellow point. The key observation is that these true costs belong to the set of feasible costs (blue region) computed via Theorem 9. Thus, Theorem 9 successfully performs IRL for the search problem and the set of feasible search costs can be reconstructed as the solution to a linear feasibility problem.

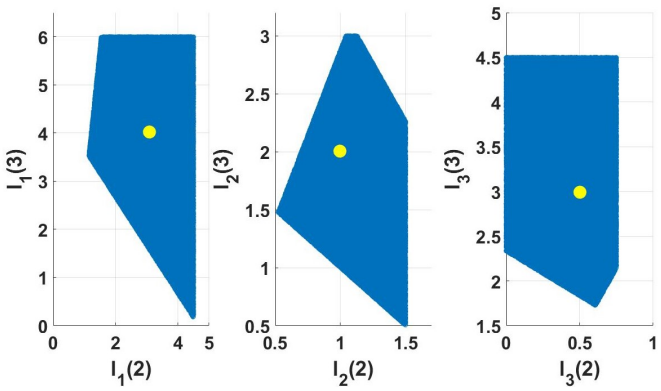

Figure 2.6: Numerical example for inverse search with parameters specified in Sec. 2.4.3. The key observation is that the true search costs (yellow points) lie in the feasible set (blue region) of costs computed via Theorem 9. This follows from the necessity proof of Theorem 9 which says if the agent is an optimal search agent, then the true costs lie in the feasible set of costs that satisfy the NIAC† inequalities.

2.5 Inverse Optimal Stopping for Predicting YouTube Commenting Behavior

In this section, we illustrate our IRL results for Bayesian stopping time problems on a real-world YouTube dataset. Although we use the same dataset in previous work [12], our IRL methodology and experimental results are new. For brevity, we discuss the key differences compared to [12] and justify our choice of Bayesian stopping for modeling user engagement on YouTube in Sec. 2.14.

We consider a YouTube dataset comprising approximately 140000 videos across $25,000$ channels spanning 18 video categories and over 9 millions users from April 2007 to May 2015. The diversity of videos in YouTube is immense; it is intuitive to exploit this diversity for understanding how groups of YouTube users exposed to different classes of video content engage differently with YouTube. Hence, by analyzing groups of YouTube users indexed by video category, our aim is to:

(1) Identify if YouTube user engagement is consistent with Bayesian optimal stopping,

and if so,

(2) Reconstruct the stopping costs of user engagement using the IRL results in this chapter, and

(3) Use the reconstructed costs to predict user engagement in videos.

Our YouTube dataset does not contain any information (visual cues) about what the human user perceives from the video webpage before choosing to engage on the YouTube platform. Recall from Theorem 3 that our IRL approach does not depend on the unobserved model dynamics that generate the IRL dataset (2.11). This makes our IRL methodology well-suited to scenarios where the parameters of the underlying decision making process are not available in the IRL dataset. Our main conclusions from our IRL analysis of the YouTube dataset can be summarized as:

- YouTube user engagement is consistent with optimal Bayesian stopping. Based on our IRL analysis on groups of YouTube users, where each group consists of approximately 3500 viewers, the YouTube dataset (described below in (2.49)) satisfies the NIAS and NIAC feasibility inequalities of Theorem 3 for optimal Bayesian stopping with a high margin.
- By choosing two representative points from the feasible set of costs generated by IRL (2.16), (2.17), namely, max-margin estimate and entropy-regularized estimate defined below, we show our reconstructed IRL costs predict user engagement with high accuracy. Figure 2.8 illustrates the predictive performance of our IRL methodology.

2.5.1 YouTube Dataset and Model Parameters

Categories in YouTube (e.g. News, Gaming, Music etc.) are numbered from $1 - 18$ (See Fig. 2.7 for the full listing). The video categories have mean numbers of users ranging from 149 to 4596 for high viewcount (greater than 10000) videos and 8 to 1801 for low viewcount videos (less than 10000). Figure 2.7 lists each video category along with the total number of views. Note that the video categories "Unavailable" or "Removed" are videos flagged by YouTube as being suspected of violating YouTube's video policies.

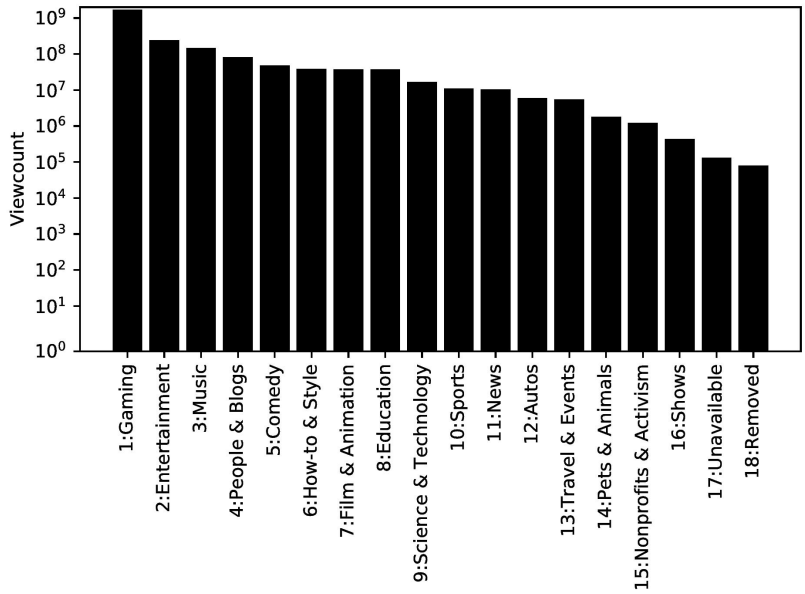

Figure 2.7: YouTube Dataset Overview. Viewcount summed over all videos (vertical axis) of $M = 18$ video categories. The 18 categories are listed on the horizontal axis.

The YouTube dataset contains the view counts, comment counts, likes, dislikes, thumbnail, title, and category of each video. To relate to our main IRL result of Theorem 3, we define the following:

1. *Agent*: Group of users interacting with videos in each video segment. User engagement in different video categories can be interpreted as the agent acting in multiple environments. In the rest of the section, we will use the terms 'user engagement' and 'commenting behavior' interchangeably.

2. *State (x)*: In the YouTube dataset, the state x of each video is the viewcount 1 day after the video was published. Specifically, state $x = 1$ is high viewcount (more than $10,000$ views) and $x = 2$ otherwise. In YouTube, video viewcount is the independent quantity which governs the commenting behavior since videos need to be viewed first before users can comment or rate the video.

3. *Terminal Action (a)*: In the YouTube dataset, the terminal action a is related to the overall commenting behavior[12] of the users, which is computed using the comment counts, like count, and dislike count 2 days after the video is published. The possible actions are: $a = 1$ denotes low comment count with negative sentiment, $a = 2$ denotes low comment count with neutral sentiment, $a = 3$ denotes low comment count with positive sentiment, $a = 4$ denotes high comment count with negative sentiment, $a = 5$ denotes high comment count with neutral sentiment, and $a = 6$ denotes high comment count with positive sentiment. Here negative sentiment occurs if the difference between the like count and dislike count is less than -25, neutral sentiment occurs if the difference lies between $-25, 25$, and positive sentiment occurs if the difference is greater than 25. A low comment count is said to occur if there are less than 100 comments, otherwise the comment count is defined to be high.

4. *Observation (y)*: The observation y for a YouTube user abstracts the visual cues a user perceives that depends on video metadata such as thumbnail, title, category etc. The observation likelihood is indicative of the attention expended by the user on a video. We note that although neither the observations y nor the observation likelihood $p(y|x)$ are contained in the YouTube dataset, our IRL algorithm abstracts away these unobserved model parameters, and still yields necessary and sufficient conditions for Bayes optimality.

4. *Environment (m)*: Environment m corresponds to each of the $M = 18$ video cate-

[12]By overall commenting behavior in YouTube, we mean both the comment count and the video ratings (likes and dislikes). Another term used in the literature [36] is "user engagement".

gories in our YouTube dataset. Fig. 2.7 lists each video category with the total number of views. Note that the video categories "Unavailable" or "Removed" are videos flagged by YouTube as being suspected of violating YouTube's video policies[13].

Recall from Sec. 2.2.3 that the inverse learner requires knowledge of the dataset $\mathcal{D}_M = (\pi_0, \boldsymbol{p})$ (2.11) for identifying optimal Bayesian stopping via Theorem 3. In the YouTube context, the variables $\pi_0, \boldsymbol{p} = \{p_m(a|x), m \in \mathcal{M}\}$ dataset $\mathcal{D}_M$ can be constructed as:

$$\pi_0(x) = \frac{1}{I}\sum_{i=1}^{I} \mathbb{1}\{x_i = x\}, \;\; p_m(a|x) = \frac{\sum_{i=1}^{I} \mathbb{1}\{x_i = x, a_i = a, \text{category}_i = m\}}{\sum_{i=1}^{I} \mathbb{1}\{x_i = x, \text{category}_i = m\}}, \tag{2.49}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, variable i indexes the YouTube videos, $I = 140000$ is the total number of YouTube videos in the dataset, and environment $m \in \{1, 2, \ldots, 18\}$ indexes the video categories. Also, $x_i, a_i, \text{category}_i$ denote the state, action and category of the YouTube video indexed by i, where the state and action interpretations for the YouTube videos are discussed above.

2.5.2 YouTube Data Analysis Results

We now discuss our experimental findings from our IRL analysis on the YouTube dataset.[14] Our main task is to predict YouTube's commenting behavior, that is, the action selection policy $p_m(a|x)$ in video category m using the IRL algorithms in this chapter. Our first observation is that the dataset $\mathcal{D}_M$ (2.49) comprising YouTube commenting behavior over $M = 18$ categories passes the convex feasibility test (2.16) and (2.17)

[13]Refer to `https://www.youtube.com/yt/about/policies/#community-guidelines` for details

[14]All our numerical results are completely reproducible and can be accessed from the GitHub repository `https://github.com/KunalP117/YouTube-Commenting-Analysis`.

of Theorem 3 with a high margin of 1.85×10^{-3}, where the margin is normalized by the maximum feasible cost $\max_{m,x,a} s_m(x,a)$. This shows that there exists a Bayesian stopping model that rationalizes YouTube commenting behavior.

We now illustrate how well the reconstructed costs from the feasibility test of Theorem 3 predict the commenting behavior of YouTube videos in different categories. For our prediction task, first, we randomly divided the YouTube dataset into two parts - training data (80%) and testing data (20%). Also, we consider only a subset of the 18 video categories for which the number of videos exceeds 200. This extra condition results in 9 out of 18 video categories considered for our IRL prediction analysis. For predicting commenting behavior via IRL, we first consider the training data and compute two point-valued estimates of the agent stopping costs that satisfy the NIAS and NIAC inequalities of Theorem 3, namely, max-margin IRL and entropy-regularized IRL defined below:

Max-Margin IRL :

$$\{\boldsymbol{S}_{\text{MM-IRL}},\ \epsilon^*\} \ = \ \underset{\epsilon \geq 0, \boldsymbol{S} \geq \boldsymbol{0}}{\text{argmax}}\, \epsilon, \text{ such that } \text{NIAS}(\mathcal{D}_M, \boldsymbol{S}) \leq -\epsilon, \ \text{NIAC}(\mathcal{D}_M, \boldsymbol{S}) \leq -\epsilon, \tag{2.50}$$

Entropy-Regularized IRL :

$$\boldsymbol{S}_{\text{Ent-IRL}} \ = \ \text{Any feasible cost } \boldsymbol{S} \equiv \{s_m(x,a), x \in \mathcal{X}, a \in \mathcal{A}\}_{m=1}^{M}, \ \text{ that satisfies} \tag{2.51}$$

$$(a) \ \ s_m(x, a_1) = 1, \ \forall\, x \in \mathcal{X} \text{ (Normalization)}, \text{ and}$$

$$(b) \ \ \text{NIAS}(\mathcal{D}_M, \boldsymbol{S}) \leq 0, \ \text{NIAC}(\mathcal{D}_M, \boldsymbol{S}) \leq 0, \ \text{SUMCOST}(\mathcal{D}_M, \boldsymbol{S}, \{MI(\pi_0; p_m(a|x))\}_{m=1}^{M}) \leq 0.$$

In (2.50) and (2.51) above, $\boldsymbol{S} = \{s_m(x,a), m \in \mathcal{M}, x \in \mathcal{X}, a \in \mathcal{A}\}$ denotes the set of stopping costs over all environments $\mathcal{M}$, states $\mathcal{X}$ and actions $\mathcal{A}$; the NIAS, NIAC and SUMCOST feasibility inequalities are defined in (2.16), (2.17) and (2.18), respectively. In (2.51), $MI(\pi_0; p_m(a|x))$ denotes the mutual information between the agent's prior π_0

and action selection policy $p_m(a|x)$ defined as:

$$MI(\pi_0; p_m(a|x)) = \sum_{x,a} \pi_0(x)\ p_m(a|x)\ \log\left(\frac{p_m(a|x)}{\sum_x \pi_0(x)\ p_m(a|x)}\right)$$

The intuition behind (2.50) is clear: choose the stopping costs that pass the feasibility inequalities of Theorem 3 with the largest margin. In (2.51), we impose the additional constraint that the expected continue cost is the mutual information between the prior and the action selection policy. The inspiration for this information-theoretic cost stems from the seminal work of [37] who modeled human attention as a limited-capacity communication channel, and from Max-Entropy IRL [28] in IRL literature. Eq. 2.51 yields a softmax structure for the feasible stopping costs (see Sec. 2.10.3 for a more detailed explanation); the key idea is that entropy-regularized IRL for Bayesian stopping yields a set of constant stopping costs, constant up to an affine monotone transformation.

For predicted cost $\{s_m(x,a), x \in \mathcal{X}, a \in \mathcal{A}, m \in \mathcal{M}\}$ and action selection policies $\{p_m(a|x), m \in \mathcal{M}\}$ from the training dataset, the predicted action selection policy $\hat{p}_m(a|x)$ for the test dataset is straightforwardly computed as:

$$\hat{p}_m(a|x) = \sum_{a'} \mathbb{1}\{a = \operatorname*{argmax}_b \sum_x \hat{p}_m(x|a')\ s_m(x,b)\}\ p_m(a'|x), \text{ where} \qquad (2.52)$$

the probability $\hat{p}_m(x|a') = \frac{\pi_{0,\text{test}}(x)\ p_m(a|x)}{\sum_x \pi_{0,\text{test}}(x)\ p_m(a|x)}$ is the predicted posterior belief of the state given action a' for the test dataset. Observe that all terms in the RHS of (2.52) pertain to the training dataset except for the prior $\pi_{0,\text{test}}$ that is empirically computed from the test dataset. Intuitively, (2.52) assumes the observation likelihood for the YouTube user in the test dataset is simply the action selection policy $p_m(a|x)$ from the training dataset. In words, the predicted action selection policy $\hat{p}_m(a|x)$ in (2.52) is obtained by simply summing the likelihoods of all actions $a' \in \mathcal{A}$ for which action a is optimal given posterior belief $\hat{p}(x|a')$.

Using (2.52), we obtained two sets of predicted action selection policies, namely,

$\hat{\boldsymbol{p}}_{\text{MM-IRL}} = \{\hat{p}_m(a|x), m \in \mathcal{M}\}_{\text{MM-IRL}}$ and $\hat{\boldsymbol{p}}_{\text{Ent-IRL}} = \{\hat{p}_m(a|x), m \in \mathcal{M}\}_{\text{Ent-IRL}}$ for the test dataset, corresponding to stopping costs $\boldsymbol{S}_{\text{MM-IRL}}$ (2.50) and $\boldsymbol{S}_{\text{Ent-IRL}}$ (2.51), respectively. To comment on the prediction accuracy, we computed the chi-squared distance and total variation distance between the true and predicted action selection policies for each video category m. [15] Figure 2.8 shows the IRL prediction results. We observed that for 7 out of the 9 video categories considered for IRL prediction analysis, the chi-squared and total variation distance for both sets of estimated action selection policies lie under 0.3. Hence, for 7 out of 9 video categories, our IRL algorithm successfully predicts the action selection policies in the test dataset with high accuracy. Another observation from Fig. 2.8 is that the max-margin IRL estimate is a more accurate predictor compared to the entropy-regularized IRL estimate and outperforms the entropy-regularized IRL in 2 out of 9 video categories.

Summary: We illustrated the predictive performance of our IRL algorithms (2.50), (2.51) on a real-world YouTube dataset. We chose two point-valued IRL estimates of stopping costs from the set of feasible costs that pass the NIAS (2.13) and NIAC (2.14) inequalities of Theorem 3, namely, max-margin IRL (2.50) and entropy-regularized IRL (2.51). We observed that both these cost estimates accurately predict YouTube commenting behavior (in terms of chi-squared and total variation distance as displayed in Fig. 2.8). Moreover, the max-margin IRL estimate yields a more accurate prediction compared to the entropy-regularized estimate.

[15] Both chi-squared and total variation distance are normalized by definition since they take values in the interval $[0, 1]$.

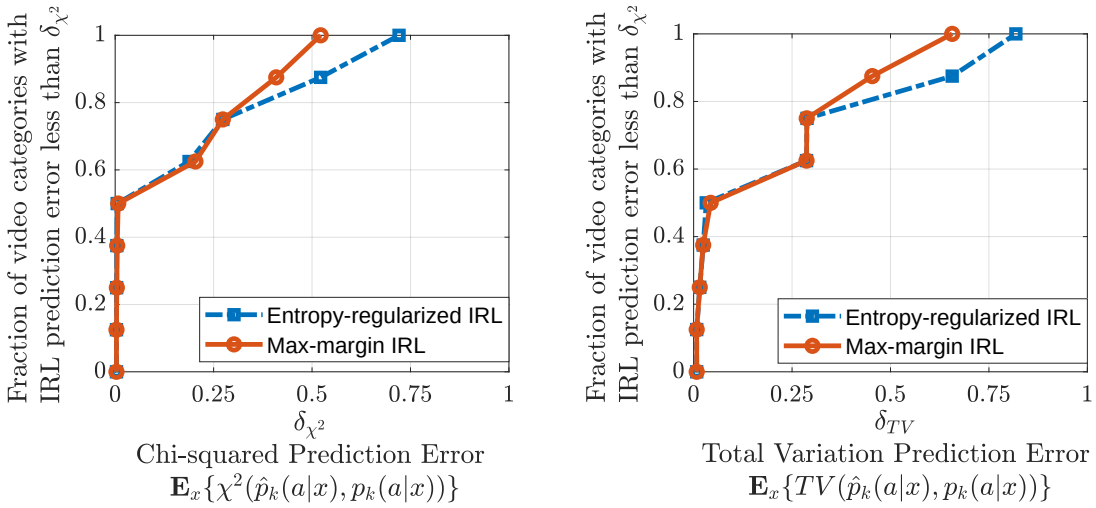

Figure 2.8: IRL Prediction Error for YouTube Dataset. The main takeaway is that point-valued IRL estimates that satisfy the feasibility test of Theorem 3 predict YouTube commenting behavior with high accuracy (low statistical distance between true and predicted distributions). For reconstructing the stopping costs, we choose two distinct point-valued stopping costs, namely, entropy-regularized IRL (2.51) and max-margin IRL (2.50) and performed our numerical experiments on 9 out of 18 video categories for which the video count exceeded 250. For both sets of estimated stopping costs, we observed that for 7 out of 9 YouTube video categories considered for analysis, both the chi-squared distance and total variation distance between the true and predicted action policy is less than 0.3.

2.6 Finite sample performance analysis of IRL decision test

Thus far, our IRL framework assumes (A1), namely, that the inverse learner has access to infinite trials of the stopping agent in M environments in order to solve the convex feasibility problem in Theorem 3. Suppose the inverse learner records only a finite number of trials and constructs its IRL dataset (2.7) comprising the agent's prior and empirically computed action selection policies in M environments. In this section, we address the following question: *How robust is the IRL decision test in Theorem 3 to finite sample datasets?* We now view Theorem 3 as a detector that takes in as input a noisy (empirical) dataset and outputs whether or not the observed agent is identified as an optimal stopping agent. Our aim is to provide bounds on the IRL detector's error

probability in terms of the number of trials recorded by the inverse learner. We then obtain finite sample IRL results for the examples of inverse SHT and inverse search.

2.6.1 Finite sample statistical test for IRL

Suppose the inverse learner observes the actions of a Bayesian stopping agent in M environments. In addition to assumption (A2), we assume the following about the inverse learner for our finite sample result stated in Theorem 11 below.

(F1) The inverse learner knows the *finite dataset*

$$\widehat{\mathcal{D}}_M(\mathcal{K}) = \{\pi_0, \{\hat{p}_m(a|x), m \in \mathcal{M}\}\}, \text{ where } \mathcal{K} = \{K_{x,m}, m \in \mathcal{M}, x \in \mathcal{X}\}. \tag{2.53}$$

In (2.53), $\mathcal{K} = \{K_{x,m}, m \in \mathcal{M}, x \in \mathcal{X}\}$, $K_{x,m}$ is the number of trials recorded by the inverse learner for environment m and state x. $\hat{p}_m(a|x)$ is the empirical action selection policy of the agent in environment m computed for $K_{x,m}$ trials via (2.28).

(F2) The finite dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ satisfies the following inequality.

$$\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K})), \varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K})) \geq \left(\sum_{x,m} \frac{A}{2K_{x,m}}\right) \left(\ln(2K_{x,m}/A) - \min_{x,m} \ln\left(2K_{x,m}/A\right)\right) \tag{2.54}$$

In (2.54), $K_m = \sum_x K_{x,m}$, $\bar{K} = K/\tau^2_{\max}$ and $\widetilde{K} = K^{-1}$. Eq. 2.54 imposes a lower bound on the number of samples needed for our sample complexity result of inverse optimal stopping. Eq. 2.54 is a sufficient condition for obtaining the constants of the sample complexity bound as the solution of a convex optimization problem; see (2.106) in the Appendix for more details. Variables $\varepsilon_1(\cdot), \varepsilon_2(\cdot)$ are the minimum perturbations needed for the finite dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ to satisfy and not

satisfy, respectively, the NIAS and NIAC inequalities in Theorem 3, and defined formally in (2.57), (2.58) for readability.

For the reader's convenience, we discuss the assumptions (F1) and (F2) after the finite sample complexity result, Theorem 11. The feasibility test of Theorem 3 given a finite number of trials $\mathcal{K}$ can be equivalently formulated as a statistical hypothesis detection test that takes as input the finite dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ and accepts one of the two hypotheses, H_0 or H_1.

- H_0: *Null hypothesis* that the observed stopping agent is identified as an optimal agent, i.e. , the true dataset $\mathcal{D}_M$ is feasible wrt the NIAS and NIAC inequalities (2.13), (2.14) in Theorem 3.
- H_1: *Alternative hypothesis* that the observed stopping agent is *not* optimal.

Definition 10 (IRL detector for inverse optimal stopping) *Consider the inverse learner with dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$. Assume (A2) and (F1) hold. The IRL decision test $\mathrm{Test}_{\mathrm{IRL}}(\cdot)$ for inverse optimal stopping is given by:*

$$\mathrm{Test}_{\mathrm{IRL}}(\widehat{\mathcal{D}}_M(\mathcal{K})) = \begin{cases} H_0, & \textit{if } \mathrm{IRL}(\widehat{\mathcal{D}}_M(\mathcal{K})) \neq \emptyset \\ H_1, & \textit{if } \mathrm{IRL}(\widehat{\mathcal{D}}_M(\mathcal{K})) = \emptyset. \end{cases} \tag{2.55}$$

Here, $\mathrm{IRL}(\mathcal{D})$ is the set of feasible solutions to the convex NIAS and NIAC inequalities (2.13), (2.14) given dataset $\mathcal{D}$.

The statistical test defined above is a detector that accepts the null hypothesis H_0 if the finite dataset passes the feasibility test of Theorem 3 and accepts the alternative hypothesis H_1 it otherwise. Our main result stated below characterizes the performance of the feasibility test in identifying optimality given finite sample constraints, namely, provide bounds on the detector's Type-I/II error probabilities.

2.6.2 Main result. Finite sample analysis for IRL

Our main result below (Theorem 11) characterizes the following error probabilities of the statistical test in Definition 10:

$$\text{Type-I error prob.}: \mathbb{P}(H_0|\,\mathrm{IRL}(\widehat{\mathcal{D}}_M(\mathcal{K}) = \emptyset), \text{ Type-II error prob.}: \mathbb{P}(H_1|\,\mathrm{IRL}(\widehat{\mathcal{D}}_M(\mathcal{K}) \neq \emptyset) \tag{2.56}$$

In (2.56), $\widehat{\mathcal{D}}_M(K)) = \emptyset$ means that the finite dataset fails the convex feasibility test for NIAC and NIAS inequalities (2.13), (2.14) and so the agent is identified as not an optimal agent. Our finite sample result in Theorem 11 below uses the dataset statistics variables $\varepsilon_1(\cdot), \varepsilon_2(\cdot), g(\cdot)$ from the finite dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ and are defined below. The quantities $\varepsilon_1(\cdot)$ and $\varepsilon_2(\cdot)$ are the minimum perturbations needed for the finite dataset to satisfy and not satisfy, respectively, the NIAS and NIAC inequalities in Theorem 3, and variable g is the constant for the error probability bounds.

Notation

Theorem 11 below uses the following variables:

$$\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K})) = \min_{\{\hat{p}'_m, m\in\mathcal{M}\}} \sum_m \|\hat{p}_m - \hat{p}'_m\|_2^2 \text{ such that } \mathrm{IRL}(\{\pi_0, \{\hat{p}'_m(a|x)\}\}) \neq \emptyset. \tag{2.57}$$

$$\varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K})) = \min_{\{\hat{p}'_m, m\in\mathcal{M}\}} \sum_m \|\hat{p}_m - \hat{p}'_m\|_2^2 \text{ such that } \mathrm{IRL}(\{\pi_0, \{\hat{p}'_m(a|x)\}\}) = \emptyset. \tag{2.58}$$

$$g(\widehat{\mathcal{D}}_M(\mathcal{K})) = \left(A\sum_{x,m}\widetilde{K}_{x,m}\right)\prod_{x.m}\left(\frac{2K_{x,m}}{A}\right)^{\frac{\widetilde{K}_{x,m}}{\sum_{x,m}\widetilde{K}_{x,m}}}, \text{ where } \widetilde{K}_{x,m} = K_{x,m}^{-1} \tag{2.59}$$

Having defined our notation for error probability bounds, let us now state our first sample complexity result for the IRL detector (2.55).

Theorem 11 (Sample complexity for IRL detector) *Consider an inverse learner with finite dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ (2.53). The inverse learner aims to detect optimality of the stopping agent's actions using the statistical test in Definition 10. Assume (A2), (F1) and (F2) hold. Then, the Type-I and Type-II error probabilities* (2.56) *of the IRL detector (Definition 10) are bounded as:*

$$\textit{Type-I error probability} \leq g(\widehat{\mathcal{D}}_M(\mathcal{K})) \exp\left(-\mathcal{K}_H \cdot \varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K}))\right), \tag{2.60}$$

$$\textit{Type-II error probability} \leq g(\widehat{\mathcal{D}}_M(\mathcal{K})) \exp\left(-\mathcal{K}_H \cdot \varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K}))\right). \tag{2.61}$$

In (2.59)*, $\mathcal{K}_H = \left(\sum_{x,m} K_{x,m}^{-1}\right)^{-1}$, and variables $\varepsilon_1(\cdot), \varepsilon_2(\cdot)$ and g are defined in* (2.57)*,* (2.58) *and* (2.59)*, respectively.*

The proof of Theorem 13 is in Sec. 2.12. Below, we provide a sketch of the proof. Theorem 11 characterizes the robustness of the IRL detector in Definition 10 to finite sample constraints. It provides an upper bound on the detector's error probabilities in terms of the number of trials recorded by the inverse learner. Observe that since $\mathcal{K}_H$ is simply the unnormalized harmonic mean of $\mathcal{K}$ (2.68), the error rate is exponential in the *harmonic mean* of the number of trials recorded over M environments and X states.

The proof of Theorem 11 uses the two-sided Dvoretzky-Kiefer-Wolfowitz (DKW) concentration inequality [38, 39] as the fundamental result to show that these error probabilities can be tightly bound in terms of the sample size $\mathcal{K}$ of the finite dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$. The DKW inequality provides a probabilistic bound on the deviation of the empirical cdf from the true cdf for i.i.d random variables. The i.i.d assumption holds for our detector in Definition 10 since the observed actions of the agent for a fixed state are independent and identically distributed over trials for all environments $\mathcal{M}$. To obtain our Type-I/II error bounds, we use the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality to probabilistically bound $\|p(a|x) - \hat{p}(a|x)\|_2$, the L_2-error between the empirical and true

action selection policy for each environment $m \in \mathcal{M}$ and state $x \in \mathcal{X}$, followed by the union bound to bound the sum of L_2-errors due to finite sample size over all states and environments.

Discussion of Assumptions.

(F1): Given the finite dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ in (2.53), (F1) says that the inverse learner checks if the convex feasibility test of Theorem 3 has a feasible solution to detect an optimal stopping agent.

(F2): Abstractly, (F2) says that the inverse learner observes sufficiently many trials of the agent over all environments $\mathcal{M}$ such that the condition (2.54) is met.

First, for a given dataset $\mathcal{D}$, note that only one out of $\varepsilon_1(\mathcal{D}), \varepsilon_2(\mathcal{D})$ is non-zero and positive. Hence, (2.54) involves only the non-zero variable out of $\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K})), \varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K}))$. Some words about the RHS of (2.54). $q(\mathcal{K}) - j(\mathcal{K})$ is a measure of how far is $K_{\min} = \min_{x,m} K_{x,m}$ from the remaining elements in $\mathcal{K} \backslash K_{\min}$. Since the RHS terms are of the form $\ln(z)/z$, it is easy to check that $q(\mathcal{K}) - j(\mathcal{K})$ decreases as the elements of $\mathcal{K}$ increase uniformly. As the number of samples go to infinity, the RHS in (2.54) tends to 0, hence the condition is almost surely satisfied for infinite samples. For finite $\mathcal{K}$, checking if (2.54) holds requires the inverse learner to solve an optimization problem (for the LHS) and perform MX multiplication operations and MX addition operations to compute the RHS of (2.54). As a practical estimate, for the inverse SHT task in Sec. 2.3.5 for 100 SHT environments, we observed that the inequality in (2.54) is satisfied if the samples exceeded $\sim 10^3$ for each environment.

2.6.3 Example 1. Finite sample effects for IRL in inverse SHT

We next turn to a finite sample analysis of IRL for inverse sequential hypothesis testing (SHT). Recall from Theorem 6 that identifying optimality of SHT is equivalent to feasibility of the linear inequalities NIAS and NIAC*. The inverse learner's SHT dataset comprises both the agent action selection policies and the expected stopping times to perform IRL compared to only the action selection policies for inverse optimal stopping. Hence, in addition to the DKW inequality, our main result, Theorem 13 also uses the Hoeffding's inequality [40] to account for the finite sample effect[16] on the computation of the expected stopping time.

Assumptions and Detection Test. Suppose the inverse learner observes the actions of the Bayesian stopping agent over M SHT environments. We assume the following about the inverse learner for our finite sample result stated below for the inverse SHT problem.

(F3) The inverse learner uses the *finite SHT dataset*

$$\widehat{\mathcal{D}}_M(\mathcal{K}) = \{\pi_0, \{\hat{p}_m(a|x), \hat{C}_m, m \in \mathcal{M}\}\} \tag{2.62}$$

to detect if the stopping agent is an optimal SHT agent or not. The variable $\mathcal{K}$ defined in (2.53) is the number of trials recorded by the inverse learner, $\hat{C}_m$ is the sample average of the agent's stopping time in the m^{th} environment. $\hat{p}_m(a|x)$ is the agent's empirical action selection policy computed for $K_{x,m}$ trials via (2.28) in the m^{th} environment.

(F4) The inverse learner knows $\tau_{\max} = \inf\ \{t > 0 \mid \mathbb{P}(\tau \leq t) = 1, \forall m \in \mathcal{M}\}$, an upper bound on the stopping time of the SHT strategies chosen by the agent in all environments $\mathcal{M}$.

[16] Hoeffding's inequality applies to bounded r.v.s , and is true for SHT since the stopping time τ is finite almost surely.

(F5) The finite dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ satisfies the following inequality.

$$\begin{aligned}
&\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K})), \varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K})) \geq q(\mathcal{K}) - j(\mathcal{K}), \text{ where} \\
&q(\mathcal{K}) = \sum_{x,m} \frac{\ln(2K_{x,m}/A)}{2K_{x,m}/A} + \sum_{m} \frac{\ln\left(2\bar{K}_m\right)}{2\bar{K}_m}, \\
&j(\mathcal{K}) = \min_{m} \left(\min_{x} \ln\left(\frac{2K_{x,m}}{A}\right), \ln\left(\frac{\bar{K}_m}{\tau^2_{\max}}\right)\right) \left(A \sum_{x,m} \frac{K^{-1}_{x,m}}{2} + \sum_{m} \frac{K^{-1}_{m}}{2} \right)
\end{aligned} \tag{2.63}$$

In (2.63), $K_m = \sum_x K_{x,m}$, $\bar{K} = K/\tau^2_{\max}$ and $\widetilde{K} = K^{-1}$. Analogous to (2.54) in assumption (F2) for finite sample complexity of IRL for optimal stopping, (2.63) imposes a lower bound on the number of samples needed for our sample complexity result of inverse SHT. Eq. 2.63 is a sufficient condition for obtaining the constants of the sample complexity bound as the solution of a convex optimization problem. $\varepsilon_1(\cdot), \varepsilon_2(\cdot)$ are the minimum perturbations needed for the finite dataset to satisfy and not satisfy, respectively, the linear NIAS and NIAC* inequalities in Theorem 6, and defined formally in 2.65). The quantities $q(\cdot), j(\cdot)$ are decreasing functions of the sample size $\mathcal{K}$. For the reader's convenience, we discuss the assumptions (F3)-(F5) after the finite sample complexity result, Theorem 13. Analogous to Definition 10, the statistical detection test for the inverse SHT problem is defined below. It takes in as input a finite (noisy) dataset and outputs one of the two hypotheses- H_0 (agent is an optimal SHT agent) or H_1 (agent is not an optimal SHT agent).

Definition 12 (IRL decision test for inverse SHT) *Consider the inverse learner with dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ (2.62). Assume (A2) (A4), (A5) and (F3) hold. The IRL detector $\mathrm{Test}_{\mathrm{IRL}}(\cdot)$ for the inverse SHT problem is given by:*

$$\mathrm{Test}_{\mathrm{IRL}}(\widehat{\mathcal{D}}_M(\mathcal{K})) = \begin{cases} H_0, & \textit{if } \mathrm{IRL}_{\mathrm{SHT}}(\widehat{\mathcal{D}}_M(\mathcal{K})) \neq \emptyset \\ H_1, & \textit{if } \mathrm{IRL}_{\mathrm{SHT}}(\widehat{\mathcal{D}}_M(\mathcal{K})) = \emptyset. \end{cases} \tag{2.64}$$

Here, $\mathrm{IRL}_{\mathrm{SHT}}(\mathcal{D})$ *is the set of feasible solutions to the linear NIAS and NIAC** *inequalities (2.16), (2.31) in Theorem 6 given dataset* $\mathcal{D}$*.*

Main Result. Finite Sample analysis for inverse SHT

We now present our finite sample result for IRL of the inverse SHT problem. It provides bounds for the Type-I/II error probabilities of the IRL detector (2.64) in terms of the sample size of $\widehat{\mathcal{D}}_M(\mathcal{K})$ (2.62).

Notation. Theorem 13 below uses the following variables:

$$\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K})): \min_{\{\hat{p}'_m, \hat{C}'_m\}} \sum_m \|\hat{p}_m - \hat{p}'_m\|_2^2 + (\hat{C}_m - \hat{C}'_m)^2,\ \mathrm{IRL}_{\mathrm{SHT}}(\{\pi_0, \{\hat{p}'_m, C'_m\}\}) \neq \emptyset$$

$$\varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K})): \min_{\{\hat{p}'_m, \hat{C}'_m\}} \sum_m \|\hat{p}_m - \hat{p}'_m\|_2^2 + (\hat{C}_m - \hat{C}'_m)^2,\ \mathrm{IRL}_{\mathrm{SHT}}(\{\pi_0, \{\hat{p}'_m, C'_m\}\}) = \emptyset$$

$$i(\widehat{\mathcal{D}}_M(\mathcal{K})) = \mathcal{K}_H(\mathrm{SHT}) \cdot h(\widehat{\mathcal{D}}_M(\mathcal{K})),\ \text{where } \mathcal{K}_H(\mathrm{SHT}) = A \sum_{x,m} K_{x,m}^{-1} + \tau_{\max}^2 \sum_m K_m^{-1}, \text{ and}$$

$$h(\widehat{\mathcal{D}}_M(\mathcal{K})) = \prod_m \left(\left(2\bar{K}_m\right)^{\bar{K}_m^{-1}} \prod_x \left(\frac{2\bar{K}_{x,m}}{A} \right)^{A\widetilde{K}_{x,m}} \right)^{\mathcal{K}_H^{-1}(\mathrm{SHT})}. \tag{2.65}$$

In (2.65), $K_m = \sum_x K_{x,m}$, $\bar{K} = K/\tau_{\max}^2$ and $\widetilde{K} = K^{-1}$. Analogous to the finite sample result for inverse optimal stopping, $\varepsilon_1(\cdot)$, $\varepsilon_2(\cdot)$ defined above are the minimum perturbations needed for the finite SHT dataset to satisfy and not satisfy, respectively, the NIAS and NIAC* inequalities in Theorem 6. Compared to the minimum perturbations defined in (2.57) and (2.58) for inverse optimal stopping, the key distinction is that $\varepsilon_1(\cdot)$ and $\varepsilon_2(\cdot)$ in (2.65) also involve perturbations in the expected continue cost of the agent. The variable i in (2.65) is the error constant for the finite sample error bounds for inverse SHT; variable $\mathcal{K}_H(\mathrm{SHT})$ can be interpreted as a weighted harmonic mean of the recorded trials $\mathcal{K}$ (2.62).

Theorem 13 (Sample complexity for inverse SHT) *Consider an inverse learner with*

dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ (2.53) detecting if the agent acting in multiple environments $\mathcal{M}$ is an optimal SHT agent using the statistical test in Definition 12. Assume (F3)-(F5) hold. Then, the Type-I and Type-II error probabilities of the IRL detector (Definition 12) are bounded as:

$$\textit{Type-I error probability} \leq i(\widehat{\mathcal{D}}_M(\mathcal{K}))\, \exp\left(-2\, \mathcal{K}_H(\mathrm{SHT}) \cdot \varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K}))\right), \quad (2.66)$$

$$\textit{Type-II error probability} \leq i(\widehat{\mathcal{D}}_M(\mathcal{K}))\, \exp\left(-2\, \mathcal{K}_H(\mathrm{SHT}) \cdot \varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K}))\right). \quad (2.67)$$

The proof of Theorem 13 is in Sec. 2.12. Theorem 13 characterizes the robustness of the linear feasibility test in Theorem 6 to finite sample constraints. Compared to Theorem 11, the finite sample result of Theorem 15 requires accounting for the empirical estimate of the agent's expected continue cost. Hence, in addition to the DKW inequality, the proof of Theorem 13 uses Hoeffding's inequality to bound the empirical estimation error for the expected continue cost.

Discussion of Assumptions.

(F3): (F3) specifies the inverse learner's dataset for the inverse SHT problem computed from a finite number of trials.

(F4) says the inverse learner knows the upper bound of the agent's stopping times over all environments. This assumption is crucial for our main result since the Hoeffding's inequality (for bounding the finite sample effect of the expected stopping time) requires this knowledge.

(F5): The condition (2.63) in (F5) is analogous to assumption (F2) for the finite sample result for IRL of optimal stopping. (F5) admits a close form expression for the error bounds in Theorem 13. Abstractly, (F5) says that the number of samples recorded by the inverse learner is sufficiently large so that the condition (2.63) is satisfied.

2.6.4 Example 2. Finite sample effects for IRL in inverse search

We now analyze the finite sample effect of IRL for inverse search. Recall from Theorem 9 that optimal search is equivalent to feasibility of the linear NIAC† inequalities. Our main result below, namely, Theorem 15, characterizes the robustness of the feasibility test (wrt the NIAC† inequality) for detecting optimal search under finite sample constraints. It turns out that Theorem 15 is a special case of Theorem 11, our finite sample complexity result for IRL of inverse optimal stopping.

Main assumptions and detection test. Suppose the inverse learner observes the actions of a Bayesian stopping agent. We assume the following about the inverse learner:

(F6) Instead of (A6), the inverse learner uses the *finite dataset*

$$\widehat{\mathcal{D}}_M(\mathcal{K}) = \{\pi_0, \{\hat{g}_m(a,x), m \in \mathcal{M}\}\} \tag{2.68}$$

to detect if the agent performs optimal search or not. $\hat{g}_m(a,x)$ is the empirical search action policy of the agent defined in (2.47) and $\mathcal{K} = \{K_{x,m}, x \in \mathcal{X}, m \in \mathcal{M}\}$ denotes the number of trials recorded by the inverse learner in state x, where m indexes the search environment.

(F7) The prior belief of the targets π_0 is a uniform prior, i.e., $\pi_0(x) = 1/X$. Also, the reveal probability $\alpha(a)$ is the same for all actions $a \in \mathcal{A}$, i.e., $\alpha(a) = \alpha$. Although the variable α is unknown to the inverse learner, it satisfies the following inequality.

$$\alpha \geq \max\left\{\alpha^*, 1 - \frac{\min_{a\in\mathcal{A}} l_m(a)}{\max_{a\in\mathcal{A}} l_m(a)},\ \forall m \in \mathcal{M}\right\} \tag{2.69}$$

For the reader's convenience, we discuss the assumptions (F6) and (F7) after the finite sample complexity result, Theorem 15. We now define the statistical detection test for the inverse search problem. It takes in as input the finite (noisy) dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ (2.68)

and detects one of the two hypotheses- H_0 (agent performs optimal search) or H_1 (agent does *not* perform optimal search).

Definition 14 (IRL decision test for inverse search) *Consider the inverse learner with dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ (2.68). Assume (A7), (F6) holds. The IRL detector $\mathrm{Test}_{\mathrm{IRL}}(\cdot)$ for the inverse search problem is given by:*

$$\mathrm{Test}_{\mathrm{IRL}}(\widehat{\mathcal{D}}_M(\mathcal{K})) = \begin{cases} H_0, & \textit{if } \mathrm{IRL}_{\mathrm{Search}}(\widehat{\mathcal{D}}_M(\mathcal{K})) \neq \emptyset \\ H_1, & \textit{if } \mathrm{IRL}_{\mathrm{Search}}(\widehat{\mathcal{D}}_M(\mathcal{K})) = \emptyset. \end{cases} \tag{2.70}$$

Here, $\mathrm{IRL}_{\mathrm{Search}}(\mathcal{D})$ is the set of feasible solutions to the linear NIAC† inequalities (2.46) in Theorem 9 given dataset $\mathcal{D}$.

Main Result. Finite Sample Result for Inverse Search.

We now present Theorem 15, our finite sample result for IRL of the inverse search problem. Theorem 15 provides bounds for the Type-I/II and posterior Type-I/II error probabilities of the IRL detector in Definition 14 in terms of the sample size of the finite search dataset and uses the following variables.

$$\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K})) = \min_{\{\hat{g}'_m, m\in\mathcal{M}\}} \sum_m \|\hat{g}_m - \hat{g}'_m\|_2^2, \ \mathrm{IRL}_{\mathrm{Search}}(\{\pi_0, \hat{g}'_m(a,x), m \in \mathcal{M}\}) \neq \emptyset.$$

$$\varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K})) = \min_{\{\hat{g}'_m, m\in\mathcal{M}\}} \sum_m \|\hat{g}_m - \hat{g}'_m\|_2^2, \ \mathrm{IRL}_{\mathrm{Search}}(\{\pi_0, \hat{g}'_m(a,x), m \in \mathcal{M}\}) = \emptyset.$$

The variables $\varepsilon_1(\cdot),\ \varepsilon_2(\cdot)$ are the minimum perturbations needed for the finite search dataset to satisfy and not satisfy, respectively, the linear NIAC† inequalities (2.46) in Theorem 9.

Theorem 15 (Sample complexity for inverse search) *Consider an inverse learner with dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ (2.53) detecting if a Bayesian stopping agent is performing optimal search by using the statistical test in Definition 14. Assume (F6) and (F7) hold. Then, the*

Type-I and Type-II error probabilities for the IRL detector (Definition 14) are bounded as:

$$\textit{Type-I error probability} \leq \frac{(1-\alpha^*)A}{\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K}))(\alpha^*)^2}\left(\sum_{x,m} K_{x,m}^{-1/2}\right)^2, \tag{2.71}$$

$$\textit{Type-II error probability} \leq \frac{(1-\alpha^*)A}{\varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K}))(\alpha^*)^2}\left(\sum_{x,m} K_{x,m}^{-1/2}\right)^2. \tag{2.72}$$

The proof of Theorem 15 is in Appendix 2.11. Theorem 15 characterizes the robustness of the linear feasibility test in Theorem 9 to finite sample constraints. It upper bounds the probability of incorrectly detecting the Bayesian agent as an optimal search agent or not an optimal search agent, in terms of the number of trials recorded by the inverse learner.

Discussion of assumptions.

(F6): Assumption (F6) specifies the inverse learner's dataset for the inverse search problem computed from a finite number of trials.

(F7): (F7) says that the agent has the same reveal probability for all locations for all environments and the inverse learner knows this reveal probability is greater than a certain value. This assumption can be viewed as an analogy of having the same instantaneous continue cost for the agent solving the SHT problem. The condition (2.69) results in the optimal search strategy of the agent to be periodic for all environments - the agent searches each location exactly once in a particular order (depends on the agent's search costs) and repeats this cycle till the target is located. This allows the search action policy $g_m(a, x)$ to be written in terms of the conditional pdf of the stopping time $\mathbb{P}_{\mu_m}(\tau|x)$ (see Sec. 2.13). By analyzing the finite sample effects of the stopping time due to the added structure, Theorem 15 results. Note that Theorem 15 does not require the inverse learner to have information about the true stopping time of the agent, but only the empirical search action policy.

Summary

For finite sample observations, this section presented an IRL detector for optimality of a Bayesian stopping agent (Definition 10) and provided error bounds of the detector (Theorems 11). We also presented finite sample IRL detectors for optimality in SHT and Bayesian search (Definitions 12, 14) and obtained error bounds of the detector in terms of the sample size (Theorem 13, 15). The key idea behind the sample complexity results is the construction of a probabilistic bound on the minimum perturbation needed to satisfy and not satisfy, respectively, the feasibility inequalities for optimality to bound the Type-I and Type-II error probabilities of the IRL detector, respectively.

2.7 Discussion and extensions

This chapter has proposed Bayesian revealed preference methods for inverse reinforcement learning (IRL) in partially observed environments. Specifically, we considered IRL for a Bayesian agent performing multi-horizon sequential stopping. The results in this chapter achieve IRL under the following restrictions on the inverse learner: (1) The inverse learner does not know if the decision maker is an optimal Bayesian stopping agent (2) The inverse learner does not know the agent's observation likelihood and (3) IRL for noisy datasets. Our IRL algorithms first identify if the agent is behaving in an optimal manner, and if so, estimate their stopping costs. The inverse learner can at best identify optimality of an agent's strategy wrt to its strategies chosen in other environments, a notion intuitively explained in the introduction and defined formally in Lemma 2. To illustrate our IRL approach, we considered two examples of sequential stopping problems, namely, sequential hypothesis testing (SHT) and Bayesian search and provided algorithms to estimate their misclassification/search costs.

Our main results were:

1. Specifying necessary and sufficient conditions for the decisions taken by a Bayesian agent in multiple environments to be identified as optimal sequential stopping and generating set-valued estimates of their stop costs (Theorem 3). To the best of our knowledge, our IRL results for Bayesian stopping time problems when the inverse learner has no knowledge of the agent's dynamics is novel.
2. Constructing convex feasibility based IRL algorithms for set-valued estimation of misclassification for an SHT agent (Theorem 6) and search costs for a search agent (Theorem 9) when decisions from infinite trials of the agent are available in multiple SHT and search environments, respectively. These results are special cases of Theorem 3 due to additional structure of the SHT and search problem compared to generic sequential stopping problems.
3. Proposing IRL detection tests for detecting optimality of sequential stopping, SHT and search when only a finite number of agent decisions are observed.
4. Providing sample complexity bounds on the Type-I/II and posterior Type-I/II error probabilities of the above detection tests under finite sample constraints (Theorems 11, 13 and 15).

Extensions

This chapter identifies optimal stopping behavior in a Bayesian agent by observing their actions without external interference. A natural extension is to consider the controlled IRL setting where the inverse learner is an *active* entity that can influence/control the actions of the agent. This leads to the question: *How to influence the agent's actions so as to better identify optimal stopping behavior and estimate the agent costs more efficiently?*

Another question is: *How to formulate conditions under which the set-valued cost estimates for an agent in finitely many environments tends to a point-valued estimate as the number of environments tend to infinity?* In classical revealed preference theory, the chapters [41, 42] characterize properties of utility functions that rationalize infinite datasets. It is worthwhile generalizing these results to a Bayesian IRL setting.

Recent advances in deep IRL [43, 44] use deep neural networks as function approximators for the underlying feature space. Our current research aims to extend the results in this chapter to deep-IRL for inverse optimal stopping where the inverse learner does not know the underlying state space and relies on neural networks for feature space approximation.

The IRL methodology of the chapter assumes the analyst has no knowledge of the agent's observation likelihood. If the inverse learner knows *a priori* that the agent must choose its observation likelihood from a finite set, the inverse learner cannot rely on NIAS and NIAC in Theorem 3 for checking optimal Bayesian stopping. Instead, one must adapt adaptive search techniques for identifying the optimal observation likelihood that is (a) consistent with the inverse learner's dataset and (b) optimizes the agent's objective. If the agent's observation likelihood is known to be multi-variate Gaussian, then one can use the tree search approach that has seen success in applications such as adversarial tracking [45] and motion planning [46]. Extending the IRL results in this chapter to tree-based adaptive search techniques is a subject of current research.

Finally, it is worthwhile exploring IRL for stopping time problems using the iterative update approach of [7] and the MCMC based sampling approach of [28].

2.8 Appendix

2.8.1 Context and Perspective. IRL for Bayesian stopping problems

2.8.2 Literature and Applications. IRL in Bayesian stopping problems

IRL methods have been successfully applied to areas like robotics [47], user engagement in multimedia social networks such as YouTube [12], autonomous navigation by [7, 28] and [48] and inverse cognitive radar [16, 49]. Below we discuss several real-world examples where an analyst aggregates data from a Bayesian stopping time agent, and has no knowledge of the agent's observation likelihood for solving the IRL problem.

- *Consumer Insights and Ad Design Research:* Online multimedia are sequential Bayesian decision makers [50, 51]; they accumulate evidence sequentially from audio-visual cues on the screen and then take an action (for example, playing a video, clicking on an ad etc.). In advertisement design, an analyst observes how an online user (stopping time agent) reacts to a pop-up advertisement in multiple environments, where the environment is characterized by the current web-page, content and position of the ad etc. In consumer research for online movie platforms, the analyst observes whether an online user clicks on a movie thumbnail or not in multiple scenarios, where the scenario depends on factors like user's past history of movies watched and neighboring movie thumbnails. The decision process of the online user (forward learner) in both these examples can be embedded into an SHT framework, where the sensing cost is the cost of attention to visual cues and the stopping (terminal) cost measures the online user's preferences for viewing

the advertisement/movie. By characterizing the content reactivity of online users in different multimedia platforms, IRL for stopping time problems is useful for targeted ad-design and content recommendation.

- *Electronic counter-countermeasures in electronic Warfare:* Sequential Bayesian jamming models are extensively used in Electronic Counter Measures (ECM) for mitigating radar systems; see [52, 53] and [54] for details. The proposed IRL algorithms can be used for Electronic Counter Counter Measures (ECCM) by the radar system to reverse engineer the adversarial ECM algorithms and avoid performance mitigation, hence extending the chapter [49] to the Bayesian case. For instance, suppose an adversarial radar uses Bayesian search to identify valuable targets like in [55]. Using IRL for inverse search, a radar analyst can use the estimated search costs of the adversarial radar for effectively designing the targets to avoid being easily detected. [56, 57] develop inverse optimal control (IOC) based IRL algorithms for reconstructing adversary intent for tracking control. Our work complements [57] since it allows one to still achieve IRL without knowledge of model dynamics, as is common to assume in the literature.
- *Interpretable ML for Smart Healthcare:* Recently, sequential Bayesian models for assisting medical diagnoses have been aggressively used in smart healthcare like in [58–61] and [62]. These trained models are usually only accessible in an abstracted black-box format in an executable software application. *Our IRL algorithm provides an interpretable Bayesian decision model for these assistive algorithms.* Interpretability in AI-enabled healthcare [63] facilitates informed decisions for the debugging and improvement of assistive diagnoses.

2.8.3 Related works in IRL

We now summarize the key IRL works in the literature and compare them to our chapter.

(a) IRL in fully observed environments: Traditional IRL [6, 7] aims to estimate an unknown deterministic reward function of an agent by observing the optimal actions of the agent in a Markov decision process (MDP) setting. The key assumption is the existence of an optimal policy. Our convex feasibility approach for IRL in stopping time problems can be viewed as a generalization of the feasibility inequalities in [6, Theorem 3]. [6, Theorem 3] compute a feasible set of rewards that ensure the agent's policy outperforms all other policies. Since the set of policies for an MDP is finite, [6, Theorem 3] comprises a finite set of linear inequalities. In comparison, the set of policies for a partially observed MDP (POMDP) is infinite. From the feasible set of rewards, [6, 26] choose the max-margin reward, i.e., the reward that maximizes the regularized sum of differences between the performance of the observed policy and all other policies. In Sec. 2.3.4, we compute a regularized max-margin estimate of costs for inverse SHT and plot the reconstruction error. [7] achieve IRL by devising iterative algorithms for estimating the agent's reward function. Abstractly, the key idea is to terminate the iterative process once the value function of the rewards converges to an ϵ interval.

[28] use the principle of maximum entropy for achieving IRL of an optimal agent, wherein the agent's policy is subject to a Shannon mutual information regularization. This regularization facilitates expressing the optimal policy in closed form; the optimal policy turns out to be softmax in terms of the Q-function of the MDP. [64] extend [28] to a more general regularization setup that also admits a closed form solution to the optimal policy in terms of strongly convex functions for regularization, for examples, the Tsallis entropy [65] that generalizes Shannon entropy. Solving the IRL task with zero dynamics knowledge has also been explored in the literature. [66] append the IRL task

with simultaneous learning of model dynamics, specifically, the agent's transition kernel. The key idea in the approach is to maximize the log-likelihood of sampled trajectories wrt the appended parameter space that parametrizes the agent's rewards and transition kernel. Our problem setting differs from [66] in that we operate in the non-parametric partially observed setting regime where the observation likelihood of the agent is unknown and not necessarily parametrizable. Indeed, our results can be specialized to parameter families of observation likelihood known to the analyst, and is a subject of current research.

[67] generalize IRL to continuous space processes and circumvent the problem of finding the optimal policy for candidate reward functions. Recently [48, 68, 69] and [70] used deep neural networks for IRL to estimate agent rewards parametrized by complicated non-linear functions. [71] achieve IRL when the agent's rewards are sampled from a prior distribution and the demonstrator's trajectories update the posterior belief of the reward. Building on the seminal work of [8], [13, 72] study identifiability of parameters for structure MDPs in IRL. In analogy, in this chapter, we provide identifiability conditions for a subset of POMDPs, namely, Bayesian stopping problems.

(b) IRL in partially observed environments: The influential works of [1, 73] and [74] are the first works on IRL in a POMDP setting. They extend traditional IRL [6, 7] to an infinite state space (space of posterior beliefs of the agent). [74] extend Bayesian IRL [71] for MDPs to the POMDP setting. In analogy to Bayesian IRL, the aim is to compute the posterior distribution of reward functions given an observation dataset. The assumption of a softmax action policy suffices to compute the likelihood of the observation dataset given a reward function, and hence, bypasses the need to compare the agent's performance with respect to other candidate policies.

Since our work is closely related to [1, 73], we briefly review their approach. In [1, 73], the inverse learner first checks if the agent chooses the optimal action given a particular

posterior belief, for *finitely many beliefs* aggregated from the observed trajectories of belief-action pairs. This is analogous to our NIAS condition (2.16) in Theorem 6, where we check if the agent's terminal action is optimal given its terminal belief. Next, the inverse learner check if the agent's policy is optimal with respect to a *finite set of policies* that deviate from the observed policy by a single step. This resembles our NIAC condition (2.31) in Theorem 3 where we check for optimality of the Bayesian decision maker's actions in multiple environments.

As [1, 73] mention, this approach to checking for optimality only gives rise to a necessary condition, and not a necessary and sufficient condition like in [6, 7], where the number of policies are finite. In other words, without prior information about the nature of the Q-function given a policy, it is impossible to check for global optimality, that is, find a reward function that outperforms *all* other policies (infinitely many policies).

Our Bayesian revealed preference based approach is *complementary* to [1]. While [1] develop IRL methods for POMDPs with no assumption on problem structure, we consider a subset of POMDPs, namely, Bayesian stopping time problems. Due to the structure of stopping time problems, we show that our IRL algorithms *do not* require knowledge of the observation likelihood of the decision maker, nor require solving a POMDP. Indeed, IRL for generic POMDPs is non-identifiable if the inverse learner does not know the model dynamics, nor can solve a POMDP. To test for optimality, our IRL algorithms rely on the decision maker's strategies from multiple environments, where every environment differs in the terminal cost. Decision strategy in multiple environments can be viewed as a surrogate for performance wrt different policies. To summarize, our work builds on the seminal work of [1] with the key discerning features of our IRL methodology being: (1) Unobservability of agent dynamics, (2) No assumptions on decision optimality, and (3) IRL generalization for empirical (noisy) datasets with performance guarantees via finite sample complexity.

(c) Inverse Rational Control (IRC). IRC [75] is a closely related field to IRL in partially observed environments. IRC models sub-optimality in decision makers as a misspecified reward function and aims to estimate this reward. The IRC task comprises two sub-tasks:

First, the inverse learner constructs a map from a continuous space of reward functions parameterized by θ to the reward's optimal policy.

Second, based on a finite observation dataset $\mathcal{D}$, the underlying hyperparameter θ is estimated as the maximum likelihood estimate $\operatorname{argmax}_\theta \mathbb{P}(\mathcal{D}|\theta)$.

In comparison, our approach bypasses the first sub-task in IRC by checking the feasibility of a finite set of convex inequalities. Given the information available to the inverse learner, these inequalities are both necessary and sufficient conditions for identifying optimality of a decision maker's decisions in multiple environments. Indeed, increasing the number of environments in which the decision maker's actions are observed decreases the size of the feasible set of rationalizing rewards, and hence increases the precision of our set-valued IRL cost estimate.[17]

(d) Revealed Preference. The key formalism used in this chapter to achieve IRL is *Bayesian revealed preferences* studied in microeconomics by [2, 14, 15, 29]. Non-parametric estimation of cost functions given a finite length time series of decisions and budget constraints is the central theme in the area of classical (non-Bayesian) revealed preferences in microeconomics, starting with [3, 76] where necessary and sufficient conditions for constrained utility maximization are given; see also [32, 77, 78] and more recently in machine learning [79].

(e) Examples of Bayesian stopping time problems. After constructing an IRL frame-

[17] Revealed preference micro-economics [41, 42] have studied the consistency of the set-valued approach to eliciting agent rewards. [42] specifies conditions under which the feasible set of utility functions reconstructed from a dataset of agent actions converges to a point for infinite datasets. [41] constructs a quasi-concave utility function that rationalizes an infinite dataset.

work for general stopping time problems, this chapter discusses two important examples, namely, inverse sequential hypothesis testing and inverse Bayesian search. Below we briefly motivate these examples.

Example 1. Inverse Sequential Hypothesis Testing (SHT). Sequential hypothesis testing (SHT) [80, 81] is widely studied in detection theory. The inverse SHT problem of estimating misclassification costs by observing the decisions of an SHT agent has not been addressed. Estimating SHT misclassification costs is useful in adversarial inference problems. For example, by observing the actions of an adversary, an inverse learner can estimate the adversary's utility and predict its future decisions.

Example 2. Inverse Bayesian Search. In Bayesian search, each agent sequentially searches locations until a stationary (non-moving) target is found. Bayesian search [81] is used in vehicular tracking [82], image processing [83] and cognitive radars [84]. IRL for Bayesian search requires the inverse learner to estimate the search costs by observing the search actions taken by a Bayesian search agent in multiple environments with different search costs.

Bayesian search is a special case of the Bayesian multi-armed bandit problem [85, 86]. A promising extension of our IRL approach would be to solve inverse Bayesian bandit problems, namely, estimate the Gittins indices of the bandit arms. Regarding the literature in inverse bandits, [87] propose a real-time assistive procedure for a human performing a bandit task based on the history of actions taken by the human. [88] solve the inverse bandit problem by assuming the inverse learner knows the variance of the stochastic reward; in comparison out setup assumes no knowledge of the rewards.

2.9 Proof of Lemma 2

Proof. Suppose $\mathcal{D}_M$ is generated by a Bayesian agent performing optimal stopping (Definition 1) in M environments. By definition, the following conditions hold:

$$\mu_m(\pi, \tau) = \operatorname*{argmin}_{a \in \mathcal{A}} \pi' \bar{s}_{m,a}, \; J(\mu_m, s_m) = \inf_{\mu \in \boldsymbol{\mu}} J(\mu, s_m), \tag{2.73}$$

where $J(\cdot)$ (2.10) is the expected cumulative cost comprising the expected stopping cost $G(\cdot)$ and expected cumulative continue cost $C(\cdot)$ Since the set of chosen strategies $\boldsymbol{\mu}_{\mathcal{M}} \subset \boldsymbol{\mu}$, the set of *all* admissible policies, the feasibility of (2.73) implies the following conditions hold:

$$\begin{aligned} \mu_m(\pi, \tau) &= \operatorname*{argmin}_{a \in \mathcal{A}} \pi' \bar{s}_{m,a}, \\ J(\mu_m, s_m) &= \min_{\mu \in \boldsymbol{\mu}_{\mathcal{M}}} J(\mu, s_m). \end{aligned} \tag{2.74}$$

Since $\boldsymbol{\mu}_{\mathcal{M}}$ is finite, the 'inf' in (2.73) can be replaced with 'min' in (2.74). The second condition in (2.74) is simply a reformulation of (2.12). Hence, the 'if' statement of Lemma 2 is proved.

We now prove the 'only if' direction. Suppose the inverse learner has access to $\mathcal{D}_M$ aggregated from a Bayesian agent's actions in M environments. Specifically, the inverse learner only knows the agent's incurred expected costs finitely many policies $\mu_m \in \boldsymbol{\mu}_{\mathcal{M}}$. Alternatively, the sole knowledge of $\mathcal{D}_M$ implies that the inverse learner *does not know* the agent's expected stopping cost, nor the expected cumulative continue cost if the agent chooses any policy $\mu \in \boldsymbol{\mu} \backslash \boldsymbol{\mu}_{\mathcal{M}}$. Condition (2.8) is independent of the agent's policy, and only depends on the agent's stopping belief. However, (2.10) requires the inverse learner to compare the expected cost of the agent's strategy μ_m in environment m against infinitely many strategies $\mu \in \boldsymbol{\mu}$. Due to inverse learner's limited knowledge, the best the inverse learner can do to check if (2.10) holds is to check the feasibility of (2.74). ■

2.10 Proof of Theorem 3

We first introduce an observation likelihood α_m over a fictitious observation space $\mathcal{Y}_\pi$ with generic element $\tilde{y}_\pi$ for stopping strategy $\mu_m, m \in \mathcal{M}$:

$$\alpha_m(\tilde{y}_\pi|x) = \sum_{\bar{y}:\pi_\tau=\pi} \left(\prod_{t=1}^{\tau} B(y_t, x) \right) \tag{2.75}$$

Here $\bar{y}$ denotes a sequence of observations $y_1, y_2, \ldots$ and τ is the random stopping time for strategy μ_m defined in (2.6). $\alpha_m(\tilde{y}_\pi|x)$ is the likelihood of all observation sequences $\bar{y}$ such that given true state $x^o = x$ and stopping strategy μ_m, the agent's belief state at the stopping time τ is π. Equivalently, $\alpha_m(\tilde{y}_\pi|x)$ is the conditional probability density of the agent's stopping belief for strategy μ_m. By definition, the mapping from $\tilde{y}_\pi$ to stopping belief π is one-to-one. Hence, $|\mathcal{Y}_\pi| = |\Delta(\mathcal{X})|$, where $\Delta(\mathcal{X})$ denotes the $X-1$ dimensional unit simplex of pmfs.

Next, we re-formulate the expected stopping cost $G(\mu_m, s_m)$ defined in (2.10) for stopping cost s_m in terms of the fictitious observation likelihood defined in (2.75).

$$G(\mu_m, s_m) = \mathbb{E}_{\mu_m}\left\{\pi_\tau' \bar{s}_{m,a_\tau}\right\} = \int_{\mathcal{Y}_\pi} \underbrace{\left(\sum_{x\in\mathcal{X}} \alpha_m(\tilde{y}_\pi|x)\pi_0(x) \right)}_{\text{Marginal distribution of } \tilde{y}_\pi} \min_{a\in\mathcal{A}} \pi' \bar{s}_{m,a}\, d\tilde{y}_\pi \tag{2.76}$$

In the above equation, the summation within the parentheses is the unconditional probability density of the stopping belief π given stopping strategy μ_m. Also, as described above, $\alpha_m(\tilde{y}_\pi|x)$ is the likelihood of all observation sequences $\bar{y}$ such that given true state $x^o = x$ and stopping strategy μ_m, the agent's belief state at the stopping time τ is π. We are now ready to prove necessity and sufficiency of the NIAS, NIAC inequalities (2.13), (2.14) in Theorem 3 for identifying an optimal stopping agent (Lemma 2).

2.10.1 Necessity of NIAS, NIAC inequalities

Recall from Theorem 3 that the analyst knows the agent's action selection policy in multiple environments. The action selection policy $p_m(a|x)$ in $\mathcal{D}_M$ is a stochastically garbled version of $\alpha_m(\tilde{y}_\pi|x)$ defined in (2.75) and $p_m(x|a)$ is a stochastic garbling of the agent's stopping belief π when the stop action is a. The action selection policy can be rewritten as follows

$$p_m(a|x) = \int_{\mathcal{Y}_\pi} p_m(a|\tilde{y}_\pi)\, \alpha_m(\tilde{y}_\pi|x)\, d\tilde{y}_\pi \tag{2.77}$$

$$\implies p_m(x|a) = \int_{\mathcal{Y}_\pi} \frac{p_m(a|\tilde{y}_\pi)\alpha_m(\tilde{y}_\pi|x)\pi_0(x)}{p_m(a)}\, d\tilde{y}_\pi = \int_{\mathcal{Y}_\pi} p_m(\tilde{y}_\pi|a)\, \pi(x)\, d\tilde{y}_\pi, \tag{2.78}$$

where π is the agent's stopping belief and $p_m(\tilde{y}_\pi|a)$ is the probability density of the fictitious observations $\tilde{y}_\pi$ conditioned on the stop action a.

NIAS

Let action a be the optimal stop action (2.8) for stopping belief π of the m^{th} agent in $\mathcal{A}$. Then,

$$\sum_{x\in\mathcal{X}} \pi(x)(s_m(x,a) - s_m(x,b)) \leq 0,\ \forall a, b \in \mathcal{A} \tag{2.79}$$

$$\implies \int_{\mathcal{Y}_\pi} \sum_{x\in\mathcal{X}} \pi(x)(s_m(x,a) - s_m(x,b))\, p_m(\tilde{y}_\pi|a) d\tilde{y}_\pi \leq 0 \tag{2.80}$$

$$\implies \sum_{x\in\mathcal{X}} \left(\int_{\mathcal{Y}_\pi} p_m(\tilde{y}_\pi|a)\, \pi(x)\, d\tilde{y}_\pi \right) (s_m(x,a) - s_m(x,b)) \leq 0 \tag{2.81}$$

$$\implies \boxed{\sum_{x\in\mathcal{X}} p_m(x|a)(s_m(x,a) - s_m(x,b)) \leq 0.} \tag{2.82}$$

Eq. 2.79 says that the expected stop cost given belief π is minimum for stop action a. Here, the expectation is taken over the finite state set $\mathcal{X}$. The LHS of (2.80) is the expected value of the LHS of (2.79) taken over the space of fictitious observations $\mathcal{Y}_\pi$

wrt the probability density $p_m(\tilde{y}_\pi|a)$. Since $|s_m(x,a) - s_n(x,a)|$ is bounded, the integral on the LHS of (2.80) is finite. Hence, by Fubini's theorem, we can change the order of summation to get (2.81). The first term in the integral of (2.81) is equal to $p_m(x|a)$ from (2.78) which results in the final NIAS inequality (2.82).

