

Boston University

OpenBU

<http://open.bu.edu>

Boston University Theses & Dissertations

Boston University Theses & Dissertations

2022

Inverse rational inattention

<https://hdl.handle.net/2144/44472>

"Downloaded from OpenBU. Boston University's institutional repository."

BOSTON UNIVERSITY
QUESTROM SCHOOL OF BUSINESS

Dissertation

INVERSE RATIONAL INATTENTION

by

ZEYU ZHU

B.A., Fudan University, 2011
M.S., New York University, 2015

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2022

© 2022 by
ZEYU ZHU
All rights reserved

Approved by

First Reader

Hao Xing, Ph.D.
Associate Professor of Finance

Second Reader

A. Max Reppen, Ph.D.
Assistant Professor of Finance

Third Reader

Steven Kou, Ph.D.
Questrom Professor in Management
Professor of Finance

INVERSE RATIONAL INATTENTION

ZEYU ZHU

Boston University, Questrom School of Business, 2022

Major Professor: Hao Xing, Associate Professor of Finance

ABSTRACT

In this thesis, we consider a rational inattentive agent who does not observe the environment perfectly and needs to acquire costly signal to make decisions. By observing agents actions, we formulate the inverse rational inattention framework to recover agents utility.

We formulate problems both in static and dynamic settings. In the static setting, we show the recovered utility is unique in equivalent classes. We propose efficient algorithms and show their convergence. We apply the model and algorithm to robo-advising problems of recovering investors utilities by observing their investment strategies in both mean-variance and target date investment settings.

CONTENTS

Abstract	iv
List of Figures	ix
List of Symbols and Abbreviations	x
1 Introduction	1
1.1 Contributions	2
1.2 Financial applications	3
1.3 Reinforcement Learning (RL) Versus Rational Inattention (RI)	5
1.4 Inverse Rational Inattention (IRI)	6
1.5 Organization of this thesis	7
2 Backgrounds	9
2.1 Decision making framework	9
2.1.1 Markov Decision Process	9
2.1.2 Choice rule/policy	10
2.1.3 Entropy & Information cost	10
2.2 Reinforcement Learning	13
2.2.1 Setup	13
2.2.2 Classic Reinforcement Learning	13
2.2.3 RL strategies	14
2.2.4 Maximum entropy RL	15
2.3 Inverse Reinforcement Learning	17

2.4	Rational Inattention	19
2.4.1	Signal-based formulation	19
2.4.2	An equivalent formulation	20
2.4.3	Dynamic RI problem	23
2.4.4	Connections between RI and RL	27
2.5	Regime Switching Model	30
2.5.1	Markov regime-switching log-normal model	31
2.5.2	Hidden Markov Model	32
3	Static inverse rational inattention problem	34
3.1	Setup	34
3.2	Formulation	34
3.3	Connection to Generative Adversarial Networks(GANs)	36
4	Solution for Static IRI problem	38
4.1	Equivalent class of utility functions	38
4.2	Interior expert policy	40
4.3	Boundary expert policy	41
5	Dynamic inverse rational inattention problem	44
5.1	Setup	44
5.2	Dynamic IRI problem	44
5.3	A Special Case	45
6	Algorithms & Experiments	49
6.1	Generate expert observations	49
6.1.1	RI algorithms	49

6.1.2	sub-optimal expert policy	50
6.2	Static IRI	50
6.2.1	Example with interior (p_E, q_E)	52
6.2.2	Example with Boundary (p_E, q_E)	54
6.2.3	sub-optimal policy	55
6.3	Dynamic IRI	58
6.3.1	Dynamic example with sub-optimal policy	59
6.4	Convergence Analysis	61
6.4.1	Setup	62
7	Financial Applications	67
7.1	Robo-advising	67
7.2	Environment Setup	68
7.2.1	State	68
7.2.2	Action	70
7.3	Mean-Variance Optimization	70
7.3.1	True Preference and policy	71
7.3.2	Recovered results	71
7.4	Target date Investment	72
7.4.1	Dynamics of priors	75
7.4.2	Recovered results	77
8	Open questions	80
8.1	Static and Continuous RI problem	80
8.2	Static and Continuous Inverse RI Problem	81
8.2.1	Algorithm	82

9	Conclusions	86
9.1	Inverse rational inattention	86
9.2	Connection to reinforcement learning	86
9.3	Applications on robo-advising	87
A	Proof	88
A.1	Proofs for Chapter 3	88
A.2	Proofs for Chapter 4	90
A.3	Proofs for Chapter 5	94
A.4	Proofs for Chapter 8	94
B	Supplementary results	97
B.1	Mean-variance example	97
B.1.1	Impact of marginal risk aversion θ	97
B.1.2	impact of marginal information cost λ	98
	Bibliography	100
	Curriculum Vitae	104

LIST OF FIGURES

1.1	Fully observed states	6
1.2	Not fully observed states	6
2.1	Reinforcement Learning	14
2.2	Trajectory	17
2.3	Hidden Markov Model	32
6.1	Relative Performance with imperfect observations (Static)	58
6.2	MSE Performance with imperfect observations (Static)	59
6.3	Relative Performance with imperfect observations (Dynamic)	61
6.4	MSE Performance with imperfect observations (Dynamic)	62
7.1	Dynamics of investor's prior distribution	76
7.2	Relative performance of target date investment	78
7.3	MSE performance of target date investment	79

LIST OF ABBREVIATIONS

IRI	Inverse Rational Inattention
IRL	Inverse Reinforcement Learning
RI	Rational Inattention
RL	Reinforcement Learning

CHAPTER 1

Introduction

People are making decisions all the time, e.g., what to take for breakfast, where to go for a vacation, which financial security to invest in, etc. Realizing it or not, our choices, to a great extent, are governed by our preference and our understanding of the current environment. For example, compared to a risk-seeking investor, a risk-averse investor is usually more likely to choose a conservative portfolio, while both of them could be more aggressive when they are in a bull market.

For economic decision making, an agent's preference is numerically represented by his utility function which maps different choices to a real-valued number. Given the utility function, optimal strategies or choice rules are determined by decision makers by solving some optimization problems. In the real world, however, specifying a utility function for a decision maker is quite challenging. It is difficult to specify an individual's preference which can explain their behaviors.

Inverse reinforcement learning, a rapidly developing machine learning framework, aims to recover agents' utility functions via their behavior demonstrations (Ng & Russell (2000), Ziebart et al. (2008), Fu et al. (2018)). Demonstrations are usually easier to access than the utility functions because people can often make choices and complete tasks efficiently without being aware of their preferences.

Inverse reinforcement learning is often used to learn utility functions for robots that will be interacting with a changing environment. For example, Shimosaka et al. (2016) applies inverse reinforcement learning to learn the reward functions for controlling cars from expert demonstrations. Also, Yu & Zhao (2019) uses Bayesian IRL to learn the reward functions of doctors managing mechanical ventilation of patients and their sedation while being ventilated.

One common assumption of the reinforcement learning framework is fully-observability which means the agent can fully observe his current state. Unfortunately, financial decision makers are often uncertain about their current environment, they thus need to acquire more information about the environment before taking real actions. For example, an investor may find it hard to specify the current market condition, so other additional information like macroeconomic indicators or an analysis of the government's policy is required to help him identify the current market condition. However, those information is usually not free, so decision makers need to balance the potential gain and cost of information acquisition.

Rational inattention (RI) is a decision making framework introduced by Sims (1998, 2003) to study the optimal choice rule under costly information acquisition. Compared to RL agents, an RI agent cannot observe his current environment perfectly, instead the agent can firstly acquires signals related to the environment, then make decisions based on the received signals.

In this thesis, we consider a rational inattentive agent who acquires costly signal to make decisions. By observing agents actions, we formulate the inverse rational inattention problem to recover agents utility.

1.1 CONTRIBUTIONS

The main contributions of this thesis are as follows:

- We consider a rational inattentive agent who acquires costly signal to make decisions, we formulate an inverse rational inattention problem to recover the agents utility function directly from his past behaviors both in static and dynamic settings.

- We provide a connection between rational inattention and reinforcement learning communities. One hand, this work extends the existing learning from demonstration methods from fully-observable states to partially-observable states; on the other hand, this work provides an inverse model under the rational inattention framework such that the preferences of rational inattention agents can be recovered from their past behaviors.
- For the static problem, we characterize the equivalent class of the recovered utility and we also identify its uniqueness.
- We propose efficient algorithms for both static and dynamic problems and prove their convergence. To evaluate our algorithms, we provide experiments with different types of input behaviors.
- As financial applications, we apply our model and algorithm to robo-advising problems with a mean-variance preference and a target date investment. In both applications, our model can recover the target utilities very well.

1.2 FINANCIAL APPLICATIONS

Robo-advisors are digital platforms that provide automated, algorithm-driven investment services with little to no human supervision. Robo-advisors can efficiently provide financial advice for clients from any level of wealth with extremely low fees (Rossi & Utkus (2019)), meanwhile, since they are based on automated algorithms, Robo-advisors can avoid the cognitive bias of human advisors (Foerster et al. (2017)).

One most common method used by many Robo-advising firms to access their clients' preference is online questionnaire, based on the collected answers, they

estimate the clients' risk preference and provide specific financial advice.

However, this questionnaire-based method have many shortcomings as discussed in Abraham et al. (2019), D'Acunto & Rossi (2020), Kaya (2017), one problem is that those one-size-fit-all questionnaires might be too simple to collect individual-specific information to serve individual client. Another problem is that people may understand the same question in different ways, for example, one person's understanding of 'high risk aversion' may be much less averse than that of another person. Thus, even if two people presented exactly the same answers to a questionnaire, their preferences could still be distinct. The third problem comes from the response bias, which means that answers provided by someone for the questionnaire do not match his behavior in reality, this may be caused by many reasons, such as the lack of concentration or emotional effect.

Meanwhile, as documented by Rossi & Utkus (2019), investors have an average of 14 years investment experience before signing up Vanguard robo-advising service, such a long history gives us the opportunities to learn from their choices to identify individual specific utilities. Therefore, we proposed a new approach to recover investors' preferences directly from their past trading behaviors.

We applied our work to robo-advising problems under two settings:

- Mean-Variance optimization: the investor has a mean-variance preference and maximizes the cumulative utility at the end of one period.
- Target investment: the investor has a power utility and maximizes the terminal utility at the end of multiple periods.

For each application, our model recover the utility function that can explain the observed behaviors. The recovered utility function can provide us a better under-

standing of the investor's preference, and further used to improve the investment performance.

1.3 REINFORCEMENT LEARNING (RL) VERSUS RATIONAL INATTENTION (RI)

Both reinforcement learning and rational inattention are decision making frameworks where an agent makes choices based on his current situation and get an reward/utility as the feedback from the environment. By interacting with the environment, as shown in Figure 1.1, the agent gradually learns how to behave optimally in order to maximize his utility.

One important feature of rational inattention is that an RI agent cannot observe the current state but some signals, then the agent undertakes actions depending on realized signal as shown in Figure 1.2. The rational inattention agent therefore is looking for the optimal choice rule to maximize the expected utility net information acquisition cost.

Whether a decision maker can observe his current situation or not is usually a case-by-case problem in practice. Caplin & Dean (2015) study when actions can be explained by costly information acquisition. They provide two conditions:

- A no improving action switches (NIAS) condition ensures that choices are optimal given current knowledge of states;
- A no improving attention cycles (NIAC) condition ensures that total utility cannot be improved by reassigning information structures.

They prove that an observed choice data can be explained by a costly information representation if and only if NIAS and NAIC are satisfied.

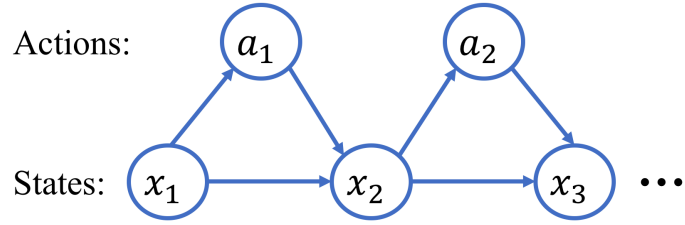


Figure 1.1: Fully observed states

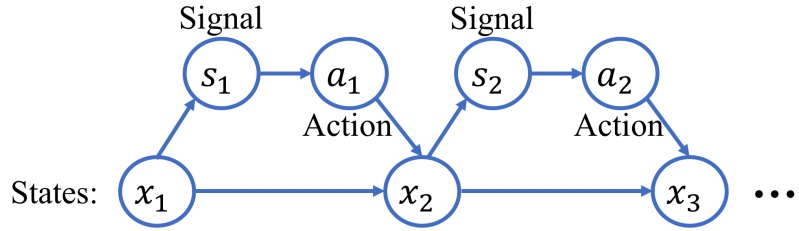


Figure 1.2: Not fully observed states

1.4 INVERSE RATIONAL INATTENTION (IRI)

Many inverse reinforcement learning (IRL) literatures have been proved to be able to successfully recover agents' preferences from their behaviors when current states are fully observable. For financial problems, however, it is not always the case. For example, when an investor is making investment decision based on the current market condition, it would be challenging for the investor, especially individual investor to tell which stage is the current market, thus the information cost coming from acquiring related signals should be taken into consideration.

When there is information cost, the observed behaviors are no longer the choices that maximize the expected utilities but compromised choices between expected utilities and information cost. Motivated by the inverse reinforcement learning, we propose the inverse rational inattention (IRI) problem in this work to recover utilities of rational inattentive agents who need to acquire costly information related to states before making decisions.

1.5 ORGANIZATION OF THIS THESIS

The organization of this thesis is as follows:

- **Part I: Preliminaries**

- **Chapter 1:** Motivations for the inverse rational inattention framework, discussions of existing methods and limitations; summary of major contributions.
- **Chapter 2:** Reviews of related literatures from reinforcement learning and rational inattention; discussion of the connection between reinforcement learning and rational inattention.

- **Part II: Main Results**

- **Chapter 3:** Formulations of static inverse rational inattention problem; discussion of the connection to Generative Adversarial Networks(GANs).
- **Chapter 4:** Definition of equivalent class of utility functions; discussion of solutions for static inverse rational inattention problem with interior or boundary expert policy.
- **Chapter 5:** Formulations of dynamic inverse rational inattention problem; discussion of a special case of the dynamic setting.
- **Chapter 6:** Algorithms and experiments for static IRI problem and dynamic IRI problem; proof of the convergence of Static IRI algorithm.

- **Part III: Financial Applications**

- **Chapter 7:** Applications of IRI model on robo-advising problems with a mean-variance preference setting and a target date investment setting; discussion of recovered results and impact of different parameters.
- **Part IV: Conclusions**
 - **Chapter 8:** Several possible future extensions of the current IRI model; potential algorithm for continuous space problem and its connection to GANs.
 - **Chapter 9:** A concluding summary of all theoretical and numerical results of this thesis.

CHAPTER 2

Backgrounds

Our inverse rational inattention framework is built on two major stands of literatures, reinforcement learning and rational inattention. In this chapter, we firstly present some fundamental concepts of decision making problem. Then we review existing works from both RL and RI communities, and discuss their connections and differences. In the end, we discuss some regime estimation methods used in specifying market regimes.

On the one hand, our work formulates an inverse problem to learn agents' preferences directly from their past behaviors, which can be applied to financial applications like robo-advising; on the other hand, although RL and RI have a lot in common, the RL and control communities remain practically disjoint Recht (2019), in this sense our work also provides a bridge between reinforcement learning and rational inattention.

2.1 DECISION MAKING FRAMEWORK

2.1.1 Markov Decision Process

Markov Decision Process (MDP) is a discrete-time stochastic control model widely used for decision making. Most reinforcement learning problems are set up in the form of MDP (Rust (1994) and Puterman (2005)).

Definition 2.1.1 (Markov Decision Process). A Markov Decision Process is a tuple $\mathcal{M}_{MDP} = (X, A, \pi(x' | x, a), u(x, a))$, where:

- X is the state space;
- A is the action space;

- $\pi(x' | x, a)$ is the transition probability that choosing action a at current state x will lead to the next state x' ;
- $u(x, a)$ is the utility or reward of choosing the action a at the state x .

2.1.2 Choice rule/policy

In the setting of reinforcement learning or static rational inattention, a choice rule $p \in \Delta(A | X)$, is a conditional distribution mapping from state space X to action space A . A stationary choice rule $p(a|x)$ means that the probability of choosing action a at state x depends only the current state x instead of time.

For dynamic rational inattention in Markovian setting, as defined in Steiner et al. (2017) and Miao & Xing (2019), a choice rule \mathbf{p} is a sequence of conditional distributions:

$$\mathbf{p} = \{p_t(a_t | x_t, a_{t-1}) \in \Delta(A | X_t \times A_{t-1}) : \text{all } (x_t, a_{t-1}), 1 \leq t \leq T\}. \quad (2.1)$$

The probability of choosing action a_t depends not only on the current state x_t but also time t and the last action a_{t-1} .

2.1.3 Entropy & Information cost

2.1.3.1 Entropy

Entropy is a measure of uncertainty of a random variable, and the information cost in rational inattention framework is modeled via the Shannon entropy-based information.

Definition 2.1.2 (Shannon Entropy). The entropy $H(X)$ of a discrete random vari-

able X is defined by

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x) \quad (2.2)$$

and if two random variables $(X, Y) \sim p(x, y)$, the conditional entropy $H(Y | X)$ is defined as

$$\begin{aligned} H(Y | X) &= - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y | x) \log p(y | x) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y | x) \end{aligned} \quad (2.3)$$

Proposition 2.1.1 (Lemma 2.1.1 in Cover & Thomas (2012)). The entropy $H(X)$ is always non-negative.

Definition 2.1.3 (Relative Entropy). The relative entropy or KL-divergence between two distributions p and q is defined as:

$$D_{KL}(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \quad (2.4)$$

Relative entropy is a measure of the distance between two distributions, and for the above definition, we use the convention $0 \log \frac{0}{0} = 0$, and the convention $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$ for any $p, q > 0$.

Proposition 2.1.2 (Theorem 2.6.3 in Cover & Thomas (2012)). $D_{KL}(p||q) \geq 0$ with equality if and only if $p = q$.

Another important concept from information theory is mutual information, which measures the amount information shared by two random variables.

Definition 2.1.4. The mutual information of two variables $(X, Y) \sim p(x, y)$ is defined by

$$I(X; Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \quad (2.5)$$

and the conditional mutual information of (X, Y) given Z is defined by

$$I(X; Y | Z) = E_{p(x,y,z)} \log \frac{p(X, Y | Z)}{p(X | Z)p(Y | Z)}. \quad (2.6)$$

Proposition 2.1.3 (Theorem 2.4.1 in Cover & Thomas (2012)).

$$\begin{aligned} I(X; Y) &= H(X) - H(X | Y) \\ &= H(Y) - H(Y | X) \\ &= H(X) + H(Y) - H(X, Y) \\ &= H(X, Y) - H(X | Y) - H(Y | X) \end{aligned} \quad (2.7)$$

2.1.3.2 Information cost

Consider a decision maker who cannot observe the current state but some signals instead, the original prior belief of the decision maker over different states is given by $\mu_0 \in \Delta(X)$, then after acquiring a signal s , the new information will reshape his belief to a posterior distribution $\mu(\cdot | s) \in \Delta(X|S)$.

The more valuable a signal s is, the less uncertainty the posterior belief will be; And if a signal is independent of the state, the prior and the posterior belief will be the same. Valuable information leads to less uncertainty and is also more costly, thus, the cost of acquiring signal or information can be measure by the uncertainty reduction (Matejka & McKay (2015)):

$$H(\mu_0) - \mathbf{E}_s[H(\mu(\cdot | s))]$$

2.2 REINFORCEMENT LEARNING

2.2.1 Setup

Consider a finite state space X , a finite action space A , $\Delta(X)$ is the set of probability distributions on X , $\mu_0 \in \Delta(X)$ is a given prior distribution, we assume that $\mu_0(x) > 0$ for all $x \in X$.

The decision maker is rewarded with flow utilities that depend on current state and action, the utility function is a bounded function $u : X \times A \rightarrow \mathbf{R}$. A choice rule (or policy) $p \in \Delta(A | X)$ represents the conditional probability of selecting each action under a given state, i.e., $p(a|x)$ is the conditional probability of choosing action a at the state x .

For $\mu \in \Delta(X)$, $H(\mu) = -\sum_x \mu(x) \ln \mu(x)$ is the entropy of $\mu(\cdot)$ which is a concave and non-negative function on $\Delta(X)$.

2.2.2 Classic Reinforcement Learning

Reinforcement learning (RL) is a subarea of machine learning, which studies how to map situations to actions in order to maximize certain cumulative reward signals. As discussed in Sutton & Barto (2018), an RL agent has a clear goal, can sense the aspects of the environment and take actions to influence the environment.¹

Unlike the supervised learning, the RL agent will not be provided with examples of correct choices, but be able to interact with the environment and learn from his experience about which action yields the most reward. And unlike unsupervised learning, the RL agent is trying to maximize the cumulative reward signals instead of discovering hidden structure from unlabeled data.

¹Image taken from Figure 1 in Sutton & Barto (2018)

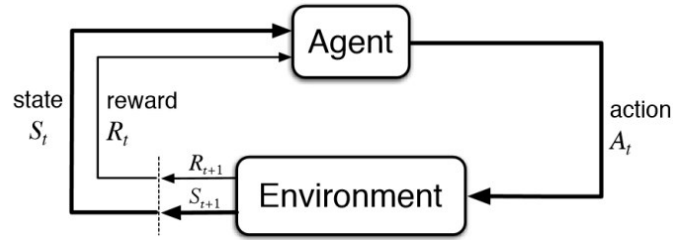


Figure 2.1: Reinforcement Learning

Given a state space X , an action space A , a state transition kernel $\pi(x'|x, a) = \Pr \{x_t = x' \mid x_{t-1} = x, a_{t-1} = a\}$, and a reward function $r(x, a) : X \times A \rightarrow \mathbf{R}$. The reinforcement learning problem is formulated as:

$$\max_{p \in \Delta(A|X)} \mathbf{E}_p \left[\sum_{t=0}^T \beta^t r(x_t, a_t) \right] \quad (2.8)$$

where $\beta \in (0, 1)$ is a discounting factor.