NIAC

Define $\tilde{G}_{m,n}$ as expected stopping cost when the fictitious observation likelihood is $p_m(a|x)$ and stopping cost is $s_n(x,a)$. Then:

$$\tilde{G}_{n,m} = \sum_{a\in\mathcal{A}} \left(\sum_{x\in\mathcal{X}} p_n(a|x)\pi_0(x) \right) \min_{b\in\mathcal{A}} \sum_{x\in\mathcal{X}} p_m(x|a) s_m(x,b). \tag{2.83}$$

It follows from Blackwell dominance [89] that:

$$\tilde{G}_{n,m} \geq G(\mu_n, s_m) \text{ for all } m,\ n, \tag{2.84}$$

since the kernel $p_m(y_{1:\tau(\mu_m)}|x)$ Blackwell dominates the action selection policy $p_m(a|x)$. A key observation is that, in (2.84), equality holds for $m = n$ and is straightforward to show using Jensen's inequality. To summarize, we have the following inequality:

$$G_{m,m} = \tilde{G}_{m,m},\ G_{n,m} \leq \tilde{G}_{n,m} \tag{2.85}$$

For any set of environment indices $\{m_1, m_2, \ldots m_I\} \subset \mathcal{M},\ (m_{I+1} = m_1)$, we have the following inequality from (2.12) in Lemma 2:

$$\sum_{i=1}^{I} G(\mu_{m_{i+1}}, s_{m_i}) - G(\mu_{m_i}, s_{m_i}) \geq \sum_{i=1}^{I} C(\mu_{m_i}) - C(\mu_{m_{i+1}}) = 0$$

Combining the inequalities (2.84) with the above inequality, we get the NIAC inequality:

$$\sum_{i=1}^{I} \tilde{G}_{m_{i+1},m_i} - \tilde{G}_{m_i,m_i} \geq 0 \tag{2.86}$$

2.10.2 Sufficiency of NIAS, NIAC inequalities for Bayes optimal stopping

The inverse learner only has access to the agent's prior π_0 over the state space $\mathcal{X}$ and action selection policy $p_m(a|x)$ induced by the agent's policy in environment m. For a finite set of fictitious observations, the sufficiency proof assumes that there exists a one-to-one correspondence between the fictitious observation $\tilde{y}_\pi$ to the terminal action a. If the observation space $\mathcal{Y}$ is continuous-valued, the sufficiency proof assumes there exist disjoint subsets $\mathcal{Y}_\pi(a) \subset \mathcal{Y}_\pi$ and pdfs f_a with support $\mathcal{Y}_\pi(a)$ such that the fictitious observation likelihood $\alpha_m(\tilde{y}_\pi|x)$ can be expressed as:

$$\alpha_m(\tilde{y}_\pi|x) = f_a(\tilde{y}_\pi)\, p_m(a|x),\ \forall x \in \mathcal{X}, \tilde{y}_\pi \in \mathcal{Y}_\pi(a),\ a \in \mathcal{A}. \tag{2.87}$$

It follows straightforwardly from the fictitious observation likelihood expression in (2.87) that, for all $\tilde{y}_\pi \in \mathcal{Y}_\pi(a)$, the belief is simply $p(\cdot|a)$:

$$p_m(x|\tilde{y}_\pi) \underset{\text{(from (2.87))}}{=} \frac{f_a(\tilde{y}_\pi)\, p_m(a|x)\pi_0(x)}{\sum_{x'} f_a(\tilde{y}_\pi)\, p_m(a|x')\pi_0(x')} = \frac{p_m(a|x)\pi_0(x)}{\sum_{x'} p_m(a|x')\pi_0(x')} = p_m(x|a) \tag{2.88}$$

The important but subtle consequence of this assumption is that the expected stopping cost (2.12) $G_{m,n}$ is *equal* to the surrogate cost $\tilde{G}_{m,n}$ (2.83) for all $m, n \in \{1, 2, \ldots, M\}$.

NIAS

Suppose NIAS inequality holds, that is, for all $m \in \mathcal{M}$,

$$a = \underset{b \in \mathcal{A}}{\operatorname{argmin}} \sum_{x \in \mathcal{X}} p_m(x|a) s_m(x, b),\ \forall a \in \mathcal{A}. \tag{2.89}$$

Since the set $\{p_m(x|a), a \in \mathcal{A}\}$ constitutes the set of all stopping beliefs when $\alpha_m(\tilde{y}_\pi|x) = p_m(a|x)$, the following condition holds from (2.89).

$$\mu_m(\pi, \tau) = \underset{a \in \mathcal{A}}{\operatorname{argmin}}\, \pi' \bar{s}_{m,a}. \tag{2.90}$$

Eq. 2.90 is precisely (2.8), which says the agent chooses the stop action that minimizes its stopping cost given its stopping belief. Hence, it only remains to show that (2.12) in Lemma 2 holds to complete our sufficiency proof.

NIAC

Assuming the NIAS condition (2.82) holds, we use the concept of KKT multipliers from duality theory [25, Sec. 5.5] to show that NIAC (2.86) is sufficient for (2.12) in Lemma 2 for optimal Bayesian stopping to hold. To do so, we use Lemma 16 below for linear assignment problems to show the existence of scalars C_m that satisfy (2.12); the feasibility inequality of interest is stated in (2.93) below. We now state Lemma 16 which can be viewed as a variational form of the NIAC inequality:

Lemma 16 *Suppose NIAC* (2.86) *holds. Then:*

(a) The solution of the following linear assignment problem is the identity map, that is, the optimal assignment map $x^*_{m,n}$ is given by $x^*_{m,n} = 1$ if $m = n$, and 0 otherwise:

$$\text{minimize}_{x_{m,n}} \sum_{m,n=1}^{M} x_{m,n}\, \tilde{G}_{m,n}, \text{ subject to:} \tag{2.91}$$

$$\sum_{n} x_{m,n} \geq 1,\ \sum_{m} x_{m,n} \geq 1,\ x_{m,n} \geq 0\ \ \forall\, m, n \in \{1, 2, \ldots, M\}.$$

(b) The KKT multipliers corresponding to the solution of the above assignment problem solve the feasibility condition of (2.12) *in Lemma 2.*

Proof.

(a) Let $x^*_{m,n}$ denote the optimal solution to the optimization problem (2.91). Indeed, since (2.91) is an LP, $x^*_{m,n} \in \{0, 1\}$. We can prove by contradiction that if NIAC (2.86) holds, then the optimal assignment variables $x^*_{m,n}$ is Kronecker delta, that is, $x^*_{m,n} = 1$ if

$m = n$ and 0:

Choose any arbitrary feasible $x_{m,n} \in \{0, 1\}$. Consider the sequence of indices $I \equiv \{1, h_x(1),\ h_x \circ h_x(1),\ \ldots,\ (h_x\circ)^{M-2}h(1)\}$, where $h_x(m) = m'$ is the unique (due to assignment constraints in (2.91)) index $m' \in \mathcal{M}$ for which $x_{m,m'} = 1$ and '$\circ$' denotes the function composition operator. From invoking NIAC (2.86) on the index sequence I, we observe that $\sum_{m,n=1}^{M} x_{m,n}\tilde{G}_{m,n} \leq \sum_m \tilde{G}_{m,m} = \sum_{m,n=1}^{M} x^*_{m,n}\tilde{G}_{m,n}$, where $x^*_{m,n} = 1$ if $m = n$ and 0 otherwise. Hence, the identity map solves the assignment problem (2.91).

(b) We now write down the Karush-Kuhn-Tucker (KKT) conditions [25, pg. 121] for the assignment problem (2.91) at the optimal solution $\{x^*_{m,n}\}_{m,n=1}^{M}$ that are first-order necessary conditions for optimality:

$$\begin{aligned}&\text{There exist scalars } \lambda_{1,m},\ \lambda_{2,m},\ \lambda_{3,m,n} \geq 0,\ m, n \in \{1, 2, \ldots M\},\ \text{such that:}\\ &\text{(i) For } n = m:\ \tilde{G}_{m,m} = \lambda_{1,m} + \lambda_{2,m},\ \text{(ii) For } n \neq m:\ \tilde{G}_{n,m} = \lambda_{1,m} + \lambda_{2,n} + \lambda_{3,n,m}.\end{aligned} \tag{2.92}$$

The scalars $\lambda_{1,m}$ and $\lambda_{2,n}$ in (2.92) correspond to KKT multipliers associated with the inequality constraints $(-\sum_n\ x_{m,n}) \leq -1$ and $(-\sum_m\ x_{m,n}) \leq -1$ in (2.91), respectively. We note that both sets of inequalities are active at $\{x^*_{m,n}\}_{m,n=1}^{M}$, the solution of (2.91). The scalar $\lambda_{m,n}$ is the KKT multiplier associated with the inequality constraint $(-x_{m,n}) \leq 0$, where the inequality is active only for $x^*_{m,n},\ m \neq n$. For any pair of environments $m, n \in \{1, 2, \ldots, M\}$, the following inequalities result due to the KKT conditions in (2.92):

$$\begin{aligned}&\tilde{G}_{m,m} - \lambda_{2,m} = \tilde{G}_{n,m} - \lambda_{2,n} - \lambda_{3,n,m} \leq \tilde{G}_{n,m} - \lambda_{2,n}\quad (\text{since } \lambda_{3,n,m} \geq 0)\\ \Leftrightarrow\ &\tilde{G}_{m,m} + (\max_{m'} \lambda_{2,m'} - \lambda_{2,m}) \leq \tilde{G}_{n,m} + (\max_{m'} \lambda_{2,m'} - \lambda_{2,n})\\ \Leftrightarrow\ &\tilde{G}_{m,m} + C_m \leq \tilde{G}_{n,m} + C_n\\ &(\text{by replacing } (\max_{m'} \lambda_{2,m'} - \lambda_{2,m}) \text{ with the variable } C_m \text{ for all } m \in \mathcal{M})\end{aligned} \tag{2.93}$$

We now reconstruct an estimate of the agent's expected continue cost $\widehat{C}$ below and

show (2.12) holds for Bayes optimal stopping. With $\boldsymbol{p}_\mu = p_\mu(a|x)$ denoting the action selection policy induced by a stopping strategy μ, consider the following reconstructed estimate of the agent's expected continue cost $\widehat{C}(\mu)$ in terms of the feasible variables $\{C_m\}_{m=1}^M$ (2.93):

$$\widehat{C}(\mu) = \max_{m=1,2,\ldots,M} \left\{ C_m + G_{m,m} - \tilde{G}(\mu, s_m) \right\}, \text{ where}$$
$$\tilde{G}(\mu, s_m) = \sum_{a\in\mathcal{A}} \left(\sum_{x\in\mathcal{X}} p_\mu(a|x)\pi_0(x) \right) \min_{b\in\mathcal{A}} \sum_{x\in\mathcal{X}} p_\mu(x|a) s_m(x,b) \tag{2.94}$$

In (2.94), $\tilde{G}(\mu, s_m)$ denotes the expected stopping cost of the Bayesian agent with strategy μ and stopping costs $s_m(x,a)$ *assuming a one-to-one map between the set of observations* $y_{1:\tau(\mu)}$ *to action* a.[18] The variable $p_\mu(x|a)$ is the posterior belief of the state computed using Bayes rule as:

$$p_\mu(x|a) = \frac{\pi_0(x)p_\mu(a|x)}{\sum_{x'} \pi_0(x')p_\mu(a|x')}$$

Indeed, if the mapping from the fictitious observation set $\mathcal{Y}_\pi$ to the action set $\mathcal{A}$ is assumed to be one-to-one, the expected stopping cost can be expressed in terms of the action selection policy $\boldsymbol{p}_\mu$ induced by the stopping strategy μ. From (2.93), it is straightforward to show that $C(\mu_m) = C_m$. Hence, replacing C_m in (2.93) with $\widehat{C}(\mu_m)$ yields the following inequalities:

$$\tilde{G}_{m,m} + \widehat{C}(\mu_m) \leq \tilde{G}_{n,m} + \widehat{C}(\mu_n)$$

$\Leftrightarrow$ A cumulative running cost can be reconstructed (2.94) such that condition (2.12) in Lemma 2 holds with expected stopping costs $\tilde{G}_{n,m},\ m,n \in \{1,2,\ldots,M\}$. ■

(2.95)

In words, for a feasible set of stopping costs $\{s_m\}_{m=1}^M$ such that NIAS and NIAC hold, the Bayesian agent's (unobserved) strategies satisfy optimal Bayesian stopping (2.12).

[18] While it may seem counter-intuitive to assume a one-to-one mapping from the fictitious observation space to action space, one can show for convex costs like entropic costs (Shannon-Gibbs, Rényi and Tsallis) that the *optimal* mapping is one-to-one. The key idea is to show that having a many-to-one map with the same expected stopping cost is sub-optimal in that the agent incurs a strictly larger expected continue cost; see [90, Lemma 1] for a more detailed explanation.

Moreover, for every feasible set of costs $\{s_m\}_{m=1}^M$, the term $(\max_{m'} \lambda_{2,m'} - \lambda_{2,m})$ denotes the expected continue cost incurred by the Bayesian agent due to choosing strategy μ_m, and $\tilde{G}_{m,n}$ denotes the agent's incurred expected stopping cost in environment n if it chooses strategy μ_m.

2.10.3 Remarks

1. *IRL for inverse SHT.* For the inverse SHT problem discussed in Sec. 2.3, the inverse learner knows C_m, the expected cumulative continue cost for the agent in environment m. Hence, the inverse learner can identify optimal SHT simply by checking if the NIAS inequality (2.82) and the following inequality is feasible:

$$\tilde{G}_{m,m} - \tilde{G}_{n,m} \leq C_n - C_m, \ \forall m, n \in \mathcal{M}, \tag{2.96}$$

where $\tilde{G}(\cdot)$ is the expected stop cost defined in (2.83) and C_m is the expected continue cost of the agent in environment m now known to the inverse learner. Due to (A5), $\tilde{G}_{m,\cdot} = G_{m,\cdot}$. Hence, (2.96) is equivalent to (2.12) in Lemma 2. We term the inequality in (2.96) as NIAC$*$ and use it in Theorem 6 for IRL for inverse SHT.

2. *Different observation likelihoods for different environments.* Theorem 3 is a purely *data-centric* approach for IRL that makes no assumptions on the agent's observation likelihood. If the NIAS and NIAC inequalities have a feasible solution, then the inverse learner's dataset $\mathcal{D}_M$ can be rationalized by a Bayesian agent that acts optimally (in the sense of Lemma 2) and has a fixed observation likelihood over all M environments. It may very well be the case that the Bayesian agent has a different observation likelihood in different environments, but Theorem 3 is opaque to this condition.

If the inverse learner knows *a priori* that the Bayesian agent uses a different observation likelihood for different environments, we need stronger conditions to achieve IRL. A

sufficient condition for identifying optimal Bayesian stopping with distinct observation likelihoods in different environments is to assume the expected cumulative continue cost of the agent is independent of the observation likelihood. One example that satisfies this assumption is the entropic continue cost:

$$c_t = \lambda \left(H(\pi_t) - \mathbb{E}\{H(\pi_{t+1})|\pi_t\} \right),\ t \geq 0,\ \lambda > 0 \tag{2.97}$$

where $H(p) = -\sum_i p_i \log(p_i)$ is the entropy of pmf p. The above choice of continue cost has two advantages:

(i) The expected *cumulative* continue cost for agent m is simply $H(\pi_0) - \mathbb{E}_a\{H(p_m(a|x))\}$, and is independent of the observation likelihood; see [90, Lemma 1] for a discussion on how conditioned on state x, the optimal mapping from the space of fictitious observations $\tilde{y}_\pi$ (2.75) to the space of actions $\mathcal{A}$ is one-to-one due to the convexity of the entropic cost.

(ii) The inverse learner can test for 'absolute optimality' (2.8), (2.9) of the Bayesian agent's decisions and does not require observing the agent's behavior in multiple environments. Using the method of Lagrange multipliers, it is straightforward to show that for environment m, the following relation holds between the agent's stopping costs and its observed decisions for optimal Bayesian stopping:

$$p_m(a|x) = \frac{p_m(a)\ \exp(-s_m(x,a)/\lambda)}{\sum_{b\in\mathcal{A}} p_m(b)\ \exp(-s_m(x,b)/\lambda)},\ \forall\ a, x, m, \tag{2.98}$$

where $\lambda > 0$ is a feasible variable that parametrizes the continue cost (2.97), and $p_m(a)$ is the marginal distribution of the action a in environment m. IRL is achieved by checking for the feasibility condition of [29, Eq. 3, Proposition 1] and solving the above set of equations (2.98) for $s_m(x,a)$; observe that there is no assumption of a fixed observation likelihood for the Bayesian agent across environments and the IRL estimate returns an ordinal estimate of the agent's stopping costs.

2.11 Proof of Theorem 9

We will show (2.46) is equivalent to the condition for identifying search optimality (2.45) in two steps. For a fixed stationary search strategy $\mu : \pi \to a$ and search cost $\{l(a), a \in \mathcal{A}\}$, we first express the expected cumulative search cost in terms of the search action policy $g(x, a)$ (2.43) and the prior π_0.

$$J(\mu, l) = \mathbb{E}_\mu \left\{ \sum_{t=1}^{\tau} l(\mu(\pi_t)) \right\} = \mathbb{E}_\mu \left\{ \sum_{a \in \mathcal{A}} l(a) \left(\sum_{t=1}^{\tau} \mathbb{1}\{\mu(\pi_t) = a\} \right) \right\}$$
$$= \sum_{a \in \mathcal{A}} l(a) \sum_{x \in \mathcal{X}} \pi_0(x) \mathbb{E}_\mu \left\{ \sum_{t=1}^{\tau} \mathbb{1}\{\mu(\pi_t) = a\} | x \right\} = \sum_{x \in \mathcal{X}, a \in \mathcal{A}} \pi_0(x) g(x, a) l(a).$$

Now, consider the set of search strategies $\{\mu_m, m \in \mathcal{M}\}$.

$$(2.45) \equiv \mu_m \in \underset{\{\mu_n, n \in \mathcal{M}\}}{\operatorname{argmin}} J(\mu_n, l_m) \iff J(\mu_m, l_m) - J(\mu_n, l_m) \leq 0, \ m, n \in \mathcal{M}.$$

$$\iff \sum_{x \in \mathcal{X}, a \in \mathcal{A}} \pi_0(x) (g_m(a, x) - g_n(a, x)) \, l_m(a) \leq 0 \equiv (2.46).$$

■

2.12 Proof of Theorem 13

We divide the proof of Theorem 11 into 4 steps:

Step 1. Using Dvoretzky-Kiefer-Wolfowitz (DKW) inequality [39] to bound the deviation of the empirically computed action selection policy $\hat{p}_m(a|x)$ from $p_m(a|x)$.

The DKW inequality [38] provides a finite sample characterization of the asymptotic result of Glivenko-Cantelli theorem by quantifying the convergence rate of the empirical cdf to the true cdf. Let $F_m(a|x)$ and $\hat{F}_m(a|x)$ denote the cdfs of $p_m(a|x)$ and $\hat{p}_m(a|x)$, respectively. From the two-sided DKW inequality, the following inequalities result:

$$1 - 2\exp\left(-2K_{x,m}\varepsilon^2\right) \leq \mathbb{P}\left(\max_{a \in \mathcal{A}} |\hat{F}_m(a|x) - F_m(a|x)| < \varepsilon \right) \leq \mathbb{P}\left(\max_{a \in \mathcal{A}} |\hat{F}_m(a|x) - F_m(a|x)| \leq \varepsilon \right)$$

$$\leq \mathbb{P}\left(|p_m(a|x) - \hat{p}_m(a|x)| \leq \varepsilon, \forall a\right) \leq \mathbb{P}\left(\sum_{a\in\mathcal{A}} |p_m(a|x) - \hat{p}_m(a|x)|^2 \leq A\varepsilon^2\right).$$

For a fixed state x and environment m, let $\varepsilon_{x,m}$ bound the error $|\hat{p}_m(a|x) - p_m(a|x)|$, $\forall a \in \mathcal{A}$. With $\varepsilon^2_{\max} = A(\sum_{x,m} \varepsilon^2_{x,m})$, we have the following probabilistic bound on the L_2-error between the true and empirical action selection policies, summed over all states, actions and environments:

$$\boxed{\mathbb{P}\left(\sum_{a,x,m} |p_m(a|x) - \hat{p}_m(a|x)|^2 \leq \varepsilon^2_{\max}\right) \geq \prod_{x,m} 1 - 2\exp\left(-2K_{x,m}\varepsilon^2_{x,m}\right)}. \quad (2.99)$$

Step 2. Using Hoeffding's Inequality [40] to bound the deviation of the sample average of the SHT stopping times $\hat{C}_m$ from the true value C_m.

The inverse learner knows the agent's stopping time $\tau \in [1, \tau_{\max}]$ for all M environments. Analogous to (2.99), for a fixed environment m, let η_m bound the error $|\hat{C}_m - C_m|$. With $\eta^2_{\max} = \sum_m \eta^2_m$, we have the following probabilistic bound on the L_2-error between the true and empirical expected stopping times of the agent, summed over all M environments via the two-sided Hoeffding's inequality:

$$\begin{aligned}
& \mathbb{P}(|\hat{C}_m - C_m| \leq \eta_m) \geq 1 - 2\exp\left(-2K_m\eta^2_m/\tau^2_{\max}\right) \\
\implies\ & \mathbb{P}(\sum_{x,m} |\hat{C}_m - C_m|^2 \leq \eta^2_{\max}) \geq \mathbb{P}(|\hat{C}_m - C_m| \leq \eta_m,\ \forall m \in \mathcal{M}) \\
\implies\ & \boxed{\mathbb{P}(\sum_{x,m} |\hat{C}_m - C_m|^2 \leq \eta^2_{\max}) \geq \prod_m 1 - 2\exp\left(-\frac{2K_m\eta^2_m}{\tau^2_{\max}}\right)}, \text{ where } K_m = \sum_x K_{x,m}
\end{aligned} \quad (2.100)$$

Step 3. Using the union bound on error bounds from steps 1 and 2 to bound the cumulative deviation of empirically computed action selection policies and expected stopping times.

Our aim is to construct a tight bound on the probability of the event $E_{pert}(\delta_{\max})$, where $E_{pert}(\delta_{\max})$ is defined as:

$$E_{pert}(\delta_{\max}) \equiv \{ \{\hat{p}_m(a|x), \hat{C}_m\} \,\big|\, \sum_{x,m} |p_m(a|x) - \hat{p}_m(a|x)|^2 + \sum_x |C_m - \hat{C}_m|^2 \leq \delta^2_{\max}\}, \tag{2.101}$$

We note that $\mathbb{P}(E_{pert}(\delta_{\max}))$ bounds the Type-I and Type-II IRL error probabilities for suitable choices of $\delta_{\max}$. The Type-I error probability is bounded by $1 - \mathbb{P}(E_{pert})$ when $\delta^2_{\max}$ in (2.101) is set to $\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K}))$. Also, the Type-II error probability is bounded by $1 - \mathbb{P}(E_{pert})$ when $\delta^2_{\max}$ in (2.101) is set to $\varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K}))$. Recall from (2.57), (2.58) that $\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K}))$ and $\varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K}))$ correspond to the minimum L_2-perturbation needed for the *finite* IRL dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ (2.53) to pass and fail, respectively, the NIAS and NIAC conditions of Theorem 3 for inverse optimal Bayesian stopping.

For a fixed error tuple $\{\varepsilon_{x,m}, \eta_m\}$, consider the surrogate event $E(\{\varepsilon_{x,m}, \eta_m\})$ defined as:

$$E(\{\varepsilon_{x,m}, \eta_m\}) = \{ \{\hat{p}_m(a|x), \hat{C}_m\} \,\big|\, \hat{p}_m(a|x) - p_m(a|x)| \leq \varepsilon_{x,m}, |\hat{C}_m - C_m| \leq \eta_m, \; \forall\, a, x, m\}. \tag{2.102}$$

Clearly, $E(\{\varepsilon_{x,m}, \eta_m\}) \;\subseteq\; E_{pert}(\delta_{\max})$ if $\delta^2_{\max}$ in (2.101) is equal to $A\sum_{x,m}\varepsilon^2_{x,m} + \sum_m \eta^2_{x,m}$. Combining the error bounds in (2.99) and (2.100) via a union bound to bound $\mathbb{P}(E(\{\varepsilon_{x,m}, \eta_m\}))$ yields the following inequality:

$$\begin{aligned}
&\mathbb{P}(E_{pert}(\delta_{\max})) \geq \mathbb{P}(E(\{\varepsilon_{x,m}, \eta_m\})) \\
&\geq \prod_{x,m} 1 - 2\exp\left(-2K_{x,m}\varepsilon^2_{x,m}\right) \prod_m 1 - 2\exp\left(-2K_m\eta^2_m/\tau^2_{\max}\right) &(2.103)\\
&\geq 1 - \sum_{x,m} 2\exp\left(-2K_{x,m}\varepsilon^2_{x,m}\right) - \sum_m 2\exp\left(-2K_m\eta^2_m/\tau^2_{\max}\right) &(2.104)\\
&\implies \boxed{\mathbb{P}(E_{pert}(\delta_{\max})) \geq 1 - \sum_{x,m} 2\exp\left(-2K_{x,m}\varepsilon^2_{x,m}\right) - \sum_m 2\exp\left(-2K_m\eta^2_m/\tau^2_{\max}\right)}, &(2.105)
\end{aligned}$$

where $\delta^2_{\max} = A\sum_{x,m}\varepsilon^2_{x,m} + \sum_m \eta^2_m$. The inequality in (2.103) is simply a union bound on the error bounds in (2.99) and (2.100). The inequality in (2.104) holds due to Assumption (F5) that says the analyst observes the agent's stopping action over sufficiently many trials. If the expected stopping time is known accurately, that is, $\hat{C}_m = C_m$ for all $m \in \mathcal{M}$, Assumption (F5) specializes to Assumption (F2).

Step 4. Obtaining a tight bound on the error probability computer in step 3.

Eq. 2.105 provides a probabilistic bound on the perturbation of the empirical dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ from the asymptotic dataset $\mathcal{D}_M$ in terms of the sample size $\mathcal{K} = \{K_{x,m}, x \in \mathcal{X}, m \in \mathcal{M}\}$, for a fixed error sequence $\{\varepsilon_{x,m}, \eta_m\}$. Our final step is to maximize the RHS in (2.105) subject to the constraint $A(\sum_{x,m}\varepsilon^2_{x,m}) + \sum_m \eta^2_m = \delta^2_{\max}$ to obtain the tightest bound on $\mathbb{P}(E_{pert}(\delta_{\max}))$ (2.105). Equivalently, our aim is to minimize the following objective function:

$$\boxed{\min_{\{\varepsilon^2_{x,m},\eta^2_m\}\geq 0} \sum_{x,m} 2\exp\left(-2K_{x,m}\varepsilon^2_{x,m}\right) + \sum_m 2\exp\left(-2K_m\eta^2_m/\tau^2_{\max}\right),\ \text{s.t. } A\sum_{x,m}\varepsilon^2_{x,m} + \sum_m \eta^2_m = \delta^2_{\max}}. \tag{2.106}$$

We observe that the terms $\exp\left(-2K_{x,m}\varepsilon^2_{x,m}\right)$ and $\exp\left(-2K_m\eta^2_m/\tau^2_{\max}\right)$ are convex in $\varepsilon^2_{x,m}$ and η^2_m, respectively. Also, Assumption (F4) ensures the values of the terms $\varepsilon^2_{x,m}, \eta^2_m$ are bounded away from 0 at the local optimum of (2.106) computed via the method of Lagrange multipliers. Assumption (F5) ensures the $\varepsilon^2_{x,m}, \eta^2_m$ in (2.107) satisfy the Slater's condition for regularity. Since the equality constraint is linear, and the objective function is convex in the feasible variables $\varepsilon^2_{x,m}$ and η^2_m, (2.106) constitutes a convex optimization problem whose solution can be computed using the method of Lagrange multipliers (since Assumption (F4) ensures inactive inequality constraints at the optimal solution). Finally, the solution of the optimization problem (2.106) can be expressed as:

$$\boxed{\varepsilon^2_{x,m} = (A\widetilde{K}_{x,m}/2)\left(\ln(\lambda) + \ln(2K_{x,m}/A)\right),\ \eta^2_{x,m} = (\widetilde{\bar{K}}_m/2)(\ln(\lambda) + \ln\big(2\bar{K}_m\big))}, \text{ where}$$

$$\ln(\lambda) = \frac{\delta^2_{\max} - A\sum_{x,m} \ln\Big((2K_{x,m}/A)^{A\widetilde{K}_{x,m}/2}\Big) - \sum_m \ln\Big(2\bar{K}_m^{\widetilde{\bar{K}}_m/2}\Big)}{(A\sum_{x,m}\widetilde{K}_{x,m} + \sum_{m\in\mathcal{M}}\widetilde{\bar{K}}_m)/2} \ . \tag{2.107}$$

In the above equations, $\bar{K}_{(\cdot)} = K_{(\cdot)}/\tau^2_{\max}$, $\widetilde{K} = K^{-1}$. Subtracting the objective function of (2.106) evaluated at the optimal values of $\varepsilon^2_{x,m}$ and η^2_m (2.107) from 1, and setting $\delta^2_{\max}$ to $\varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K}))$ and $\varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K}))$, respectively, yield lower bounds for Type-I and Type-II error probabilities of the IRL detector, respectively. ■

Remark. The proof of Theorem 11 for finite sample complexity of Theorem 3 is identical to the above proof structure (except that there is no step 2) and hence, omitted for brevity.

2.13 Proof of Theorem 15

To prove Theorem 15, we first state and prove an auxiliary result, namely, Proposition 17, below. Theorem 15 is a special case of Proposition 17 as discussed below.

Proposition 17 *Given dataset $\widehat{\mathcal{D}}_M(\mathcal{K})$ and (F7), the deviation of the finite sample search action policy $\hat{g}_m(a,x)$ and the true search action policy $g_m(a,x)$ can be bounded in terms of the number of samples $\mathcal{K} = \{K_{x,m}\}$ as follows.*

$$\mathbb{P}\left(\sum_{a,x,m} |g_m(a,x) - \hat{g}_m(a,x)|^2 \le \epsilon\right) \ge 1 - \sum_{x,m} \frac{u(\widehat{\mathcal{D}}_M(\mathcal{K}))}{\epsilon K^{1/2}_{x,m}}, \tag{2.108}$$

where $u(\widehat{\mathcal{D}}_M(\mathcal{K})) = \frac{(1-\alpha^*)}{(\alpha^*)^2} X \sum_{x,m} K^{-1/2}_{x,m}$ and $\hat{g}_m(a,x)$ is the sample average of the number of times the agent searches location a given state x in environment m.

Proof. Assumption (F7) implies that for any environment $m \in \mathcal{M}$, given prior π_0, the agent's optimal search sequence $a_0, a_1, a_2, \ldots$ is periodic, i.e., $a_t = a_{t+A}$. In other

words, in any interval of A time steps, the agent searches each location exactly once in a particular (unknown to the inverse learner) order.

Consider the agent's search policy $g_m(a,x)$ defined in (2.43). Below we express $g_m(a,x)$ in terms of the pdf of the stopping time of the search process.

$$g_m(a,x) = \mathbb{E}_{\mu_m}\left\{\sum_{t=1}^{\tau}\mathbb{1}\{\mu_m(\pi_t)=a\}|x^o=x\right\} = \sum_{t=1}^{\infty}\mathbb{P}_{\mu_m}\left(\tau=t|x^o=x\right)\left(\lfloor t/X\rfloor + r(x,a)\right).$$
$$= \sum_{t=1}^{\infty}\lfloor t/X\rfloor\mathbb{P}_{\mu_m}\left(\tau=t|x^o=x\right) + r(x,a) = \mathbb{E}_{\mu_m}\{\lfloor\tau/X\rfloor\,|x\} + r_m(x,a) = \frac{1}{\alpha} + r_m(x,a).$$

Here, α denotes the reveal probability of the agent and $\lfloor\cdot\rfloor$ denotes the floor function. $r(x,a)=1$ if agent searches location a prior to location x in one search cycle from time $t=0\rightarrow X-1$ in environment m, and 0 otherwise. The final equality follows from the fact that conditioned on the true state $x^o=x$, the random variable $\lfloor\tau/X\rfloor$ follows a geometric distribution with parameter α (unknown) due to (F7).

Consider now the quantity $|\hat{g}_m(a,x) - g_m(a,x)|$. Define $\hat{\mathbb{E}}_{\mu_m}\{\tau/X\} = \frac{\sum_{k=1}^{K_{x,m}}\lfloor\tau_{x,m,k}/X\rfloor}{K_{x,m}}$, the sample average of the normalized stopping time $\lfloor\tau/X\rfloor$ computed from $\widehat{\mathcal{D}}_M(\mathcal{K})$. Then,

$$|\hat{g}_m(a,x) - g_m(a,x)| = |\mathbb{E}_{\mu_m}\{\lfloor\tau/X\rfloor\,|x\} + r_m(x,a) - \hat{\mathbb{E}}_{\mu_m}\{\lfloor\tau/X\rfloor\,|x\} - r_m(x,a)|$$
$$= \left|\frac{1}{\alpha} - \hat{\mathbb{E}}_{\mu_m}\{\lfloor\tau/X\rfloor\,|x\}\right| \text{ (equal for all } a \text{ for a fixed } x).$$

$\hat{\mathbb{E}}_{\mu}\{\lfloor\tau/X\rfloor|x\}$ is an unbiased estimator of $\mathbb{E}_{\mu}\{\lfloor\tau/X\rfloor|x\}$ with variance $(1-\alpha)/K_{x,m}\alpha^2$. Using Chebyshev's inequality for random variables with finite variance to bound $|\hat{g}_m(a,x) - g_m(a,x)|$ for fixed a,x,m, the following inequality results

$$\mathbb{P}\left(|\hat{g}_m(a,x) - g_m(a,x)| \le \varepsilon\right) \ge 1 - \frac{(1-\alpha)}{K_{x,m}(\alpha\varepsilon)^2} \tag{2.109}$$

For any set of positive reals $\{\varepsilon_{x,m}, x\in\mathcal{X}, m\in\mathcal{M}\}$ s.t. $|g_m(a,x) - \hat{g}_m(a,x)| \le \varepsilon_{x,m}$ and $(A\sum_{x,m}\varepsilon_{x,m}^2) \le \varepsilon_{\max}^2$, we have

$$\mathbb{P}(\sum_{x,m}|\hat{g}_m(a,x) - g_m(a,x)|^2 \le \varepsilon_{\max}^2) \ge \prod_{x,m} 1 - \frac{(1-\alpha)}{K_{x,m}(\alpha\varepsilon_{x,m})^2}$$

$$\geq 1 - \sum_{x,m} \frac{(1-\alpha)}{K_{x,m}(\alpha\varepsilon_{x,m})^2} \geq 1 - \sum_{x,m} \frac{(1-\alpha^*)}{K_{x,m}(\alpha^*\varepsilon_{x,m})^2}. \tag{2.110}$$

Since $\frac{(1-\alpha)}{K_{x,m}(\alpha\varepsilon_{x,m})^2}$ is decreasing in $\varepsilon_{x,m}$, the tightest lower bound is achieved for the above inequality when $\sum_{x,m} \varepsilon_{x,m}^2 = \varepsilon_{\max}^2$ and is the solution to the following constrained optimization problem.

$$\min_{\{\varepsilon_{x,m}, x\in\mathcal{X}, m\in\mathcal{M}\}} \sum_{x,m} \frac{(1-\alpha^*)}{K_{x,m}(\alpha^*\varepsilon_{x,m})^2} \text{ s.t. } A\sum_{x,m} \varepsilon_{x,m}^2 = \varepsilon_{\max}^2. \tag{2.111}$$

Moreover, since the objective function in (2.111) is convex in $\varepsilon_{x,m}^2$ and constraint is affine in $\varepsilon_{x,m}^2$, the method of Lagrange multipliers [25] yields necessary and sufficient conditions for an optimal solution to the above optimization problem if the solution obtained is positive for all $x \in \mathcal{X}, m \in \mathcal{M}$. The optimal value of $\varepsilon_{x,m}^2 = \frac{\varepsilon_{\max}^2 K_{x,m}^{-1/2}}{A\sum_{x,m} K_{x,m}^{-1/2}} > 0$ and thus minimizes the objective function in (2.111). Plugging this value in (2.110) and setting $\varepsilon_{\max}^2 = \epsilon$ yields the bound in the RHS of (2.108) and completes the proof for Proposition 17.

To obtain the error bounds (2.71), note that setting $\varepsilon = \varepsilon_1(\widehat{\mathcal{D}}_M(\mathcal{K}))$ in (2.108) and subtracting the objective function from 1 bounds from below the Type-I error probability (see Sec. 2.6.2 for a detailed explanation). Similarly, setting $\varepsilon = \varepsilon_2(\widehat{\mathcal{D}}_M(\mathcal{K}))$ in (2.108) and subtracting the objective function from 1 bounds from below the Type-II error probability of the IRL detector which completes the proof. ■

2.14 Context. IRL For Predicting YouTube Commenting Behavior

Our previous work [12] analyzes YouTube user engagement from a behavioral economics viewpoint. Although we use the same dataset for our numerical experiments in this chapter, we emphasize that the IRL approach in this chapter is new and differs from [12] as:

(1) In [12], we check if YouTube engagement is consistent with rationally inattentive utility maximization behavior [15], a *static* decision model studied widely in behavioral and information economics. In comparison, our aim here is to test if the YouTube dataset satisfies Bayes optimal stopping, a *dynamic* decision model.

(2) The inference algorithms in [12] considers *pairs* of video categories to reconstruct the underlying utility function of the YouTube user. In this chapter, our IRL approach considers *all* 18 *YouTube video categories (described in Sec. 2.5.1 below) simultaneously in the feasibility test* for reconstructing the underlying stopping costs of the YouTube user, and hence, fully exploits the diversity in engagement behavior.

(3) In [12], we perform a naive prediction analysis of YouTube user engagement using a *maximum a posteriori* (MAP) approach. In this chapter, we predict the distribution of user engagement behavior via two representative point estimates of the recovered stopping costs and show the statistical similarity of the predicted distribution to the true engagement distribution.

YouTube user engagement and Bayesian stopping.

YouTube is a social multimedia platform where human users interact with video content on YouTube channels by posting comments and rating videos. Empirical studies ([36, 91–93]) show that the comments and ratings from users are influenced by the thumbnail, title, category, and perceived popularity of each video. Models for human decision making in the context of online multimedia platforms have been studied extensively in the literature. Two widely-used classes of models that motivate us to understand YouTube user engagement from the lens of Bayesian stopping are 'parallel constraint satisfaction models' and 'evidence accumulation models'. *Parallel constraint satisfaction models* ([94, 95]) assume that information is screened sequentially to highlight salient alternatives and final choice is made when the decision maker reaches sufficient internal coherence. *Evidence accumulation models* ([50, 51]) model consumers' attention by

drift-diffusion models that accumulate evidence based on whether they are fixating their gaze on either the product or its price. The decision is taken when any of the alternatives' evidence threshold level is achieved.

Both classes of models described above have one aspect in common - the decision maker makes a final choice *after* sequentially accumulating information, and naturally fits our Bayesian stopping time framework. In terms of YouTube webpage parameters, we hypothesize the YouTube user is a Bayesian agent that sequentially consumes webpage cues such as thumbnail, title and perceived popularity and incurs a cost of attention, followed by engaging on the YouTube platform and incurring a terminal cost. Our IRL aim in this section is to identify using the YouTube dataset, if YouTube users engage 'optimally' in a Bayesian stopping sense.

CHAPTER 3

IRL FOR IDENTIFICATION AND MITIGATION OF ADVERSARY SENSING SYSTEMS

3.1 Introduction

Cognitive sensors are reconfigurable sensors that optimize their sensing mechanism and transmit functionalities. The concept of cognitive radar [96–99] has evolved over the last two decades and a common aspect is the sense-learn-adapt (SLA) paradigm. A cognitive fully adaptive radar enables joint optimization of the adaptive transmit and receive functions by sensing (estimating) the radar channel that includes clutter and other interfering signals [100, 101].

Objectives

This chapter addresses the next step and achieves the following objectives schematically shown in Figure 3.1. The framework in this chapter involves an adversarial signal processing problem comprising "us" and an "adversary". "Us" refers to an asset such as a drone/UAV or electromagnetic signal that probes an "adversary" cognitive radar. Figure 3.2 shows the schematic setup. A cognitive sensor observes our kinematic state x_k in noise as the observation y_k. It then uses a Bayesian tracker to update its posterior distribution π_k of our state x_k and chooses an action u_k based on this posterior. We observe the sensor's action in noise as a_k. Given knowledge of "our" state sequence $\{x_k\}$ and the observed actions $\{a_k\}$ taken by the adversary's sensor, we focus on the following inter-related aspects:

1. Inverse tracking and estimating the Adversary's Sensor Gain. Suppose the adversary radar observes our state in noise; updates its posterior distribution π_k of our state x_k using a Bayesian tracker, and then chooses an action u_k based on this posterior. Given knowledge of "our" state and sequence of noisy measurements $\{a_k\}$ of the adversary's actions $\{u_k\}$, how can the adversary radar's posterior distribution (random measure) be estimated? We will develop an inverse Bayesian filter for tracking the radar's posterior belief of our state and present an example involving the Kalman filter where the inverse filtering problem admits a finite dimensional characterization.

A related question is: How to remotely estimate the adversary radar sensor's conditional pdf of observation given the state when it is estimating us? This is important because it tells us how accurate the adversary's sensor is; in the context of Figure 3.2 it tells us, how accurately the adversary tracks our drone. The data we have access to is our state (probe signal) sequence $\{x_k\}$ and measurements of the adversary's radar actions $\{a_k\}$. Estimating the adversary's sensor accuracy is non-trivial with several challenges. First, even though we know our state and state dynamics model (transition law), the adversary does not. The adversary needs to estimate our state and state transition law based on our trajectory; and we need to estimate the adversary's estimate of our state transition law. Second, computing the MLE of the adversary's sensor gain also requires inverse filtering.

2. Revealed Preferences and Identifying Cognitive Radars. Suppose the cognitive radar is a constrained utility maximizer that optimizes its actions u_k subject to physical level (Bayesian filter) constraints. How can we detect this utility maximization behavior? The actions u_k can be viewed as resources the radar adaptively allocates to maximize its utility. We consider two such resource allocation problems, namely,

- *Beam Allocation*: The radar adaptively switches its beam while tracking multiple

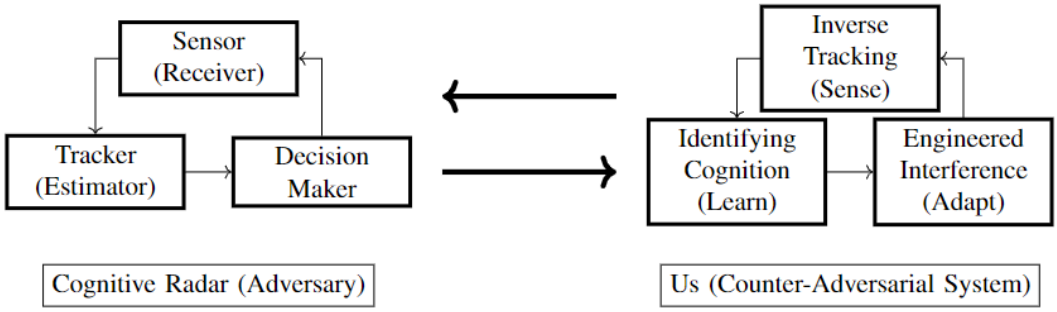

Figure 3.1: Schematic illustrating the main ideas in the chapter. The three components on the right are inter-related and constitute the sense-learn-adapt paradigm of the observer ("us") reacting to a reactive system such as the cognitive radar (on the left). This chapter considers the above schematic and proposes counter-adversarial schemes against cognitive radars for different levels of abstraction, i.e., interference design based on Wiener filters at the pulse/waveform level, inverse Kalman filters at the Bayesian tracking level, and revealed preference techniques for estimating adversary's utility function at the systems level.

targets.

- *Waveform Design*: The radar adaptively designs its waveform while ensuring the signal-to-interference-plus-noise ratio (SINR) exceeds a pre-defined threshold.

Nonparametric detection of utility maximization behavior is the central theme of *revealed preference* in microeconomics. A remarkable result is *Afriat's theorem*: it provides a necessary and sufficient condition for a finite dataset to have originated from a utility maximizer. We will develop constrained set-valued utility estimation methods that account for signal processing constraints introduced by the Bayesian tracker for performing adaptive beam allocation and waveform design respectively.

3. Smart Signal Dependent Interference. We next consider the adversary radar choosing its transmit waveform for target tracking by implementing a Wiener filter to maximize its signal-to-clutter-plus-noise ratio (SCNR[1]). By observing the optimal waveform chosen by the radar, our aim is to develop a strategy to estimate the adversary

[1]The terms SCNR and SINR are used interchangeably in the chapter.

cognitive radar channels and then construct signal dependent interference generation to confuse the adversary radar.

Perspective

The adversarial dynamics considered in this chapter fit naturally within the so called Dynamic Data and Information Processing (DDIP) paradigm. The adversary's radar senses, adapts and learns from us. In turn we adapt, sense and learn from the adversary. So in simple terms we are modeling and analyzing the interaction of two DDIP systems. In this context, this chapter has three major themes schematically shown in Figure 3.1: inverse filtering which is a Bayesian framework for interacting DDIP systems, inverse cognitive sensing which is a non-parametric approach for utility estimation for interacting DDIP systems, and interference design to confuse the adversarial DDIP system.

This work is also motivated by the design of counter-autonomous systems: given measurements of the actions of an autonomous adversary, how can our counter-autonomous system estimate the underlying belief of the adversary, identify if the adversary is cognitive (constrained utility maximizer) and design appropriate probing signals to confuse the adversary. This chapter generalizes and contextualizes recent works in adversarial signal processing [102, 103] which only deal with specific radar functionalities. Instead, this chapter views the cognitive radar as a holistic system operating at three stages of sophistication unifies the three inter-related aspects of adversarial signal processing, namely, inverse tracking, identifying cognition and designing interference. The three components complement one another and constitute this chapter's adversarial signal processing sense-learn-adapt (SLA) paradigm of Figure 3.1.

Organization

We conclude this section with a brief outline of the key results of the following sections, and their relevant to the sense, learn and adapt elements of the SLA paradigm of Figure 3.1.

Sense: In Sec. 3.2, we discuss inverse tracking techniques to estimate the sensor accuracy of an adversary radar. We focus mainly on the inverse Kalman filter and illustrate in carefully chosen examples how the adversary sensor's accuracy can be estimated. This constitutes the 'sensing' aspect of the SLA paradigm.

Learn: In Sec. 3.3, we abstractly view the adversarial radar as a cognitive decision maker that maximizes a utility function subject to physical resource constraints. Specifically, we show that if the cognitive radar optimizes its waveform to maintain its SINR above a threshold, then we can identify (and hence, 'learn') the utility function of the radar. The utility function provides deeper knowledge of the radar's behavior and constitutes the 'learn' element of the SLA paradigm.

Adapt: In Sec. 3.4, we consider a slightly modified setup where the radar chooses its waveform to maximize its SCNR. We show that by intelligently probing the radar with interference signals and observing the changes in the radar's waveform, we can confuse the adversary's radar by decreasing its SCNR. This adaptive signal processing algorithm is justified only if the 'sense' and 'learn' aspect of the SLA paradigm function properly, that is, the counter-adversarial system knows how the radar will react to changes in its environment.

Finally, we emphasize that the three main aspects of inverse tracking (sensing the estimate of the adversary), identifying utility maximization (learning the adversary's utility function) and adaptive interference (adapting our response) are instances of the

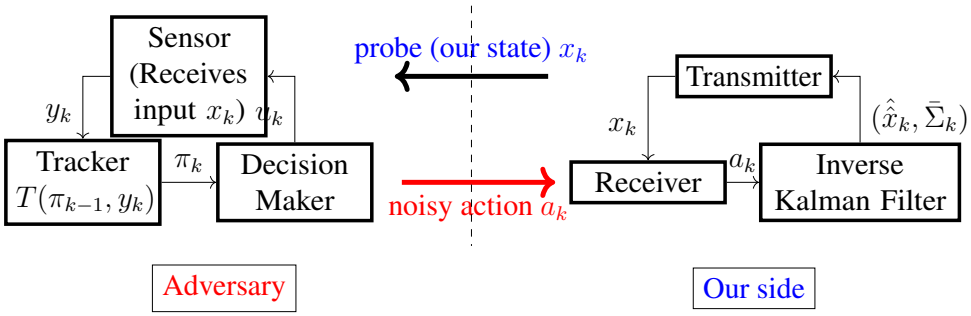

Figure 3.2: Schematic of Adversarial Inference Problem. Our side is a drone/UAV or electromagnetic signal that probes the adversary's cognitive radar system. Based on the action a_k of the adversary, our side computes the estimate of the adversary's estimate of our state x_k using the inverse Kalman filter outlined in Sec. 3.2.2.

general paradigm of sense-learn-adapt in counter-adversarial systems. As mentioned above, our formulation deals with the interaction of two such sense-learn-adapt systems.

3.2 Inverse Tracking and Estimating Adversary's Sensor

This section discusses inverse tracking in an adversarial system schematically illustrated in Figure 3.2. Our main ideas involve estimating the adversary's estimate of us and estimating the adversary's sensor conditional pdf of observation given the state.

3.2.1 Background and Preliminary Work

We start by formulating the problem which involves two entities; "us" and "adversary". With $k = 1, 2, \ldots$ denoting discrete time, the model has the following dynamics:

$$\begin{aligned} &x_k \sim P_{x_{k-1},x} = p(x|x_{k-1}), \quad x_0 \sim \pi_0,\ y_k \sim B_{x_k,y} = p(y|x_k) \\ &\pi_k = T(\pi_{k-1}, y_k) = p(x_k|y_{1:k}),\ a_k \sim G_{\pi_k,a} = p(a|\pi_k) \end{aligned} \tag{3.1}$$

Let us explain the notation in (3.1):

- $x_k \in \mathcal{X}$ is our Markovian state with transition kernel $P_{x_{k-1},x}$, prior π_0 and state space $\mathcal{X}$.
- y_k is the adversary's noisy observation of our state x_k; with conditional pdf of observation given the state $B_{x_k,y}$.
- π_k is the adversary's belief (posterior) of our state x_k where $y_{1:k}$ denotes the observation sequence $y_1, \ldots, y_k$. The operator T in (3.1) is the classical Bayesian optimal filter that computes the posterior belief of the state given observation y and current belief π.

$$T(\pi, y) = \text{vec}\left(\frac{B_{x,y} \int_{\mathcal{X}} P_{\zeta,x}\, \pi(\zeta)\, d\zeta}{\int_{\mathcal{X}} B_{x,y} \int_{\mathcal{X}} P_{\zeta,x}\, \pi(\zeta)\, d\zeta dx}, x \in \mathcal{X} \right) \tag{3.2}$$

  Let Π denote the space of all such beliefs. When the state space $\mathcal{X}$ is finite, then Π is the unit $X-1$ dimensional simplex of X-dimensional probability mass functions.
- a_k denotes our measurement of the adversary's action based on its current belief π_k. The adversary chooses an action u_k as a (possibly) stochastic function of π_k and we obtain a noisy measurement of u_k as a_k. We encode this as G_{π_k,a_k}, the conditional probability of observing action a_k given the adversary's belief π_k. Although not explicitly shown, G abstracts two stochastic maps: 1) the map from the adversary's belief π_k to its action u_k, and 2) the map from the adversary's action u_k to our noisy measurement a_k of this action.

Figure 3.2 displays a schematic and graphical representation of the model (3.1). The schematic model shows "us" and the adversary's variables.

Aim: Referring to model (3.1) and Figure 3.2, we address the following questions in this section:

1. How to estimate the adversary's belief given measurements of its actions (which are based on its filtered estimate of our state)? In other words, assuming probability

distributions P, B, G are known[2], we aim to estimate the adversary's belief π_k at each time k, by computing posterior $p(\pi_k \mid \pi_0, x_{0:k}, a_{1:k})$.

2. How to estimate the adversary's observation kernel B, i.e its sensor gain? This tells us how accurate the adversary's sensor is.

From a practical point of view, estimating the adversary's belief and sensor parameters allows us to calibrate its accuracy and predict (in a Bayesian sense) future actions of the adversary.

Related Works. In recent works [104–106], the mapping from belief π to adversary's action u was assumed deterministic. In comparison, our proposed research here assumes a probabilistic map between π and a and we develop Bayesian filtering algorithms for estimating the posterior along with MLE (Maximum Likelihood Estimation) algorithms for estimating the underlying model. Estimating/reconstructing the posterior given decisions based on the posterior is studied in microeconomics under the area of social learning [107] and game-theoretic generalizations [108]. There are strong parallels between inverse filtering and Bayesian social learning [107], [109–111]; the key difference is that social learning aims to estimate the underlying state given noisy posteriors, whereas our aim is to estimate the posterior given noisy measurements of the posterior and the underlying state. Recently, [112] used cascaded Kalman filters for LQG control over communication channels. This work motivates the design of the function ϕ in (3.8) below that maps the adversary's belief to its action; see also footnote 5. Finally, in [113], the authors investigate the inverse problem of trajectory identification based on target measurements, where the target is assumed to follow a constant velocity model.

[2]As mentioned in footnote 6, this assumption simplifies the setup; otherwise we need to estimate the adversary's estimate of us, which makes our task substantially complex.

3.2.2 Inverse Tracking Algorithms

How to estimate the adversary's posterior distribution of us?

Here we discuss inverse tracking for the model (3.1). Define the posterior distribution $\rho_k(\pi_k) = p(\pi_k|a_{1:k}, x_{0:k})$ of the adversary's posterior distribution given our state sequence $x_{0:k}$ and actions $a_{1:k}$. Note that the posterior $\rho_k(\cdot)$ is a *random measure* since it is a posterior distribution of the adversary's posterior distribution (belief) π_k. By using a discrete time version of Girsanov's theorem and appropriate change of measure[3] [115] (or a careful application of Bayes rule) we can derive the following functional recursion for ρ_k (see [102])

$$\rho_{k+1}(\pi) = \frac{G_{\pi,a_{k+1}} \int_\Pi B_{x_{k+1},y_{\pi_k,\pi}} \rho_k(\pi_k) d\pi_k}{\int_\Pi G_{\pi,a_{k+1}} \int_\Pi B_{x_{k+1},y_{\pi_k,\pi}} \rho_k(\pi_k) d\pi_k \, d\pi} \tag{3.3}$$

Here $y_{\pi_k,\pi}$ is the observation such that $\pi = T(\pi_k, y)$ where T is the adversary's filter (3.2). We call (3.3) the *optimal inverse filter* since it yields the Bayesian posterior of the adversary's belief given our state and noisy measurements of the adversary's actions.

Example: Inverse Kalman Filter

We consider a special case of (3.3) where the inverse filtering problem admits a finite dimensional characterization in terms of the Kalman filter. Consider a linear Gaussian state space model

$$\begin{aligned} x_{k+1} &= A\, x_k + w_k, \quad x_0 \sim \pi_0 \\ y_k &= C\, x_k + v_k \end{aligned} \tag{3.4}$$

[3]This chapter deals with discrete time. Although we will not pursue it here, the recent chapter [114] uses a similar continuous time formulation. This yields interesting results involving Malliavin derivatives and stochastic calculus.

where $x_k \in \mathcal{X} = \mathbb{R}^X$ is "our" state with initial density $\pi_0 \sim \mathcal{N}(\hat{x}_0, \Sigma_0)$, $y_k \in \mathcal{Y} = \mathbb{R}^Y$ denotes the adversary's observations, $w_k \sim \mathcal{N}(0, Q_k)$, $v_k \sim \mathcal{N}(0, R_k)$ and $\{w_k\}$, $\{v_k\}$ are mutually independent i.i.d. processes. Here, $\mathcal{N}(\mu, C)$ denotes the normal distribution with mean μ and covariance matrix C.

Based on observations $y_{1:k}$, the adversary computes the belief $\pi_k = \mathcal{N}(\hat{x}_k, \Sigma_k)$ where $\hat{x}_k$ is the conditional mean state estimate and Σ_k is the covariance; these are computed via the classical Kalman filter equations:[4]

$$\begin{aligned}
\Sigma_{k+1|k} &= A\Sigma_k A' + Q_k,\ S_{k+1} = C\Sigma_{k+1|k}C' + R_k \\
\hat{x}_{k+1} &= A\,\hat{x}_k + \Sigma_{k+1|k}C'S_{k+1}^{-1}(y_{k+1} - C\,A\,\hat{x}_k) \\
\Sigma_{k+1} &= \Sigma_{k+1|k} - \Sigma_{k+1|k}C'S_{k+1}^{-1}C\Sigma_{k+1|k}
\end{aligned} \tag{3.7}$$

The adversary then chooses its action as $\bar{a}_k = \phi(\Sigma_k)\,\hat{x}_k$ for some pre-specified function[5] ϕ. We measure the adversary's action as

$$a_k = \phi(\Sigma_k)\,\hat{x}_k + \epsilon_k, \quad \epsilon_k \sim \text{ iid } \mathcal{N}(0, \sigma_\epsilon^2) \tag{3.8}$$

The Kalman covariance Σ_k is deterministic and fully determined by the model parameters. Hence, we only need to estimate $\hat{x}_k$ at each time k given $a_{1:k}, x_{0:k}$ to estimate the belief $\pi_k = \mathcal{N}(\hat{x}_k, \Sigma_k)$. Substituting (3.4) for y_{k+1} in (3.7), we see that (3.7), (3.8)

[4]For localization problems, we will use the information filter form:

$$\Sigma_{k+1}^{-1} = \Sigma_{k+1|k}^{-1} + C'R^{-1}C, \quad \psi_{k+1} = \Sigma_{k+1}C'R^{-1} \tag{3.5}$$

Similarly, the inverse Kalman filter in information form reads

$$\bar{\Sigma}_{k+1}^{-1} = \bar{\Sigma}_{k+1|k}^{-1} + \bar{C}'_{k+1}\bar{R}^{-1}\bar{C}_{k+1},\ \bar{\psi}_{k+1} = \bar{\Sigma}_{k+1}\bar{C}'_{k+1}\bar{R}^{-1}. \tag{3.6}$$

[5]In general the action a_k is a function of the state estimate and covariance matrix. Choosing the action a_k as a linear function of the state estimate is for convenience and motivates the inverse Kalman filter discussed below. Moreover it mimics linear quadratic Gaussian (LQG) control where the feedback is a linear function of the state estimate. In LQG control, the feedback gain is obtained from backward Riccati equation. Here we weigh by a nonlinear function of the Kalman covariance matrix (forward Riccati equation) to allow for incorporating uncertainty of the estimate into the choice of the action a_k.

constitute a linear Gaussian system with unobserved state $\hat{x}_k$, observations a_k, and known exogenous input x_k:

$$\begin{aligned}
\hat{x}_{k+1} &= (\boldsymbol{I} - \psi_{k+1}C)\, A\hat{x}_k + \psi_{k+1}v_{k+1} + \psi_{k+1}Cx_{k+1} \\
a_k &= \phi(\Sigma_k)\, \hat{x}_k + \epsilon_k, \quad \epsilon_k \sim \text{ iid } \mathcal{N}(0, \sigma_\epsilon^2), \\
&\text{where } \psi_{k+1} = \Sigma_{k+1|k}C' S_{k+1}^{-1}.
\end{aligned} \tag{3.9}$$

ψ_{k+1} is called the Kalman gain and $\boldsymbol{I}$ is the identity matrix.

To summarize, our filtered estimate of the adversary's filtered estimate $\hat{x}_k$ given measurements $a_{1:k}, x_{0:k}$ is achieved by running "our" Kalman filter on the linear Gaussian state space model (3.9), where $\hat{x}_k, \psi_k, \Sigma_k$ in (3.9) are generated by the adversary's Kalman filter. Therefore, our Kalman filter uses the parameters

$$\begin{aligned}
&\bar{A}_k = (\boldsymbol{I} - \psi_{k+1}C)A,\ \bar{F}_k = \psi_{k+1}C,\ \bar{C}_k = \phi(\Sigma_k), \\
&\bar{Q}_k = \psi_{k+1}\, \psi_{k+1}',\ \bar{R}_k = \sigma_\epsilon^2
\end{aligned} \tag{3.10}$$

The equations of our inverse Kalman filter for estimating the adversary's estimate of our state are:

$$\begin{aligned}
\bar{\Sigma}_{k+1|k} &= \bar{A}_k\bar{\Sigma}_k\bar{A}_k' + \bar{Q}_k,\ \bar{S}_{k+1} = \bar{C}_{k+1}\bar{\Sigma}_{k+1|k}\bar{C}_{k+1}' + \bar{R}_k \\
\hat{\bar{x}}_{k+1} &= \bar{A}_k\, \hat{\bar{x}}_k + \bar{\Sigma}_{k+1|k}\bar{C}_{k+1}'\bar{S}_{k+1}^{-1} \times \left[a_{k+1} - \bar{C}_{k+1}\left(\bar{A}_k\hat{\bar{x}}_k + \bar{F}_k x_{k+1}\right)\right] + \bar{F}_k x_{k+1} \\
\bar{\Sigma}_{k+1} &= \bar{\Sigma}_{k+1|k} - \bar{\Sigma}_{k+1|k}\bar{C}_{k+1}'\bar{S}_{k+1}^{-1}\bar{C}_{k+1}\bar{\Sigma}_{k+1|k}
\end{aligned} \tag{3.11}$$

Note $\hat{\bar{x}}_k$ and $\bar{\Sigma}_k$ denote our conditional mean estimate and covariance of the adversary's conditional mean $\hat{x}_k$. The computational cost of the inverse Kalman filter is identical to the classical Kalman filter, namely $O(X^2)$ computations at each time step.

Remarks:

1. As discussed in [102], inverse Hidden Markov model (HMM) filters and inverse particle filters can also be derived to solve the inverse tracking problem. For

example, the inverse HMM filter deals with the case when π_k is computed via an HMM filter and the estimates of the HMM filter are observed in noise. In this case the inverse filter has a computational cost that grows exponentially with the size of the observation alphabet.

2. A general approximate solution for (3.3) involves sequential Markov chain Monte-Carlo (particle filtering). In particle filtering, cases where it is possible to sample from the so called optimal importance function are of significant interest [116, 117]. In inverse filtering, [102] shows that the optimal importance function can be determined explicitly due to the structure of the inverse filtering problem. Specifically, in our case, the "optimal" importance density is $\pi^* = p(\pi_k, y_k|\pi_{k-1}, y_{k-1}, x_k, a_k)$. Note that in our case

$$\pi^* = \; p(\pi_k|\pi_{k-1}, y_k)\, p(y_k|x_k, a_k) = \delta\big(\pi_k - T(\pi_{k-1}, y_k)\big)\, p(y_k|x_k) \qquad (3.12)$$

is straightforward to sample from. There has been a substantial amount of recent research in finite sample concentration bounds for the particle filter [118, 119]. In future work such results can be used to evaluate the sample complexity of the inverse particle filter.

3.2.3 Estimating the Adversary's Sensor Gain

In this section, we discuss how to estimate the adversary's sensor observation kernel B in (3.1) which quantifies the accuracy of the adversary's sensors. We assume that B is parameterized by an M-dimensional vector $\theta \in \Theta$ where Θ is a compact subset of $\mathbb{R}^M$. Denote the parameterized observation kernel as B^θ. Assume that both us and the adversary know [6]) P (state transition kernel and G (probabilistic map from adversary's

[6]Otherwise the adversary estimates P as $\hat{P}$ and we need to estimate the adversary's estimate of us, namely $\hat{\hat{P}}$. This makes the estimation task substantially more complex. In future work we will examine

belief to its action). As mentioned earlier, the stochastic kernel G in (3.1) is a composition of two stochastic kernels: 1) the map from the adversary's belief π_k to its action u_k, and 2) the map from the adversary's action u_k to our measurement a_k of this action.

Then, given our state sequence $x_{0:N}$ and adversary's action sequence $u_{1:N}$, our aim is to compute the maximum likelihood estimate (MLE) of θ. That is, with $L_N(\theta)$ denoting the log-likelihood, the aim is to compute

$$\theta^* = \underset{\theta\in\Theta}{\operatorname{argmax}}\, L_N(\theta),\ L_N(\theta) = \log p(x_{0:N}, a_{1:N}|\theta). \tag{3.13}$$

The likelihood can be evaluated from the un-normalized inverse filtering recursion (3.3)

$$\begin{aligned} L_N(\theta) &= \log \int_\Pi q_N^\theta(\pi) d\pi, \\ q_{k+1}^\theta(\pi) &= G_{\pi,a_{k+1}} \int_\Pi B^\theta_{x_{k+1},y^\theta_{\pi_k,\pi}}\, q_k^\theta(\pi_k) d\pi_k, \end{aligned} \tag{3.14}$$

initialized by setting $q_0^\theta(\pi_0) = \pi_0$. Here $y^\theta_{\pi_k,\pi}$ is the observation such that $\pi = T(\pi_k, y)$ where T is the adversary's filter (3.2) with variable B parametrized by θ. Given (3.14), a local stationary point of the likelihood can be computed using a general purpose numerical optimization algorithm.

3.2.4 Example. Estimating Adversary's Gain in Linear Gaussian case

The aim of this section is to provide insight into the nature of estimating the adversary's sensor gain via numerical examples. Consider the setup in Sec.3.2.2 where our dynamics are linear Gaussian and the adversary observes our state linearly in Gaussian noise (3.4). The adversary estimates our state using a Kalman filter, and we estimate the adversary's

conditions under which the MLE in this setup is identifiable and consistent.

estimate using the inverse Kalman filter (3.9). Using (3.9), (3.10), the log-likelihood for the adversary's observation gain matrix $\theta = C$ based on our measurements is[7]

$$L_N(\theta) = \text{const} - \frac{1}{2}\sum_{k=1}^{N} \log|\bar{S}_k^\theta| - \frac{1}{2}\sum_{k=1}^{N} \iota_k' \, (\bar{S}_k^\theta)^{-1} \, \iota_k$$

$$\iota_k = a_k - \bar{C}_k^\theta \, \bar{A}_{k-1}^\theta \hat{\hat{x}}_{k-1} - \bar{F}_{k-1}^\theta x_{k-1} \tag{3.15}$$

where ι_k are the innovations of the inverse Kalman filter (3.11). In (3.15), our state x_{k-1} is known to us and therefore is a known exogenous input. Also note from (3.10) that $\bar{A}_k, \bar{F}_k$ are explicit functions of C, while $\bar{C}_k$ and $\bar{Q}_k$ depend on C via the adversary's Kalman filter.

The log-likelihood for the adversary's observation gain matrix $\theta = C$ can be evaluated using (3.15). To provide insight, Figure 3.3 displays the log-likelihood versus adversary's gain matrix C in the scalar case for 1000 equally spaced data points over the interval $C = (0, 10]$. The four sub-figures correspond to true values $C^o = 2.5, 3.5$ of C, respectively.

Each sub-figure in Figure 3.3 has two plots. The plot in red is the log-likelihood of $\hat{C} \in (0, 10]$ evaluated based on the adversary's observations using the standard Kalman filter. (This is the classical log-likelihood of the observation gain of a Gaussian state space model.) The plot in blue is the log-likelihood of $C \in (0, 10]$ computed using our measurements of the adversary's action using the inverse Kalman filter (where the adversary first estimates our state using a Kalman filter) - we call this the inverse case.

Figure 3.3 shows that the log-likelihood in the inverse case (blue plots) has a less pronounced maximum compared to the standard case (red plots). Therefore, numerical algorithms for computing the MLE of the adversary's gain C^o using our observations of the adversary's actions (via the inverse Kalman filter) will converge much more slowly than the classical MLE (based on the adversary's observations). This is intuitive since

[7]The variable θ is introduced only for notational clarity.

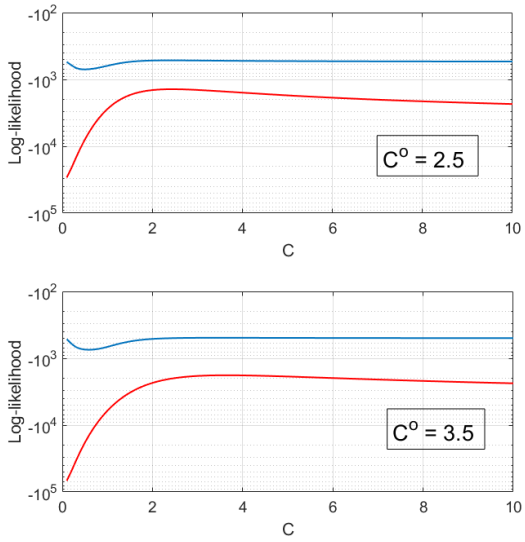

Figure 3.3: Log-Likelihood as a function of adversary's gain $C \in (0, 10]$ when true value is C^o. The red curves denote the log-likelihood of C given the adversary's measurements of our state. The blue curves denote the log-likelihood of C using the inverse Kalman filter given our observations of the adversary's action u_k. The plots show that it is more difficult to compute the MLE (3.13) for the inverse filtering problem due to the almost flat likelihood (blue curves) compared to red curves.

our estimate of the adversary's parameter is based on the adversary's estimate of our state and so has more noise.

Sensitivity of MLE. It is important to evaluate the sensitivity of the MLE of C wrt covariance matrices $Q_k,\ R_k$ in the state space model (3.4). For example, the sensitivity wrt Q_k reveals how sensitive the MLE is wrt our maneuver covariance since from (3.4), Q_k determines our maneuvers. Our sensitivity analysis evaluates the variation of the second derivative of the log-likelihood of C computed at the true gain C^o to small changes in Q_k and R_k. Table 3.1 displays our sensitivity results wrt the scalar setup of Figure 3.3. Table 3.1 comprises two sensitivity values,

$$\begin{aligned} \eta_Q &= \frac{\partial}{\partial Q_k}\left(\frac{\partial^2 L_N(\theta)}{\partial \theta^2}\right)\Bigg|_{\theta=C^o} \text{ and} \\ \eta_R &= \frac{\partial}{\partial R_k}\left(\frac{\partial^2 L_N(\theta)}{\partial \theta^2}\right)\Bigg|_{\theta=C^o}, \end{aligned} \tag{3.16}$$

evaluated for both the inverse case (that uses the inverse Kalman filter (3.15)) and the

	C^o	Classic	Inverse
η_Q	2.5	-43.45	-6.46
	3.5	-25.16	-2.77
η_R	2.5	-189.39	-50.04
	3.5	-65.27	-30.55

Table 3.1: Comparison of sensitivity values (3.16) for log-likelihood of C wrt noise covariances $Q_k,\ R_k$ (3.4) - classical model vs inverse Kalman filter model.

classic case where the adversary's observations are known. $\eta_{(\cdot)}$ measures the change in the sharpness of the log-likelihood plot around the true sensor gain wrt change in the noise covariance. Note that the experimental setup of Figure 3.3 assumes the covariances Q_k, R_k are constant over time index k, hence we drop the subscript in the LHS of (3.16).

Table 3.1 shows that the second derivative of the log-likelihood is more sensitive (in magnitude) to the adversary's observation covariance R_k than the maneuver covariance Q_k. Also, it is observed that the sensitivity of the log-likelihood is higher for lower sensor gain C^o. This observation is consistent with intuition since a larger gain C implies a larger SNR (signal-to-noise ratio) of the observation y_k which intuitively suggests the estimate of C is more robust to changes in maneuver covariance and observation noise covariance.

Cramér-Rao (CR) bounds. It is instructive to compare the CR bounds for MLE of C for the classic model versus that of the inverse Kalman filter model. Table 3.2 displays the CR bounds (reciprocal of expected Fisher information) for the four examples considered above evaluated using via the algorithm in [120]. It shows that the covariance lower bound for the inverse case is substantially higher than that for the classic case. This is consistent with the intuition that estimating the adversary's parameter based on its actions (which is based on its estimate of us) is more difficult than directly estimating C in a classical state space model based on the adversary's observations of our state that determines its actions.

C^o	Classic	Inverse
0.5	0.24×10^{-3}	5.3×10^{-3}
1.5	1.2×10^{-3}	37×10^{-3}
2	2.1×10^{-3}	70×10^{-3}
3	4.6×10^{-3}	336×10^{-3}

Table 3.2: Comparison of Cramér-Rao bounds for C - classical model vs inverse Kalman filter model.

Consistency of MLE. The above example (Figure 3.3) shows that the likelihood surface of $L_N(\theta) = \log p(x_{0:N}, a_{1:N}|\theta)$ is flat and hence computing the MLE numerically can be difficult. Even in the case when we observe the adversary's actions perfectly, [104] shows that non-trivial observability conditions need to be imposed on the system parameters.

For the linear Gaussian case where we observe the adversary's Kalman filter in noise, strong consistency of the MLE for the adversary's gain matrix C can be established fairly straightforwardly. Specifically, if we assume that state matrix A is stable, and the state space model is an identifiable minimal realization, then the adversary's Kalman filter variables converge to steady state values geometrically fast in k [121] implying that asymptotically the inverse Kalman filter system is stable linear time invariant. Then, the MLE θ^* for the adversary's observation matrix C is unique and strongly consistent [122].

3.3 Identifying Utility Maximization in a Cognitive Radar

The previous section was concerned with estimating the adversary's posterior belief and sensor accuracy. This section discusses detecting utility maximization behavior and estimating the adversary's utility function in the context of cognitive radars. As described in the introduction, inverse tracking, identifying utility maximization and designing interference to confuse the radar constitute our adversarial setting.

Cognitive radars [123] use the perception-action cycle of cognition to sense the environment and learn from it relevant information about the target and the environment. The cognitive radars then tune the radar sensor to optimally satisfy their mission objectives. Based on its tracked estimates, the cognitive radar adaptively optimizes its waveform, aperture, dwell time and revisit rate. In other words, a cognitive radar is a constrained utility maximizer.

This section is motivated by the next logical step, namely, *identifying a cognitive radar* from the actions of the radar. The adversary cognitive radar observes our state in noise; it uses a Bayesian estimator (target tracking algorithm) to update its posterior distribution of our state and then chooses an action based on this posterior. From the intercepted emissions of an adversary's radar, we address the following question: Are the adversary sensor's actions consistent with optimizing a monotone utility function (i.e., is the cognitive sensor behavior rational in an economics sense)? If so how can we estimate a utility function of the adversary's cognitive sensor that is consistent with its actions? The main synthesis/analysis framework we will use is that of revealed preferences [32, 124, 125] from microeconomics which aims to determine preferences by observing choices. The results presented below are developed in detail in the recent work [103]; however, the SINR constraint formulation in Sec. 3.3.3 for detecting waveform optimization is new. Related work that develops adversarial inference strategies at a higher level of abstraction than tracking level (like the Bayesian filter level in Sec. II) includes [126]. [126] places counter unmanned autonomous systems at a level of abstraction above the physical sensors/actuators/weapons and datalink layers; and below the human controller layer.

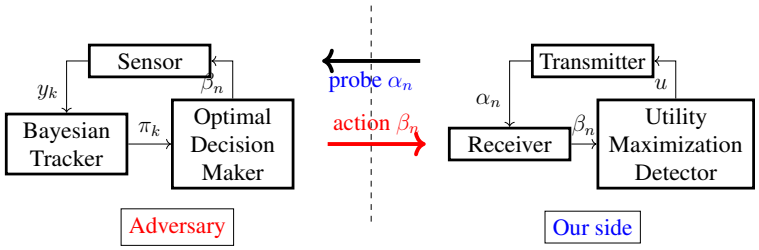

Figure 3.4: Schematic of Adversarial Inference Problem. Our side is a drone/UAV or electromagnetic signal that probes the adversary's cognitive radar system. k denotes a fast time scale and n denotes a slow time scale. Our state x_k, parameterized by α_n (purposeful acceleration maneuvers), probes the adversary radar. Based on the noisy observation y_k of our state, the adversary radar responds with action β_n. Our aim (in the Utility Maximization Detector block) is to detect if the adversary radar is economic rational, i.e., is its response β_n generated by constrained optimizing a utility function u, and if so, estimate the utility function.

3.3.1 Background. Revealed Preferences and Afriat's Theorem

Non-parametric detection of utility maximization behavior is studied in the area of revealed preferences in microeconomics. A key result is the following:

Definition 18 ([127, 128]) *A system is a* utility maximizer *if for every probe* $\alpha_n \in \mathbb{R}^m_+$, *the response* $\beta_n \in \mathbb{R}^m$ *satisfies*

$$\beta_n \in \underset{\alpha_n'\beta \leq 1}{\operatorname{argmax}}\, U(\beta) \tag{3.17}$$

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

In economics, α_n is the price vector and β_n the consumption vector. Then $\alpha_n'\beta \leq 1$ is a natural budget constraint[8] for a consumer with 1 dollar. Given a dataset of price and consumption vectors, the aim is to determine if the consumer is a utility maximizer (rational) in the sense of (3.17).

The key result is the following theorem due to Afriat [32, 125, 127–129]

[8]The budget constraint $\alpha_n'\beta \leq 1$ is without loss of generality, and can be replaced by $\alpha_n'\beta \leq c$ for any positive constant c. A more general nonlinear budget incorporating spectral constraints will be discussed below.

Theorem 19 (Afriat's Theorem [127]) *Given a data set* $\mathcal{D} = \{(\alpha_n, \beta_n), n \in \{1, 2, \ldots, N\}\}$, *the following statements are equivalent:*

1. *The system is a utility maximizer and there exists a monotonically increasing, continuous, and concave utility function that satisfies (3.17).*
2. *There exist positive reals* $u_t, \lambda_t > 0,\ t = 1, 2, \ldots, N$, *such that the following inequalities hold.*

$$u_s - u_t - \lambda_t \alpha_t'(\beta_s - \beta_t) \leq 0 \ \forall t, s \in \{1, 2, \ldots, N\}. \tag{3.18}$$

   The monotone, concave utility function[9] *given by*

$$U(\beta) = \min_{t \in \{1,2,\ldots,N\}} \{u_t + \lambda_t \alpha_t'(\beta - \beta_t)\} \tag{3.19}$$

   constructed using u_t *and* λ_t *defined in* (3.18) *rationalizes the dataset by satisfying* (3.17).
3. *The data set* $\mathcal{D}$ *satisfies the Generalized Axiom of Revealed Preference (GARP) also called cyclic consistency, namely for any* $t \leq N$, $\alpha_t'\beta_t \geq \alpha_t'\beta_{t+1} \quad \forall t \leq k-1 \implies \alpha_k'\beta_k \leq \alpha_k'\beta_1$.

Afriat's theorem tests for economics-based rationality; its remarkable property is that it gives a *necessary and sufficient condition* for a system to be a utility maximizer based on the system's input-output response. Although GARP in statement 4 in Theorem 19 is not critical to the developments in this chapter, it is of high significance in micro-economic theory and is stated here for completeness. The feasibility of the set of inequalities (3.18) can be checked using a linear programming solver; alternatively GARP can be checked using Warshall's algorithm with $O(N^3)$ computations [130, 131].

[9] As pointed out in [130], a remarkable feature of Afriat's theorem is that if the dataset can be rationalized by a monotone utility function, then it can be rationalized by a continuous, concave, monotonic utility function. Put another way, continuity and concavity cannot be refuted with a finite datasaset.

The recovered utility using (3.19) is not unique; indeed any positive monotone increasing transformation of (3.19) also satisfies Afriat's Theorem; that is, the utility function constructed is ordinal. This is the reason why the budget constraint $\alpha_n'\beta \leq 1$ is without generality; it can be scaled by an arbitrary positive constant and Theorem 19 still holds. In signal processing terminology, Afriat's Theorem can be viewed as set-valued system identification of an *argmax* system; set-valued since (3.19) yields a set of utility functions that rationalize the finite dataset $\mathcal{D}$.

3.3.2 Beam Allocation: Revealed Preference Test

This section constructs a test to identify a cognitive radar that switches its beam adaptively between targets. This example is based on [103] and is presented here for completeness The setup is schematically shown in Figure 3.4. We view each component i of the probe signal $\alpha_n(i)$ as the trace of the information matrix (inverse covariance) of target i. We use the trace of the information matrix of each target in our probe signal – this allows us to consider multiple targets. Since the adversarial radar is assumed to be stationary, the target covariance used to define the probe for the radar is indeed the maneuver covariance.

The setup in Figure 3.4 differs significantly from the setup of Figure 3.2 considered in the previous section. First, the adversary in the current setup is an economically rational agent. In Figure 3.2, the adversary is only specified at a lower level of abstraction as using a Bayesian filter to track our maneuvers. Second, this section abstracts adversary's actions at the fast time scale indexed by k by an appropriately defined response at the slow time scale indexed by n. The previous section's analysis was confined to the actions generated only at the fast time scale k. Lastly, Figure 3.4 assumes the abstracted response β_k of the adversary is measured accurately by us as opposed to a noisy measurement a_k

of the adversary's action u_k in Figure 3.2.

Suppose a radar adaptively switches its beam between m targets where these m targets are controlled by us. As in (3.4), on the fast time scale indexed by k, each target i has linear Gaussian dynamics and the adversary radar obtains linear Gaussian measurements:

$$\begin{aligned} x_{k+1}^i &= A\, x_k^i + w_k^i, \quad x_0 \sim \pi_0 \\ y_k^i &= C\, x_k^i + v_k^i, \quad i = 1, 2, \ldots, m \end{aligned} \tag{3.20}$$

Here $w_k^i \sim \mathcal{N}(0, Q_n(i))$, $v_k^i \sim \mathcal{N}(0, R_n(i))$. Recall from Figure 3.4 that n indexes the epoch (slow time scale) and k indexes the fast time scale within the epoch. We assume that both $Q_n(i)$ and $R_n(i)$ are known to us and the adversary.

The adversary's radar tracks our m targets using Kalman filter trackers. The fraction of time the radar allocates to each target i in epoch n is $\beta_n(i)$. The price the radar pays for each target i at the beginning of epoch n is the trace of the predicted *accuracy* of target i. Recall that this is the trace of the inverse of the predicted covariance at epoch n using the Kalman predictor

$$\alpha_n(i) = \mathrm{tr}(\Sigma_{n|n-1}^{-1}(i)), \quad i = 1, \ldots, m \tag{3.21}$$

The predicted covariance $\Sigma_{n|n-1}(i)$ is a deterministic function of the maneuver covariance $Q_n(i)$ of target i. So the probe[10] $\alpha_n(i)$ is a signal that we can choose, since it is a deterministic function of the maneuver covariance $Q_n(i)$ of target i. We abstract the target's covariance by its trace denoted by $\alpha_n(i)$. Note also that the observation noise covariance $R_n(i)$ depends on the adversary's radar response $\beta_n(i)$, i.e., the fraction of time allocated to target i. We assume that each target i can estimate the fraction of time $\beta_n(i)$ the adversary's radar allocates to it using a radar detector.

[10]In comparison to (3.4), the velocity and acceleration elements of x_k^i in (3.20) must be multiplied by normalization factors Δt and $(\Delta t)^2$ respectively, for (3.21) to be dimensionally correct, where Δt is the time duration between two discrete time instants on the fast time scale.

Given the time series α_n, β_n, $n = 1, \ldots, N$, our aim is to detect if the adversary's radar is cognitive. We assume that a cognitive radar optimizes its beam allocation as the following constrained optimization:

$$\beta_n = \underset{\beta}{\operatorname{argmax}}\ U(\beta), \ \text{ s.t. } \beta' \alpha_n \leq p_*, \tag{3.22}$$

where $U(\cdot)$ is the adversary radar's utility function (unknown to us) and $p_* \in \mathbb{R}_+$ is a pre-specified average accuracy of all m targets.

The economics-based rationale for the budget constraint is natural: For targets that are cheaper (lower accuracy $\alpha_n(i)$), the radar has incentive to devote more time $\beta_n(i)$. However, given its resource constraints, the radar can achieve at most an average accuracy of p_* over all targets.

The setup in (3.22) is directly amenable to Afriat's Theorem 19. Thus (3.18) can be used to test if the radar satisfies utility maximization in its beam scheduling (3.22) and also estimate the set of utility functions (3.19). Furthermore (as in Afriat's theorem) since the utility is ordinal, p_* can be chosen as 1 without loss of generality (and therefore does not need to be known by us).

3.3.3 Waveform adaptation: Revealed Preference Test for Non-linear budgets

In the previous subsection, we tested for cognitivity of a radar by viewing it as an abstract system that switches its beam adaptively between targets. Here, we discuss cognitivity with respect to waveform design. Specifically, we construct a test to identify cognitive behavior of an adversary radar that optimizes its waveform based on the SINR of the target measurement. By using a generalization of Afriat's theorem (Theorem 19) to

non-linear budgets, our main aim is to detect if a radar intelligently chooses its waveform to maximize an underlying utility subject to signal processing constraints. Our setup below differs from [103] since we introduce the SINR as a nonlinear budget constraint; in comparison [103] uses a spectral budget constraint.

We start by briefly outlining the generalized utility maximization setup.

Definition 20 ([124]) *A system is a* generalized utility maximizer *if for every probe* $\alpha_n \in \mathbb{R}_+^m$*, the response* $\beta_n \in \mathbb{R}^m$ *satisfies*

$$\beta_n \in \underset{g_n(\beta)\leq 0}{\operatorname{argmax}} U(\beta) \tag{3.23}$$

where $U(\beta)$ *is a* monotone *utility function and* $g_n(\cdot)$ *is monotonically increasing in* β*.*

The above utility maximization model generalizes Definition 18 since the budget constraint $g_n(\beta) \leq 0$ can accommodate non-linear budgets and includes the linear budget constraint of Definition 18 as a special case. The result below provides an explicit test for a system that maximizes utility in the sense of Definition 20 and constructs a set of utility functions that rationalizes the decisions β_n of the utility maximizer.

Theorem 21 (Test for rationality with nonlinear budget [124]) *Let* $B_n = \{\beta \in \mathbb{R}_+^m | g_n(\beta) \leq 0\}$ *with* $g_n : \mathbb{R}^m \to \mathbb{R}$ *an increasing, continuous function and* $g_n(\beta_n) = 0$ *for* $n = 1, \ldots N$*. Then the following conditions are equivalent:*

1) There exists a monotone continuous utility function U *that rationalizes the data set* $\{\beta_n, B_n\}, n = 1, \ldots N$*. That is*

$$\beta_n = \underset{\beta}{\operatorname{argmax}} U(\beta), \quad g_n(\beta) \leq 0$$

2) There exist positive reals $u_t, \lambda_t > 0,\ t = 1, 2, \ldots, N$*, such that the following inequalities hold.*

$$u_s - u_t - \lambda_t g_t(\beta_s) \leq 0 \quad \forall t, s \in \{1, 2, \ldots N\} \tag{3.24}$$

The monotone, concave utility function given by

$$U(\beta) = \min_{t \in \{1,\dots N\}} \{u_t + \lambda_t g_t(\beta)\} \tag{3.25}$$

constructed using u_t and λ_t defined in (3.24) rationalizes the data set by satisfying (3.23).

3) The data set $\{\beta_n, B_n\}, n = 1, \dots, N$ satisfies $GARP$:

$$g_t(\beta_j) \leq g_t(\beta_t) \implies g_j(\beta_t) \geq 0 \tag{3.26}$$

Like Afriat's theorem, the above result provides a *necessary and sufficient condition* for a system to be a utility maximizer based on the system's input-output response. In spite of a non-linear budget constraint, it can be easily verified that the constructed utility function $U(\beta)$ (3.25) is ordinal since any positive monotone increasing transformation of (3.25) satisfies the GARP inequalities (3.26).

We now justify the non-linear budget constraint in (3.23) in the context of the cognitive radar by formulating an optimization problem the radar solves equivalent to Definition 20. Suppose we observe the radar over $n = 1, 2, \dots, N$ time epochs (slow varying time scale). At the n^{th} epoch, we probe the radar with an interference vector $\alpha_n \in \mathbb{R}^M$. The radar responds with waveform $\beta_n \in \mathbb{R}^M_+$. We assume the chosen waveform β_n maximizes the radar's underlying utility function while ensuring the radar's SINR exceeds a particular threshold $\delta > 0$, where the SINR of the radar given probe α and response β is defined as

$$\text{SINR}(\alpha, \beta) = \frac{\beta' Q \beta}{\beta' P(\alpha) \beta + \gamma}. \tag{3.27}$$

In (3.27), the radar's signal power (numerator) and interference power (first term in denominator) are assumed to be quadratic forms of $Q, P(\alpha)$ respectively, where $Q, P(\alpha) \in \mathbb{R}^{M \times M}$ are positive definite matrices known to us. The term $\gamma > 0$ is the noise power. The SINR definition in (3.27) is a more general formulation of the

SCNR (3.33) of a cognitive radar derived in Sec. 3.4 using clutter response models [132]. The matrices $Q, P(\alpha)$ are analogous to the covariance of the channel impulse response matrices $H_t(\cdot)$ and $H_p(\cdot)$ corresponding to the target and clutter (external interference) channels, respectively (see Sec. 3.4.1 for a discussion).

Having defined the SINR above in (3.27), we now formalize the radar's response β_n given probe α_n, $n = 1, 2, \ldots$ as the solution of the following constrained optimization problem.

$$\beta_n \in \underset{\beta}{\operatorname{argmax}}\, U(\beta), \text{ s.t. } \mathrm{SINR}(\alpha_n, \beta) \geq \delta \tag{3.28}$$

Clearly, the above setup falls under the non-linear utility maximization setup in Definition 20 by defining the non-linear budget $g_n(\cdot)$ as $g_n(\beta) = \delta - SIR(\alpha_n, \beta)$ where $SIR(\cdot)$ is defined in (3.27). It only remains to show that this definition of $g_n(\beta)$ is monotonically increasing in β. Theorem 22 stated below establishes two conditions that are sufficient for $g_n(\beta)$ to be monotonically increasing in β.

Theorem 22 *Suppose that the adversary radar uses the SINR constraint* (3.28)*. Then* $g_n(\beta) = \delta - \mathrm{SIR}(\alpha_n, \beta)$ *is monotonically increasing in* β *if the following two conditions hold.*

1. *The matrix* Q *is a diagonal matrix with off-diagonal elements equal to zero.*
2. *The matrix* $\left(\frac{c_{P(\alpha_n)}}{d_Q} P(\alpha_n) - Q\right)$ *is component-wise less than* 0 *for all* $n \in \{1, 2, \ldots N\}$*, where* $c_{P(\alpha_n)} > 0$ *and* $d_Q > 0$ *denote the smallest and largest eigenvalues of* $P(\alpha_n)$ *and* Q *respectively.*

The proof of Theorem 22 follows from elementary calculus and omitted for brevity. Hence, assuming the two conditions hold in Theorem 22 above, we can use the results from Theorem 21 to test if the radar satisfies utility maximization in its waveform

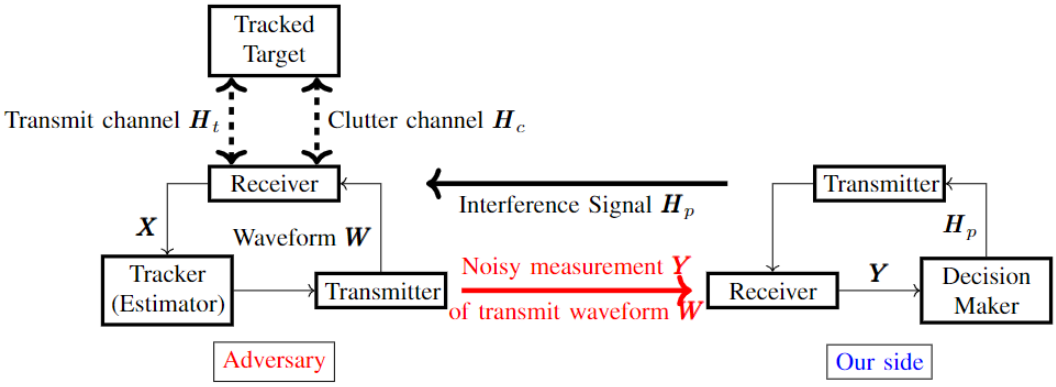

Figure 3.5: Schematic of transmit channel H_t, clutter channel H_c and interference signal $\boldsymbol{H}_p$ involving an adversarial cognitive radar and us. We observe the radar's waveform $\boldsymbol{W}$ in noise. The aim is to engineer the interference signal $\boldsymbol{H}_p$ to confuse the adversary cognitive radar.

design (3.28) and also estimate the set of feasible utility functions $U(\cdot)$ (3.28) that rationalizes the radar's responses $\{\beta_n\}$.

3.4 Designing Smart Interference To Confuse Cognitive Radar

This section discusses how we can engineer (design) external interference (a probing signal) at the physical layer level to confuse a cognitive radar. By abstracting the probing signal to a channel in the frequency domain, our objective is to minimize the signal power of the interference generated by us while ensuring the SCNR of the radar does not exceed a pre-defined threshold. The setup is schematically shown in Figure 3.5. Note that the level of abstraction used in this section is at the Wiener filter pulse/waveform level; whereas the previous two sections were at the systems level (which uses the utility maximization framework) and tracker level (which uses the Kalman filter formalism), respectively. This is consistent with the design theme of sense globally (high level of abstraction) and act locally (lower level of abstraction).

As can be seen in the SCNR expression (3.33), the interference signal power manifests as additional clutter perceived by the radar in the denominator thus forcing the SCNR to go down. The radar then re-designs its waveform to maximize its SCNR given our interference signal. We observe the adversarial radar's chosen waveform in noise. Our task can thus be re-formulated as choosing the interference signal with minimal power while ensuring that with probability at least $1-\epsilon$, the optimized SCNR lies below a threshold level Δ. Here , ϵ and Δ are user-defined parameters. This approach closely follows the formulation in Sec. 3.3.3 where the cognitive radar chooses the optimal waveform while ensuring the SINR exceeds a threshold value. Further, the SCNR of the adversary's radar defined in (3.34) below can be interpreted as a monotone function of the radar's utility function in the abstracted setup of Sec. 3.3.3, since in complete analogy to the utility maximization model of Sec. 3.3.3, this section assumes the radar maximizes its SCNR in the presence of smart interference signals (probes).

3.4.1 Interference Signal Model

We first characterize how a cognitive radar optimally chooses its waveform based on its perceived interference. The radar's objective is to choose the optimal waveform that maximizes its signal-to-interference-plus-noise (SINR) ratio.

Suppose we observe the adversarial radar over $l = 1, 2 \ldots L$ pulses, where each pulse comprises $n = 1, 2, \ldots N$ discrete time steps. A single-input single-output (SISO) radar system has two channel impulse responses, one for the target and the other for clutter. Let $w(n)$ denote the radar transmit waveform and $h_t(n)$, $h_c(n)$ denote the target and clutter channel impulse responses, respectively. Then, the radar measurements corresponding to

the l-th pulse can be expressed as

$$x(n,l) = h_t(n,l) \circledast w(n,l) + h_c(n,l) \circledast w(n,l) + e_r(n,l) \tag{3.29}$$

where $\circledast$ represents a convolution operator and $e_r(n,l)$ is the radar measurement noise modeled as an i.i.d random variable with zero mean and known variance σ_r^2. We model the radar's measurement using the stochastic Green's function impulse response model presented in [132], where the radar's electromagnetic channel is modeled using a physics based impulse response.

Since convolution in the time domain can be expressed as multiplication in the frequency domain (with notation in upper case), we can express the measurements in the frequency domain as follows:

$$X(k,l) = H_t(k,l)W(k,l) + H_c(k,l)W(k,l) + E_r(k,l) \tag{3.30}$$

where $k \in \mathcal{K} = \{1,\ldots,K\}$ is the frequency bin index. Eq. (3.30) can be extended to an $I \times J$ MIMO radar and the received signal at the j-th receiver is given by

$$X_j(k,l) = \sum_{i=1}^{I} H_{t_{ij}}(k,l)W_i(k,l) + H_{c_{ij}}(k,l)W_i(k,l) + E_{r,j}(k,l), \tag{3.31}$$

$\forall k \in \{1,\ldots K\}$. Using matrices and vectors obtained by stacking and concatenating (3.31) for all i, j, and k, the MIMO radar measurement model at the l^{th} pulse in vector-matrix form can be expressed as

$$\boldsymbol{X}(l) = \boldsymbol{H}_t(l)\boldsymbol{W}(l) + \boldsymbol{H}_c(l)\boldsymbol{W}(l) + \boldsymbol{E}_r(l) \tag{3.32}$$

where $\boldsymbol{X}(l) \in \mathbb{C}^{(J\times K)\times 1}$ is the received signal vector, $\boldsymbol{H}_c(l), \boldsymbol{H}_t(l) \in \mathbb{C}^{(J\times K)\times(I\times J\times K)}$ are the effective transmit and clutter channel impulse response matrices respectively, $\boldsymbol{W}(l) \in \mathbb{C}^{(I\times J\times K)\times 1}$ is the radar's effective waveform vector. $\boldsymbol{E}_r(l) \in \mathbb{C}^{(J\times K)\times 1}$ is the effective additive noise vector modeled as a zero mean i.i.d random variable (independent over pulses) with covariance matrix $C_r \in \mathbb{R}^{(J\times K)\times(J\times K)}, C = (\sigma_r^2/K)\boldsymbol{I} = \tilde{\sigma}_r^2\boldsymbol{I}$. The block diagram in Fig. 3.5 shows the entire procedure for this model.

3.4.2 Smart Interference to confuse the adversary radar

The aim of this section is to design optimal interference signals (to confuse the adversary cognitive radar) by solving a probabilistically constrained optimization problem.

At the beginning of the l^{th} pulse, the adversary radar transmits a pilot signal to estimate the transmit and clutter channel impulse responses $\boldsymbol{H}_t(l)$ and $\boldsymbol{H}_c(l)$ respectively. Assuming it has a perfect estimate of $\boldsymbol{H}_t(l)$ and $\boldsymbol{H}_c(l)$, the radar then chooses the optimal waveform $\boldsymbol{W}^*(l)$ such that SCNR defined below in (3.33) is maximized. The radar's waveform $\boldsymbol{W}^*(l)$ is the solution to the following optimization problem

$$\boldsymbol{W}^*(l) = \underset{\boldsymbol{W}(l):\|\boldsymbol{W}(l)\|_2=1}{\operatorname{argmax}} \mathrm{SCNR}(\boldsymbol{H}_t(l), \boldsymbol{H}_c(l), \boldsymbol{W}(l)), \tag{3.33}$$

where the SCNR is defined as

$$\mathrm{SCNR}(\boldsymbol{H}_t, \boldsymbol{H}_c, \boldsymbol{W}) = \frac{\left\| \boldsymbol{H}_t \boldsymbol{W} \right\|_2^2}{\mathbb{E}\Big\{ \left\| \boldsymbol{H}_c \boldsymbol{W} + \boldsymbol{E}_r \right\|_2^2 \Big\}}. \tag{3.34}$$

Denote the maximum SCNR achieved in (3.33) as

$$\mathrm{SCNR}_{\max}(\boldsymbol{H}_t(l), \boldsymbol{H}_c(l), \sigma_r^2) = \mathrm{SCNR}(\boldsymbol{H}_t(l), \boldsymbol{H}_c(l), \boldsymbol{W}^*(l)). \tag{3.35}$$

Given $\boldsymbol{H}_t(l), \boldsymbol{H}_c(l)$ and the radar's measurement noise power σ_r^2, the radar generates an optimal waveform at the l^{th} pulse using (3.33) as the solution to the following eigenvector problem [100]:

$$\begin{aligned} \boldsymbol{A}\, \boldsymbol{W}^\star(l) &= \lambda_l \boldsymbol{W}^\star(l) \\ \boldsymbol{A} &= \Big(\big(\boldsymbol{H}_c(l)' \boldsymbol{H}_c(l) + \tilde{\sigma}_r^2 \boldsymbol{I} \big)^{-1} \boldsymbol{H}_t(l)' \boldsymbol{H}_t(l) \Big), \end{aligned}$$

Here $(\cdot)'$ denotes the Hermitian transpose operator.

As an external observer, we send a sequence of probe signals $P = \{\boldsymbol{H}_p(l), l \in \{1, 2, \ldots L\}\}$ over L pulses to confuse the adversary radar and degrade its performance.

The interference signal $\boldsymbol{H}_p(l-1)$ at the $(l-1)^{th}$ affects only radar's clutter channel impulse response $\boldsymbol{H}_c(l)$ at the l^{th} pulse which subsequently results in change of optimal waveform (3.33) chosen by the radar $\boldsymbol{W}^*(l)$. We measure the optimal waveform at the l^{th} pulse in noise as $\boldsymbol{Y}(l)$. We assume constant transmit and clutter channel impulse responses $\boldsymbol{H}_t, \boldsymbol{H}_c$ in the absence of the probe signals P. The dynamics of our interaction with the adversary radar due to probe P are as follows and schematically shown in Fig. 3.6:

$$\boldsymbol{H}_c(l) = \boldsymbol{H}_c + \boldsymbol{H}_p(l-1) \tag{3.36}$$

$$\boldsymbol{H}_t(l) = \boldsymbol{H}_t \tag{3.37}$$

$$\Big(\big(\boldsymbol{H}_c(l)'\boldsymbol{H}_c(l) + \tilde{\sigma}_r^2\boldsymbol{I}\big)^{-1}\boldsymbol{H}_t(l)'\boldsymbol{H}_t(l)\Big)\boldsymbol{W}^*(l)$$

$$= \lambda_l \boldsymbol{W}^*(l) \tag{3.38}$$

$$\boldsymbol{Y}(l) = \boldsymbol{W}^*(l) + \boldsymbol{E}_o(l). \tag{3.39}$$

In (3.39), $E_o(l)$ is our measurement noise modeled as a zero mean i.i.d random variable (independent over pulses) sampled from a known pdf f_o with zero mean and covariance $C_o = (\sigma_o^2/K)\boldsymbol{I} = \tilde{\sigma}_o^2\boldsymbol{I}$.

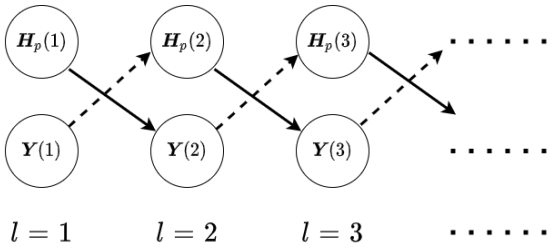

Figure 3.6: Schematic of smart interference design to confuse the cognitive radar. The interference signal at the l^{th} pulse affects the waveform choice of the radar in the $(l+1)^{th}$ pulse. We record the noisy waveform measurement $\boldsymbol{Y}(l+1)$ and generate the interference signal for the $(l+2)^{th}$ pulse.

Our objective is to optimally design the probe signals $P^* = \{\boldsymbol{H}_p^*(l), l \in \{1, \dots L\}\}$ that minimizes the interference signal power such that for a pre-defined $\Delta > 0$, there

exists $\epsilon \in [0,1)$ such that the probability the SCNR of the radar lies below Δ exceeds $(1-\epsilon)$, for all $l = 1, 2, \ldots L$.

$$\begin{aligned} &\min_{\{\boldsymbol{H}_p(l), l\in\{1,2,\ldots L\}\}} \sum_{l=1}^{L} \boldsymbol{H}_p(l)'\boldsymbol{H}_p(l) \\ &\text{s.t. } \mathbb{P}_{f_o}(\mathrm{SCNR}(\boldsymbol{H}_t(l), \boldsymbol{H}_c(l), \boldsymbol{Y}(l)) \leq \Delta) \geq 1-\epsilon, \\ &\qquad \forall l \in \{1,2,3,\ldots L\}. \end{aligned} \tag{3.40}$$

Here, $\mathbb{P}_f(\cdot)$ denotes the probability wrt pdf f. The design parameter Δ is the SCNR (performance) upper bound of the cognitive radar. To confuse the radar, our task is to ensure the SCNR of the radar is less than Δ with probability at least $1-\epsilon$. Hence, ϵ is the maximum probability of failure to confuse the radar with our smart interference signals. Although not explicitly shown, the $\mathrm{SCNR}_{\max}$ expression in (3.40) depends on our interference signal $\boldsymbol{H}_p$ as depicted in (3.36).

Solving the non-convex optimization problem (3.40) is challenging except for trivial cases. It involves two inter-related components (i) Estimating the transmit and clutter channel impulse responses $\boldsymbol{H}_t, \boldsymbol{H}_c$ from observation $\boldsymbol{Y}(l)$ and (ii) Using the estimated value of channel impulse responses to generate the interference signal $\boldsymbol{H}_p(l)$. Moreover, solving for $\boldsymbol{H}_c$ and $\boldsymbol{H}_t$ from recursive equations (3.36) through (3.39) for $l = 1, \ldots, L$ is a challenging problem since it does not have an analytical closed form solution.

With the above formulation, we can now discuss construction of smart inference to confuse the adversary radar. The cognitive radar maximizes its energy in the direction of its target impulse response and transfer function. As soon as we have an accurate estimate of the target channel transfer function from the L pulses, we can immediately generate signal dependent interference that nulls the target returns. Even if the clutter channel impulse response changes after we perform our estimation, since the target channel is stationary for longer durations, the signal dependent interference will work successfully

for several pulses after we compute the estimate. The main take away from this approach is that we are exploiting the fact that the cognitive radar provides information about its channel by optimizing the waveform with respect to its environment.

3.4.3 Numerical example illustrating design of smart interference

We conclude this section with a numerical example that illustrates the smart interference framework developed above. The simulation setup is as follows:

- $L = 2$ pulses (optimization horizon in (3.40)).
- Impulse response matrices for transmit channel $\boldsymbol{H}_t = [7\ 7]$, clutter channel $\boldsymbol{H}_c = [1\ 1]$, and adversary radar noise covariance $\tilde{\sigma}_r^2 = 1$ (3.32).
- *Design parameters:* SCNR upper bound $\Delta = \{2.8, 3, 3.2\}$, minimum probability of success
  $\epsilon = 0.2, 0.3$ (3.40).
- Probe signals for pulse index:

$$l = 1,\ \boldsymbol{H}_p(1) = [0.2r\ 0.5r],\ l = 2,\ \boldsymbol{H}_p(2) = [0.4r\ 0.4r]. \tag{3.41}$$

  The smart interference parameter $r > 0$ parametrizes the magnitude of the probe signals . The aim is to find the optimal probe signals $\boldsymbol{H}_p(l),\ l = 1, 2$ parametrized by r in (3.41) that solves (3.40). The corresponding value of r is our optimal smart interference parameter.
- Our measurement noise covariance is $\tilde{\sigma}_o^2 = 0.1$ (3.39).

Figure 3.7 displays the performance of the cognitive radar as our smart interference parameter r is varied. It shows that increasing r leads to increased confusion (worse SCNR performance) of the cognitive radar. Specifically, we plot the LHS of (3.40),

namely, $\mathbb{P}_{f_o}(\mathrm{SCNR}(\boldsymbol{H}_t(l), \boldsymbol{H}_c(l), \boldsymbol{Y}(l)) \leq \Delta)$, for SCNR upper bound $\Delta \in \{2.8, 3, 3.2\}$. Recall that this is the probability with which the maximum SCNR of the radar (3.35) lies below Δ.

To glean insight from Figure 3.7, let $r^*(\Delta, \epsilon)$ denote the optimal smart interference parameter that solves (3.40) for design parameters Δ and ϵ. Figure 3.7 shows that $r^*(\Delta, \epsilon)$ decreases with both design parameters Δ and ϵ. This can be justified as follows. For a fixed value of failure probability ϵ, increasing the upper bound Δ implies the constraint (3.40) is satisfied for smaller r, hence the optimal interference parameter $r^*(\Delta, \epsilon)$ decreases with Δ. Recall ϵ upper bounds the probability with which the maximum SCNR of the radar exceeds Δ. Increasing ϵ (or equivalently, relaxing the maximum probability of failure) allows us to decrease the magnitude of the probe signals without violating the constraint in (3.40) for a fixed Δ. Hence, $r^*(\Delta, \epsilon)$ decreases with both Δ and ϵ.

3.5 Conclusion

This chapter considered three important inter-related aspects of adversarial inference involving cognitive radars. First we discussed inverse tracking (estimating the adversary tracker's estimate based on the radar's actions) and calibration of the adversary's sensor accuracy. Then we presented a revealed preferences methodology for identifying cognitive radars; i.e. , identifying a constrained utility maximizer. Finally, we discussed designing interference to confuse the cognitive radar. The above three aspects are inter-related as depicted in Figure 3.1. The levels of abstraction range from smart interference design based on Wiener filters (at the pulse/waveform level), inverse Kalman filters at the tracking level and revealed preferences for identifying utility maximization at the systems level.

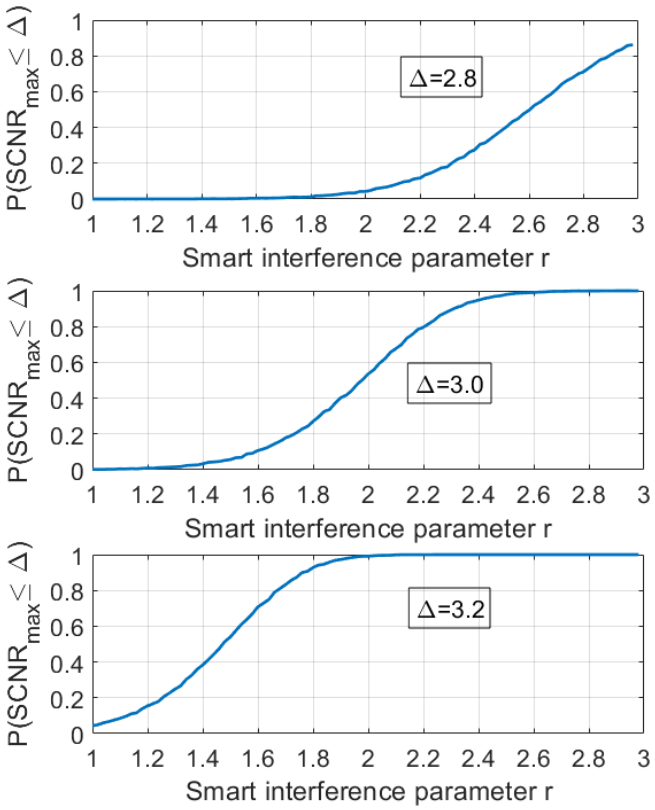

Figure 3.7: The figure illustrates the performance of the cognitive radar as our smart interference parameter r in (3.41) is varied. The plots display the LHS of (3.40), namely, probability that the radar's maximum SCNR (which depends on r (3.41)) is smaller than threshold Δ. The probability curves are plotted for $\Delta = 2.8, 3, 3.2$ and signify the extent of SCNR degradation as a function of the magnitude of the probe signal.

Extensions

The results in this chapter lead to several interesting future extensions. There is strong motivation to determine analytic performance bounds for inverse tracking/filtering and estimation of the adversary's sensor gain. Another aspect (not considered here) is when the adversary does not know the transition kernel of our dynamics; the adversary then needs to estimate this transition kernel, and we need to estimate the estimate of this transition kernel. In future work we will design the smart interference problem (3.40) as a stochastic control problem; since dynamic programming is intractable we will explore limited look-ahead policies and open-loop feedback control.

Regarding identifying cognitive radars, it is worthwhile developing statistical tests for utility maximization when the response of the utility maximizing adversarial radar is observed in noise; see Varian's work [131] on noisy revealed preference. Ongoing research in developing a dynamic revealed preference framework will be used to extend the beam allocation problem of Sec. 3.3.2 to a multi-horizon setup where we analyze batches of adversary responses over multiple slow time scale epochs. Another natural extension is to a Bayesian context, namely, identifying a radar that is a Bayesian utility maximizer. We refer to [2] for seminal work in this area stemming from behavioral economics.

Finally, in the design of controlled inference, it is worthwhile considering a game-theoretic setting where the cognitive radar (adversary) and us interact dynamically. In previous work [133] we studied simpler versions of the problem in the context of cross-polarized jamming. Also, in future work, it is worthwhile to develop a stochastic gradient algorithm for estimating the optimal probe signal.

CHAPTER 4

UNIFICATION OF ECONOMICS-BASED IRL FOR BAYESIAN AND NON-BAYESIAN DECISION SYSTEMS

4.1 Introduction

Afriat's theorem [3, 76, 77, 134] in revealed preference theory gives necessary and sufficient conditions for a finite sequence of linear budget constraints and consumption bundles to be rationalized by a monotone concave utility function. [4] generalized Afriat's theorem to general (non-linear) budget sets and provided feasibility conditions for utility maximization under monotone budget constraints. More recently, in a Bayesian context in information economics, authors in [2] address costly information acquisition and give necessary and sufficient conditions for a finite sequence of utility functions and action selection policies to be consistent with expected utility maximization with an information acquisition cost. In this chapter, we refer to testing for costly information acquisition in [2] as "*revealed rational inattention*".

Revealed preference and revealed rational inattention, respectively, test for economics based rationality (optimal decision making under resource constraints) in a non-Bayesian and Bayesian sense, respectively. So it is intuitively plausible that there exists a one-to-one correspondence between the two results. *Our key finding is that the NIAC (No Improving Attention Cycles) condition of [2, Theorem 1] in revealed rational inattention is a special case of the General Axiom of Revealed Preference (GARP) [77] used widely*

in revealed preference.[1] To the best of our knowledge, this result is new[2], and stated formally in Theorem 27. To prove this result, we first develop a revealed preference to test for cost minimization subject to utility constraints, and then extend the test to probability vectors in the unit simplex equipped with the Blackwell partial order [138].

Theorem 27 states that GARP for non-linear budgets is a generalization of the NIAC condition [2]. Indeed, GARP is an acyclic condition due to unconstrained Lagrange multipliers (marginal utility values) and is a less restrictive condition compared to the the cyclical monotonicity structure of NIAC (4.21). To complete the connection between NIAC and GARP, we generalize the revealed rational inattention result of [2] to test for expected utility maximization subject to a bound on the information acquisition cost (the result Lagrange multipliers need not be a constant across decision problems unlike [2]). The NIAC generalizes to a condition we term 'GARRI' (Generalized Axiom of Revealed Rational Inattention), and show GARRI is equivalent to GARP under the variable map of Theorem 27.

Since we will unify revealed preference and revealed rational inattention, the reader might wonder: how to abstract Bayes rule into the revealed preference formulation? It is here that the usage of Blackwell partial order is critical. In the Bayesian framework, the decision maker computes the posterior belief of the state of nature via Bayes rule using a private measurement unknown to the analyst, and then takes an action observed by the analyst. From the analyst's perspective, the decision maker chooses analyst-observable

[1] Specifically, the NIAC condition [2] is equivalent (under an appropriate variable map) to the feasibility of Afriat inequalities [3] for GARP with the additional constraint that the feasible Lagrange multipliers are constant across all problem instances. On a related note, in Sec. 4.3, we also discuss the equivalence between the NIAC condition (4.21) and the cyclical monotonicity condition for testing quasi-linear utility maximization [135].

[2] [2] allude to the cyclical monotonicity condition of [136] in the discussion of the NIAC condition. This provided us with additional motivation to investigate the connection between revealed preference and revealed rational inattention. Also, [137] generalize the setup in [2] and consequently propose a GARP-type condition for testing rational inattention. In Sec. 4.4, we argue how our unification result differs from that of [137].

probabilistic information structures that map the state to a distribution over actions. The observed information structures are termed as action selection policies in this chapter, that link the decision maker's prior belief to its posterior belief given a chosen action. As a result, the action selection policy also determines the decision maker's expected utility. We will show that the consumption bundle in the revealed preference test translates to the action selection strategy (which is a probability distribution) in the revealed rational inattention test. In revealed preference, an element-wise higher consumption good yields a larger utility for the decision maker. Thus, the utility function is a monotonically increasing function of the consumption bundle with respect to the natural (element-wise) partial order on the Euclidean space (space of consumption bundles). *In complete analogy, for the Bayesian case, a more accurate action selection policy (in the Blackwell sense) results in a higher expected utility of the Bayesian decision maker. Equivalently, the expected utility in the Bayesian setup is monotone in the action selection policy with respect to the* ***Blackwell order***.[3] This analogy is crucial for the main unification result of this chapter, Theorem 27 and is schematically shown in Fig. 4.1. We formally discuss the Blackwell order and the monotonicity of expected utility wrt the Blackwell order in Lemma 29. To convey the key ideas early on in the chapter, we present below an information version of our unification result, Theorem 27 below:

Theorem 23 (Unification Result (Informal)) *(S1.) The NIAC condition [2] for revealed rational inattention is a special case of GARP [4] under the Blackwell partial order [138] and an appropriate variable map.*
(S2.) The minimum modification needed in the decision model of [2] so that a GARP-type condition is necessary and sufficient for revealed rational inattention is the addition of a multiplier that scales the decision maker's expected utility. We term the ratio-

[3]For revealed rational inattention, it suffices to ensure weak monotonicity of the expected utility with respect to attention strategies. One well-known partial order that satisfies this condition is the Blackwell [138] order.

nal inattention analog of the GARP condition as 'GARRI' (4.27)*, defined formally in Corollary 30.*

Several reasons motivate our chapter. Revealed preference and revealed rational inattention are developed largely independently in the literature; an exception being works like [129] that use revealed preference ideas to identify maximization of the mean (not Bayesian) utility. However, unlike [2], the expected utility maximization problem assumes the probability distribution over states of nature as an exogenous variable. With our unification result, the results in these two areas can enrich each other. In Sec. 4.5, we extend the concept of robustness measures for goodness-of-fit in revealed preference literature to revealed rational inattention. We also illustrate the rational inattention analog of robustness measures to show rationally inattentive user engagement behavior in a massive YouTube dataset.[4] Apart from applications in economics, revealed preference methods have also been applied in areas like machine learning, specifically for inverse reinforcement learning [6, 12, 141–143], adversarial signal processing [144] and interpretable machine learning [145].

Related Work

Extending the revealed preference test of [3] to more general partially ordered sets of consumption bundles dates back to [146], and more recently, to [147] where the consumption bundles are partially ordered via first-order stochastic dominance. [148, 149] generalize the revealed preference test to the partial order over probability distributions (mixed strategies). Unlike the problem setting in this chapter, the decision maker in [149] does not update its belief via Bayes rule. The subtle distinction between [149] and our

[4]The term 'user engagement' is used widely in the literature [36, 139, 140] to describe user interaction on online multimedia platforms.

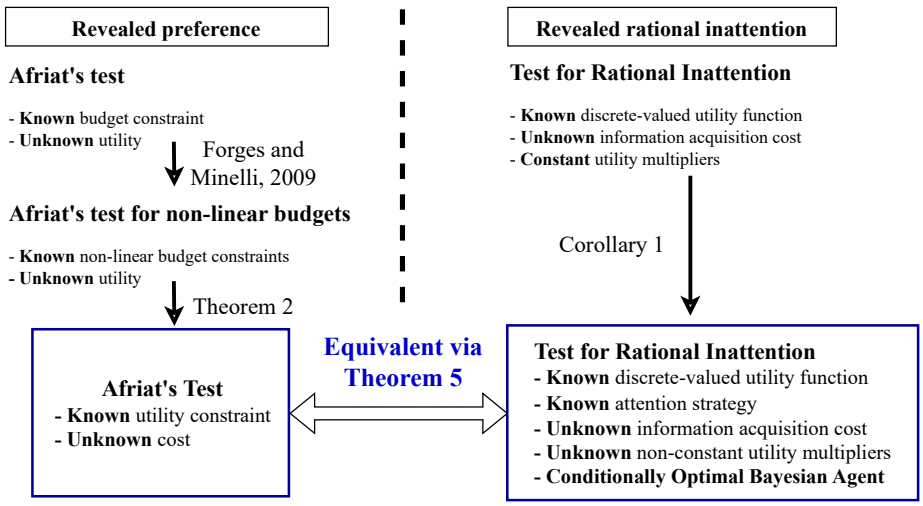

Figure 4.1: Schematic illustration of the main result in this chapter. In Theorem 25, we devise a revealed preference test for known utility constraints but unknown cost. In Corollary 30, we propose necessary and sufficient conditions for a costly information acquisition of a decision maker, where the decision maker follows a generalized decision model compared to that [2]. Finally, in our main result, Theorem 27, we construct a one-to-one equivalence map between the revealed preference test of Theorems 25 and the revealed rational inattention test of Corollary 30, and show that the NIAC condition for revealed rational inattention is a special case of the GARP condition in revealed preference.

work is that the decision maker's choice in this chapter lies in the Cartesian product of probability simplices and thus requires a different partial order. [150] consider a generalized decision model (compared to [2]) for the Bayesian agent, and give necessary and sufficient conditions for Bayesian rationality, namely, NIAS and GACI (Generalized Axiom of Costly Information) that generalize Theorem 1 in [2]. In this chapter, we focus primarily on the result of [2] and its connection to revealed preference. In spite of a unification flavor in the result of [150] where GACI and GARP are discussed in a similar vein, the variable map in the unification result of this chapter is distinct from that used by [150] to formulate GACI. Finally, [151] unify multiple approaches in revealed preference theory under an algebraic axiom of revealed preference. Our result builds on [151] in that we connect revealed preference to revealed rational inattention [2, 15] where the consumer's response takes the form of attention strategies and action selection

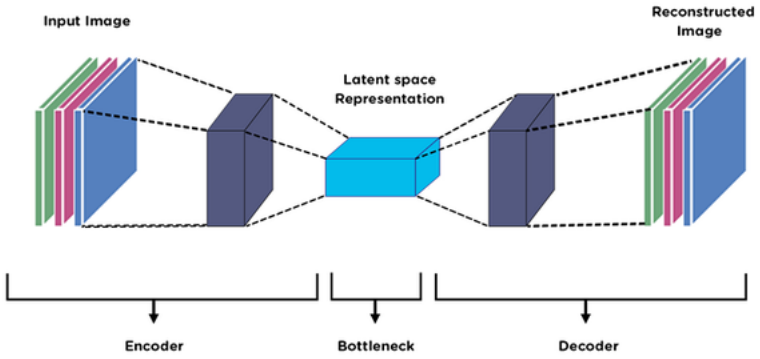

Figure 4.2: Schematic of the deep auto-encoder architecture (Image taken from Medium). The auto-encoder takes in as input and outputs an RGB image (high-dimensional data). The key feature of the auto-encoder is the 'latent space representation' or the 'bottleneck' in the above architecture, and is interpreted as a low-dimensional embedding of the input image. Auto-encoders are widely used in image processing and machine learning to generate compact representations of high-dimensional data [152–155]. In this chapter, we train a auto-encoder to convert YouTube videos' thumbnails into 16 dimensional features, that are further mapped to one of 6 feature bins. The video thumbnail's feature bin, appended with the video title's sentiment (positive, negative,neutral) generated by an NLP-based sentiment analyzer is the 'state' of the YouTube user in the rationally inattentive utility maximization model (4.16), (4.17).

policies, and the aim is to test for costly information acquisition. Also, we discuss relevant works that bridge revealed preference and revealed rational inattention in Sec. 4.4 in more detail.

Terminology

We use the terms 'costly information acquisition', 'rationally inattentive utility maximization', 'rational inattention' and 'Bayesian rationality' interchangeably in the chapter. We also use the terms 'decision maker' and 'agent' interchangeably.

Outline of Results

• *Revealed preference background:*

In Sec. 4.2, we introduce the key results of revealed preference and revealed rational inattention. We also propose: (i) the test $\mathrm{RP}_{\mathrm{mod}}$, an extension of the classical revealed preference test to identify the existence of a rationalizing cost subject to utility constraints on the decision maker, and (ii) the test, $\mathrm{RI}_{\mathrm{mod}}$, a modification of the revealed rational inattention that introduces one extra degree of freedom in the objective function of the Bayesian decision maker.

• *Unification result - Relating GARP and NIAC:*

Theorem 27 in Sec. 4.3 is our key unification result that says that $\mathrm{RP}_{\mathrm{mod}}$ is equivalent to $\mathrm{RI}_{\mathrm{mod}}$ when the Bayesian decision maker is assumed to be conditionally optimal - it chooses the action that maximizes its expected utility conditioned on its posterior belief. Our second result, Corollary 30, introduced the minimal set of assumptions that facilitates testing for rational inattention via a GARP-type inequality, we term this inequality as *GARRI* (Generalized Axiom of Revealed Rational Inattention).

• *Discussion of related works:*

Sec. 4.4 discusses existing works in the literature that bridge the methodologies of revealed preference and revealed rational inattention, and also highlights the differences between existing unification results and the key results of Sec. 4.3.

• *Extending robustness measures in revealed preference to revealed rational inattention:*

Building on the unification results in Sec. 4.3, Sec. 4.5 introduces robustness measures for revealed rational inattention that measure how far a dataset is from being consistent with rationally inattentive behavior. We introduce Bayesian analogs of well-studied robustness measures in the literature, for example, the Afriat's efficiency index [156], the Houtman-Maks index [157] and the minimal perturbation test [158]. To the best of our knowledge, robustness measures for revealed rational inattention have not been explored

in the literature.

• *Robustness analysis of revealed rational inattention on a real-world YouTube dataset:* Finally, in Sec. 2.5, we perform a revealed rational inattention test on a real-world YouTube dataset comprising meta-data from approximately $140,000$ videos. The meta-data in the YouTube dataset comprises the video thumbnail, title, viewcount, number of comments, and number of likes and dislikes on each video, recorded after 20 days of posting the video. The numerical experiments on the YouTube dataset extend our recent works [12, 143] in the following aspects: (i) we present a novel deep auto-encoder and natural language processing (NLP)-based feature extraction procedure for YouTube metadata (video thumbnail and title), (ii) we use the equivalence results developed in the chapter to conduct a systematic robustness analysis of YouTube metadata by computing robustness measures adapted from revealed preference theory, and (iii) we test for a generalized rational inattention model compared to that proposed by [2]. Testing YouTube metadata for a generalized model of rational inattention is useful because YouTube user engagement literature [159–161] shows that different groups of YouTube users have different attention spans. In the rational inattention context, different attention spans translates to different marginal costs of information acquisition. The forward optimization model of [2] is restrictive in that the decision maker has the same marginal cost in all decision problems (video categories in the YouTube context). The generalized model proposed in Corollary 30 allows for non-constant attention spans in different decision problems. In the context of the YouTube dataset for our numerical experiments, we use the terms 'user engagement' and 'commenting behavior' interchangeably. Our numerical results show that YouTube metadata features output pass the revealed rational inattention test by a large margin, where the goodness-of-fit to the rational inattention model is measured by computing the robustness metrics defined in Sec. 4.5. All our numerical results are completely reproducible and can be accessed from the GitHub repository

https://github.com/KunalP117/YouTube-Commenting-Analysis.

4.2 Background

To set the stage for our unification result that relates revealed preference and revealed rational inattention, we start with a review of the key results of [4] (revealed preference for non-linear budgets) and [2] (revealed rational inattention).

4.2.1 Revealed preference (non-linear budget)

Theorem 24 ([4]) *Consider a decision maker that, at time $k = 1, 2, \ldots, K$, chooses a consumption bundle $\beta_k \in \mathbb{R}^m_+$ subject to a non-linear budget constraint $g_k(\beta) \leq 0$. Assume that $g_k(\cdot)$ is continuous, monotone and known to the external analyst, the set $\{\beta | g_k(\beta) \leq 0\}$ of feasible bundles is compact, and the budget constraint is active at β_k, that is, $g_k(\beta_k) = 0$. Then, the following statements are equivalent:*

1. *There exists a monotone, continuous utility function $u : \mathbb{R}^m_+ \to \mathbb{R}$ that rationalizes the data set $\{\beta_k, g_k(\cdot) \leq 0\}_{k=1}^K$:*

$$\beta_k \in \underset{\beta \in \mathbb{R}^m_+}{\operatorname{argmax}}\, u(\beta),\ \textit{s.t.}\ g_k(\beta) \leq 0. \tag{4.1}$$

2. *The data set $\{\beta_k, g_k(\cdot) \leq 0\}_{k=1}^K$ satisfies GARP:*

$$\beta_k \geq_H \beta_j \Rightarrow g_j(\beta_k) \geq 0,\ \forall k, j \in \{1, 2, \ldots, K\},\ k \neq j, \tag{4.2}$$

where the relation $\beta_k \geq_H \beta_j$ ('revealed preferred to') means there exists indices $i_1, i_2, \ldots, i_L$ such that $g_k(\beta_{i_1}) \leq 0,\ g_{i_1}(\beta_{i_2}) \leq 0,\ \ldots, g_{i_L}(\beta_j) \leq 0$.

3. *There exist positive scalars* $u_k, \lambda_k > 0,\ k = 1, 2, \ldots, K$ *such that the following inequalities hold:*

$$u_s - u_t - \lambda_t g_t(\beta_s) \leq 0 \quad \forall t, s \in \{1, 2, \ldots K\} \tag{4.3}$$

The reconstructed utility function $\hat{u}$ *defined in* (4.4) *below in terms of the feasible variables in* (4.3) *is monotone, continuous and rationalizes* $\{\beta_k, g_k(\cdot) \leq 0\}_{k=1}^{K}$ (4.1)*:*

$$u(\beta) = \min_{k \in \{1,\ldots,K\}} \{u_k + \lambda_k g_k(\beta)\} \tag{4.4}$$

Theorem 24 says that a sequence of budget constraints and consumption bundles are rationalized by a utility function if and only if a set of linear inequalities (4.3) has a feasible solution. The GARP condition (4.2) is equivalent (due to [77]) to the cyclical consistency condition proposed in [3]. For completeness, we remark that constraining λ_k in the feasibility test (4.3) to be a constant for all k, and assuming a linear budget $g_k(\cdot)$ in Theorem 24 is equivalent to testing for quasi-linear utility maximization [135], that is, the cyclical monotonicity condition of [135, Theorem 2.2] holds. Indeed, cyclical monotonicity for quasi-linear utility maximization is a stronger condition than GARP.

Theorem 24 is a standard result in revealed preference literature. In Sec. 4.2.2 below, we extend Theorem 24 to the scenario when the the analyst knows the decision maker's utility, and tests for the existence of budget constraints that rationalize the decision maker's actions.

4.2.2 Modifying the revealed preference test: Observed utility, unobserved budget constraint

Our key aim in this chapter is to establish a correspondence between revealed preference and revealed rational inattention. However, both results differ in what the analyst knows

about the decision maker's decisions. Revealed rational inattention assumes the analyst knows the decision maker's utility function and tests for the existence of a rationalizing information acquisition cost. Hence, to establish the correspondence, we need to modify Theorem 24 to the case where the analyst knows the decision maker's utility function and tests for the existence of a rationalizing cost.[5] We formalize this departure from the problem setting in Theorem 24 in Assumption 1 below.

Assumption 1 *Consider the decision maker and analyst described in Theorem 24.*
(A1.1) The analyst's aim is to test if the decision maker chooses its consumption bundles optimally by solving the following optimization problem:

$$\beta_k \in \underset{\beta \in \mathbb{R}^m_+}{\operatorname{argmin}} \; g(\beta), \; u_k(\beta) \geq u^*_k, \; \forall k = 1, 2, \ldots, K. \tag{4.5}$$

In (4.5)*, the decision maker minimizes the cost of choosing bundle β subject to a lower bound on the utility.*
(A1.2) The utility constraint in (4.5) *is active, that is, $u_k(\beta_k) = u^*_k$ for all $k = 1, 2, \ldots, K$.*
(A1.3) The analyst knows the dataset $\mathcal{D}_{RP}$ defined as:

$$\mathcal{D}_{RP} = \{u_k, u^*_k, \beta_k\}_{k=1}^K. \tag{4.6}$$

(A1.4) The analyst's aim is to identify if there exists a cost $g(\beta)$ that rationalizes the analyst's dataset $\mathcal{D}_{RP}$ (4.6)*, that is, the observed responses $\{\beta_k\}_{k=1}^K$ solve the optimization problem* (4.5)*.*

In (4.5), we refer to $g(\beta)$ in (4.5) as the 'cost' of purchasing consumption bundle β, in comparison to a budget *constraint* in classical revealed preference where the utility

[5] In classical revealed preference, the decision maker maximizes its utility (unknown to the analyst) subject to an upper bound on their budget (known to the analyst). Assumption 1 modifies the classical setup to the case where the decision maker minimizes a cost (unknown to the analyst) subject to a lower bound on their utility (known to the analyst). Under mild conditions on the decision maker's cost and utility, both optimization problems are equivalent. However, for notational convenience, we pose the decision problem in such a way that the revealed preference estimand is the decision maker's objective function.

function is unknown. The optimization problem in (4.6) is a cost minimization problem subject to a lower bound on the utility. A related model is studied in [162] where the decision maker at time $k = 1, 2, \ldots, K$ minimizes a cost function but is constrained to choose its response from a specified compact set. In [162], analogous to WARP, the *Weak Axiom of Cost Minimization* (WACM) is proposed as a necessary and sufficient condition that rationalizes the dataset. In this chapter, we focus on relating GARP from revealed preference theory to revealed rational inattention results. Hence, in Theorem 25 below, our necessary and sufficient condition for rationalizability is expressed in terms of the GARP condition (4.7) even though it is straightforward to express (4.7) in the style of [162, Theorem 1]. We are now ready to state Theorem 25. In complete analogy to Theorem 24, Theorem 25 below states necessary and sufficient conditions for utility maximization when the analyst knows the decision maker's utility constraints (4.5).

Theorem 25 (Revealed Preference (Unknown Cost, Known Utility Constraints)) *Consider a decision maker that, at time $k = 1, 2, \ldots, K$, chooses a consumption bundle $\beta_k \in \mathbb{R}^m_+$ subject to a utility constraint $u_k(\beta) \geq u^*_k$. Suppose Assumption 1 holds and the set $\{\beta | u_k(\beta) \geq u^*_k\}$ of feasible bundles is compact for all k. Then, the following statements are equivalent:*

1) There exists a monotone, continuous cost $g : \mathbb{R}^m_+ \to \mathbb{R}_+$ that rationalizes the dataset $\mathcal{D}_{RP}$, that is, (4.5) *holds.*

2) The data set $\{\beta_k, u_k(\beta_k) - u_k(\cdot)\}_{k=1}^K$ satisfies GARP (4.2)*:*

$$\beta_k \geq_H \beta_j \Rightarrow u_j(\beta_j) \leq u_j(\beta_k), \; \forall \, k \neq j, \tag{4.7}$$

where the relation $\beta_k \geq_H \beta_j$ implies there exists indices $i_1, i_2, \ldots, i_L$ such that $u_k(\beta_{i_1}) \geq u_k(\beta_k), \; u_{i_1}(\beta_{i_2}) \geq u_{i_1}(\beta_{i_1}), \; \ldots, u_{i_L}(\beta_j) \geq u_{i_L}(\beta_{i_L})$.

3) There exist positive scalars $g_k, \lambda_k > 0, \; k = 1, 2, \ldots, K$ such that the following

inequalities hold:

$$g_j - g_k - \lambda_k \left(u_k(\beta_j) - u_k(\beta_k)\right) \geq 0 \; \forall k, j \in \{1, 2, \ldots, K\}, \; k \neq j.$$

Or equivalently, $\lambda_k \, u_k(\beta_k) - g_k \geq \lambda_k \, u_k(\beta_j) - g_j, \forall k, j \in \{1, 2, \ldots, K\}, \; k \neq j$ (4.8)

We denote the inequalities (4.8) *in abstract notation as* $\mathrm{RP}(\{u_k, \beta_k\})$. *The reconstructed cost* $\hat{g}$ *defined in* (4.9) *below in terms of the feasible variables in* (4.8) *is monotone, continuous and rationalizes* $\{\beta_k, u_k(\cdot) \geq u_k^*\}_{k=1}^K$ (4.5)*:*

$$\hat{g}(\beta) = \max_{k \in \{1,2,\ldots K\}} \{g_k + \lambda_k \left(u_k(\beta) - u_k(\beta_k)\right)\}. \tag{4.9}$$

Theorem 25 above yields necessary and sufficient conditions for utility maximization behavior when the decision maker's utility function is observed by the analyst; its budget (cost) is unobserved and must be reconstructed by the analyst by testing for feasibility of a set of Afriat-type inequalities (4.8) for rationalizability. The proof of Theorem 25 is in the Appendix. At first sight, (4.5) in Theorem 25 appears to be a dual statement to the optimization problem (4.1) in Theorem 24. However, the proof does not use duality. Also, in comparison to the reconstruction procedure (4.4) that yields a point-wise *minimum* of piece-wise monotone functions, notice the reconstructed cost g in (4.9) is a point-wise *maximum* of piece-wise monotone functions. Indeed, if $u_k(\cdot)$ is differentiable for all k, then the cost $g(\beta) = \max_{k \in \{1,\ldots,K\}} \{g_k + \lambda_k \nabla u_k(\beta_k)'(\beta - \beta_k)\}$ also rationalizes the decision maker's actions. This result follows from [4, Proposition 2]. Notice how in comparison to a piece-wise linear, concave utility reconstruction in Afriat's theorem for linear budget constraints, we now have a piece-wise linear *convex* cost that rationalizes the decision maker's actions if u_k is differentiable.

Recall that our key objective is to establish a correspondence between revealed preference and revealed rational inattention [2]. In revealed rational inattention, the

analyst knows the agent's utility and tests for the existence of a rationalizing information acquisition cost. In Sec. 4.3, we will show that the setup in [2] is a Bayesian analog of Theorem 25 and relate the reconstructed cost in (4.9) to the information acquisition cost. The key takeaway of Theorem 25 is that we have a revealed preference test for unobserved costs in terms of GARP. Authors in [137] have generalized the forward optimization problem in revealed rational inattention [2] so that the inverse learner uses a Bayesian analog of GARP instead of NIAC to test for rational inattention. However, as will be discussed in Theorem 27, the variable map for relating NIAC and GARP in this chapter is distinct from [137]; we rely on the auxiliary revealed preference result of Theorem 25 to establish the one-to-one correspondence.

4.2.3 Revealed Rational Inattention

Having introduced revealed preference results for non-linear budgets, we now turn our attention to the Bayesian utility maximization setup of [2]. Since our aim is to relate Theorem 25 to revealed rational inattention, we review the key revealed rational inattention result of [2].