As discussed in Recht (2019), RL problem can be viewed as a special case of the classic optimal control problem:

$$\begin{aligned} & \text{maximize} && \mathbb{E} \left[\sum_{t=0}^T \beta^t r_t(x_t, a_t) \right] \\ & \text{subject to} && x_{t+1} = \pi_t(x_t, a_t), a_t = p_t(x^t, a^{t-1}) \end{aligned} \quad (2.9)$$

where x^t, a^t is the history of states and actions up to time t . For RL problem, the transition kernel π_t and control policy p_t are usually assumed to be stationary and action a_t depends only on current state, i.e. $\pi_t \equiv \pi$ and $p_t \equiv p(\cdot|x_t)$.

2.2.3 RL strategies

Given the transition kernel, RL problem can be solved as a standard optimization problem, if the transition kernel is unknown, model-free RL is trying to learn a map from state to action directly. Two major approaches to solve model-free RL

problem are value-function approach (or approximate dynamic programming in control community) and direct policy search (Sutton & Barto (2018), Recht (2019)).

The value-function approach uses Bellman's principle of optimality to estimate the so-called Q-function and the value-function with previous observations, and it solves the problem with techniques from dynamic programming. The direct policy approach, on the other hand, is trying to find the policy function directly, which further turns into a gradient-based method like policy gradient or a gradient-free method like REINFORCE (Williams (1992)).

To extend the above methods to high-dimensional or continuous space, we can use the neural network to approximate the Q-function and the value-function in a value-function approach or the policy function in a policy search approach.

2.2.4 Maximum entropy RL

One important feature of reinforcement learning is exploration, which means instead of taking actions which currently yields highest current reward, the agent also chooses other actions which lead to better understanding of the environment and better choice of actions in the future.

Even though the optimal policy for classic reinforcement learning problem under full observability is always deterministic which can be viewed as a special case of (2.9), a stochastic policy is preferred for exploration. Also, Ziebart (2010) showed that the maximizing-entropy policies are robust to the model and estimation errors.

Haarnoja et al. (2017) proposed the Maximum entropy reinforcement learning problem which augmented the reward with an entropy term to obtain stochastic

policies. The problem is formulated as

$$RL(r) = \max_{p \in \Delta(A|X)} \mathbf{E}_p \left[\sum_{t=0}^T \beta^t r(x_t, a_t) \right] + \lambda \mathcal{H}(p) \quad (2.10)$$

When the state and action spaces are finite, the expectation \mathbf{E}_p above can be written as

$$\mathbf{E}_p \left[\sum_{t=0}^T \beta^t r(x_t, a_t) \right] = \sum_{t=0}^T \beta^t \sum_{a_t, x_t} \mu_t(x_t) p(a_t | x_t) r(x_t, a_t)$$

where the state distribution μ_t at time t is given recursively by

$$\mu_t(x_t) = \sum_{x_{t-1}, a_{t-1}} \pi(x_t | x_{t-1}, a_{t-1}) p(a_{t-1} | x_{t-1}) \mu_{t-1}(x_{t-1}), \quad t \geq 1,$$

with given prior μ_0 .

In (2.10), the discounted casual entropy of the policy p is

$$\mathcal{H}(p) \equiv \mathbf{E}_p \left[- \sum_{t=0}^T \beta^t \log p(a_t | x_t) \right] = - \sum_{t=0}^T \beta^t \sum_{a_t, x_t} \mu_t(x_t) p(a_t | x_t) \log p(a_t | x_t).$$

Ziebart (2010) and Haarnoja et al. (2017) proved that the optimal policy of problem (2.10) has general energy-based form $p(\mathbf{a}_t | \mathbf{x}_t) \propto \exp(-\mathcal{E}(\mathbf{x}_t, \mathbf{a}_t))$, and can be extended to high dimensional or continuous space by approximating the energy function $\mathcal{E}(\mathbf{x}_t, \mathbf{a}_t)$ with deep neural networks; Haarnoja et al. (2018b) provided an off-policy actor-critic algorithm for the maximizing entropy framework, which has substantial improvement in sample efficiency and stability.

2.3 INVERSE REINFORCEMENT LEARNING

The problem of deriving a reward function from observed behaviors is referred to inverse reinforcement learning (Ng & Russell (2000)).

Consider demonstrations $\mathcal{D}_E = \{\tau_1, \dots, \tau_N\}$ collected from n decision maker, where each τ_i is a trajectory consist of state-action pairs. The goal of the IRL problem is to learn the utility function $u_\theta(x, a)$, parameterized by θ , that can explain the observed demonstrations \mathcal{D}_E .

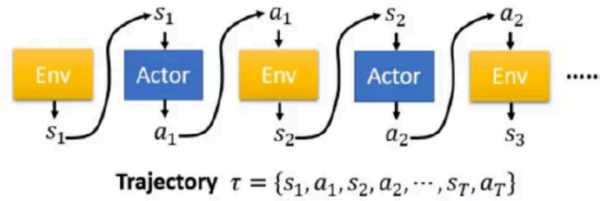


Figure 2.2: Trajectory

There are several formulations for IRL problems, Ratliff et al. (2006) casts the problem as one of structured maximum margin prediction problems which learns the utility function via the the structured margin method; Abbeel & Ng (2004) assume the agent has the linear utility function in some features and solve the IRL problem by matching feature expectations between learner's behavior and observed demonstrations.

However, the matching of feature counts is ambiguous because many different policies can lead to the same feature counts. When sub-optimal policies are observed, mixtures of policies are needed to match the feature counts, and again many different mixtures can satisfy the matching.

To deal with this ambiguity, Ziebart et al. (2008) proposed the Maximum Entropy Inverse Reinforcement learning (MaxEnt IRL), by assuming the optimal pol-

icity has an exponential form

$$P(\tau_i | \theta) = \frac{1}{Z(\theta)} e^{r_\theta(\tau_i)} = \frac{1}{Z(\theta)} e^{\theta^\top \mathbf{f}_{\tau_i}}$$

the IRL problem is then solved by employing the principle of maximum entropy which is equivalent to maximizing the likelihood of the observed behavior:

$$\max_{\theta} E_{\tau \sim \mathcal{D}_E} [\log p_{\theta}(\tau)] \quad (2.11)$$

As a result, the model provides probabilistic support for all possible behaviors and non-zero probability for demonstrated behavior.

The partition function Z introduces heavy computational burden for the maximum entropy framework, especially in the high-dimensional or continuous space. Building on the maximum entropy framework, Finn et al. (2016a) use an adaptive sampling distribution to estimate the partition function Z and approximate the reward function in the high-dimensional domains via neural networks; Finn et al. (2016b) presented that MaxEnt IRL is an energy-based model and sample-based MaxEnt-IRL algorithms are mathematically equivalent to generative adversarial networks (Goodfellow et al. (2014)).

Ho & Ermon (2016) proposed a model-free imitation learning algorithm which is also connected to generative adversarial networks (GANs):

$$\max_{c \in \mathcal{C}} \left(\min_{p \in \Delta(A|S)} -H(p) + \mathbf{E}_p[c(s, a)] \right) - \mathbf{E}_{p_E}[c(s, a)], \quad (2.12)$$

where c is the cost function and p_E is empirical distribution from D_E .

2.4 RATIONAL INATTENTION

One common assumption of most RL problems is full observability, which means the agent can observe the current state accurately without any cost. In practice, however, decision makers may be uncertain about their current environment and needs to acquire costly information before making decision.

Rational inattention (RI) is a decision making framework introduced by Sims (1998, 2003) to study the optimal choice rule under costly information acquisition. An RI agent needs to first design a signal structure about the states and then undertakes actions depending on the realized signal. The agent determines the signal structure and the action rule to maximize the expected utility net information acquisition cost.

2.4.1 Signal-based formulation

Consider a finite state space X , action space A , and a signal space S , The decision process for an RI agent has two stages (see e.g. Matejka & McKay (2015)):

In the first stage, the decision maker with prior belief $\mu_0 \in \Delta(X)$ designs an information strategy $f : X \rightarrow \Delta(S)$, where $f(s|x)$ is the probability of observing signal realization s given the state realization x . The information strategy f solves

$$\max_{f \in \Delta(S|X)} \sum_x \sum_s V(\mu(\cdot|s)) f(s|x) \mu_0(x) - \lambda I(f), \quad (2.13)$$

where $\mu(\cdot|s)$ is the conditional distribution of state given signal s given by

$$\mu(x|s) = \frac{f(s|x) \mu_0(x)}{\sum_{x'} f(s|x') \mu_0(x')}.$$

In (2.13), the first term is the expected payoff associated to the belief $\mu(\cdot | s)$ and the parameter λ is the marginal cost of uncertainty reduction.

The uncertainty reduction $I(f)$ is given by

$$I(f) \equiv H(\mu_0) - \mathbf{E}[H(\mu(\cdot | s))], \quad (2.14)$$

which measures the difference between the entropy of the prior and the expected entropy of posterior. After observing an informative signal, the decision maker may have a more accurate estimation of the state described by the posterior which has a lower entropy comparing to the entropy of the prior. More reduction of the uncertainty means the signal is more informative, hence introducing a larger information cost in (2.13).

In the second stage, after the signal s is observed and the posterior belief $\mu(\cdot | s)$ is formed, the decision maker chooses an action strategy $\sigma : S \rightarrow A$ to maximize the expected payoff:

$$V(\mu(\cdot | s)) \equiv \max_{\sigma} \mathbb{E}_{\mu(\cdot | s)} [u(x, a)] = \max_{\sigma} \sum_x u(x, \sigma(s)) \mu(x | s) \quad (2.15)$$

2.4.2 An equivalent formulation

Receiving distinct signals that lead to the same action is inefficient, because such signals have no impact on the action choice. It would save information cost by combining all signals which lead to the same action, let $S_a \equiv \{\mathbf{s} \in \mathbb{R}^N : a(F(\mathbf{x} | \mathbf{s})) = a\}$ and define

$$p(a|x) \equiv \sum_{s \in S_a} f(s|x),$$

to be the probability of action a when the state realization is x , and define

$$q(a) \equiv \sum_x p(a|x)\mu_0(x).$$

as the unconditional probability of taking action a . We call $p(\cdot|x)$ the choice rule and $q(\cdot)$ the default rule.

The following lemma transforms the two stages problem (2.13) and (2.15) to an equivalent formulation:

Lemma 2.4.1 (Matejka & McKay (2015), Lemma 1). The optimal value of the static rational inattention problem (2.13) is the same as the optimal value of the following problem:

$$\max_p \mathbf{E}_p[u(x, a)] - \lambda I(x, a) = \max_p \sum_{x,a} p(a|x)\mu(x) \left[u(x, a) - \lambda \ln \frac{p(a|x)}{q(a)} \right], \quad (2.16)$$

subject to the constraint

$$q(a) = \sum_x p(a|x)\mu(x). \quad (2.17)$$

Now, we denote

$$F(p, q) \equiv \sum_{x,a} p(a|x)\mu(x) \left[u(x, a) - \lambda \ln \frac{p(a|x)}{q(a)} \right].$$

F is jointly concave in p and q , and it follows from Blahut (1972, Theorem 4) that, for fixed $p \in \Delta(A|X)$, $\max_{q \in \Delta(A)} F(p, q)$ is given by $q(a) = \sum_x \mu(x)p(a|x)$. Therefore the problem (2.16) is equivalent to the following convex optimization problem:

$$\max_{p \in \Delta(A|X), q \in \Delta(A)} F(p, q) = \max_{p \in \Delta(A|X), q \in \Delta(A)} \sum_{x,a} p(a|x)\mu(x) \left[u(x, a) - \lambda \ln \frac{p(a|x)}{q(a)} \right]. \quad (2.18)$$

Compared to the formulation (2.16) and (2.17), problem (2.18) introduces another control variable q and the resulting optimization problem is jointly concave in p and q .

To see the connection between (2.18) and the signal approach in section 2.4.1, the choice rule $p(a|x)$ is the probability of signal s_a under which the action a is taken once s_a is observed, i.e. for $\forall a \in A$, we have the information strategy $f(s_a|x) = p(a|x)$, the action strategy $\sigma(a) = a$, and the posterior distribution $\mu(x|s_a) = \frac{p(a|x)\mu(x)}{q(a)}$ which is posterior of state once the signal s_a is observed.

The optimal policy for (2.18) is reported below.

Proposition 2.4.1. (Matejka & McKay (2015), Theorem 1) When $\lambda > 0$, the optimal choice rule for (2.18) is

$$p(a|x) = \frac{q(a) \exp(u(x, a)/\lambda)}{\sum_{a'} q(a') \exp(u(x, a')/\lambda)}. \quad (2.19)$$

and

$$q(a) = \sum_x p(a|x)\mu(x). \quad (2.20)$$

When $\lambda = 0$, then the decision maker selects the action(s) with the highest payoff with probability 1.

The optimal choice rule in (2.19) also has a energy-based form, compared to the optimal policy in maximum entropy reinforcement learning, $p(a|x)$ in (2.19) is weighted by the exponential utility and the endogenously determined default rule $q(a)$ together.

2.4.3 Dynamic RI problem

For dynamic case with horizon $T < \infty$, the decision maker takes flow utilities which depend on current state and action only (Steiner et al. (2017), Miao & Xing (2019)), and the utility function is time-independent, given by $u : X \times A \rightarrow \mathbf{R}$, and we may also have a terminal utility $u_{T+1} : X_{T+1} \rightarrow \mathbf{R}$.

The choice rule \mathbf{p} under this setting is a sequence of conditional distributions:

$$\mathbf{p} = \{p_t(a_t | x_t, a_{t-1}) \in \Delta(A | X_t \times A_{t-1}) : \text{all } (x_t, a_{t-1}), 1 \leq t \leq T\} \quad (2.21)$$

Here we take the Markovian choice rule as an approximation of the choice rules that depend on the entire history of states and actions, for each period the choice a_t depends only on current state x_t and last action a_{t-1} .

Here we take the Markovian choice rule as an approximation of the choice rules that depends on the entire history of states and actions, for each period, the choice rule p_t depends only on current state x_t and last action a_{t-1} .

The joint prior distribution of the state and the action for each period is denoted by $\mu_t(x_t, a_{t-1})$, which is uniquely determined by a prior distribution $\mu_1(x_1)$, a transition kernel $\pi(x_{t+1} | x_t, a_t)$ and a choice rule \mathbf{p} by a recursive formula (see also Equation 1 in Miao & Xing (2019)):

$$\mu_{t+1}(x_{t+1}, a_t) = \sum_{x_t, a_{t-1}} \pi(x_{t+1} | x_t, a_t) p_t(a_t | x_t, a_{t-1}) \mu_t(x_t, a_{t-1}) \quad \text{for } t \geq 1 \quad (2.22)$$

The choice rule \mathbf{p} generates flow utility, and with $\beta \in (0, 1)$, the cumulative

discounted expected utility is given by:

$$J(\mathbf{p}) = \mathbf{E}_{\{x_t\}_{t=1}^T \sim \pi, \{a_t\}_{t=1}^T \sim \mathbf{p}} \left[\sum_{t=1}^T \beta^{t-1} u(x_t, a_t) + \beta^T u_{T+1}(x_{T+1}) \right] \quad (2.23)$$

Prior to choosing an action in each period t , the decision maker acquires costly information with entropy-based cost:

$$I_\beta(\mathbf{x}^T \rightarrow \mathbf{a}^T; \mathbf{p}) = \sum_{t=1}^T \beta^{t-1} I(\mathbf{x}_t; \mathbf{a}_t \mid \mathbf{a}_{t-1}) \quad (2.24)$$

The conditional mutual information $I(\mathbf{x}_t; \mathbf{a}_t \mid \mathbf{a}_{t-1})$ is defined as

$$I(\mathbf{x}_t; \mathbf{a}_t \mid \mathbf{a}_{t-1}) = H(\mu_t(\cdot \mid \cdot)) - H(\mu_t(\cdot \mid \cdot, a_{t-1})) = H(q(\cdot \mid a_{t-1})) - H(p(\cdot \mid \cdot, a_{t-1}))$$

where $H(\cdot)$ is the Shannon entropy. Therefore $I(\mathbf{x}_t; \mathbf{a}_t \mid \mathbf{a}_{t-1})$ measures the reduction of uncertainty about state x_t after obtaining more information.

Then, the dynamic RI problem can be formulated as:

$$\max_{\mathbf{p} \in \Pi} J(\mathbf{p}) - \lambda I_\beta(\mathbf{x}^T \rightarrow \mathbf{a}^T; \mathbf{p}) \quad (2.25)$$

where $\Pi \equiv \prod_{t=1}^T \Delta(A \mid X_t \times A_{t-1})$ and $\lambda > 0$.

To illustrate solution of problem (2.25), we consider a two-period case ($T = 2$) with $u_{T+1} = 0$, which is then formulated by:

$$\begin{aligned} & \sum_{a_1, x_1} \mu_1(x_1) p_1(a_1 \mid x_1) \left[u(x_1, a_1) - \ln \frac{p_1(a_1 \mid x_1)}{q_1(a_1)} \right] \\ & + \beta \sum_{a_1, a_2, x_2} \mu_2(x_2, a_1) p_2(a_2 \mid x_2, a_1) \left[u(x_2, a_2) - \ln \frac{p_2(a_2 \mid x_2, a_1)}{q_2(a_2 \mid a_1)} \right] \end{aligned} \quad (2.26)$$

where

$$q_1(a_1) = \sum p_1(a_1 | x_1) \mu_1(x_1), \quad (2.27)$$

$$q_2(a_2|a_1) = \sum_{x_2} p_2(a_2 | x_2, a_1) \mu_2(x_2|a_1), \quad (2.28)$$

$$\mu_2(x_2, a_1) = \sum_{x_1} \pi(x_2 | x_1, a_1) p_1(a_1 | x_1) \mu_1(x_1). \quad (2.29)$$

Miao & Xing (2019) reduces the dynamic RI problem to a collection of static problems using the Bellman equation, following their method, the dynamic RI problem (2.26) is equivalent to the following control problem:

$$V_1(\mu_1) \equiv \max_{p_1, p_2, q_1, q_2} F(p_1, p_2, q_1, q_2), \quad (2.30)$$

where $p_1 \in \Delta(A | X)$, $p_2 \in \Delta(A | X \times A)$, $q_1 \in \Delta(A)$, $q_2 \in \Delta(A | A)$, and

$$\begin{aligned} F(p_1, p_2, q_1, q_2) = & \sum_{a_1, x_1} \mu_1(x_1) p_1(a_1 | x_1) \left[u(x_1, a_1) - \lambda \ln \frac{p_1(a_1 | x_1)}{q_1(a_1)} \right] \\ & + \beta \sum_{a_1, a_2, x_2} \mu_2(x_2, a_1) p_2(a_2 | x_2, a_1) \left[u(x_2, a_2) - \lambda \ln \frac{p_2(a_2 | x_2, a_1)}{q_2(a_2 | a_1)} \right]. \end{aligned} \quad (2.31)$$

At period 2, we solve the problem

$$W_2(\mu_2) \equiv \sum_{a_1} q_1(a_1) \widetilde{W}_2(a_1),$$

where

$$\widetilde{W}_2(a_1) \equiv \max_{p_2, q_2} \sum_{a_2, x_2} \mu_2(x_2 | a_1) p_2(a_2 | x_2, a_1) \left[u(x_2, a_2) - \lambda \ln \frac{p_2(a_2 | x_2, a_1)}{q_2(a_2 | a_1)} \right], \quad (2.32)$$

which is equivalent to a static RI problem with prior belief $\mu_2(\cdot | a_1)$.

By the Proposition 2 in Miao & Xing (2019), the value function satisfies

$$\widetilde{W}_2(a_1) = \sum_{x_2} \mu_2(x_2 | a_1) \widetilde{V}_2(x_2, a_1) \quad (2.33)$$

where

$$\widetilde{V}_2(x_2, a_1) = \lambda \ln \sum_{a_2} q_2(a_2 | a_1) \exp[u(x_2, a_2) / \lambda] \quad (2.34)$$

By dynamic programming, at the period 1, we solve the problem

$$\sum_{a_1, x_1} \mu_1(x_1) p_1(a_1 | x_1) \left[u(x_1, a_1) + \sum_{x_2} \pi(x_2 | x_1, a_1) \widetilde{V}_2(x_2, a_1) - \lambda \ln \frac{p_1(a_1 | x_1)}{q_1(a_1)} \right] \quad (2.35)$$

which is equivalent to a static RI problem with exogenous payoff

$$v_1(x_1, a_1) \equiv u(x_1, a_1) + \sum_{x_2} \pi(x_2 | x_1, a_1) \widetilde{V}_2(x_2, a_1) \quad (2.36)$$

Necessary and sufficient conditions for Markovian solutions of (2.25) are provided in Proposition 6 and 7 in Miao & Xing (2019). The optimal solution for the dynamic RI problem is given by the following Proposition.

Proposition 2.4.2 (Proposition 6 in Miao & Xing (2019)). Let $\beta \in (0, 1)$. Then the Markovian solution to the dynamic RI problem in (2.25), $\{p_t(a_t | x_t, a_{t-1})\}_{t=1}^T$ and $\{q_t(a_t | a_{t-1})\}_{t=1}^T$, is characterized by the following system of difference equations for $t = 1, 2, \dots, T$:

$$p_t(a_t | x_t, a_{t-1}) = \frac{q_t(a_t | a_{t-1}) \exp(v_t(x_t, a_t) / \lambda)}{\sum_{a'_t} q_t(a'_t | a_{t-1}) \exp(v_t(x_t, a'_t) / \lambda)} \text{ for } \mu_t(x_t, a_{t-1}) > 0 \quad (2.37)$$

$$q_t(a_t | a_{t-1}) = \sum_{x_t} p_t(a_t | x_t, a_{t-1}) \mu_t(x_t | a_{t-1}) \text{ for } \mu_t(a_{t-1}) > 0 \quad (2.38)$$

where

$$v_t(x_t, a_t) = u(x_t, a_t) + \beta \sum_{x_{t+1}} \pi(x_{t+1} | x_t, a_t) \tilde{V}_{t+1}(x_{t+1}, a_t), \quad (2.39)$$

$$\tilde{V}_t(x_t, a_{t-1}) = \lambda \ln \sum_{a_t} q_t(a_t | a_{t-1}) \exp(v_t(x_t, a_t) / \lambda), \quad (2.40)$$

$$\mu_{t+1}(x_{t+1}, a_t) = \sum_{x_t, a_{t-1}} \pi(x_{t+1} | x_t, a_t) p_t(a_t | x_t, a_{t-1}) \mu_t(x_t, a_{t-1}), \quad (2.41)$$

$$\mu_t(x_t | a_{t-1}) = \frac{\mu_t(x_t, a_{t-1})}{\mu_t(a_{t-1})} \text{ for } \mu_t(a_{t-1}) = \sum_{x_t} \mu_t(x_t, a_{t-1}) > 0. \quad (2.42)$$

2.4.4 Connections between RI and RL

Both Markovian RI and RL are decision-making frameworks built on Markov Decision Processes (MDPs) where the agents make choices based on the current state and aim to maximize the expected reward/utility, the resulting optimal policies have a similar energy-based form.