Before we state the key result, we describe the decision model of the Bayesian agent. Suppose the decision maker acts in a sequence of decision problems $k = 1, 2, \ldots, K$, where the decision problem parametrizes the decision maker's utility. The decision maker in [2] is a Bayesian agent - it has a prior probability distribution π_0 over a finite set of states $\mathcal{X}$. In every decision problem k, the agent chooses an information structure $\boldsymbol{\alpha}_k(\cdot|x)$ that maps every state $x \in \mathcal{X}$ to a probability distribution over a finite set of subjective signals $\mathcal{Y}$; the kernel $\boldsymbol{\alpha}_k$ is termed as the agent's *attention strategy* in decision problem k:

$$\textit{Attention Strategy} \quad \boldsymbol{\alpha}_k : \mathcal{X} \to \Delta(\mathcal{Y}), \tag{4.10}$$

where $\Delta(\mathcal{Y})$ denotes the space of probability distributions over the set $\mathcal{Y}$. The agent then observes a realized signal (random variable) $y \in \mathcal{Y}$ (and not the ground truth x) and updates its belief (posterior) of the unobserved state x using Bayes rule:

$$\pi_y(x) = \frac{\pi_0(x)\boldsymbol{\alpha}_k(y|x)}{\sum_{x' \in \mathcal{X}} \pi_0(x')\boldsymbol{\alpha}_k(y|x')} \tag{4.11}$$

After computing the belief π_y, the agent then chooses an action a from a finite set of actions $\mathcal{A}$ according the *conditional action policy*:

$$\textit{Conditional Action Policy} \quad \eta_k : \Delta(\mathcal{X}) \to \mathcal{A} \tag{4.12}$$

The agent's attention strategy $\boldsymbol{\alpha}_k$ and conditional action policy η_k induce the *action selection policy* $p_k(a|x)$ and is defined as the probability of choosing action a if the true state is x:

$$p_k(a|x) = \sum_{y \in \mathcal{Y}} \boldsymbol{\alpha}_k(y|x)\ \eta_k(a|\pi_y), \tag{4.13}$$

where π_y is defined in (4.11) and *conditional action selection policy* $\eta_k(\cdot|\pi)$ is the probability the agent chooses action a given belief π in decision problem k. In the revealed rational inattention problem, we assume the analyst knows the dataset:

$$\mathcal{D}_{RRI} = \{\pi_0, \{p_k(a|x), \boldsymbol{U}_k\}_{k=1}^K\}, \tag{4.14}$$

In (4.14), π_0 is the prior probability distribution over the state space $\mathcal{X}$ and $p_k(a|x)$ is the agent's action selection policy defined in (4.13). The variable $\boldsymbol{U}_k$ is the agent's discrete-valued utility in decision problem k that depends on the state $x \in \mathcal{X}$ and action $a \in \mathcal{A}$:

$$\boldsymbol{U}_k \ \equiv \{\boldsymbol{U}_k(x,a) \in \mathbb{R},\ x \in \mathcal{X}, a \in \mathcal{A}\}. \tag{4.15}$$

The analyst's aim in revealed rational inattention is to test if the Bayesian agent acts optimally and maximizes its *expected* utility maximization less a non-negative information cost $L(\boldsymbol{\alpha})$ that depends only the attention strategy $\boldsymbol{\alpha}$. The forward optimization problem

is termed as 'rationally inattentive utility maximization' in the literature:

Rationally Inattentive Utility Maximization:

$$(a)\ \eta_k(a|\pi_y) \in \underset{p\in\Delta(\mathcal{A})}{\operatorname{argmax}} \sum_{x\in\mathcal{X}} \pi_y(x)\boldsymbol{U}_k(x,a),\ \forall y \in \mathcal{Y} \tag{4.16}$$

$$(b)\ \boldsymbol{\alpha}_k \in \underset{\boldsymbol{\alpha}:\mathcal{X}\to\Delta(\mathcal{Y})}{\operatorname{argmax}}\ J_{\pi_0}(\boldsymbol{\alpha},\boldsymbol{U}_k) - L(\boldsymbol{\alpha}),\ \text{where} \tag{4.17}$$

$$J_{\pi_0}(\boldsymbol{\alpha},\boldsymbol{U}_k) = \sum_{y\in\mathcal{Y}} p_k(y)\ \left(\max_{a'\in\mathcal{A}} \sum_x\ \pi_y(x)\boldsymbol{U}_k(x,a')\right),\ p_k(y) = \sum_x \pi_0(x)\boldsymbol{\alpha}_k(y|x). \tag{4.18}$$

In (4.16), π_y denotes the agent's belief computed using Bayes rule in (4.11) after observing signal $y \in \mathcal{Y}$. In (4.17), $\Delta(S)$ denotes the space of probability distributions over a set S. Eq. 4.16 ensures that the agent maximizes its conditional expected utility given any posterior probability distribution π computed using (4.11). Assuming (4.16) is true, (4.17) ensures that agent's chosen attention strategy $\boldsymbol{\alpha}_k$ maximizes its objective function, namely, expected utility (4.17) minus an information acquisition cost. The analyst's aim is to test for the existence of an information cost L (4.18) that rationalizes the dataset $\mathcal{D}_{RRI}$ (4.14), that is, the *unobserved* agent variables $\{\boldsymbol{\alpha}_k, \eta_k\}_{k=1}^K$ satisfy the optimality conditions (4.16), (4.17), which makes the revealed rational inattention result of [2] stated below remarkable.

Theorem 26 (Revealed rational inattention [2]) *Consider a Bayesian agent that faces K decision problems, where decision problem k is parameterized by a finite set of states $\mathcal{X}$, signals $\mathcal{Y}$, actions $\mathcal{A}$ and utility $\boldsymbol{U}_k$ (4.15). Suppose an analyst knows the dataset $\mathcal{D}_{RRI}$ (4.14) that comprises the agent's action selection policies $\{p_k(a|x)\}_{k=1}^K$ and its utility functions $\{\boldsymbol{U}_k\}_{k=1}^K$ in the K decision problems. Then, the following statements are equivalent:*

1) There exists a monotone information acquisition cost $L(\boldsymbol{\alpha})$ that rationalizes the dataset $\mathcal{D}_{RRI}$. That is, the agent's unobserved attention strategy $\boldsymbol{\alpha}_k$ (4.10) and conditional action

policy η_k (4.12) *solve the nested optimization problem* (4.16)*,* (4.17) *for all decision problems* $k = 1, 2, \ldots, K$.

2) The dataset $\mathcal{D}_{RRI}$ *satisfies the 'No-Improving-Action-Switches' (NIAS) and the 'No-Improving-Action-Cycles' (NIAC) conditions:*

$$\textit{NIAS: } \sum_{x \in \mathcal{X}} p_k(x|a) \ (\ \boldsymbol{U}_k(x,a) - \boldsymbol{U}_k(x,b) \ ) \geq 0, \ \forall \ a \neq b, \ a, b \in \mathcal{A}, \ k = 1, 2, \ldots, K \tag{4.19}$$

NIAC: For any sequence of distinct indices $i_1, i_2, \ldots, i_M \in \{1, 2, \ldots, K\}$ $(M \leq K)$, (4.20)

the following inequality holds :

$$\sum_{m=1}^{M} \left(\sum_{x \in \mathcal{X}, a \in \mathcal{A}} \pi_0(x) \ p_{i_m}(a|x) \ \boldsymbol{U}_{i_m}(x,a) - G(p_{i_{m+1}}(a|x), \boldsymbol{U}_{i_m}) \right) \geq 0,$$

where $i_{M+1} = i_1$ *and* : (4.21)

$$G(p_j(a|x), \boldsymbol{U}_i) = \sum_a p_j(a) \ \max_{b \in \mathcal{A}} \sum_x p_j(x|a) \ \boldsymbol{U}_i(x,b) \tag{4.22}$$

The variable $G(\cdot, \cdot)$ *in* (4.22) *is the Bayesian agent's surrogate expected utility. In* (4.21)*, the variable* $p_k(a) = \sum_x \pi_0(x) p_k(a|x)$ *is the marginal distribution of the action* a*, the variable* $p_k(x|a) = \pi_0(x) p_k(a|x) / p_k(a)$ *is the posterior belief of the state when action* a *is realized.*

3) The dataset $\mathcal{D}_{RRI}$ *satisfies the data-matching condition:*

$$\textit{There exists } \eta_k : \mathcal{Y} \to \Delta(\mathcal{A}) \textit{ s.t. } p_k(a|x) = \sum_y \eta_k(a|y) \ \boldsymbol{\alpha}_k(y|x), \ \forall \ k. \tag{4.23}$$

Theorem 26 is well-known in the information economics literature [2]; see [2, Sec. 10.2] and [143, Sec. C.2.2] for the proof. The 'No-Improving-Action-Switches' (NIAS) (4.19)

and 'No-Improving-Action-Cycles' (NIAC) (4.21) conditions in Theorem 26 are necessary and sufficient for the existence of an information acquisition cost L that rationalizes the dataset $\mathcal{D}_{RRI}$, that is, conditions (a) (4.16) and (b) (4.17) hold. The first term in the LHS in (4.21) is the expected utility of the agent in decision problem k. The second term in the LHS is the agent's *surrogate expected utility* $G(p_{i_{m+1}}(a|x), \boldsymbol{U}_{i_m})$, surrogate since the expectation is in terms of the action selection policy of the agent, and not its attention strategy. It is straightforward to show that the attention strategy Blackwell dominates the action selection policy. We make the notion of Blackwell dominance precise in Lemma 29 below. Due to Blackwell dominance and Lemma 29, we further have the following inequality:

$$G(p_{i_{m+1}}(a|x), \boldsymbol{U}_{i_m}) \leq G(\boldsymbol{\alpha}_{i_{m+1}}, \boldsymbol{U}_{i_m}) = J_{\pi_0}(\boldsymbol{\alpha}_{i_m}, \boldsymbol{U}_{i_m}) \text{ (4.18)}, \tag{4.24}$$

where equality holds when $i_{m+1} = i_m$. A second observation that is crucial for the unification result of Theorem 27 below is that $G(p_{i_m}(a|x), \boldsymbol{U}_{i_m}) = J_{\pi_0}(\boldsymbol{\alpha}_{i_m}, \boldsymbol{U}_{i_m}) = \sum_{x\in\mathcal{X}, a\in\mathcal{A}} \pi_0(x)\, p_{i_m}(a|x)\, \boldsymbol{U}_{i_m}(x,a)$ if and only if NIAS holds. Proving this relation is straightforward and omitted for brevity. Intuitively, the term $G(p_{i_{m+1}}(a|x), \boldsymbol{U}_{i_m})$ is a surrogate for the agent's expected utility since the inverse learner does not know the agent's attention strategies. While necessity for optimality is straightforwardly determined, the sufficiency proof assumes η_k to be a one-to-one map; the surrogate expected utility matches the true expected utility, that is, $G(p_{i_{m+1}}(a|x), \boldsymbol{U}_{i_m}) = G(\boldsymbol{\alpha}_{i_{m+1}}, \boldsymbol{U}_{i_m})$.

Abstractly, the NIAS condition is true if and only if condition (a) (4.16) is true, and the NIAC condition is true if and only if condition (b) (4.17) is true. Finally, the data-matching condition (4.23) ensures that the dataset $\mathcal{D}_{RRI}$ is indeed generated from a *Bayesian* decision maker that makes an action based on its realized posterior belief.

Discussion of Theorem 26.

1. Classical revealed rational inattention [2] assumes that there exists a *single* utility

function $\boldsymbol{U}(x,a),\ x \in \mathcal{X},\ a \in \mathcal{A}$, and the Bayesian agent's action choice in decision problem k is restricted to a subset $\mathcal{A}_k \subset \mathcal{A}$. This restriction can be equivalently modeled as the decision maker having a utility function $\boldsymbol{U}_k$ in decision problem k without any restriction on the choice of actions.[6]

2. Abstractly, Theorem 26 says that the analyst can test for rationally inattentive utility maximization (4.16), (4.17) even if the analyst only has access to a stochastically garbled [7] version of the attention strategy, namely, the action selection policy. We discuss this concept of stochastic garbling in more detail later in the chapter in the context of Blackwell [138] partial order on the space of probability distributions.

4.2.4 Revealed rational inattention and observability of Bayesian agent's decisions by the analyst

In revealed preference (Theorem 25), the analyst observes the agent decisions accurately. In revealed rational inattention (Theorem 26), the analyst observes a noisy version of the Bayesian agent's decisions. A key yet unusual takeaway of revealed rational inattention is that the analyst can test for rational inattention by treating the noisy measurement of the agent decision as the true decision. We justify this claim below.

In the rational inattention model (4.16), (4.17), the decision maker, in decision problem k, chooses its attention strategy $\boldsymbol{\alpha}_k$ and conditional action policy η_k. The quantities of interest to the Bayesian decision maker, namely, the information acquisition cost L and expected utility J only depend on $\boldsymbol{\alpha}_k$ and η_k, in addition to the prior π_0 that

[6]The problem setting in [2] is equivalent to setting $\boldsymbol{U}_k(x,a) = \boldsymbol{U}(x,a)$ if $a \in \mathcal{A}_k$ and $-\infty$ otherwise, where $\boldsymbol{U}$ is the agent's fixed utility over decision problems.

[7]Indeed, $p_k(a|x) = \sum_{y\in\mathcal{Y}} \eta(a|\pi_y)\boldsymbol{\alpha}_k(y|x)$. Hence, the matrix $Q \in [0,1]^{|\mathcal{A}|\times|\mathcal{Y}|}$ with elements $Q_{a,y} = \eta(a|\pi_y)$ can be viewed as a noisy channel that takes as input the attention strategy $\boldsymbol{\alpha}_k$ and outputs the action selection policy $p_k(a|x)$.

is assumed known to the analyst. Hence, it is intuitive to expect that test for optimal Bayesian decision-making is possible only if the chosen attention strategies $\{\boldsymbol{\alpha}_k\}_{k=1}^K$ and conditional action policies $\{\eta_k\}_{k=1}^K$ are known to the analyst performing revealed rational inattention.

However, unlike revealed preference, the analyst only observes the action selection policy $p_k(a|x) = \sum_{y\in\mathcal{Y}} \eta_k(y)\boldsymbol{\alpha}_k(y|x)$ that is a noisy (less informative) version of the attention strategy $\boldsymbol{\alpha}_k$. Theorem 26 indicates that the knowledge of the action selection policies suffices for testing Bayesian rationality. In the context of our equivalence result stated in Theorem 27 below, testing if rational inattention (4.16), (4.17) holds is *equivalent* to testing if rational inattention (4.16), (4.17) holds when the attention strategy $\boldsymbol{\alpha}_k$ is replaced by the action selection policy $p_k(a|x)$. Hence, for the purpose of our equivalence result, we can treat the effective Bayesian decision maker's 'response' as simply its action selection policy $p_k(a|x)$ in decision problem $k = 1, 2, \ldots, K$. Let us briefly elaborate on the above claims:

- If the agent is Bayes rational (4.16), (4.17), then the optimality conditions (4.16), (4.17) also hold when the attention strategy is replaced with the action selection policy, a noisy version of the attention strategy. Indeed, the proof of necessity of NIAS and NIAC for rational inattention shows that replacing the attention strategy in (4.16), (4.17) with the action selection policy, and testing for a weaker version of (4.17) to ensure optimality over a finite number of strategies yields the NIAS and NIAC inequalities. The key component in the necessity of the NIAS and NIAC conditions for rational inattention is Blackwell dominance discussed in Lemma 29; at a deeper level, NIAS and NIAC is necessary for (4.16) and (4.17) to hold since the attention strategy 'Blackwell dominates' the action selection policy.

- The sufficiency proof of NIAS and NIAC for rational inattention assumes a one-to-one map from the observation y to the action a. Since the attention strategy is not observed, the analyst can assume that the *observed* action selection policy is the same as the *unobserved* attention strategy without loss of generality. Afriat [3] computes a set-valued utility function that rationalizes the finite dataset $\mathcal{D}_{RP}$ (4.6). In complete analogy, [2] exploit a result from quadratic assignment problems [163] to construct a set-valued rational inattention cost that is non-zero at the observed finitely many action selection policies (or equivalently, the attention strategies) in the K environments, and ∞ elsewhere. Since the reconstruction of the rational inattention cost only occurs in the sufficiency part of the proof, it thus suffices to replace the attention strategy with the action selection policy and simply check for Bayesian rationality of the chosen action selection policies. For clarity, we also express the reconstructed information acquisition cost as a function of the action selection policy in the generalization of the revealed rational inattention test of [2] stated in Corollary 30 below and implicitly assume a one-to-one map from the observations to the actions.

To summarize, in this section we justify how an analyst can test for Bayesian rationality (4.16), (4.17) by *assuming the observed action selection policy is the same as the unobserved attention strategy*. The key idea is that since the attention strategy is not observed, the analyst can, without loss of generality, assume a one-to-one mapping from the observation y to the action a. As a result, in the equivalence result below, we will show that the Bayesian decision maker's equivalent response is the action selection policy, and not the unobserved attention strategy and conditional action policy. Also, in Corollary 30 (a generalization of Theorem 26), the reconstructed information acquisition cost from the revealed rational inattention test is expressed in terms of the action selection policy; it is assumed that the action selection policy is the same as the unobserved attention strategy.

4.3 Main Result. Unification of Revealed preference and Revealed rational inattention

We present our first key result in this section that unifies revealed preference and revealed rational inattention. Our unification result is Theorem 27 below. Informally, the key takeaway of Theorem 27 is as follows:
The NIAC condition of [2] is a special case of GARP (4.7) *if NIAS holds. If NIAC* (4.21) *holds, then the GARP condition* (4.7) *for utility maximization is true under an equivalent variable map. If NIAS holds, and the Afriat-type feasibility inequalities* (4.8) *are feasible with the Lagrange multipliers* λ_k *set to* 1 *under the variable map, then the NIAC condition is true.*

Let us now formalize the above takeaways in Theorem 27 below.

Theorem 27 (Unification of revealed preference and revealed rational inattention) *Consider the revealed rational inattention result of Theorem 26 and the revealed preference result of Theorem 25. Recall that the analyst uses dataset* $\mathcal{D}_{RRI}$ (4.14) *to test for rational inattention* (4.16)*,* (4.17) *in Theorem 26, and uses dataset* $\mathcal{D}_{RP}$ (4.6) *to test for utility maximization* (4.5)*. Also, suppose the NIAS condition* (4.19) *holds for the dataset* $\mathcal{D}_{RRI}$*. Then:*

1. *The NIAC condition* (4.21) *in Theorem 26 is a special case of GARP* (4.7) *under the variable map below:*

Revealed Preference	$\Leftrightarrow$	*Revealed Rational Inattention*
• *Time step* k	$\Leftrightarrow$	*Decision problem* k
• *Response* β_k	$\Leftrightarrow$	*Action selection policy* $p_k(a\|x)$
• *Utility Function* $u_k(\beta)$	$\Leftrightarrow$	*Expected Utility* $G(p(a\|x), \boldsymbol{U}_k)$ (4.22)

- *Utility Bound* u_k^* $\Leftrightarrow$ *Expected Utility* $G(p_k(a|x), \boldsymbol{U}_k)$ (4.22)
- *Cost* $g(\beta)$ $\Leftrightarrow$ *Information Acquisition Cost* L (4.18)
- *Element-wise partial order on the space of consumption bundles (Euclidean space)* $\Leftrightarrow$ *Blackwell partial order on the space of attention strategies (pmfs)*

2. *The following modification of rationally inattentive utility maximization* (4.16)*,* (4.17) *generalizes NIAC in the revealed rational inattention test of Theorem 26 to a GARP-type condition:*

   Modified Rationally Inattentive Utility Maximization $\equiv$ (4.16) *and the following modification of* (4.17) *holds:*

$$\boldsymbol{\alpha}_k \in \underset{\boldsymbol{\alpha}:\mathcal{X}\to\Delta(\mathcal{Y})}{\operatorname{argmax}} \ \lambda_k \, J_{\pi_0}(\boldsymbol{\alpha}, \boldsymbol{U}_k) - L(\boldsymbol{\alpha}), \ \textit{or equivalently,} \tag{4.25}$$

$$\boldsymbol{\alpha}_k \in \underset{\boldsymbol{\alpha}:\mathcal{X}\to\Delta(\mathcal{Y})}{\operatorname{argmin}} \ L(\boldsymbol{\alpha}), \ \textit{subject to } J_{\pi_0}(\boldsymbol{\alpha}, \boldsymbol{U}_k) \geq J_k^*, \tag{4.26}$$

   where $J_{\pi_0}(\cdot)$ *is defined in* (4.18)*, and* $\lambda_k > 0$ *in* (4.25) *is a utility multiplier. We formalize the GARP-type generalization of NIAC in Corollary 30 below. Also, the setup in* (4.25) *is the "minimum" modification needed wrt the forward decision model in [2] specified by conditions* (4.16)*,* (4.17) *that allows checking for optimality of the chosen attention strategy via a GARP-type condition.*

We prove Theorem 27 in Sec. 4.7.2 and briefly discuss the intuition behind the proof below. A key aspect of the unification result (statement (1)) in Theorem 27 is that the equivalent variables in the Bayesian decision setup comprise only the variables observed by the external analyst. For example, although the analyst knows the Bayesian decision maker chooses the attention strategy $\boldsymbol{\alpha}_k$ and the action selection policy η_k (4.12), the equivalent response under the variable map is only the observed action selection policy $p_k(a|x)$ that depends on $\boldsymbol{\alpha}_k$ and η_k. We justify this unusual claim in Sec. 4.2.4.

The key idea behind relating NIAC (4.21) and GARP (4.7) is to first express the NIAC condition for revealed rational inattention test in Theorem 26 as a feasibility inequality (4.45) (see Sec. 4.7.2 for the proof), and then compare the feasibility inequality to the Afriat-type inequality (4.8) for the modified revealed preference test in Theorem 24. Under the variable map outlined in statement (1) in Theorem 27 above, we observe in the proof that the inequality (4.45) is the same as (4.8) with the Lagrange multipliers λ_k in (4.8) set to 1. As a result, we show that NIAC is a special case of GARP, when NIAS is true, under the variable map of Theorem 27.

4.3.1 Discussion of Theorem 27

Theorem 27 presents three key results on the unification of revealed preference and revealed rational inattention discussed in more detail below:

1. *Assuming NIAS holds for the unification result.* Recall from Theorem 26 that the first term in the summation in the LHS of the NIAS feasibility condition is the expected utility under the joint distribution $\pi_0(x)p_{i_m}(a|x)$. The variable map in statement (1) of Theorem 27 requires that the surrogate expected utility $G(p_{i_m}(a|x), \boldsymbol{U}_{i_m})$ be equal to the expected utility $\sum_{x,a}\ \pi_0(x)p_{i_m}(a|x)\boldsymbol{U}_{i_m}(x,a)$ that holds only if NIAS is true. In words, the revealed rational inattention test of Theorem 26 checks (a) if the decision maker chooses the optimal action given its posterior belief from a realized observation, and (b) if the decision maker's expected utility from the chosen attention strategy less the information acquisition cost exceeds that for any other attention strategy chosen in the remaining $K-1$ decision problems. Assuming NIAS is true ensures the decision's expected utility $\sum_{x,a}\ \pi_0(x)p_{i_m}(a|x)\boldsymbol{U}_{i_m}(x,a)$ is the maximum possible utility for the decision

maker, where the maximum is taken over all conditional action policies η_{i_m}. Put differently, assuming NIAS to be true only requires the analyst to check NIAC for testing rational inattention, and hence, enables a one-to-one comparison with revealed preference.

2. *Treating the action selection policy as the effective response for the Bayesian agent.* The variable map in statement (1) in Theorem 27 states the the equivalent response in the Bayesian setup is the action selection policy, a noisy version of the unobserved attention strategy chosen by the agent. Although unusual, it suffices for the analyst to treat the action selection policy to be the same as the attention strategy for testing Bayesian rationality; we discuss this in more detail in Sec. 4.2.4.
3. *Relating NIAC and GARP via a variable map.* Statement (1) in Theorem 27 establishes a one-to-one correspondence between revealed preference and revealed rational inattention and relates both approaches via a variable map. The key takeaway is that NIAC is a special case of GARP, when NIAS is true. We see from the variable map in Theorem 27 that in the Bayesian decision framework, the "effective" utility function from revealed preference is the surrogate expected utility G (4.22) that encodes both the prior pmf π_0 and utility $\boldsymbol{U}$, and depends on the observed action selection policy. The cost g in revealed preference translates to the information acquisition cost L in revealed rational inattention.

   Statement (2) introduces a generalization of the forward optimization model considered in [2] for which the NIAC condition generalizes to a GARP-type condition. We formalize the revealed rational inattention test for the generalized model in Corollary 30 below. The generalization of NIAC, namely, GARRI defined in (4.27) in Corollary 30 is equivalent to a Bayesian analog of GARP, thus completely unifying revealed rational inattention and revealed preference.

4. *Generalizing the rational inattention model of [2].* The key distinction between the generalized model (4.25) and the classical rational inattention model considered in [2] specified by (4.16) and (4.17) is the free variable λ_k. Analogous to Theorem 24 where the Lagrange multipliers (4.3) can be interpreted as the marginal utility of the decision maker, λ_k in (4.25) can be interpreted as the marginal cost in the constrained cost minimization problem (4.26). Eq. 4.17 is equivalent to (4.25) with λ_k set to a constant. As a result, the revealed rational inattention test of Theorem 26 yields the cyclic NIAC condition for checking Bayesian rationality (4.17). However, the generalized model of (4.25) has λ_k as a free variable and must be estimated by the external analyst in addition to testing for the existence of an information acquisition cost. The revealed rational inattention test for the generalized model yields the acyclic GARRI condition defined in (4.27) in Corollary 30 below, and is equivalent to GARP under the above variable map.
5. *Change of partial order from revealed preference to revealed rational inattention.* In the variable mapping of Theorem 27, the "response" $\boldsymbol{\alpha}_k$ in the Bayesian setup lies in the unit simplex of probability mass functions. More precisely, the response belongs to the space $\Delta(\mathcal{Y})^{|\mathcal{X}|}$, where $\Delta(\mathcal{Y})$ is the unit simplex of pmfs over the set of signals $\mathcal{Y}$. Clearly, with respect to the natural element-wise partial order of Euclidean spaces, the expected utility $J_{\pi_0}(\boldsymbol{\alpha}, \boldsymbol{U})$ and information acquisition cost $L(\boldsymbol{\alpha})$ are not monotonically increasing in $\boldsymbol{\alpha}$. Hence, the unification result of Theorem 27 involves equipping the space of attention strategies with a different partial order, namely, the Blackwell partial order [138] for probability measures discussed in more detail below.

4.3.2 Change of partial order from revealed preference to revealed rational inattention

The change of partial order from the natural element-wise partial ordering of positive vectors in the Euclidean space for revealed preference to the Blackwell [138] partial ordering of attention strategies is a key component in establishing the one-to-one correspondence in Theorem 27 above. Let us discuss the Blackwell order in detail.

Definition 28 (Blackwell order [138]) *Consider two attention strategies* $\boldsymbol{\alpha}, \bar{\boldsymbol{\alpha}} \in \Delta(\mathcal{Y})^{|\mathcal{X}|}$*, where* $\mathcal{X}$ *and* $\mathcal{Y}$ *denote the finite set of states and private signals, respectively, in the rationally inattentive utility maximization framework of Theorem 26. Then,* $\boldsymbol{\alpha}$ *Blackwell dominates* $\bar{\boldsymbol{\alpha}}$ *(denoted as* $\boldsymbol{\alpha} \geq_{\mathcal{B}} \bar{\boldsymbol{\alpha}}$*) if there exists a row-stochastic matrix* $\boldsymbol{Q}$ *such that* $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}\, \boldsymbol{Q}$*.*

The Blackwell order introduces the notion of monotonicity in the space of attention strategies (probability distributions). The Blackwell order is a partial order, since there exist attention strategy pairs that cannot be ordered via the Blackwell relation (Definition 28). Intuitively, attention strategy $\boldsymbol{\alpha}$ Blackwell dominates $\bar{\boldsymbol{\alpha}}$ if $\bar{\boldsymbol{\alpha}}$ is a noisy (garbled) version of $\boldsymbol{\alpha}$. In classical revealed preference results, the decision maker's response belongs to the Euclidean space. The standard assumption (and key to establishing revealed preferences results) is to impose a monotonicity condition on the decision maker's budget constraint with respect to the element-wise partial ordering for the Euclidean space. The Blackwell partial order can be viewed as a Bayesian analog of the element-wise ordering for rational inattention. Recall from the equivalence result of Theorem 27 that a constraint on the expected utility is the rational inattention analog of the decision maker's budget constraint in revealed preference. For the reader's clarity, we show below the expected

utility is monotone with respect to the Blackwell partial order.[8]

Lemma 29 *Consider the rationally inattentive utility maximization setup in Theorem 25. Suppose NIAS holds, that is, the decision maker chooses the optimal action given its posterior belief. Also, suppose the space of attention strategies* $\Delta(\mathcal{Y})^{|\mathcal{X}|}$ *(probability simplices) are equipped with the Blackwell partial order* $\mathcal{B}$ *(Definition 28). Then, the decision maker's expected utility* (4.22) *is monotonically increasing and convex in the attention strategy.*

The proof of Lemma 29 is in Sec. 4.7.3. Lemma 29 facilitates the one-to-one correspondence between the revealed preference results of Theorem 25 (element-wise partial order over consumption vectors) and the revealed rational inattention result of Corollary 30 (Blackwell partial order over attention strategies).

Remark. Lemma 29 states the expected utility is monotone is the Bayesian decision maker's 'response', namely, the action selection policy under the Blackwell order. However, it is straightforward to show *any convex functional* is monotone wrt the Blackwell order. Indeed, the expected utility (4.22) is convex in the action selection strategy.

To summarize, in revealed preference, the cost of consumption is monotone with respect to the natural element-wise partial order on the Euclidean space. The reconstructed utility function [3] is a monotone function of the consumption cost, and hence, is also monotone with respect to the natural element-wise partial order. In complete analogy, in revealed rational inattention, the *expected* utility is monotone with respect to the Blackwell order on the space of attention strategies. Corollary 30 below presents

[8]That a convex functional in monotone with respect to the Blackwell partial order is well-known in the literature, and stated in the main text for completeness. In future work, we will investigate more general partial orders such as interval dominance and integral precision dominance and how they affect revealed rational inattention results.

an Afriat-type reconstruction of a valid information acquisition cost. The reconstructed information acquisition cost is a monotone function of the expected utility, and hence, is also monotone with respect to the Blackwell order.

Finally, Lemma 29 is particularly useful in justifying why testing for rational inattention (Theorem 26) is possible when only the action selection policies are known; hence, the analyst can treat the action selection policy $p_k(a|x)$ as the Bayesian decision maker's response in decision problem k.

4.3.3 Generalizing Revealed Rational Inattention to Variable Attention Spans

Having stated our equivalence result in Theorem 27 and Lemma 29 above, we now state our second theoretical result, Corollary 30. Recall from statement (1) in Theorem 27 that NIAC is a special case of GARP. Corollary 30 generalizes the revealed rational inattention result of [2, Th. 1] to a decision model with added degrees of freedom to accommodate variable attention spans of the Bayesian decision maker in different decision problems, or equivalent, different marginal costs of information acquisition in the rational inattention setup of (4.17). The key idea is to introduce minimum additional degrees of freedom in the rationally inattentive utility maximization model of [2] so that the rationalizability condition for the *inverse* task generalizes from NIAC to a GARP-type condition. We term the rational inattention analog of GARP as *GARRI*, defined in (4.27) below.

Motivation to generalize the rational inattention model of [2]

Generalizing the revealed rational inattention test of [2] is useful because empirical studies [159–161] on online user engagement data show that online multimedia users have *different* attention spans in different decision problems. In the rational inattention context, different attention spans translate to different marginal costs of information acquisition. The forward optimization model of [2] is restrictive in that the decision maker has the same marginal cost in all decision problems (video categories in the YouTube context). In comparison, the generalized revealed rational inattention test proposed in Corollary 30 below allows for non-constant attention spans of the decision maker in different decision problems. Also, cognitive psychology literature [164–166] suggests the human attention span (hence, the information acquisition cost) is task-dependent (decision problem-dependent), in contrast to the rational inattention model of [2] where the expected utility and information acquisition cost are weighed equally for decision making. The above works serve as motivation to generalize the rational inattention model of [2] to test for non-constant values for the margin cost of information acquisition in datasets aggregated from human decisions.

Corollary 30 *Consider the Bayesian decision maker in Theorem 26. Suppose the analyst knows the dataset* $\mathcal{D}_{RRI}$ (4.14) *and tests for generalized rationally inattentive utility maximization, namely, conditions* (4.16) *and* (4.25)*. In* (4.25)*, the positive utility multiplier* λ_k *is unknown to the analyst. Then, the following statements are equivalent:*
1) There exists a monotone (wrt Blackwell order (Lemma 29)) information acquisition cost $L(\boldsymbol{\alpha})$ *that rationalizes the dataset* $\mathcal{D}_{RRI}$*. That is, the dataset* $\mathcal{D}_{RRI}$ *satisfies the data-matching condition* (4.23)*, and the agent's unobserved attention strategy* $\boldsymbol{\alpha}_k$ (4.10) *and conditional action policy* η_k (4.12) *solve the nested optimization problem* (4.16) *and* (4.25) *for all decision problems* $k = 1, 2, \ldots, K$.

2) The dataset $\mathcal{D}_{RRI}$ satisfies the data-matching condition (4.23)*, NIAS* (4.19) *and the Generalized Axiom of Revealed Rational Inattention (GARRI) defined below. GARRI* (4.27) *is equivalent to GARP* (4.7) *under the variable map of statement (1) in Theorem 27.*

$$\textit{GARRI: } p_k(a|x) \geq_H p_j(a|x) \Rightarrow G(p_k(a|x), \boldsymbol{U}_j) \leq G(p_j(a|x), \boldsymbol{U}_j), \tag{4.27}$$

where the relation $p_k(a|x) \geq_H p_j(a|x)$ means there exists indices $i_1, i_2, \ldots, i_L$ s.t. $G(p_{i_1}(a|x), \boldsymbol{U}_k) \geq G(p_k(a|x), \boldsymbol{U}_k)$, $G(p_{i_2}(a|x), \boldsymbol{U}_{i_1}) \geq G(p_{i_1}(a|x), \boldsymbol{U}_{i_1}), \ldots, G(p_j(a|x), \boldsymbol{U}_{i_L}) \geq G(p_{i_L}(a|x), \boldsymbol{U}_{i_L})$. The surrogate expected utility G in (4.27) *is defined in* (4.22)*.*

3) The dataset $\mathcal{D}_{RRI}$ satisfies the data-matching condition (4.23) *and NIAS* (4.19)*. Also, there exist positive scalars $\lambda_k,\ c_k,\ k = 1, 2, \ldots, K$ such that the following inequalities hold for all pairs of decision variables $k,\ j,\ k \neq j$:*

$$\lambda_k \sum_{x \in \mathcal{X}, a \in \mathcal{A}} p_k(a|x)\, \pi_0(x)\, \boldsymbol{U}_k(x, a) \ - \ c_k \ \geq \ \lambda_k \sum_{a \in \mathcal{A}} p_j(a) \ \left(\max_{b \in \mathcal{A}} \sum_{x \in \mathcal{X}} p_j(x|a)\, \boldsymbol{U}_k(x, b) \right) \ - \ c_j, \tag{4.28}$$

where $p_k(a) = \sum_{x \in \mathcal{X}} \pi_0(x) p_k(a|x)$ is the marginal distribution of the agent's action a in decision problem k, $p_k(x|a) = \pi_0(x) p_k(a|x) / p_k(a)$ is the posterior distribution of the agent's state in decision problem k.

4) If NIAS and GARRI hold, the following reconstructed information acquisition cost $\widehat{L}$ rationalizes the dataset $\mathcal{D}_{RRI}$:

$$\widehat{L}(p(a|x)) = \max_{k \in \{1, 2, \ldots, K\}} \{c_k + \lambda_k (G(p(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k))\} \tag{4.29}$$

where the positive scalars c_k, λ_k are feasible solutions of (4.28)*, and $G(\cdot)$ is the decision maker's surrogate expected utility defined in* (4.22)*.*

The Afriat-type [3] reconstructed cost $\widehat{L}$ in (4.29) *is a point-wise maximum of monotone convex functions, monotone with respect to the Blackwell order on the space of action*

selection policies. The reconstructed cost also satisfies the axiomatic properties of weak monotonicity and mixture feasibility as postulated in [2, Theorem 2].

The proof of Corollary 30 is in Sec. 4.7.4.

Equivalence between GARRI and GARP. Corollary 30 assumes the same problem setting as that of Theorem 26. The only difference in the problem setting between Theorem 26 and Corollary 30 is the additional scalar multiplier λ_k in (4.25). The key takeaway is that the NIAC condition for checking optimality of attention strategies across K decision problems is replaced by the generalized axiom of revealed rational inattention (GARRI) (4.27). Indeed, NIAC is a special case of GARRI since (4.17) is a special case of (4.25) with $\lambda_k = 1$ for all k. In fact, the addition of the non-constant multiplier λ_k is the minimum modification needed in the decision model of [2] for checking the optimality of the chosen attention strategies via the cyclical consistency of condition of GARP, instead of the more restrictive cyclical monotonicity condition of NIAC.

Afriat-type reconstruction of information acquisition cost. Corollary 30 builds on [2] and reconstructs an Afriat-type monotone convex cost of information acquisition (4.29) that rationalizes the dataset $\mathcal{D}_{RRI}$ (4.14). Afriat's [3] utility reconstruction involves 'stitching' a piece-wise linear, concave utility function that rationalizes the agent's actions and is expressed in terms of the utility value earned by the agent at every time step and the budget constraints. In complete analogy, the piece-wise convex monotone cost (4.29) rationalizes the dataset $\mathcal{D}_{RRI}$ (4.14) and is expressed in terms of the information acquisition cost incurred by the agent in every decision problem and the surrogate expected utility functional G. Recall from Sec. 4.2.4 that the analyst performing revealed rational inattention can treat the observed action selection policy as the unobserved attention strategy chosen by the Bayesian agent. Hence, the *reconstructed* cost of information acquisition is a function of the observed variable, namely, the action selection policy.

The reconstructed utility function in [3] is locally non-satiated, monotone and concave, and rationalizes the observed price and consumption bundles. In complete analogy, the reconstructed cost (4.29) is weakly monotonic (in information (Blackwell partial order [138])), mixture feasible (convex) and normalized, and rationalizes the observed utility functions and action selection strategies in $\mathcal{D}_{RRI}$ (4.14).

Remark. The authors of [2] generalize their results to posterior separable costs of information acquisition in [167]. Specifically, [167, Theorem 2] provides a constructive procedure for recovering a posterior separable cost that satisfies the rationalizability axioms for optimality in decision making under posterior cost constraints. Apart from the Afriat-style of cost construction, the construction style in [167, Theorem 2] differs from that in Theorem 27 in one key aspect. Our reconstruction procedure assumes the inverse learner only has access to a finite set of agent utility functions from finitely many environments, that is, the completeness axiom in [167, Axiom A4] is *not* assumed.

4.4 Related Works

There are several works in the economics literature that generalize classical results of revealed preference and revealed rational inattention. In this section, we compare GARRI (4.27), the Bayesian analog of GARP, in Corollary 30 with related existing works.

GARRI and GAPP. [168] test for the existence a preference relation over *prices* set by the buyer, instead of a preference relation over consumption bundles. The key testable axiom is Generalized Axiom of Price Preference (GAPP). Corollary 30 proposes a GARP-type condition, GARRI that generalizes NIAC. In complete analogy, GARP is a generalization of the GAPP condition of [168]. A key point is that GAPP generalizes

(and is not equivalent to) the test for quasi-linear utility maximization, even though the above relations might suggest drawing this conclusion. We discuss the relation between GARI and revealed preference test for quasi-linear utility maximization below.

GARRI, NIAC and Strong law of demand. We now relate revealed rational inattention (Corollary 30) and revealed preference for quasi-linear utility maximization [135]. Theorem 2.2 in [135] proposes testable conditions for quasilinear utility maximization and term this rationalization as strong law of demand. Test for quasilinear utility maximization checks for the existence of a utility function U such that the following condition holds for $k = 1, 2, \ldots, K$:

$$\beta_k \in \underset{\beta:\nu_k'\beta\leq 1}{\operatorname{argmax}}\, u(\beta) - \nu_k'\beta - 1,\ \nu_k'\beta_k = 1. \tag{4.30}$$

The objective function of (4.30) is a non-Bayesian analog of the decision framework of [2], with the response β_k replaced with $p_k(a|x)$, utility $u_k(\cdot)$ replaced with $J_{\pi_0}(\cdot, \boldsymbol{U}_k)$, and cost $\nu_k'(\cdot)$ replaced with the information acquisition cost $L(\cdot)$. Under the variable map of Theorem 27, the cyclical monotonicity condition of [135, Definition 3], a special case of GARP (4.7), is equivalent to NIAC (4.21), a special case of GARRI (4.27). Testing for quasilinear utility maximization [135, Theorem 2.2] is equivalent to checking for the feasibility of Afriat's inequalities with constant Lagrange multipliers. In complete analogy, testing for the classical rational inattention setup of [2] is equivalent to checking for the feasibility of (4.28) with the Lagrange multipliers set to a constant.

GARRI and GACI. [137,169] generalize [2] to the case where the decision maker maximizes a non-separable objective function. The key axiom that generalizes NIAC to the non-separable case is the *Generalized Axiom of Costly Information* (GACI) [137, Condition 1], that possesses the cyclical monotonicity structure of GARP. However, on careful examination, we observed that the variable map from the Bayesian to non-Bayesian decision framework in [137] is distinct from the one proposed in Theorem 27 in spite

of the GARP flavor in both GACI and our proposed generalization of NIAC, namely, GARRI (4.27) in Corollary 30. [137] relate the revealed attention strategy $p_k(a|x)$ to the price of a good, and the expected utility functional $J_{\pi_0}(\cdot, \boldsymbol{U}_k)$ to the response of the decision maker in revealed preference. However, in our equivalence result, the expected utility functional is analogous to the utility constraint (4.5), and attention strategy is analogous to the decision maker's response in the modified revealed preference setup in Sec. 4.2.2. The difference in the variable map between GARRI and [137] to GARP can be attributed to Theorem 25 that yields a GARP-type condition for testing if the decision maker minimizes an unobserved cost subject to a lower bound on its utility. We remark that the decision model of [137] accommodates both models of [2] and Corollary 30 as special cases. However, the addition of the scalar multiplier in (4.25) is the *minimum* modification required in the decision framework of [2] for the dataset to be rationalized by a GARP-type condition.

4.5 Extending Robustness Measures in Revealed Preference to Revealed Rational Inattention

We now exploit the equivalence result of Theorem 27 and the generalized revealed rational inattention result of Corollary 30 to construct robustness measures for the revealed rational inattention test. The key idea is to compute the minimum perturbation needed for a dataset to pass the revealed rational inattention test, namely, the feasibility of NIAS and GARRI conditions in Corollary 30. There are several works in the revealed preference literature that characterize how far a sequence of budget constraints and consumption bundles is from satisfying GARP (4.2). To the best of our knowledge, there

is no formal approach in the literature to measure how well a dataset $\mathcal{D}_{RRI}$ (4.14) fits the rational inattention model.

Abstractly, the key idea behind the robustness measures in revealed preference is to minimally perturb the observed dataset so that GARP holds. A few notable robustness measures include:

1. The 'Afriat Efficiency Index (AEI)' [156] that yields the minimum relaxation (expenditure wastage) needed in the budget constraints to rationalize the data.

2. The chi-squared 'Minimal Perturbation Test (MPT)' [158] where the analyst assumes an additive measurement error in the observed response, and performs a chi-squared test on the minimum $\mathcal{L}_2$-deviation from the observed responses such that the perturbed responses rationalize the dataset.

3. The 'Money Pump Index (MPI)' [170] that yield the maximum profit a seller can make from a dataset violating GARP.

4. The 'Minimum Cost Index (MCI)' [171] that yields the lowest normalized cost of breaking all revealed preference cycles from a dataset.

5. The 'Houtman-Maks Index (HMI)' [157] that, for a specified rationalizability tolerance, outputs the largest subset that satisfies GARP.

In this section, we exploit the equivalence result of Theorem 27 and extend the robustness measures described above to revealed rational inattention. For brevity, we only discuss the rational inattention analogs of robustness measures 1-3 above, namely, the Afriat Efficiency Index (AEI), Minimal Perturbation test (MPT) and the Money Pump Index (MPI). However, we emphasize that *any* robustness measure from the revealed preference literature can be extended to the revealed rational inattention test via the unification result of Theorem 27.

4.5.1 Afriat Efficiency Index for Rational Inattention (RI-AEI)

In the classical revealed preference setup with linear budget constraints, the Afriat Efficiency Index (AEI) [156] is a uniform lower bound on the scalar multiplier e so that GARP holds for a dataset $\{\nu_k, \beta_k\}_{k=1}^K$ The variables ν_k and β_k denote the price vector and consumption bundle at time k, the consumer's budget constraint is given by $\nu_k'\beta_k \leq 1$. AEI is defined as:

$$\text{Afriat Efficiency Index } (\text{AEI}) = \underset{e \in \geq 0}{\text{argmin}} \ e, \text{ such that GARP}(e) \text{ holds}, \tag{4.31}$$

where GARP(e) is a generalization of GARP defined as:

$$\mathbf{1}\{e\ \nu_k'\beta_k \geq \nu_k'\beta_j\}\ \nu_j'(e\ \beta_j - \beta_k) \leq 0,\ \forall\ j, k,\ j \neq k,\ e \geq 0 \tag{4.32}$$

In (4.32) above, $\mathbf{1}\{\cdot\}$ is the indicator function. In words, the constraint in (4.32) says that, for a relaxation level e, if β_j is e-affordable at time k, then it must be that β_k must not be e-affordable at time j. Clearly, setting $e = 1$ in (4.32) yields the classical GARP condition (4.2) for linear budget constraints. The parameter e can be viewed as a relaxation of the GARP condition. Hence, AEI measures the *minimum* relaxation in budget constraints needed for the dataset to satisfy utility maximization behavior.

It is straightforward to show that, for a fixed value of e, checking if GARP(e) (4.32) holds is equivalent to checking for the feasibility of the following set of linear inequalities:

Find non-negative scalars λ_k, u_k such that $u_j - u_k - \lambda_k\ \nu_k'(\beta_j - e\ \beta_k) \leq$

for all index pairs $j, k,\ j \neq k$.

(4.33)

Hence, AEI for the dataset $\{\nu_k, \beta_k\}_{k=1}^K$ can be computed as:

$$\boxed{\text{AEI} = \underset{e, \{\lambda_k, u_k\}_{k=1}^K \geq 0}{\text{argmin}} \ e, \ \text{ such that } u_j - u_k - \lambda_k\ \nu_k'(\beta_j - e\ \beta_k) \leq 0}, \text{ for all index pairs } j, k,\ j \neq k. \tag{4.34}$$

If the dataset $\{\nu_k, \beta_k\}_{k=1}^K$ satisfies Afriat's inequalities for utility maximization behavior, then AEI≥ 1 when computed via (4.34).[9] We now extend AEI to the revealed rational inattention setup of Corollary 30 by invoking the equivalence result of Theorem 27. For clarity, we term AEI for the revealed rational inattention case as *Rationally Inattentive*-AEI (RI-AEI).

Definition 31 (Rationally Inattentive Afriat Efficiency Index (RI-AEI)) *Consider an external analyst with the stochastic choice dataset* $\mathcal{D}_{RRI}$ (4.14)*. The rationally inattentive Afriat efficiency index (RI-AEI) is the* minimum *relaxation in expected utility constraints* (4.26) *required for the dataset to be consistent with rationally inattentive utility maximization behavior. RI-AEI is defined as:*

$$\text{RI-AEI} = \underset{e\geq 1, \{\lambda_k, c_k\}_{k=1}^K \geq 0}{\text{argmin}} \; e, \quad \textit{such that NIAS (4.28) and the following set of inequalities hold:}$$

$$c_j - c_k - \lambda_k \left(\sum_{a\in\mathcal{A}} p_j(a) \left(\max_{b\in\mathcal{A}} \sum_{x\in\mathcal{X}} p_j(x|a)\, \boldsymbol{U}_k(x,b) \right) - e \sum_{x\in\mathcal{X}, a\in\mathcal{A}} p_k(a|x)\, \pi_0(x)\, \boldsymbol{U}_k(x,a) \right) \geq 0,$$

for all index pairs $j, k \in \{1, 2, \ldots, K\},\ j \neq k.$

(4.35)

If $e = 1$ satisfies the feasibility inequality in (4.35) above, then the dataset is consistent with rationally inattentive utility maximization. If not, then the minimum value of e for which (4.35) has a feasible solution is bounded from below by 1, in contrast to AEI (4.31), where the minimum perturbation needed for a dataset to satisfy GARP is less than unity. This difference arises due to the decision maker's constraint in the rationally inattentive case. Recall from (4.1) that for the non-Bayesian decision model in revealed preference, the decision maker faces an *upper* bound on its budget, whereas in the rational

[9] Indeed, GARP(e) is equivalent to the square matrix $A(e) = [\nu_k'(\beta_j - e\ k)]_{k,j=1}^K$ satisfying GARP. Using the relation between the elements of the square GARP matrix to the Afriat inequalities [172, Th. 2], the constraint in (4.34) results as an equivalent formulation for GARP(e).

inattention setup of Corollary 30, the decision maker faces a *lower* bound on the expected utility (4.26).

4.5.2 Minimum Perturbation Test for Rational Inattention (RI-MPT)

The minimum perturbation test (MPT) introduced in [158] assumes the decision maker's chosen consumption bundles are measured in noise, and computes the *minimum* perturbation needed in the consumption bundles for the dataset to be consistent with utility maximization behavior.

Suppose the analyst has a noisy dataset $\widehat{\mathcal{D}}_{RP} = \{\nu_k, z_k\}_{k=1}^K$, where $z_k = \beta_k + n_k$ is a noisy version of the true response β_k unobserved by the analyst, and the measurement error $n_k \sim f_n$ is an i.i.d. random variable with pdf f_n. Assume, WLOG, that the decision maker's budget constraint at time k is given by $\nu_k'\beta_k \leq 1$. Let us now introduce the null and alternate hypotheses H_0 and H_1:
H_0: The true (noiseless) dataset $\mathcal{D}_{RP} = \{\nu_k, \beta_k\}_{k=1}^K$ satisfies GARP (4.2), H_1: The true dataset $\mathcal{D}_{RP}$ does NOT satisfy GARP (4.2).
The analyst then performs MPT on the noisy dataset $\widehat{\mathcal{D}}_{RP}$, namely, computes a test statistic defined below and performs a hypothesis test to reject or accept the null hypothesis

H_0:

$$\text{MPT} : \phi(\widehat{\mathcal{D}}_{RP}) \underset{H_1}{\overset{H_0}{\lessgtr}} \eta_f, \text{ where the test statistic } \phi(\cdot) \text{ is defined as:}$$

$$\phi(\widehat{\mathcal{D}}_{RP}) = \min_{\varepsilon_{1,2,\ldots,K} \in \mathbb{R}^m, \{\lambda_k, u_k\}_{k=1}^K \geq 0} \sum_{k=1}^K \|\varepsilon_k\|_2^2, \text{ such that}$$

$$(i)\ \beta_k + \varepsilon_k \geq \mathbf{0},\ \nu_k'(\beta_k + \varepsilon_k) = 1,\ (ii)\ u_j - u_k - \lambda_k\, \nu_k'(\beta_j + \varepsilon_j - \beta_k - \varepsilon$$

for all index pairs $j, k,\ j \neq k$,

(4.36)

where η is a parameter that bounds the detector's Type-I error probability $\mathbb{P}(H_1|H_0)$. The rationale behind (4.36) is that if H_0 holds, then $\phi(\widehat{\mathcal{D}}_{RP})$ is a lower bound on the measurement error $\sum_{k=1}^K \|\beta_k - z_k\|_2^2$. This observation further implies (see [36, 173] for details) that the Type-I error probability of the hypothesis test is upper bounded by $F_w^{-1}(\eta_f)$, where $F_w(\cdot)$ is the cdf of the noise pdf f_w.

We now extend MPT (4.36) to the rationally inattentive utility maximization setup of [2] by exploiting the equivalence result of Theorem 27. Suppose the analyst has a noisy dataset $\widehat{\mathcal{D}}_{RRI} = \{\pi_0, \{\hat{p}_k(a|x), \boldsymbol{U}_k\}_{k=1}^K\}$, where $\hat{p}_k(a|x)$ is a noisy version of the true action selection policy $p_k(a|x)$.[10] For clarity, we term MPT for the revealed rational inattention case as *Rationally Inattentive*-MPT (RI-MPT).

Definition 32 (Rationally Inattentive Minimum Perturbation Test (RI-MPT)) *Consider an external analyst with the stochastic choice dataset $\mathcal{D}_{RRI}$* (4.14)*. The rationally inattentive minimum perturbation test (RI-MPT) is the* minimum *perturbation in the observed action selection policies required for the dataset to be consistent with rationally inatten-*

[10] Noise in probability mass functions may arise due to multiple factors such as misspecification error, or computing the action selection policy empirically from a finite number of samples; see [143] for a finite sample analysis of the revealed preference test.

tive utility maximization behavior. The RI-MPT is defined as:

$$\text{RI-MPT}: \phi(\widehat{\mathcal{D}}_{RRI}) \underset{H_1}{\overset{H_0}{\lessgtr}} \eta_f, \textit{ where the test statistic } \phi(\cdot) \textit{ is defined below:}$$

$$\phi(\widehat{\mathcal{D}}_{RRI}) = \min_{\{\tilde{p}_m(a|x),\lambda_k,u_k\}_{k=1}^K \geq 0} \sum_{k=1}^{K} \sum_{x\in\mathcal{X},a\in\mathcal{A}} \|\tilde{p}_m(a|x) - p_m(a|x)\|_2^2, \textit{ such that}$$

$$(i) \sum_{a\in\mathcal{A}} \tilde{p}_k(a|x) = 1 \; \forall, x, k \; (\textit{Valid pmf})$$

$$(ii)\; NIAS: \sum_{x\in\mathcal{X}} \tilde{p}_k(x|a)(\boldsymbol{U}_k(x,a) - \boldsymbol{U}_k(x,b)) \geq 0 \; \forall \; k, a, b,$$

$$(iii)\; GARRI: c_j - c_k - \lambda_k \Bigg(\sum_{a\in\mathcal{A}} \tilde{p}_j(a) \max_{b\in\mathcal{A}} \sum_{x\in\mathcal{X}} \tilde{p}_j(x|a)\, \boldsymbol{U}_k(x,b) - \sum_{x\in\mathcal{X},a\in\mathcal{A}} \tilde{p}_k(a|x)\, \pi_0(x)\, \boldsymbol{U}_k(x,a) \Bigg) \geq 0, \textit{ for all index pairs } j,k,\; j \neq k. \tag{4.37}$$

In (4.37) *above,* $\tilde{p}_k(a) = \sum_{x\in\mathcal{X}} \tilde{p}_m(a|x)\pi_0(x)$ *is the marginal action probability, and* $\tilde{p}_k(x|a) = \frac{\tilde{p}_k(a|x)\pi_0(x)}{\sum_{x'\in\mathcal{X}} \tilde{p}_k(a|x')\pi_0(x')}$ *is the posterior state distribution given action selection policy* $\tilde{p}_k(a|x)$.

RI-MPT defined in (4.37) above is a hypothesis test that considers the minimum perturbation needed in the action selection policies for the feasibility of NIAS and GARRI conditions as the sufficient statistic for the test. In complete analogy to (4.36), the variable η_f in (4.37) controls the Type-I error probability of detecting Bayesian rationality, that is, NIAS and GARRI conditions hold for the noise-less dataset $\mathcal{D}_{RRI}$ (4.14).

4.5.3 Money Pump Index for Rational Inattention (RI-MPI)

The money pump index (MPI) introduced in [170] quantifies the severity of violations of GARP. If the decision maker's choices are observed in noise, the computed value of MPI can be used to test if the decision maker is rational or not. In this section, we consider

the case where the decision maker's responses are measured accurately. MPI is defined for a sequence of tuples of prices and consumption bundles that violate GARP. Consider the classical revealed preference setup in Theorem 24 with linear budget constraints $\nu_k'\beta_k \leq 1$. Suppose the sequence $\{\nu_{k_i}, \beta_{k_i}\}_{i=1}^{l}$ violates GARP ($l \leq K$). Then MPI for the violating sequence is defined as:

$$\boxed{\text{MPI}_{\{\nu_{k_i},\beta_{k_i}\}_{i=1}^{l}} = \frac{1}{l}\sum_{i=1}^{l}(\nu_{k_{i+1}} - \nu_{k_i})'(\beta_{k_{i+1}} - \beta_{k_i})} \qquad (k_{l+1} \equiv k_1) \qquad (4.38)$$

If GARP fails for the sequence $\{\nu_{k_i}, \beta_{k_i}\}_{i=1}^{l}$, it is straightforward to show that MPI$(\{\nu_{k_i}, \beta_{k_i}\}_{i=1}^{l}) > 0$ in (4.38). Intuitively, MPI measures the profit a malicious arbitrager can make by buying the consumption bundles in the GARP-violating sequence at a lower price, and selling the same bundles to the non-rational decision maker at a higher price.

We now extend MPI (4.38) to the rationally inattentive utility maximization setup of [2] by exploiting the unification result of Theorem 27. We term MPI for the revealed rational inattention case as *Rationally Inattentive*-MPI (RI-MPI).

Definition 33 (Rationally Inattentive Money Pump Index (RI-MPI)) *Consider an external analyst with the stochastic choice dataset $\mathcal{D}_{RRI}$* (4.14)*. The rationally inattentive money pump index RI-MPI is defined as:*

$$\text{RI-MPI} = \max_{\{\nu_{k_{1:l}},\beta_{k_{1:l}},\, l\leq K\}} \frac{\sum_{i=1}^{l} \bar{J}_{\pi_0,k_i,k_{i+1}}(p_{k_i}(a|x), p_{k_{i+1}}(a|x))}{\sum_{i=1}^{l} J_{\pi_0}(p_{k_i}(a|x), \boldsymbol{U}_{k_i})} \qquad (4.39)$$

where $k_{l+1} \equiv k_1$ and $\bar{J}_{\pi_0,k_i,k_{i+1}}(p_{k_i}(a|x), p_{k_{i+1}}(a|x))$ defined below is the net expected utility a malicious arbitrager can gain by exploiting the fact that the Bayesian decision maker's choices fail the GARRI (4.27) *condition:*

$$\bar{J}_{\pi_0,k_i,k_{i+1}}(p_{k_i}(a|x), p_{k_{i+1}}(a|x)) = J_{\pi_0}(p_{k_{i+1}}(a|x), \boldsymbol{U}_{k_i}) - J_{\pi_0}(p_{k_i}(a|x), \boldsymbol{U}_{k_i})$$
$$+ J_{\pi_0}(p_{k_i}(a|x), \boldsymbol{U}_{k_{i+1}}) - J_{\pi_0}(p_{k_{i+1}}(a|x), \boldsymbol{U}_{k_{i+1}}) \qquad (4.40)$$

In (4.40)*, $J_{\pi_0}(\cdot)$ is the expected utility functional defined in* (4.18)*.*

In [170], the money pump index is defined for a sequence of indices for which GARP fails. In complete analogy, (4.39) in Definition 33 computes the *maximum* 'profit' a malicious arbitrager can make over all possible sequences of decision problems combinations, normalized by the sum of expected utilities of the Bayesian decision maker in the decision problems. The extent of irrationality of the Bayesian maker that facilitates arbitrage is captured by the variable $\bar{J}_{\pi_0,k_i,k_{i+1}}(p_1(a|x), p_2(a|x))$ in (4.39) and discussed below in more detail.

Let us briefly discuss the intuition behind RI-MPI (4.39). Without loss of generality, suppose GARRI fails for indices $1, 2$ (sequence of length 2), which implies the following set of inequalities hold:

$$\begin{aligned} & J_{\pi_0}(p_2(a|x), \boldsymbol{U}_1) \ \geq \ J_{\pi_0}(p_1(a|x), \boldsymbol{U}_1) \ \text{ and } J_{\pi_0}(p_1(a|x), \boldsymbol{U}_2) \ \geq \ J_{\pi_0}(p_2(a|x), \boldsymbol{U}_2) \\ \implies & \underbrace{J_{\pi_0}(p_2(a|x), \boldsymbol{U}_1) - J_{\pi_0}(p_1(a|x), \boldsymbol{U}_1)}_{\geq 0} + \underbrace{J_{\pi_0}(p_1(a|x), \boldsymbol{U}_2) \ \geq \ J_{\pi_0}(p_2(a|x), \boldsymbol{U}_2)}_{\geq 0} \geq 0 \\ \implies & \bar{J}_{\pi_0,1,2}(p_1(a|x), p_2(a|x)) \geq 0 \end{aligned} \tag{4.41}$$

The term $\bar{J}_{\pi_0,1,2}(p_1(a|x), p_2(a|x))$ measures the excess expected utility a malicious arbitrager can gain by 'buying' choices $p_2(a|x), \ p_1(a|x)$ when presented with utilities $\boldsymbol{U}_1, \ \boldsymbol{U}_2$ in decision problems $1, 2$, respectively, and selling choices $p_2(a|x), \ p_1(a|x)$ to the Bayesian decision maker in decision problems $2, 1$, respectively.

Summary. In this section, we extended three robustness measures from revealed preference to the revealed rational inattention result of Corollary 30. Specifically, we extended the Afriat Efficiency Index (AEI) [156], Varian's [158] Minimum Perturbation Test (MPT) and the Money Pump Index (MPI) [170] to the Bayesian case. We now illustrate the Bayesian analogs of AEI, MPT and MPI on a real-world YouTube metadata comprising user engagement from approximately 140,000 videos. We characterize, using the robustness measures for rational inattention defined above, the goodness-of-fit to the

YouTube dataset to the rationally inattentive utility maximization model.

4.6 Conclusion and Future Work

In this chapter, we established the connection between revealed preference and revealed rational inattention. Our main finding is that the NIAC condition [2] in revealed rational inattention is a special case of GARP [77] in revealed preference under a different partial order and a different state space (probability simplex). We exploit this result to construct a monotone convex information acquisition cost in revealed rational inattention. The construction procedure resembles that of the utility function reconstructed from consumer data in [3]. Due to the equivalence result, we adapt goodness-of-fit measures from revealed preference to the revealed rational inattention case and characterize how well a dataset fits the rational inattention model. Finally, we characterize the goodness-of-fit of a massive dataset of YouTube metadata from 140,000 videos using the adapted robustness measures. To the best of our knowledge, our numerical experiments are a novel exercise on systematically analyzing the goodness-of-fit of decision data from Bayesian decision makers to the rational inattention model.

In future work it is worthwhile exploiting this unification to study revealed preference in Bayesian versions of potential games building on [174, 175], market games building on [4], inverse reinforcement learning building on [6, 143], and dynamic revealed preference building on [176].

4.7 Appendix

4.7.1 Proof of Theorem 25

Statement (1)⇒(2).

Fix indices j, k and assume $\beta_k \geq_H \beta_j$. From the definition of the relation '$\geq_H$' (4.7), there exist indices $i_1, i_2, \ldots, i_L$ such that such that $u_k(\beta_{i_1}) \geq u_k(\beta_k),\ u_{i_1}(\beta_{i_2}) \geq u_{i_1}(\beta_{i_1}),\ \ldots, u_{i_L}(\beta_j) \geq u_{i_L}(\beta_{i_L})$. Since $g(\cdot)$ rationalizes the decision maker's dataset, we must have $g(\beta_k) \leq g(\beta_{i_1}) \leq g(\beta_{i_2}) \leq \ldots \leq g(\beta_{i_L}) \leq g(\beta_j)$ which, in turn, implies $g(\beta_k) \leq g(\beta_j)$.

Our aim is to show $u_j(\beta_j) \geq u_j(\beta_k)$. We prove this by contradiction. Suppose $u_j(\beta_j) < u_j(\beta_k)$. Then, from the continuity of $u_j(\cdot)$ and monotonicity of $g(\cdot)$, we could find a $\beta \in \mathbb{R}^m_+$ in the neighborhood of β_k such that the β satisfies the two conditions below:

(a) $u_j(\beta_j) < u_j(\beta) \leq u_j(\beta_k)$ (β satisfies utility constraint (4.5) for time step j) and

(b) $g(\beta) < g(\beta_k) \leq g(\beta_j)$, or equivalently, $g(\beta) \leq g(\beta_j)$ (β *strictly* costs lesser compared to β_j).

Clearly, if both (a) and (b) hold, then β_j does not rationalize the decision maker's action at time j, i.e. , g does not rationalize the analyst's dataset, hence our assumption is false. Therefore, it must be the case that $\boxed{u_j(\beta_j) \geq u_j(\beta_k)}$.

Statement (2)⇒(3). Construct a matrix $A \in \mathbb{R}^{K \times K}$ with elements $A_{j,k} = u_k(\beta_k) - u_k(\beta_j)$. Since GARP (4.7) holds, it is trivial to show the matrix A is cyclically consistent. From [4, Lemma 2] and [172, Sections 2 and 3], there exist scalars $\bar{g}_k$ and $\lambda_k > 0$ that satisfy the following set of Afriat-type (4.3) feasibility inequalities:

$$\bar{g}_j - \bar{g}_k - \lambda_k(u_k(\beta_k) - u_k(\beta_j)) \leq 0 \quad \forall\, k, j. \tag{4.42}$$

Although (4.42) resembles the Afriat inequalities in Theorem 25, we cannot reconstruct the decision maker's cost that rationalizes its choices (4.5) without modifying the feasible variables $\bar{g}_k,\ \lambda_k$. However, if we were to reconstruct the cost g using $\bar{g}_k$ in (4.42), we would obtain a decreasing function that violates the properties of the cost. To alleviate this issue and have a monotone reconstruction of the cost, we perform the following change of variables:

Without loss of generality, we restrict $\bar{g}_k$ to be finite for all $k = 1, 2, \ldots, K$. Let $M < \infty$ denote an arbitrary positive scalar that uniformly bounds $\{\bar{g}_k\}_{k=1}^K$ from above. Note that since $\bar{g}_k$ is bounded for all k, such an M exists and $\bar{g}_k > 0$. We now define the variable $g_k = M - \bar{g}_k$ for all k and rewrite (4.42) in terms of the new variables $\{g_k\}_{k=1}^K$:

$$\begin{aligned}
&\bar{g}_j - \bar{g}_k - \lambda_k(u_k(\beta_k) - u_k(\beta_j)) \leq 0,\ \forall\, k, j \\
\Leftrightarrow &-\bar{g}_j - (-\bar{g}_k) - \lambda_k(u_k(\beta_j) - u_k(\beta_k)) \geq 0,\ \forall\, k, j \\
\Leftrightarrow &(M - \bar{g}_j) - (M - \bar{g}_k) - \lambda_k(u_k(\beta_j) - u_k(\beta_k)) \geq 0,\ \forall\, k, j \\
\Leftrightarrow &\boxed{g_j - g_k - \lambda_k(u_k(\beta_j) - u_k(\beta_k)) \geq 0,\ \forall\, k, j \equiv (4.8)}.
\end{aligned} \tag{4.43}$$

Consider the reconstructed cost $g(\beta) = \max_k\{g_k + \lambda_k(u_k(\beta) - u_k(\beta_k))\}$. Clearly, g is monotone and continuous since it is a point-wise maximum of monotone continuous functions. We now show that $g_{recon}(\beta) = \max_k\{g_k + \lambda_k(u_k(\beta) - u_k(\beta_k))\}$ rationalizes the dataset $\{\beta_k, u_k(\cdot) \geq u_k^*\}_{k=1}^K$:

Fix index k. Then, $g_{recon}(\beta_k) \geq g_k$ by definition. Also, from (4.43), we have that:

$$g_k \geq g_j + \lambda_j(u_j(\beta_k) - u_j(\beta_j))\ \forall j \neq k$$

Therefore, it is clear that $g_{recon}(\beta_k) = g_k$ (4.43). Now, let β denote any feasible consumption bundle at time k, that is, $u_k(\beta) \geq u_k^* = u_k(\beta_k)$. Then:

$$g_{recon}(\beta) = \max_k \{g_k + \lambda_k(u_k(\beta) - u_k(\beta_k))\}$$

$$\geq \underbrace{g_k}_{=g_{recon}(\beta_k)} + \underbrace{\lambda_k(u_k(\beta) - u_k(\beta_k))}_{\geq 0} \geq g_{recon}(\beta_k)$$

$$\Rightarrow \boxed{\beta_k = \operatorname*{argmin}_{\beta} \; g_{recon}(\beta) \text{ s.t. } \; u_k(\beta) \geq u_k^*} \tag{4.44}$$

Since (4.44) holds for all $k = 1, 2, \ldots, K$, the reconstructed cost $g_{recon}(\beta) = \max_k\{g_k + \lambda_k(u_k(\beta) - u_k(\beta_k))\}$ rationalizes the analyst's dataset $\{\beta_k, u_k(\cdot) \geq u_k^*\}_{k=1}^K$.