In our work, we try to provide a connection between rational inattention and reinforcement learning communities, find their similarities and differences, discuss the possibilities that each can learn from the other.

2.4.4.1 Entropy term

Both (2.18) and (2.10) have an entropy term in their formulations, which further leads to the similar form of the optimal policies, but the entropy term is actually a major difference between RL and RI problems.

The entropy term in (2.10) is introduced to encourage the stochastic policy and exploration while the entropy term in (2.18) is used to measure information cost.

Fox et al. (2016) also use the entropy to measure the information cost, the target

problem of their work is formulated as

$$\max_p \sum_{x,a} p(a|x)\mu(x) \left[u(x,a) - \lambda \ln \frac{p(a|x)}{q_{fix}(a)} \right]. \quad (2.43)$$

The difference is that the entropy term in (2.43) penalizes the divergence of policy p from a default policy q_{fix} which is fixed and exogenous. While the default policy q in (2.18), as explained in 2.4.1 and 2.4.2, is determined endogenously by the information cost, and thus q in (2.18) is a control variable that introduces the trade-off between expected utility and information cost.

2.4.4.2 Utility/Reward function

Both utility function u and reward function r are functions mapping from a state-action pair (x, a) to a real-valued number.

Reinforcement learning interprets the reward as a feedback from the environment, every time the agent choose action a at state x , he takes the $r(x, a)$ as a feedback, then by interacting with the environment the agent will gradually learn how to behave optimally.

In economics, utility function is a numerical representation of the agent's preference over a set of choices, and if we further extend this preference to different states, the utility $u(x, a)$ measures the preference of an agent to choose action a at state x .

Reward function in RL problem can be manually designed in some cases to obtain desired behavior (Ng et al.,1999), for example, we can assign a large negative number to a certain action as a penalty so that the agent will try to avoid this action when making choices; while utility function in RI problem can represents some endogenous features of the agent, like the risk attitude of an investor.

In most works, utility functions are assumed to have some parametric form such as mean-variance utility $u(x, a) = \text{return}(x, a) - \gamma \text{Variance}(x, a)$ with risk aversion parameter γ . In finite and discrete state and action space, we can use a matrix form:

$$u(x, a) = \begin{pmatrix} u(x_1, a_1) & u(x_1, a_2) & \dots & u(x_1, a_n) \\ u(x_2, a_1) & u(x_2, a_2) & \dots & u(x_2, a_n) \\ \vdots & \vdots & \ddots & \vdots \\ u(x_m, a_1) & u(x_m, a_2) & \dots & u(x_m, a_n) \end{pmatrix}$$

to model utility function. Therefore, utility functions in our setting is non-parametric.

2.4.4.3 Policy

Many existing works in rational inattention (Matejka & McKay (2015), Steiner et al. (2017), Miao & Xing (2019)) have shown that the optimal choice rule in both static and dynamic settings has a multinomial-logit or Maxwell-Boltzmann distribution,

$$p_t^*(a_t | x_t, a_{t-1}) \propto q_t^*(a_t | a_{t-1}) \exp(u_t(x_t, a_t) / \lambda).$$

The same form of optimal policy, also called maximum entropy stochastic policy was also derived in reinforcement learning literatures such as Ziebart et al. (2008), Toussaint (2009), Fox et al. (2016), etc. The mathematical connection between the logit distribution and entropy was derived in Jaynes (1957) and Shannon (1959).

The policy in reinforcement learning is a stationary policy where the agent's choice depends only on the current state and is time-independent. While the choice

rule in dynamic RI problem as presented in (2.21), is a sequence of conditional distributions, i.e. for the same state, the agent may react differently at different periods. Also, the action a_t in dynamic RI problem depends not only on the current state x_t but also on the last action a_{t-1} , Miao & Xing (2019) model the dynamic choice rule in a more general setting, they allow the action a_t depends on the entire history of states and actions up to time t . In this paper, we follow their work in a Markovian setting.

2.4.4.4 Algorithm

The optimal choice rule has an analytic solution for rational inattention problem, thus, it can also be efficiently solved by the forward-backward Arimoto-Blahut algorithm (Tanaka et al. (2018)). However, this algorithm still has some limitations.

Firstly, to solve the RI problem, we need to know the prior distribution μ_0 and transition kernel $\pi(x_{t+1}|x_t, a_t)$ while model-free RL algorithms, like Q-learning, REINFORCE and etc., are able to directly learn a map from states to actions without a known dynamics.

Another limitation comes from problem setting, so far, both static and dynamic rational inattention are defined on finite and discrete state space and action space, by contrast, RL algorithms, such as TRPO (Schulman et al. (2015)), PPO (Schulman et al. (2017)), SAC (Haarnoja et al. (2018a)), allow the reinforcement learning framework to be applied to high-dimensional or continuous space.

2.5 REGIME SWITCHING MODEL

As discussed previously, both rational inattention and reinforcement learning are built on a predefined environment which consists of all possible states, actions,

and the transition probability between different states.

We use the regime-switching model to identify the current market regime from real financial data Hardy (2001), Kim et al. (2019), Wang et al. (2020).

Hamilton (1989) and Kim and Nelson (2017) introduced the econometrics of state-space models with regime-switching, and several works Vo & Maurer (2013), Kim et al. (2019), Wang et al. (2020), etc. have shown that dynamic asset allocation based on the identification of market regimes has substantially better performance than strategies that ignore the regime.

2.5.1 Markov regime-switching log-normal model

Consider a market environment with K regimes, each regime is characterized by a set of different model parameters; regimes switch between each other randomly, and the transition probability only depends on the current state instead of the history.

In this paper, we assume the market condition has 3 regimes: x_1 (Boom), x_2 (Normal) and x_3 (Recession), and as in Hardy (2001), the market log-return under each state follows a normal distribution:

$$\log \frac{S_{t+1}}{S_t} | x \sim N(\mu_x, \sigma_x^2). \quad (2.44)$$

Thus, for this three-regime model we have 12 parameters in total:

$$\Theta = \{\mu_{x_1}, \mu_{x_2}, \mu_{x_3}, \sigma_{x_1}, \sigma_{x_2}, \sigma_{x_3}, \pi\},$$

where

$$\pi = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix}$$

is the transition probabilities.

Then regimes can be estimated by fitting this mode to real market returns via the maximum likelihood estimation method.

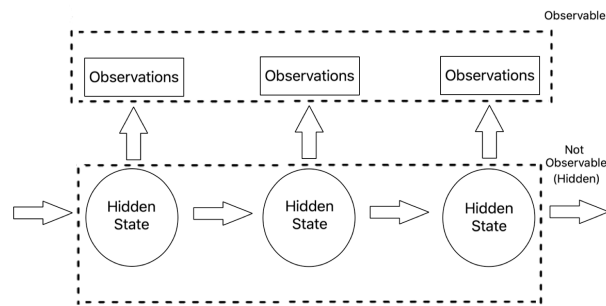


Figure 2.3: Hidden Markov Model

2.5.2 Hidden Markov Model

Another model used to detect market regime is Hidden Markov Model(HMMs) Kim et al. (2019), Wang et al. (2020). One major feature of HMMs is that the states in HMMs are hidden, we can only observe the outputs of the hidden states instead of the states themselves. For example, as an investor we cannot directly observe the market regime but we can observe its return and volatility.

As shown in Wang et al. (2020), given the number of hidden states, we assume the observations follow the Gaussian distribution, the Gaussian HMM model is able to estimate the transition probability between different hidden states, the mean and variance for observations with respect to each respective hidden state,

and an initial distribution over hidden states.²

²Image taken from Figure 1 in Wang et al. (2020)

CHAPTER 3

Static inverse rational inattention problem

3.1 SETUP

For finite space X and A , we assume the prior distribution $\mu \in \Delta(X)$ satisfies $\mu(x) > 0$ for all $x \in X$. The decision maker (or expert) with a prior belief μ_0 is endowed with a unknown utility function $u_E(x, a) : X \times A \rightarrow \mathbf{R}$. Expert choice rule is denoted as $p_E(a|x)$ where $p_E \in \Delta(A|X)$.

In practice, p_E is usually not directly accessible, the agent's behaviors are provided as demonstrations $\mathcal{D}_E = \{\tau_1, \tau_2, \dots\}$ as stated in 2.3, then p_E is obtained as the empirical conditional distribution from \mathcal{D}_E .

3.2 FORMULATION

Let p_E and q_E be the optimal solution for the static RI problem:

$$RI(u_E) \triangleq \max_{p,q} \sum_{x,a} p(a|x)\mu(x) \left(u_E(x, a) - \lambda \ln \frac{p(a|x)}{q(a)} \right). \quad (3.1)$$

Given (p_E, q_E) , we formulate the static inverse rational inattention problem as:

$$IRI(p_E, q_E) \triangleq \min_u \left\{ \max_{p,q} \left[\sum_{x,a} p(a|x)\mu(x) \left(u(x, a) - \lambda \ln \frac{p(a|x)}{q(a)} \right) \right] - \left[\sum_{x,a} p_E(a|x)\mu(x) \left(u(x, a) - \lambda \ln \frac{p_E(a|x)}{q_E(a)} \right) \right] \right\}. \quad (3.2)$$

Compared to Abbeel & Ng (2004) and Ziebart et al. (2008) where the utility function is a linear combination of some given features, or Alsabab et al. (2019) where utility function is assumed to follow a parametric form, we do not impose

any given form on the utility function u .

To understand the above min-max problem, let's first focus on the inner rational inattention problem for a given u . Because p_E and q_E may not be optimal for the inner problem, then

$$\begin{aligned} & \max_{p,q} \left[\sum_{x,a} p(a|x)\mu(x) \left(u(x,a) - \lambda \ln \frac{p(a|x)}{q(a)} \right) \right] \\ & - \left[\sum_{x,a} p_E(a|x)\mu(x) \left(u(x,a) - \lambda \ln \frac{p_E(a|x)}{q_E(a)} \right) \right] \geq 0. \end{aligned}$$

The left-hand side has the minimum value zero and zero is attained when $u = u_E$.

Now we denote

$$H(p, q) \triangleq - \sum_{x,a} \mu(x)p(a|x) \ln \frac{p(a|x)}{q(a)}. \quad (3.3)$$

where $\sum_a p(a|x) \ln \frac{p(a|x)}{q(a)}$ is the relative entropy between $p(\cdot|x)$ and $q(\cdot)$, by Theorem 2.7.2 in Cover & Thomas (2012) $H(p, q)$ is jointly concave in (p, q) .

Lemma 3.2.1. For finite action space A , $H(p_E, q_E)$ has finite value.

Proof. See Appendix A.1. □

Ignore the constant term $H(p_E, q_E)$ in (6.3) and rescale the problem by dividing λ throughout, we reformulate (6.3) as

$$IRI(p_E, q_E) \triangleq \min_u \left\{ \max_{p,q} H(p, q) + \sum_{x,a} u(x, a)\mu(x) \left(p(a|x) - p_E(a|x) \right) \right\}, \quad (3.4)$$

where, for simplicity of notation, u is the scaled utility u/λ .

Proposition 3.2.1. The optimizer for (3.4) must be (p_E, q_E) . We can view (3.4) as

the dual problem for the following constrained problem

$$\max_{p,q} H(p, q) \quad \text{subject to } p(a|x) = p_E(a|x) \quad \text{for } \forall a \in A, x \in X, \quad (3.5)$$

whose Lagrangian multiplier is $u(x, a)\mu(x)$.

Proof. See Appendix A.1. □

Given the above results, we are able to derive an analytic solution for the static IRI problem in the next chapter.

3.3 CONNECTION TO GENERATIVE ADVERSARIAL NETWORKS(GANS)

To avoid over-fitting, we introduce a convex regularizer ψ as in Finn et. al. (2016a) and Ho and Ermon (2016) into the IRI problem, then the static IRI problem with a convex regularizer $\psi : \mathbb{R}^{X \times A} \rightarrow \overline{\mathbb{R}}$ is formulated as

$$IRI_{\psi}(p_E, q_E) \triangleq \min_u \left\{ \psi(u) + \max_{p,q} \left\{ H(p, q) + \sum_{x,a} u(x, a)\mu(x) \left(p(a|x) - p_E(a|x) \right) \right\} \right\}. \quad (3.6)$$

Given a convex regularizer $\psi : \mathbb{R}^{X \times A} \rightarrow \overline{\mathbb{R}}$, we introduce its convex conjugate $\psi^* : \mathbb{R}^{X \times A} \rightarrow \overline{\mathbb{R}}$ via $\psi^*(x) = \sup_{y \in \mathbb{R}^{X \times A}} x^T y - \psi(y)$, then a similar result as the Proposition 3.2 in Ho and Ermon (2016) is shown as follows.

Proposition 3.3.1. For each optimizer u^* of (3.6), the set of optimizers of the RI associated to u^* is the same as the set of optimizers for the following problem

$$\max_{p,q} H(p, q) + \psi^*(\pi_{p_E} - \pi_p), \quad (3.7)$$

where $\pi_p(x, a) = p(a|x)\mu(x)$, for any choice rule p and $(x, a) \in X \times A$.

Proof. See Appendix A.1.

□

CHAPTER 4

Solution for Static IRI problem

4.1 EQUIVALENT CLASS OF UTILITY FUNCTIONS

Before solving the inverse rational inattention (IRI) problem, it is important to realize that different utility functions may lead to the same optimal behavior, which leads to an ambiguous problem for recovering the utility functions because the same observed behaviors could come from multiple utility functions.

To deal with this ambiguity, we define an equivalent class of utility functions and prove that utility functions from the same equivalent class lead to the same optimal choice rule.

For a utility function u and a policy (p, q) , we define the following concepts:

- Exponential utility: $v(x, a) \triangleq e^{u(x,a)/\lambda}$;
- Consideration set: $B \triangleq \{a \in A | q(a) > 0\}$, which contains all actions which are chosen with positive probabilities;
- Posterior belief: $\gamma^a(x) = \frac{p(a|x)\mu(x)}{q(a)}$, given prior μ , there is a one-to-one mapping between (p, q) and (q, γ) .

The following result from Caplin et al. (2018) provides a necessary and sufficient condition for a policy (q, γ) to be an optimal solution to the static RI problem $RI(u)$ defined by (3.1).

Proposition 4.1.1 (Proposition 2 in Caplin et al. (2018)). For a RI problem with prior distribution μ , a policy (q, γ) is optimal if and only if

(a) $\sum_{a \in A} q(a)\gamma^a(x) = \mu(x)$ for all $x \in X$;

(b) Invariant Likelihood Ratio (ILR) Equations for Chosen Options: for $a, b \in B$ and $x \in X$,

$$\frac{\gamma^a(x)}{v(x, a)} = \frac{\gamma^b(x)}{v(x, b)};$$

(c) Likelihood Ratio Inequalities for Unchosen Options: for $a \in B$ and $c \in A \setminus B$,

$$\sum_{x \in X} \frac{\gamma^a(x)}{v(x, a)} v(x, c) \leq 1.$$

As discussed in Caplin & Dean (2013) and Caplin et al. (2018), the intuition for ILR condition comes from the idea that the net utilities (utility minus information cost) of two actions have the same slope for all states, thus optimal behavior is invariant to a constant change to utility over all actions.

Motivated by Proposition 4.1.1, we propose the following definition for equivalent class of utility functions.

Definition 4.1.1. [Equivalent Class of Utility Functions] Given a prior μ and a policy (p, q) , we say two utility functions u_1 and u_2 are equivalent if for any $x \in X$, there exists a constant C_x such that

$$\begin{aligned} u_2(x, a) &= u_1(x, a) + C_x, \quad \text{for any } a \in B(q); \\ \sum_{x \in X} \frac{\gamma^a(x)}{v_1(x, a) e^{C_x/\lambda}} v_2(x, c) &\leq 1, \quad \text{for any } a \in B(q), c \in A \setminus B(q), \end{aligned}$$

where $v_i(x, a) = e^{u_i(x, a)/\lambda}$, $i = 1, 2$ and γ is the optimal posterior associated to u_1 .

For equivalent u_1 and u_2 , we denote $u_1 \sim u_2$. Given a utility function u , we define its equivalent class as $[u] = \{\tilde{u} \mid u \sim \tilde{u}\}$.

The first condition claims that if a utility function is shifted by a constant depending only on state, then the resulting utility function has the same optimal

policy as the original utility function. Similar idea was also discussed in Ng et al. (1999) as shaping rewards.

Lemma 4.1.1. For given prior μ , optimal policy (q, γ) one-to-one corresponds to the equivalent class of utility functions.

Proof. See Appendix A.2. □

Therefore, for a given expert's optimal policy (p_E, q_E) , if we find one utility function u^E that is optimal for the problem $IRI(p_E, q_E)$, any utility function in the equivalent class $[u^E]$ is also optimal for the problem $IRI(p_E, q_E)$, and by Lemma 4.1.1, this equivalent class is unique. In the next two sub-sections, we will show how to find the equivalent class $[u^E]$ for different types of (p_E, q_E) .

4.2 INTERIOR EXPERT POLICY

Definition 4.2.1 (Interior and boundary optimal policy). For an optimal solution (p_E, q_E) for $RI(u)$, we say (p_E, q_E) is an interior policy if $q_E(a) > 0$ for all $a \in A$. In other words, all actions in A are chosen with positive probabilities and the consideration set $B(q_E) = A$;

And we say (p_E, q_E) is a boundary policy if $q_E(a) = 0$ for some $a \in A$. In other words, some actions in A are not chosen and $B(q_E) \subset A$.

Suppose the observed policy (p_E, q_E) is an interior optimal policy, the following proposition claims that we can uniquely recover the equivalent class $[u^E]$.

Proposition 4.2.1. (Utility function for interior policy) For an interior optimal expert policy (p_E, q_E) , the inverse rational inattention problem (3.4) has a unique solution $[u^E]$.

Proof. See Appendix A.2. □

So far we have proved that for an interior optimal policy (p_E, q_E) , we can uniquely recover the equivalent class $[u_E]$, can we move one step further to find the exact utility function u_E . To achieve this goal, we need the help of another concept called anchor action.

Definition 4.2.2 (Anchor Action). An action $a^A \in A$ is called anchor action for utility function u , if value of $u(x, a^A)$ is known for all $x \in X$.

One natural example of anchor action is an action that provides no utility at any state, such as a firm selling no goods (Bajari et al. (2010)) or a portfolio with only risk-free asset. With the assumption of anchor action, Geng et al. (2020) develops an efficient inverse reinforcement learning algorithm that can uniquely recover the target reward function.

Corollary 4.2.1. If an anchor action a^A exist for the target utility u_E , then values of $u_E(x, a)$ for all $x \in X, a \in A$ can be uniquely recovered.

Proof. After we find the unique equivalent class $[u_E]$ and a certain utility function $u \in [u_E]$ via Proposition 4.2.1, we can compute the value of translator $C_x = u_E(x, a^A) - u(x, a^A)$ for all $x \in X$, then recover $u_E(x, a)$ for all $(x, a) \in X \times A$ by the condition 1 in Definition 4.1.1: $u_E(x, a) = u(x, a) + C_x$ \square

4.3 BOUNDARY EXPERT POLICY

For an interior policy (p_E, q_E) , since all actions are chosen with positive probability, only the first condition in 4.1.1 has effect when we recover the target utility function u_E . For boundary expert policy (p_E, q_E) , the consideration set $B(q_E)$ is not equal to the entire action space A , thus the second condition in 4.1.1 will also influence the recovered result.

In this case, we will recover utility function $u(\cdot, a)$ for $a \in B(q_E)$ and $u(\cdot, c)$ for $c \in A \setminus B(q_E)$ separately and the following Lemma claims that for an utility function u , the problem $RI(u)$ has a boundary optimal policy (p, q) , we can change the value of $u(\cdot, c)$ for $c \in A \setminus B(q_E)$ to construct a new utility function \hat{u} , then (p, q) is also an optimal solution for $RI(\hat{u})$.

Lemma 4.3.1. For a given utility function $u(x, a)$ and prior μ , if the RI problem $RI(u)$ has an optimal boundary policy (p, q) . we can construct new utility function \hat{u} by choosing the value of $\hat{u}(x, a)$ such that:

- Condition 1: $\hat{u}(x, a) = u(x, a)$, for $\forall a \in B(q_E)$;
- Condition 2: $\sum_{x \in X} \frac{\gamma^a(x)}{v(x, a)} \hat{v}(x, c) < 1$, for $\forall a \in A$ and $\forall c \in A \setminus B(q_E)$

Then, (p, q) is also optimal for $RI(\hat{u})$.

Proof. See Appendix A.2. □

The lemma above provides us a method to construct one utility function in $[u_E]$ and thus recover $[u_E]$ when (p_E, q_E) is an optimal boundary policy. The basic idea is to recover the target utility function as two parts, firstly we restrict the action set to $B(q_E)$ and recover the value of $u_E(x, a)$ on $X \times B(q_E)$ via Proposition 4.2.1; Then we choose the rest values of $u_E(x, c)$ on $X \times A \setminus B(q_E)$ such that for $\forall a \in B(q_E), c \in A \setminus B(q_E)$

$$\sum_{x \in X} \frac{\gamma^a(x)}{v_E(x, a)} v_E(x, c) < 1.$$

Proposition 4.3.1. For a boundary optimal policy (p_E, q_E) , the inverse rational inattention problem (3.4) has a unique solution $[u^E]$.

Proof. See Appendix A.2. □

For a boundary optimal policy (p_E, q_E) , if an anchor action $a^A \in B(q_E)$ exists for u_E , we can uniquely determine the values of $u^E(x, a)$ for $\forall a \in B(q_E)$, but the values of $u^E(x, c)$ for $c \in A \setminus B(q_E)$ has infinite choices, for example, we can take

$$u^E(x, c) = \min_{a \in B(q_E)} u^E(x, a) - \epsilon \quad \text{for } \forall \epsilon > 0.$$

and the resulting utility functions are belong to $[u_E]$.