Statement (3)⇒(1). Consider the reconstructed cost $g_{recon}(\beta) = \max_k\{g_k + \lambda_k(u_k(\beta) - u_k(\beta_k))\}$. By construction, the cost g_{recon} is monotone, continuous (since it is a pointwise maximum of finitely many monotone continuous segments) and rationalizes the decision maker's actions (4.44). ■

4.7.2 Proof of Theorem 27

The proof of our unification result, namely, Theorem 27, comprises two steps. First, we show the NIAC condition can be expressed as an equivalent feasibility inequality. Second, under the variable map of statement 1 of Theorem 1, we compare the equivalent inequality with the Afriat-type inequality (4.8) in Theorem 25 for testing if non-Bayesian cost minimization (4.5) holds. Finally, statement (2) in Theorem 27 is proved in Sec. 4.7.4.

Expressing the NIAC condition as an equivalent feasibility inequality

Suppose NIAC (4.21) is true. Then, one can show using the concept of KKT conditions from duality theory [25, Sec. 5.5] that there exist non-negative scalars $\{c_k\}_{k=1}^K$ that satisfy the following inequalities:

$$\boxed{G(p_k(a|x), \boldsymbol{U}_k) - c_k \geq G(p_j(a|x), \boldsymbol{U}_k) - c_j, \; \forall \, j, k = 1, 2, \ldots, K}. \tag{4.45}$$

In (4.45), the scalars $\{c_k\}_{k=1}^K$ are Lagrange multipliers corresponding to an equivalent linear assignment problem [163] that is solved by the identity map if the NIAC condition is true. We refer the reader to [143, Sec. C.2.2] and [2, Sec. 10.2] for a more elaborate discussion on the existence of the scalars $\{c_k\}_{k=1}^K$ that satisfy (4.45) if NIAC is true.

To prove equivalence between (4.45) and NIAC, it suffices to show that if there exist non-negative scalars $\{c_k\}_{k=1}^K$ that satisfy (4.45), then NIAC holds. Fix a sequence of indices $k_1, k_2, \ldots, k_m,\ m \leq K$, where $k_i \in \{1, 2, \ldots, K\}$, $\forall i = 1, 2, \ldots, m$. Since there exists a feasible solution $\{c_k\}_{k=1}^K$ to the inequality (4.45), the following inequalities result:

$$
\begin{aligned}
&G(p_{k_1}(a|x), \boldsymbol{U}_{k_1}) - c_{k_1} \geq G(p_{k_2}(a|x), \boldsymbol{U}_{k_1}) - c_{k_2}, \\
&G(p_{k_2}(a|x), \boldsymbol{U}_{k_2}) - c_{k_2} \geq G(p_{k_3}(a|x), \boldsymbol{U}_{k_2}) - c_{k_3}, \\
&\qquad\qquad \ldots \\
&G(p_{k_{m-1}}(a|x), \boldsymbol{U}_{k_{m-1}}) - c_{k_{m-1}} \geq G(p_{k_m}(a|x), \boldsymbol{U}_{k_{m-1}}) - c_{k_m}, \\
&G(p_{k_m}(a|x), \boldsymbol{U}_{k_m}) - c_{k_m} \geq G(p_{k_1}(a|x), \boldsymbol{U}_{k_m}) - c_{k_1} \\
&\Rightarrow \sum_{i=1}^{m} \Big(G(p_{k_i}(a|x), \boldsymbol{U}_{k_i}) - G(p_{k_{i+1}}(a|x), \boldsymbol{U}_{k_i}) \Big) \geq 0, \text{ where } k_{m+1} = k_1 \\
&\underset{\text{(if NIAS holds)}}{\Leftrightarrow} \sum_{i=1}^{m} \Bigg(\sum_{x,a} \pi_0(x) p_{k_i}(a|x) \boldsymbol{U}_{k_i}(x,a) - G(p_{k_{i+1}}(a|x), \boldsymbol{U}_{k_i}) \Bigg) \geq 0, \text{ where } k_{m+1} = k_1 \\
&\equiv \text{NIAC (4.21)}
\end{aligned}
$$

Hence, $\boxed{\text{NIAC (4.21)} \equiv \text{ there exists feasible non-negative scalars } \{c_k\}_{k=1}^K \text{ that satisfy (4.45)}}$.

■

Relating the feasibility inequalities for NIAC (4.45) and GARP (4.8) under the variable map of Theorem 27

We now relate the feasibility inequality (4.45) to the Afriat-type revealed preference inequality (4.8):

$$\text{NIAC} \underset{\text{if NIAS holds}}{\Leftrightarrow} (4.45):$$

Given dataset $\mathcal{D}_{RRI}$ (4.14), there exist $\{c_k\}_{k=1}^K \geq 0$ such that (4.45) is feasible:

$$G(p_k(a|x), \boldsymbol{U}_k) - c_k \geq G(p_j(a|x), \boldsymbol{U}_k) - c_j, \ \forall\, j, k \in \{1, 2, \ldots, K\}, \ k \neq j.$$

$$\text{GARP} \Leftrightarrow (4.8):$$

Given dataset $\mathcal{D}_{RP}$ (4.6), there exist $\{\lambda_k, g_k\}_{k=1}^K \geq 0$ such that (4.8) is feasible:

$$\lambda_k \, u_k(\beta_k) - g_k \geq \lambda_k \, u_k(\beta_j) - g_j, \ \forall\, j, k \in \{1, 2, \ldots, K\}, \ k \neq j. \tag{4.46}$$

Clearly, from (4.46), we observe that the feasibility of (4.45) is equivalent to the feasibility of (4.8) with $\lambda_k = 1$ for all k under the variable map of statement (1) in Theorem 27. Hence, NIAC is a special case of GARP if NIAS holds, under the variable map of statement (1) of Theorem 27. This concludes the proof of Theorem 27. ■

4.7.3 Proof of Lemma 29

We first show the expected utility (4.40) is monotone wrt the Blackwell partial order. Consider two attention strategies $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \Delta(\mathcal{Y})^{|\mathcal{X}|}$, where $\boldsymbol{\alpha}_1 \geq_{\mathcal{B}} \boldsymbol{\alpha}_2$ without loss of generality[11]. From the definition of Blackwell dominance [25], there exists a matrix $Q \in [0,1]^{|\mathcal{Y}|\times|\mathcal{Y}|}$ such that $\boldsymbol{\alpha}_2(y|x) = \sum_{y' \in \mathcal{Y}} \boldsymbol{Q}_{y',y} \, \boldsymbol{\alpha}_1(y'|x)$ and $\sum_{y \in \mathcal{Y}} \boldsymbol{Q}_{y',y} = 1$ for

[11] Although implicitly assumed in the proof, showing monotonicity of the expected utility wrt Blackwell dominance does not require both attention strategies $\boldsymbol{\alpha}_1, \ \boldsymbol{\alpha}_2$ to be defined on the same space of observations. Adapting the proof to the case where $\boldsymbol{\alpha}_1 \in \Delta(\mathcal{Y}_1)^{|\mathcal{X}|}, \ \boldsymbol{\alpha}_2 \in \Delta(\mathcal{Y}_2)^{|\mathcal{X}|}, \ \mathcal{Y}_1 \neq \mathcal{Y}_2$ is straightforward, and hence, omitted.

all $y, y' \in \mathcal{Y},\ x \in \mathcal{X}$. We now prove that the expected utility functional $J_{\pi_0}(\cdot, \boldsymbol{U})$ for a utility function $\boldsymbol{U}$ is monotone with respect to the Blackwell partial order, that is, $J_{\pi_0}(\boldsymbol{\alpha}_1, \boldsymbol{U}) \geq J_{\pi_0}(\boldsymbol{\alpha}_2, \boldsymbol{U})$:

$$
\begin{aligned}
&J_{\pi_0}(\boldsymbol{\alpha}_2, \boldsymbol{U}) = \sum_{y \in \mathcal{Y}} p_2(y)\ \max_{a \in \mathcal{A}} \pi'_{y,2} \boldsymbol{U}(\cdot, a) = \sum_{y \in \mathcal{Y}} \max_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} p_2(y)\ \pi_{y,2}(x)\ \boldsymbol{U}(x, a) \\
&= \sum_{y \in \mathcal{Y}} \max_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{\alpha}_2(y|x)\ \pi_0(x) \boldsymbol{U}(x, a) = \sum_{y \in \mathcal{Y}} \max_{a \in \mathcal{A}} \sum_{y' \in \mathcal{Y}} Q_{y',y} \left(\sum_x \boldsymbol{\alpha}_1(y'|x)\ \pi_0(x) \boldsymbol{U}(x, a) \right) \\
&\leq \sum_{y \in \mathcal{Y}} \left(\sum_{y' \in \mathcal{Y}} Q_{y',y}\ \max_{a \in \mathcal{A}} \left(\sum_x \boldsymbol{\alpha}_1(y'|x)\ \pi_0(x) \boldsymbol{U}(x, a) \right) \right) \\
&= \sum_{y' \in \mathcal{Y}} \underbrace{\left(\sum_{y \in \mathcal{Y}} Q_{y',y} \right)}_{=1 \forall\ y' \in \mathcal{Y}} \max_{a \in \mathcal{A}} \sum_x \boldsymbol{\alpha}_1(y'|x)\ \pi_0(x) \boldsymbol{U}(x, a) = J_{\pi_0}(\boldsymbol{\alpha}_1, \boldsymbol{U})
\end{aligned}
$$

Therefore, $\boxed{\boldsymbol{\alpha}_1 \geq_{\mathcal{B}} \boldsymbol{\alpha}_2 \implies J_{\pi_0}(\boldsymbol{\alpha}_1, \boldsymbol{U}) \geq J_{\pi_0}(\boldsymbol{\alpha}_2, \boldsymbol{U}) \text{ for any utility function } \boldsymbol{U}}$

(4.47)

Eq. 4.47 shows that the expected utility functional $J_{\pi_0}(\cdot, \boldsymbol{U})$ is monotone in the Blackwell partial order. We now prove the expected utility is convex in the attention strategy. Fix a scalar $\theta \in [0, 1]$. Define $\boldsymbol{\alpha}_\theta = \theta\ \boldsymbol{\alpha}_1 + (1 - \theta)\ \boldsymbol{\alpha}_2$. Then:

$$
\begin{aligned}
&J_{\pi_0}(\boldsymbol{\alpha}_\theta, \boldsymbol{U}) = \sum_{y \in \mathcal{Y}} p_\theta(y) \max_{a \in \mathcal{A}} \pi'_{y,\theta} \boldsymbol{U}(\cdot, a) = \sum_{y \in \mathcal{Y}} \max_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} p_\theta(y) \pi_{y,\theta}(x) \boldsymbol{U}(x, a) \\
&= \sum_{y \in \mathcal{Y}} \max_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{\alpha}_\theta(y|x) \pi_0(x) \boldsymbol{U}(x, a) = \sum_{y \in \mathcal{Y}} \max_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} (\theta \boldsymbol{\alpha}_1(y|x) + (1 - \theta) \boldsymbol{\alpha}_2(y|x))\ \pi_0(x) \boldsymbol{U}(x, a) \\
&\leq \theta \sum_{y \in \mathcal{Y}} \max_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{\alpha}_1(y|x)\ \pi_0(x)\ \boldsymbol{U}(x, a) + (1 - \theta) \sum_{y \in \mathcal{Y}} \max_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{\alpha}_2(y|x)\ \pi_0(x)\ \boldsymbol{U}(x, a) \\
&= \theta\ J_{\pi_0}(\boldsymbol{\alpha}_1, \boldsymbol{U}) + (1 - \theta)\ J_{\pi_0}(\boldsymbol{\alpha}_2, \boldsymbol{U})
\end{aligned}
$$

(4.48)

Therefore, $\boxed{J_{\pi_0}(\boldsymbol{\alpha}_\theta, \boldsymbol{U}) \leq \theta\ J_{\pi_0}(\boldsymbol{\alpha}_1, \boldsymbol{U}) + (1 - \theta)\ J_{\pi_0}(\boldsymbol{\alpha}_2, \boldsymbol{U})\ \forall\ \theta \in [0, 1], \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{U}.}$

■

Remark. Lemma 29 is crucial in the proof of the revealed rational inattention result of Theorem 26 since the attention strategy $\boldsymbol{\alpha}_k$ in decision problem k Blackwell dominates

the action selection policy $p_k(a|x)$ observed by the analyst:

$$p_k(a|x) = \sum_{y \in \mathcal{Y}} p_k(a, y|x) = \sum_{y \in \mathcal{Y}} p_k(a|y, x)\boldsymbol{\alpha}_k = \sum_{y \in \mathcal{Y}} \eta_k(a|y)\boldsymbol{\alpha}_k \tag{4.49}$$

Hence, $G(p_j(a|x), \boldsymbol{U}_k)) = J_{\pi_0}(p_j(a|x), \boldsymbol{U}_k) \leq J_{\pi_0}(\boldsymbol{\alpha}_j, \boldsymbol{U}_k)$, where equality holds when $k = j$, where the surrogate expected utility $G(\cdot)$ is defined in (4.22).

4.7.4 Proof of Corollary 30

The proof of Corollary 30 is identical to that of Theorem 25 (identical via the variable map of statement (1) of Theorem 27) and exploits the unification result of Theorem 27. Since the external analyst does not observe the attention strategy $\boldsymbol{\alpha}$, we can assume WLOG that the mapping from observation y to action a is one-to-one[12]. Hence, $\boldsymbol{\alpha}_k$ can be replaced with $p_k(a|x)$.

Statement (1)⇒(2): Fix indices j, k and assume $p_k(a|x) \geq_H p_j(a|x)$. From the definition of the relation '$\geq_H$' (4.27), there exist indices $i_1, i_2, \ldots, i_L$ such that such that $G(p_{i_1}(a|x), \boldsymbol{U}_k) \geq G(p_k(a|x), \boldsymbol{U}_k)$, $G(p_{i_2}(a|x), \boldsymbol{U}_{i_1}) \geq G(p_{i_1}(a|x), \boldsymbol{U}_{i_1})$, $\ldots, G(p_{i_L}(a|x), \boldsymbol{U}_{i_2}) \geq G(p_{i_L}(a|x), \boldsymbol{U}_{i_L})$. Since there exists an information acquisition cost L that rationalizes the decision maker's dataset, we must have $L(\boldsymbol{\alpha}_k) \leq L(\boldsymbol{\alpha}_{i_1}) \leq L(\boldsymbol{\alpha}_{i_2}) \leq \ldots \leq L(\boldsymbol{\alpha}_{i_L}) \leq L(\boldsymbol{\alpha}_j)$ which, in turn, implies $L(\boldsymbol{\alpha}_k) \leq L(\boldsymbol{\alpha}_j)$.

Our aim is to show $G(p_j(a|x), \boldsymbol{U}_j) \geq G(p_k(a|x), \boldsymbol{U}_j)$. We prove this by contradiction. Suppose $G(p_j(a|x), \boldsymbol{U}_j) < G(p_k(a|x), \boldsymbol{U}_j)$. We note here that the surrogate expected utility G (4.22) is continuous in its first argument, namely, the action selection policy. Hence, from the continuity of $G(\cdot, \boldsymbol{U}_j)$ and monotonicity of the cost L, we could find an action selection policy $p(a|x)$ in the neighborhood of $p_k(a|x)$ such that:

[12]This assumption is also used by [2] to prove the sufficiency of NIAS and NIAC for rationally inattentive utility maximization

(a) $G(p_j(a|x), \boldsymbol{U}_j) < G(p(a|x), \boldsymbol{U}_j) \leq G(p_k(a|x), \boldsymbol{U}_j)$ ($p(a|x)$ satisfies utility constraint (4.26) for decision problem j) and

(b) $L(p(a|x)) < L(\boldsymbol{\alpha}_k) \leq L(\boldsymbol{\alpha}_j)$, or equivalently, $L(p(a|x)) \leq L(\boldsymbol{\alpha}_j)$ ($p(a|x)$ *strictly* costs lesser compared to $p_j(a|x)$).

Clearly, if both (a) and (b) hold, then $p_j(a|x)$ does not rationalize the Bayesian decision maker's response in decision problem j, i.e., L does not rationalize the analyst's dataset, hence our assumption is false. Therefore, it must be the case that $\boxed{G(p_j(a|x), \boldsymbol{U}_j) \geq G(p_k(a|x), \boldsymbol{U}_j)}$, that is, GARRI (4.27) holds.

Equivalence of GARRI and GARP

On closely examining GARRI (4.27) and GARP (4.7) together, we observe that they are both equivalent under the variable map of statement (1) in Theorem 27, if NIAS holds. We require the NIAS condition to be true for the equivalence between GARRI and GARP since the 'effective' utility in the revealed rational inattention case via the variable map is the surrogate expected utility G (4.22) that, by definition, is the 'maximum' expected utility generated from an action selection policy. The maximum is attained when the Bayesian decision maker chooses the optimal action given the posterior belief, or in other words, NIAS holds.

Statement (2)⇒(3). Construct a matrix $A \in \mathbb{R}^{K \times K}$ with elements $A_{j,k} = G(p_k(a|x), \boldsymbol{U}_k) - G(p_j(a|x), \boldsymbol{U}_k)$. Since GARRI (4.27) holds, or equivalently, GARP (4.7) holds under the variable map of Theorem 27, it is trivial to show the matrix A is cyclically consistent. From [4, Lemma 2] and [172, Sections 2 and 3], there exist scalars $\bar{c}_k$ and $\lambda_k > 0$ that satisfy the following set of Afriat-type (4.3) feasibility inequalities:

$$\bar{c}_j - \bar{c}_k - \lambda_k \left(G(p_k(a|x), \boldsymbol{U}_k) - G(p_j(a|x), \boldsymbol{U}_k) \right) \leq 0 \quad \forall\, k, j. \tag{4.50}$$

In complete analogy to (4.42) in Sec. 4.7.1, we modify (4.50) into a form that resembles (4.8) whose feasibility is equivalent to GARP (4.7) by performing the following change of variables:

Without loss of generality, we restrict $\bar{c}_k$ to be finite for all $k = 1, 2, \ldots, K$. Let $M_c < \infty$ denote an arbitrary positive scalar that uniformly bounds $\{\bar{c}_k\}_{k=1}^K$ from above. Note that since $\bar{c}_k$ is bounded for all k, such an M_c exists and $\bar{c}_k > 0$. We now define the variable $c_k = M - \bar{c}_k$ for all k and rewrite (4.50) in terms of the new variables $\{c_k\}_{k=1}^K$:

$$\begin{aligned}
&\bar{c}_j - \bar{c}_k - \lambda_k(G(p_k(a|x), \boldsymbol{U}_k) - G(p_j(a|x), \boldsymbol{U}_k)) \leq 0, \ \forall\, k, j \\
\Leftrightarrow &- \bar{c}_j - (-\bar{c}_k) - \lambda_k(G(p_j(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k)) \geq 0, \ \forall\, k, j \\
\Leftrightarrow &(M_c - \bar{c}_j) - (M_c - \bar{c}_k) - \lambda_k(G(p_j(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k)) \geq 0, \ \forall\, k, j \\
\Leftrightarrow &c_j - c_k - \lambda_k(G(p_j(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k)) \geq 0, \ \forall\, k, j \\
\overset{\Leftrightarrow}{\text{(if NIAS holds)}} &\boxed{c_j - c_k - \lambda_k(G(p_j(a|x), \boldsymbol{U}_k) - \sum_{x,a} \pi_0(x) p_k(a|x) \boldsymbol{U}_k(x, a)) \geq 0, \ \forall\, k, j \equiv (4.28)}.
\end{aligned} \tag{4.51}$$

Statement (3) $\Rightarrow$ (4). Consider the reconstructed information acquisition cost $L_{recon}(p(a|x)) = \max_k\{c_k + \lambda_k(G(p(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k))\}$. Clearly, L_{recon} is monotone and continuous since it is a point-wise maximum of monotone continuous functions. We now show that $L_{recon}(p(a|x)) = \max_k\{c_k + \lambda_k(G(p(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k))\}$ rationalizes the dataset $\mathcal{D}_{RRI}$ (4.14):

Fix index k. Then, $L_{recon}(p_k(a|x)) \geq c_k$ by definition. Also, from (4.51), we have that:

$$c_k \geq c_j + \lambda_j(G(p_k(a|x), \boldsymbol{U}_j) - G(p_j(a|x), \boldsymbol{U}_j)) \ \forall j \neq k$$

Therefore, it is clear that $L_{recon}(p_k(a|x)) = c_k$ (4.51). Now, let $p(a|x)$ denote any feasible response in decision problem k, that is, $G(p(a|x), \boldsymbol{U}_k) \geq G(p_k(a|x), \boldsymbol{U}_k)$. Then:

$$
\begin{aligned}
L_{recon}(p(a|x)) &= \max_k \{c_k + \lambda_k(G(p(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k))\} \\
&\geq \underbrace{c_k}_{=L_{recon}(p_k(a|x))} + \underbrace{\lambda_k(G(p(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k))}_{\geq 0} \geq L_{recon}(p_k(a|x)) \\
&\Rightarrow \lambda_k\, G(p_k(a|x), \boldsymbol{U}_k) - L_{recon}(p_k(a|x)) \geq \lambda_k\, G(p(a|x), \boldsymbol{U}_k) - L_{recon}(p(a|x)),\ \forall\, p(a|x) \\
&\Rightarrow \boxed{p_k(a|x) = \underset{p(a|x)}{\operatorname{argmax}} \quad \lambda_k\, G(p(a|x), \boldsymbol{U}_k) - L_{recon}(p(a|x))}
\end{aligned}
\tag{4.52}
$$

Since (4.52) holds for all $k = 1, 2, \ldots, K$, the reconstructed cost $L_{recon}(p(a|x)) = \max_k\{c_k + \lambda_k(G(p(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k))\}$ rationalizes the analyst's dataset $\mathcal{D}_{RRI}$ (4.14).

Statement (4)⇒(1). The reconstructed information acquisition cost $L_{recon}(p(a|x)) = \max_k\{c_k + \lambda_k(G(p(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k))\}$ is identical to $\widehat{L}$ defined in (4.29). By construction, the cost L_{recon} is monotone, continuous (since it is a point-wise maximum of finitely many monotone continuous segments) and rationalizes the Bayesian decision maker's actions (4.52). ■

Showing reconstructed information acquisition cost is weakly monotone, mixture feasible and normalized

Consider the reconstructed information acquisition cost $\widehat{L}$ (4.29). The cost $\widehat{L}$ is ordinal, i.e., any monotone transformation of $\widehat{L}$ and λ_k rationalizes the dataset $\mathcal{D}_{RRI}$ (4.14) equally well. Hence, without loss of generality, we normalize $\widehat{L}$ (4.29) as:

$$
\begin{aligned}
\widehat{L}_{norm}(p(a|x)) &= \max_k\ \{c_k + \lambda_k\,(G(p(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k))\} - \widehat{L}^*, \text{ where} \\
\widehat{L}^* &= \max_k\ \{c_k + \lambda_k\,(G(p_0(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k))\}.
\end{aligned}
\tag{4.53}
$$

In (4.53), G is the surrogate expected utility defined in (4.22) and $p_0(a|x)$ is the non-informative (uniform conditional probability) action selection policy, i.e., $p_0(a|x) = 1/|\mathcal{A}|$ for all a, x. Combining (4.28) and (4.53) above gives $\widehat{L}_{norm}(p_k(a|x)) = c_k - \widehat{L}^*$. To show $\widehat{L}_{norm}$ (4.53) rationalizes dataset $\mathcal{D}_{RRI}$ (4.14) in Corollary 30, fix index k and consider any action selection policy $p(a|x)$ such that $\widehat{L}_{norm}(p(a|x)) \leq \widehat{L}_{norm}(p_k(a|x))$. By definition (4.53), $0 \geq \widehat{L}_{norm}(p(a|x)) - \widehat{L}_{norm}(p_k(a|x)) \geq \lambda_k(G(p(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k))$, which implies $G(p(a|x), \boldsymbol{U}_k) \leq G(p_k(a|x), \boldsymbol{U}_k)$. This inequality holds for all k. Hence, $\widehat{L}_{norm}$ (4.53) rationalizes the $\mathcal{D}_{RRI}$ (4.14).

We now use Lemma 29 to show that the reconstructed cost $\widehat{L}_{norm}$ (4.53) is: (i) weakly monotone in information, mixture feasible and normalized as theorized in [2, Theorem 2].

K1. Weak monotonicity in information. *The cost $\widehat{L}_{norm}$ is weakly monotonic in information if for any two action selection policies $p(a|x), \hat{p}(a|x)$, we have $\widehat{L}_{norm}(\hat{p}(a|x)) \leq \widehat{L}_{norm}(p(a|x))$, when $p(a|x) \geq_{\mathcal{B}} \hat{p}(a|x)$, where $\geq_{\mathcal{B}}$ stands for 'Blackwell dominates'.*

Condition K1 can be viewed as a monotonicity condition with respect to the Blackwell order.

Proof. Since $p(a|x) \geq_{\mathcal{B}} \hat{p}(a|x)$, Lemma 29 ensures the following inequalities hold:

$$\begin{aligned}
\widehat{L}_{norm}(\hat{p}(a|x)) &= \max_k \{c_k + \lambda_k (G(\hat{p}(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k))\} - \widehat{L}^* \\
&\leq \max_k \{c_k + \lambda_k (G(p(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k))\} - \widehat{L}^* = \widehat{L}_{norm}(p(a|x)) \\
\Rightarrow & \boxed{\widehat{L}_{norm}(\hat{p}(a|x)) \leq \widehat{L}_{norm}(p(a|x))}
\end{aligned}$$

1. Mixture feasibility. *The cost $\widehat{L}_{norm}$ is mixture feasible if for action selection policies $p(a|x), p'(a|x), p''(a|x)$ related as $p(a|x) = \theta p'(a|x) + (1-\theta)p''(a|x),\ \theta > 0$, cost $\widehat{L}_{norm}$ satisfies $\widehat{L}_{norm}(p(a|x)) \leq \theta\, \widehat{L}_{norm}(p'(a|x)) + (1-\theta)\, \widehat{L}_{norm}(p''(a|x))$.*

Proof.

$$\begin{aligned}
\widehat{L}_{norm}(p(a|x)) + \widehat{L}^* &= \max_k \ \{c_k + \lambda_k \left(G(p(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k)\right)\} \\
&= \max_k \ \{c_k + \lambda_k \left(G(\theta p'(a|x) + (1-\theta)p''(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k)\right)\} \\
&\leq \max_k \ \{c_k + \lambda_k \left(\theta\, G(p'(a|x), \boldsymbol{U}_k) + (1-\theta) G(p''(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}\right. \\
&\quad \text{(since the surrogate expected utility } G(\cdot, \boldsymbol{U}_k) \text{ is convex in } p(a|x)) \\
&\leq \theta \ \max_k \ \{c_k + \lambda_k \left(G(p'(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k)\right)\} \\
&\quad + (1-\theta) \ \max_k \ \{c_k + \lambda_k \left(G(p''(a|x), \boldsymbol{U}_k) - G(p_k(a|x), \boldsymbol{U}_k)\right)\} \\
&\quad \text{(since the max operation is convex)} \\
&\Rightarrow \widehat{L}_{norm}(p(a|x)) \leq \theta(\widehat{L}_{norm}(p'(a|x)) + \widehat{L}^*) + (1-\theta)(\widehat{L}_{norm}(p''(a|x)) + \widehat{L}^*) \\
&\Rightarrow \boxed{\widehat{L}_{norm}(p(a|x)) \leq \theta\ \widehat{L}_{norm}(p'(a|x)) + (1-\theta)\ \widehat{L}_{norm}(p''(a|x))}
\end{aligned}$$

2. Normalization. The cost $\widehat{L}_{norm}$ is normalized if $\widehat{L}_{norm}(p_0(a|x)) = 0$, where $p_0(a|x) = 1/|\mathcal{A}|$ (uninformative action selection policy).

   Proof. This holds true from the definition of $\widehat{L}_{norm}$ in (4.53).

CHAPTER 5

IRL AS A PRINCIPLED APPROACH FOR INTERPRETING DEEP NEURAL NETWORKS

5.1 Introduction

This chapter studies interpretable models for deep image classification. We propose a set-valued system identification approach to explain deep image classification. We show that image classification using deep Convolutional Neural Networks (CNNs) can be interpreted as a Bayesian utility maximization where the observation likelihood is optimized. Such rationally inattentive Bayesian utility maximization models have recently been used to explain human decision-making in microeconomics. As discussed below, our surprising finding based on extensive analysis of data, is that deep image classification satisfies the necessary and sufficient conditions for Bayesian utility maximization by a large robustness margin.

In micro- and behavioral economics[1], a fundamental question relating to human decision making is: *How to model attention spans in humans (agents)?* The area of rational inattention [37, 177], pioneered by Nobel laureate Christopher Sims models human attention in information-theoretic terms. The key hypothesis is that agents are "boundedly rational"- their perception of the environment is modeled as a Shannon capacity limited channel. In simple terms, rational inattention assigns a mutual information cost for human attention spans.

Building on the rational inattention model, the next key concept is that of a Bayesian

[1]Micro-economics models the interaction of individual agents pursuing their private interests. Behavioral economics models human decision making in terms of subjective probabilities via prospect theory and framing. In the rest of this chapter, we will use the term 'agent' to denote a Bayesian decision-maker.

agent with rational inattention that maximizes its expected utility. Such models are studied extensively in micro-economics [78, 178, 179]. The intuition is this: more attentive decisions yield a higher expected utility at the expense of a larger attention cost. Hence, the Bayesian agent optimally trades off between minimizing its sensing cost and maximizing its expected utility. An important question is: *How to test for rationally inattentive utility maximization given the decisions of a Bayesian agent?* In the last decade, necessary and sufficient conditions have been developed in the area of Bayesian revealed preference [2, 15] to test if the decisions of a Bayesian agent are consistent with rationally inattentive utility maximization. In this chapter we use the necessary and sufficient conditions of [2, 15] to construct interpretable models for deep classification.

This non-parametric data-driven approach embeds the image classification task as a Bayesian utility maximization problem constrained by an information acquisition cost. We construct set-valued estimates of utility functions and information acquisition costs that rationalize deep image classification. In a signal processing context, the information cost often referred to as the rational inattention cost in the literature is analogous to the sensing cost incurred by a radar in controlled sensing [180, 181]. This approach to deep image classification can be viewed as an *inverse optimization* problem. Recently, neural networks have been used successfully to solve inverse problems in imaging [182–185]. However, to the best of our knowledge, an economics-based inverse optimization analysis of deep neural networks has not been explored in the literature.

5.1.1 Summary of Results.

The question we address is: *Can the decisions of deep CNNs in image classification be explained by a rationally inattentive Bayesian utility maximizer?*

This chapter uses a *data-driven* micro-economics based system identification approach for interpretable deep classification. The key ideas stem from Bayesian revealed preference [2, 15]. Bayesian revealed preference is a set-valued system identification algorithm for argmax non-linearity (in signal processing terms) that describes a Bayesian decision maker. Bayesian revealed preference is a *post-hoc* analysis of agent decisions. It constructs a generative[2] explanatory model for the agent decisions, parameterized by utility functions and an information acquisition cost. As a practical application, the interpretable model can also be used to predict the classification accuracy of the neural network trained on arbitrary training parameters; we discuss this in more detail in Sec. 5.3.2. Our approach draws important parallels between human decision making and deep neural networks; namely that deep neural networks satisfy economics based rationality.

Why set-valued estimates of utility?

The aim of interpretable deep image classification is to construct feasible utility functions and information costs that rationalize neural network image label predictions over a finite set of training parameters. Estimating a utility function is an ill-posed problem (in the sense of Hadamard) since any non-negative increasing function of the utility is also a valid utility. From a statistical signal processing perspective, a point-valued estimate is not useful for rationalizing a Bayesian decision maker's actions: (i) every point in the reconstructed set of feasible utilities and costs explain the actions equally well; hence the problem is ill-posed, and (ii) a least squares estimate of the decision maker's utility function and information cost does not rationalize its actions. Bayesian revealed preference reconstructs a *set* of feasible utility functions and information acquisition costs

[2] A generative model is image-independent, and hence provides a global explanation for deep image classification. In contrast, local approximation models for deep image classification are image-specific; they approximate model decisions via tractable functionals in a δ-neighborhood of every input.

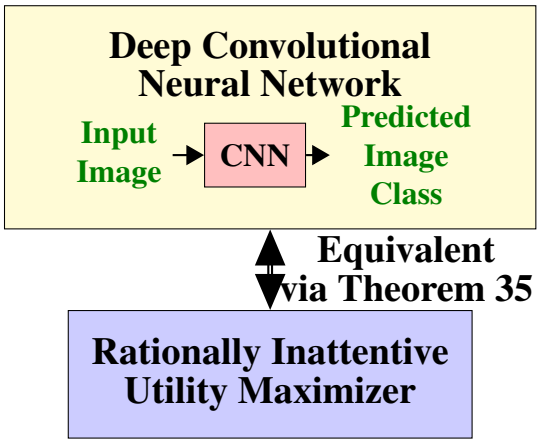

Figure 5.1: Schematic illustration of rationally inattentive Bayesian utility maximization based interpretable image classification by deep CNNs. Theorem 35 establishes equivalence between the image classification behavior of a deep CNN and the decisions of a rationally inattentive maximizer. Hence, the deep CNN's image classification behavior can be parsimoniously represented by a utility function and an information acquisition cost.

that rationalize a Bayesian decision maker's actions in a finite number of environments. Every element in the feasible set explains the deep CNN decisions equally well. In Bayesian revealed preference, the utility functions are indexed by the environment; the information cost is invariant across environments. The computed utility function induces a preference ordering on the set of image classes. That is, how much a deep CNN prioritizes accurate classification over an inaccurate classification. The information acquisition cost abstracts the penalty incurred by the deep CNN to 'learn' an accurate latent feature representation and can be interpreted as the training cost to achieve a desired accuracy of image classification.

The key results in this chapter are:

1. We show that the image classification decisions of deep CNNs satisfy the necessary and sufficient conditions for rationally inattentive utility maximization by a large margin as displayed in Table 5.1. Our findings are based on approximately 500 experiments on 5 widely used neural network architectures for image classification.

This result establishes that the rationally inattentive utility maximization widely used to explain human decisions explains deep image classification remarkably well. This result is schematically shown in Fig. 5.1.

2. To aid visualization of our interpretable model, we provide a sparsity-enhanced decision test that computes the sparsest utility function and information acquisition cost which rationalizes deep CNN decisions. The sparsest solution yields a parsimonious representation of hundreds of thousands of layer weights of the deep CNNs in terms of a few hundred parameters. The utility function of the sparsest interpretable model also induces a useful preference ordering amongst the set of hypotheses (image labels) considered by the CNN; for example, how much additional priority is allocated to the classification of a cat as a cat compared to a cat as a dog. In classical deep learning, this preference ordering is not explicitly generated. The sparsity results for various deep CNN architectures are displayed in Table 5.2 and Fig. 5.3.
3. Our final result demonstrates the usefulness of our interpretable model. We show that, via interpolation, the interpretable model computed from CNN decisions can predict the classification accuracy of a CNN trained with arbitrary parameters with high accuracy. The prediction results are displayed in Table 5.4.

The above results are backed by approximately 500 experiments performed on several deep CNN architectures on the CIFAR-10 [186] image dataset over 3 learning rates, 200 training epochs and 20 values of noise variance for corrupting the original images. The first two results use deep CNN decisions aggregated over varying training epochs. The third (prediction) result uses deep CNN decisions trained on noisy image datasets parameterized by the noise variance. Due to space constraints, we only consider the CIFAR-10 image dataset for our experiments.[3]. However, our algorithms for reconstructing interpretable models for deep image classification are independent of the dataset and can be

[3]The neural network classification accuracy as learning rates and training epochs are varied can be downloaded from `zerenzhang2022.github.io`

straightforwardly extended to larger and more granular datasets like CIFAR-100 [186] and ImageNet at the cost of greater computational resources.

5.1.2 Related Works

Since we study interpretable deep learning using behavioral and micro- economics, we briefly discuss related works in these areas.

Bayesian revealed preference and Rational inattention. Estimating utility functions given a finite sequence of decisions and budget constraints is the central theme of revealed preference in micro-economics. The seminal work of [3, 187] (see also [77]) give necessary and sufficient conditions for the existence of a utility function that rationalizes a finite time series of consumption bundles of a decision-maker. Rationally inattentive models for Bayesian decision making have been studied extensively in [78, 178, 179]. In the last decade, the area of Bayesian revealed preference [2, 15] develops necessary and sufficient conditions to test for rationally inattentive Bayesian utility maximization.

Interpretable ML. Providing transparent models for de-obfuscating 'black-box' ML algorithms under the area of interpretable machine learning is a subject of extensive research [188–190]. Interpretable machine learning is defined in [191] as "the use of machine-learning models for the extraction of relevant knowledge about domain relationships contained in data".

Since the literature is enormous, we only discuss a subset of works pertaining to interpretability of deep neural networks for image classification [192, 193]. One prominent approach, namely, saliency maps, reconstructs the most preferred or typical image pertaining to each image class the deep neural network has learned [194, 195]. Related work

includes creating hierarchical models for determining the importance of image features that determine its label [196]. In this chapter, this feature importance is encoded into the utility function that parametrizes our interpretable model. Another approach seeks to provide local approximations to the trained model, local w.r.t the input image [197, 198]. In contrast, our generative interpretable model provides a global black-box approximation for deep image classification. A third approach approximates the decisions of the deep neural networks by a linear function of simplified individual image features [5, 198–200]. In contrast, our interpretable model fits a stochastic non-linear map that relates the true and predicted image labels. The parameters of the map are obtained by solving a convex feasibility problem parameterized by the deep CNN decisions. Finally, deep neural networks have also been modeled by Bayesian inference frameworks using probabilistic graphical methods [201].

To the best of our knowledge, an economics based approach for the post-hoc analysis of deep neural networks has not been explored in literature. However, we note that behavioral economics based interpretable models have been applied to domains outside interpretable machine learning, for example, in online finance platforms for efficient advertising [202, 203], training neural networks [204] and more recently in YouTube to rationalize user commenting behavior [205]. Finally, due to our recent equivalence result [145], our behavioral economics approach to interpretable deep image classification can be related to classical revealed preference methods [3, 187] in microeconomics.

5.2 Bayesian Revealed preference with Rational Inattention

This section describes the key ideas behind Bayesian revealed preference. Despite the abstract formulation below, the reader should keep in mind the deep learning context. In

Sec. 5.3, we will use Bayesian revealed preference theory to construct an interpretable deep learning representation by showing that deep CNNs are equivalent to rationally inattentive Bayesian utility maximizers.

5.2.1 Utility Maximization with Rational Inattention (UMRI)

Bayesian revealed preference aims to determine if the decisions of a Bayesian agent are consistent with expected utility maximization subject to a rational inattention sensing cost. We start by describing the **u**tility **m**aximization model with **r**ational **i**nattention (henceforth called UMRI) for a ***collection*** of Bayesian decision makers/agents.

Abstractly, the UMRI model is parameterized by the tuple

$$\Theta = (\mathcal{K}, \mathcal{X}, \mathcal{Y}, \mathcal{A}, \pi_0, L, \{\boldsymbol{\alpha}_k, u_k, k \in \mathcal{K}\}). \tag{5.1}$$

With respect to the abstract parametrization of the UMRI model for a collection of Bayesian agents, the following elements constitute the tuple Θ defined in (5.1).

Agents: $\mathcal{K} = \{1, 2, \ldots, K\}$ $(K \geq 2)$ indexes the finite set of Bayesian agents.

State: $\mathcal{X}$ is the finite set of ground truths with prior probability distribution π_0. With respect to our image classification context, $\mathcal{X} = \{1, 2, \ldots 10\}$ is the set of image classes in the CIFAR-10 dataset and π_0 is the empirical probability distribution of the image classes in the test dataset of CIFAR-10.

Observation and attention strategy: Agent $k \in \mathcal{K}$ chooses attention strategy $\boldsymbol{\alpha}_k : \mathcal{X} \rightarrow \Delta(\mathcal{Y})$, a stochastic mapping from $\mathcal{X}$ to a finite set of observations $\mathcal{Y}$. Given state x and attention strategy $\boldsymbol{\alpha}_k$, the agent samples observation y with probability $\boldsymbol{\alpha}_k(y|x)$. The agent then computes the posterior probability distribution $p(x|y)$ via Bayes formula as

$$p(x|y) = \frac{\pi_0(x)\boldsymbol{\alpha}_k(y|x)}{\sum_{x' \in \mathcal{X}} \pi_0(x')\boldsymbol{\alpha}_k(y|x')}. \tag{5.2}$$

The observation and attention strategy are latent variables that abstractly represent the learned feature representations in the deep image classification context. Bayesian revealed preference theory tests their existence via the convex feasibility test in Theorem 35 below.
Action: Agent $k \in \mathcal{K}$ chooses action a from a finite set of actions $\mathcal{A}$ after computing the posterior probability distribution $p(x|y)$. In the image classification context, a is the image class predicted by the neural network, hence $\mathcal{A} = \mathcal{X}$.
Utility function: Agent $k \in \mathcal{K}$ has a utility function $u_k(x, a) \in \mathbb{R}^+$, $x \in \mathcal{X}, a \in \mathcal{A}$ and aims to maximize its expected value, with the expectation taken wrt the random state x and random observation y. A key feature in our approach is to show that the utility function rationalizes the decisions of the deep CNNs (made precise in Definition 34).
Information Acquisition Cost: The information acquisition cost $L(\boldsymbol{\alpha}, \pi_0) \in \mathbb{R}^+$ depends on attention strategy $\boldsymbol{\alpha}$ and prior pmf π_0. It is the sensing cost the agent incurs in order to estimate the underlying state (5.2). In the context of machine learning, $L(\cdot)$ abstractly captures the 'learning' cost incurred during training of the deep neural networks. In rational inattention theory from behavioral economics, a higher information acquisition cost is incurred for more accurate attention strategies (equivalently, more accurate state estimates (5.2) given observation y). We refer the reader to the influential work of [37, 177].

Each Bayesian agent $k \in \mathcal{K}$, aims to maximize its expected utility while minimizing its cost of information acquisition. Hence, the action a given observation y, and attention strategy $\boldsymbol{\alpha}_k$ are chosen as follows:

Definition 34 (Rationally Inattentive Utility Maximization) *Consider a collection of Bayesian agents $\mathcal{K}$ parameterized by Θ in* (5.1) *under the UMRI model. Then,*
(a) **Expected Utility Maximization:** *Given posterior probability distribution $p(x|y)$, every agent $k \in \mathcal{K}$ chooses action a that maximizes its expected utility. That is, with $\mathbb{E}$*

denoting mathematical expectation, the action a satisfies

$$a \in \underset{a' \in \mathcal{A}}{\operatorname{argmax}} \ \mathbb{E}_x\{u_k(x, a')|y\} = \sum_{x \in \mathcal{X}} p(x|y) u_k(x, a') \tag{5.3}$$

(b) **Attention Strategy Rationality:** *For agent k, the attention strategy $\boldsymbol{\alpha}_k$ optimally trades off between maximizing the expected utility and minimizing the information acquisition cost.*

$$\boldsymbol{\alpha}_k \in \underset{\boldsymbol{\alpha}'}{\operatorname{argmax}} \mathbb{E}_y\{ \max_{a \in \mathcal{A}} \mathbb{E}_x\{u_k(x, a)|y\}\} - L(\boldsymbol{\alpha}', \pi_0) \tag{5.4}$$

Eq. 5.3,5.4 in Definition 34 constitute a nested optimization problem. The lower-level optimization task is to choose the the 'best' action for any observation y based on the computed posterior belief of the state. The upper-level optimization task is to sample the observations optimally by choosing the 'best' attention strategy.

Remark. The multiple Bayesian agents in Θ have the same state space $\mathcal{X}$, observation space $\mathcal{Y}$, action space $\mathcal{A}$, prior π_0 and cost of information acquisition L, but only differ in their utility functions. Bayesian revealed preference theory relies on this crucial constraint on the optimization variables in (5.3), (5.4) for detecting optimal behavior in a finite number of agents.

5.2.2 Bayesian Revealed Preference (BRP) Test for Rationally Inattentive Utility Maximization

Having described the UMRI model (collection of rationally inattentive utility maximizers), we are now ready to state our key result. Theorem 35 below says that the decisions of a collection of Bayesian agents is rationalized by a UMRI tuple Θ *if and only if* a set of convex inequalities have a feasible solution. These inequalities comprise our

Bayesian **R**evealed **P**reference (henceforth called BRP) test for rationally inattentive utility maximization.

For notational convenience, the decisions of the Bayesian agents in the UMRI model are compacted into the dataset $\mathbb{D}$ defined as:

$$\mathbb{D} = \{\pi_0, p_k(a|x), x \in \mathcal{X}, a \in \mathcal{A}, k \in \mathcal{K}\}. \tag{5.5}$$

In (5.5), $\pi_0 \in \Delta^{|\mathcal{X}|-1}$ denotes the prior pmf over the set of states $\mathcal{X}$ in Θ (5.1). The variable $p_k(a|x)$ is the conditional probability that agent $k \in \mathcal{K} = \{1, 2, \ldots, K\}$ takes action a given state x. $\mathbb{D}$ characterizes the input-output behavior of the collection of Bayesian agents and serves as the input for BRP feasibility test described below.

Theorem 35 (BRP Test for Rationally Inattentive Utility Maximization [2]) *Given the dataset* $\mathbb{D}$ (5.5) *obtained from a collection of Bayesian agents* $\mathcal{K}$*. Then,*

1. *Existence: There exists a UMRI tuple* $\Theta(\mathbb{D})$ (5.1) *that rationalizes dataset* $\mathbb{D}$ *if and only if there exists a feasible solution that satisfies the set of convex inequalities*

$$\textit{BRP}(\mathbb{D}, \{u_k, c_k\}_{k=1}^K) \leq \mathbf{0},\ u_k \in \mathbb{R}_+^{|\mathcal{X}|\times|\mathcal{A}|},\ c_k > 0. \tag{5.6}$$

In (5.6)*,* $\textit{BRP}(\cdot)$ *corresponds to a set of convex (in the variables* $\{u_k, c_k\}_{k=1}^K$*) inequalities, stated in Algorithm 1.*

2. *Reconstruction: Given a feasible solution* $\{u_k, c_k\}_{k=1}^K$ *to* $BRP(\mathbb{D}, \cdot)$*,* u_k *is the* k^{th} *Bayesian agent's utility function in the feasible model tuple* $\Theta(\mathbb{D})$*. The set of observations* $\mathcal{Y} = \mathcal{A}$*, the set of actions in* $\mathbb{D}$*. The feasible cost of information acquisition* L *in* $\Theta(\mathbb{D})$ *is defined in terms of* c_k *as:*

$$\begin{aligned} L(\boldsymbol{\alpha}) = \max_{k\in\mathcal{K}} c_k + \sum_a \max_{b\in\mathcal{A}} \sum_x p(x,a)u_k(x,b) \\ - \sum_{x,a} p_k(x,a)u_k(x,a),\ \boldsymbol{\alpha} = \{p(a|x)\}. \end{aligned} \tag{5.7}$$

The proof of Theorem 35 is in Sec. 5.5.1. Before launching into a detailed discussion, we stress the "iff" in Theorem 35. Put simply: if the inequalities in (5.6) are not feasible, then the Bayesian agents that generate the dataset $\mathbb{D}$ are not rationally inattentive utility maximizers. If (5.6) has a feasible solution, then there exists a reconstructable family of viable utility functions and information acquisition costs that rationalize $\mathbb{D}$[4]. A key feature of Theorem 35 is that the estimated utilities (and information costs) are set-valued; every utility and cost function in the feasible set explains $\mathbb{D}$ equally well. The estimated UMRI model parameters are set-valued due to the finite number of Bayesian agents whose decisions constitute the dataset $\mathbb{D}$. The estimated parameter set converges to a point if and only if the inequality (5.6) holds as $|\mathcal{K}| \to \infty$.

Computational Aspects of BRP Test. Suppose the dataset $\mathbb{D}$ is obtained from K Bayesian agents. Then, BRP($\mathbb{D}$) comprises a feasibility test with K $(|\mathcal{X}||\mathcal{A}| + 1)$ free variables and $K^2 + K$ $(|\mathcal{A}|^2 - |\mathcal{A}| - 1)$ convex inequalities. Thus, the number of free variables and inequalities in the BRP feasibility test scale linearly and quadratically, respectively, with the number of observed Bayesian agents.

5.2.3 Relating UMRI and BRP test to Interpretable Deep Image Classification

We now discuss how the above BRP test relates to interpretable image classification using deep CNNs. The BRP convex feasibility test in Theorem 35 comprises two sets of inequalities, namely, the *NIAS* (No-Improving-Action-Switches) (5.8) and *NIAC* (No-

[4]In terms of interpretable deep learning, of all parameters in the UMRI tuple, we are only interested in the utility functions of the agents and the cost of information acquisition, since the remaining parameters are immediately deduced from the decision dataset $\mathbb{D}$.

Algorithm 1 BRP Convex Feasibility Test of Theorem 35

Require: Dataset $\mathbb{D} = \{\pi_0, p_k(a|x), x, a \in \mathcal{X}, k \in \mathcal{K}\}$ from a collection of Bayesian agents $\mathcal{K}$.

Find: Positive reals $c_k,\ u_k(x,a) \in (0,1]$ for all $x \in \mathcal{X}$, $a \in \mathcal{A},\ k \in \mathcal{K}$ that satisfy the following inequalities:

$$\underline{\textbf{NIAS}}: \sum_x p_k(x|a)\ (u_k(x,b) - u_k(x,a)) \leq 0, \tag{5.8}$$
$$\forall a, b \in \mathcal{A},\ k \in \mathcal{K},$$

$$\underline{\textbf{NIAC}}: \sum_a \left(\max_b \sum_x p_j(x,a) u_k(x,b) \right) - c_j \tag{5.9}$$
$$- \sum_{x,a} p_k(x,a) u_k(x,a) + c_k \leq 0,\ \forall j,k \in \mathcal{K},$$

where $p_k(x,a) = \pi_0(x) p_k(a|x),\ p_k(x|a) = \frac{p_k(x,a)}{\sum_{x'} p_k(x',a)}$.

Return: Set of feasible utility functions u_k and information acquisition costs c_k incurred by agents $k \in \mathcal{K}$.

Improving-Action-Cycles) (5.9) inequalities (Algorithm 1). NIAS ensures that the agent takes the best action given a posterior pmf. NIAC ensures that every agent chooses the best attention strategy. BRP test checks if there exist K utility functions and K positive reals that, together with $\mathbb{D}$, satisfy the NIAS and NIAC inequalities.

Toy Example with 2 CNNs

The following discussion gives additional insight into our approach. Consider the simplest case involving two trained deep CNNs N_1 and N_2; so $\mathcal{K} = \{1, 2\}$ in the above notation. Assume N_1 and N_2 have the same network architecture. Suppose an analyst observes that N_1 makes accurate decisions on a rich input image dataset while N_2 makes less accurate decisions on the same dataset.

Our UMRI model first abstracts the accuracy of the feature representations of the input image data learned by N_1 and N_2 via attention strategies $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ in (5.4). Second,

the information acquisition cost function $L(\cdot)$ abstracts the computational resources expended for learning the representations. The rationale is that learning an accurate latent feature representation is costly, and this is abstracted by the information acquisition cost.

Let the training cost incurred by N_1 and N_2 be $L(\boldsymbol{\alpha}_1)$ and $L(\boldsymbol{\alpha}_2)$ respectively. If the decisions of N_1 and N_2 can be explained by the UMRI model (and Theorem 35 below will give necessary and sufficient conditions for this), then there exist utility functions u_1 and u_2 for N_1 and N_2, that satisfy:

$$\mathbb{E}_{\boldsymbol{\alpha}_i}\{u_i\} - L(\boldsymbol{\alpha}_i) \geq \mathbb{E}_{\boldsymbol{\alpha}_j}\{u_i\} - L(\boldsymbol{\alpha}_j),\ i,j \in \{1,2\} \tag{5.10}$$

The above inequality says that CNNs N_1 and N_2 would be worse off (in an expected utility sense) if they make decisions based on swapping each other's learned representations. That is, both N_1 and N_2 learn the 'best' feature representation of the input images given their training parameters.

Discussion

(i) *Parsimonious Interpretable Representation of deep CNNs.* In the deep image classification context, due to the UMRI model's parsimonious parametrization in (5.1), the decisions of K CNNs can be rationalized by just K utility functions and an information acquisition cost function, thus bypassing the need of several million parameters to describe the deep CNNs.

(ii) *Identifiability.* The BRP feasibility test requires the dataset $\mathbb{D}$ to be generated from $K > 2$ Bayesian agents. If $K = 1$, then (5.6) holds trivially since any information acquisition cost satisfies the convex inequalities of BRP. Another intuitive way of motivating a collection of agents for the BRP is as follows. Reconstructing a feasible UMRI model tuple Θ that rationalizes the decisions of the deep CNNs is analogous to fitting a line to

a finite number of points. One can fit infinitely many lines through a single point. The task becomes non-trivial if the number of points exceeds 2. In the Bayesian revealed preference context, the points correspond to the decisions from each Bayesian agent. The slope and intercept of the fitted line, in our case, corresponds to the utility function and cost of information acquisition that rationalize the agent decisions.

(iii) *Relative Optimality implies Global Optimality.* In the setting involving $K > 2$ deep CNNs (agents), the NIAS and NIAC inequalities of BRP test check for relative optimality - *given utility function u_k, does deep CNN k performs at least as well as any other observed deep CNN in $\mathcal{K}\backslash\{k\}$?* Clearly, testing for relative optimality is weaker than testing for global optimality (5.4) which ideally requires access to decisions from an infinite number of deep CNNs. Setting the cost of information acquisition as a free variable bridges this gap. The proof of Theorem 35 shows that if the deep CNN decisions satisfy relative optimality, then there exists a cost of information acquisition such that the decisions are globally optimal. That is, Theorem 35 ensures relative optimality is sufficient for global optimality.

(iii) *Generalization of [2].* Theorem 35 generalizes [2, Theorem 1] in two ways. (1) In [2], the utilities u_k in UMRI model tuple Θ are assumed known, and only the information acquisition costs c_k are estimated, whereas Theorem 35 estimates both parameters. (2) The expression for the reconstructed model tuple $\Theta(\mathbb{D})$ is novel; the discussion in [2] is only confined to the existence of such a tuple.

(vi) *Single Utility UMRI (S-UMRI).* In Sec. 5.5.2, we propose a sparse version of UMRI , namely, the S-UMRI model in (5.21). The key distinction of this model is that all agents have the same utility function u and thus can be represented with substantially fewer parameters. In complete analogy to Theorem 35, we outline a decision test in Theorem 39 that states necessary and sufficient conditions for agent decisions to be consistent with the S-UMRI model of rationally inattentive utility maximization. We discuss this sparse

parametrization in the Appendix so as not to interrupt the flow of the main text.

(vii) *Degenerate solution to BRP and S-BRP tests.* The degenerate utility function of all zeros and cost of information acquisition $L = 0$ trivially satisfy the BRP and S-BRP tests and lie at the boundary of the feasible set of parameters.

Summary

This section formulated an economics-based decision-making model. Since this model may not be familiar to a machine learning reader, we summarize the main ideas. We introduced the rationally inattentive utility maximization model, namely, the UMRI model for a collection of Bayesian agents (decision makers). Our main result Theorem 35 outlines a decision test BRP for rationally inattentive utility maximization given decisions from a collection of agents. This BRP test comprises a set of convex inequalities that have a feasible solution *if and only if* the collection of agents are rationally inattentive utility maximizers. Theorem 35 also provides an explicit reconstruction of the feasible UMRI model parameters that rationalize input agent decisions. The set of feasible utility functions and information acquisition costs thus parsimoniously explain the decisions generated by the Bayesian agents. In Sec. 5.5.2, we propose a single utility version of the UMRI model with fewer parameters. Due to fewer parameters, the decision test for this sparse model, given in Theorem 39, is computationally less expensive yet more restrictive than the BRP test for rationality in Theorem 35.

The rest of the chapter focuses on computing interpretable UMRI models that rationalize deep CNN decisions. We will investigate through extensive experiments how well the UMRI fits the deep CNN decisions via robustness tests. We will also investigate how well the computed interpretable models, namely, UMRI and S-UMRI , predict the deep

CNNs' decisions when the training parameters are varied.

5.3 Bayesian Revealed Preference explains CIFAR-10 Image Classification by Deep CNNs

The experimental results in this section are divided into two parts: First, we show that the deep CNNs decisions pass the BRP and S-BRP tests formulated in Theorems 35 and 39 by a large margin. This implies that the rationally inattentive utility maximization model is a robust fit to the deep CNN decisions.

Our second result demonstrates an application of the reconstructed interpretable model. Training datasets are often noisy. We show that in such a noisy setting, the reconstructed interpretable model from Theorem 35 can accurately predict (with accuracy exceeding 94%) the image classification performance of the deep CNNs. This bypasses the need to train the deep CNN for various noise variances that corrupt the training dataset.

Experimental Setup: Deep CNN Architectures, Training Parameters and Construction of Dataset

Image Dataset. For our numerical experiments, we trained and validated the deep CNNs using the CIFAR-10 benchmark image dataset [186]. This public dataset consists of 60000 32x32 colour images in 10 distinct classes (for example, airplane, automobile, ship, cat, dog etc.), with 6000 images per class. There are 50000 training images and 10000

test images. We will use the terms image classes and image labels interchangeably.[5]

Network Architecture and Training Parameters. In this chapter, we use 5 well-known deep CNN architectures for our experiments. 1. LeNet [206], 2. AlexNet [207] 3. VGG16 [208] 4. ResNet-50 [193] 5. Network-in-Network (NiN) [209] The deep CNNs are trained and validated on the CIFAR-10 image dataset, using 3 learning rate schedules, namely, L.R. 1, L.R. 2 and L.R. 3. All 3 schedules use the RMSprop optimizer [210] with the decay parameter and maximum training epochs (full passes of the training dataset) set to 10^{-6} and 200, respectively, and initial step size set to 0.01. The step size is halved every $20, 30, 40$ epochs, respectively, for L.R. 1, 2 and 3.

Relation to Bayesian revealed preference. We now relate the deep CNN setup to the Bayesian revealed preference framework in Sec. 5.2. For each CNN architecture, we use the decisions of $K = 20$ CNNs, i.e. , 20 Bayesian agents in the terminology of Sec. 5.2, for our BRP and S-BRP decision tests. The CNN decisions from K CNNs on the test image dataset of CIFAR-10 are aggregated into dataset $\mathbb{D}$ (5.5). The results of the decision tests are discussed below. In the deep image classification context, the parameter $p_k(a|x)$ in (5.5) is the probability that the k^{th} deep CNN classifies an image from category x into category a in the CIFAR-10 test image dataset. The prior π_0 in $\mathbb{D}$ (5.5) is the empirical pmf over the set of image categories in the CIFAR-10 test dataset. Constructing $\mathbb{D}$ from raw CNN decisions is discussed in Sec. 5.5.3.

Network Architecture	Learning Rate	$\mathcal{R}$ $(\times 10^{-4})$	$\mathcal{R}_{\text{S-BRP}}$ $(\times 10^{-4})$
LeNet	L. R. 1	30.34	4.72
	L. R. 2	35.14	4.65
	L. R. 3	37.97	5.11
AlexNet	L. R. 1	32.10	3.21
	L. R. 2	34.98	3.91
	L. R. 3	40.60	4.62
VGG16	L. R. 1	96.36	4.09
	L. R. 2	107.4	4.02
	L. R. 3	119.8	4.44
ResNet-50	L. R. 1	126.2	2.82
	L. R. 2	129.2	3.45
	L. R. 3	132.3	3.83
Network-In-Network (NiN)	L. R. 1	108.3	3.59
	L. R. 2	132.1	3.36
	L. R. 3	149.1	5.57

Table 5.1: How does increasing the complexity of the network architecture improve robustness of fit to the CNN decisions to the interpretable model? We see that $\mathcal{R}$ (5.11) is substantially higher (by an order of magnitude) than $\mathcal{R}_{\text{S-BRP}}$ (5.12) for all CNN architectures. We conclude that the UMRI model fits CNN decisions substantially better than the S-UMRI model, but with larger computing cost for evaluating the parameters of the interpretable model. Thus, if there are no computational constraints, we recommend using the UMRI model for interpreting CNN decisions.

5.3.1 BRP and S-BRP Tests for deep CNN datasets. Results and Insights

A. Robustness Results on Deep CNN datasets

Our first key result is that image classifications of all 5 deep CNN architectures listed in Sec. 5.3 pass the BRP and S-BRP tests by a large margin. The results are tabulated in Table 5.1. The robustness values $\mathcal{R}$ and $\mathcal{R}_{\text{S-BRP}}$ in Table 5.1 are defined in Definition 36 below which formalizes the notion of margin for the decision tests.

[5]Our experiments are confined to the CIFAR-10 dataset for clarity of exposition. Our approach to interpretable deep learning can be easily extended to richer benchmark image datasets like ImageNet and CIFAR-100 (that comprise over a 100 image labels).

Network Architecture	Learning Rate (L.R.)	airplane	auto	bird	cat	deer	dog	frog	horse	ship	truck
LeNet	L.R. 1	17.61	3.55	20.06	1.88	17.19	21.42	42.00	27.79	1.91	9.55
	L.R. 2	4.13	5.20	7.82	1.90	13.18	18.66	23.84	8.16	2.48	2.47
	L.R. 3	10.79	8.27	18.62	22.67	19.91	25.01	47.71	73.52	2.65	1.01
AlexNet	L.R. 1	210.78	41.84	49.77	59.71	51.24	68.31	83.94	211.61	60.43	125.73
	L.R. 2	85.51	47.89	17.38	1.00	25.34	202.78	21.30	35.01	533.62	248.57
	L.R. 3	18.00	49.55	58.25	28.31	135.54	29.24	224.91	214.51	8.29	264.20
VGG16	L.R. 1	164.48	154.77	15.42	33.67	6.28	123.89	62.83	26.21	1.43	170.69
	L.R. 2	88.73	154.10	45.63	297.61	131.08	136.52	57.34	229.80	145.99	11.90
	L.R. 3	24.33	10.78	93.90	11.11	91.96	56.64	77.30	110.60	20.28	17.09
ResNet-50	L.R. 1	50.83	17.55	16.09	4.66	17.92	3.67	4.92	3.95	15.46	4.88
	L.R. 2	7.51	8.40	72.70	30.72	32.43	83.65	221.27	74.59	99.04	20.51
	L.R. 3	14.61	367.59	31.61	9.20	16.35	11.58	41.44	243.95	222.67	483.91
Network-in-Network	L.R. 1	5.02	30.95	9.91	71.38	63.69	45.88	31.39	67.86	17.03	21.41
	L.R. 2	40.17	60.32	4.40	55.67	95.02	88.72	91.15	15.98	176.75	10.27
	L.R. 3	10.47	75.32	55.97	24.17	17.41	8.94	23.02	71.27	29.94	80.91

Figure 5.2: The utility function of the sparsest interpretable model is a diagonal matrix. The diagonal elements yield a natural preference ordering amongst the set of image classes (classification hypotheses). For example, consider the VGG16 architecture trained using learning rate 1 (third row, first sub-row of table). The maximum utility is for trucks (170.69, last column) and the minimum is for ships (1.43, second last column). This shows the sparsest interpretable model induces the following preference ordering for the VGG16 architecture: classifying trucks correctly is prioritized 100 times more than classifying ships. Such a preference ordering is not explicitly generated by a CNN.

Definition 36 (Robustness (Goodness-of-fit) of BRP and S-BRP Tests.) *Given dataset* $\mathbb{D}$ (5.5) *aggregated from a collection of Bayesian agents,* $\mathcal{R}(\mathbb{D})$ *and* $\mathcal{R}_{\textit{S-BRP}}(\mathbb{D})$ *measure the largest perturbation so that* $\mathbb{D}$ *passes the BRP and S-BRP decision tests:*

$$\mathcal{R}(\mathbb{D}) = \max_{\varepsilon>0} \frac{\varepsilon\ K}{\sum_{k=1}^{K} \|u_k\|_2^2},\ \textit{BRP}(\mathbb{D}, \{u_k, c_k\}_{k=1}^{K}) \leq -\varepsilon. \tag{5.11}$$

$$\mathcal{R}_{\textit{S-BRP}}(\mathbb{D}) = \max_{\varepsilon>0} \frac{\varepsilon}{\|u\|_2^2}, \textit{S-BRP}(\mathbb{D}, u, \{c_k, \lambda_k\}_{k=1}^{K}) \leq -\varepsilon. \tag{5.12}$$

In Definition 36, robustness values $\mathcal{R}$ and $\mathcal{R}_{\text{S-BRP}}$ measure, respectively, the smallest perturbation needed for $\mathbb{D}$ to fail the BRP and S-BRP decisions tests. Both $\mathcal{R}$ and $\mathcal{R}_{\text{S-BRP}}$ are normalized wrt the row-wise $\mathcal{L}_2$ norm of the feasible utility functions. Higher robustness values imply a better fit of the UMRI , S-UMRI models to the decision dataset

[6].

Discussion and Insights. Robustness Results of Table 5.1

(i) *Deep CNN dataset:* The deep CNN datasets used for the robustness tests (5.11), (5.12) comprise decisions of $K = 20$ deep CNNs for every network architecture, where CNN k was trained for $10\ k$ training epochs, $k = 1, 2, \ldots, K$.

(ii) *Comparison between $\mathcal{R}$ and $\mathcal{R}_{S\text{-}BRP}$ values for deep CNN datasets:* The average value of $\mathcal{R}_{\text{S-BRP}}$ (5.12) over all 3 learning rate schedules and 5 network architectures was found to be 4.09×10^{-4}. In contrast, the average value of $\mathcal{R}$ (5.11) was found to be 87.45×10^{-4}, almost 20 times the average value of $\mathcal{R}_{\text{S-BRP}}$. This result shows that the UMRI model fits deep CNN decisions substantially better than the S-UMRI model. This result is expected since S-UMRI is parameterized using much fewer variables compared to the UMRI and hence, S-BRP test is more restrictive than BRP .

(iii) *Sensitivity of $\mathcal{R}, \mathcal{R}_{S\text{-}BRP}$ to Network Architecture:* The average value of $\mathcal{R}$ is 122.29×10^{-4} for the LeNet and AlexNet architectures, which is approximately 3.5 times the the average value of $\mathcal{R}$ for the VGG16, ResNet-50 and NiN architectures which is 35.18×10^{-4}. The variation of $\mathcal{R}_{\text{S-BRP}}$ with network architecture is negligible compared to $\mathcal{R}_{\text{S-BRP}}$. This shows the robustness test for UMRI model is more sensitive to network architecture compared to that for the S-UMRI model.

(iv) *Computational aspects of $\mathcal{R}$ and $\mathcal{R}_{S\text{-}BRP}$.* The computation time for $\mathcal{R}$ is almost 30 times that for $\mathcal{R}_{\text{S-BRP}}$. This is expected since the UMRI model is parameterized by K utility functions compared to a single utility function in S-UMRI .

[6]The robustness value for the non-informative dataset of uniformly distributed pmfs is 0. Hence, the robustness value measures the informativeness of the attention strategies in $\mathbb{D}$ relative to the uniform probability distribution.

B. Sparsity-enhanced Interpretable Model

Our next task is to determine the sparsest possible interpretable model that satisfies the decision tests BRP and S-BRP. The motivation is three fold:

1. The sparsest interpretable model explains the deep CNN decisions using the fewest number of parameters.
2. The sparsest interpretable model induces a useful preference ordering amongst the set of hypotheses (image labels) considered by the CNN; for example, how much additional priority is allocated to the classification of a cat as a cat compared to a cat as a dog. In classical deep learning, this preference ordering is not explicitly generated.
3. Third, the sparsest solution is a point valued estimate. Recall the BRP and S-BRP decision tests yield a set-valued estimate of feasible utility functions and cost of information acquisition that explain the deep CNN datasets. While every element in the set explains the dataset equally well, it is useful to have a single representative point.

Theorem 37 below computes the sparsest utility function out of all feasible utility functions.

Theorem 37 (Sparsity Enhanced BRP and S-BRP Tests for Deep CNN datasets) *Given dataset* $\mathbb{D}$ (5.5) *from a collection of* K *Bayesian agents. The sparsest solutions to the BRP and S-BRP tests minimize the sum of row-wise* $\mathcal{L}_1$ *norm of the feasible utility functions of the* K *agents that generate* $\mathbb{D}$*.*

$$
\begin{aligned}
(u_{1:K})^* &= \underset{u_{1:K}}{\operatorname{argmin}} \sum_{k=1}^{K} \|u_k\|_1, \textit{BRP}(\mathbb{D}, \cdot) \leq \mathbf{0}, \ \sum_{k=1}^{K} \|u_k\|_2^2 = K. \\
u^* &= \underset{u}{\operatorname{argmin}} \|u\|_1, \textit{S-BRP}(\mathbb{D}, \cdot) \leq \mathbf{0}, \ \|u\|_2^2 = 1. \qquad (5.13)
\end{aligned}
$$

where $\|\cdot\|_1$ denotes the row-wise $\mathcal{L}_1$ norm.

Results and Discussion. Sparsity Test for deep CNN datasets

The sparsest utility function from the S-BRP test are tabulated in Table 5.2 for all 5 deep CNN architectures. The corresponding information acquisition cost for all 5 architectures averaged over learning rates $1, 2, 3$ are shown in Fig. 5.3. Together, the sparsest utility and information cost constitute the sparsest S-UMRI interpretable model[7] for the deep CNN decisions.

(i) *Preference ordering induced from sparsest utility.* The sparsest utility function for the S-UMRI model induces a useful preference ordering among the predicted image classes. That is, they measure how the deep CNN's priority for accurate classification varies across image classes. For instance, consider the VGG16 architecture trained using learning rate schedule 1. Of all image categories, the maximum utility is observed for trucks (170.69) and the minimum for ships (1.43). This shows the VGG16 architecture prioritizes classifying trucks correctly about 100 times more than classifying ships.

(ii) *Penalty for learning image features accurately.* The computed information acquisition costs in Fig. 5.3 can be understood as the training cost the CNN incurs to learn latent image features accurately. The interpretable model cannot explain the variation in CNN classification accuracy versus variation in training parameters without an information acquisition cost. From Fig. 5.3, we can conclude that learning accurate image features is the most and least costly, respectively, for the AlexNet and ResNet architectures, respectively.

[7]For brevity, we have only included the sparsity results for the S-UMRI model. The sparsest utility functions of the UMRI model that explains deep CNN decisions are included in our public GitHub repository that contains all test results and codes.

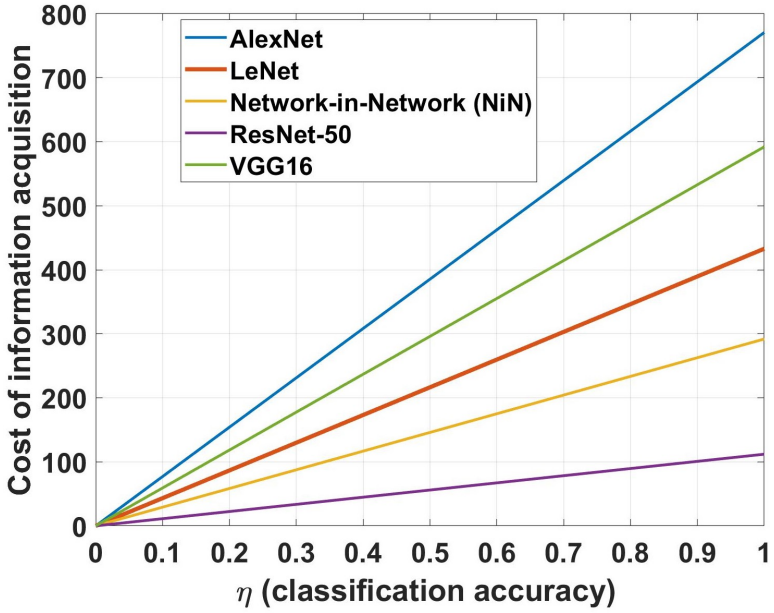

Figure 5.3: The figure illustrates an important property of our approach to interpretable deep learning: in addition to the utility function (Table 5.2), we also need a rational inattention term (cost of learning latent image features) to explain CNN decisions. Put differently, we cannot explain the variation in CNN classification accuracy versus variation in training parameters without an information acquisition cost. The figure displays the information acquisition cost L (5.7) evaluated for the sparsest interpretable model. We also observe that learning accurate image features is most expensive for AlexNet, and least expensive for ResNet architectures.

5.3.2 Predicting deep CNN classification accuracy using our Interpretable Models

Training datasets are often noisy; for example, [211] considers noisy datasets for handwritten character recognition. We now exploit the proposed interpretable model to predict how the deep CNN will perform with a noisy training dataset without actually implementing the deep CNN.

Our predictive procedure is as follows. We first train the CNNs on noisy datasets that are generated by adding simulated Gaussian noise with noise variances chosen from

Network Architecture	airplane	auto	bird	cat	deer	dog	frog	horse	ship	truck
LeNet	0.042	0.042	0.041	0.027	0.046	0.025	0.049	0.034	0.040	0.042
AlexNet	0.025	0.031	0.034	0.021	0.046	0.032	0.049	0.039	0.045	0.036
VGG16	0.033	0.035	0.043	0.041	0.048	0.048	0.035	0.046	0.037	0.048
ResNet-50	0.030	0.031	0.027	0.031	0.020	0.027	0.040	0.015	0.023	0.024
Network-in-Network	0.051	0.029	0.025	0.028	0.056	0.059	0.030	0.058	0.045	0.036

Figure 5.4: How well does our interpretable model predict CNN classification accuracy? The table displays the prediction error $\delta_\eta(x)$ defined in (5.14). Recall $\delta_\eta(x)$ is the error between the true CNN performance and the predicted performance using the interpretable model with Algorithm 2. The maximum error across all image classes and architectures was found to be 5.9%. Hence, our interpretable model predicts CNN classification performance with accuracy exceeding 94%.

a finite set. [8] Then given the CNN decisions, we compute our interpretable model over this finite set of noise variances. Finally, to predict how the CNN will perform for a noise variance not in the set, we *interpolate* the utility function of the interpretable model at this noise variance. Then given the interpolated utility function and information acquisition cost from our interpretable model, the predicted classification performance is computed by solving convex optimization problem (5.4). The above procedure is formalized in Algorithm 2. Hence, our interpretable model serves as a computationally efficient method for predicting the performance of a CNN without implementing the CNN. The interpretable model can be viewed as a low dimension projection of the high-dimension CNN with predictive accuracy exceeding 94%.

Remark. An alternative procedure is to directly interpolate the performance over the space of CNN weights (several hundreds of thousands). Due to the high dimensionality, this is an intractable interpolation. In comparison, interpolation over the utility functions in our interpretable model is over a few hundred variables.

[8] Injecting artificial noise in training datasets is also used in variational auto-encoders for robust feature learning [212, 213].

Prediction Results of Algorithm 2 on Deep CNN Performance

Table 5.4 displays the prediction errors (difference between the true and predicted classification accuracy) for the deep CNNs for all 5 architectures and all image classes in CIFAR-10. For a fixed CNN architecture and noise variance $\eta > 0$, the prediction error $\delta_\eta(x)$ for image class x is defined as:

$$\delta_\eta(x) = |\hat{p}(x|x) - p_{\text{CNN}}(x|x)|. \tag{5.14}$$

In (5.14), $\hat{p}(\cdot|\cdot)$ is the predicted CNN performance generated from Algorithm 2 and $p_{\text{CNN}}(\cdot|\cdot)$ is the true CNN performance obtained by implementing the CNN. Recall that $p(x|x)$ is the probability that the CNN correctly classifies an image belonging to class x.

Algorithm 2 Predicting Deep CNN Classification Accuracy via the S-UMRI model using Theorem 37.

Require: Dataset $\mathbb{D}$ (5.26) from K deep CNNs from a fixed network architecture. The k^{th} CNN is trained on a noisy dataset with added Gaussian noise with variance $\eta_k = 1 + 0.1 \times (k-1)$.

Step 1: *Constructing Interpretable Model.* The most robust utility functions $\{u_k^*\}_{k=1}^K$ and information acquisition cost L^* are computed by solving the following convex optimization problem.

$$\{u_k^*, c_k^*\}_{k=1}^K = \underset{u_{1:K}}{\text{argmax}} \frac{\varepsilon\ K}{\sum_{k=1}^K \|u_k\|_2^2},\ \text{BRP}(\mathbb{D}, \cdot) \leq -\varepsilon.$$
$$L^*(p(a|x)) = \max_{k=1} c_k^* + \sum_{x,a} \pi_0(x)(p(a|x) - p_k(a|x))u^*(x,a).$$

Step 2: *Predicting Classification Accuracy.* For an arbitrary noise variance $\eta \in [\eta_1, \eta_K]$, obtain index $g \in \mathbb{Z}_+,\ g \leq K$ such that $\eta \in [\eta_g, \eta_{g+1}]$. Then, the predicted classification accuracy $\hat{p}(a|x)$ for noise variance η is computed as follows:

$$\hat{p}(a|x) = \underset{p(a|x)}{\text{argmax}} \sum_a \max_b \sum_x \pi_0(x) p(a|x) \hat{u}(x,a) - L^*(p),$$
$$\hat{u} = 10 \times \{(\eta_{g+1} - \eta)u_g^* + (\eta - \eta_g)u_{g+1}^*\}. \tag{5.15}$$

Return: Predicted performance $\hat{p}(a|x)$ for noise variance η.

Discussion and Insights

(i) Our interpretable model can predict CNN classification performance with high accuracy (see below).

(ii) The interpretable model (utility functions and information acquisition cost) for our predictive procedure (Algorithm 2) is evaluated on the set of noise variances $G_1 = \{1 + 0.1 \times (k-1), k = 1, 2, \ldots, 11\}$. The predictive procedure of Algorithm 2 is applied on the the set of noise variances given by $G_2 = \{1.05 + 0.1 \times (k-1), k = 1, 2, \ldots, 10\}$. Table 5.4 displays the prediction errors $\delta_\eta(x)$ averaged over all $\eta \in G_2$.

(iii) From Table 5.4, the prediction error $\delta_\eta(x)$ averaged over all image classes x for the 5 CNN architectures are:

1. LeNet - 0.038
2. AlexNet - 0.036
3. VGG16 - 0.041
4. ResNet-50 - 0.027
5. NiN- 0.035

So the least accuracy is 95.9%, and highest accuracy is 97.3%.

(iv) The prediction error averaged over the network architectures was observed to be minimum for image class 'cat' (98.1%) and maximum for image class 'deer' (95.7%) over all image classes.

(iv) *Statistical Similarity between Deep CNNs and Interpretable Model.* We computed the Kullback-Leibler (KL) divergence between the true and predicted classification performances $p_{\text{imp}}(a|x)$ and $\hat{p}(a|x)$. Recall $\hat{p}(a|x)$ is computed from the interpretable model via Algorithm 2 and $p_{\text{CNN}}(a|x)$ is obtained from the CNN. The KL divergence values for the 5 CNN architectures are:

1. LeNet - 0.015
2. AlexNet - 0.012
3. VGG16 - 0.016
4. Resnet-50 - 0.006

5. NiN - 0.018.

Thus the decisions made by the deep CNNs are statistically similar to decisions generated by our interpretable model.

Remark. Although our numerical experiments only consider the CIFAR-10 image dataset, our results are straightforward to extend to larger and more granular datasets like CIFAR-100 [186] and ImageNet at the cost of greater computational resources.

5.4 Conclusions and Extensions

This chapter proposed a data-driven micro-economics based system identification approach for interpretable deep classification. The key results stem from Bayesian revealed preference. By embedding deep image classification in a constrained Bayesian utility maximization framework, interpretable deep image classification is equivalent to set-valued system identification of an argmax non-linearity (in signal processing terms). Based on approximately 500 experiments on 5 popular CNN architectures, we showed that deep CNNs can be explained remarkably well by Bayesian utility maximization constrained by an information cost.

Our main results were the following:
1. Using the theory of Bayesian revealed preference, Theorem 35 gave a necessary and sufficient condition for the actions of a collection of decision makers to be consistent with rationally inattentive Bayesian utility maximization. We showed that deep CNNs operating on the CIFAR-10 dataset satisfy these necessary and sufficient conditions.
2. Next we studied the robustness margin by which the deep CNNs satisfy Theorem 35; we found that the margins were sufficiently large implying robustness of the results. Our

robustness results are summarized in Table 5.1.

3. In Theorem 37, we constructed the sparsest interpretable model from the feasible set generated using Theorem 35. The sparsest interpretable model explains deep CNN decisions using the least number of parameters. The sparsest interpretable model introduces a useful preference ordering amongst the set of hypotheses (image labels) considered by the deep neural network; for example, how much additional priority is allocated to the classification of a cat as a cat compared to a cat as a dog. In classical deep learning, this preference ordering is not explicitly generated

4. Finally, we showed that our interpretable model can predict CNN performance with accuracy exceeding 94%, and the decisions generated by our interpretable model are statistically similar to that of a deep CNN. At a more conceptual level, our results suggest that deep CNNs for image classification are equivalent to an economics-based constrained Bayesian decision system (used in micro-economics to model human decision making).

Extensions. An immediate extension of this work is to use an auto-encoder to extract features and replace the image class label with the image features as the state in the constrained utility maximization model. This would result in a richer descriptive model of the CNN due to more degrees of freedom in the utility function.

Our proposed interpretable model generates a concave utility function by design. This is an important feature of the revealed preference framework; even though the actual deep learner's utility may not be convex. To quote Varian [77]: "If data can be rationalized by any non-trivial utility function, then it can be rationalized by a nice utility function. Violations of concavity cannot be detected with only a finite number of observations." A more speculative extension is to investigate the asymptotic behavior of the BRP and S-BRP decision tests for rationally inattentive utility maximization-do the tests pass when the number of deep CNNs tend to infinity? Recent results [214] show that an infinite

dataset can at best be rationalized by a quasi-concave utility function.

Reproducibility: The computer programs and deep image classification datasets needed to reproduce all the results in this chapter can be obtained from the public GitHub repository `https://github.com/KunalP117/DL_RI`.

5.5 Appendix

5.5.1 Proof of Theorem 35

Proof of necessity of NIAS and NIAC:

1. NIAS (5.8): For agent $k \in \mathcal{K}$, define the subset $\mathcal{Y}_a \subseteq \mathcal{Y}$ so that for any observation $y \in \mathcal{Y}_a$, given posterior pmf $p_k(x|y)$, the optimal choice of action is a (5.3). We define the revealed posterior pmf given action a as $p_k(x|a)$. The revealed posterior pmf is a stochastically garbled version of the actual posterior pmf $p_k(x|y)$, that is,

$$p_k(x|a) = \sum_{y \in \mathcal{Y}} \frac{p_k(x,y,a)}{p_k(a)} = \sum_{y \in \mathcal{Y}} p_k(y|a) p_k(x|y) \tag{5.16}$$

Since the optimal action is a for all $y \in \mathcal{Y}_a$, (5.3) implies:

$$\begin{aligned}
&\sum_{x \in \mathcal{X}} p_k(x|y)(u_k(x,b) - u_k(x,a)) \leq 0 \\
\implies &\sum_{y \in \mathcal{Y}_a} p_k(y|a) \sum_{x \in \mathcal{X}} p_k(x|y)(u_k(x,b) - u_k(x,a)) \leq 0 \\
\implies &\sum_{y \in \mathcal{Y}} p_k(y|a) \sum_{x \in \mathcal{X}} p_k(x|y)(u_k(x,b) - u_k(x,a)) \leq 0 \\
&(\text{since } p_k(y|a) = 0,\ \forall y \in \mathcal{Y} \backslash \mathcal{Y}_a) \\
\implies &\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_k(y|a) p_k(x|y)(u_k(x,b) - u_k(x,a)) \leq 0
\end{aligned}$$

$$\implies \sum_{x \in \mathcal{X}} p_k(x|a)(u_k(x,b) - u_k(x,a)) \leq 0 \text{ (from (5.16))}$$

This is precisely the NIAS inequality (5.8).

2. NIAC (5.9): Let $c_k = L(\boldsymbol{\alpha}_k) > 0$, where $L(\cdot)$ denotes the information acquisition cost of the collection of agents $\mathcal{K}$. Also, let $J(\boldsymbol{\alpha}_k, u_k)$ denote the expected utility of the k^{th} agent given attention strategy $\boldsymbol{\alpha}_k$ (first term in RHS of (5.4)). Here, the expectation is taken wrt both the state x and observation y. It can be verified that $J(\cdot, u_k)$ is convex in the first argument. Finally, for the k^{th} agent, we define the revealed attention strategy $\boldsymbol{\alpha}'_k$ over the set of actions $\mathcal{A}$ as

$$\boldsymbol{\alpha}'_k(a|x) = p_k(a|x), \ \forall a \in \mathcal{A},$$

where the variable $p_k(a|x)$ is obtained from the dataset $\mathbb{D}$. Clearly, the revealed attention strategy is a stochastically garbled version of the true attention strategy since

$$\boldsymbol{\alpha}'_k(a|x) = p_k(a|x) = \sum_{y \in \mathcal{Y}} p_k(a|y)\boldsymbol{\alpha}_k(y|x) \tag{5.17}$$

From Blackwell dominance [138] and the convexity of the expected utility functional $J(\cdot, u_k)$, it follows that:

$$J(\boldsymbol{\alpha}'_k, u_j) \leq J(\boldsymbol{\alpha}_k, u_j), \tag{5.18}$$

when $\boldsymbol{\alpha}_k$ Blackwell dominates $\boldsymbol{\alpha}'_k$. The above relationship holds with equality if $k = j$ (this is due to NIAS (5.8)). We now turn to condition (5.4) for optimality of attention strategy. The following inequalities hold for any pair of agents $j \neq k$:

$$\begin{aligned} J(\boldsymbol{\alpha}'_k, u_k) - c_k &\overset{(5.18)}{=} J(\boldsymbol{\alpha}_k, u_k) - c_k \\ \overset{(5.4)}{\geq} J(\boldsymbol{\alpha}_j, u_k) - c_j &\overset{(5.18)}{\geq} J(\boldsymbol{\alpha}'_j, u_k) - c_j. \end{aligned} \tag{5.19}$$

This is precisely the NIAC inequality (5.9).

Proof for sufficiency of NIAS and NIAC: Let $\{u_k, c_k\}_{k=1}^K$ denote a feasible solution to the NIAS and NIAC inequalities of Theorem 35. To prove sufficiency, we construct an UMRI tuple as a function of dataset $\mathbb{D}$ and the feasible solution that satisfies the optimality conditions (5.3),(5.4) of Definition 34.

Consider the following UMRI model tuple:

$$\Theta = (\mathcal{K}, \mathcal{X}, \mathcal{Y} = \mathcal{A}, \mathcal{A}, \pi_0, L, \{p_k(a|x), u_k, k \in \mathcal{K}\}), \text{ where}$$
$$L(p(a|x)) = \max_{k \in \mathcal{K}} c_k + J(p(a|x), u_k) - J(p_k(a|x), u_k). \tag{5.20}$$

In (5.20), $C(\cdot)$ is a convex cost since it is a point-wise maximum of monotone convex functions. Further, since NIAC is satisfied, (5.20) implies $L(\boldsymbol{\alpha}_k) = c_k$. It only remains to show that inequalities (5.3) and (5.4) in Definition 34 are satisfied for all agents in $\mathcal{K}$.

1. *NIAS implies* (5.3) *holds.* This is straightforward to show since the observation and action sets are identical.
2. *Information Acquisition Cost* (5.20) *implies* (5.4) *holds.* Fix agent $j \in \mathcal{K}$. Then, for any attention strategy $p(a|x)$, the following inequalities hold.

$$\begin{aligned}
&L(p(a|x)) = \max_{k \in \mathcal{K}} \; c_k + J(p(a|x), u_k) - J(p_k(a|x), u_k) \\
\implies &\; J(p_j(a|x)) - c_j \geq J(p(a|x)) - C(p(a|x)), \; \forall \, p(a|x) \\
\implies &\; p_k(a|x) \in \underset{p(a|x)}{\operatorname{argmax}} \, J(p(a|x), u_k) - L(p(a|x)) = \text{(5.4)}.
\end{aligned}$$

5.5.2 S-UMRI (Sparse UMRI) Model for Rationally Inattentive Bayesian Utility Maximization

In Sec. 5.2.1, we outlined the UMRI model for rationally inattentive utility maximization of K Bayesian agents parameterized by K utility functions and a cost of information

acquisition. This section proposes a sparse version of the UMRI model, namely, the S-UMRI model that is parameterized by a *single* utility function that rationalizes the decisions of K Bayesian agents. Abstractly, the S-UMRI model is described by the tuple

$$\Theta = (\mathcal{K}, \mathcal{X}, \mathcal{Y}, \mathcal{A}, \pi_0, L, u, \{\boldsymbol{\alpha}_k, \lambda_k, k \in \mathcal{K}\}). \tag{5.21}$$

All parameters in (5.21) are identical to that in (5.1) except for the additional parameter $\lambda_k \in \mathbb{R}_+$. λ_k can be interpreted as the sensitivity to information acquisition of the k^{th} agent. We discuss the significance of λ_k in more detail below. In complete analogy to Definition 34, Definition 38 below specifies the optimal action and attention strategy policy of the Bayesian agents in $\mathcal{K}$.

Definition 38 (Rationally Inattentive Utility Maximization for S-UMRI) *Consider a collection of Bayesian agents $\mathcal{K}$ parameterized by Θ in* (5.21) *under the S-UMRI model. Then,*

(a) **Expected Utility Maximization:** *Given posterior pmf $p(x|y)$, agent $k \in \mathcal{K}$ chooses action a that maximizes its expected utility.*

$$a \in \underset{a' \in \mathcal{A}}{\operatorname{argmax}} \ \mathbb{E}_x\{u_k(x, a')|y\} = \sum_{x \in \mathcal{X}} p(x|y)u(x, a') \tag{5.22}$$

(b) **Attention Strategy Rationality:** *Agent k chooses attention strategy $\boldsymbol{\alpha}_k$ that optimally trades off between utility maximization and cost minimization.*

$$\boldsymbol{\alpha}_k \in \underset{\boldsymbol{\alpha}'}{\operatorname{argmax}} \, \mathbb{E}_y\{ \max_{a \in \mathcal{A}} \mathbb{E}_x\{u(x, a)|y\}\} - \lambda_k L(\boldsymbol{\alpha}', \pi_0) \tag{5.23}$$

Remarks. 1. *Role of λ_k.* In (5.23), λ_k is the differentiating parameter across agents. Even though all agents have the same utility function, different values of λ_k result in different optimal strategies $\boldsymbol{\alpha}_k$ (5.23).

2. *Sparsity of S-UMRI*. The UMRI model tuple for K Bayesian agents is parameterized

using $K(|\mathcal{X}||\mathcal{A}|+1)$ variables. In comparison, the S-UMRI tuple is parameterized via $|\mathcal{X}||\mathcal{A}|+K$ variables. The difference in variables for parametrization is linear in K.

Finally, in complete analogy to Theorem 35, we now state Theorem 39 that states necessary and sufficient conditions for the decisions of a collection of Bayesian agents to be rationalized by the S-UMRI model.

Theorem 39 (S-BRP Test for Rationally Inattentive Utility Maximization) *Given the dataset* $\mathbb{D}$ (5.5) *obtained from a collection of Bayesian agents* $\mathcal{K}$*. Then,*
1. *Existence: There exists a S-UMRI tuple* $\Theta(\mathbb{D})$ (5.1) *that rationalizes dataset* $\mathbb{D}$ *if and only if there exists a feasible solution that satisfies the set of inequalities*

$$\textit{S-BRP}(\mathbb{D}) \leq \mathbf{0}. \tag{5.24}$$

In (5.6)*,* $\textit{S-BRP}(\cdot)$ *corresponds to a set of inequalities stated in Algorithm 3 below. The set-valued estimate of* Θ *that rationalizes* $\mathbb{D}$ *is the set of all feasible solutions to* (5.6)*.*
2. *Reconstruction: Given a feasible solution* $\{u, \lambda_k, c_k\}_{k=1}^K$ *to* $S-BRP(\mathbb{D},\cdot)$*,* u *is the* k^{th} *Bayesian agent's utility function, for all* $k=1,2,\ldots,K$*. The set of observations* $\mathcal{Y}=\mathcal{A}$*, the set of actions in* $\mathbb{D}$*. The feasible cost of information acquisition* L *in* $\Theta(\mathbb{D})$ *is defined in terms of the feasible variables* c_k, λ_k *as:*

$$L(\boldsymbol{\alpha}) = \max_{k\in\mathcal{K}} c_k + \lambda_k \sum_{x,a} (p(x,a) - p_k(x,a))u(x,a),$$
$$(\boldsymbol{\alpha} = \{p(a|x), a\in\mathcal{A}, x\in\mathcal{X}\}) \tag{5.25}$$

The proof of Theorem 39 closely follows the proof of Theorem 35 and hence, omitted. In comparison to the BRP test of Theorem 35, the S-BRP test has the same number of inequalities but fewer decision variables. Hence, the set of feasible parameters generated from Algorithm 3 is smaller compared to Algorithm 1.

Algorithm 3 S-BRP Convex Feasibility Test of Theorem 39

Require: Dataset $\mathbb{D} = \{\pi_0, p_k(a|x), x, a \in \mathcal{X}, k \in \mathcal{K}\}$ from a collection of Bayesian agents $\mathcal{K}$.

Find: Positive reals $c_k, \lambda_k, u \in (0, 1]$ for all $x \in \mathcal{X}$, $a \in \mathcal{A}$, $k \in \mathcal{K}$ that satisfy the following inequalities:

$$1.\ \sum_x p_k(x|a)\ (u(x,b) - u(x,a)) \leq 0, \forall a, b, k,$$

$$2.\ \sum_{x,a} (p_j(x,a) - p_k(x,a))u(x,a) + \lambda_k(c_k - c_j) \leq 0, \forall j, k,$$

where $p_k(x,a) = \pi_0(x)p_k(a|x),\ p_k(x|a) = \frac{p_k(x,a)}{\sum_{x'} p_k(x',a)}$.

Return: Set of feasible utility function u, scalars λ_k and information acquisition costs c_k incurred by agents $k \in \mathcal{K}$.

5.5.3 Construction of Deep CNN Dataset

We now explain how the decisions of the deep CNNs are incorporated into our main theorems Theorems 35 and 39. Suppose K deep CNNs indexed by $k = 1, \ldots, K$ with different training parameters are trained on the CIFAR-10 dataset. For every trained deep CNN k, given test image i from CIFAR-10 test dataset with image class s_i, let the vector $f_{i,k} \in \Delta^9$ denote the corresponding softmax output of the deep CNN. The vector $f_{i,k}$ is a 10-dimensional probability vector where $f_{i,k}(j)$ is the probability that deep CNN k classifies test image i into image class j.

The decisions of all K deep CNNs on the CIFAR-10 test dataset are aggregated into dataset $\mathbb{D}$ for compatibility with Theorems 35 and 39 as follows:

$$\begin{aligned}
&\mathbb{D} = \{\pi_0, p_k(a|x), x, a \in \mathcal{X}, k \in \{1, 2, \ldots, K\}\},\ \text{where} \\
&\pi_0(x) = \sum_{i=1}^{N} \frac{\mathbb{1}\{s_i = x\}}{N},\ p_k(a|x) = \frac{\sum_{i=1}^{N} \mathbb{1}\{s_i = x\} f_{i,k}(a)}{\sum_{i=1}^{N} \mathbb{1}\{s_i = x\}}, \\
&N = 10000,\ \mathcal{X} = \mathcal{A} = \{1, 2, \ldots 10\}.
\end{aligned} \tag{5.26}$$

Here $\pi_0(x)$ is the empirical probability that the image class of a test image in the CIFAR-

10 test dataset is x. Since the output of the CNN is a probability vector, we compute $p_k(a|x)$ for the k^{th} CNN by averaging the a^{th} component of the output over all test images in image class x. Finally, N is the number of test images in the CIFAR-10 test dataset, and the set of true and predicted image classes are the same, i.e., $\mathcal{X} = \mathcal{A}$. Although implicit in the above description, our Bayesian revealed preference approach to interpretable deep image classification assumes the deep CNN's (agent's) ground truth is the true image label, and its decision a is the predicted image label.

CHAPTER 6

INVERSE-INVERSE REINFORCEMENT LEARNING. SPOOFING AN INVERSE REINFORCEMENT LEARNER

6.1 Introduction

In abstract terms, a cognitive radar is a constrained utility maximizer with multiple sets of utility functions and constraints that allow the radar to deploy different strategies depending on changing environments. Cognitive radars adapt their waveform scheduling and beam allocation by optimizing their utility functions in different situations. If a smart adversary can estimate the utility function or constraints of the cognitive radar, then it can exploit this information to mitigate the radar's performance (e.g., jam the radar with purposefully designed interference). A natural question is: *how can a cognitive radar hide its cognition from an adversary?* Put simply, how can a smart sensor hide its strategy by acting dumb? We term this cognition-masking functionality as meta-cognition.[1] A meta-cognitive radar [215] switches between two modes of cognition; one mode to achieve a high quality estimate of a target, the other mode to hide its utility function (plan).

A meta-cognitive radar pays a penalty for stealth - it deliberately transmits sub-optimal responses to keep its strategy hidden from the adversary resulting in performance degradation. This chapter investigates how a cognitive radar hide its strategy when the adversary observes the radar's responses. Our meta-cognition results are inspired by privacy-preserving mechanisms in differential privacy and adversarial obfuscation in deep learning with related works discussed below. Although this chapter is radar-centric,

[1] "Meta-cognition" [215] is used to describe a sensing platform that switches between multiple objectives (constrained utility functions).

we emphasize that the problem formulation and algorithms also apply to adversarial inverse reinforcement learning in general machine learning applications, namely, how to purposefully choose sub-optimal actions to hide a strategy.

Related Works

Cognitive radars are widely studied in the literature [216–248][2]; see [216–218, 236] for comprehensive discussions on the cognitive radar literature. More recently, our chapters [103, 249] deal with inverse reinforcement learning (IRL) algorithms for cognitive radars, namely, how can an adversary estimate the utility function of a cognitive radar by observing its decisions. Reconstructing a decision maker's utility function by observing its actions is the main focus of IRL [6, 28, 250] in machine learning and revealed preference [2, 127] in micro-economics literature. In the radar literature, such IRL based adversarial actions to mitigate the radar's operations are called electronic countermeasures (ECM) [103, 126, 251]. This chapter builds on [103, 249, 252] and develops electronic counter countermeasures (ECCM) [253–255] to mitigate ECM. This chapter assumes that adversary's ECM is unaware if the radar has ECCM capability, which is consistent with state-of-the-art ECCM literature. The central theme of this chapter is to apply results from revealed preference in micro-economics theory [32, 127]. To the best of our knowledge, this approach for ECCM to hide cognition is not explored in literature.

Several works in literature [256–258] highlight how an adversary benefits from learning the radar's utility function. In [256], the adversary optimize its probes to increase the power of its statistical hypothesis test for utility maximization. [257, 258]

[2]We discuss cognitive radars in more detail in Sec. 6.2.2 and contextualize conventional models of radar cognition to the abstract constrained utility maximization framework assumed throughout this chapter.

show how revealed preference-based IRL techniques can be used to manipulate consumer behavior.

In the radar context, [259] uses the Laplacian mechanism for meta-cognition; the cognitive radar anonymizes its trajectories via additive Laplacian noise. Differential privacy-based adversarial obfuscation has seen success in applications such as ML [260], user data sharing [261] and recommendation systems [262]. In our cognition masking approach, the radar mitigates adversarial IRL via purposeful perturbations from optimal strategy, where the perturbations are computed via stochastic gradient algorithms (see Algorithm 5 in Sec. 6.4.2).

Outline and Organization of Results

(i) Background. Inverse reinforcement learning (IRL): In Sec. 6.2, we formulate the interaction between a cognitive radar and an adversary target. We first discuss several cognitive radar models studied in optimal waveform design and sensor management in Sec. 6.2.2. We then review the main idea of revealed preference-based adversarial IRL algorithms, namely, Theorems 41 and 50 in Sec. 6.2.3, that the adversary uses to reconstruct the radar's strategy from its actions. Then we outline two examples of cognitive radar functionalities, namely, waveform adaptation and beam allocation. Theorem 51 stated in Sec. 6.7.5 in the supplementary document extends adversarial IRL to the case where the cognitive radar faces multiple constraints. Theorem 51 is omitted from the main text for readability.

(ii) Masking Radar's Strategy from Adversarial IRL: Sec. 6.3 contains our main meta-cognition results, namely, Theorems 44 for mitigating adversarial IRL by masking the radar's strategy. The key idea is for the radar to deliberately deviate from its optimal

(naive) response to ensure:

(1) its true strategy almost fails to rationalize its perturbed responses (masked from adversarial IRL), and

(2) its performance degradation due to sub-optimal responses does not exceed a particular threshold. Theorem 53 in Sec. 6.7.5 extends Theorem 44 to the case where the cognitive radar has multiple constraints. Theorem 54 provides performance bounds on the cognition masking scheme of Theorem 44 when the adversary has misspecified measurements of the radar's response.

(iii) Masking Radar's Strategy from Adversarial IRL Detectors in noise. Sec. 6.4 extends our IRL and cognition masking results to the case where the adversary has noisy measurements of the radar's response. First, we define IRL detectors (Definition 45) that *detect* radar's cognition in noise. Then, we enhance our cognition masking scheme of Theorem 44 to mitigate the IRL detectors. The radar's cognition masking objective now is to maximize the detectors' conditional Type-I error probability, subject to a bound on its deliberate performance degradation.

(iv) Numerical illustration of masking cognition by meta-cognitive radars. Sec. 6.5 illustrates our meta-cognition results on two target tracking functionalities, namely, waveform adaptation and beam allocation. Our numerical experiments show that the meta-cognition algorithms in this chapter can effectively mask both the radar's utility function and resource constraint when the cognitive radar is probed by the adversarial target. Our main finding is that a small deliberate performance loss of the meta-cognitive radar suffices to mask the radar's strategy from the adversary to a large extent.

Running Example. Since the concept of ECCM via cognition masking is somewhat abstract, for the reader's convenience, we relate each assumption, definition and theorem introduced in this chapter at an implementation level to a real-world cognitive radar example. Specifically, we consider a cognitive radar [234] tracking an adversarial target.

6.2 Background. IRL to Estimate Cognitive Radar

Since this chapter investigates how to construct a cognitive radar that hides its utility from an adversarial IRL system, this section gives the background on how an adversarial system can use IRL to estimate the radar's utility. An important aspect of the IRL framework below is that it is a necessary and sufficient condition for identifying cognition (utility maximization behavior); hence it can be considered as an optimal IRL scheme. Sec. 6.7.7 and 6.7.6 discuss cognition masking when the adversary performs sub-optimal IRL.

6.2.1 Radar-Adversary Dynamics

Model 1 (Radar-Target Interaction) *The cognitive radar-adversary interaction has the following dynamics:*

$$\begin{aligned}
\textit{target probe: } & \alpha_k \in \mathbb{R}_+^d \\
\textit{radar action: } & \beta_k \in \mathbb{R}_+^d \\
\textit{target state: } & x_k = \{x_k(t),\ t = 1, 2, \ldots\}, \\
& x_k(t+1) \sim p_{\alpha_k}(x|x_k(t)),\ x_0 \sim \pi_0 \\
\textit{radar observation: } & y_k \sim p_{\beta_k}(y|x_k) \\
\textit{radar tracker: } & \pi_k = T(\pi_{k-1}, y_k) \\
\textit{observed radar action: } & z_k = \beta_k + \omega_k,\ \omega_k \sim f_\omega
\end{aligned} \tag{6.1}$$

Remarks. We now give examples for the abstract model (6.1).

1. A widely used example [263, 264] for the radar-adversary dynamics model (6.1) is that of linear Gaussian dynamics for target kinematics and linear Gaussian measurements:

$$x_k(t+1) = Ax_k(t) + w_t(\alpha_k), \quad x_k(0) \sim \pi_0 = \mathcal{N}(\hat{x}_0, \Sigma_0)$$

$$y_k(t) = Cx_k(t) + v_t(\beta_k), \quad k = 1, 2, \ldots, N \tag{6.2}$$

Here $x_k(t) \in \mathcal{X} = \mathbb{R}^X$, $y_k(t) \in \mathcal{Y} = \mathbb{R}^Y$. A is a block diagonal matrix [265] when the target state represents its position and velocity in Euclidean space. The variables $w_t \sim \mathcal{N}(0, Q(\alpha_k))$ and $v_t \sim \mathcal{N}(0, R(\beta_k))$ are mutually independent Gaussian noise processes.

2. In this chapter, we are only concerned with the asymptotic statistics of the radar tracker T (6.1) for our cognition-masking algorithms. One example is that of a Bayesian tracker (Kalman filter) where the asymptotic covariance of the state estimate is the unique positive semi-definite solution of the algebraic Riccati equation (ARE). Other tracker examples include the particle filter, interacting multiple model (IMM) filter etc.

We now proceed to define a cognitive radar which we assume in this chapter to be a constrained utility maximizer.

Definition 40 (Cognitive Radar) *Consider the radar-adversary interaction dynamics of Model 1. The cognitive radar chooses its response* β_k^* (6.1) *at time* k *by maximizing a utility function* $\mathbf{u}(\alpha_k, \cdot)$ *subject to constraint* $\mathbf{g}(\alpha_k, \cdot) \leq 0$*:*

$$\begin{aligned} \beta_k^* &\in \operatorname{argmax}\ \mathbf{u}(\alpha_k, \beta), \\ \mathbf{g}(\alpha_k, \beta) &\leq 0. \end{aligned} \tag{6.3}$$

We assume that $\mathbf{g}(\cdot)$ *is an increasing function of* β*.*

From a radar practitioner's perspective, let us briefly relate the parameters in Definition 40 to a cognitive radar-adversary interaction. Consider a cognitive radar as modeled in [234, Sec. 4B] tracking an adversarial target. The response β parametrizes the radar's transmitted waveform, the probe α_k parametrizes the state noise covariance matrix due to the adversary's maneuvers. In the cognitive radar context of [234], the utility function

$u(\cdot)$ is equivalent to the inverse of the transmitted signal power (radar minimizes its transmission power); the constraints $g(\alpha_k, \cdot) \leq 0$ can be interpreted as posterior Crámer-Rao bound (PCRB) constraints on the radar's estimate of the target's state[3]

Remarks.

1. In the main text of this chapter, we consider a single constraint. This is consistent with most works in cognitive radar literature which also assume a single operating constraint. For example, in [266], the cognitive radar is constrained by a bound on the target dwell time (monotone in the time the radar spends tracking each target). In [267], the radar's constraint is a bound on the receiver sensor processing cost (monotone in the radar's choice of sensor accuracy for target tracking). Hence, we only consider the operating cost of the radar in the main text which is reflected in the radar's scalar-valued constraint g in (6.3).

2. *Multiple resource constraints.* Our IRL methodology discussed below can be extended to multiple resource constraints (g is vector-valued). However, for readability, we only consider a scalar-valued constraint g in the main text of this chapter. We consider multiple resource constraints in Sec. 6.7.5. The notation for IRL and cognition masking results is complicated for vector-valued cost $\mathbf{g}(\cdot)$ and hence omitted from the main text and discussed in the supplementary document.

6.2.2 Radar Cognition as Constrained Utility Maximization

Cognitive radars have been studied extensively in the literature [216–248]. In this section, we discuss relevant works from the cognitive radar literature, and contextualize widely used models of radar cognition to the abstract constrained utility maximization framework

[3]It is straightforward to show that PCRB is inversely proportional to the radar sensor's signal-to-noise ratio (SNR) that depend on the target's maneuvers, hence $g(\alpha_k)$ can be viewed as SNR constraints with explicit dependence on the adversarial probe α_k.

proposed in Definition 40.

Cognitive Radars

The term "cognition" in cognitive radars is used to describe a number of functionalities such as optimal waveform design, knowledge-aided radar detection and tracking for minimizing response times, and sensor management. A cognitive radar [99, 123, 268] uses the perception-action cycle of cognition to sense the environment and learn from it relevant information about the target and the environment. Cognitive radars have also been modeled as reinforcement learners in the literature that maximize their utility [103, 249, 269–271] and tune their sensing resources to optimally satisfy mission objectives.[4]

Table 6.1 displays works in the cognitive radar literature related to the constrained utility maximizer framework of Definition 40. For brevity, we limit our discussion of cognitive radars to waveform design, sensor management and joint waveform-receiver filter design.

• *Radar cognition for optimal waveform design:* The signal-to-interference-noise ratio (SINR) is a widely used objective maximized by cognitive radars for waveform adaptation [219–224]. In [219, 220], the radar is constrained by the maximum peak-to-average ratio (PAR) of the transmission code that controls the variation of the code about its mean value, and hence, controls the transmission bandwidth. In [221–223], the radar is constrained by the total contiguous bandwidth available for transmission, and the resulting optimization problem results in the well-studied Sense-React-Notch paradigm. The cognitive radar in [224] faces multiple constraints, namely, bounds on the total transmission power, Hamming/Manhattan distance with respect to a reference code,

[4] In the context of DFIG (Data Fusion Information Group) process model [272], sensor adaptation by the radar can be viewed as Level 4-Process Refinement in the DFIG model.

and the interference power spilled over in undesirable frequency bands. We extend our cognition masking result of Theorem 44 to vector-valued constraints in Theorem 53 in the supplementary document.

The cognitive radar discussed in [225] minimizes a convex combination of two metrics, namely, the Spectral Integrated Level Ratio (SILR), a variable that is inversely proportional to the SINR, and the Integrated Cross-Correlation Level (ICCL) that measures the cross-correlation of the transmitted waveforms across multiple antennae. The transmitted waveform is constrained to be either constant modulus, or discrete phase (equivalent to M-ary Phase Shift Keying with a pre-specified alphabet size). In [226], the radar minimizes the $\mathcal{L}_2$-norm between the ambiguity function of the transmitted waveform and that of a reference waveform constrained by the total transmission power. The waveform design scheme in [230] resembles that of [226] in that the cognitive radar minimizes a convex combination of the interference power and the side lobe correlation, subject to a bound on the transmission power. In [227–229], the radar maximizes the mutual information (based on differential entropy) between the received signal and the impulse response of the target subject to a bound on the transmission power. In [235], the cognitive radar minimizes the posterior Crámer-Rao lower bound (CRLB) of the target estimate subject to a bound on the transmission power. The Crámer-Rao lower bound for the target estimate is also widely used in cognitive radars performing optimal sensor management as discussed below.

• *Radar cognition for optimal sensor management:* Analogous to optimal waveform design, SINR is also a widely used objective for optimal sensor management in cognitive radars [231, 232] where the radar is constrained by sensing constraints such as cost of changing the tracked cell in Euclidean space [231] and bound on downlink interference power [233]. The posterior and predicted Crámer-Rao lower bounds (CRLB) for the

target estimate are also widely used optimization metrics for cognitive radar performing optimal sensor deployment [235, 236] where the radar faces constraints such as bounds on the communication cost with the central processing unit [235] and bounds on the sensing and processing cost [236]. A similar model is proposed in [237] where the radar optimizes its sensor deployment locations, and the number of active sensors. The radar minimizes the mean squared tracking error subject to constraints on the number of sensors deployed. [238–240] addresses optimal beamsteering for cognitive radars. To choose the optimal cell for focusing its transmit beam, the radar maximizes the entropy of the target's location. For target tracking, the optimal sensor parameters minimize the target's tracking entropy (based on location and velocity of the target). Finally, [241, 242] design cognitive radars for adaptive target detection that maximize the likelihood of target emission on a 2-dimensional grid, subject to block sparsity constraints on the target location.

• *Radar cognition for joint waveform-receiver filter optimization:* Joint optimization of waveform and receiver filter design is well explored in the cognitive radar literature; we discuss a few notable works below. Note that the radar optimizes over two variables, namely, the receiver filter and the transmitted waveform. In [243, 244], the radar minimizes the clutter/interference power at the receiver subject to the well-known Capon [273] constraint, namely, a normalization constraint on the received signal power. In addition, the radar in [243] is subject to an equality (normalization) constraint on the received signal power, and a bound on the transmission power in [243]. The radar in [244] faces an addition waveform constraint (identical to [226]), namely, the transmission waveform is constrained to be either constant modulus or discrete-phase. On a related note, robust constrained Capon beamforming is investigated in [274–276]. [245] considers a bi-static cognitive radar transmitting two waveforms. The joint waveform-receiver filter optimization is done in two steps: first, the optimum receiver filters are computed that

maximize the receiver SINR. Then, the optimal waveforms are computed that maximize the signal power at the radar receiver subject to (i) orthogonality constraint on the two waveforms, (ii) transmission power constraints and (iii) bounds on $\mathcal{L}_2$-deviation from a set of reference waveforms. The authors of [245] generalize their work to multi-static radars in [246] and to cognitive radar networks in [247]. Finally, [248] maximize the signal-to-clutter-noise-ratio (SCNR) at the receiver subject to a bound on the transmission power. The key idea is that the introduction of a physics-based scattering model for the clutter environment makes the maximization of SCNR tractable unlike traditional approaches.

Meta-cognitive Radars

A meta-cognitive radar [216, 277–279]transcends conventional notions of 'cognition' in radars. In this chapter, we view meta-cognition as the radar's sensing ability to identify an adversary in its environment, and strategic ability to spoof the adversary using inverse-inverse reinforcement learning techniques to 'mask' its cognition. The working assumption of the chapter is that an adversary can (a) identify the cognitive ability of a radar, and (b) mitigate the radar's operations based on this information. Recent works address how to identify cognitive radars by analysing a finite time series of emission exchanges with the radar [103, 249, 252]; a summary of the strategy identification results is presented in Sec. 6.2.3 below.

Radar functionalities that mitigate adversarial systems are termed as ECCM in radar literature; see [280] for a comprehensive discussion. Low-probability-of-intercept (LPI) transmission design [281–283] achieves stealth for cognitive radars and avoid cognition detection. Waveform adaptation schemes to counter barrage jamming are studied in [253, 254]. Frequency diversity for stealth-based ECCM in multi-target and moving target

Works	Category	Utility	Constraint
[219, 220]	Waveform	Minimum SINR of finitely many users	Bound on Peak-to-Average Ratio (PAR)
[221–223]	Waveform	SINR (Sense-React-Notch Paradigm)	Bound on contiguous transmission bandwidth
[224]	Waveform	SINR	Bounds on interference power in restricted frequency bands, total transmission power, and lower bound on similarity wrt a reference code
[225]	Waveform	Negative of convex combination of SILR (interference), ICCL (waveform correlation)	Baseband transmission codes are constrained to be either constant modulus or discrete phase
[226]	Waveform	Negative of $\mathcal{L}_2$-deviation from a desired ambiguity function	Bound on transmission power
[227–229]	Waveform	Mutual information (M.I.) between measured signal and target impulse response	Bound on transmission power
[230]	Waveform	Negative of convex combination of interference power in restricted frequency bands and side-lobe correlation	Bound on transmission power
[231–233]	Tracked cell	SINR	Bounds on Euclidean distance between current and next tracked cell (cost of shifting target cell), downlink interference power
[234, 235]	Sensor, Waveform	Negative of predicted posterior Crámer-Rao lower bound (CRLB) for target estimate	Bounds on transmission power, communication cost
[236]	Sensor	Negative of predicted conditional CRLB of target estimate	Bounds on sensing cost, processing cost
[237]	Sensor	Negative of root mean squared error (RMSE) between target state and estimate	(i) *For sensor deployment locations:* unconstrained (global optimum achieved in finitely many steps), (ii) *For number of sensors to be deployed:* Bound on deployed sensors
[238–240]	Beamsteering	Target position entropy (as a function of target cell)	Bound on target tracking entropy
[241, 242]	Sensor	Likelihood of emission on a 2-D grid under Gaussian measurement model	Emission activity norm (block-sparsity constraint)
[243]	Joint waveform-receiver filter	Negative of interference and clutter power at the receiver	Normalization constraint on the received signal power (Capon constraint), Bound on transmission power
[244]	Joint waveform-receiver filter	Negative of interference power at the receiver	Capon constraint, code constraint from [226]
[245–247]	Joint waveform-receiver filter	SINR, signal power at the receiver	Orthogonality constraint between waveforms, bounds on transmission power, bounds on $\mathcal{L}_2$-distance from a set of reference waveforms
[248]	Joint waveform-receiver filter	SCNR	Bound on transmission power

Table 6.1: Cognitive radars as constrained utility maximizers. In the table above, we contextualize several notable works in the cognitive radar literature according to the abstract constrained utility maximization setup of Definition 40. For every cognitive radar model discussed in Sec. 6.2.2, we list the optimization type or 'category', the equivalent utility being maximized by the radar, and the resource constraint faced by the radar. The meta-cognition algorithms in this chapter provide a principled approach to spoof an adversary that can identify the radar's plan and mitigate the radar's operations.

tracking applications is studied in [284–286].

While the works discussed above mitigate an adversary, the ECCM measures do not necessarily mask the radar's cognition. The meta-cognitive radar's aim in this chapter is to *confuse* the adversary's detector and *hide* its cognition, i.e. , ensure the adversary incorrectly reconstructs the radar's strategy with high probability, by deliberately transmitting sub-optimal responses. Specifically, this chapter contributes to anti-stealth and anti-ARM ECCM [287] by ensuring that adversarial mitigation is ineffective with a large probability.

6.2.3 Adversarial IRL for Identifying Strategy of Cognitive Radar

We now review the main results for adversarial IRL, namely, how an adversary can identify and reconstruct the radar's strategy by observing the radar's responses. The adversarial IRL system is schematically shown in Fig. 6.1. The key idea is to formulate the adversary's task of identifying the radar's strategy as a linear feasibility problem in terms of the radar's responses. This chapter considers two distinct scenarios in terms of the dependency of the adversary's probe α_k on the radar's utility $\mathbf{u}$ and resource constraint $\mathbf{g}$ in (6.3). The two scenarios are formalized in Assumptions 2 and 3 below in our IRL results, Theorems 41 and 50, and justified in Sec. 6.2.4 in the tracking examples of waveform adaptation and beam allocation.

IRL for Identifying Radar's Utility Function

In works such as [225,230], the adversary can mitigate the cognitive radar if the adversary knows the utility weights. Theorem 41 below provides a set-valued reconstruction

algorithm to estimate the radar's utility function when the adversary controls the radar's resource constraint. Such scenarios where the adversary knows the radar's resource constraint is formalized below in Assumption 2:

Assumption 2 *The radar's resource constraint* $\mathbf{g}(\cdot)$ *in* (6.3) *is linear in the adversary's probe* α_k *and the radar's utility* $\mathbf{u}(\cdot)$ *is independent of* α_k*:*

$$\mathbf{g}(\alpha_k, \beta) = \alpha_k' \beta - 1, \ \mathbf{u}(\alpha_k, \beta) \equiv \mathbf{u}(\beta) \tag{6.4}$$

IRL objective. The adversary aims to reconstruct the radar's utility $\mathbf{u}(\cdot)$ *using the dataset* $\mathcal{D}_g$*, where* $\mathcal{D}_g$ *is defined as:*

$$\mathcal{D}_g = \{\mathbf{g}(\alpha_k, \cdot), \beta_k\}_{k=1}^N, \tag{6.5}$$

where $\mathbf{g}(\alpha_k, \cdot)$ *is defined in* (6.4)*.*

In spite of its linear structure, the constraint in (6.4) can model non-linear radar constraints via a suitable definition of the radar's response β and the adversary's probe α. For example, an upper bound on the asymptotic precision of the radar's state estimate (inverse of the solution of the algebraic Riccati equation (ARE)) can be expressed as a linear constraint in terms of the eigenvalues of the state and noise covariance matrix; see [103, Lemma 3] for a detailed exposition. Let us now state Theorem 41 for achieving IRL when assumption 2 holds.

Theorem 41 (IRL for Identifying Radar's Utility Function) *Consider the cognitive radar described in Model 1. Suppose assumption 2 holds. Then:*

(a) The adversary checks for the existence of a feasible utility function that satisfies (6.3) *by checking the feasibility of a set of linear inequalities:*

$$\begin{aligned} &\textit{There exists a feasible } \theta \in \mathbb{R}_+^{2N} \textit{ s.t. } \mathcal{A}(\theta, \mathcal{D}_g) \leq \mathbf{0}, \\ \Leftrightarrow &\exists \ u \textit{ s.t. } \beta_k \in \operatorname{argmax} u(\beta), \ \alpha_k' \beta \leq 1 \ \forall k, \end{aligned} \tag{6.6}$$

where dataset $\mathcal{D}_g$ is defined in (6.5) *and the set of inequalities $\mathcal{A} \leq \mathbf{0}$ is defined in Sec. 6.7.1.*

(b) If $\mathcal{A}(\cdot, \mathcal{D}_g) \leq \mathbf{0}$ has a feasible solution, the set-valued IRL estimate of the radar's utility $\mathbf{u}$ *is given by:*

$$u_{\mathrm{IRL}}(\beta) \equiv \{u_{\mathrm{IRL}}(\beta;\theta) : \mathcal{A}(\theta, \mathcal{D}_g) \leq \mathbf{0}\},$$
$$u_{\mathrm{IRL}}(\beta;\theta) = \min_{k \in \{1,2,..,N\}} \{\theta_k + \theta_{k+N}\ \alpha_k'(\beta - \beta_k)\}. \tag{6.7}$$

Theorem 41 is well known in micro-economics as Afriat's theorem [32, 127] and widely used for set-valued estimation of consumer utilities from offline data. In complete analogy, the adversary also performs IRL on a batch of probe-response exchanges with the cognitive radar to reconstruct the radar's utility [5]. Abstractly, Theorem 41 says that given a finite dataset, the adversary can at best construct a polytope of feasible strategies that rationalize the adversary's dataset. Theorem 41 achieves IRL when the radar faces a single operating constraint. We discuss adversarial IRL for multiple resource constraints in Theorem 51 in Sec. 6.7.5. Then the linear feasibility test of (6.6) generalizes to a mixed-integer linear feasibility test, linear in the real-valued feasible variables in the multi-constraint case.

The important aspects of Theorem 41 to a practitioner are the following: Unlike typical *reactive* ECM systems, the adversarial target in this chapter is assumed to be a *cognitive* entity [288]. The cognitive ECM entity has the capability to: (1) estimate the radar's strategy encoded in its utility function $\mathbf{u}$, and then (2) perform adversarial maneuvers $(\alpha_{1:N})_{\mathrm{Adv}}$ that minimize the radar's utility:

$$(\alpha_{1:N})_{\mathrm{Adv}} \in \operatorname*{argmin}_{\alpha_{1:N}} \sum_{k=1}^{N} \max_{\beta_{1:N}} \mathbf{u}(\beta_k),\ \mathbf{g}(\alpha_k, \beta_k) \leq 0 \tag{6.8}$$

[5] Afriat's theorem with linear constraints (6.4) has been generalized to non-linear monotone constraints in literature [124]. For the radar context in this chapter, it suffices to assume a linear constraint when the adversary is trying to estimate the radar's utility

In the context of the cognitive radar modeled in [234], the utility function could be a Quality-of-Service (QoS) metric [289] the radar maximizes to yield the optimal waveform parameter (instead of simply minimizing the transmission power). The ECM objective in this scenario would be to identify the radar's QoS function for mitigating its operations. Through the reconstructive procedure of (6.47) in Theorem 41, the adversary can estimate the radar's utility, and then use (6.8) to design optimal maneuvers that minimize the radar's QoS.[6]

IRL for Identifying Radar's Resource Constraints

In certain scenarios, the utility of the radar is well known (e.g., signal-to-noise ratio (SNR)), but the operational constraints of the radar are not known, for example, bound on the Peak-to-Average Ratio (PAR) [219, 220]. We formalize such scenarios where the adversary knows the radar's utility function below as Assumption 3:

Assumption 3 *The radar's utility function* $\mathbf{u}(\cdot)$ (6.3) *is controlled by the adversary's probe* α_k*, the radar's resource constraint* $\mathbf{g}$ *is independent of* α_k *and has the following form:*

$$\mathbf{g}(\alpha_k, \beta) \equiv \mathbf{g}(\beta) - \gamma_k,\ \gamma_k > 0, \tag{6.9}$$

where γ_k, g *are independent of* α_k*.*

IRL objective. The adversary aims to reconstruct $\mathbf{g}(\cdot)$ *using the dataset* $\mathcal{D}_u$*, where* $\mathcal{D}_u$ *is defined as:*

$$\mathcal{D}_u = \{\mathbf{u}(\alpha_k, \cdot), \beta_k\}_{k=1}^{N}. \tag{6.10}$$

[6] A popular framework to study radar-adversary interactions of the form in (6.8) is the principal agent problem (PAP). We refer the reader to [290, 291] where the authors design ECCM strategies using a PAP framework for adversarial mitigation.

IRL for estimating the radar resource constraints has the same structure as that of Theorem 41 and is discussed in the Appendix. IRL for Assumption 3 is formally stated in Theorem 50 in Sec. 6.7.2 and summarized below:

$$g_{\mathrm{IRL}}(\beta) \equiv \{g_{\mathrm{IRL}}(\beta;\theta) : \mathcal{A}(\theta, \mathcal{D}_u) \geq \mathbf{0}\}, \tag{6.11}$$
$$g_{\mathrm{IRL}}(\beta;\theta) = \max_{k\in\{1,2,\ldots,N\}} \{\theta_k + \theta_{N+k}(\mathbf{u}(\alpha_k,\beta) - \mathbf{u}(\alpha_k,\beta_k))\},$$

where g_{IRL} is the adversary's set-valued estimate of the radar's constraint $\mathbf{g}$, dataset $\mathcal{D}_u$ is defined in (6.10) and $\theta \in \mathbb{R}^{2N}_{+}$ is a feasible vector wrt the feasibility test $\mathcal{A}(\cdot, \mathcal{D}_u) \geq \mathbf{0}$. Note how the IRL feasibility inequalities in (6.11) are identical to that of (6.6) in Theorem 41 but with the inequality direction reversed.

Theorem 50 is useful when the adversary is interested in evaluating the radar's constraints. Consider the cognitive radar in [234]. The adversary knows the radar's utility, for example, the signal-to-noise ratio (SNR). The adversary's aim instead is to estimate the radar's constraints on the cost of communication [234, Eq. 40] with the central processing unit. Knowledge of the radar's communication cost facilitates adversarial maneuver selection as:

$$(\alpha_{1:N})_{\mathrm{Adv}} \in \operatorname*{argmin}_{\alpha_{1:N}} \sum_{k=1}^{N} \max_{\beta_{1:N}} \mathbf{u}(\alpha_k,\beta_k),\ \mathbf{g}(\beta_k) \leq \gamma_k, \tag{6.12}$$

where the utility function is simply the radar sensor's SNR that indeed depends on the adversary's maneuvers (parametrized by probe α_k), $g(\cdot)$ is the radar's communication cost, and γ_k is the cost threshold at time step k.

6.2.4 Examples of IRL for Identifying Radar Cognition

Below, we discuss two examples of cognitive radar functionalities, namely, waveform adaptation and beam allocation. Throughout this chapter, we will use the two examples

below for contextualizing our cognition masking results.[7]

Example 1. Waveform Adaptation for Cognitive Radar

Waveform adaptation [292–297] is a crucial functionality of a cognitive radar. Consider a cognitive radar with linear Gaussian dynamics and measurements (6.2). The cognitive radar's aim is to choose the optimal sensor mode (observation noise covariance) based on the target's maneuvers. A more accurate sensor results in more precise observations, but is also costlier to deploy. Sec. 3.3.3 formalizes the optimal waveform adaptation and abstracts the problem as the constrained utility maximization problem of (6.3). In simple terms, the cognitive radar maximizes its observation noise covariance (least accurate sensing mode) subject to a lower bound on the radar's SNR. The key idea is to assume that adversary's probe α_k and radar's response β_k are the eigenvalues of covariance matrices Q and R^{-1}, respectively, and hence, parameterize the state and observation noise covariance in the state space model of (6.2). We encourage the reader to refer to Sec. 3.3.3 that shows the equivalence between an upper bound on the radar's asymptotic covariance $(\Sigma^*(\alpha_k, \beta_k))^{-1}$ and the linear constraint $\alpha_k'\beta \leq 1$. In summary, the cognitive radar's optimal waveform adaptation strategy can be abstracted as:

$$\beta_k \in \operatorname{argmax}\ \mathbf{u}(\beta),\ \alpha_k'\beta \leq 1, \tag{6.13}$$

where u is the radar's utility, and the linear constraint $\alpha_k'\beta \leq 1$ equivalently bounds the *asymptotic precision* of the radar.

Let us briefly discuss the state-of-the-art in waveform design in the radar literature, and show how optimal waveform design can be embedded in the abstract constrained

[7] In Appendices 6.7.3, we formally relate the variables in (6.14) to tracker-level parameters of the cognitive radar. We encourage the reader to refer to Sec. 3.3.3 for a detailed discussion of waveform optimization, and contextualization of the variables in (6.13).

utility maximization setup of (6.13). In [292], the constraint in (6.13) is a bound on the waveform power; the utility function is either the conditional mutual information between the target impulse response and the reflected waveforms given the knowledge of transmitted waveform, or simply the negative of the mean squared error between the true and estimated location of the target being tracked, with both choices of utility function yielding the same optimal waveform choice. In [294], the authors study waveform design in omni-directional radars where the radar's utility function (6.13) is the negative of the downlink multi-user interference and the resource constraint is simply a bound on the transmitted power. In [296], the radar's utility is the negative of the Crámer-Rao bound on the variance of the radar's state estimate; the radar's resource constraint is a bound on its transmission power. The authors in [295] design optimal waveforms with an added ECCM functionality to mitigate ECM. The key idea is to first send a pilot waveform to estimate the parameters of the adversary's ECM, followed by intra-pulse frequency coding with appropriate parameters to deceive the adversary's ECM. Our ECCM approach is similar to that of [295] with the only difference that instead of increasing the bandwidth of our transmitted signal to combat smart noise jamming, the cognitive radar transmits sub-optimal waveforms to avoid its strategy from being reconstructed by the adversary.

IRL for optimal waveform adaptation. The adversary's aim is to identify the radar's utility function u. Also, the setup of (6.13) falls under Assumption 2. Hence, the adversary uses the IRL test of (6.6) in Theorem 41 for identifying u.

Example 2. Beam allocation for Cognitive Radar

Sec. 6.7.3 discusses optimal beam allocation [98, 298–301]. [8] The cognitive radar's aim is to allocate its beam intensity optimally between multiple targets. Compared to a target with less jerky maneuvers, a target with unpredictable maneuvers requires a more focused beam for the SNR to lie above a certain threshold. Sec. 6.7.3 formalizes the beam allocation problem and abstracts the problem as a constrained utility maximization problem (6.3). The key idea is to relate the adversary's probe α_k to the asymptotic *predicted* precision of the radar tracker. In summary, the cognitive radar's optimal waveform adaptation problem can be abstracted as:

$$\beta_k \in \text{argmax } \mathbf{u}(\alpha_k, \beta) \equiv \prod_{i=1}^{m} \beta(i)^{\alpha_k(i)}, \ \|\beta\|_\kappa \leq \gamma_k, \tag{6.14}$$

where the radar maximizes a Cobb-Douglas utility subject to a bound γ_k on the total transmit beam intensity (κ-norm of intensity vector) for all k.

IRL for optimal beam allocation. Since the adversary knows the radar's utility (Assumption 3), its aim is to identify the radar's constraint $\mathbf{g}(\cdot) - \gamma_k \leq 0$ using the IRL test (6.50) in Theorem 50.

Summary. This section discussed how an adversary can deploy IRL to estimate a cognitive radar's utility and constraint. While IRL with a single operational constraint is discussed in [103], the IRL algorithm for multiple constraints (in Sec. 6.7.5) is new. This section also related Theorem 41 for identifying radar cognition to the parameters of a cognitive radar [234].

[8] Although similar to the abstract setup in Sec. 3.3.2, a key difference here is that the cognitive radar knows the adversary's utility, but does not know the resource constraint. Hence, the classical Afriat's theorem needs to be modified to estimate the adversary's resource constraint instead of the utility.

6.3 Inverse IRL. Masking Radar Utility and Constraints from Adversarial IRL

Having discussed how an IRL system can detect a cognitive radar, we are now ready to design a cognitive radar that is aware of the adversary's IRL motives and hides its strategy (utility function and resource constraints) from the IRL system. In radar terminology, IRL for mitigating a radar system falls under the field of electronic counter measures (ECM). Since meta-cognition deals with spoofing adversarial IRL, it can be viewed as a form of electronic counter-counter measure (ECCM) against ECM.

Rationale. How to hide cognition? Recall that the feasibility of (6.6), (6.50) is both necessary and sufficient for identifying utility maximization behavior (6.3); see [32, 127] for the proof. Hence, a cognitive radar's true strategy lies within the polytope of feasible strategies computed by the adversary (Fig. 6.1). The cognition masking rationale in this chapter is to transmit purposefully designed perturbed responses that ensure the radar's true strategy lies close to the edge of the polytope of feasible strategies. The distance from the edge of the feasibility polytope is a measure of goodness-of-fit of the strategy to the radar's responses; see Definitions 42, 43 below. In other words, the radar deliberately sacrifices performance to ensure its strategy poorly rationalizes its perturbed responses, hence hiding its strategy from adversarial IRL.

Main Result. How a radar can mask its utility/constraints

Theorem 44 below is our main result for cognition masking. Theorem 44 uses the concept of feasibility margin - how far is a strategy from failing the IRL feasibility tests (6.6) or (6.50). We define two margins - $\mathcal{M}_u$ and $\mathcal{M}_g$ for the feasibility margins of feasible

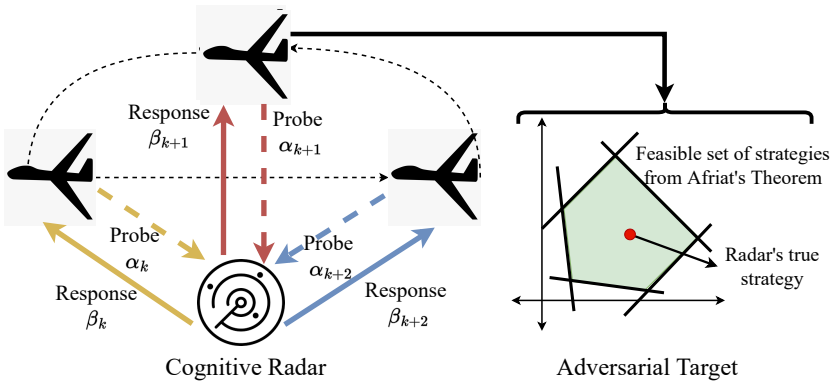

Figure 6.1: Schematic of adversarial IRL against cognitive radars. The adversary observes a sequence of decisions of the cognitive radar in response to a sequence of adversarial probes. Revealed preference-based adversarial IRL (Afriat's Theorem) [32, 127] is equivalent to checking the existence of a feasibility polytope for a set of inequalities (Afriat's Theorem [32, 127]). Our aim in this chapter is to make adversarial IRL cumbersome - how to purposefully distort radar responses meta-cognition objective in this chapter is to spoof adversarial IRL, namely, how to make checking linear feasibility difficult.

utilities u and constraints g, respectively.

Definition 42 (Feasibility Margin for Reconstructed Utility (6.6)) *Consider the dataset* $\mathcal{D}_g$ *defined in* (6.5). *The feasibility margin* $\mathcal{M}_u(\mathcal{D}_g)$ *defined below measures how far is the utility* u *from failing the IRL feasibility test* (6.6).

$$\mathcal{M}_u(\mathcal{D}_g) = \min_{\varepsilon \geq 0} \ \varepsilon, \ \ \mathcal{A}(u, \mathcal{D}_g) + \varepsilon \mathbf{1} \geq \mathbf{0}, \tag{6.15}$$

where $\mathbf{1}$ *is the column vector of all ones.*

Definition 43 (Feasibility Margin for Reconstructed Constraints (6.50)) *Consider the dataset* $\mathcal{D}_u$ *defined in* (6.10). *The feasibility margin* $\mathcal{M}_g(\mathcal{D}_u)$ *defined below measures how far the constraint* g *is from failing the IRL feasibility test* (6.50)*:*

$$\mathcal{M}_g(\mathcal{D}_u) = \min_{\varepsilon \geq 0} \varepsilon, \ \ \mathcal{A}(g, \mathcal{D}_u) - \varepsilon \mathbf{1} \leq \mathbf{0}, \tag{6.16}$$

where $\mathbf{1}$ *is the column vector of all ones.*

The margin (6.15), (6.16) is a measure of goodness-of-fit for the IRL feasibility inequalities (6.6) and (6.50), respectively, for any feasible strategy.[9] If u is a feasible utility that rationalizes $\mathcal{D}_g$ (6.5), we have $\mathcal{A}(u, \mathcal{D}_g) \leq \mathbf{0}$ from (6.6). Hence, the margin for u is the minimum *non-negative* perturbation so that the IRL test of (6.6) fails, that is, $\mathcal{A}(\cdot, \mathcal{D}_g) + \varepsilon \mathbf{1} \geq \mathbf{0}$. Similarly, if g is a feasible resource constraint that rationalizes $\mathcal{D}_u$ (6.10), we have $\mathcal{A}(u, \mathcal{D}_g) \geq \mathbf{0}$ from (6.50). Hence, the margin for u is the minimum *non-positive* perturbation so that the IRL test of (6.6) fails, that is, $\mathcal{A}(\cdot, \mathcal{D}_g) - \varepsilon \mathbf{1} \geq \mathbf{0}$. Equivalently, the margin measures how far a strategy lies from the edge of the polytope of feasible strategies. [10] The concept of margins arises in many prominent areas of machine learning, for example, in support vector machines (SVM) [307] for classification tasks and also max-margin IRL [26]. In the radar context, a strategy with a large feasible margin is a high-confidence point estimate of the radar's strategy and hence, at higher risk of getting exposed.

We are now ready to state our first cognition masking result, Theorem 44. Theorem 44 ensures the radar's true strategy has a low feasibility margin wrt the IRL tests of Theorems 41, 50 by deliberately perturbing the radar's naive responses (6.3). In a sense, the radar optimally switches between maximizing its performance and maximizing the privacy of its plan.