CHAPTER 5

Dynamic inverse rational inattention problem

5.1 SETUP

For finite space X and A , we consider a dynamic setting with horizon $T < \infty$. The prior distribution at period 1 is given by $\mu_1 \in \Delta(X)$ which satisfies $\mu_1(x_1) > 0$ for all $x_1 \in X$. The decision maker (or expert) with a prior belief μ_1 is endowed with a unknown utility function $u_E(x, a) : X \times A \rightarrow \mathbf{R}$, which is time-independent.

An expert choice rule under this setting is a sequence of conditional distributions:

$$\mathbf{p}^E = \{p_t^E(a_t | x_t, a_{t-1}) \in \Delta(A | X_t \times A_{t-1}) : \text{all } (x_t, a_{t-1}), 1 \leq t \leq T\}. \quad (5.1)$$

To formulate the inverse problem, we assume the decision maker is making dynamic choices under the dynamic rational inattention framework (2.25). At each period t , the choice a_t depends on the current state x_t and last action a_{t-1} .

5.2 DYNAMIC IRI PROBLEM

One thing special about dynamic IRI problem is that given the observed behavior \mathbf{p}^E , a known transition kernel $\pi(x_{t+1} | x_t, a_t)$ and the prior distribution at the first period $\mu_1(x_1)$, we can obtain the dynamics of an agent's prior distributions $\{\mu_t\}_1^T$ by:

$$\mu_{t+1}(x_{t+1}, a_t) = \sum_{x_t, a_{t-1}} \pi(x_{t+1} | x_t, a_t) p_t(a_t | x_t, a_{t-1}) \mu_t(x_t, a_{t-1}), \quad \text{for } t \geq 1. \quad (5.2)$$

In Chapter 7, when we apply IRI model to robo-advising applications, we will

show how the priors evolve with time if an investor has an optimistic or a pessimistic initial belief.

Then the default rule $\mathbf{q}^E = \{q_t^E(a_t|a_{t-1}) \in \Delta(A|A_{t-1}) : \text{all } a_{t-1}, 1 \leq t \leq T\}$ is given by:

$$q_t^E(a_t | a_{t-1}) = \sum_{x_t} p_t^E(a_t | x_t, a_{t-1}) \mu_t(x_t | a_{t-1}) \text{ when } \mu_t(a_{t-1}) > 0 \quad (5.3)$$

Under the above setting, we define the dynamic inverse rational inattention problem as:

$$\begin{aligned} \min_u \psi(u) + \max_{\mathbf{p}, \mathbf{q}} \{ & H(\mathbf{p}, \mathbf{q}) + \mathbb{E}_{\{x_t\}_{t=1}^T \sim \pi, \{a_t\}_{t=1}^T \sim \mathbf{p}} \left[\sum_{t=1}^T \beta^{t-1} u(x_t, a_t) \right] \\ & - \mathbb{E}_{\{x_t\}_{t=1}^T \sim \pi, \{a_t\}_{t=1}^T \sim \mathbf{p}^E} \left[\sum_{t=1}^T \beta^{t-1} u(x_t, a_t) \right] \}, \end{aligned} \quad (5.4)$$

where

$$H(\mathbf{p}, \mathbf{q}) = -\mathbb{E}_{\{x_t\}_{t=1}^T \sim \pi, \{a_t\}_{t=1}^T \sim \mathbf{p}} \left[\sum_{t=1}^T \beta^{t-1} \ln \frac{p_t(a_t|x_t, a_{t-1})}{q(a_t|a_{t-1})} \right] \quad (5.5)$$

is the entropy-based information cost over all periods.

For this dynamic IRI problem, we propose an efficient numerical algorithm which can be applied to both optimal expert policy and sub-optimal expert policy. In Chapter 6, we present details of the algorithm and evaluate its performance under different settings.

5.3 A SPECIAL CASE

In this section, we consider a special case of the above dynamic problem (5.4), where the transition kernel π does not depend on actions, i.e. $\pi \equiv \pi(x_{t+1}|x_t)$ and

the next choice only depends on the current state instead of the past choice, i.e. $p_t \equiv p_t(a_t|x_t)$.

As an example of this special case, we can consider an individual investor making portfolio choice based on the macroeconomic condition, his choice is too negligible to change the whole market's transition and he can easily change his holding at each period based only on the current situation.

To formulate the IRI problem under this special case, we assume the agent is making choices via a corresponding dynamic RI problem (for notation simplicity we consider a two-period case), which is a special case of problem (2.25) (See also in Miao & Xing (2019)):

$$\begin{aligned} & \sum_{a_1, x_1} \mu_1(x_1) p_1(a_1 | x_1) \left[u(x_1, a_1) - \ln \frac{p_1(a_1 | x_1)}{q_1(a_1)} \right] \\ + \beta & \sum_{a_2, x_2, a_1} \mu_2(x_2, a_1) p_2(a_2 | x_2) \left[u(x_2, a_2) - \ln \frac{p_2(a_2 | x_2)}{q_2(a_2)} \right], \end{aligned} \quad (5.6)$$

where

$$q_1(a_1) = \sum p_1(a_1 | x_1) \mu_1(x_1), \quad (5.7)$$

$$q_2(a_2) = \sum_{x_2} p_2(a_2 | x_2) \mu_2(x_2), \quad (5.8)$$

$$\mu_2(x_2) = \sum_{x_1} \pi(x_2 | x_1) \mu_1(x_1), \quad (5.9)$$

$$\mu_2(x_2, a_1) = \sum_{x_1} \pi(x_2 | x_1) p_1(a_1 | x_1) \mu_1(x_1). \quad (5.10)$$

Note that in the second line of (5.6), $\mu_2(x_2, a_1)$ is the only term depends on a_1 , so it can be replaced by $\sum_{a_1} \mu_2(x_2, a_1) = \mu_2(x_2)$, then the problem (5.6) is equivalent

to

$$\begin{aligned} & \sum_{a_1, x_1} \mu_1(x_1) p_1(a_1 | x_1) \left[u(x_1, a_1) - \ln \frac{p_1(a_1 | x_1)}{q_1(a_1)} \right] \\ & + \beta \sum_{a_2, x_2} \mu_2(x_2) p_2(a_2 | x_2) \left[u(x_2, a_2) - \ln \frac{p_2(a_2 | x_2)}{q_2(a_2)} \right]. \end{aligned} \quad (5.11)$$

Because $p_1(a_1|x_1)$ does not impact $\mu_2(x_2)$ at period 2, the problem (5.11) can be viewed as two separate static RI problems:

$$V_1(\mu_1) = \sup_{p_1, q_1} \sum_{a_1, x_1} \mu_1(x_1) p_1(a_1 | x_1) \left[u(x_1, a_1) - \ln \frac{p_1(a_1 | x_1)}{q_1(a_1)} \right], \quad (5.12)$$

$$V_2(\mu_2) = \sup_{p_2, q_2} \sum_{a_2, x_2} \mu_2(x_2) p_2(a_2 | x_2) \left[u(x_2, a_2) - \ln \frac{p_2(a_2 | x_2)}{q_2(a_2)} \right], \quad (5.13)$$

and the optimal value of (5.11) is $V_1(\mu_1) + \beta V_2(\mu_2)$.

Under this special setting, we formulate the following special-case dynamic IRI problem:

$$\min_u \psi(u) + \max_{\mathbf{p}, \mathbf{q}} \left\{ H(\mathbf{p}, \mathbf{q}) + \sum_{t=1}^T \beta^{t-1} \sum_{x_t, a_t} \mu_t(x_t) u(x_t, a_t) (p_t(a_t|x_t) - p_t^E(a_t|x_t)) \right\}, \quad (5.14)$$

where

$$H(\mathbf{p}, \mathbf{q}) = - \sum_{t=1}^T \beta^{t-1} \sum_{x_t, a_t} \mu_t(x_t) p_t(a_t|x_t) \ln \frac{p_t(a_t|x_t)}{q_t(a_t)}. \quad (5.15)$$

The major difference between the general dynamic IRI (5.4) and the special case (5.14) is that in the special case the policy p_t does not impact the prior μ_t in each period, thus the entire dynamic IRI problem can be viewed as a sequence of static IRI problems with different priors $\{\mu_t\}_{t=1}^T$.

Also, since μ_t and p_t are independent of each period, the information cost $H(\mathbf{p}, \mathbf{q})$ in this special setting is jointly concave in (\mathbf{p}, \mathbf{q}) , and we can easily extend

Proposition 3.3.1 to the dynamic setting:

Proposition 5.3.1. For each optimizer u^* of (5.14) with convex penalty function $\psi(u)$, the set of optimizers of the RI associated to u^* is the same as the set of optimizers for the following problem:

$$\max_{\mathbf{p}, \mathbf{q}} H(\mathbf{p}, \mathbf{q}) + \psi^*\left(\rho_{dy}^E - \rho_{dy}\right) \quad (5.16)$$

where $\rho_{dy}(x, a) = \sum_{t=1}^T \beta^{t-1} \mu_t(x) p_t(a|x)$ and $\rho_{dy}^E(x, a) = \sum_{t=1}^T \beta^{t-1} \mu_t(x) p_t^E(a|x)$ for any choice rule p and $(x, a) \in X \times A$.

Proof. See Appendix A.3. □

For this special case IRI problem, since it can be viewed as a sequence of static problems, we can either solve it via the dynamic IRI algorithm, or solve it backwards by applying the static algorithm for T times.

CHAPTER 6

Algorithms & Experiments

In this chapter, we propose efficient algorithms to solve the static and dynamic IRI problems, these algorithms can be applied to both optimal expert policies and sub-optimal expert policies with a penalty function. For each case, we implement several experiments to evaluate the performance. At the end, we provide a proof of the convergence of the algorithms.

6.1 GENERATE EXPERT OBSERVATIONS

Before applying the IRI algorithms, we firstly generate expert's behavior observations which are used as inputs for the inverse problem. To simulate expert's behavior, we assume an agent with utility function u^E and prior μ_1 is making choices by solving either a static RI problem (3.1) or a dynamic RI problem (2.25) with horizon T .

6.1.1 RI algorithms

For static case, we obtain the optimal policy p^E, q^E by solving the static RI problem with a given expert utility function u^E via the following Arimoto-Blahut Algorithm (Arimoto (1972), Blahut (1972)):

For the dynamic case, we obtain the expert policy $\{p_t^E, q_t^E\}_{t=1}^T$ by numerically solving the dynamic RI problem via the extended Forward-Backward Arimoto-Blahut Algorithm proposed in Miao & Xing (2019):

Algorithm 1 Arimoto-Blahut

Input: Utility function $u(x, a)$, prior distribution $\mu(x)$ and marginal information cost λ .

Output: Optimal policy $p(a|x)$ and $q(a)$.

- 1: Initialize $p^{(0)}(a|x)$ randomly;
- 2: for iteration $k \geq 1$, compute:
- 3:

$$q^{(k)}(a) = \sum_x \mu(x) p^{(k-1)}(a | x)$$

- 4:

$$p^{(k)}(a | x) = \frac{q^{(k)}(a) \exp(u(x, a)/\lambda)}{\sum_{a'} q^{(k)}(a') \exp(u(x, a')/\lambda)}$$

- 5: Iterate on k until convergence.
-

6.1.2 sub-optimal expert policy

After generating the expert policy via RI algorithms, the expert policy $\{p_t^E, q_t^E\}_{t=1}^T$ can be used as the input for the inverse RI problem, and we assume that we can directly observe the optimal expert policy.

In practice, it's common to have a sub-optimal expert policy especially in case of high-dimensional state and action space due to limited amount of observations available.

To simulate this situation, we firstly use the optimal policy p^E to randomly generate a set of observations $D^{emp} = \{(x_1, a_1), \dots, (x_N, a_N)\}$ of size N , then an empirical policy (p^{emp}, q^{emp}) can be computed based on D^{emp} . When N is small, (p^{emp}, q^{emp}) can be viewed as the sub-optimal expert policy.

6.2 STATIC IRI

Given the optimal policy p^E, q^E (or sub-optimal policy p^{emp}, q^{emp}), we can solve the following static IRI problem (denote as $IRI(p^E, q^E, \mu)$) by the StaticIRI Algorithm

Algorithm 2 DynamicRI Algorithm (Miao & Xing (2019))

]

Input: Utility function $u^E(x, a)$, prior distribution $\mu_1(x)$ and marginal information cost λ , horizon T .

Output: Optimal policy $\{p_t^E(a|x), q_t^E(a)\}_{t=1}^T$.

Initialize:

$$\mu_1^{(T,1)}(x_1, a_0) = \mu_1(x_1)$$

$$p_t^{(T,0)}(a_t | x_t, a_{t-1}) = p^{(T,0)}(a_t | x_t, a_{t-1}) > 0 \text{ for } t = 1, 2, \dots, T$$

$$\phi_{T+1}^{(k)} = 1 \text{ for } k = 1, 2, \dots, K$$

for all $x_t \in X$ and $a_t, a_{t-1} \in A$, K is a large integer.

2: For iteration $k = 1, 2, \dots, K$ until convergence:

forward path, for $t = 1, 2, \dots, T$:

$$\mu_{t+1}^{(T,k)}(x_{t+1}, a_t) = \sum_{x_t, a_{t-1}} \pi(x_{t+1} | x_t, a_t) p_t^{(T,k-1)}(a_t | x_t, a_{t-1}) \mu_t^{(T,k)}(x_t, a_{t-1})$$

$$q_t^{(T,k)}(a_t | a_{t-1}) = \sum_{x_t} p_t^{(T,k-1)}(a_t | x_t, a_{t-1}) \mu_t^{(T,k)}(x_t | a_{t-1})$$

$$\mu_t^{(T,k)}(x_t | a_{t-1}) = \frac{\mu_t^{(T,k)}(x_t, a_{t-1})}{\sum_{x_t} \mu_t^{(T,k)}(x_t, a_{t-1})} \text{ if } \sum_{x_t} \mu_t^{(T,k)}(x_t, a_{t-1}) > 0$$

4: backward path, for $t = T, T - 1, \dots, 1$:

$$v_t^{(T,k)}(x_t, a_t) = u(x_t, a_t) + \beta \sum_{x_{t+1}} \pi(x_{t+1} | x_t, a_t) \lambda \ln \phi_{t+1}^{(T,k)}(x_{t+1}, a_t)$$

$$\phi_t^{(T,k)}(x_t, a_{t-1}) = \sum_{a_t} q_t^{(T,k)}(a_t | a_{t-1}) \exp\left(v_t^{(T,k)}(x_t, a_t) / \lambda\right)$$

$$p_t^{(T,k)}(a_t | x_t, a_{t-1}) = \frac{q_t^{(T,k)}(a_t | a_{t-1}) \exp\left[v_t^{(T,k)}(x_t, a_t) / \lambda\right]}{\phi_t^{(T,k)}(x_t, a_{t-1})}$$

For infinite horizon, increase T until convergence.

below:

$$\min_u \max_{p,q} \left\{ H(p, q) + \sum_{x,a} u(x, a) \left[\mu(x) (p(a|x) - p^E(a|x)) \right] \right\}, \quad (6.1)$$

Algorithm 3 StaticIRI

Input: Expert policy p^E, q^E or p^{emp}, q^{emp} , prior distribution $\mu(x)$.

Output: Recovered utility function u .

- 1: Initialize $u^{(0)}, p^{(0)}, q^{(0)}$ randomly;
- 2: For iteration $k \geq 1$ and all x, a :
- 3: Update $p(a|x), q(a)$ from $(p^{(k-1)}, q^{(k-1)})$ to $(p^{(k)}, q^{(k)})$ by the Arimoto-Blahut Algorithm with utility $u^{(k-1)}$
- 4: Update $u(x, a)$ by

$$u^{(k)}(x, a) = u^{(k-1)}(x, a) - \alpha \left(\psi'(u^{(k-1)}(x, a)) + \left[\mu(x) (p^{(k)}(a|x) - p^E(a|x)) \right] \right)$$

- 5: Iterate on k until convergence.
-

For iteration k , policy $(p^{(k-1)}, q^{(k-1)})$ is updated via the Arimoto-Blahut algorithm with temporary utility $u^{(k-1)}$, then $u^{(k-1)}$ is updated by the gradient descent algorithm with policy $(p^{(k)}, q^{(k)})$.

To evaluate the performance of StaticIRI algorithm, we apply it to two examples, one with interior expert policy (p_E, q_E) and the other with boundary policy (p_E, q_E) . For both experiments, we will firstly find the equivalent class $[u_E]$, furthermore, if an anchor action exists, we can either find the exact utility function u_E for interior policy case or construct infinite u_E for boundary policy case.

6.2.1 Example with interior (p_E, q_E)

For this example, we consider an environment with 5 states and 5 actions, i.e., $|X| = |A| = 5$; prior belief μ is an uniform distribution over X ; the marginal information cost $\lambda = 1$.

The agent has a utility function function u^E which is an Identity matrix:

$$\mathbf{u}^E(\mathbf{x}, \mathbf{a}) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Next, take the expert policy (p^E, q^E) obtained by solving the problem $RI(u^E, \mu, \lambda)$ as inputs, we apply the staticIRI algorithm, the recovered utility function $u^{recover}(x, a)$ is given by:

$$u^{recover}(x, a) = \begin{pmatrix} 1.187 & 0.187 & 0.187 & 0.187 & 0.187 \\ 0.169 & 1.169 & 0.169 & 0.169 & 0.169 \\ 0.315 & 0.315 & 1.315 & 0.315 & 0.315 \\ 0.469 & 0.469 & 0.469 & 1.469 & 0.469 \\ 0.289 & 0.289 & 0.289 & 0.289 & 1.289 \end{pmatrix}.$$

It's easy to find that $u_{recover}(x, a)$ is equivalent to $u^E(x, a)$ with the constant translators:

$$[C_{x_1}, C_{x_2}, C_{x_3}, C_{x_4}, C_{x_5}] = [0.187, 0.169, 0.315, 0.469, 0.289]$$

Thus, if there is an anchor action exist, we can compute the above translators and further shift $u_{recover}(x, a)$ to get exact $u^E(x, a)$.

6.2.2 Example with Boundary (p^E, q^E)

For this example, we use the same environment settings but the agent has a different utility function u^E given by:

$$\mathbf{u}^E(\mathbf{x}, \mathbf{a}) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0.2 \\ 0 & 1 & 0 & 0 & 0.2 \\ 0 & 0 & 1 & 0 & 0.2 \\ 0 & 0 & 0 & 1 & 0.2 \\ 0 & 0 & 0 & 0 & 0.2 \end{pmatrix}$$

With this utility function, the simulated optimal policy (p^E, q^E) of this agent is a boundary policy with:

$$q^E(a) = (0.25, 0.25, 0.25, 0.25, 0)$$

which means the last action will never be chosen.

In this case, the consideration set $B(q^E) = a_1, a_2, a_3, a_4$, and we use the truncated optimal policy (\hat{p}^E, \hat{q}^E) defined on $X \times B(q^E)$

$$\hat{p}^E(a | x) = p^E(a | x), \quad \hat{q}^E(a) = q^E(a)$$

to recover $u^E(x, a)$ for $a \in B(q^E)$, the recovered results is given by

$$\hat{u}^{rec}(x, a) = \begin{pmatrix} 1.433 & 0.433 & 0.433 & 0.433 \\ 0.438 & 1.438 & 0.438 & 0.438 \\ 0.262 & 0.262 & 1.262 & 0.262 \\ 0.186 & 0.186 & 0.186 & 1.186 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}$$

which is equivalent to the first four columns of the true utility function

$$[u^E(x, a_1), u^E(x, a_2), u^E(x, a_3), u^E(x, a_4)]$$

with translation constant $[C_{x_1}, C_{x_2}, C_{x_3}, C_{x_4}, C_{x_5}] = [0.433, 0.438, 0.262, 0.186, 0.25]$.

As for the last column $[u(x, a_5)]$, we have infinite choices of its value, for example, we can take

$$u^E(x, a_5) = \min_{a_1, a_2, a_3, a_4} u^E(x, a_i) - \epsilon \text{ for } \forall \epsilon > 0.$$

Furthermore, if we have an anchor-action $a^A \in B(q^E)$, then we can find the exact values of $u^E(x, a)$ for all $a \in B(q^E)$, but for the value of $u^E(x, c)$ with $c \in A/B(q^E)$, there infinite possibilities.

6.2.3 sub-optimal policy

In the above two examples, we use the optimal policy (p^E, q^E) as the inputs for the IRI problem, but in practice, the optimal policy may not be accessible because of limited amount of observations.

In this example, we use the empirical distribution (p^{emp}, q^{emp}) as the inputs for

the IRI problem. To avoid over-fitting when search a non-parametric and high-dimensional utility function with limited samples, we introduce a convex penalty function as stated in (3.6).

6.2.3.1 Experiment settings

In the following example, we consider a 10×10 space, an agent with uniform prior μ and the utility function u^E is an identity matrix. We aim to show the impact of ψ on the convergence of $u^{recover}$ and $p^{recover}$ as N increasing.

We consider three groups of ψ -function:

- Group 1 : fixed L2 penalty $\psi(u) = 0.1u^2, 0.01u^2, 0.001u^2$;
- Group 2 : Adaptive L2 penalty $\psi(u) = \frac{1}{N}u^2$;
- Group 3 : no penalty function i.e. $\psi(u) \equiv 0$.

For all experiments, we use two metrics, mean squared error (MSE) and relative error (RE), to evaluate the performance of our inverse models and algorithms, The MSE and RE for utility and policy are defined as follows:

$$MSE(u) = \frac{1}{|X| \times |A|} \sum_{x,a} (u^E(x, a) - u^{recovered}(x, a))^2,$$

$$MSE(p) = \frac{1}{|X| \times |A|} \sum_t \sum_{x,a} (p^E(a|x) - p^{recovered}(a|x))^2,$$

$$RE(u) = \frac{1}{|X| \times |A|} \sum_{x,a} \frac{|u^E(x, a) - u^{recovered}(x, a)|}{|u^E(x, a)|},$$

$$RE(p) = \frac{1}{|X| \times |A|} \sum_{x,a} \frac{|p^E(x, a) - p^{recovered}(x, a)|}{|p^E(x, a)|}.$$

To compute the difference between recovered utility $u^{recover}$ and true utility u^E , we first shift the values of one chosen action in $u^{recover}$ and u^E (one column for the utility matrix), then we compute MSE and RE with shifted values of other chosen actions. If there is only one chosen action, as proved in Chapter 4, there are infinite value choices of the not-chosen actions, which makes MSE and RE invalid. As for policy performance, we compute the difference between $p^{recover}$ and p^E for all actions.

Since the MSE metric is scale dependent, it would be difficult to evaluate the recovered performance by using MSE metric only, especially when values of u^E are in different scale. In this case, the RE metric can be a more direct metric to observe the percentage difference between recovered values and true values.

However, if true values of u^E or p^E are very close to zero, the relative error will be very sensitive, even small difference between recovered values and true values may lead to huge relative error, in such case, we will use the MSE to help us evaluate the performance.

6.2.3.2 Performance

In the following figures, we present the performance of models with different penalty, the vertical axis is RE or MSE between the true values and recovered values, and the horizontal axis is the number of observations.

As shown in Figure 6.1 and Figure 6.2, when there are enough observations, all models have a similar performance except the model with $\psi(u) = 0.1u^2$ where the model didn't learn anything because the penalty effect is dominant; while when there are only limited observations, the model with appropriate penalty has much better and stable performance than the original model.

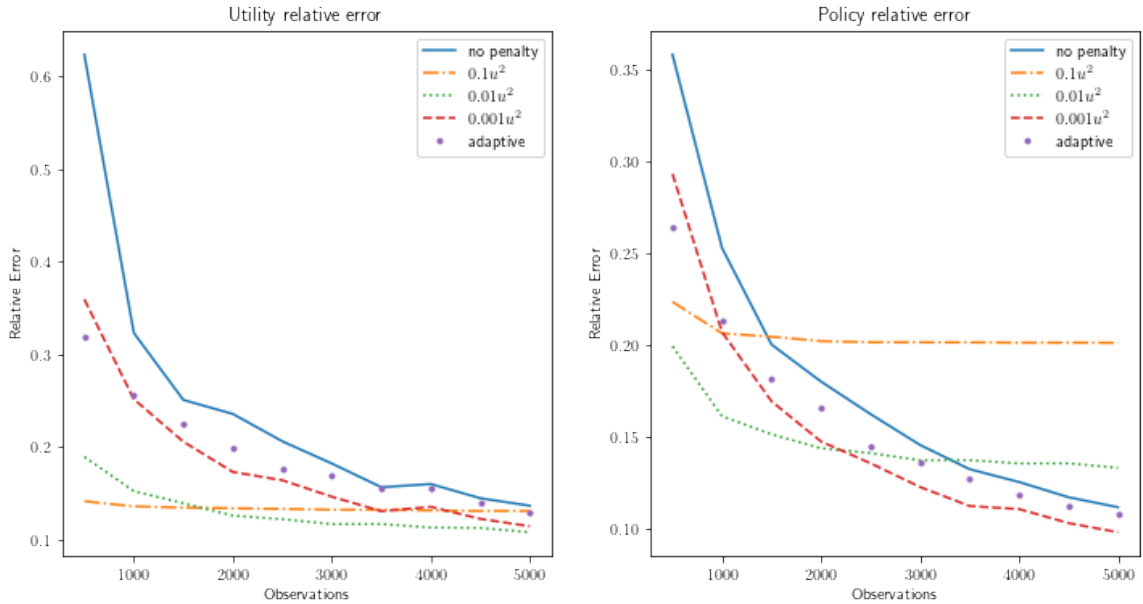


Figure 6.1: Relative Performance with imperfect observations (Static)

6.3 DYNAMIC IRI

In the last section, we have shown that the staticIRI algorithm can successfully recover the target utility function for both interior and boundary expert policies. And when there are only a limited amount observations, an appropriate convex regularizer can obviously improve the performance and stability.

In this section, we propose the DynamicIRI algorithm which can efficiently solve the general case in the dynamic IRI problem formulated in (5.4). As for the special case in the IRI problem (5.14), it can be either solved as a special case of (5.4) via DynamicIRI algorithm or solved as a sequence of static IRI problems via StaticIRI algorithm.

As a numerical method for the general-case dynamic IRI problem defined in (5.4), we propose the following Dynamic-IRI algorithm:

For iteration k , the policy $(p_t^{(k-1)}, q_t^{(k-1)})$ for each period is updated via the Dy-

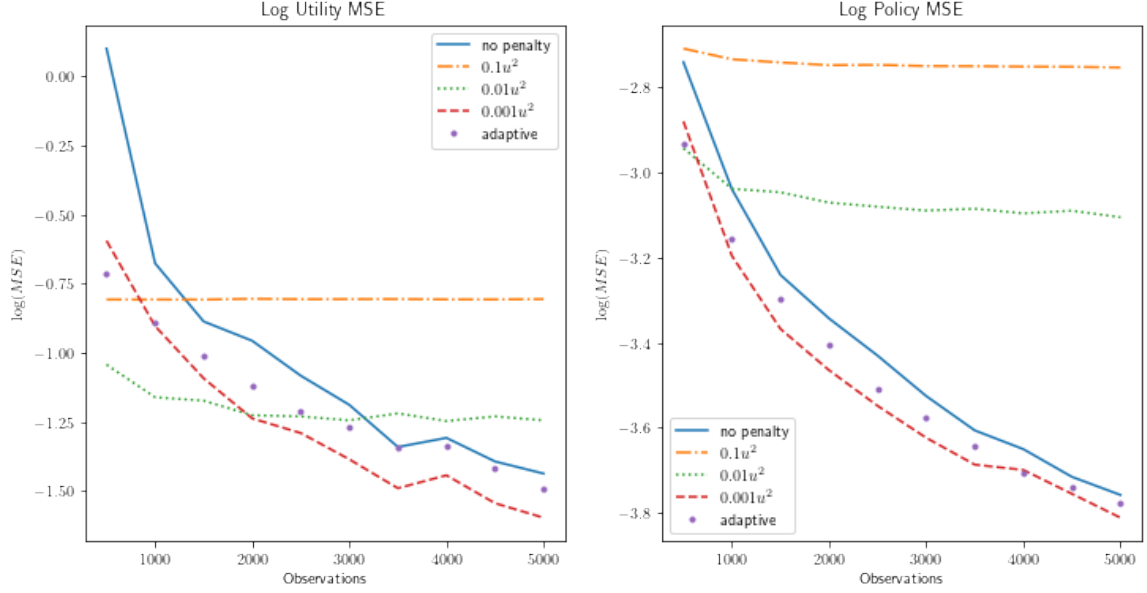


Figure 6.2: MSE Performance with imperfect observations (Static)

dynamicRI algorithm with temporary utility $u^{(k-1)}$, then $u^{(k-1)}$ is updated by the gradient descent algorithm with all periods policies $\{p_t^{(k)}, q_t^{(k)}\}_{t=1}^T$.

6.3.1 Dynamic example with sub-optimal policy

To evaluate the performance of DynamicIRI algorithm, we consider an environment with 3 states and 3 actions, each state has 0.4 probability to stay at the current state and equal probability (0.3) to transfer to other states.

An agent with a uniformly distributed prior μ_1 , is assumed to make decision by solving a dynamic RI problem with horizon $T = 5$. The agent's utility function is an identity matrix and the marginal information cost $\lambda = 1$.

To simulate the sub-optimal policy from limited observations, we firstly use the real optimal policy p_t^E to sample an observation set of size $N \times T$, then we use these datasets to compute the empirical policy $\{p_t^{emp}\}$.

We compare the performance of models with and without penalty function in

Algorithm 4 DynamicIRI

Input: Expert policy $\{p_t^E, q_t^E\}_{t=1}^T$, transition probability π , prior belief μ_1 .

Output: Recovered utility function $\{u_t^E\}_{t=1}^T$.

- 1: Initialize $u^{(0)}$ and $\{p_t^{(0)}, q_t^{(0)}\}_{t=1}^T$ randomly;
- 2: For iteration $k \geq 1$ and all \mathbf{x}, \mathbf{a} :
- 3: Use current utility $u^{(k-1)}$ to update $\{p_t, q_t\}_{t=1}^T$ from $\{p_t^{(k-1)}, q_t^{(k-1)}\}_{t=1}^T$ to $\{p_t^{(k)}, q_t^{(k)}\}_{t=1}^T$ by the Dynamic-RI Algorithm.
- 4: Use current policy $\{p_t^{(k)}\}$, and joint distribution $\{\mu_t^{(k)}(x_t, a_{t-1})\}$ to update $u(x, a)$ by

$$u^{(k)}(x, a) = u^{(k-1)}(x, a) - \alpha \left(\psi'(u^{(k-1)}(x, a)) + \sum_{t=1}^T \sum_{a'} \mu_t^{(k)}(x, a') (p_t^k(a|x, a') - p_t^E(a|x, a')) \right)$$

- 5: Iterate on k until convergence.
-

the following Figure 6.3 and Figure 6.4, respectively, the shaded area represents standard deviation of probabilities among 30 runs for each number of observations. It's easy to find that results from the model with appropriate penalty function are much more stable and accurate than those from no penalty model.

To get a closer look at the recovered results, the recovered utility from the model with penalty function $\psi(u) = 0.1u^2$ and 500 observations is shown below:

$$\mathbf{u}_{recover}(\mathbf{x}, \mathbf{a}) = \begin{pmatrix} 1.00000 & 0.08000 & 0.05000 \\ 0.00000 & 0.91000 & -0.04000 \\ 0.00000 & 0.01000 & 0.96000 \end{pmatrix}$$

The recovered utility function $\mathbf{u}_{recover}$ successfully matches each element of the true utility u^E and can perfectly explain the observed behavior.

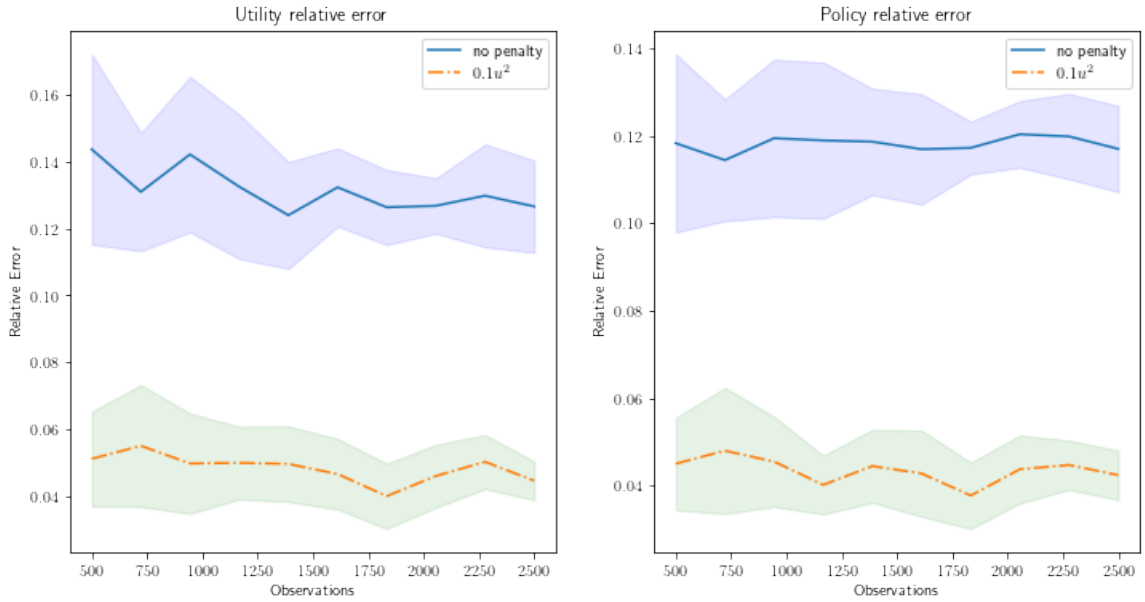


Figure 6.3: Relative Performance with imperfect observations (Dynamic)

6.4 CONVERGENCE ANALYSIS

The static IRI problem formulated in (3.6) is a minmax problem with convex-concave objective function. Nedic & Ozdaglar (2009) provided a sub-gradient method to solve the problem and a proof of its convergence.

The StaticIRI algorithm proposed in this chapter is a combination of Block Coordinate Descent (BCD) algorithm and gradient descent/ascent algorithm (GDA). In this section, we provide a proof of the convergence of StaticIRI algorithm, and the result can be easily extended to special-case dynamic IRI which can be viewed as a sequence of static problem.

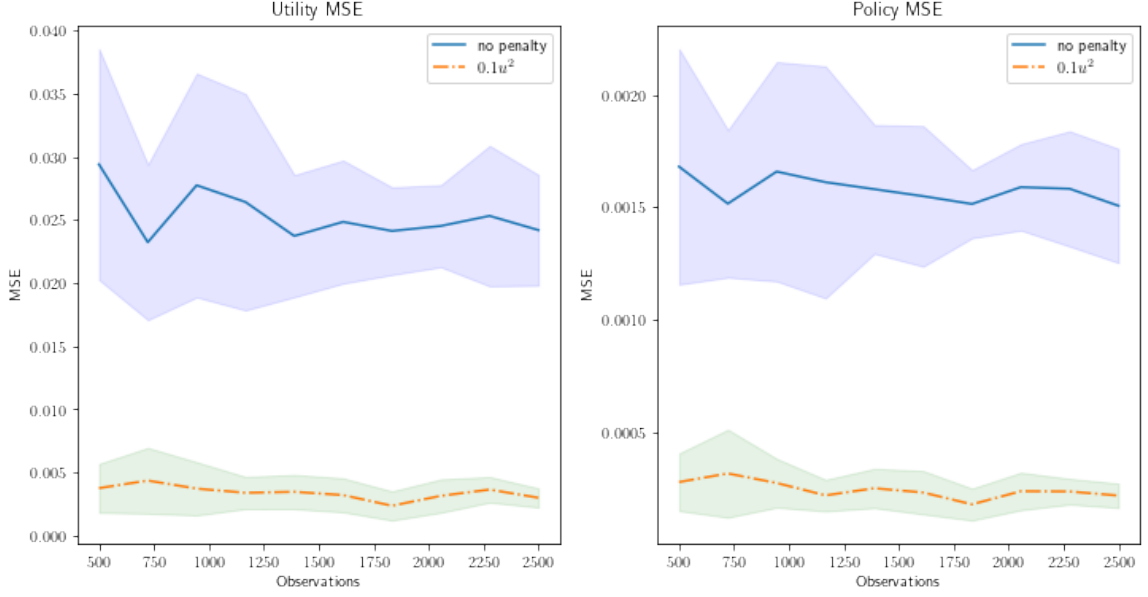


Figure 6.4: MSE Performance with imperfect observations (Dynamic)

6.4.1 Setup

Consider the following inverse rational inattention problem in Static Case

$$\min_u \psi(u) + \left\{ \max_{p,q} \left[\sum_{x,a} p(a|x) \mu(x) \left(u(x,a) - \lambda \ln \frac{p(a|x)}{q(a)} \right) \right] - \left[\sum_{x,a} p_E(a|x) \mu(x) \left(u(x,a) - \lambda \ln \frac{p_E(a|x)}{q_E(a)} \right) \right] \right\}. \quad (6.2)$$

Given $\mu(x)$ and expert policy p_E , and q_E , we define the objective function as:

$$L(u, p, q) = \psi(u) + \sum_{x,a} u(x,a) \mu(x) \left(p(a|x) - p_E(a|x) \right) - \sum_{x,a} \mu(x) p(a|x) \ln \frac{p(a|x)}{q(a)} \quad (6.3)$$

where $L(u, p, q)$ is convex in u and jointly concave in (p, q) .

The StaticIRI algorithm can be briefly summarized as:

- 1) Initialize $u_{(0)}, p_{(0)}^B$ and $q_{(0)}^B$;

For $k = 0, 1, 2, 3, \dots$

2) update (p, q) with block coordinate descent algorithm(BCD) and update u with gradient descent algorithm(GDA):

$$p_k^B = \arg \max_p L(u_k, p, q_k^B) \quad (6.4)$$

$$q_k^B = \arg \max_q L(u_{(k)}, p_{k+1}^B, q)$$

$$u_{(k+1)} = u_{(k)} - \alpha L_u(u_k, p_{k+1}^B, q_{k+1}^B) \quad (6.5)$$

where (p_{k+1}^B, q_{k+1}^B) are the updates from BCD, and meanwhile we denote (p_{k+1}^G, q_{k+1}^G) as the updates from GDA as in Nedic & Ozdaglar (2009):

$$p_{k+1}^G = p_k^G - \alpha L_p(u_k, p_k^G, q_k^G) \quad (6.6)$$

$$q_{k+1}^G = q_k^G - \alpha L_q(u_k, p_k^G, q_k^G)$$

To prove the convergence, we introduce the following assumption:

Assumption 6.4.1.

$$\begin{aligned} \|\mathcal{L}_u(u_k, p_{k+1}, q_{k+1})\| &\leq L_u \\ \|\mathcal{L}_p(u_k, p_k, q_k)\| &\leq L_{pq} \quad \|\mathcal{L}_q(u_k, p_k, q_k)\| \leq L_{pq} \quad \text{for all } k \geq 0 \end{aligned} \quad (6.7)$$

One setting to have the above assumption hold is that we consider a bounded utility functions u and policy q which is bounded away from zero, i.e. all actions are chosen with positive probability.

Under Assumption 6.4.1, we can prove the following results:

Lemma 6.4.1. For any u and (p, q) , we have

$$\mathcal{L}(u_k, p_{k+1}^B, q_{k+1}^B) - \mathcal{L}(u, p_{k+1}^B, q_{k+1}^B) \leq \frac{1}{2\alpha} (\|u_k - u\|^2 - \|u_{k+1} - u\|^2) + \frac{\alpha}{2} L_u^2 \quad (6.8)$$

And,

$$\begin{aligned} & \frac{1}{2\alpha} (\|p - p_{k+1}^G\|^2 - \|p - p_k^G\|^2 + \|q - q_{k+1}^G\|^2 - \|q - q_k^G\|^2) - \frac{\alpha}{2} L_{pq}^2 \\ & \leq \mathcal{L}(u_k, p_{k+1}^B, q_{k+1}^B) - \mathcal{L}(u_k, p, q) \end{aligned} \quad (6.9)$$

Proof. For any value of u , we have

$$\begin{aligned} \|u_{k+1} - u\|^2 &= \|u_k - \alpha \mathcal{L}_u(u_k, p_{k+1}^B, q_{k+1}^B) - u\|^2 \\ &= \|u_k - u\|^2 - 2\alpha \mathcal{L}_u(u_k, p_{k+1}^B, q_{k+1}^B)(u_k - u) + \alpha^2 \|\mathcal{L}_u(u_k, p_{k+1}^B, q_{k+1}^B)\|^2 \end{aligned} \quad (6.10)$$

By the convexity of \mathcal{L} in u ,

$$-\mathcal{L}_u(u_k, p_{k+1}^B, q_{k+1}^B)(u_k - u) \leq -(\mathcal{L}(u_k, p_{k+1}^B, q_{k+1}^B) - \mathcal{L}(u, p_{k+1}^B, q_{k+1}^B))$$

thus,

$$\begin{aligned} \|u_{k+1} - u\|^2 &\leq \|u_k - u\|^2 - 2\alpha (\mathcal{L}(u_k, p_{k+1}^B, q_{k+1}^B) - \mathcal{L}(u, p_{k+1}^B, q_{k+1}^B)) \\ &\quad + \alpha^2 \|\mathcal{L}_u(u_k, p_{k+1}^B, q_{k+1}^B)\|^2 \end{aligned}$$

by Assumption 6.4.1,

$$\mathcal{L}(u_k, p_{k+1}^B, q_{k+1}^B) - \mathcal{L}(u, p_{k+1}^B, q_{k+1}^B) \leq \frac{1}{2\alpha} (\|u_k - u\|^2 - \|u_{k+1} - u\|^2) + \frac{\alpha}{2} L_u^2$$

For the second result, by the joint concavity of \mathcal{L} in (p, q) and Assumption 6.4.1, a similar argument as above obtains that for any value of (p, q) ,

$$\begin{aligned} & \frac{1}{2\alpha} (\|p - p_{k+1}^G\|^2 - \|p - p_k^G\|^2 + \|q - q_{k+1}^G\|^2 - \|q - q_k^G\|^2) \\ & \quad - \frac{\alpha}{2} L_{pq}^2 \leq \mathcal{L}(u_k, p_{k+1}^G, q_{k+1}^G) - \mathcal{L}(u_k, p, q) \end{aligned}$$

then by the fact $L(u_k, p_{k+1}^G, q_{k+1}^G) \leq L(u_k, p_{k+1}^B, q_{k+1}^B)$, we have

$$\begin{aligned} & \frac{1}{2\alpha} (\|p - p_{k+1}^G\|^2 - \|p - p_k^G\|^2 + \|q - q_{k+1}^G\|^2 - \|q - q_k^G\|^2) - \frac{\alpha}{2} L_{pq}^2 \\ & \leq \mathcal{L}(u_k, p_{k+1}^B, q_{k+1}^B) - \mathcal{L}(u_k, p, q) \end{aligned}$$

□

Proposition 6.4.1. Assume (u^*, p^*, q^*) is a saddle point of \mathcal{L} , we have:

$$\begin{aligned} & -\frac{1}{2K\alpha} (\|p^* - p_0\|^2 + \|q^* - q_0\|^2) - \frac{\alpha}{2} L_{pq}^2 \leq \frac{1}{K} \sum_k \mathcal{L}(u_k, p_{k+1}^B, q_{k+1}^B) - \mathcal{L}(u^*, p^*, q^*) \\ & \leq \frac{1}{2\alpha K} \|u_0 - u^*\|^2 + \frac{\alpha}{2} L_u^2 \end{aligned} \quad (6.11)$$

Proof. Firstly, we replace u by u^* in (6.8) and add all the results for $k = 1, \dots, K$, we obtain

$$\frac{1}{K} \sum_k \mathcal{L}(u_k, p_{k+1}^B, q_{k+1}^B) - \frac{1}{K} \sum_k \mathcal{L}(u^*, p_{k+1}^B, q_{k+1}^B) \leq \frac{1}{2\alpha K} \|u_0 - u^*\|^2 + \frac{\alpha}{2} L_u^2 \quad (6.12)$$

Take $\hat{p}_K^B, \hat{q}_K^B = (\frac{1}{K} \sum_k p_{k+1}^B, \frac{1}{K} \sum_k q_{k+1}^B)$, which is the average of BCD iterations, then by the joint concavity of \mathcal{L} in (p, q) , we have $\frac{1}{K} \sum_k \mathcal{L}(u^*, p_{k+1}^B, q_{k+1}^B) \leq \mathcal{L}(u^*, \hat{p}_K^B, \hat{q}_K^B)$. Thus,

$$\frac{1}{K} \sum_k \mathcal{L}(u_k, p_{k+1}^B, q_{k+1}^B) - \mathcal{L}(u^*, \hat{p}_K^B, \hat{q}_K^B) \leq \frac{1}{2\alpha K} \|u_0 - u^*\|^2 + \frac{\alpha}{2} L_u^2 \quad (6.13)$$

Similarly, we have

$$-\frac{1}{2K\alpha} (\|p^* - p_0\|^2 + \|q^* - q_0\|^2) - \frac{\alpha}{2} L_{pq}^2 \leq \frac{1}{K} \sum_k \mathcal{L}(u_k, p_{k+1}^B, q_{k+1}^B) - \mathcal{L}(\hat{u}_K, p^*, q^*) \quad (6.14)$$

where $\hat{u}_K = \frac{1}{K} \sum_k u_k$ is the average of iterations.