Theorem 44 (Masking Cognition from Adversarial IRL Feasibility Tests.) *Consider the cognitive radar* (6.3) *in Definition 40. Let* $\{\beta_k^*\}_{k=1}^N$ *denote the naive response sequence* (6.3) *that maximizes the cognitive radar's utility. Then:*

[9] Strictly speaking, the margin (6.15) is the minimum perturbation so that $\mathcal{A}(u_{\mathcal{A}}, \mathcal{D}_u)$ is infeasible, where $u_{\mathcal{A}}$ is the finite-dimensional projection of u for the IRL feasibility test defined in (6.48) in Sec. 6.7.1. However, we abuse notation and express the feasibility test as $\mathcal{A}(u, \mathcal{D}_u)$ for the sake of simplicity of exposition. We abuse notation in a similar way for (6.16)

[10] There exist several robustness measures in literature [302–306] that check how well a *dataset* satisfies economic based rationality. Our cognition masking aim is more subtle - our aim is to ensure a particular *strategy* rationalizes a dataset poorly by minimizing its feasibility margin (6.15) (6.16).

(i) Masking Utility Function from IRL. Suppose Assumption 2 holds. The response sequence $\{\tilde{\beta}^*_{1:N}\}$ *defined below masks the radar's utility* $\mathbf{u}$ *from the adversary by ensuring* $\mathbf{u}$ *passes the IRL feasibility test* (6.6) *with a sufficiently low margin* (6.15) *parametrized by* $\eta \in [0,1]$*:*

$$\{\tilde{\beta}^*_{1:N}\} = \underset{\{\beta_k \geq \mathbf{0},\ \alpha'_k \beta_k \leq 1\}}{\operatorname{argmin}} \sum_{k=1}^{N} \mathbf{u}(\beta^*_k) - \mathbf{u}(\beta_k), \tag{6.17}$$

$$\mathcal{M}_{\mathbf{u}}(\mathcal{D}_g) \leq (1-\eta)\ \mathcal{M}_{\mathbf{u}}(\mathcal{D}^*_g), \tag{6.18}$$

where dataset $\mathcal{D}^*_g = \{\alpha'_k(\cdot) - 1, \beta^*_k\}_{k=1}^N$ *is the adversary's dataset when the radar transmits naive responses* $\{\beta^*_k\}_{k=1}^N$*, and* $\mathcal{D}_g$ *is defined in* (6.5)*.*

(ii) Masking Resource Constraint from IRL. Suppose Assumption 3 holds. The response sequence $\{\tilde{\beta}^*_{1:N}\}$ *defined below masks the radar's resource constraint* $\mathbf{g}$ *from the adversary by ensuring* $\mathbf{g}$ *passes the IRL feasibility test* (6.50) *with a sufficiently low margin* (6.16) *parametrized by* $\eta \in [0,1]$*:*

$$\{\tilde{\beta}^*_{1:N}\} = \underset{\{\beta_k \geq \mathbf{0},\ g(\beta_k) \leq \gamma_k\}}{\operatorname{argmin}} \sum_{k=1}^{N} \mathbf{u}(\beta^*_k) - \mathbf{u}(\beta_k), \tag{6.19}$$

$$\mathcal{M}_{\mathbf{g}}(\mathcal{D}_u) \leq (1-\eta)\, \mathcal{M}_{\mathbf{g}}(\mathcal{D}^*_u), \tag{6.20}$$

where dataset $\mathcal{D}^*_u = \{\mathbf{u}(\alpha_k, \cdot), \beta^*_k\}_{k=1}^N$ *is the adversary's dataset when the radar transmits naive responses* $\{\beta^*_k\}_{k=1}^N$*, and* $\mathcal{D}_u$ *is defined in* (6.10)*.*

Theorem 44 is our first result for masking cognition; see Algorithm 4 for a stepwise procedure for masking the radar's utility (6.17). This is schematically illustrated in Fig. 6.3. Theorem 44 computes the *optimal* sub-optimal response of the radar that sufficiently mitigates adversarial IRL. The radar minimizes its performance degradation (maximizes *Quality-of-service (QoS)*) due to sub-optimal responses, subject to a bound (6.18), (6.20) on the feasibility margin of the radar's strategy (maximizes *adversarial confusion*). Theorem 44 can be viewed as an *inverse IRL* (I-IRL) scheme that mitigates an IRL system and is a critical feature of a *meta-cognitive* radar that switches between

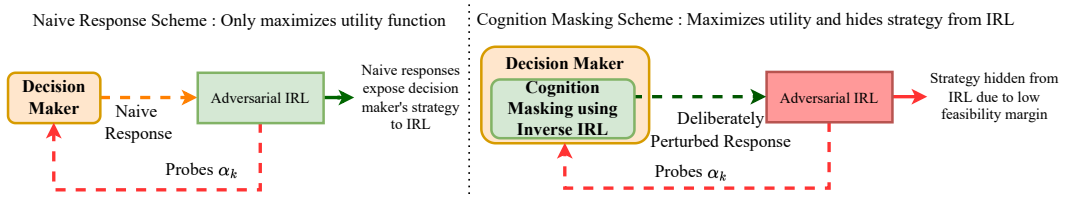

Figure 6.2: Schematic of the cognitive radar masking its strategy from adversarial IRL (via Theorem 44).
Naive response scheme (Left): The adversary sends a sequence of probe signals to the radar and records its responses to the adversary's probes. The radar's strategy passes the IRL feasibility test of Theorem 41 with a large margin if the radar transmits naive responses (6.3) and can be reconstructed by IRL.
Cognition masking scheme (Right): If the radar is aware of adversarial IRL, the radar deliberately perturbs its responses according to Theorem 44 to hide its strategy from the adversary at the cost of performance degradation. In Sec. 6.5, we illustrate via numerical examples how small deliberate perturbations in the radar's naive responses mask the radar's strategy from adversarial IRL to a large extent.

different plans. For completeness, Sec. 6.7.5 extends cognition masking to the case where the cognitive radar faces multiple constraints. Theorem 53 generalizes the cognition masking scheme of Theorem 44 to the multi-constraint case where the adversary uses Theorem 51 for optimal IRL. Also, Sec. 6.7.6 discusses cognition masking when the adversary has *mis-specified* measurements of the radar's responses. Our key result is Theorem 54 that provides a performance bound on the cognition masking scheme of Theorem 44 in terms of the misspecification error magnitude.

Extent of cognition-masking η in Theorem 44. A smaller value of η implies a larger extent of cognition masking from adversarial IRL and also a greater degradation in the radar's performance. One extreme case is setting $\eta = 0$. This results in maximal masking of the radar's strategy. That is, the IRL feasibility inequalities (6.6), (6.50) are no more feasible and there exists *no feasible strategy* that rationalizes the radar's responses. Setting $\eta = 0$ also causes the radar to deviate maximally from its naive responses (6.3), and hence results in a large performance degradation. The other extreme case is setting $\eta = 1$. In this case, the radar simply transmits its naive response (6.3) and

its strategy is not hidden from the adversary.

Algorithm 4 Masking Radar's Utility via Theorem 44 from IRL Feasibility Test (6.6)

Step 1. Compute radar's naive response sequence $\beta^*_{1:N}$ by solving the convex optimization problem (6.3):

$$\beta^*_k = \operatorname{argmin} \mathbf{u}(\beta),\ \mathbf{g}(\alpha_k, \beta) \leq 0,\ \beta \geq \mathbf{0}\ \forall k \in \{1, 2, \ldots, N\},$$

where $\mathbf{u}$ is concave monotone in β and $\mathbf{g}(\alpha_k, \beta)$ is convex monotone in β.
Step 2. Choose $\eta \in [0, 1]$ (extent of cognition masking from IRL feasibility test).
Step 3. Compute upper bound $\mathcal{M}_{\text{thresh}}$ on desired margin (6.15) after cognition masking: $\mathcal{M}_{\text{thresh}} = (1 - \eta)\ \mathcal{M}_{\mathbf{u}}(\{\alpha_k, \beta^*_k\}_{k=1}^N)$, where $\mathcal{M}_{\mathbf{u}}$ is defined in (6.15).
Step 4. Compute the cognition-making responses by solving the following optimization problem:

$$\begin{aligned} \{\tilde{\beta}^*_{1:N}\}_{\text{MASK-U}} = \operatorname{argmin}\ & \sum_{k=1}^{N} \mathbf{u}(\beta^*_k) - \mathbf{u}(\beta_k), \\ & \beta_k \geq \mathbf{0},\ \alpha_k' \beta_k \leq 1\ \forall k \in \{1, 2, \ldots, N\}, \\ & \mathcal{M}_{\mathbf{u}}(\{\alpha_k, \beta_k\}_{k=1}^N) \leq \mathcal{M}_{\text{thresh}}\,. \end{aligned} \tag{6.21}$$

Due to the non-linear margin constraint in (6.21), the optimization problem can be solved using a general purpose non-linear programming solver, for example, `fmincon` in MATLAB, to obtain a local minimum.

Let us briefly explain the essence of the cognition masking algorithm in Theorem 44 through our running cognitive radar example from [234]. We first assume the naive cognitive radar maximizes its QoS subject to constraints on its posterior Crámer-Rao bound (PCRB). The adversary exploits the ECM scheme of Theorem 41 to estimate the radar's QoS function and generates malicious probes (6.8). As an ECCM measure, the radar intentionally chooses a sub-optimal waveform that trades off between maximizing the radar's QoS (6.17) and ensuring a poor reconstruction of the radar's strategy by the adversary (margin constraint (6.18)). Let us consider the second scenario where the cognitive radar's utility is the inverse of the posterior Crámer-Rao bound (PCRB), that is, the radar minimizes its PCRB [234, Eq. 40] subject to a constraint on its communication cost with the central processing unit. The adversary can use Theorem 50 to estimate the radar's communication cost, and can then use (6.12) to generate malicious probes. As

an ECCM measure, the radar intentionally violates the communication cost constraint that trades off between minimizing the radar's transmission power (6.19) and ensuring a poor reconstruction of the radar's communication cost by the adversary (margin constraint (6.20)).

Summary

In this section, we introduced our key cognition masking result, namely, Theorem 44 that mitigates the ECM attempts of the adversary (Theorems 41 and 50) to estimate the radar's strategy (utility function/resource constraint). From a practitioner's perspective, we also related the cognition masking scheme to a formal model of a cognitive radar [234] that chooses its waveform by solving a constrained optimization problem. This section sets the stage to address cognition masking from an adversary under noisy measurements. In the remainder of the chapter, we motivate our cognition masking results using two radar functionalities, namely, optimal waveform adaptation and optimal beam allocation, instead of the cognitive radar model of [234].

6.4 How to Mask Cognition from Detector?

The framework considered in Theorem 44 was deterministic; we assumed that the adversary had accurate measurements of the radar's responses. In this section, we generalize Theorem 44 to the case where the adversary has *noisy* measurements of the radar's decisions. That is, the noise term ω_k in the radar's response measurement z_k in (6.1) of Model 1 is a non-zero random variable with pdf f_ω. If the adversary

deploys a Neyman-Pearson[11] type detector, how can we design our cognition masking strategy to spoof this detector so that the radar can hide its utility and constraints? Before generalizing Theorem 44 to the noisy case, we first address the following question: *How do the adversary's IRL algorithms, Theorems 41 and 50, adapt to noisy measurements?*

6.4.1 Noisy Adversarial IRL Detectors against Cognitive Radars

Our key IRL results for noisy radar measurements are outlined in Definition 45 below. Recall from Sec. 6.2 that the adversary's IRL algorithm in Theorem 41 comprises a linear feasibility test to identify a feasible strategy that rationalizes the radar's responses. When the adversary has noisy measurements of the radar's response, the deterministic feasibility test generalizes to a *feasibility hypothesis test* to detect the existence of feasible strategies (utilities and constraints) so that the radar responses satisfy utility maximization (6.3).

For our hypothesis tests below, let H_0 and H_1 denote the null and alternate hypotheses that the adversary's noise-less datasets defined in (6.5) and (6.10) pass, and not pass, respectively, the IRL feasibility tests (6.6) and (6.50), respectively.

$$\begin{aligned} H_0 &: \text{Radar is a constrained utility maximizer (6.3)} \\ H_1 &: \text{Radar is NOT a constrained utility maximizer (6.3)} \end{aligned} \tag{6.22}$$

The two types of error that arise in hypothesis testing are Type-I and Type-II errors. In the radar context, the Type-I and Type-II errors have the following interpretation:

$$\begin{aligned} &\textbf{Type-I:} \text{ Classify a cognitive radar as non-cognitive} \\ &\textbf{Type-II:} \text{ Classify a non-cognitive radar as cognitive} \end{aligned} \tag{6.23}$$

[11] By Neyman Pearson's lemma [308], it is impossible to maximize the Type-I and Type-II error of a detector simultaneously. In this chapter, we focus on mitigating the detector by maximizing its *conditional* Type-I error probability.

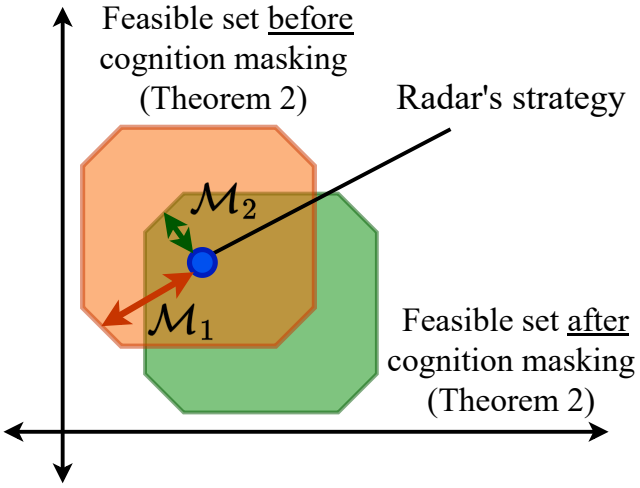

Figure 6.3: Cognition Masking for mitigating adversarial IRL. The radar's naive responses pass the IRL feasibility tests in Theorems 41 and 50 with a large feasibility margin $\mathcal{M}_1$. Cognition masking distorts the feasibility polytope so that the radar's true strategy is almost infeasible (low margin $\mathcal{M}_2$) wrt the IRL feasibility inequalities (close to the edge of feasibility polytope). Hence, the true strategy is a low-confidence estimate for IRL and successfully hidden from the adversary.

In analogy to Theorems 41 and 50, our IRL detectors defined below assume two scenarios, namely, Assumptions 4 and 5 that generalize Assumptions 2 and 3, respectively, to the case where the adversary has noisy response measurements.

Assumption 4 *Consider the radar-adversary interaction scenario of Assumption 2. The adversary has access to the noisy dataset $\widehat{\mathcal{D}}_g$ defined as:*

$$\widehat{\mathcal{D}}_g = \{\mathbf{g}(\alpha_k, \cdot), z_k\}_{k=1}^N,\ z_k = \beta_k + \omega_k,\ \omega_g \sim f_\omega \tag{6.24}$$

where $\mathbf{g}(\alpha_k, \cdot)$ is defined in (6.4)*, β_k is the radar's response and ω_k is the adversary sensor's measurement noise* (6.1) *with pdf f_ω known to the radar.*

IRL objective. *The adversary uses the IRL detector* (6.27) *in Definition 45 to detect if the noise-free dataset $\mathcal{D}_g$* (6.5) *passes the IRL test* (6.6) *of Theorem 41*

Assumption 5 *Consider the radar-adversary interaction scenario of Assumption 3. The*

adversary has access to the noisy dataset $\widehat{\mathcal{D}}_u$ *defined as:*

$$\widehat{\mathcal{D}}_u = \{\mathbf{u}(\alpha_k, \cdot), z_k\}_{k=1}^N,\ z_k = \beta_k + \omega_k,\ \omega_g \sim f_\omega \tag{6.25}$$

where β_k *is the radar's response,* ω_k *is the adversary sensor's measurement noise* (6.1) *with pdf* f_ω *known to the radar.*

IRL objective. *The adversary uses the IRL detector* (6.27) *in Definition 45 to detect if the noise-free dataset* $\mathcal{D}_u$ (6.10) *passes the IRL test* (6.50) *of Theorem 50*

Our IRL hypothesis tests for detecting radar's cognition (feasible utilities and resource constraints) for noisy radar response measurements are stated in Definition 45 below.

Definition 45 (IRL Detectors for Noisy Response Measurements) *Consider the cognitive radar* (6.3) *from Definition 40 and the radar-adversary interaction from Model 1.*

1. IRL for detecting feasible utilities. Suppose Assumption 4 holds. Then, the statistical test below detects if the radar's responses satisfy utility maximization behavior (6.3)*:*

$$\boxed{\mathbb{P}(\phi_u^*(\widehat{\mathcal{D}}_g) \leq L_g) \lesseqgtr_{H_0}^{H_1} \gamma}. \tag{6.26}$$

2. IRL for detecting feasible resource constraints. Suppose Assumption 5 holds. Then, the statistical test below detects if the radar's responses satisfy utility maximization behavior (6.3)*:*

$$\boxed{\mathbb{P}(\phi_g^*(\widehat{\mathcal{D}}_u) \leq L_u) \lesseqgtr_{H_0}^{H_1} \gamma}. \tag{6.27}$$

In the statistical tests (6.26) *and* (6.27)*:*

- $\gamma \in [0, 1]$ *is the 'significance level' of the test.*
- L_g *and* L_u *are random variables defined as:*

$$L_g \equiv \max_{s,k} \alpha_k'(\omega_k - \omega_s) \tag{6.28}$$

$$L_u \equiv \max_{s,k} \left(\mathbf{u}(\alpha_k, z_k) - \mathbf{u}(\alpha_k, z_s)\right) - \\ \left(\mathbf{u}(\alpha_k, z_k - \omega_k)\right) - \mathbf{u}(\alpha_k, z_s - \omega_s)), \tag{6.29}$$

where $\omega_k \sim f_\omega$ is the measurement noise in the adversary's measurement of the radar's response (6.1)*.*

- *The test statistics $\phi_g^*(\cdot)$ and $\phi_u^*(\cdot)$ are the minimum perturbations required for the noisy datasets $\widehat{\mathcal{D}}_g$ and $\widehat{\mathcal{D}}_u$, respectively, to pass the IRL feasibility tests* (6.6) *and* (6.50)*:*

$$\phi_u^*(\widehat{\mathcal{D}}_g) = \min_{\epsilon,\theta>0} \epsilon, \;\; \mathcal{A}(\theta, \widehat{\mathcal{D}}_g + \epsilon) \leq 0, \tag{6.30}$$

$$\phi_g^*(\widehat{\mathcal{D}}_u) = \max_{\epsilon,\theta>0} \epsilon, \;\; \mathcal{A}(\theta, \widehat{\mathcal{D}}_u - \epsilon) \geq 0, \tag{6.31}$$

Remarks. 1. The random variable L_g (6.28) bounds the perturbation needed for $\widehat{\mathcal{D}}_g$ to pass the IRL test (6.6), if H_0 holds:

$$H_0: \; \exists \theta > 0 \text{ s.t. } \mathcal{A}(\theta, \mathcal{D}_g) \leq 0 \implies \mathcal{A}(\theta, \widehat{\mathcal{D}}_g + L_g) \leq 0,$$

where $\mathcal{D}_g$ is the noise-free version of the noisy dataset $\widehat{\mathcal{D}}_g$. Similarly, the random variable L_u (6.29) bounds the perturbation needed for $\widehat{\mathcal{D}}_u$ to pass the IRL test (6.50), if H_0 holds:

$$H_0: \; \exists \theta > 0 \text{ s.t. } \mathcal{A}(\theta, \mathcal{D}_u) \geq 0 \implies \mathcal{A}(\theta, \widehat{\mathcal{D}}_u + L_u) \geq 0,$$

where $\mathcal{D}_u$ is the noise-free version of the noisy dataset $\widehat{\mathcal{D}}_u$.

2. The IRL detectors (6.26) and (6.27) classify the radar as a utility maximizer if the perturbation needed for the feasibility of the IRL inequalities lies under a particular threshold, and vice versa. Consider the statistical test of (6.26). Eq. 6.26 can be expressed differently as:

$$\phi_u^*(\widehat{\mathcal{D}}_g) \lessgtr_{H_1}^{H_0} F_{L_\alpha}^{-1}(1-\gamma), \tag{6.32}$$

where the RHS term in (6.32) is the test threshold for test statistic $\phi_u^*(\cdot)$. Intuitively, the larger the perturbation needed for the feasibility of the IRL inequalities, the less confidence the adversary has to classify the radar as a utility maximizer.

Computational Complexity of IRL Detectors. The constrained optimization problems (6.30) and (6.31) are non-convex since the RHS of the constraint is bilinear in the feasible variable. However, since the objective function depends only on a scalar, a 1-dim. line search algorithm can be used to solve for $\phi_u^*(\cdot)$ in (6.30) and $\phi_g^*(\cdot)$ in (6.31). That is, for any fixed value of ϵ, the constraints in (6.30), (6.31) specialize to a set of linear inequalities for which feasibility is straightforward to check.

We now discuss a key feature of the statistical tests (6.26) and (6.27) in Theorem 46 that bounds the Type-I error probability $\mathbb{P}(H_1|H_0)$ of the IRL detectors. Recall that the Type-I error probability is the probability of incorrectly classifying the radar as non-cognitive, when the radar's response is the solution of a constrained utility maximization problem (6.3).

Theorem 46 (Performance of IRL Detectors (Definition 45)) *Consider the statistical tests* (6.26) *and* (6.27) *in Definition 45. The Type-I error probability of the tests is bounded by the significance level of the tests* γ*:*

$$\mathbb{P}(H_1|H_0) \leq \gamma \textit{ for both detectors } (6.26) \textit{ and } (6.27). \tag{6.33}$$

The proof of Theorem 46 is in Sec. 6.7.4. The key idea in the proof is to show that, given that the null hypothesis H_0 holds, the random variables L_g and L_u dominate the test statistics $\phi_g^*(\widehat{\mathcal{D}}_u)$ and $\phi_u^*(\widehat{\mathcal{D}}_g)$, respectively. Since the IRL detectors have a bounded Type-I probability, our cognition masking rationale for the noisy case discussed below is to maximize their conditional Type-I error probability.

6.4.2 Masking Cognition from IRL Detectors

In the previous section, we generalize the IRL results of Theorem 41 and 50 in Sec. 6.2 to the case where the adversary has noisy measurements of the radar's responses. The key idea is that the IRL feasibility tests (6.6) and (6.50) generalize to IRL detectors (6.26) and (6.27) in Definition 45, respectively, that detect utility maximization behavior. This section addresses cognition masking when the adversary uses the IRL detectors of Definition 45: *How to mitigate the statistical tests of* (6.26) *and* (6.27) *and make utility maximization detection difficult?*

Intuition for hiding cognition from IRL Detectors. Suppose the radar follows the cognition masking scheme of Theorem 44 for the noisy case. Indeed, the radar's strategy is hidden from the IRL feasibility tests of Theorems 41 and 50, but does not affect the performance of the IRL detectors of Definition 45. To do so, the radar maximizes the *conditional Type-I error probability*[12] of the IRL detectors by deliberately deviates from its naive responses (6.3). The conditional Type-I error probability can be viewed as the noisy analog of the inverse of the feasibility margin in the noise-less case.

Definition 47 (Conditional Type-I error probability for IRL Detectors (Definition 45)) *Consider the datasets $\mathcal{D}_g$ and $\mathcal{D}_u$ defined in* (6.5) *and* (6.10)*, and their corresponding noisy versions $\widehat{\mathcal{D}}_g$ and $\widehat{\mathcal{D}}_u$ defined in* (6.24) *and* (6.25)*, respectively. Let $\phi_u(\widehat{\mathcal{D}}_g, \mathbf{u})$ and $\phi_g(\widehat{\mathcal{D}}_u, \mathbf{g})$ denote the minimum perturbations required for the tuples $(\widehat{\mathcal{D}}_g, \mathbf{u})$ and $(\widehat{\mathcal{D}}_u, \mathbf{g})$, respectively, to pass the IRL feasibility tests* (6.6)*,* (6.50)*:*

$$\begin{aligned} \phi_u^*(\widehat{\mathcal{D}}_g, \mathbf{u}) &= \min_{\varepsilon \geq 0} \varepsilon, \;\; \mathcal{A}(\mathbf{u}, \widehat{\mathcal{D}}_g + \varepsilon) \leq 0 \\ \phi_g^*(\widehat{\mathcal{D}}_u, \mathbf{g}) &= \min_{\varepsilon \geq 0} \varepsilon, \;\; \mathcal{A}(\mathbf{g}, \widehat{\mathcal{D}}_u - \varepsilon) \geq 0, \end{aligned} \tag{6.34}$$

[12]The radar can at best maximize the conditional Type-I error probability to mitigate the IRL detectors as the Type-I error probability is bounded by the detectors' significance level γ due to Theorem 46.

where $\mathbf{u}$ *and* $\mathbf{g}$ *are the radar's utility and resource constraint, respectively. Then:*

1. For IRL detector (6.26)*, the conditional Type-I error probability, conditioned on* $\widehat{\mathcal{D}}_g$ (6.24) *and radar's utility* $\mathbf{u}$ *is given by* $\mathbb{P}(H_1|\mathcal{D}_g, \mathbf{u})$*, and defined as:*

$$\mathbb{P}(H_1|\mathcal{D}_g, \mathbf{u}) = \mathbb{P}(\ \phi_u^*(\widehat{\mathcal{D}}_g, \mathbf{u}) > F_{L_g}^{-1}(1-\gamma)\ ) \tag{6.35}$$

2. For IRL detector (6.27)*, the conditional Type-I error probability conditioned on* $\widehat{\mathcal{D}}_u$ (6.25) *and radar's constraint* $\mathbf{g}$ *is given by* $\mathbb{P}(H_1|\widehat{\mathcal{D}}_u, \mathbf{g})$*, and defined as:*

$$\mathbb{P}(H_1|\mathcal{D}_u, \mathbf{g}) = \mathbb{P}(\ \phi_g^*(\widehat{\mathcal{D}}_u, \mathbf{g}) > F_{L_u}^{-1}(1-\gamma)\ ), \tag{6.36}$$

In (6.35)*,* (6.36)*, the alternate hypothesis event* H_1 *is expressed differently in the equivalent representation form of* (6.32)*, and the random variables* $L_u,\ L_g$ *are defined in* (6.28) *and* (6.29)*.*

Remarks.

1. The test statistics of the IRL detectors defined in (6.35) and (6.36) are computed via an optimization over $\mathbb{R}_+^{2N+1}$, whereas the optimization in (6.36) is over $\mathbb{R}_+$. Hence, $\phi_u^*(\widehat{\mathcal{D}}_g, \mathbf{u})$ and $\phi_g^*(\widehat{\mathcal{D}}_u, \mathbf{g})$ (6.36) are cheaper to compute than the test statistics $\phi_u^*(\widehat{\mathcal{D}}_g)$ (6.35) and $\phi_g^*(\widehat{\mathcal{D}}_u)$ (6.36), respectively.

2. The IRL detector performance is already constrained due to Theorem 46 (bounded Type-I error probability). Hence, to mitigate the IRL detector, the best the radar can do is to maximize its conditional Type-I error probability using the statistics defined in (6.34).

We are now ready to state our cognition masking result, Theorem 48, that mitigates IRL detectors (Definition 45). In analogy to Theorem 44 for mitigating the IRL feasibility tests of Theorems 41 and 50, the radar deliberately degrades it performance to maximize the IRL detectors' conditional Type-I error probability defined in (6.35) and (6.36).

Theorem 48 (Masking Cognition from Adversarial IRL Detectors) *Consider the cognitive radar* (6.3) *from Definition 40. Let* $\{\beta_k^*\}_{k=1}^N$ *denote the naive response sequence* (6.3) *that maximizes the cognitive radar's utility. Then:*

1. Masking Utility Function from Detector. Suppose Assumption 4 holds. Then, the response sequence defined below makes cognition detection difficult by ensuring the detector (6.26) *has a sufficiently large conditional Type-I error probability:*

$$\{\tilde{\beta}_{1:N}^*\} = \underset{\{\beta_k \geq \mathbf{0},\ \alpha_k' \beta_k \leq 1\}}{\operatorname{argmin}} \sum_{k=1}^{N} \mathbf{u}(\beta_k^*) - \mathbf{u}(\beta_k) - \lambda\, \mathbb{P}(H_1 | \mathcal{D}_g, \mathbf{u}) \tag{6.37}$$

2. Masking Resource Constraint from Detector. Suppose Assumption 5 holds. Then, the response sequence below makes cognition detection difficult by ensuring the detector (6.27) *has a sufficiently large conditional Type-I error probability:*

$$\{\tilde{\beta}_{1:N}^*\} = \underset{\{\beta_k \geq \mathbf{0},\ \mathbf{g}(\beta_k) \leq \gamma_k\}}{\operatorname{argmin}} \sum_{k=1}^{N} \mathbf{u}(\beta_k^*) - \mathbf{u}(\beta_k) - \lambda\, \mathbb{P}(H_1 | \mathcal{D}_u, \mathbf{g}) \tag{6.38}$$

In (6.37)*,* (6.38)*, the positive scalar* λ *parametrizes the extent of mitigation of the IRL detector.*

Theorem 48 is our second result for cognition masking; see Algorithm 5 for a step-wise procedure for masking cognition in noise (6.37) when the adversary knows the radar's constraints. Eq. 6.37 and 6.38 compute the *optimal* sub-optimal radar response that sufficiently hides the radar's cognition from being detected by the IRL hypothesis tests of Definition 45. The parameter λ in Theorem 48 is analogous to parameter η in Theorem 44. A larger value of λ (6.37) results in a larger conditional Type-I error probability for the IRL detector (larger adversarial confusion) while increasing the radar's deviation from its optimal response (greater performance degradation), and vice versa.

The optimization problems (6.37) and (6.38) can be solved by stochastic gradient algorithms. Algorithm 5 outlines a constrained simultaneous perturbation stochastic approximation (SPSA) [309, 310] implementation for computing the cognition masking

scheme of Theorem 48. The objective function is non-convex in the radar's responses, hence SPSA converges to a local optimum. SPSA is a generalization of adaptive algorithms where the gradient computation in (6.37) requires only two empirical estimates of the objective function per iteration, i.e. , the number of evaluations is independent of the dimension of the radar's response. For decreasing step size $\eta = 1/i$ (6.42), the SPSA algorithm converges with probability one to a local stationary point. For constant step size η, it converges weakly (in probability).

Summary: In this section, we generalized our cognition masking results of Theorem 44 to the case where the adversary has noisy measurements of the radar's responses. We first generalized our adversarial IRL feasibility tests of Theorems 41, 50 to IRL hypothesis tests (6.26) and (6.27) in Definition 45 to detect utility maximization behavior given noisy radar response measurements. We then present Theorem 48 that masks the radar's cognition by making utility maximization detection erroneous by maximizing the conditional Type-I error probability of the IRL detectors via purposeful sub-optimal responses. Our cognition masking results can be extended WLOG to any sub-optimal IRL algorithm as discussed in Sec. 6.7.7.

6.5 Numerical Results for I-IRL

In this section, we illustrate how our cognition masking results of Theorems 44 and 48 successfully confuse adversarial IRL via the two radar tracking functionalities, namely, waveform adaptation and beam allocation discussed in Sec. 6.2.

6.5.1 Cognition Masking via Theorem 44 for Noise-less Adversary Measurements

Consider the scenario where the adversary has accurate measurements of the radar's responses. Recall from Sec. 6.2.4 that the adversary knows the radar's constraints for waveform adaptation and the radar's utility function for beam allocation. For waveform adaptation, the probe signal parametrizes the state covariance matrix of the radar's Kalman filter due to the adversary's maneuvers, and the response signal parametrizes the sensory accuracy chosen by the radar. Recall that the probe signal α_k is the diagonal of the state noise covariance matrix: $Q_k = \text{diag}[\alpha_k(1), \alpha_k(2)]$. For beam allocation, the i^{th} component of the probe signal $\alpha_k(i)$ is the asymptotic predicted precision of the radar tracker for target i. The probe α_k parametrizes the radar's Cobb-Douglas utility for beam allocation. Our simulation parameters for our numerical experiments are listed below in Table 6.2.

Parameters for Numerical Experiments

Table 6.2: Parameters for Numerical Experiments

Masking Smart Waveform Adaptation	
Time Horizon	$N = 50$
Probe/response dimension	$m = 4$
Probe	$\alpha_k(i) \overset{\text{i.i.d}}{\sim} \text{Unif}(0.2, 2.5),\ i = 1, 2, \ldots, m$
Utility function	$\mathbf{u}_1(\beta) = \sum_{i=1}^{m} \sqrt{\beta(i)}, \mathbf{u}_2(\beta) = \sum_{i=1}^{m} \beta(i)^2$
Resource constraint	$\mathbf{g}(\alpha_k, \beta) = \alpha_k' \beta - 1$
Masking Smart Beam Allocation	
Time Horizon	$N = 50$
Probe/response dimension	$m = 4$
Probe	$\alpha_k(i) \overset{\text{i.i.d}}{\sim} \text{Unif}(0.1, 0.7),\ i = 1, 2, \ldots, m$
Utility function	$\mathbf{u}(\alpha_k, \beta) = \prod_{i=1,2,\ldots,m} \beta(i)^{\alpha_k(i)}$
Resource constraint	$\mathbf{g}(\alpha_k, \beta) = \|\beta\|_\kappa - \gamma_k$, $\gamma_k \overset{\text{i.i.d}}{\sim} \text{Unif}(0.5, 2)$.

In Table 6.2, $\text{Unif}(a, b)$ denotes uniform pdf with support (a, b). The elements of the probes α_k (6.3), and intensity thresholds γ_k (6.14) are generated randomly and independently over time $k = 1, 2, \ldots, N$. For waveform adaptation, we conduct our numerical experiments for two distinct utility functions $\mathbf{u}_1$ and $\mathbf{u}_2$. Given the probe sequence $\{\alpha_k\}_{k=1}^N$, the cognitive radar chooses its smart response sequence via (6.17) for masking optimal waveform adaptation, and via (6.19) for masking optimal beam allocation. Recall from Sec. 6.2 that response β_k is the diagonal of the inverse of radar's observation noise covariance matrix for waveform adaptation. For beam allocation, $\beta_k(i)$ is the beam intensity directed towards target i at time k.

Figures 6.4 and 6.5 show the performance loss (minimum perturbation from optimal response computed via (6.17), (6.19) in Theorem 44) of the cognitive radar as a function of η (extent of cognition masking) when the cognitive radar performs waveform adaptation and beam allocation, respectively. We see that for both functionalities, both the

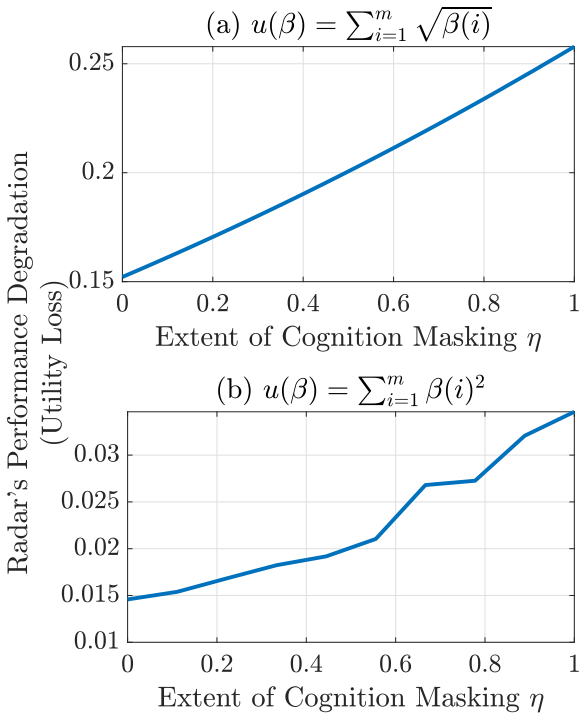

Figure 6.4: Masking Waveform Adaptation Strategy from Adversarial IRL. Small deliberate performance loss (vertical axis) of the cognitive radar results in large performance mitigation of the adversary (horizontal axis). The figure illustrates a cognitive radar operating with two distinct utility functions.
(i) $\eta = 1$ corresponds to maximum cognition masking and hence results in maximum performance loss. (ii) For a fixed value of η, the quadratic utility (sub-figure (b)) requires smaller perturbation (≈ 10 times) from the optimal response compared to the sub-linear utility of sub-figure (a).

radar's performance loss and adversarial IRL mitigation increase with η. This is expected since larger η implies a larger shift of the set of feasible strategies computed via IRL to ensure the radar's strategy is sufficiently close to the edge of the feasible set, at the cost of greater deviation from the radar's optimal strategy.

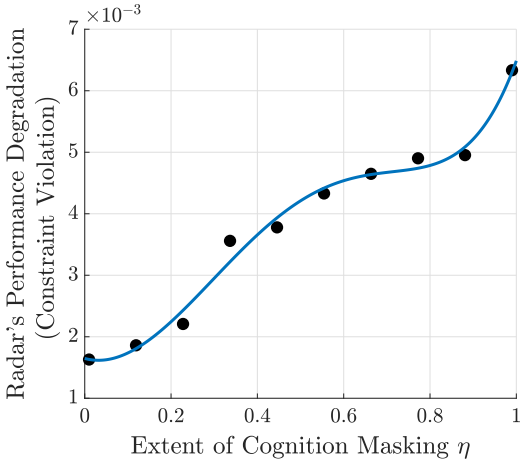

Figure 6.5: Masking Beam Allocation Strategy from Adversarial IRL: Small deliberate utility loss of the radar (vertical axis) results in large performance loss (extent of strategy masking η) of the adversarial IRL algorithm (horizontal axis). $\eta = 0$ corresponds to zero strategy masking, and $\eta = 1$ corresponds to complete strategy masking by the decision maker. As expected, the deliberate utility loss of the radar increases with η.

6.5.2 Cognition Masking via Theorem 48 for Noisy Adversary Measurements

We now consider the scenario where the adversary has noisy measurements of the radar's response. Consider the simulation parameters of Table 6.2. For our second set of numerical experiments for both waveform adaptation and beam allocation, we set the noise pdf f_ω (6.1) to $\mathcal{N}(0, 0.3I)$, where $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate normal distribution with mean μ and covariance Σ, and I denotes the identity matrix. In Theorem 48.

For the noisy case, we consider only a single utility function for waveform adaption, namely, $\mathbf{u}(\beta) = \sum_{i=1}^{m} \sqrt{\beta(i)}$. We performed our numerical experiments for three values of $\gamma = \{0.05, 0.1, 0.2\}$ for both waveform adaptation and beam allocation. Recall from Sec. 6.4 that γ is the significance level of the adversary's IRL detectors (6.26) and (6.27) in Definition 45.

Given the probe sequence $\{\alpha_k\}_{k=1}^{N}$, we generated the cognition masking response se-

quence via (6.37) for waveform adaption and (6.38) for beam allocation by varying the parameter λ (6.37) over the interval $[10^0, 10^5]$. Recall from Theorem 48 that the radar minimizes the detectors' conditional Type-I error probability (6.35), (6.36) to mitigate adversarial IRL while deliberately compromising on its performance (utility).

Our SPSA algorithm [309, 310] (Algorithm 5) for stochastic gradient descent was executed over 10^4 iterations for all pairs of (λ, γ), $\lambda \in \{10^0, 10^1, 10^2, 10^3, 10^4, 10^5\}$ and $\gamma \in \{0.05, 0.1, 0.2\}$. Figure 6.6 shows the conditional Type-I error probability (adversarial IRL mitigation) of the detector and performance loss of the radar as the parameter λ is varied for three different values of the significance level α of the adversary's detector. Recall from Theorem 48 that the parameter λ controls the extent of cognition masking for noisy inverse IRL. From Fig. 6.6, we see that both the conditional Type-I error probability of the IRL detectors and radar's performance loss increase with λ as well as γ.

If $\lambda = 0$, the radar simply transmits its naive response that maximizes its utility (no performance loss) and also results in zero adversarial mitigation. For the limiting case of $\lambda \to \infty$, the radar's cognition masking response computed via Theorem 48 degenerates to a constant for all time k, hence maximizing the conditional Type-I error probability of the detector at the cost of maximal performance loss for the radar.

Let us briefly discuss the variation of the radar performance and adversarial mitigation as the parameter γ is varied. γ (6.26) can be viewed as the risk-aversion tendency of the adversary's IRL system since it bounds the detector's Type-I error probability. Recall from (6.22) that the Type-I error is the probability of detecting a cognitive radar as non-cognitive. Higher γ implies the detector is *risk-seeking* and a lower γ implies the detector is *risk-averse*. Naturally, a larger deviation from optimal strategy is required to mitigate a risk-averse detector compared to a risk-seeking detector to the same extent.

Algorithm 5 SPSA for Mitigating Utility Maximization Detection for Adversarial IRL Detector (6.26) ((6.37) in Theorem 48)

Step 1. Set $\boldsymbol{\beta}_0 = \{\beta^*_{1:N}\}$, the naive response sequence (6.13) that maximizes the radar's utility (6.3).
Step 2. Choose $\lambda > 0$ (extent of cognition masking).
Step 3. For iterations $n = 0, 1, 2, \ldots$,
(i) Compute $\widehat{\mathbb{P}}(H_1|\{\alpha_k\}_{k=1}^N, \boldsymbol{\beta}_i, \mathbf{u})$, the empirical probability estimate of the conditional Type-I error probability of the detector (6.26) defined in (6.35) using $R \times N$ fixed realizations $\{\omega_{r,k}\}_{r,k=1}^{R,N}$ of adversary's measurement noise $\omega_k \sim f_\omega$ (6.1):

$$\frac{1}{R}\sum_{r=1}^{R} \mathbb{1}\left\{\phi^*_u(\{\alpha_k, \beta_{n,k} + \omega_{r,k}\}_{k=1}^N, \mathbf{u}) > F^{-1}_{L_g}(1-\gamma)\right\} \tag{6.39}$$

In (6.39):
• $\boldsymbol{\beta}_i \equiv \{\beta_{i,1:N}\} \geq \mathbf{0}$ is a vector of responses
• $\mathbb{1}\{\cdot\}$ denotes the indicator function
• R controls the accuracy of the empirical probability estimate
• $F_{L_g}(\cdot)$ is the distribution function of the r.v. L_g (6.26)
• The statistic $\phi^*_u(\cdot, \mathbf{u})$ is defined in (6.34).

Let $J(\boldsymbol{\beta}_i)$ denote the objective being maximized in (6.37):

$$J(\boldsymbol{\beta}_i) = \sum_{k=1}^{N} \mathbf{u}(\beta_{i,k}) - \mathbf{u}(\beta_{i,k}) - \lambda \mathbb{P}(H_1|\{\alpha_k\}_{k=1}^N, \boldsymbol{\beta}_i, \mathbf{u}) \tag{6.40}$$

Then: (ii) Compute empirical estimate $\widehat{J}(\boldsymbol{\beta}_i)$ of objective $J(\boldsymbol{\beta}_i)$ (6.40):

$$\widehat{J}(\boldsymbol{\beta}_n) = \sum_{k=1}^{N} \mathbf{u}(\beta^*_k) - \mathbf{u}(\beta_{n,k}) - \lambda\, \widehat{\mathbb{P}}(H_1|\{\alpha_k\}_{k=1}^N, \boldsymbol{\beta}_i, \mathbf{u}), \tag{6.41}$$

where $\widehat{\mathbb{P}}(H_1|\{\alpha_k\}_{k=1}^N, \boldsymbol{\beta}_i, \mathbf{u})$ is computed in (6.39).
(ii) Compute the estimate of the gradient $\nabla_{\boldsymbol{\beta}}\, J(\boldsymbol{\beta}_n)$ as:

$$\widehat{\nabla}_{\boldsymbol{\beta}}(\widehat{J}(\boldsymbol{\beta}_n)) = \frac{\Delta_n}{\omega \|\Delta_n\|_F^2}\, \widehat{J}(\boldsymbol{\beta}_n + \delta\, \Delta_n) - \widehat{J}(\boldsymbol{\beta}_n - \delta\, \Delta_n),$$

where δ is the gradient step size, $||\cdot||_2$ denotes the Frobenius norm, and $\Delta_n \in \{-1, +1\}^{m \times N}$ is a random perturbation vector whose each element is ± 1 with probability $1/2$.

(iii) Update the radar's response as:

$$\boldsymbol{\beta}_{n+1} = \mathrm{Proj}_{S_\alpha}\left(\boldsymbol{\beta}_n + \eta\, \frac{\Delta_n}{\|\Delta_n\|_F} \widehat{\nabla}_{\boldsymbol{\beta}} \widehat{J}(\boldsymbol{\beta}_n)\right), \tag{6.42}$$

where η is the response update step size and Proj_{S_α} is the projection operator to the hyperplane $S_\alpha = \{\beta_{1:N} : \alpha_k' \beta_k = 1, \beta_k \geq \mathbf{0}\}$

Step 4. Set $n \leftarrow n + 1$ and go to Step 3.

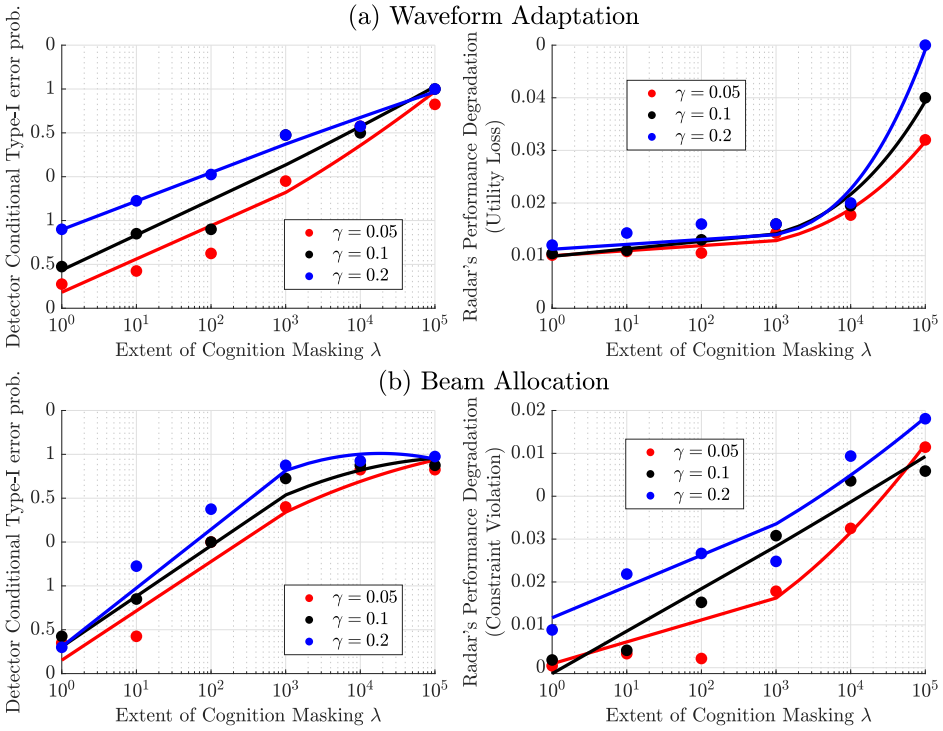

Figure 6.6: Masking Cognition from IRL Detectors. Performance of meta-cognitive radar for waveform adaptation (sub-figure (a)) and beam allocation (sub-figure (b)) when the adversary deploys an IRL detector (6.26), (6.27) for cognition detection. The key takeaway is that a small sacrifice in performance of the radar results in large performance loss of adversary's IRL detector. The performance loss of both the radar and the adversary due to meta-cognition increase with scaling factor λ (6.37) and significance level γ of the adversary's IRL detectors (6.26), (6.27).

6.6 Conclusion and Extensions

This chapter investigated how a cognitive radar can hide its cognition from an adversary, when the adversary performs inverse reinforcement learning (IRL) to estimate the radar's utility function by observing its actions. The adversary's IRL estimate of the radar's strategy is a polytope of feasible solutions to a set of convex inequalities. Our first cognition masking result is Theorem 44. When the adversary has accurate measurements of the radar's response, cognition masking via Theorem 44 ensures the radar's true strategy lies close to the edge of the feasibility polytope computed via adversarial IRL

(true strategy poorly rationalizes adversary's dataset). When the adversary has noisy measurements of the radar's response, adversarial IRL generalizes to a cognition detector defined in Definition 45. Our second cognition masking result is Theorem 48. The key idea is to maximize the probability of the radar being classified as non-cognitive by the detector subject to a bound on the radar's performance loss. Finally, in Sec. 6.5, we illustrate our cognition masking results on a cognitive radar that performs waveform adaptation and beam allocation for target tracking. We show that small purposeful deviations from the optimal strategy of the radar suffice to significantly confuse the adversarial IRL system.

This chapter builds significantly on our previous work [103] on ECM for identifying cognitive radars, and [173, 311, 312] on ECCM for masking radar cognition. Theorem 51 extends IRL for cognitive radars [103] when the radar faces multiple resource constraints. The linear IRL feasibility test for a single constraint case generalizes to a mixed integer feasibility test. Theorem 53 generalizes the cognition masking result of [144] to multiple constraints. Our previous works [144, 311, 312] assume optimal adversarial IRL via Afriat's theorem. This chapter generalizes cognition masking to sub-optimal adversarial IRL algorithms. Algorithm 6 outlines a cognition scheme when the adversary uses an arbitrary IRL algorithm to estimate the radar's strategy. Theorem 54 provides performance bounds for our cognition masking scheme when the adversary has misspecified measurements of the radar's response. Although this chapter is radar-centric, we emphasize that the problem formulation and algorithms developed also apply to adversarial inverse reinforcement learning in general machine learning applications.

Finally, a useful extension of this chapter would be to study cognition masking in a dynamic radar-adversary interaction environment in comparison to the batch-wise probe-response exchange considered in this chapter. Also, how to mask cognition when the adversary knows of the radar's ECCM capability? Such an approach warrants a game-

theoretic discussion in terms of a Stackelberg game where the adversary moves first and the radar responds to the adversary's probes. It is also worthwhile exploring state-of-the-art concepts in chance constrained optimization [313] and robust optimization [314, 315] to achieve cognition masking under uncertainty - when the radar has noisy measurements of the adversary's probes.

6.7 Appendix

The appendix comprises auxiliary results and lemmas that complement the main cognition masking results presented in the main text.

6.7.1 Feasibility Test for Adversarial IRL

Definition 49 (IRL Feasibility Test) *Consider a dataset $\mathcal{D} = \{g_k(\cdot), \beta_k\}_{k=1}^N$ of monotone functions g_k and responses $\beta_k \geq \mathbf{0}$. The set of IRL feasibility inequalities $\mathcal{A}(\cdot, \mathcal{D})$ is defined as:*

$$\mathcal{A}(\theta, \mathcal{D}) \leq \mathbf{0} \tag{6.43}$$

$$\equiv \theta_s - \theta_k - \theta_{k+N}(g_k(\beta_s) - g_k(\beta_k)) \leq 0, \text{ for all } s, k \in \{1, 2, \ldots, N\},\ s \neq k, \tag{6.44}$$

where the feasible variable $\theta \in \mathbb{R}_+^{2N}$.

Remarks.

1. The IRL feasibility inequalities are linear in the feasible variable θ. Hence, $\mathcal{A}(\cdot, \mathcal{D})$ is a linear feasibility test whose feasibility can be checked using a linear programming solver.

2. The set of inequalities $\mathcal{A}(\cdot)$ checks for relative optimality [316] between any pair of indices $s, k \in \{1, 2, \ldots, N\}$. Consider the abstract setup of the cognitive radar in

Definition 40 and suppose Assumption 2 holds. In this context, we define relative optimality as:

$$\begin{aligned} &\text{If } \alpha_k'\beta_s \leq \alpha_k'\beta_k, \text{ then } u(\beta_k) \geq u(\beta_s), \\ &\text{for all } s,k \in \{1,2,\ldots,N\},\ s \neq k. \end{aligned} \tag{6.45}$$

Eq. 6.45 states that β_k is the optimal response choice from the *finite* set $\{\beta_k\}_{k=1}^N$ and hence a weaker notion of optimality wrt (6.3). Hence, if Assumption 2 holds, we say wrt utility $u(\cdot)$, the dataset $\{\alpha_k'(\cdot) - 1, \beta_k\}_{k=1}^N$ satisfies relative optimality. The feasibility of the IRL inequalities (6.43) is equivalent (due to [127]) to the existence of a utility function such that relative optimality holds.

3. Finally, let us provide some intuition on the feasible variable θ in (6.43). Suppose $\mathcal{A}(\theta, \mathcal{D}) \leq \mathbf{0}$ (6.43). On inspecting (6.44) in more detail, the first N components $\theta(1),\ \theta(2), \ldots,\ \theta(N)$ of the vector θ can be viewed as the utility function u_θ corresponding to the feasible variable θ evaluated at responses $\beta_1, \beta_2, \ldots, \beta_N$. The last N components $\theta(N+1),\ \theta(N+2), \ldots,\ \theta(2N)$ can be viewed as the Lagrange multipliers associated with the Lagrangian of the optimization problem $\max_{\beta \geq \mathbf{0}}\ u(\beta),\ g_k(\beta) \leq 0$ over all k. In summary, we have the following correspondence between the feasible variable θ and the feasible utility function u_θ:

$$\begin{aligned} &\theta(k) = u_\theta(\beta_k), \\ &\theta(N+k) = \frac{\nabla_\beta\ u_\theta(\beta)|_{\beta=\beta_k}}{\nabla_\beta\ g_k(\beta)|_{\beta=\beta_k}},\ k \in \{1,2,\ldots,N\}, \end{aligned} \tag{6.46}$$

where the division operation in (6.46) is element-wise.

4. The interpretation of the feasibility vector θ in (6.46) facilitates us to define the mapping $\theta \to u_\theta(\cdot)$ as:

$$u_\theta(\beta) = \min_{k \in \{1,2,\ldots,N\}} \left\{ \theta(k) + \theta(N+k)(g_k(\beta) - g_k(\beta_k)) \right\} \tag{6.47}$$

It is straightforward to show that both relative optimality (6.45) and absolute optimality (6.3) hold for u_θ (6.47) if $\mathcal{A}(\theta, \mathcal{D}) \leq \mathbf{0}$. Also, for any utility function u and dataset

$\mathcal{D}$ (6.43), define the reverse mapping $u \to u_{\mathcal{A}}$ as:

$$\begin{aligned} u_{\mathcal{A}}(k) &= u(\beta_k), \\ u_{\mathcal{A}}(N+k) &= \frac{\nabla_\beta \, u(\beta)|_{\beta=\beta_k}}{\nabla_\beta \, g_k(\beta)|_{\beta=\beta_k}}, \; k \in \{1, 2, \ldots, N\}. \end{aligned} \tag{6.48}$$

Then, it is straightforward to show:

$$\beta_k \in \underset{\beta \geq \mathbf{0}}{\operatorname{argmax}} \; u(\beta), \; g_k(\beta) \leq 0 \implies \mathcal{A}(u_{\mathcal{A}}, \mathcal{D}) \leq \mathbf{0}, \tag{6.49}$$

where $u_{\mathcal{A}}$ is defined in (6.48) and can be interpreted as the finite dimensional projection of $u(\cdot)$ for the IRL feasibility test $\mathcal{A}(\cdot, \mathcal{D}) \leq \mathbf{0}$.

6.7.2 IRL for Identifying Radar's Resource Constraint

Theorem 50 (IRL for Identifying Resource Constraint) *Suppose Assumption 3 holds. With $\mathcal{D}_u = \{u_k(\cdot), \beta_k\}_{k=1}^N$ denoting the utility-response dataset, the adversary can identify if there exists feasible constraints that satisfies* (6.3) *for all k by checking the feasibility of the linear inequalities $\mathcal{A}(\cdot, \mathcal{D}_u) \geq \mathbf{0}$* (6.43)*:*

$$\begin{aligned} &\exists \theta \in \mathbb{R}_+^{2N} \textit{ s.t. } \mathcal{A}(\theta, \mathcal{D}_u) \geq \mathbf{0}, \\ \iff &\exists g, \gamma_k \textit{ s.t. } \beta_k \in \operatorname{argmax} u_k(\beta), \; g(\beta) \leq \gamma_k \forall k \end{aligned} \tag{6.50}$$

The set-valued IRL estimate of the constraint g is given by:

$$\begin{aligned} g_{\mathrm{IRL}}(\beta) &\equiv \{g_{\mathrm{IRL}}(\beta; \theta) : \mathcal{A}(\theta, \mathcal{D}_u) \geq \mathbf{0})\} \\ g_{\mathrm{IRL}}(\beta; \theta) &= \max_{k \in \{1,2,\ldots,N\}} \{\theta_k + \theta_{k+N} \, (u_k(\beta) - u_k(\beta_k))\}, \end{aligned} \tag{6.51}$$

The proof of Theorem 50 follows from [317, Theorem 3]. Intuitively, Afriat's theorem (Theorem 41) reconstructs a monotone concave utility function. In complete analogy, Theorem 50 reconstructs a monotone *convex* budget constraint for the radar, hence the change in sign in the feasibility test. Although Theorem 50 appears to be a dual statement of Theorem 41 the proof does not use duality.

6.7.3 Example. Optimal Beam Allocation

For abstracting beam allocation into a constrained utility maximization setup (6.3), we work at a higher level of abstraction compared to waveform adaptation. Specifically, we assume the adversary comprises multiple targets. At this higher level of abstraction, we view each component i of the adversarial probe signal $\alpha_k(i)$ as the trace of the predicted precision matrix (inverse covariance) of target i. Recall from the previous section that we used the probe signal (6.13) to parametrize the maneuver covariance matrix. In comparison, we now use the trace of the precision of each target in our probe signal – this allows us to consider multiple targets.

Multiple Target-tracking. For the optimal beam allocation example, we assume the adversary comprises a *collection* of m adversarial targets indexed by $i = 1, 2, \ldots, m$. We assume the cognitive radar adaptively switches its beam between the m targets. As in (6.2), on the fast time scale indexed by k, target i has linear Gaussian dynamics and the radar obtains linear Gaussian measurements of the targets' maneuvers:

$$\begin{aligned} x^i_{n+1} &= Ax^i_n + w^i_n, \quad x_0 \sim \pi_0 \\ y^i_n &= C\, x^i_n + v^i_n, \quad i = 1, 2, \ldots, m \end{aligned} \tag{6.52}$$

Here $w^i_n \sim \mathcal{N}(0, Q_k(i))$, $v^i_n \sim \mathcal{N}(0, R_k(i))$. We assume that both $Q_k(i)$ and $R_k(i)$ are known to the radar and adversary. As in the previous sub-section, k indexes the slow time scale and n indexes the fast time scale. The enemy's radar tracks our m targets using m Kalman filter trackers.

Probe-Response Parametrization. The i^{th} target's predicted *precision* parametrizes the i^{th} element of the adversary's probe. Specifically, the price the radar pays at the start of epoch k for tracking target i is the trace of the inverse of the predicted covariance at

epoch k using the Kalman predictor:

$$\alpha_k(i) = \mathrm{tr}(\Sigma_k^{-1}(i)), \quad i = 1, \ldots, m,$$
$$\Sigma_k(i) = \lim_{n\to\infty} A^n \Sigma_{0,t}(i) A'^n + \sum_{l=0}^{n-1} A^l Q_k(i) A'^l, \tag{6.53}$$

where $\Sigma_{0,t}(i)$ is the covariance of the i^{th} target's covariance at time (fast time scale) $n = 0$ in epoch t. Clearly, the predicted covariance $\Sigma_k(i)$ (6.53) is a deterministic function of the maneuver covariance $Q_k(i)$ of target i. The radar's response $\beta_k(i)$ to the adversary's probe is the beam intensity allocated to target i in epoch k. Intuitively, at the start of every epoch k, the radar computes the predicted precision of the state estimate of target i, and then chooses the beam intensity towards target i during epoch k. The radar can at best compute the *predicted* precision for its decision (since it has no access to observations $y_1, y_2, \ldots$ at the beginning of the epoch).

Optimal beam allocation. We assume that the radar, at epoch k, faces a resource constraint $\mathbf{g}(\beta) \leq \gamma_k$ and directs beam intensities $\beta(1), \beta(2), \ldots, \beta(m)$ towards targets $1, 2, \ldots, m$, respectively. We assume the radar's resource constraint can be expressed as $\mathrm{g} = \|\beta\|_k$, the k-norm of the radar's response. The radar's aim is to maximize the Cobb-Douglas utility of the transmitted intensities:

$$\mathbf{u}_k(\beta) = \prod_{i=1,\ldots,m} \beta(i)^{\alpha_k(i)}. \tag{6.54}$$

The exponents for the Cobb-Douglas utility function (6.54), referred to as elasticity parameters in consumer economics literature, parameterize the marginal utility per consumer good. In complete analogy, the elasticity parameters in our cognitive radar context parameterize the incentive for the radar to focus its beam towards a particular target. The economic rationale for the cognitive radar is as follows - a higher predicted precision $\alpha_k(i)$ for target i implies a *better* state estimate, and thus a higher incentive for the radar to direct its transmission intensity towards target i. To summarize, the cognitive radar's beam allocation functionality can be abstracted as:

$$\boxed{\beta_k = \underset{\beta}{\operatorname{argmax}}\ \mathbf{u}_k(\beta),\ \mathbf{g}(\beta) \leq \gamma_k,} \tag{6.55}$$

$$\mathbf{u}_k = \prod_{i=1,\ldots,m} \beta(i)^{\alpha_k(i)}, \mathbf{g}(\beta) = \|\beta\|_\kappa,\ \kappa > 1$$

Eq. 6.55 abstracts the beam allocation functionality of the cognitive radar; at time k the radar maximizes a Cobb-Douglas[13] utility $\mathbf{u}_k(\beta)$ (utility specified by the adversarial target) subject to a constraint on the k-norm of its response (beam intensity allocation). In the beam allocation case, the aim of adversarial IRL is to estimate the scalar κ that parametrizes the radar's budget constraint $\mathbf{g}$.

In (6.55), notice how the adversary's probe parametrizes the radar's utility function instead of its budget constraint. Also, observe that both the utility function and cost are monotone in the transmission intensities - higher beam intensity yields a larger utility but is also more costly. We assume that: (1) each target i is equipped with a radar detector and can estimate the beam intensity $\beta_k(i)$ the enemy's radar directs towards the target - this assumption facilitates the targets to carry out adversarial IRL attacks on the radar, and (2) the adversarial targets know the radar is maximizing the Cobb-Douglas utility function (6.54), but does not know the radar's budget constraint $\mathbf{g}(\beta) \leq \gamma_k$ and is the adversary's IRL objective in the beam allocation context as discussed below.

IRL for optimal beam allocation. We now present Theorem 41, a revealed preference-based IRL algorithm for the adversary. Unlike Theorem 41, the adversary parametrizes (and hence, knows) the utility function of the cognitive radar, but does not know its budget constraint. Hence, the IRL objective of the adversarial target in Theorem 41 is to estimate the radar's budget $\mathbf{g}(\cdot)$.

[13]The Cobb-Douglas utility function is widely used in microeconomics to model human satisfaction from buying consumer goods. In the radar context, this utility function measures the performance of the cognitive radar.

6.7.4 Proof of Theorem 46

Proof The Type-I error probability is given by $\mathbb{P}(H_1|H_0)$, that is, probability that the adversary incorrectly classifies a utility maximizer as not a utility maximizer. Let us first consider the statistical test (6.26) in Definition 45. Assume H_0 holds. Then, the Afriat inequalities (6.6) for the true dataset $\mathcal{D}_\alpha$ to have a feasible solution. Let $\{u_t, \lambda_k\}$ denote any feasible solution to Afriat's inequalities (6.6). Then, the following inequalities result:

$$\begin{aligned}
& u_s - u_t - \lambda_k \alpha_k'(\beta_s - \beta_k) \leq 0 \\
\Leftrightarrow\ & u_s - u_k - \lambda_k \alpha_k'(\beta_s - z_s + z_s - \beta_k + z_k - z_k) \leq 0 \\
\Leftrightarrow\ & u_s - u_k - \lambda_k \left(\alpha_k'(z_s - z_k) + \alpha_k'(\omega_k - \omega_s)\right) \leq 0 \\
\Leftrightarrow\ & u_s - u_k - \lambda_k \left(\alpha_k'(z_s - z_k) + \max_{s,k} \alpha_k'(\omega_k - \omega_s)\right) \leq 0
\end{aligned} \tag{6.56}$$

Since the test statistic $\phi^*(\widehat{\mathcal{D}}_\alpha)$ (6.30) is the minimum perturbation needed for the feasibility of Afriat's inequalities (6.6), (6.56) yields the following inequality:

$$\phi^*(\widehat{\mathcal{D}}_\alpha) \leq \max_{s,k} \alpha_k'(\omega_k - \omega_s) \equiv L_\alpha \tag{6.57}$$

The Type-I error probability can now be bounded as:

$$\begin{aligned}
\mathbb{P}(H_1|H_0) &= \mathbb{P}(\mathbb{P}(\phi^*(\widehat{\mathcal{D}}_\alpha) \leq L_\alpha) \leq \gamma) \\
&= \mathbb{P}(\phi^*(\widehat{\mathcal{D}}_\alpha) \geq F_{L_\alpha}^{-1}(1-\gamma)) \\
&\leq \mathbb{P}(L_\alpha \geq F_{L_\alpha}^{-1}(1-\gamma)) \text{ from (6.57)} \\
&= 1 - \mathbb{P}(L_\alpha \leq F_{L_\alpha}^{-1}(1-\gamma)) = 1 - (1-\gamma) = \gamma
\end{aligned}$$

Hence, the Type-I error probability of the statistical test (6.26) is bounded by its significance level γ. Showing the Type-I error probability of the detector (6.27) is bounded is identical to the steps outlined above, and thus, omitted. ■

6.7.5 Masking Radar's Utility Function for Multiple Constraints

Our IRL results for estimating the radar's utility $\mathbf{u}$ (6.3) assumes a scalar-valued budget constraint for the radar. In general, the radar faces multiple constraints, or equivalently, the constraint $\mathbf{g}(\cdot) \leq 0$ is vector-valued. *How to generalize Theorem 41 to vector-valued constraints?* In this section, we generalize our IRL algorithm (Theorem 41) and cognition masking results of Theorem 44 to vector-valued $\mathbf{g}$ when the adversary knows the radar's constraints and estimates the radar's utility $\mathbf{u}$ - this scenario is formalized below in assumption 6. Generalizing IRL for identifying a vector-valued constraint $\mathbf{g}(\cdot) \leq 0$ is non-identifiable and hence, omitted.

Assumption 6 *The radar's resource constraint* $\mathbf{g}(\alpha_k, \beta_k) : \mathbb{R}_+^{2m} \to \mathbb{R}_+^I$ *in* (6.3) *is vector-valued, and the radar's utility* $\mathbf{u}(\cdot)$ *is independent of* α_k*:*

$$\mathbf{g}(\alpha_k, \beta) \equiv \{\mathbf{g}_i(\alpha_k, \beta)\}_{i=1}^I, \ \mathbf{u}(\alpha_k, \beta) \equiv \mathbf{u}(\beta), \tag{6.58}$$

where m *is the dimension of the probe/response, and* $g_i(\cdot)$ *is a scalar-valued constraint.* *IRL objective. The adversary aims to reconstruct the radar's utility* $\mathbf{u}(\cdot)$ *using the dataset* $\mathcal{D}_g$*, where* $\mathcal{D}_g$ *is defined in* (6.5)*.*

Assumption 6 specializes to assumption 2 when $\mathbf{g}$ is scalar-valued and linear in both the probe and response vectors. Let us now state Theorem 51 for achieving IRL for vector-valued constraints when assumption 6 holds.

Theorem 51 (IRL for Identifying Radar's Utility Function for Vector-Valued Constraints) *Consider the cognitive radar described in Model 1. Suppose assumption 2 holds. Then:*
(a) The adversary checks for the existence of a feasible utility function that satisfies (6.3) *by checking the feasibility of the following set of inequalities:*

There exists a feasible $\theta \in \mathbb{R}^{(1+I)N}$ *such that:*

$$(i) \ \ \theta_k - \theta_s - \sum_{i=1}^{I} \theta_{N+(s-1)I+i} \, g_i(\alpha_s, \beta_k) \leq 0, \ \forall \ s, k, \tag{6.59}$$

$$(ii)\ \theta_{1:N} > \mathbf{0},\ \theta_{N+(k-1)I+1:N+kI} \geq \mathbf{0} \textit{ and not all zeros } \forall k \tag{6.60}$$
$$\Leftrightarrow \exists\ u \textit{ s.t. } \beta_k \in \operatorname{argmax} u(\beta),\ \mathbf{g}(\alpha_k, \beta_k) \leq \mathbf{0}\ \forall k,$$

where dataset $\mathcal{D}_g$ is defined in (6.5)*.*

(b) If (6.59) *has a feasible solution, the set-valued IRL estimate of the radar's utility* $\mathbf{u}$ *is given by:*

$$u_{\text{IRL}}(\beta) \equiv \{u_{\text{IRL}}(\beta;\theta) : \text{(6.59) } \textit{is feasible}\},$$
$$u_{\text{IRL}}(\beta;\theta) = \min_k \{\theta_k + \sum_{i=1}^{I} \theta_{N+(k-1)I+i}\ (\ g_i(\alpha_k, \beta) - g_i(\alpha_k, \beta_k)\ )\}, \tag{6.61}$$

where θ is any feasible solution to the inequalities (6.59)*,* (6.60)*.*

Remarks.