Since (u^*, p^*, q^*) is a saddle point,

$$\mathcal{L}(u^*, \hat{p}_K^B, \hat{q}_K^B) \leq \mathcal{L}(u^*, p^*, q^*) \leq \mathcal{L}(\hat{u}_K, p^*, q^*)$$

The result follows by replacing $\mathcal{L}(u^*, \hat{p}_K^B, \hat{q}_K^B)$ and $\mathcal{L}(\hat{u}_K, p^*, q^*)$ by $\mathcal{L}(x^*, p^*, q^*)$ in (6.13) and (6.14). \square

The above result shows that the averaged value $\frac{1}{K} \sum_k \mathcal{L}(u_k, p_{k+1}^B, q_{k+1}^B)$ obtained from StaticIRI algorithm will converge to the saddle point value $\mathcal{L}(u^*, p^*, q^*)$ with rate $\frac{1}{K}$ and error level $\max(\frac{\alpha}{2} L_u^2, \frac{\alpha}{2} L_{pq}^2)$

CHAPTER 7

Financial Applications

In this chapter, we apply our IRI framework to robo-advising problems in a mean-variance setting and a target date investment, in both cases, we are able to recover the investors' utility functions, the recovered utility can help us to better understand their preference and further improve the investment performance.

7.1 ROBO-ADVISING

In the last decades, Robo-advisors have emerged as a popular alternative to traditional financial advisors because of the benefits such as low service fee, low open balance, less human bias etc. (Foerster et al. (2017), D'Acunto & Rossi (2020)).

Robo-advisers collect and use individual-specific information to estimate their clients' preferences and construct specific financial plans and advice for clients. Most existing robo-advisors categorize clients based on features such as risk attitude, age, income and etc., and clients that fall into the same buckets obtain the same advice (D'Acunto & Rossi (2020)). This kind of method may fail to provide individual-specific advice because clients in the same category can have distinct preferences.

Alsabah et al. (2019) proposed a reinforcement learning framework in which the robo-advisor can learn investors' risk preferences over time by observing portfolio choices under different market conditions.

Our model also learns investors' preferences by observing their choices, compared to Alsabah et al. (2019), we assume the investors are making choices with information cost because market conditions are often not observable, especially for individual investors. In most cases, the investors cannot directly observe the

current state, their behaviors may not be the results of maximizing utility but a balance between utility and information cost. Besides, we formulate the preference reveal problem in a more general setting with non-parametric utility function and non-stationary choice rule.

7.2 ENVIRONMENT SETUP

As the first step for both applications, we set up the environment where investors make choices and take utilities.

The environment is modeled as a Markov Decision Process $(X, A, \pi, u^E(x, a))$, where X is the state space and each state is a market regime; A is the action space, each action is a investment strategy; π is the transition kernel, for these two applications, we assume $\pi \equiv \pi(x'|x)$, the next state depends only on the current. We discuss each component in details in this section.

7.2.1 State

We consider a market with 3 regimes: x_0 (Boom), x_1 (Normal) and x_2 (recession), and as modeled in Hardy (2001), we assume the market log-return under each regime follows a normal distribution:

$$\log \frac{S_{t+1}}{S_t} | x \sim N(\mu_x, \sigma_x^2) \quad (7.1)$$

Thus, for this three-regime model we have 12 parameters in total:

$$\Theta = \{\mu_{x_1}, \mu_{x_2}, \mu_{x_3}, \sigma_{x_1}, \sigma_{x_2}, \sigma_{x_3}, \pi\}$$

where

$$\pi = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix}$$

is the transition probabilities.

Then we fit this regime-switch model to monthly return data of *S&P500* index from 1980 to 2020 via maximum likelihood estimation with the likelihood function:

$$L(\Theta) = P(x_1 | \Theta) P(x_2 | \Theta, x_1) P(x_3 | \Theta, x_1, x_2) \dots P(x_n | \Theta, x_1, \dots, x_{n-1}) \quad (7.2)$$

The estimated mean and variance of the return in each regime are given by

State	return μ_x	variance σ_x^2
x_0	0.022	0.0002
x_1	0.007	0.0026
x_2	-0.01	0.00035

And the estimated transition kernel is given by:

$$\pi(x_{t+1}|x_t) = \begin{pmatrix} 0.51 & 0.03 & 0.46 \\ 0.018 & 0.98 & 0.002 \\ 0.77 & 0.04 & 0.19 \end{pmatrix}$$

Based on the above transition kernel, a boom market has half probability to continue to be a boom market and the other half probability to move to a recession market; while a recession market has high probability to move to a boom market in the next period. The normal market is very likely to continue to be normal, thus

we can usually observe a normal market last for a quite long time.

7.2.2 Action

For the actions, we assume each action is an investment strategy where the investor holds a portfolio consist of a risk-free asset and a risky asset. So different actions represent portfolios with different weights of the risky asset.

Here we consider three actions: portfolios with 30%, 50% and 80% risky asset, which represent a conservative strategy, a balanced strategy, and an aggressive strategy separately.

7.3 MEAN-VARIANCE OPTIMIZATION

Given the above setting, in this section, we consider an investor with mean-variance utility function, and the investor is making investment decisions at each period to optimize the objective:

$$\max_{\{p_t\}_{t=1}^T, \{q_t\}_{t=1}^T} \sum_{t=1}^T \beta^{t-1} \mathbb{E}[u^E(x_t, a_t)] - \lambda \beta^{t-1} I(x_t; a_t | a_{t-1}) \quad (7.3)$$

The first term in the above objective function is the expected utility:

$$\mathbb{E}[u^E(x_t, a_t)] = \sum_{x_t, a_t} \mu_t(x_t) p_t(a_t | x_t) (W_0 R(x_t, a_t) - \theta W_0^2 \sigma^2(x_t, a_t))$$

where θ is a constant risk aversion parameter and W_0 is the initial wealth.

The second term

$$I(x_t; a_t | a_{t-1}) = \mu_t(x_t) p_t(a_t | x_t) \log \frac{p_t(a_t | x_t)}{q_t(a_t)}$$

is the information cost cause by agent's acquisition of information at each period.

By choosing the prior distribution μ as the stationary prior $\mu = [0.23, 0.63, 0.14]$, the above problem is equivalent to a static problem.

7.3.1 True Preference and policy

Given parameters $\theta = 0.05$, $\lambda = 0.5$, and the initial wealth $W_0 = 100$, we can obtain the true utility function u^E and optimal choice policy (p^E, q^E) as follows:

$$\mathbf{u}^E(\mathbf{x}, \mathbf{a}) = \begin{pmatrix} 0.69304 & 1.10540 & 1.70924 \\ 0.13482 & 0.05478 & -0.26036 \\ -0.27438 & -0.51523 & -0.90356 \end{pmatrix}$$

$$\mathbf{p}^E(\mathbf{a}|\mathbf{x}) = \begin{pmatrix} 0.240 & 0.000 & 0.760 \\ 0.840 & 0.000 & 0.160 \\ 0.890 & 0.000 & 0.110 \end{pmatrix}$$

In the Appendix B, we present experiment results with different choices of θ and λ to show their impacts on agents' optimal behavior p^E .

7.3.2 Recovered results

Given the optimal policy (p^E, q^E) and stationary prior μ , we recover the investor's utility function by solving the corresponding IRI problem.

Recovered utility function $u^{recover}$ (with the help of an anchor action) and the recovered policy $p^{recover}$ are shown below:

$$\mathbf{u}^{recover}(\mathbf{x}, \mathbf{a}) = \begin{pmatrix} 0.69304 & 0.99948 & 1.70924 \\ 0.13482 & -0.50013 & -0.26036 \\ -0.27438 & -0.71982 & -0.90356 \end{pmatrix}$$

$$\mathbf{p}^{recover}(\mathbf{a}|\mathbf{x}) = \begin{pmatrix} 0.239 & 0.000 & 0.761 \\ 0.841 & 0.000 & 0.159 \\ 0.894 & 0.000 & 0.106 \end{pmatrix}$$

which perfectly recovered the values of all chosen actions of the true utility function u^E and all values of the observed policy p^E .

The above experiment takes the optimal policy p^E as the input, while in practice, the observed behaviors are usually sub-optimal because of either human mistakes or limited amount of observation data.

To simulate this situation, we firstly use 500 samples generated from optimal policy p^E to compute the empirical policy p^{emp} , then we use this sub-optimal policy p^{emp} as the inputs for the inverse problem to evaluate the performance of our model.

As shown in Table B.2 and Table B.4 in Appendix B, for most experiments with sub-optimal input policies, our model can still successfully recover the target utility, and meanwhile the model with selected penalty function always has a more stable and accurate performance than the model without penalty function.

7.4 TARGET DATE INVESTMENT

In this example, we consider a target investment problem with horizon T where the investor aims to maximize the expected utility of his wealth at time T net the

information cost:

$$\max_{\pi_t, d_t} \mathbb{E} \left[\beta^T u(W_T) \right] - \lambda \mathbb{E} \left[\sum_{t=0}^T \beta^t W_t^{1-\gamma} I(\mu_t, d_t) \right] \quad (7.4)$$

where β is the discounting factor, u is the CRRA utility $u(c) = \frac{c^{1-\gamma}}{1-\gamma}$, and μ_t is the prior distribution at time t of the current market regime, i.e., $\mu_t \equiv \mu_t(x_t)$.

π_t is the action strategy, where $\pi_t(s)$ is the proportion of wealth invested in the risky asset when observe signal s at time t in order to maximize the expected terminal utility. Since the current state x_t is not observable at time t , the agent also designs an information strategy $d_t(s_t|x_t)$ which describes the probability of observing signal realization s_t at time t given the state realization x_t .¹ The mutual information

$$I(\mu_t, d_t) = H(\mu_t) - \mathbb{E}[H(\mu_t(\cdot | s))]$$

measures the uncertainty reduction over the state after observing the signal.

For a constant interest rate r , agent's wealth W follows:

$$W_{t+1} = (1+r)W_t + (R_{t+1} - r)\pi_t W_t, \quad t = 0, 1, \dots, T \quad (7.5)$$

where π_t is the proportion of wealth invested in the risky asset and it is measurable to the signal value s_t at time t .

To maintain the homothetic property of the utility function, we consider the following modified version of (7.4):

$$\max_{\pi_t, d_t} \mathbb{E} \left[\beta^T \frac{W_T^{1-\gamma}}{1-\gamma} \right] - \lambda \left[\sum_{t=0}^T \beta^t W_t^{1-\gamma} I(\mu_t, d_t) \right]. \quad (7.6)$$

¹If the signal distribution also depends on the historic returns, d_t would condition on $\{R_1, \dots, R_{t+1}\}$. In this case, p_t would also condition on past investment strategies $\{\pi_1, \dots, \pi_{t-1}\}$.

When $\gamma > 1$, richer agent has smaller information cost. Similar modification to preserve the homothetic property is also used in the model ambiguity literature.

Following the Lemma 1 in Matejka and McKay (2015), problem (7.4) can be transformed to the following problem

$$\max_{p_t} \mathbb{E} \left[\beta^T \frac{W_T^{1-\gamma}}{1-\gamma} \right] - \lambda \left[\sum_{t=0}^T \beta^t W_t^{1-\gamma} I(p_t, q_t) \right], \quad (7.7)$$

where p_t is the conditional probability of investment strategy π_t given x_t and

$$I(p_t, q_t) = \sum_{\pi_t, x_t} p_t(\pi_t | x_t) \mu_t(x_t) \log \frac{p_t(\pi_t | x_t)}{q_t(\pi_t)}, \quad q_t(\pi_t) = \sum_{x_t} p_t(\pi_t | x_t) \mu_t(x_t).$$

We solve the above problem (7.7) backwards:

Terminal period: At the last period, we solve the problem:

$$\max_{p_{T-1}} \mathbb{E} \left[\beta \frac{W_T^{1-\gamma}}{1-\gamma} \right] - W_{T-1}^{1-\gamma} I(\mu_{T-1}, p_{T-1}), \quad (7.8)$$

Since $W_T = (1+r)W_{T-1} + (R_T - r)\pi_{T-1}W_{T-1}$, it's equivalent to

$$\begin{aligned} \max_{p_{T-1}} W_{T-1}^{1-\gamma} \{ & \sum_{x_{T-1}, \pi_{T-1}} p_{T-1}(\pi_{T-1} | x_{T-1}) \mu_{T-1}(x_{T-1}) \{ \mathbb{E}_{x_{T-1}} \left[\frac{\beta}{1-\gamma} (1+r) \right. \\ & \left. + (R_T - r)\pi_{T-1} \right]^{1-\gamma} \} - \lambda \log \frac{p_{T-1}(\pi_{T-1} | x_{T-1})}{q_{T-1}(\pi_{T-1})} \} \} \end{aligned} \quad (7.9)$$

where $q_{T-1}(\pi_{T-1}) = \sum_{x_{T-1}} p_{T-1}(\pi_{T-1} | x_{T-1}) \mu_{T-1}(x_{T-1})$. This can be viewed as a static RI problem with utility function

$$u(x_{T-1}, \pi_{T-1}) = \mathbb{E}_{x_{T-1}} \left[\frac{\beta}{1-\gamma} (1+r + (R_T - r)\pi_{T-1})^{1-\gamma} \right] \quad (7.10)$$

where given regime x_{T-1} at time T-1, the return $R_T | x_{T-1}$ follows the normal distri-

bution $N(\mu(x_{T-1}), \sigma(x_{T-1}))$.

We call the optimal value as $W_{T-1}^{1-\gamma} V_{T-1}(\mu_{T-1})$, where the argument of V_{T-1} is the prior distribution of x_{T-1} at time $T-1$.

Other periods: Given the results of time $t+1$, the problem at time t is

$$\max_{p_t} \mathbb{E} \left[\beta W_{t+1}^{1-\gamma} V_{t+1}(\mu_{t+1}) \right] - W_t^{1-\gamma} I(\mu_t, p_t), \quad (7.11)$$

substitute $W_{t+1} = (1+r)W_t + (R_{t+1} - r)\pi_t W_t$, we obtain:

$$\begin{aligned} & \max_{p_t} W_t^{1-\gamma} \left\{ \sum_{x_t, \pi_t} p_t(\pi_t | x_t) \mu_t(x_t) \left[\mathbb{E}_{x_t} \left[\beta V_{t+1}(\mu_{t+1}) \left(1 + r + (R_{t+1} - r)\pi_t \right)^{1-\gamma} \right] \right. \right. \\ & \left. \left. - \lambda \log \frac{p_t(\pi_t | x_t)}{q_t(\pi_t)} \right] \right\}, \end{aligned} \quad (7.12)$$

where $q_t(\pi_t) = \sum_{x_t} p_t(\pi_t | x_t) \mu_t(x_t)$ and given regime x_t at time t , the return $R_{t+1} | x_t$ follows the normal distribution $N(\mu(x_t), \sigma(x_t))$.

Given that μ_t does not depend on p_t , problem (7.12) can be viewed as another separate static problem with the utility function u_t given by:

$$u_t^E(x_t, \pi_t) = \mathbb{E}_{x_t} \left[\beta V_{t+1}(\mu_{t+1}) \left(1 + r + (R_{t+1} - r)\pi_t \right)^{1-\gamma} \right] \quad (7.13)$$

Thus, the original optimization problem (7.4) can be viewed as a sequence of static RI problems with different prior distributions μ_t and utility function u_t^E for $t = 1, 2, \dots, T$.

7.4.1 Dynamics of priors

As mentioned in Chapter 6, for dynamic inverse IRI problem, the dynamic of the agent's prior distribution can be obtained before we solve the inverse problem. For

this application, the dynamic of priors with initial prior μ_1 is given by:

$$\mu_{t+1}(x_{t+1}) = \sum_{x_t} T(x_{t+1}|x_t)\mu_t(x_t). \quad (7.14)$$

where $T(x_{t+1}|x_t)$ is the transition probability from the regime estimation.

The top two plots in the following Figure 8.1 show that how the investor's prior and optimal behavior will evolve if the investor has a optimistic prior belief ($\mu_1 = [0.6, 0.2, 0.2]$, i.e. the investor believes the current state is more likely to be a boom market) at time 1, and the bottom two plots show the same dynamics if the investor has a pessimistic prior belief ($\mu_1 = [0.2, 0.2, 0.6]$, i.e. the investor believes the current state is more likely to be a recession market at time 1).

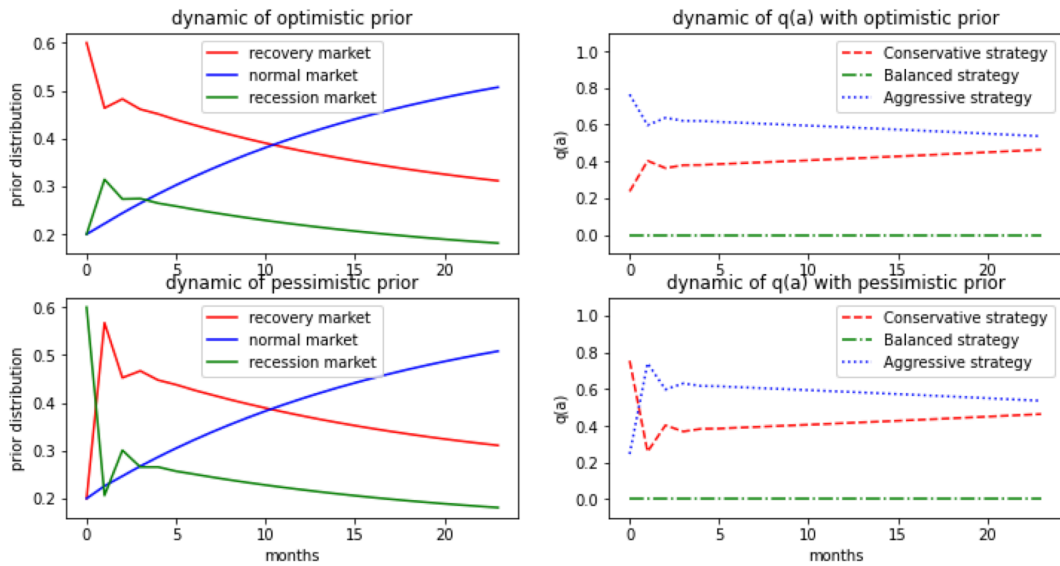


Figure 7.1: Dynamics of investor's prior distribution

We can find that as time goes by, an optimistic investor starts to guess whether a boom period is followed by a recession period, thus at period 1, which can be viewed as a transition period, the investor starts to put less weight on the aggres-

sive strategy and more weight on the conservative strategy.

Meanwhile, since the transition kernel shows high chance of transferring from a recession market to a boom market, a pessimistic agent tend to believe a boom market is coming after the current recession market, thus the investor starts to put less weight on the conservative strategy and more weight on the aggressive strategy.

After the first several periods, the investor gradually knows more about the dynamic of the market and his optimal policy will then converge to a long-term equilibrium.

7.4.2 Recovered results

Since the target date investment problem (7.4) can be viewed as a sequence of separate static problems, we solve the inverse problem as a sequence of inverse static problems to recover the intermediate utility functions given by (7.10) and (7.13).

In Figure 7.2 and Figure 7.3, we present the performance of a model with penalty function and a model without penalty function. The input policy is a sub-optimal policy from limited observation data and the performance is measured by the averaged MSE and RE over all time steps.

Compared with other examples, the relative error of recovered utilities from this example is quite high in this example. This is because, for this example, values of the true utility $u^E(x, a)$ for different (x, a) are similar, after shifting all values by one column, the shifted values are close to zero, thus even very small difference between the recovered value and the true value can lead to a huge relative error.

Therefore, in this example, the MSE metric is a more valid metric to evaluate

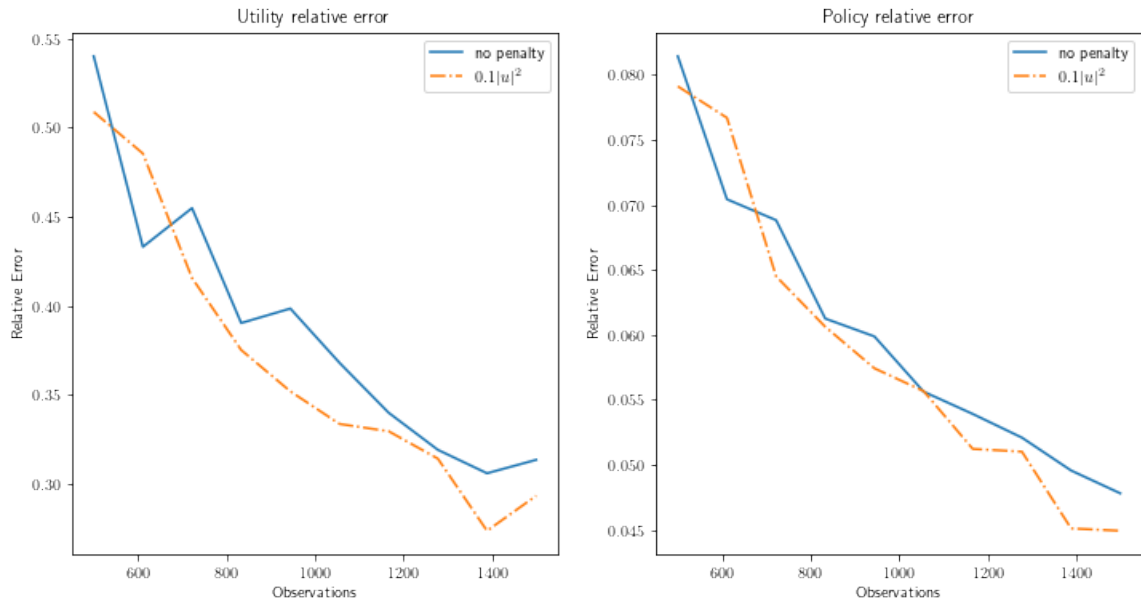


Figure 7.2: Relative performance of target date investment

the performance. As shown in Figure 7.3, both models have the utility MSE around 10^{-5} to 10^{-6} , which is quite small considering the average value of the true utility u^E is around 0.2.

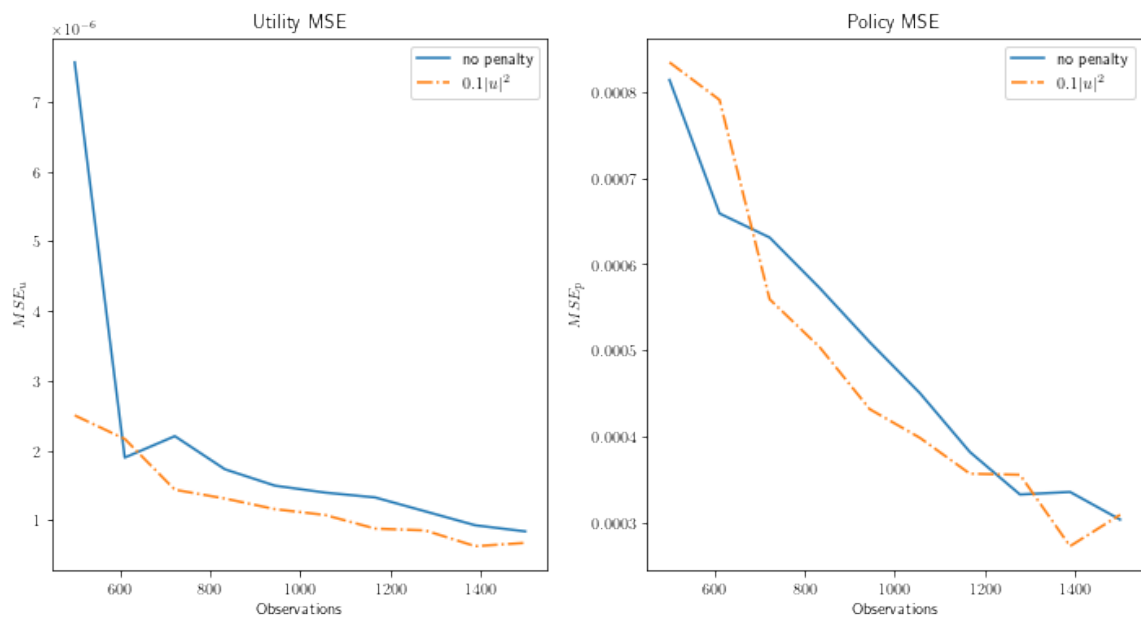


Figure 7.3: MSE performance of target date investment

CHAPTER 8

Open questions

In this thesis, we formulate an inverse rational inattention problems to recover agents' utility function by observing their behaviors, so far the inverse problem is built on finite and discrete state space and action space. When it comes to high-dimensional or continuous space, the existing model and algorithm may be inefficient because of the computation burden from the partition function and the non-parametric setting.

In this chapter, we provide several extensions to the current work in order to apply the IRI framework to more general settings.

8.1 STATIC AND CONTINUOUS RI PROBLEM

Consider a state space \mathcal{X} and an action space \mathcal{A} , both \mathcal{X} and \mathcal{A} are assumed to be continuous. A choice rule is a conditional distribution with probability density function $p(a|x)$, $\mu(x)$ is the probability density function of the prior distribution over state, and the probability density of the endogenous default rule $q(a)$ is determined by the constraint:

$$q(\mathbf{a}) = \int_{\mathcal{X}} p(\mathbf{a}|\mathbf{x}) \mu(\mathbf{x}) d\mathbf{x} \quad (8.1)$$

The agent take utilities that depend on current state and action, the utility function is a bounded function $u : \mathcal{X} \times \mathcal{A} \rightarrow [u_{min}, u_{max}]$. Then, the objective of an rational inattention problem can be defined as:

$$RI(u) \equiv \max_{p,q} \mathbb{E}_{(\mathbf{x},\mathbf{a}) \sim p(\cdot|x), \mu(\cdot)} [u(\mathbf{x}, \mathbf{a})] - \lambda \mathcal{I}(\mathbf{x}; \mathbf{a}) \quad (8.2)$$

where

$$\mathcal{I}(\mathbf{x}; \mathbf{a}) = \mathcal{H}(\mathbf{a}) - \mathcal{H}(\mathbf{a} | \mathbf{x}) = \int_{\mathcal{A}} \int_{\mathcal{X}} \mu(\mathbf{x}) p(\mathbf{a} | \mathbf{x}) \log \frac{p(\mathbf{a} | \mathbf{x})}{q(\mathbf{a})} d\mathbf{x} d\mathbf{a} \quad (8.3)$$

is the information cost.

Given the above formulation of the continuous RI problem, we have two important questions:

- First, whether the optimal policy $p(a|x)$ still has the energy-based form $p(a|x) = \frac{q(a)e^{u_\theta(x,a)}}{Z(x)}$?
- Second, is there any algorithm that can efficiently compute the partition function $Z(x)$ defined in continuous space.

8.2 STATIC AND CONTINUOUS INVERSE RI PROBLEM

Given a set of demonstrations $\mathcal{D}^E = \{(x_1, a_1), \dots, (x_N, a_N)\}$, the inverse RI problem in continuous space is trying to infer the utility function $u_\theta(x, a)$, which is parameterized by θ , by solving a maximum likelihood problem:

$$\max_{\theta} E_{(x,a) \sim \mathcal{D}^E} [\log p_\theta(x, a)] \quad (8.4)$$

where the joint distribution $p_\theta(x, a) = \mu(x) p_\theta(a|x)$, and we formula the stochastic policy by the Energy-based model:

$$p_\theta(a|x) = \frac{q(a)e^{u_\theta(x,a)}}{Z(x)} \quad (8.5)$$

and the constraint (8.1).

8.2.1 Algorithm

To solve this continuous IRI problem, we have two possible approaches:

8.2.1.1 IRL-based algorithm

By setting $\hat{u}(x, a) = u(x, a) + \log q(a)$, the RI objective (8.2) can be written as

$$RL(\hat{u}) \equiv \max_p \mathbb{E}_{(\mathbf{x}, \mathbf{a}) \sim p(\cdot|x), \mu(\cdot)} [\hat{u}(\mathbf{x}, \mathbf{a})] + \lambda \mathcal{H}(p(\cdot|x)) \quad (8.6)$$

which is equivalent to a maximum entropy reinforcement learning problem with reward function \hat{u} , where $\mathcal{H}(p(\cdot|x)) = \mathbb{E}[-\log p(\mathbf{a}|x)]$.

Lemma 8.2.1. For an RI problem with true utility function u^E , optimal policies (p^E, q^E) , set $\hat{u}^E \equiv u^E + \log q^E$, if $RL(\hat{u}^E)$ has the optimal solution \hat{p}^E , then $p^E = \hat{p}^E$

Proof.

$$\hat{p}^E(a|x) = \frac{e^{\hat{u}^E(x,a)}}{\int_{\mathcal{A}} e^{\hat{u}^E(x,a)}} = \frac{q^E(a)e^{u^E(x,a)}}{\int_{\mathcal{A}} q^E(a)e^{u^E(x,a)}} = p^E(a|x)$$

□

Therefore, given a policy (p^E, q^E) , we can recover the function \hat{u}^E by solving $IRL(p^E)$ via any continuous IRL algorithm, then compute the original utility function by $u^E(x, a) = \hat{u}^E(x, a) - \log q^E(a)$. The basic idea is summarized as follow:

8.2.1.2 Continuous RI-based approach

Instead of solving the IRI problem as an IRL problem, we can also directly solve the continuous space IRI problem. One major advantage of this approach is that solving the corresponding RI problem may provide us insights of the relationship between utility function and optimal policy as in the discrete space.

-
- Step 1: Given prior density $\mu(x)$ and expert data $\mathcal{D}^E = \{(x_1, a_1), \dots, (x_N, a_N)\}$;
 - Step 2: Solve for $\hat{u}^{recover}$ and $\hat{p}^{recover}$ by any continuous IRL algorithm like AIRL(Fu(2018));
 - Step 3: Compute $q^{recover}(\mathbf{a}) = \int_{\mathcal{X}} p^{recover}(\mathbf{a}|\mathbf{x}) \mu(\mathbf{x}) d\mathbf{x}$;
 - Step 4: Compute $u^{recover}(x, a) = \hat{u}^{recover}(x, a) - \log q^{recover}(a)$.
-

For the continuous IRI problem, we consider the Energy-based model which modeling the joint distribution $p_\theta(x, a)$ as:

$$p_\theta(x, a) = \frac{\mu_0(x)q(a)e^{u_\theta(x,a)}}{Z} \quad (8.7)$$

Then, objective of the IRI problem is to maximize the log-likelihood:

$$J(\theta) = \mathbb{E}_{(x,a) \sim \mathcal{D}^E} [\log p_\theta(\tau)] = \mathbb{E}_{(x,a) \sim \mathcal{D}^E} [u_\theta(x, a) + \log (\mu_0(x)q(a))] - \log Z \quad (8.8)$$

subject to (8.1).

Unlike the finite discrete space, the partition function Z in high dimensional or continuous space could be very computationally challenging, thus always estimated via some sample-based methods in practice. As in Finn et al. (2016a), Z was estimated via importance sampling with a mixed distribution $\mu = \frac{1}{2}D + \frac{1}{2}\tilde{p}$, where \tilde{p} is a new sampling distribution and D is an estimate of the density of demonstrations, for example, we can use current p_θ in each iteration.

Here we use the same estimation method, let $\mu = \frac{1}{2}p_\theta + \frac{1}{2}\tilde{p}$, we have

$$\log Z = \log \mathbb{E}^\mu \left[\frac{\mu_0(x)q(a)e^{u_\theta(x,a)}}{\frac{1}{2}\frac{\mu_0(x)q(x)e^{u_\theta(x,a)}}{Z} + \frac{1}{2}\tilde{p}} \right] \quad (8.9)$$

Take the derivative of $J(\theta)$, we have:

$$\frac{\partial}{\partial \theta} J(\theta) = E^{\mathcal{D}} \left[\frac{\partial}{\partial \theta} u_\theta(x, a) \right] - E^{p_\theta} \left[\frac{\partial}{\partial \theta} u_\theta(x, a) \right] \quad (8.10)$$

replace p_θ with the importance sampling distribution $\tilde{\mu}$:

$$\frac{\partial}{\partial \theta} J(\theta) = E^{\mathcal{D}} \left[\frac{\partial}{\partial \theta} u_\theta(x, a) \right] - E^{\tilde{\mu}} \left[\frac{p_\theta}{\tilde{\mu}} \frac{\partial}{\partial \theta} u_\theta(s_t, a_t) \right] \quad (8.11)$$

where $\tilde{\mu} = \frac{1}{2}p_\theta + \tilde{p}$.

Lemma 8.2.2. Training a GANs model with following form of discriminator

$$D_\theta(x, a) = \frac{\frac{\mu_0(x)q(a) \exp\{u_\theta(x,a)\}}{Z}}{\frac{\mu_0(x)q(a) \exp\{u_\theta(x,a)\}}{Z} + p(x, a)} \quad (8.12)$$

is equivalent to solve a IRI problem defined in (8.8)

Proof. See Appendix A.4. □

Given the above lemma, a potential algorithm for the IRI problem is provided below:

Algorithm 5 Continuous-space IRI Algorithm

Input: Expert data D^E , prior distribution $\mu(x)$.

Output: Recovered utility function u .

- 1: Initialize policy $p^{(0)}(a|x)$, $q^{(0)}(a)$ and $\theta^{(0)}$ for the NN u_θ .
- 2: For iteration $k \geq 1$:
- 3: Collect expert trajectories τ_k^E from D^E and sample trajectories τ_k from $p^{(k)}$.
- 4: Compute the discriminator's loss

$$\mathcal{L}_{\theta^{(k)}} = E^D [\log D_{\theta^{(k-1)}}(\tau_{k-1}^E)] + E^{\tilde{p}} [\log (1 - D_{\theta^{(k-1)}}(\tau_{k-1}))]$$

using $p^{(k-1)}(a|x)$, $q^{(k-1)}(a)$ and $u_{\theta^{(k-1)}}$.

- 5: Update θ from $\theta^{(k-1)}$ to $\theta^{(k)}$ by the loss function $\mathcal{L}_{\theta^{(k)}}$.
- 6: Update $p(a|x)$, $q(a)$ from $(p^{(k-1)}, q^{(k-1)})$ to $(p^{(k)}, q^{(k)})$ by solving an RI problem with utility function $u_{\theta^{(k)}}$ and constraint

$$q^{(k)}(a) = \sum_x \mu_0(x) p^{(k)}(a|x)$$

- 7: Iterate on k until convergence.
-

CHAPTER 9

Conclusions

In this chapter, we conclude by summarizing all theoretical and numerical results presented in this thesis. Related to rational inattention, we formulate an inverse problem to recover agents' utility function from their past behavior; Related to reinforcement learning, our work introduce the costly information acquisition for agent's decision-making problem.

9.1 INVERSE RATIONAL INATTENTION

Assuming that an agent need to acquire costly information before making decisions, we formulate inverse rational inattention problems for both static and dynamic settings. To avoid the ambiguous problem that one optimal policy can be explained by multiple utility functions, we define an equivalent class of utility functions and prove that for one optimal policy, there is a unique equivalent class that can explain it.

We also provide algorithms to solve static and dynamic IRI problems, with an convex penalty function, these algorithms can recover the target utility function with limited amount of observations.

9.2 CONNECTION TO REINFORCEMENT LEARNING

Both RI and RL are decision-making frameworks built on Markov Decision Processes (MDPs), they have similar structures and targets. Each of them has its own advantages, RI framework allows the state to be only partially observed and can be applied to more general settings like non-markovian and non-stationary policy,

while RL algorithms can learn the policy directly without knowing the dynamics of the states and can be applied efficiently to high-dimensional or continuous space.

We hope our work can provide a connection between rational inattention and reinforcement learning communities, distinguish the differences between them, and discuss the advantages that one can learn from the other.

9.3 APPLICATIONS ON ROBO-ADVISING

Compared to the questionnaire-based method currently used by many robo-advising firms to collect clients' preference, our model can learn clients' utility functions directly from their past trading behaviors, thus the recovered utility function should be more accurate and individual-specific.

As presented in Chapter 7, our work can be applied to robo-advising problems with different settings, and the recovered utility functions can help the robo-advisors to better understand their clients' preference and further used to improve the investment performance when there is less information cost.

APPENDIX A

Proof

A.1 PROOFS FOR CHAPTER 3

Proof of Lemma 3.2.1. If there exist some $a \in A$ such that $q_E(a) = 0$, by equation (2.19), $p_E(a|x) = 0$ for all x . Then, by the convention in footnote 1, we have $p_E(a|x) \ln \frac{p_E(a|x)}{q_E(a)} = 0$ which does not contribute to the value of $H(p_E, q_E)$.

For all actions a with $q_E(a) > 0$, because A is a finite space, we can find the constant $C_q \triangleq \min_{a \in A} q_E(a)$. Rewrite $H(p_E, q_E)$ as

$$H(p_E, q_E) = - \sum_{x,a} \mu(x) p_E(a|x) \log p_E(a|x) + \sum_{x,a} \mu(x) p_E(a|x) \log(q_E(a))$$

the first term is always positive and the second term $\sum_{x,a} \mu(x) p_E(a|x) \log(q_E(a)) \geq \log C_q$, thus $H(p_E, q_E) \geq \log C_q$.

On the other hand,

$$H(p_E, q_E) = - \sum_x \mu(x) \sum_a p(a|x) \ln \frac{p(a|x)}{q(a)}$$

since relative entropy $\sum_a p(a|x) \ln \frac{p(a|x)}{q(a)}$ is always non-negative (Theorem 2.6.3 in Cover & Thomas (2012)), $H(p_E, q_E) \leq 0$. Therefore, $H(p_E, q_E)$ always has finite value.

□

Proof of Proposition 3.2.1. When $p = p_E$, Theorem 4 in Blahut (1972) implies that $\operatorname{argmax}_q H(p_E, q) = q_E$, therefore the value of (3.5) is equal to $H(p_E, q_E)$. Meanwhile, because the value of (6.3) is non-negative, the value of (3.4) is bigger or equal to the value of (3.5).

Then, for any optimal solution (p, q) of (3.4), if $p(a|x) > p_E(a|x)$, we can choose $u(x, a) = -\infty$; if $p(a|x) < p_E(a|x)$, we can choose $u(x, a) = \infty$, in either case, the value of (3.4) is $-\infty$, which contradicts the facts that the value of (3.4) is bigger or equal to $H(p_E, q_E)$ and $H(p_E, q_E)$ is finite in Lemma 3.2.1.

□

Proof of Proposition 3.3.1.

$$\text{For any } (p_A, q_A) \in \arg \max_{p, q} H(p, q) + \psi^*(\pi_{p_E} - \pi_p) \quad (\text{A.1})$$

$$= \arg \max_{p, q} \min_u H(p, q) + \psi(u) + \sum_{x, a} (\mu(x)p(a|x) - \mu(x)_E(a|x)) u(x, a) \quad (\text{A.2})$$

Define the Lagrangian $L(p, q, u)$ as

$$L(p, q, u) = H(p, q) + \psi(u) + \sum_{x, a} (\mu(x)p(a|x) - \mu(x)p_E(a|x)) u(x, a).$$

Observe that $L(\cdot, \cdot, u)$ is concave because $H(p, q)$ is joint concave in (p, q) , and $L(p, q, \cdot)$ is convex because ψ is convex. Moreover, $\Delta(X) \times \Delta(A)$ is compact and convex, $\mathbb{R}^{X \times A}$ is convex, it follows from the minimax theorem (?) that

$$\max_{p, q} \min_u L(p, q, u) = \min_u \max_{p, q} L(p, q, u). \quad (\text{A.3})$$

Let $\tilde{u} \in \text{IRI}_\psi(p_E, q_E)$, then \tilde{u} is an optimizer for the right-hand side of (A.3) and $\max_{p, q} L(p, q, \tilde{u}) = \min_u L(p_A, q_A, u)$. However, $\max_{p, q} L(p, q, \tilde{u}) \geq L(p_A, q_A, \tilde{u}) \geq \min_u L(p_A, q_A, u)$, thus these three quantities must be equal. Therefore

$$(p_A, q_A) \in \arg \max_{p, q} L(p, q, \tilde{u}).$$

In other words, (p_A, q_A) is an optimizer for a RI problem with the utility \tilde{u} . This shows that any optimizer for (3.7) is also an optimizer for (3.6).

To prove the reverse statement, take any optimizer (\tilde{p}, \tilde{q}) for the RI problem associated to \tilde{u} , we have $L(\tilde{p}, \tilde{q}, \tilde{u}) = \max_{p,q} L(p, q, \tilde{u})$. Therefore, $(\tilde{p}, \tilde{q}, \tilde{u})$ is a saddle point of (A.3) and $(\tilde{p}, \tilde{q}) \in \arg \max_{p,q} H(p, q) - \psi^*(\pi_{p_E} - \pi_p)$, implying that (\tilde{p}, \tilde{q}) is also an optimizer for (3.7). \square

A.2 PROOFS FOR CHAPTER 4

Proof of Lemma 4.1.1. Suppose utility functions u_1, u_2 are equivalent and (q, γ) is an optimal policy for $RI(u_1)$, we first prove that (q, γ) is also optimal for $RI(u_2)$ by checking the three conditions in Proposition 2.

The first condition is satisfied automatically because it depends on the policy (q, γ) and the prior μ only. To check the second condition, take any $a, b \in B$ and $x \in X$, we have

$$\frac{\gamma^a(x)}{e^{u_2(x,a)/\lambda}} = \frac{\gamma^a(x)}{e^{u_1(x,a)/\lambda} e^{C_x/\lambda}} = \frac{\gamma^b(x)}{e^{u_1(x,b)/\lambda} e^{C_x/\lambda}} = \frac{\gamma^b(x)}{e^{u_2(x,b)/\lambda}}.$$

For $a \in B$ and $c \in A \setminus B$,

$$\sum_{x \in X} \frac{\gamma^a(x)}{e^{u_2(x,a)/\lambda}} e^{u_2(x,c)/\lambda} = \sum_{x \in X} \frac{\gamma^a(x)}{e^{u_1(x,a)/\lambda} e^{C_x/\lambda}} e^{u_2(x,c)/\lambda} \leq 1.$$

Therefore (q, γ) is also optimal for $RI(u_2)$, hence we confirm equivalent utility functions corresponds to the same optimal policy.

To prove the reverse statement, suppose that there are two non-equivalent util-

ity functions u_1 and u_2 which yield the same optimal policy (q, γ) . Either there exist $x \in X, a, b \in B$, and constants $C_{x,a} \neq C_{x,b}$ such that

$$u_2(x, a) = u_1(x, a) + C_{x,a} \quad \text{and} \quad u_2(x, b) = u_1(x, a) + C_{x,b}, \quad (\text{A.4})$$

or there exists $a \in B$ and $c \in A \setminus B$ such that

$$\sum_{x \in X} \frac{\gamma^a(x)}{v_1(x, a) e^{C_x/\lambda}} v_2(x, c) > 1. \quad (\text{A.5})$$

In the first situation, because (q, γ) is optimal for both $RI(u_1)$ and $RI(u_2)$, we have

$$\begin{aligned} \frac{\gamma^a(x)}{e^{u_1(x,a)/\lambda + C_{x,a}/\lambda}} &= \frac{\gamma^a(x)}{e^{u_2(x,a)/\lambda}} = \frac{\gamma^b(x)}{e^{u_2(x,b)/\lambda}} = \frac{\gamma^b(x)}{e^{u_1(x,b)/\lambda + C_{x,b}/\lambda}}, \\ \frac{\gamma^a(x)}{e^{u_1(x,a)/\lambda}} &= \frac{\gamma^b(x)}{e^{u_1(x,b)/\lambda}}. \end{aligned}$$

Combining the two equations above, we obtain

$$\frac{e^{u_1(x,a)/\lambda + C_{x,a}/\lambda}}{e^{u_1(x,b)/\lambda + C_{x,b}/\lambda}} = \frac{e^{u_1(x,a)/\lambda}}{e^{u_1(x,b)/\lambda}},$$

which yields $C_{x,a} = C_{x,b}$ and contradicts with the assumption. Therefore, the first condition in Definition 4.1.1 must hold.