1. In comparison to the linear feasibility test (6.6) of Theorem 41, (6.59) in Theorem 51 is a mixed-integer linear feasibility test, mixed-integer due to the second set of inequalities (6.60) in the feasibility test.

2. Afriat's theorem 6.6 requires that the constraint be active at the solution, meaning $\alpha_k'\beta_k = 1$ for all k. This requirement is implicitly satisfied for the scalar case due to the monotonicity of both the utility $\mathbf{u}(\cdot)$ and constraint $\alpha_k'(\beta) \leq 1$. For vector-valued $\mathbf{g}$, however, requiring all constraints to be active at the solution is highly restrictive. That is, $\mathbf{g}_i(\alpha_k, \beta_k) = 0$ for all time steps k and constraint indices i is not true in general for a cognitive radar. Hence, the IRL inequalities for multiple constraints must account for the inactive constraints for all time steps k. More precisely, the inverse learner needs to check for at least one active constraint out of all I resource constraints for all k. This is ensured by the feasibility of (6.60) in Theorem 51. At a deeper level, (6.60) tests for complementary slackness in the KKT conditions [318, Sec. 5.5] for first-order optimality of the radar's responses.

3. In Afriat's theorem (Theorem 41), the reconstructed utility function (6.7) is a point-wise minimum of scaled and shifted versions of the radar's linear constraints

$\alpha_k'\beta,\ k = 1, 2, \ldots, N$. Intuitively, the basis functions for the adversary's estimate of the radar's utility are the radar's constraints $\{\alpha_k'\beta\}_{k=1}^{N}$. For the multiple constraints case in Theorem 51 above, the adversary's utility estimate has a richer representation due to a larger set of basis functions $\{\mathbf{g}_i(\alpha_k, \beta)\}_{i,k=1}^{I,N}$.

Having defined our IRL algorithm for multiple constraints in Theorem 51 above, we now present our cognition masking result for mitigating the IRL procedure of Theorem 51.

Definition 52 (Feasibility Margin for Reconstructed Utility (6.6) **for Multiple Constraints)** *Consider the dataset $\mathcal{D}_g$ defined in* (6.5)*. Suppose the radar's constraint* $\mathbf{g}$ *is vector-valued. The feasibility margin $\mathcal{M}_u(\mathcal{D}_g)$ defined below measures how far is the utility u is from failing the IRL feasibility inequalities* (6.59)*,* (6.60)*:*

$$\mathcal{M}_u(\mathcal{D}_g) = \min_{\boldsymbol{\lambda}_{1:N}} \Big\{ \min_{s,k}\ u(\beta_s) - u(\beta_t) + \\ (\nabla\ \mathbf{g}(\alpha_s, \beta_s)\ \boldsymbol{\lambda}_s)'\ (\beta_t - \beta_s) \Big\},$$

$$\boldsymbol{\lambda}_{1:N} \in \mathbb{R}^I,\ \boldsymbol{\lambda}_k \in \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\ G_k(\boldsymbol{\lambda}),\ \boldsymbol{\lambda} \geq 0, \tag{6.62}$$

where $G_k(\cdot)$ is the dual of the optimization problem (6.3) *at time k.*

The margin definition in (6.62) above is a multi-constraint generalization of Definition 42. To glean some insight into the notation in (6.62) above, consider the simple case where $I = 1$. Then, the solution to $\operatorname{argmax}\ G_k(\boldsymbol{\lambda})$ is simply the Lagrange multiplier associated with the single operating constraint in the optimization problem (6.3). For the multiple constraint case, the solution to $\operatorname{argmax}\ G_k(\boldsymbol{\lambda})$ is the vector of Lagrange multipliers for the constraints at β_k (6.3), the optimal response at time k. Denoting the IRL feasibility test wrt inequalities (6.59) and (6.60) as $\boldsymbol{\mathcal{A}}(\theta, \mathcal{D}_u)$, the margin $\mathcal{M}_u(\mathcal{D}_u)$ for any utility u when $\mathbf{g}$ is vector-valued can be compactly defined as:

$$\mathcal{M}_u(\mathcal{D}_g) = \min_{\varepsilon \geq 0}\ \varepsilon,\ \boldsymbol{\mathcal{A}}(\theta, \mathcal{D}_g) + \varepsilon \mathbf{1} \geq \mathbf{0}. \tag{6.63}$$

Having generalized the margin definition of (6.15) in Definition 42 to the multiple constraint case, we now state our cognition masking result, Theorem 53 for vector-valued

$\mathbf{g}$. The cognition masking rationale for vector-valued $\mathbf{g}$ remains the same as that in Theorem 44: add engineered noise to the radar's optimal responses, and ensure the radar's utility $\mathbf{u}$ lies sufficiently close to the edge of the feasibility polytope of viable utilities computed via IRL.

Theorem 53 (Masking Utility from Adversarial IRL for Multiple Resource Constraints.) *Consider the cognitive radar* (6.3) *from Definition 40 with multiple resource constraints (assumption 6 holds). Let* $\{\beta_k^*\}_{k=1}^N$ *denote the naive response sequence* (6.3) *that maximizes the cognitive radar's utility. The response sequence* $\{\tilde{\beta}_{1:N}^*\}$ *defined below masks the radar's utility* $\mathbf{u}$ *from the adversary by ensuring* $\mathbf{u}$ *satisfies the IRL inequalities* (6.59)*,* (6.60) *with a sufficiently low margin* (6.62)*:*

$$\{\tilde{\beta}_{1:N}^*\} = \underset{\{\beta_k \geq \mathbf{0},\ \mathbf{g}(\alpha_k,\beta_k) \leq \mathbf{0}\}}{\operatorname{argmin}} \sum_{k=1}^{N} \mathbf{u}(\beta_k^*) - \mathbf{u}(\beta_k), \tag{6.64}$$

$$\mathcal{M}_{\mathbf{u}}(\mathcal{D}_g) \leq (1-\eta)\ \mathcal{M}_{\mathbf{u}}(\mathcal{D}_g^*), \tag{6.65}$$

where dataset $\mathcal{D}_g^* = \{\mathbf{g}(\cdot), \beta_k^*\}_{k=1}^N$ *is the adversary's dataset when the radar transmits naive responses* $\{\beta_k^*\}_{k=1}^N$*, and* $\mathcal{D}_g$ *is defined in* (6.5)*.*

The proof of Theorem 53 is straightforward and omitted due to brevity. The only distinguishing factor between Theorems 53 and 44 is the generalized definition of the margin $\mathcal{M}_{\mathbf{u}}$ (6.62) for vector-valued constraints.

6.7.6 Cognition Masking Performance under Misspecified Radar Response Measurements

In this section, we investigate how the performance of the radar's cognition masking algorithm, Theorem 44, changes when the adversary has misspecified measurements of the radar's responses. By misspecified responses, we mean the true radar response β_k is corrupted by an additive deterministic perturbation ζ_k, $k = 1, 2, \ldots, N$. Misspecified

response measurements formalized in assumption 7 below are different from noisy measurements since the perturbation ζ_k is a deterministic vector and not a random variable like in the noisy case considered in Sec. 6.4.

Assumption 7 (Misspecified Radar Response Measurements) *Suppose the adversary has misspecified measurements $\bar{\beta}_k = \beta_k + \zeta_k,\ \zeta_k \in \mathbb{R}^m$ of the radar's response β_k. Assume the misspecifications ζ_k have a bounded $\mathcal{L}_2$-norm:*

$$\|\zeta_k\|_2 \leq \zeta \tag{6.66}$$

The adversary's misspecified datasets $\bar{\mathcal{D}}_g$ and $\bar{\mathcal{D}}_u$ are defined as:

$$\begin{aligned}\bar{\mathcal{D}}_g &\equiv \{\mathbf{g}(\alpha_k, \cdot), \beta_k + \zeta_k\}_{k=1}^N,\ \zeta_k \in \mathbb{R}^m \\ \bar{\mathcal{D}}_u &\equiv \{\mathbf{u}(\alpha_k, \cdot), \beta_k + \zeta_k\}_{k=1}^N,\ \zeta_k \in \mathbb{R}^m\end{aligned} \tag{6.67}$$

where β_k is the radar's response at time k, $\mathbf{u}$ and $\mathbf{g}$ are the utility and constraint, respectively, of the cognitive radar (6.3)*.*

Recall from Theorem 44 that the positive scalar η parametrizes the extent of cognition masking by the cognitive radar. Our key objective is to derive a lower bound on the effective extent of cognition masking η_{eff} defined as:

$$\begin{aligned}\eta_{\text{eff}} &= \mathcal{M}_{\mathbf{u}}(\tilde{\bar{\mathcal{D}}}_g)/\,\mathcal{M}_{\mathbf{u}}(\bar{\mathcal{D}}_g)\ (\text{when assumption 2 holds}), \\ \eta_{\text{eff}} &= \mathcal{M}_{\mathbf{g}}(\tilde{\bar{\mathcal{D}}}_u)/\,\mathcal{M}_{\mathbf{g}}(\bar{\mathcal{D}}_u)\ (\text{when assumption 3 holds}),\end{aligned} \tag{6.68}$$

where $\bar{\mathcal{D}}_g, \bar{\mathcal{D}}_u$ (6.67) are the misspecified datasets of the adversary if the radar transmits naive responses (6.3), and $\tilde{\bar{\mathcal{D}}}_g, \bar{\mathcal{D}}_u$ are the misspecified datasets when the radar transmits cognition masking responses computed via (6.17) and (6.19), respectively. The variable η_{eff} (6.68) is the ratio between the margins of the radar's strategy (utility $\mathbf{u}$ or constraint $\mathbf{g}$) (6.15), (6.16) with and without the radar's cognition masking scheme (Theorem 44) when the adversary has misspecified response measurements. It is easy to see that $\eta_{\text{eff}} = \eta$

if the misspecification error ζ_k (6.67) is 0 for all k. For non-zero ζ_k, Theorem 54 below yields a lower bound for η_{eff} and uses the following variables (given assumption 7 holds):

$$d_{1,\mathbf{u}} = \min_k \nabla \mathbf{u}(\beta_k)' \zeta_k \ , \ d_{2,\mathbf{u}} = \max_k \nabla \mathbf{u}(\beta_k)' \zeta_k,$$
$$d_{1,\mathbf{g}} = \min_k \nabla \mathbf{g}(\beta_k)' \zeta_k \ , \ d_{2,\mathbf{g}} = \max_k \nabla \mathbf{g}(\beta_k)' \zeta_k, \tag{6.69}$$

The variables defined in (6.69) measure the deviation in the radar's utility and constraint values evaluated at the misspecified radar responses measured by the adversary, compared to the utility and constraint evaluations at the true radar responses. We are now ready to state Theorem 54.

Theorem 54 (Performance of Cognition Masking (Theorem 44) for Misspecified Responses) *Consider the cognition masking scheme of Theorem 44. Assume the adversary has misspecified radar response measurements (assumption 7 holds). Then:*
(i) Suppose assumption 2 holds, i.e., the adversary knows the radar's constraint $\mathbf{g}$. *Then, the effective extent of cognition masking* η_{eff} *is bounded from below as:*

$$\eta_{\text{eff}} \geq \eta - \left(\frac{(1-\eta)\ (d_{2,\mathbf{u}} - d_{1,\mathbf{u}})}{\mathcal{M}_{\mathbf{u}}(\mathcal{D}_g) - d_{2,\mathbf{u}}} \right) \tag{6.70}$$

(ii) Suppose assumption 3 holds, i.e., the adversary knows the radar's constraint $\mathbf{u}$. *Then, the effective extent of cognition masking* η_{eff} *is bounded from below as:*

$$\eta_{\text{eff}} \geq \eta - \left(\frac{(1-\eta)\ (d_{2,\mathbf{g}} - d_{1,\mathbf{g}})}{\mathcal{M}_{\mathbf{g}}(\mathcal{D}_u) - d_{2,\mathbf{g}}} \right) \tag{6.71}$$

The variables $d_{1,\mathbf{u}}, d_{2,\mathbf{u}}, d_{1,\mathbf{g}}, d_{2,\mathbf{g}}$ *measure the distortion in the adversary's dataset due to misspecified measurements and defined in* (6.69).

Theorem 54 computes a lower bound on the effectiveness of the cognitive masking scheme of Theorem 44 when the adversary has misspecified measurements of the radar's response. The proof for Theorem 54 is omitted for brevity. Observe that the bounds in (6.70), (6.71) are inversely proportional to the quantities $(d_{2,\mathbf{u}} - d_{1,\mathbf{u}})$ and $(d_{2,\mathbf{g}} - d_{1,\mathbf{g}})$. These quantities can be interpreted as the 'spread' in the utility and constraint evaluations

at the radar's true responses due to the misspecification errors ζ_k (6.67) and, in turn, are proportional to ζ (6.66), the maximum $\mathcal{L}_2$ norm of $\{\zeta_k\}_{k=1}^N$. Hence, we can conclude the lower bound for the effectiveness of the radar's cognition masking scheme worsens with the magnitude of the misspecification errors in the adversary's measurements.

6.7.7 Cognition Masking for Arbitrary IRL Algorithm

Our cognition masking results of Theorems 44 and 48 assume the adversary performs optimal IRL via Afriat's theorem (Theorems 41 and 50) to reconstruct the radar's strategy. However, our cognition masking results can be straightforwardly extended to any IRL algorithm. Any IRL algorithm can be expressed WLOG as a set-valued estimation algorithm that generates a set of feasible strategies given a finite dataset of adversary probes $\{\alpha_k\}_{k=1}^N$ and radar responses $\{\beta_k\}_{k=1}^N$:

$$\mathrm{IRL}(\theta, \{\alpha_k, \beta_k\}_{k=1}^N) \leq \mathbf{0}, \text{where } \mathrm{IRL} : \Theta \times \mathbb{R}_+^{2mN} \to \mathbb{R}^L \tag{6.72}$$

In (6.72), $\theta \in \Theta$ parametrizes the reconstructed utility, $\{\alpha_k, \beta_k\}_{k=1}^N$ is the adversary's dataset and L is the number of IRL feasibility inequalities. In Afriat's theorem, for example, $\Theta = \mathbb{R}_+^{2N}$ and $L = N^2 - N$. Algorithm 6 below outlines the steps for mitigating an arbitrary IRL algorithm $\mathrm{IRL}(\cdot, \{\alpha_k, \beta_k\}_{k=1}^N)$. Recall Theorem 44 minimizes the feasibility margin of the radar's strategy wrt the Afriat inequalities (6.6), (6.50) by deliberately perturbing the radar's responses. In complete analogy, a radar can hide its strategy from any set-valued IRL estimation scheme by minimizing the feasibility margin defined below in (6.73) wrt the IRL feasibility inequalities $\mathrm{IRL}(\cdot, \{\alpha_k, \beta_k\}_{k=1}^N)$ by purposefully injecting noise in the radar's responses. Due to the non-linear margin constraint in (6.74), the optimization problem can be solved using a general purpose non-linear programming solver, for example, `fmincon` in MATLAB, to obtain a local minimum.

Algorithm 6 Masking Radar Utility from Arbitrary IRL algorithm $\mathrm{IRL}(\cdot, \{\alpha_k, \beta_k\}_{k=1}^N) \leq \mathbf{0}$

Step 1. Compute radar's naive response sequence $\beta_{1:N}^*$ by solving the convex optimization problem (6.3):

$$\beta_k^* = \operatorname{argmin} \mathbf{u}(\beta),\ \mathbf{g}(\alpha_k, \beta) \leq 0,\ \beta \geq \mathbf{0}\ \forall k \in \{1, 2, \ldots, N\},$$

where $\mathbf{u}$ is concave monotone in β and $\mathbf{g}(\alpha_k, \beta)$ is convex monotone in β.
Step 2. Choose $\eta \in [0, 1]$ (extent of cognition masking from IRL feasibility test).
Step 3. Compute the margin of the naive responses wrt the IRL algorithm:

$$\begin{aligned}\mathcal{M}_u(\{\alpha_k, \beta_k^*\}_{k=1}^N; \mathrm{IRL}) = \min_{\varepsilon \geq 0}\ \varepsilon,\\ \mathrm{IRL}(\mathbf{u}, \{\alpha_k, \beta_k^*\}_{k=1}^N) + \varepsilon \mathbf{1} \geq \mathbf{0},\end{aligned} \tag{6.73}$$

where $\mathbf{1}$ is a vector of all ones and $\mathbf{u}$ is the radar's utility.
Step 3. Compute upper bound $\mathcal{M}_{\mathrm{thresh}}$ on desired margin (6.15) after cognition masking: $\mathcal{M}_{\mathrm{thresh}} = (1 - \eta)\ \mathcal{M}_{\mathbf{u}}(\{\alpha_k, \beta_k^*\}_{k=1}^N; \mathrm{IRL})$.
Step 4. Compute the cognition-making response sequence $\{\tilde{\beta}_{1:N}^*\}$ by solving the following optimization problem:

$$\begin{aligned}\min\ \sum_{k=1}^N \mathbf{u}(\beta_k^*) - \mathbf{u}(\beta_k),\\ \beta_k \geq \mathbf{0},\ \alpha_k' \beta_k \leq 1\ \forall k \in \{1, 2, \ldots, N\},\\ \mathcal{M}_{\mathbf{u}}(\{\alpha_k, \beta_k\}_{k=1}^N; \mathrm{IRL}) \leq \mathcal{M}_{\mathrm{thresh}}\,.\end{aligned} \tag{6.74}$$

6.7.8 Context. ECCM and Meta-Cognition

System level ECCM vs Pulse level ECCM. Our cognition masking algorithm is implemented at the system level (Bayesian tracker level) and not the pulse level (Wiener filter level). Pulse-level ECCM [319–321] accomplishes LPI-type functionalities for cognitive radars. Cognition masking hides the radar's *strategy* from the adversary instead of mitigating the adversary's detection of the radar's transmission. Hence, cognition masking ECCM is deployed at a higher level of abstraction than pulse level ECCM.

Hiding Cognition against Optimal IRL vs Sub-optimal IRL. The cognition masking results in this chapter assume the adversary performs optimal IRL using Afriat's theorem [32, 127]. Afriat's theorem achieves optimal IRL for non-parametric utility estimation of a

cognitive radar as it generates a polytope of *all viable utilities* that rationalizes a finite dataset of adversarial probes and radar responses. However, our cognition masking results can be extended to *any* potentially sub-optimal IRL algorithm that generates a set-valued estimate of the radar's utility, as long as the radar has knowledge of the IRL algorithm being used by the adversary. Algorithm 6 in the appendix outlines how a cognitive radar can mask its cognition for an arbitrary IRL algorithm. At an abstract level, cognition masking simply obfuscates a set-valued mapping from the adversary's dataset to a set of feasible utilities by intelligently distorting the radar's responses and hence, is not affected by the optimality of adversarial IRL.

At a deeper level, this chapter also quantifies cognition masking performance when the adversary has misspecified measurements of the radar's response, and performs sub-optimal IRL. Theorem 54 provides performance guarantees for cognition masking when the radar does not know the misspecification errors and provides performance bounds in terms of the error magnitude.

Why not an MDP or non-cooperative game? In machine learning based IRL [6,26,28], the aim is to reconstruct the rewards of a Markov decision process (MDP) subject to entropic constraints on the policy. This requires complete knowledge of the transition dynamics of the adversary's probes. In comparison, our radar-adversary interaction is batch-wise - the adversary transmits a batch of probe signals, and then the radar responds with a batch of responses. This non-parametric identification of the radar's strategy is agnostic to transition dynamics in the adversary's probes. Hence, a static utility maximization setup is more realistic for IRL and inverse IRL involving cognitive radar.

We consider a radar-adversary interaction where the adversary is not aware of the radar's cognition masking strategy. A more general formulation is a Stackelberg game between the radar and the adversary, with the adversary as the leader and the radar as the follower. However, such an approach for computing the optimal meta-cognition strategy for the

radar is ill-posed since the existence of a pure and unique Nash equilibrium is not guaranteed. Finally, from an inverse game theoretic perspective, identifying if the radar-adversary behavior is consistent with Nash equilibrium is intractable since the analyst needs to know both the radar's and adversary's utility function. Addressing these issues is beyond the scope of this chapter, and the subject of future work.

CHAPTER 7

CONCLUDING REMARKS

In this thesis, we discuss inverse reinforcement learning from a micro-economics lens. Traditionally, inverse reinforcement learning has been studied as a subset of machine learning and estimates reward functions from an agent's behavior, where the agent acts optimally in an MDP setting. In economics, a similar question is addressed in the fields of revealed preference and revealed rational inattention. This thesis attempts to unify and enrich the two areas. We first formulate necessary and sufficient conditions for IRL for Bayes optimal stopping using results from the revealed preference literature. Second, we unified revealed preference and costly information acquisition via an equivalence parameter map. This unification helps in extending robustness tests for utility maximization from revealed preference literature to revealed rational inattention. Third, we use the constrained expected utility maximization framework of revealed rational inattention to construct simple intuitive interpretable models that rationalize the decisions of deep neural networks for image classification. Finally, we propose inverse IRL (I-IRL) where the key aim is to spoof an adversarial IRL system. The key idea is to deliberately choose sub-optimal responses that trade-off between maximizing the system utility and the probability of correct reconstruction of the system utility via IRL. We illustrate our I-IRL algorithm in the context of meta-cognitive radars mitigating an adversarial target that intends to learn the radar utility.

Future Work and Extensions

An immediate extension of the work in this thesis would be to generalize IRL algorithms to multiple cooperative agents that have a shared utility (social welfare). Revealed preference-based IRL is an offline method - the utility can be estimated *after* responses from multiple time stamps have been aggregated by an analyst. One interesting extension

currently being investigated is to convert the offline IRL method to an online technique using stochastic approximation algorithms. The key idea is to realize that rationalizability in the revealed preference sense is equivalent to minimizing the expected hinge loss of Afriat inequalities, which can be formulated into a stochastic gradient-type algorithm.

BIBLIOGRAPHY

[1] J. Choi and K. Kim. Inverse reinforcement learning in partially observable environments. *Journal of Machine Learning Research*, 12:691–730, 2011.

[2] A. Caplin and M. Dean. Revealed preference, rational inattention, and costly information acquisition. *The American Economic Review*, 105(7):2183–2203, 2015.

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

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

[5] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

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

[7] 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.

[8] J. Rust. Structural estimation of Markov decision processes. *Handbook of econometrics*, 4:3081–3143, 1994.

[9] P. Rolland, L. Viano, N. Schürhoff, B. Nikolov, and V. Cevher. Identifiability and generalizability from multiple experts in inverse reinforcement learning. *arXiv preprint arXiv:2209.10974*, 2022.

[10] G. Lee, M. Luo, F. Zambetta, and X. Li. Learning a super mario controller from examples of human play. In *2014 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2014.

[11] M. Gombolay, R. Jensen, J. Stigile, S.-H. Son, and J. Shah. Apprenticeship scheduling: learning to schedule from human experts. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 826–833, 2016.

[12] W. Hoiles, V. Krishnamurthy, and K. Pattanayak. Rationally inattentive inverse reinforcement learning explains YouYube commenting behavior. *The Journal of Machine Learning Research*, 21(1):6879–6917, 2020.

[13] K. Kim, S. Garg, K. Shiragur, and S. Ermon. Reward identification in inverse reinforcement learning. In *International Conference on Machine Learning*, pages 5496–5505. PMLR, 2021.

[14] D. Martin. Bayesian revealed preferences. *Available at SSRN 2393035*, 2014.

[15] A. Caplin and D. Martin. A testable theory of imperfect perception. *The Economic Journal*, 125(582):184–202, 2015.

[16] V. Krishnamurthy. Adversarial radar inference. from inverse tracking to inverse reinforcement learning of cognitive radar. *arXiv preprint arXiv:2002.10910*, 2020.

[17] C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of Markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

[18] D. P. Bertsekas. Dynamic programming and optimal control 4th edition, volume ii. *Athena Scientific*, 2015.

[19] V. Krishnamurthy. *Partially observed Markov decision processes*. Cambridge University Press, 2016.

[20] W. S. Lovejoy. On the convexity of policy regions in partially observed systems. *Operations Research*, 35(4):619–621, 1987.

[21] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

[22] V. Krishnamurthy and B. Wahlberg. POMDP multiarmed bandits – structural results. *Mathematics of Operations Research*, 34(2):287–302, May 2009.

[23] J. Hong, B. Kveton, M. Zaheer, and M. Ghavamzadeh. Hierarchical Bayesian bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 7724–7741. PMLR, 2022.

[24] X. Fontaine, Q. Berthet, and V. Perchet. Regularized contextual bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2144–2153. PMLR, 2019.

[25] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[26] 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.

[27] C. P. Chamley. *Rational herds: Economic models of social learning*. Cambridge University Press, 2004.

[28] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy Inverse Reinforcement Learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

[29] A. Caplin, M. Dean, and J. Leahy. Rational inattention, optimal consideration sets, and stochastic choice. *The Review of Economic Studies*, 86(3):1061–1094, 2019.

[30] D. Brown, W. Goo, P. Nagarajan, and S. Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, pages 783–792. PMLR, 2019.

[31] L. Le Cam. On some asymptotic properties of maximum likelihood estimates and related bayes' estimates. *Univ. Calif. Publ. in Statist.*, 1:277–330, 1953.

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

[33] M. Houtman and J. Maks. Determining all maximal data subsets consistent with revealed preference. *Kwantitatieve methoden*, 19(1):89–104, 1985.

[34] S. N. Afriat. Efficiency estimation of production functions. *International economic review*, pages 568–598, 1972.

[35] H. R. Varian et al. *Goodness-of-fit for revealed preference tests*. Department of Economics, University of Michigan Ann Arbor, 1991.

[36] M. L. Khan. Social media engagement: What motivates user participation and consumption on YouTube? *Computers in Human Behavior*, 66:236–247, 2017.

[37] C. Sims. Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690, 2003.

[38] A. Van Der Vaart and J. A. Wellner. Preservation theorems for glivenko-cantelli and uniform glivenko-cantelli classes. In *High dimensional probability II*, pages 115–133. Springer, 2000.

[39] M. R. Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media, 2007.

[40] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[41] P. J. Reny. A characterization of rationalizable consumer behavior. *Econometrica*, 83(1):175–192, 2015.

[42] A. Mas-Colell. On revealed preference analysis. *The Review of Economic Studies*, 45(1):121–131, 1978.

[43] M. Wulfmeier, P. Ondruska, and I. Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015.

[44] C. You, J. Lu, D. Filev, and P. Tsiotras. Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning. *Robotics and Autonomous Systems*, 114:1–18, 2019.

[45] X. Lan and M. Schwager. Planning periodic persistent monitoring trajectories for sensing robots in Gaussian random fields. In *2013 IEEE International Conference on Robotics and Automation*, pages 2415–2420. IEEE, 2013.

[46] A. Bry and N. Roy. Rapidly-exploring random belief trees for motion planning under uncertainty. In *2011 IEEE International Conference on Robotics and Automation*, pages 723–730. IEEE, 2011.

[47] H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard. Socially compliant mobile robot navigation via inverse reinforcement learning. *The International Journal of Robotics Research*, 35(11):1289–1307, 2016.

[48] S. Sharifzadeh, I. Chiotellis, R. Triebel, and D. Cremers. Learning to drive using inverse reinforcement learning and deep q-networks. *arXiv preprint arXiv:1612.03653*, 2016.

[49] V. Krishnamurthy, D. Angley, R. Evans, and B. Moran. Identifying cognitive radars-inverse reinforcement learning using revealed preferences. *IEEE Transactions on Signal Processing*, 68:4529–4542, 2020.

[50] R. Ratcliff and P. L. Smith. A comparison of sequential sampling models for two-choice reaction time. *Psychological review*, 111(2):333, 2004.

[51] I. Krajbich, C. Armel, and A. Rangel. Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience*, 13(10):1292–1298, 2010.

[52] M. Arik and O. B. Akan. Enabling cognition on electronic countermeasure systems against next-generation radars. In *MILCOM 2015-2015 IEEE Military Communications Conference*, pages 1103–1108. IEEE, 2015.

[53] H. Song, M. Xiao, J. Xiao, Y. Liang, and Z. Yang. A POMDP approach for scheduling the usage of airborne electronic countermeasures in air operations. *Aerospace Science and Technology*, 48:86–93, 2016.

[54] I. Arasaratnam, S. Haykin, T. Kirubarajan, and F. A. Dilkes. Tracking the mode of operation of multi-function radars. In *2006 IEEE Conference on Radar*, pages 6–pp. IEEE, 2006.

[55] F. Bourgault, T. Furukawa, and H. F. Durrant-Whyte. Optimal search for a lost target in a Bayesian world. In *Field and service robotics*, pages 209–222. Springer, 2003.

[56] R. Self, M. Harlan, and R. Kamalapurkar. Online inverse reinforcement learning for nonlinear systems. In *2019 IEEE conference on control technology and applications (CCTA)*, pages 296–301. IEEE, 2019.

[57] W. Xue, P. Kolaric, J. Fan, B. Lian, T. Chai, and F. L. Lewis. Inverse reinforcement learning in tracking control based on inverse optimal control. *IEEE Transactions on Cybernetics*, 2021.

[58] M. Nishio, M. Nishizawa, O. Sugiyama, R. Kojima, M. Yakami, T. Kuroda, and K. Togashi. Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. *PloS one*, 13(4):e0195875, 2018.

[59] A. Oniśko and M. J. Druzdzel. Impact of precision of Bayesian network parameters on accuracy of medical diagnostic systems. *Artificial intelligence in medicine*, 57(3):197–206, 2013.

[60] K. P. Exarchos, T. P. Exarchos, C. V. Bourantas, M. I. Papafaklis, K. K. Naka, L. K. Michalis, O. Parodi, and D. I. Fotiadis. Prediction of coronary atherosclerosis progression using dynamic Bayesian networks. In *2013 35th Annual International*

Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 3889–3892. IEEE, 2013.

[61] J. Jack Lee and C. T. Chu. Bayesian clinical trials in action. *Statistics in medicine*, 31(25):2955–2972, 2012.

[62] N. V. Thakor, A. Natarajan, and G. F. Tomaselli. Multiway sequential hypothesis testing for tachyarrhythmia discrimination. *IEEE Transactions on Biomedical Engineering*, 41(5):480–487, 1994.

[63] M. A. Ahmad, C. Eckert, and A. Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 559–560, 2018.

[64] W. Jeon, C.-Y. Su, P. Barde, T. Doan, D. Nowrouzezahrai, and J. Pineau. Regularized inverse reinforcement learning. *arXiv preprint arXiv:2010.03691*, 2020.

[65] K. Lee, S. Kim, S. Lim, S. Choi, M. Hong, J. I. Kim, Y.-L. Park, and S. Oh. Generalized Tsallis entropy reinforcement learning and its application to soft mobile robots. In *Robotics: Science and Systems*, volume 16, pages 1–10, 2020.

[66] M. Herman, T. Gindele, J. Wagner, F. Schmitt, and W. Burgard. Inverse reinforcement learning with simultaneous estimation of rewards and dynamics. In *Artificial intelligence and statistics*, pages 102–110. PMLR, 2016.

[67] S. Levine and V. Koltun. Continuous inverse optimal control with locally optimal examples. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 475–482, 2012.

[68] J. Fu, K. Luo, and S. Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018.

[69] M. Wulfmeier, P. Ondruska, and I. Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015.

[70] C. Finn, S. Levine, and P. Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning*, pages 49–58, 2016.

[71] D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. In

Proceedings of the 20th international joint conference on Artifical intelligence, pages 2586–2591, 2007.

[72] H. Cao, S. Cohen, and L. Szpruch. Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12362–12373, 2021.

[73] J. Choi and K.-E. Kim. Inverse reinforcement learning in partially observable environments. In *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[74] T. Makino and J. Takeuchi. Apprenticeship learning for model parameters of partially observable environments. *arXiv preprint arXiv:1206.6484*, 2012.

[75] M. Kwon, S. Daptardar, P. Schrater, and X. Pitkow. Inverse rational control with partially observable continuous nonlinear dynamics. *arXiv preprint arXiv:2009.12576*, 2020.

[76] P. A. Samuelson. A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61–71, 1938.

[77] H. R. Varian. The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, pages 945–973, 1982.

[78] M. Woodford. Inattentive valuation and reference-dependent choice. *Unpublished Manuscript, Columbia University*, 2012.

[79] M. Lopes, F. Melo, and L. Montesano. Active learning for reward estimation in inverse reinforcement learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 31–46. Springer, 2009.

[80] H. Poor. *An Introduction to Signal Detection and Estimation*. Springer-Verlag, New York, 2 edition, 1993.

[81] S. M. Ross. *Introduction to stochastic dynamic programming*. Academic press, 2014.

[82] E.-M. Wong, F. Bourgault, and T. Furukawa. Multi-vehicle Bayesian search for multiple lost targets. In *Proceedings of the 2005 ieee international conference on robotics and automation*, pages 3169–3174. IEEE, 2005.

[83] O. Pele and M. Werman. Robust real-time pattern matching using Bayesian

sequential hypothesis testing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1427–1443, 2008.

[84] N. A. Goodman, P. R. Venkata, and M. A. Neifeld. Adaptive waveform design and sequential hypothesis testing for target recognition with active sensors. *IEEE Journal of Selected Topics in Signal Processing*, 1(1):105–113, 2007.

[85] J. C. Gittins. *Multi–armed Bandit Allocation Indices*. Wiley, 1989.

[86] S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[87] L. Chan, D. Hadfield-Menell, S. Srinivasa, and A. Dragan. The assistive multi-armed bandit. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 354–363. IEEE, 2019.

[88] R. Noothigattu, T. Yan, and A. D. Procaccia. Inverse reinforcement learning from like-minded teachers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35-10, pages 9197–9204, 2021.

[89] D. Blackwell. Equivalent comparisons of experiments. *The annals of mathematical statistics*, pages 265–272, 1953.

[90] F. Matejka and A. Kay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–98, 2015.

[91] H. K. Kwon and A. Gruzd. Is offensive commenting contagious online? examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research*, 27(4):991–1010, 2017.

[92] S. Alhabash, J.-h. Baek, C. Cunningham, and A. Hagerstrom. To comment or not to comment?: How virality, arousal level, and commenting behavior on YouTube videos affect civic behavioral intentions. *Computers in human behavior*, 51:520–531, 2015.

[93] A. Aprem and V. Krishnamurthy. Utility change point detection in online social media: A revealed preference framework. *IEEE Transactions on Signal Processing*, 65(7), April 2017.

[94] A. Glöckner and T. Betsch. Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *MPI Collective Goods Preprint*, 2(2008), 2008.

[95] J. L. McClelland and D. E. Rumelhart. *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. MIT press, 1989.

[96] E. K. P. Chong, C. Kreucher, and A. Hero. Partially observable Markov decision process approximations for adaptive sensing. *Discrete Event Dynamic Systems*, 19(3):377–422, 2009.

[97] V. Krishnamurthy and D. Djonin. Structured threshold policies for dynamic sensor scheduling–a partially observed Markov decision process approach. *IEEE Transactions on Signal Processing*, 55(10):4938–4957, Oct. 2007.

[98] V. Krishnamurthy and D. Djonin. Optimal threshold policies for multivariate POMDPs in radar resource management. *IEEE Transactions on Signal Processing*, 57(10), 2009.

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

[100] J. S. Bergin, J. R., R. M. Guerci, and M. Rangaswamy. MIMO clutter discrete probing for cognitive radar. In *IEEE International Radar Conference*, pages 1666–1670, April 2015.

[101] J. R. Guerci, J. S. Bergin, R. J. Guerci, M. Khanin, and M. Rangaswamy. A New MIMO Clutter Model for Cognitive Radar. In *IEEE Radar Conference*, May 2016.

[102] V. Krishnamurthy and M. Rangaswamy. How to calibrate your adversary's capabilities? inverse filtering for counter-autonomous systems. *IEEE Transactions on Signal Processing*, 67(24):6511–6525, 2019.

[103] V. Krishnamurthy, D. Angley, R. Evans, and B. Moran. Identifying cognitive radars - inverse reinforcement learning using revealed preferences. *IEEE Transactions on Signal Processing*, 68:4529–4542, 2020.

[104] R. Mattila, C. Rojas, V. Krishnamurthy, and B. Wahlberg. Inverse filtering for hidden Markov models. In *Advances in Neural Information Processing Systems*, pages 4204–4213, 2017.

[105] R. Mattila, C. Rojas, V. Krishnamurthy, and B. Wahlberg. Inverse filtering for

linear Gaussian state-space models. In *Proceedings of IEEE Conference on Decision and Control*, 2018.

[106] R. Mattila, I. Lourenço, C. R. Rojas, V. Krishnamurthy, and B. Wahlberg. Estimating private beliefs of Bayesian agents based on observed decisions. *IEEE Control Systems Letters*, 2019.

[107] C. Chamley. *Rational herds: Economic Models of Social Learning*. Cambridge University Press, 2004.

[108] G. Angeletos, C. Hellwig, and A. Pavan. Dynamic global games of regime change: Learning, multiplicity, and the timing of attacks. *Econometrica*, 75(3):711–756, 2007.

[109] V. Krishnamurthy. *Partially Observed Markov Decision Processes. From Filtering to Controlled Sensing*. Cambridge University Press, 2016.

[110] V. Krishnamurthy. Quickest detection POMDPs with social learning: Interaction of local and global decision makers. *IEEE Transactions on Information Theory*, 58(8):5563–5587, 2012.

[111] V. Krishnamurthy. Bayesian sequential detection with phase-distributed change time and nonlinear penalty – a lattice programming POMDP approach. *IEEE Transactions on Information Theory*, 57(3):7096–7124, Oct. 2011.

[112] C.-C. Huang, B. Amini, and R. R. Bitmead. Predictive coding and control. *IEEE Transactions on Control of Network Systems*, 6(2):906–918, 2018.

[113] D. Ciuonzo, P. K. Willett, and Y. Bar-Shalom. Tracking the tracker from its passive sonar ml-pda estimates. *IEEE Transactions on Aerospace and Electronic Systems*, 50(1):573–590, 2014.

[114] V. Krishnamurthy, E. Leoff, and J. Sass. Filterbased stochastic volatility in continuous-time hidden Markov models. *Econometrics and statistics*, 6:1–21, 2018.

[115] R. J. Elliott, L. Aggoun, and J. B. Moore. *Hidden Markov Models – Estimation and Control*. Springer-Verlag, New York, 1995.

[116] B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech, 2004.

[117] O. Cappe, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models*. Springer-Verlag, 2005.

[118] P. Del Moral, E. Rio, et al. Concentration inequalities for mean field particle models. *Annals of Applied Probability*, 21(3):1017–1052, 2011.

[119] J. Marion. *Finite Sample Bounds and Path Selection for Sequential Monte Carlo*. PhD thesis, Duke University, 2018.

[120] J. Cavanaugh and R. Shumway. On computing the expected fisher information matrix for state space model parameters. *Statistics & Probability Letters*, 26:347–355, 1996.

[121] B. D. O. Anderson and J. B. Moore. *Optimal filtering*. Prentice Hall, Englewood Cliffs, New Jersey, 1979.

[122] P. Caines. *Linear Stochastic Systems*. Wiley, 1988.

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

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

[125] W. Diewert. Afriat's theorem and some extensions to choice under uncertainty. *The Economic Journal*, 122(560):305–331, 2012.

[126] A. Kuptel. Counter unmanned autonomous systems (cuaxs): Priorities. policy. future capabilities. *Policy. Future Capabilities (May 5, 2017). Multinational Capability Development Campaign (MCDC)*, pages 15–16, 2017.

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

[128] S. Afriat. *Logic of choice and economic theory*. Clarendon Press Oxford, 1987.

[129] H. R. Varian. Nonparametric tests of models of investor behavior. *Journal of Financial and Quantitative Analysis*, pages 269–278, 1983.

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

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

[132] J. Guerci, J. Bergin, R. Guerci, M. Khanin, and M. Rangaswamy. A new MIMO clutter model for cognitive radar. In *2016 IEEE Radar Conference (RadarConf)*, pages 1–6. IEEE, 2016.

[133] M. Maskery and V. Krishnamurthy. Network-enabled missile deflection: Games and correlation equilibrium. *IEEE Transactions on Aerospace and Electronic Systems*, 43(3):843–863, July 2007.

[134] H. S. Houthakker. Revealed preference and the utility function. *Economica*, 17(66):159–174, 1950.

[135] D. J. Brown and C. Calsamiglia. The nonparametric approach to applied welfare analysis. *Economic Theory*, 31(1):183–188, 2007.

[136] R. T. Rockafellar. *Convex analysis*. Princeton university press, 2015.

[137] C. P. Chambers, C. Liu, and J. Rehbeck. Nonseparable costly attention and revealed preference. Technical report, working paper, 2017.

[138] D. Blackwell. Equivalent comparisons of experiments. *The annals of mathematical statistics*, pages 265–272, 1953.

[139] Y. Hu, S. Farnham, and K. Talamadupula. Predicting user engagement on twitter with real-world events. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9-1, pages 168–177, 2015.

[140] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, pages 49–62, 2009.

[141] M. Lopes, F. Melo, and L. Montesano. Active learning for reward estimation in inverse reinforcement learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 31–46. Springer, 2009.

[142] C. Dimitrakakis and C. A. Rothkopf. Bayesian multitask inverse reinforcement learning. In *European workshop on reinforcement learning*, pages 273–284. Springer, 2011.

[143] K. Pattanayak and V. Krishnamurthy. Necessary and sufficient conditions for inverse reinforcement learning of Bayesian stopping time problems. *Journal of Machine Learning Research*, 24(52):1–64, 2023.

[144] K. Pattanayak, V. Krishnamurthy, and C. Berry. How can a cognitive radar mask its cognition? In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5897–5901. IEEE, 2022.

[145] K. Pattanayak and V. Krishnamurthy. Behavioral economics approach to interpretable deep image classification. rationally inattentive utility maximization explains deep image classification. *arXiv preprint arXiv:2102.04594*, 2021.

[146] M. K. Richter. Revealed preference theory. *Econometrica: Journal of the Econometric Society*, pages 635–645, 1966.

[147] H. Nishimura, E. A. Ok, and J. K.-H. Quah. A comprehensive approach to revealed preference theory. *American Economic Review*, 107(4):1239–63, 2017.

[148] M. Freer and C. Martinelli. A representation theorem for general revealed preference. *GMU Working Paper in Economics No. 16-21, Available at SSRN: https://ssrn.com/abstract=2791906 or http://dx.doi.org/10.2139/ssrn.2791906*, 2016.

[149] M. Freer and C. Martinelli. A utility representation theorem for general revealed preference. *Mathematical Social Sciences*, 111:68–76, 2021.

[150] C. P. Chambers, C. Liu, and J. Rehbeck. Costly information acquisition. *Journal of Economic Theory*, 186:104979, 2020.

[151] M. Freer and C. Martinelli. An algebraic approach to revealed preference. *Economic Theory*, pages 1–26, 2022.

[152] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017.

[153] J. Liu, Y. Pan, M. Li, Z. Chen, L. Tang, C. Lu, and J. Wang. Applications of deep learning to mri images: A survey. *Big Data Mining and Analytics*, 1(1):1–18, 2018.

[154] J. Zabalza, J. Ren, J. Zheng, H. Zhao, C. Qing, Z. Yang, P. Du, and S. Marshall.

Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging. *Neurocomputing*, 185:1–10, 2016.

[155] C. Hong, J. Yu, J. Wan, D. Tao, and M. Wang. Multimodal deep autoencoder for human pose recovery. *IEEE Transactions on Image Processing*, 24(12):5659–5670, 2015.

[156] S. N. Afriat. Efficiency estimation of production functions. *International economic review*, pages 568–598, 1972.

[157] M. Houtman and J. Maks. Determining all maximal data subsets consistent with revealed preference. *Kwantitatieve methoden*, 19(1):89–104, 1985.

[158] H. R. Varian. Non-parametric analysis of optimizing behavior with measurement error. *Journal of Econometrics*, 30(1-2):445–458, 1985.

[159] A. Brodersen, S. Scellato, and M. Wattenhofer. Youtube around the world: geographic popularity of videos. In *Proceedings of the 21st international conference on World Wide Web*, pages 241–250, 2012.

[160] G. Chatzopoulou, C. Sheng, and M. Faloutsos. A first step towards understanding popularity in youtube. In *2010 INFOCOM IEEE Conference on Computer Communications Workshops*, pages 1–6. IEEE, 2010.

[161] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 1–14, 2007.

[162] H. R. Varian. The nonparametric approach to production analysis. *Econometrica: Journal of the Econometric Society*, pages 579–597, 1984.

[163] T. C. Koopmans and M. Beckmann. Assignment problems and the location of economic activities. *Econometrica: journal of the Econometric Society*, pages 53–76, 1957.

[164] S. Qin and C. Basak. Age-related differences in brain activation during working memory updating: an fmri study. *Neuropsychologia*, 138:107335, 2020.

[165] J. G. McCoy and R. E. Strecker. The cognitive cost of sleep lost. *Neurobiology of learning and memory*, 96(4):564–582, 2011.

[166] S. Musslick and J. D. Cohen. Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences*, 25(9):757–775, 2021.

[167] A. Caplin, M. Dean, and J. Leahy. Rational inattention, optimal consideration sets, and stochastic choice. *The Review of Economic Studies*, 86(3):1061–1094, 2019.

[168] R. Deb, Y. Kitamura, J. K. Quah, and J. Stoye. Revealed price preference: theory and empirical analysis. *The Review of Economic Studies*, 90(2):707–743, 2023.

[169] C. P. Chambers, C. Liu, and J. Rehbeck. Costly information acquisition. *Journal of Economic Theory*, 186:104979, 2020.

[170] F. Echenique, S. Lee, and M. Shum. The money pump as a measure of revealed preference violations. *Journal of Political Economy*, 119(6):1201–1223, 2011.

[171] M. Dean and D. Martin. Measuring rationality with the minimum cost of revealed preference violations. *Review of Economics and Statistics*, 98(3):524–534, 2016.

[172] A. Fostel, H. E. Scarf, and M. J. Todd. Two new proofs of Afriat's theorem. *Economic Theory*, 24(1):211–219, 2004.

[173] K. Pattanayak, V. Krishnamurthy, and C. M. Berry. Meta-cognitive radar. masking cognition from an inverse reinforcement learner. *IEEE Transactions on Aerospace and Electronic Systems*, pages 1–18, 2023.

[174] R. Deb. Interdependent preferences, potential games and household consumption. MPRA Paper 6818, University Library of Munich, Germany, Jan 2008.

[175] R. Deb. A testable model of consumption with externalities. *Journal of Economic Theory*, 144(4):1804–1816, 2009.

[176] I. Crawford. Habits revealed. *The Review of Economic Studies*, 77(4):1382–1402, 2010.

[177] C. Sims. Rational inattention and monetary economics. *Handbook of Monetary Economics*, 3:155–181, 2010.

[178] F. Matějka and A. McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–98, 2015.

[179] H. De Oliveira, T. Denti, M. Mihm, and K. Ozbek. Rationally inattentive preferences and hidden information costs. *Theoretical Economics*, 12(2):621–654, 2017.

[180] S. Nitinawarat and V. V. Veeravalli. Controlled sensing for sequential multihypothesis testing with controlled markovian observations and non-uniform control cost. *Sequential Analysis*, 34(1):1–24, 2015.

[181] V. Krishnamurthy and H. V. Poor. A tutorial on interactive sensing in social networks. *IEEE Transactions on Computational Social Systems*, 1(1):3–21, 2014.

[182] A. Lucas, M. Iliadis, R. Molina, and A. K. Katsaggelos. Using deep neural networks for inverse problems in imaging: beyond analytical methods. *IEEE Signal Processing Magazine*, 35(1):20–36, 2018.

[183] D. Liang, J. Cheng, Z. Ke, and L. Ying. Deep magnetic resonance image reconstruction: Inverse problems meet neural networks. *IEEE Signal Processing Magazine*, 37(1):141–151, 2020.

[184] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

[185] M. T. McCann, K. H. Jin, and M. Unser. Convolutional neural networks for inverse problems in imaging: A review. *IEEE Signal Processing Magazine*, 34(6):85–95, 2017.

[186] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[187] W. E. Diewert. Afriat and revealed preference theory. *The Review of Economic Studies*, 40(3):419–425, 1973.

[188] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, et al. Interpretability of deep learning models: a survey of results. In *2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*, pages 1–6. IEEE, 2017.

[189] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[190] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):1–42, 2018.

[191] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019.

[192] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. ImageNet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[193] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[194] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[195] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *arXiv preprint arXiv:1605.09304*, 2016.

[196] P. Hase, C. Chen, O. Li, and C. Rudin. Interpretable image recognition with hierarchical prototypes. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7-1, pages 32–40, 2019.

[197] T. Lei, R. Barzilay, and T. Jaakkola. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*, 2016.

[198] S. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[199] M. T. Ribeiro, S. Singh, and C. Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[200] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. Not just a black box:

Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

[201] H. Wang and D.-Y. Yeung. Towards Bayesian deep learning: A framework and some existing methods. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3395–3408, 2016.

[202] P. Milgrom. Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, 12(2):380–391, 1981.

[203] L. Huang and H. Liu. Rational inattention and portfolio selection. *The Journal of Finance*, 62(4):1999–2040, 2007.

[204] S. E. Mirsadeghi, A. Royat, and H. Rezatofighi. Unsupervised image segmentation by mutual information maximization and adversarial regularization. *IEEE Robotics and Automation Letters*, 2021.

[205] W. Hoiles, V. Krishnamurthy, and K. Pattanayak. Rationally Inattentive Inverse Reinforcement Learning Explains YouTube commenting behavior. *The Journal of Machine Learning Research*, 21(170):1–39, 2020.

[206] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

[207] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[208] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[209] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[210] G. Hinton, N. Srivastava, and K. Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8), 2012.

[211] R. Anand, T. Shanthi, R. Sabeenian, and S. Veni. Real time noisy dataset implementation of optical character identification using CNN. *International Journal of Intelligent Enterprise*, 7(1-3):67–80, 2020.

[212] Y. Bengio. *Learning deep architectures for AI*. Now Publishers Inc, 2009.

[213] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[214] P. J. Reny. A characterization of rationalizable consumer behavior. *Econometrica*, 83(1):175–192, 2015.

[215] K. V. Mishra, M. B. Shankar, and B. Ottersten. Toward metacognitive radars: Concept and applications. In *2020 IEEE International Radar Conference (RADAR)*, pages 77–82. IEEE, 2020.

[216] X. Wang, Z. Fei, J. A. Zhang, J. Huang, and J. Yuan. Constrained utility maximization in dual-functional radar-communication multi-uav networks. *IEEE Transactions on Communications*, 69(4):2660–2672, 2020.

[217] A. F. Martone, K. D. Sherbondy, J. A. Kovarskiy, B. H. Kirk, R. M. Narayanan, C. E. Thornton, R. M. Buehrer, J. W. Owen, B. Ravenscroft, S. Blunt, et al. Closing the loop on cognitive radar for spectrum sharing. *IEEE Aerospace and Electronic Systems Magazine*, 36(9):44–55, 2021.

[218] A. F. Martone. Cognitive radar demystified. *URSI Radio Science Bulletin*, 2014(350):10–22, 2014.

[219] L. Zhao and D. P. Palomar. Maximin joint optimization of transmitting code and receiving filter in radar and communications. *IEEE Transactions on Signal Processing*, 65(4):850–863, 2016.

[220] L. Wu, P. Babu, and D. P. Palomar. Cognitive radar-based sequence design via SINR maximization. *IEEE Transactions on Signal Processing*, 65(3):779–793, 2016.

[221] A. F. Martone, K. D. Sherbondy, J. A. Kovarskiy, B. H. Kirk, J. W. Owen, B. Ravenscroft, A. Egbert, A. Goad, A. Dockendorf, C. E. Thornton, et al. Practical aspects of cognitive radar. In *2020 IEEE Radar Conference (RadarConf20)*, pages 1–6. IEEE, 2020.

[222] S. D. Blunt, J. K. Jakabosky, C. A. Mohr, P. M. McCormick, J. W. Owen, B. Ravenscroft, C. Sahin, G. D. Zook, C. C. Jones, J. G. Metcalf, et al. Principles and applications of random fm radar waveform design. *IEEE Aerospace and Electronic Systems Magazine*, 35(10):20–28, 2020.

[223] B. Ravenscroft, J. W. Owen, J. Jakabosky, S. D. Blunt, A. F. Martone, and K. D. Sherbondy. Experimental demonstration and analysis of cognitive spectrum sensing and notching for radar. *IET Radar, Sonar & Navigation*, 12(12):1466–1475, 2018.

[224] A. Aubry, A. De Maio, M. Piezzo, M. M. Naghsh, M. Soltanalian, and P. Stoica. Cognitive radar waveform design for spectral coexistence in signal-dependent interference. In *2014 IEEE radar conference*, pages 0474–0478. IEEE, 2014.

[225] M. Alaee-Kerahroodi, E. Raei, S. Kumar, and B. S. Mysore Rama Rao. Cognitive radar waveform design and prototype for coexistence with communications. *IEEE Sensors Journal*, 22(10):9787–9802, 2022.

[226] H. Esmaeili-Najafabadi, H. Leung, and P. W. Moo. Unimodular waveform design with desired ambiguity function for cognitive radar. *IEEE Transactions on Aerospace and Electronic Systems*, 56(3):2489–2496, 2019.

[227] M. R. Bell. Information theory and radar waveform design. *IEEE Transactions on Information Theory*, 39(5):1578–1597, 1993.

[228] R. Romero and N. A. Goodman. Information-theoretic matched waveform in signal dependent interference. In *2008 IEEE Radar Conference*, pages 1–6. IEEE, 2008.

[229] N. A. Goodman, P. R. Venkata, and M. A. Neifeld. Adaptive waveform design and sequential hypothesis testing for target recognition with active sensors. *IEEE Journal of Selected Topics in Signal Processing*, 1(1):105–113, 2007.

[230] H. He, P. Stoica, and J. Li. Waveform design with stopband and correlation constraints for cognitive radar. In *2010 2nd International Workshop on Cognitive Information Processing*, pages 344–349. IEEE, 2010.

[231] W. Melvin, M. Wicks, P. Antonik, Y. Salama, P. Li, and H. Schuman. Knowledge-based space-time adaptive processing for airborne early warning radar. *IEEE Aerospace and Electronic Systems Magazine*, 13(4):37–42, 1998.

[232] W. W. Howard, A. F. Martone, and R. M. Buehrer. Distributed online learning for coexistence in cognitive radar networks. *IEEE Transactions on Aerospace and Electronic Systems*, 2022.

[233] S. Maleki, S. Chatzinotas, B. Evans, K. Liolis, J. Grotz, A. Vanelli-Coralli, and

N. Chuberre. Cognitive spectrum utilization in ka band multibeam satellite communications. *IEEE Communications Magazine*, 53(3):24–29, 2015.

[234] P. Chavali and A. Nehorai. Scheduling and power allocation in a cognitive radar network for multiple-target tracking. *IEEE Transactions on Signal Processing*, 60(2):715–729, 2012.

[235] J. Li, L. Xu, P. Stoica, K. W. Forsythe, and D. W. Bliss. Range compression and waveform optimization for MIMO radar: A cramér–rao bound based study. *IEEE Transactions on Signal Processing*, 56(1):218–232, 2007.

[236] K. L. Bell, C. J. Baker, G. E. Smith, J. T. 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.

[237] M. Hernandez, T. Kirubarajan, and Y. Bar-Shalom. Multisensor resource deployment using posterior cramér-rao bounds. *IEEE Transactions on Aerospace and Electronic Systems*, 40(2):399–416, 2004.

[238] V. C. Vannicola and J. A. Mineo. Applications of knowledge based systems to surveillance. In *Proceedings of the 1988 IEEE National Radar Conference*, pages 157–164. IEEE, 1988.

[239] R. A. Romero and N. A. Goodman. Cognitive radar network: Cooperative adaptive beamsteering for integrated search-and-track application. *IEEE Transactions on Aerospace and Electronic Systems*, 49(2):915–931, 2013.

[240] R. A. Romero and N. A. Goodman. Adaptive beamsteering for search-and-track application with cognitive radar network. In *2011 IEEE RadarCon (RADAR)*, pages 1091–1095. IEEE, 2011.

[241] A. Aubry, A. De Maio, and M. Govoni. Two-dimensional spectrum sensing for cognitive radar. In *2018 IEEE Radar Conference (RadarConf18)*, pages 0815–0820. IEEE, 2018.

[242] A. Aubry, V. Carotenuto, A. De Maio, and M. A. Govoni. Multi-snapshot spectrum sensing for cognitive radar via block-sparsity exploitation. *IEEE Transactions on Signal Processing*, 67(6):1396–1406, 2018.

[243] P. Setlur and M. Rangaswamy. Waveform design for radar stap in signal dependent interference. *IEEE Transactions on Signal Processing*, 64(1):19–34, 2015.

[244] E. Raei, M. Alaee-Kerahroodi, and M. B. Shankar. ADMM based transmit waveform and receive filter design in cognitive radar systems. In *2020 IEEE radar conference (RadarConf20)*, pages 1–6. IEEE, 2020.

[245] G. Rossetti and S. Lambotharan. Waveform optimization techniques for bi-static cognitive radars. In *2016 IEEE 12th International Colloquium on Signal Processing & Its Applications (CSPA)*, pages 115–118. IEEE, 2016.

[246] G. Rossetti, A. Deligiannis, and S. Lambotharan. Waveform design and receiver filter optimization for multistatic cognitive radar. In *2016 IEEE Radar Conference (RadarConf)*, pages 1–5. IEEE, 2016.

[247] G. Rossetti and S. Lambotharan. Coordinated waveform design and receiver filter optimization for cognitive radar networks. In *2016 IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 1–5. IEEE, 2016.

[248] J. Guerci, J. Bergin, R. Guerci, M. Khanin, and M. Rangaswamy. A new MIMO clutter model for cognitive radar. In *2016 IEEE Radar Conference (RadarConf)*, pages 1–6. IEEE, 2016.

[249] 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.

[250] 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.

[251] J. A. Boyd, D. B. Harris, D. D. King, and H. Welch Jr. Electronic countermeasures. *Electronic Countermeasures*, 1978.

[252] L. Snow, V. Krishnamurthy, and B. M. Sadler. Identifying coordination in a cognitive radar network-a multi-objective inverse reinforcement learning approach. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[253] 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.

[254] F. A. Butt, I. H. Naqvi, and U. Riaz. Hybrid phased-MIMO radar: A novel approach with optimal performance under electronic countermeasures. *IEEE Communications Letters*, 22(6):1184–1187, 2018.

[255] S. Gong, X. Wei, and X. Li. ECCM scheme against interrupted sampling repeater jammer based on time-frequency analysis. *Journal of Systems Engineering and Electronics*, 25(6):996–1003, 2014.

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

[257] 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 29-1, 2015.

[258] A. Roth, J. Ullman, and Z. S. Wu. Watch and learn: Optimizing from revealed preferences feedback. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 949–962, 2016.

[259] Y. Sakuma, T. P. Tran, T. Iwai, A. Nishikawa, and H. Nishi. Trajectory anonymization through laplace noise addition in latent space. In *2021 Ninth International Symposium on Computing and Networking (CANDAR)*, pages 65–73, 2021.

[260] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

[261] R. Shokri. Privacy games: Optimal user-centric data obfuscation. *Proceedings on Privacy Enhancing Technologies*, 2015(2):1–17, 2015.

[262] G. Beigi, A. Mosallanezhad, R. Guo, H. Alvari, A. Nou, and H. Liu. Privacy-aware recommendation with private-attribute protection using adversarial learning. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 34–42, 2020.

[263] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan. *Estimation with applications to tracking and navigation*. John Wiley, New York, 2008.

[264] X. R. Li and V. P. Jilkov. Survey of maneuvering target tracking. part i. dynamic models. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4):1333–1364, 2003.

[265] S. Blackman and R. Popoli. *Design and Analysis of Modern Tracking Systems*. Artech House, 1999.

[266] S. Haykin. Cognitive radar: a way of the future. *IEEE signal processing magazine*, 23(1):30–40, 2006.

[267] K. L. Bell, C. J. Baker, G. E. Smith, J. T. 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.

[268] 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.

[269] F. Smits, A. Huizing, W. van Rossum, and P. Hiemstra. A cognitive radar network: Architecture and application to multiplatform radar management. In *2008 European Radar Conference*, pages 312–315. IEEE, 2008.

[270] M. Kozy, J. Yu, R. M. Buehrer, A. Martone, and K. Sherbondy. Applying deep-q networks to target tracking to improve cognitive radar. In *2019 IEEE Radar Conference (RadarConf)*, pages 1–6. IEEE, 2019.

[271] C. E. Thornton, R. M. Buehrer, A. F. Martone, and K. D. Sherbondy. Experimental analysis of reinforcement learning techniques for spectrum sharing radar. In *2020 IEEE International Radar Conference (RADAR)*, pages 67–72. IEEE, 2020.

[272] E. Blasch, I. Kadar, J. Salerno, M. M. Kokar, S. Das, G. M. Powell, D. D. Corkill, and E. H. Ruspini. Issues and challenges of knowledge representation and reasoning methods in situation assessment (level 2 fusion). In *Signal Processing, Sensor Fusion, and Target Recognition XV*, volume 6235, page 623510. International Society for Optics and Photonics, 2006.

[273] J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418, 1969.

[274] J. Li, P. Stoica, and Z. Wang. On robust capon beamforming and diagonal loading. *IEEE transactions on signal processing*, 51(7):1702–1715, 2003.

[275] J. Li, P. Stoica, and Z. Wang. Doubly constrained robust capon beamformer. *IEEE Transactions on Signal Processing*, 52(9):2407–2423, 2004.

[276] P. Stoica, Z. Wang, and J. Li. Robust capon beamforming. In *Conference Record of*

the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 2002., volume 1, pages 876–880. IEEE, 2002.

[277] A. Stringer, G. Dolinger, D. Hogue, L. Schley, and J. G. Metcalf. A meta-cognitive approach to adaptive radar detection. *IEEE Transactions on Aerospace and Electronic Systems*, 2023.

[278] G. T. Capraro and M. C. Wicks. Metacognition for waveform diverse radar. In *2012 International Waveform Diversity & Design Conference (WDD)*, pages 348–351. IEEE, 2012.

[279] A. F. Martone, K. D. Sherbondy, J. A. Kovarskiy, B. H. Kirk, C. E. Thornton, J. W. Owen, B. Ravenscroft, A. Egbert, A. Goad, A. Dockendorf, et al. Metacognition for radar coexistence. In *2020 IEEE International Radar Conference (RADAR)*, pages 55–60. IEEE, 2020.

[280] 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.

[281] 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.

[282] D. Schleher. LPI radar: fact or fiction. *IEEE Aerospace and Electronic Systems Magazine*, 21(5):3–6, 2006.

[283] P. E. Pace. *Detecting and classifying low probability of intercept radar*. Artech house, 2009.

[284] 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.

[285] 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.

[286] Z. Zhang, S. Salous, H. Li, and Y. Tian. Optimal coordination method of opportunistic array radars for multi-target-tracking-based radio frequency stealth in clutter. *Radio Science*, 50(11):1187–1196, 2015.

[287] L. Neng-Jing. Radar ECCMs new area: anti-stealth and anti-ARM. *IEEE Transactions on Aerospace and Electronic Systems*, 31(3):1120–1127, 1995.

[288] M. Arik and O. B. Akan. Enabling cognition on electronic countermeasure systems against next-generation radars. In *MILCOM 2015-2015 IEEE Military Communications Conference*, pages 1103–1108. IEEE, 2015.

[289] A. Charlish, F. Hoffmann, C. Degen, and I. Schlangen. The development from adaptive to cognitive radar resource management. *IEEE Aerospace and Electronic Systems Magazine*, 35(6):8–19, 2020.

[290] A. Gupta and V. Krishnamurthy. Principal–agent problem as a principled approach to electronic counter-countermeasures in radar. *IEEE Transactions on Aerospace and Electronic Systems*, 58(4):3223–3235, 2022.

[291] S. Jain, K. Pattanayak, V. Krishnamurthy, and C. Berry. Adaptive eccm for mitigating smart jammers. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[292] Y. Yang and R. S. Blum. MIMO radar waveform design based on mutual information and minimum mean-square error estimation. *IEEE Transactions on Aerospace and electronic systems*, 43(1):330–343, 2007.

[293] F. Gini, A. De Maio, and L. Patton. *Waveform design and diversity for advanced radar systems*. Institution of engineering and technology London, UK, 2012.

[294] F. Liu, L. Zhou, C. Masouros, A. Li, W. Luo, and A. Petropulu. Toward dual-functional radar-communication systems: Optimal waveform design. *IEEE Transactions on Signal Processing*, 66(16):4264–4279, 2018.

[295] Z. Wei, Z. Liu, B. Peng, and R. Shen. ECCM scheme against interrupted sampling repeater jammer based on parameter-adjusted waveform design. *Sensors*, 18(4):1141, 2018.

[296] Y. Liu, G. Liao, Z. Yang, and J. Xu. Multiobjective optimal waveform design for OFDM integrated radar and communication systems. *Signal Processing*, 141:331–342, 2017.

[297] E. Grossi and M. Lops. MIMO radar waveform design: a divergence-based approach for sequential and fixed-sample size tests. In *3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, pages 165–168, 2009.

[298] V. Krishnamurthy and R. Evans. Hidden Markov model multi-arm bandits: A methodology for beam scheduling in multi-target tracking. *IEEE Transactions on Signal Processing*, 49(12):2893–2908, December 2001.

[299] M. Xie, W. Yi, L. Kong, and T. Kirubarajan. Receive-beam resource allocation for multiple target tracking with distributed MIMO radars. *IEEE Transactions on Aerospace and Electronic Systems*, 54(5):2421–2436, 2018.

[300] J. Wang and H. Zhu. Beam allocation and performance evaluation in switched-beam based massive MIMO systems. In *2015 IEEE International Conference on Communications (ICC)*, pages 2387–2392. IEEE, 2015.

[301] J. Wang, H. Zhu, L. Dai, N. J. Gomes, and J. Wang. Low-complexity beam allocation for switched-beam based multiuser massive MIMO systems. *IEEE Transactions on Wireless Communications*, 15(12):8236–8248, 2016.

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

[303] M. Dean and D. Martin. Measuring rationality with the minimum cost of revealed preference violations. *Review of Economics and Statistics*, 98(3):524–534, 2016.

[304] F. Echenique, S. Lee, and M. Shum. The money pump as a measure of revealed preference violations. *Journal of Political Economy*, 119(6):1201–1223, 2011.

[305] M. Dean and D. Martin. Measuring rationality with the minimum cost of revealed preference violations. *Review of Economics and Statistics*, 98(3):524–534, 2016.

[306] B. Smeulders, F. C. Spieksma, L. Cherchye, and B. De Rock. Goodness-of-fit measures for revealed preference tests: Complexity results and algorithms. *ACM Transactions on Economics and Computation (TEAC)*, 2(1):1–16, 2014.

[307] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[308] H. V. Trees. *Detection, Estimation and Modulation Theory*. John Wiley & Sons, 1968.

[309] J. Spall. *Introduction to Stochastic Search and Optimization*. Wiley, 2003.

[310] I.-J. Wang and J. C. Spall. A constrained simultaneous perturbation stochastic approximation algorithm based on penalty functions. In *Proceedings of the 1999*

American Control Conference (Cat. No. 99CH36251), volume 1, pages 393–399. IEEE, 1999.

[311] K. Pattanayak, V. Krishnamurthy, and C. Berry. Meta-cognition. an inverse-inverse reinforcement learning approach for cognitive radars. In *2022 25th International Conference on Information Fusion (FUSION)*, pages 01–08. IEEE, 2022.

[312] K. Pattanayak, V. Krishnamurthy, and C. Berry. Inverse-inverse reinforcement learning. how to hide strategy from an adversarial inverse reinforcement learner. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 3631–3636. IEEE, 2022.

[313] A. Nemirovski and A. Shapiro. Scenario approximations of chance constraints. *Probabilistic and randomized methods for design under uncertainty*, pages 3–47, 2006.

[314] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

[315] H.-G. Beyer and B. Sendhoff. Robust optimization–a comprehensive survey. *Computer methods in applied mechanics and engineering*, 196(33-34):3190–3218, 2007.

[316] K. Pattanayak, V. Krishnamurthy, and E. Blasch. Inverse sequential hypothesis testing. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–7. IEEE, 2020.

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

[318] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[319] J. Akhtar. Orthogonal block coded ECCM schemes against repeat radar jammers. *IEEE Transactions on Aerospace and Electronic Systems*, 45(3):1218–1226, 2009.

[320] M. Soumekh. SAR-ECCM using phase-perturbed LFM chirp signals and DRFM repeat jammer penalization. In *IEEE International Radar Conference, 2005.*, pages 507–512. IEEE, 2005.

[321] D. S. Garmatyuk and R. M. Narayanan. ECCM capabilities of an ultrawideband

bandlimited random noise imaging radar. *IEEE Transactions on Aerospace and Electronic Systems*, 38(4):1243–1255, 2002.