In the second situation,

$$\sum_{x \in X} \frac{\gamma^a(x)}{v_2(x, a)} v_2(x, c) = \sum_{x \in X} \frac{\gamma^a(x)}{v_1(x, a) e^{C_x/\lambda}} v_2(x, c) > 1,$$

which contradicts with assumption that (q, γ) is optimal for $RI(u_2)$. \square

Proof of Proposition 4.2.1 . The Lagrange function of (3.4) is given by

$$L(u, p, q, \xi) = H(p, q) + \sum_{x,a} u(x, a) \mu(x) (p(a | x) - p_E(a | x)) + \xi (\sum_a q(a) - 1)$$

where ξ is the Lagrangian multiplier of the constraint $\sum_a q(a) = 1$.

Since $\mu(x) > 0$ for all x , by the first order condition, we have:

$$p^*(a|x) = p^E(a|x) \quad \text{for} \quad \forall x \in X, a \in A; \quad (\text{A.6})$$

with respect to $p(x, a)$, we have:

$$u^*(x, a) = \ln \frac{p^*(a|x)}{q^*(a)} + 1 \quad \text{for} \quad \forall x \in X, a \in A; \quad (\text{A.7})$$

with respect to $q(a)$, we have:

$$q^*(a) = \sum_x \mu(x) p^*(a|x) \quad \text{for} \quad \forall a \in A. \quad (\text{A.8})$$

When $p^*(a|x) = p^E(a|x)$, $q^*(a) = \sum_x \mu(x) p^E(a|x) = q^E(a)$, thus $u^*(x, a)$ is given by

$$u^*(x, a) = \ln \frac{p^E(a|x)}{q^E(a)} + 1 \quad \text{for} \quad \forall x \in X, a \in A$$

For any u that is equivalent to u^* , i.e., $u(x, a) = u^*(x, a) + C_x$, since $p^*(a|x) = p^E(a|x)$, we have $L(u, p^*, q^*, \xi^*) = L(u^*, p^*, q^*, \xi^*)$, so u is also optimal for (3.4). Then, by Lemma 8, the equivalent class $[u^E]$ is unique for the IRI problem $IRI(p_E, q_E)$.

Therefore, for the interior expert policy (p_E, q_E) , the IRI problem $IRI(p_E, q_E)$

has an unique solution $[u^E]$, where $u^E = \ln \frac{p^E(a|x)}{q^E(a)} + 1$.

□

Proof of Lemma 4.3.1. We will prove the optimality of (p, q) for $RI(\hat{u})$ by checking those three conditions in Proposition 4.1.1.

The first condition holds automatically because the prior μ is unchanged;

The second condition can be verified by the fact that $\hat{u}(\cdot, a) = u(\cdot, a)$ and $\hat{u}(\cdot, b) = u(\cdot, b)$ for $\forall a, b \in B(q_E)$;

The last condition holds thanks to $u(x, a) = \hat{u}(x, a)$ and the choice of $\hat{u}(x, c)$ in condition 2. □

Proof of Proposition 4.3.1. Without the loss of generality, we assume the first $|B(q_E)|$ actions are in the considerate set $B(q_E)$, then we can write the target utility function u_E as two parts:

$$u^E = \left[\underbrace{u_B^E}_{|X| \times |B(q_E)|}, \quad \underbrace{u_{A/B}^E}_{|X| \times (|A| - |B(q_E)|)} \right]$$

and define a truncated policy $(\hat{p}(a|\cdot), \hat{q}(a))$ on $X \times B(q_E)$ as

$$(\hat{p}(a|\cdot), \hat{q}(a)) = (p^E(a|\cdot), q^E(a)) \quad \text{for } a \in B(q_E)$$

Since $(\hat{p}(a|\cdot), \hat{q}(a))$ is an interior expert policy for u_B^E , we can uniquely find $[u_B^E]$ by Proposition 10, and choose $u_{A/B}^E$ such that for any $c \in A \setminus B$ such that

$$\sum_{x \in X} \frac{\gamma^a(x)}{u_B^E(x, a)} u_{A/B}^E(x, c) < 1.$$

By Lemma 12, the constructed u^E is a solution for $IRI(p_E, q_E)$ and all the solutions constructed in this way are equivalent by Definition 7. By Lemma 8, the equivalent class $[u^E]$ is unique for $IRI(p_E, q_E)$.

□

A.3 PROOFS FOR CHAPTER 5

Proof of Proposition 5.3.1. Firstly, we define the Lagrangian $L_{dy}(\mathbf{p}, \mathbf{q}, u)$ as:

$$L_{dy}(\mathbf{p}, \mathbf{q}, u) = \psi(u) + H(\mathbf{p}, \mathbf{q}) + \sum_{x,a} \sum_{t=1}^T \beta^{t-1} \mu_t(x) u(x, a) (p_t(a|x) - p_t^E(a|x))$$

Since $L_{dy}(\mathbf{p}, \mathbf{q}, u)$ is jointly concave in (\mathbf{p}, \mathbf{q}) , by the minimax theorem

$$\max_{\mathbf{p}, \mathbf{q}} \min_u L_{dy}(\mathbf{p}, \mathbf{q}, u) = \min_u \max_{\mathbf{p}, \mathbf{q}} L_{dy}(\mathbf{p}, \mathbf{q}, u)$$

Given the convex conjugate function

$$\begin{aligned} & \psi^* \left(\sum_{t=1}^T \beta^{t-1} \mu_t^E(x, a) - \sum_{t=1}^T \beta^{t-1} \mu_t(x, a) \right) \\ &= \max_u -\psi(u) - \sum_{x,a} u(x, a) \left(\sum_{t=1}^T \beta^{t-1} \mu_t(x, a) - \sum_{t=1}^T \beta^{t-1} \mu_t^E(x, a) \right) \end{aligned}$$

then the result follows by the same argument as in Proposition 3.3.1. □

A.4 PROOFS FOR CHAPTER 8

Proof of Lemma 8.2.2. We take the discriminator with a special form of:

$$D_\theta(x, a) = \frac{\frac{\mu_0(x)q(a) \exp\{u_\theta(x, a)\}}{Z}}{\frac{\mu_0(x)q(a) \exp\{u_\theta(x, a)\}}{Z} + p(x, a)} \quad (\text{A.9})$$

where $\frac{\mu_0(x)q(a)\exp\{u_\theta(x,a)\}}{Z}$ is the estimated data density which is assumed to follow the Boltzmann distribution, and $p(x, a)$ is a fixed generator density.

The negative discriminator's loss is given by:

$$-\mathcal{L}_{\text{discriminator}}(D_\theta) = E^{\mathcal{D}}[\log D_\theta(\tau)] + E^p[\log(1 - D_\theta(\tau))] \quad (\text{A.10})$$

$$\begin{aligned} -\mathcal{L}_{\text{discriminator}}(D_\theta) &= E^{\mathcal{D}}[\log D_\theta(\tau)] + E^p[\log(1 - D_\theta(\tau))] \\ &= E^{\mathcal{D}}\left[\log \frac{\mu_0(x)q(a)\exp\{u_\theta(x,a)\}}{Z}\right] \\ &\quad - E^{\mathcal{D}}\left[\log \frac{\mu_0(x)q(a)\exp\{u_\theta(x,a)\}}{Z} + p(x,a)\right] + E^p[\log p(a|s)] \\ &\quad - E^p\left[\log \frac{\mu_0(x)q(a)\exp\{u_\theta(x,a)\}}{Z} + p(x,a)\right] \\ &= -\log Z + E^{\mathcal{D}}[\log(\mu_0(x)q(a)\exp\{u_\theta(x,a)\})] + E^p[\log p(a|s)] \\ &\quad - 2E^\mu[\log \mu(x,a)] \end{aligned} \quad (\text{A.11})$$

where $\mu = \frac{1}{2} \frac{\mu_0(x)q(a)e^{u_\theta(x,a)}}{Z} + \frac{1}{2}p$.

By F.O.C with respect to Z , we have

$$\frac{1}{Z} + E^\mu\left[\frac{-\frac{1}{Z^2}\mu_0(x)q(a)\exp\{u_\theta(x,a)\}}{\mu}\right] = 0$$

$$Z = E^\mu\left[\frac{\mu_0(x)q(a)\exp\{u_\theta(x,a)\}}{\mu}\right]$$

Thus, the optimal Z is exactly the importance sampling estimate of the partition function we got in equation (8.9).

With this optimal value of Z , we can show that:

$$\frac{\partial}{\partial \theta} - \mathcal{L}(\theta) = E^{\mathcal{D}} \left[\frac{\partial}{\partial \theta} u_{\theta}(x, a) \right] - E^{\mu} \left[\frac{\frac{\mu_0(x)q(a)e^{u_{\theta}(x,a)}}{Z}}{\mu} \frac{\partial}{\partial \theta} u_{\theta}(x, a) \right] \quad (\text{A.12})$$

which matches gradient shown in (8.11).

The generator's loss is given by

$$\mathcal{L}_{\text{generator}}(p) = \mathbb{E}^p[\log(1 - D_{\theta}) - \log(D_{\theta})] = \mathbb{E}^p \left[\log\left(\frac{p}{\mu}\right) - \log\left(\frac{\mu_0(x)q(a)e^{u_{\theta}(x,a)}}{Z\mu}\right) \right] \quad (\text{A.13})$$

Since $\log Z$ is fixed during the optimization of the generator, it can be eliminated to get:

$$\begin{aligned} \mathcal{L}_{\text{generator}}(p) &= \mathbb{E}^p [\log(p(x, a)) - \log(\mu_0(x)q(a)) - u_{\theta}(x, a)] \\ &= \mathbb{E}^p \left[\log \frac{p(a|x)}{q(a)} - u_{\theta}(x, a) \right] \end{aligned} \quad (\text{A.14})$$

Thus minimizing this loss is equivalent to maximize a RI problem with utility function u_{θ} .

□

APPENDIX B

Supplementary results

B.1 MEAN-VARIANCE EXAMPLE

B.1.1 Impact of marginal risk aversion θ Table B.1: Impact of marginal risk aversion θ , true values

Setting	u^E	p^E
$\lambda = 0.5, \theta = 0.02$	$\begin{pmatrix} 0.69833 & 1.12010 & 1.74687 \\ 0.20504 & 0.24985 & 0.23903 \\ -0.26464 & -0.48815 & -0.83425 \end{pmatrix}$	$\begin{pmatrix} 0.007 & 0.000 & 0.993 \\ 0.050 & 0.000 & 0.950 \\ 0.151 & 0.000 & 0.849 \end{pmatrix}$
$\lambda = 0.5, \theta = 0.04$	$\begin{pmatrix} 0.69480 & 1.11030 & 1.72178 \\ 0.15822 & 0.11980 & -0.09390 \\ -0.27114 & -0.50620 & -0.88046 \end{pmatrix}$	$\begin{pmatrix} 0.159 & 0.000 & 0.841 \\ 0.709 & 0.000 & 0.291 \\ 0.833 & 0.000 & 0.167 \end{pmatrix}$
$\lambda = 0.5, \theta = 0.05$	$\begin{pmatrix} 0.69304 & 1.10540 & 1.70924 \\ 0.13482 & 0.05478 & -0.26036 \\ -0.27438 & -0.51523 & -0.90356 \end{pmatrix}$	$\begin{pmatrix} 0.240 & 0.000 & 0.760 \\ 0.840 & 0.000 & 0.160 \\ 0.890 & 0.000 & 0.110 \end{pmatrix}$
$\lambda = 0.5, \theta = 0.08$	$\begin{pmatrix} 0.68775 & 1.09070 & 1.67161 \\ 0.06459 & -0.14030 & -0.75975 \\ -0.28413 & -0.54230 & -0.97287 \end{pmatrix}$	$\begin{pmatrix} 0.400 & 0.000 & 0.600 \\ 0.961 & 0.000 & 0.039 \\ 0.950 & 0.000 & 0.050 \end{pmatrix}$
$\lambda = 0.5, \theta = 0.1$	$\begin{pmatrix} 0.68422 & 1.08090 & 1.64652 \\ 0.01777 & -0.27035 & -1.09268 \\ -0.29063 & -0.56035 & -1.01908 \end{pmatrix}$	$\begin{pmatrix} 0.469 & 0.000 & 0.531 \\ 0.982 & 0.000 & 0.018 \\ 0.963 & 0.000 & 0.037 \end{pmatrix}$

Table B.2: Impact of marginal risk aversion θ , performance

Setting	$MSE(u)$	$RE(u)$	$MSE(p^{rec})$	$RE(p^{rec})$	$RE(p^{emp})$
$\lambda = 0.5, \theta = 0.02$	0.0078	46%	10^{-5}	8.2%	7.8%
$\lambda = 0.5, \theta = 0.04$	0.003	6.9%	10^{-4}	3.0%	2.9%
$\lambda = 0.5, \theta = 0.05$	0.0039	6.6%	10^{-5}	3.4%	3.3%
$\lambda = 0.5, \theta = 0.08$	0.0078	7.2%	10^{-4}	4.5%	4.4%
$\lambda = 0.5, \theta = 0.1$	0.012	7.6%	10^{-4}	5.7%	5.5%

A risk seeking investor ($\theta = 0.02$) will choose to take an aggressive strategy in

almost all states, in this case, since some action can barely be observed with limited observations (for example, the action of choosing the conservative strategy in a boom market), it would be very hard to recover the corresponding utility, thus the recovered utility has relatively higher error as shown in Table B.2;

As the risk aversion increasing, the investor will start to choose between a conservative strategy and a aggressive strategy based on his understanding of the current market. As the investor becomes more risk averse ($\theta = 0.1$), he is very likely to choose a conservative strategy, especially in a normal or recession market.

As shown in Table B.2, for most cases, our recovered utility could be quite close to the true utility; And the recovered policies in all cases are at the same accuracy-level as the input empirical policies.

B.1.2 impact of marginal information cost λ

Table B.3: Impact of marginal information cost λ , true values

Setting	u^E	p^E
$\lambda = 0.05, \theta = 0.05$	$\begin{pmatrix} 0.69304 & 1.10540 & 1.70924 \\ 0.13482 & 0.05478 & -0.26036 \\ -0.27438 & -0.51523 & -0.90356 \end{pmatrix}$	$\begin{pmatrix} 0.000 & 0.000 & 1.000 \\ 1.000 & 0.000 & 0.000 \\ 1.000 & 0.000 & 0.000 \end{pmatrix}$
$\lambda = 0.2, \theta = 0.05$	$\begin{pmatrix} 0.69304 & 1.10540 & 1.70924 \\ 0.13482 & 0.05478 & -0.26036 \\ -0.27438 & -0.51523 & -0.90356 \end{pmatrix}$	$\begin{pmatrix} 0.013 & 0.000 & 0.987 \\ 0.889 & 0.000 & 0.111 \\ 0.979 & 0.000 & 0.021 \end{pmatrix}$
$\lambda = 0.5, \theta = 0.05$	$\begin{pmatrix} 0.69304 & 1.10540 & 1.70924 \\ 0.13482 & 0.05478 & -0.26036 \\ -0.27438 & -0.51523 & -0.90356 \end{pmatrix}$	$\begin{pmatrix} 0.240 & 0.000 & 0.760 \\ 0.840 & 0.000 & 0.160 \\ 0.890 & 0.000 & 0.110 \end{pmatrix}$
$\lambda = 1, \theta = 0.05$	$\begin{pmatrix} 0.69304 & 1.10540 & 1.70924 \\ 0.13482 & 0.05478 & -0.26036 \\ -0.27438 & -0.51523 & -0.90356 \end{pmatrix}$	$\begin{pmatrix} 0.534 & 0.003 & 0.463 \\ 0.823 & 0.003 & 0.174 \\ 0.855 & 0.002 & 0.143 \end{pmatrix}$
$\lambda = 5, \theta = 0.05$	$\begin{pmatrix} 0.69304 & 1.10540 & 1.70924 \\ 0.13482 & 0.05478 & -0.26036 \\ -0.27438 & -0.51523 & -0.90356 \end{pmatrix}$	$\begin{pmatrix} 0.000 & 1.000 & 0.000 \\ 0.000 & 1.000 & 0.000 \\ 0.000 & 1.000 & 0.000 \end{pmatrix}$

Table B.4: Impact of marginal information cost λ , performance

Setting	$MSE(u)$	$RE(u)$	$MSE(p^{rec})$	$RE(p^{rec})$	$RE(p^{emp})$
$\lambda = 0.05, \theta = 0.05$	0.069	47%	10^{-6}	0%	0%
$\lambda = 0.2, \theta = 0.05$	0.0036	6.3%	10^{-6}	7.2%	7.0%
$\lambda = 0.5, \theta = 0.05$	0.0039	6.6%	10^{-5}	3.4%	3.3%
$\lambda = 1, \theta = 0.05$	0.018	11.6%	10^{-4}	4.4%	4.3%
$\lambda = 5, \theta = 0.05$	N/A	N/A	0	0%	0%

When marginal cost is negligible ($\lambda = 0.005$), the investor simply chooses the action with highest utility; When marginal cost is too high ($\lambda = 5$), it becomes too hard for the investor to tell the current market condition, and he tends to choose a balanced strategy in all states, in this case, since we only have one chosen action, the MSE and RE cannot be defined.

As the marginal information cost λ increasing but still within a reasonable range (0.2 - 1), the investor will choose between the conservative strategy and the aggressive strategy.

Again, for most cases, our model can recover the investor's utility and policy very well.

BIBLIOGRAPHY

- Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *ICML '04: Proceedings of the Twenty-First International Conference on Machine Learning*. <https://dl.acm.org/doi/proceedings/10.1145/1015330>.
- Abraham, F., Schmukler, S., & Tessada, J. (2019). *Robo-Advisors: Investing Through Machine*. <https://ssrn.com/abstract=3360125>.
- Alsabah, H., Capponi, A., Lacedelli, O. R., & Stern, M. (2019). Robo-advising: Learning investors risk preferences via portfolio choices. In *Journal of Financial Econometrics*, vol. 19, (pp. 369–392). <https://doi.org/10.1093/jjfinec/nbz040>.
- Arimoto, S. (1972). An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18, 14–20.
- Bajari, P., Hong, H., & Ryan, S. P. (2010). Identification and estimation of a discrete game of complete information. *Econometrica*, 78.
- Blahut, E., Richard (1972). Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18, 460–473.
- Caplin, A., & Dean, M. (2013). Revealed preference, rational inattention, and costly information acquisition. In *NBER working paper*, vol. 19318. <https://www.nber.org/papers/w19876>.
- Caplin, A., & Dean, M. (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7), 2183–2203.
- Caplin, A., Dean, M., & Leahy, J. (2018). *Rational Inattention, Optimal Consideration Sets, and Stochastic Choice, forthcoming in Review of Economic Studies*.
- Cover, T. M., & Thomas, J. A. (2012). *Elements of Information Theory, 2th edition*. John Wiley Sons.
- D'Acunto, F., & Rossi, A. G. (2020). *Robo-Advising*. <http://dx.doi.org/10.2139/ssrn.3545554>.
- Finn, C., Christiano, P., Abbeel, P., & Levine, S. (2016a). A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *abs/1611.03852*.
- Finn, C., Levine, S., & Abbeel, P. (2016b). Guided cost learning: Deep inverse optimal control via policy optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, vol. 48, (pp. 49–58).

- Foerster, S., Linnainmaa, J. T., Melzer, B. T., & Previtro, A. (2017). Retail financial advice: Does one size fit all? In *The Journal of Finance*, vol. 72, (pp. 1441–1482). <https://doi.org/10.1111/jofi.12514>.
- Fox, R., Pakman, A., & Tishby, N. (2016). Taming the noise in reinforcement learning via soft updates. In *UAI'16: Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, (pp. 202–211).
- Fu, J., Luo, K., & Levine, S. (2018). Learning robust rewards with adversarial inverse reinforcement learning. In *International Conference on Learning Representations (ICLR)*.
- Geng, S., Nassif, H., Manzanares, C. A., Reppen, A. M., & Sircar, R. (2020). Deep pqr: Solving inverse reinforcement learning using anchor actions. *Proceedings of the 37th International Conference on Machine Learning*.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*. <https://papers.nips.cc/paper/5423-generative-adversarial-nets>.
- Haarnoja, T., Hartikainen, K., Abbeel, P., & Levine, S. (2018a). Latent space policies for hierarchical reinforcement learning. In *ICML'18: Proceedings of the 35th International Conference on Machine Learning*. <https://arxiv.org/pdf/1801.01290>.
- Haarnoja, T., Tang, H., Abbeel, P., & Levine, S. (2017). Reinforcement learning with deep energy-based policies. In *ICML'17: Proceedings of the 34th International Conference on Machine Learning*, vol. 70, (pp. 1352–1361). <https://dl.acm.org/doi/10.5555/3305381.3305521>.
- Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018b). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *ICML'18: Proceedings of the 35th International Conference on Machine Learning*. <https://arxiv.org/pdf/1804.02808>.
- Hardy, M. R. (2001). Information theory and statistical mechanics. *North American Actuarial Journal*, 5, 41–53.
- Ho, J., & Ermon, S. (2016). Generative adversarial imitation learning. In *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, (pp. 4572–4580). <https://dl.acm.org/doi/10.5555/3157382.3157608>.
- Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106, 620–630.

- Kaya, O. (2017). *Robo-advice: a true innovation in asset management*. <https://www.dbresearch.com>.
- Kim, E.-c., Jeong, H.-w., & Lee, N.-y. (2019). Global asset allocation strategy using a hidden markov model. vol. 12. <https://doi.org/10.3390/jrfm12040168>.
- Matejka, F., & McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105, 272–298.
- Miao, J., & Xing, H. (2019). Dynamic rationally inattentive discrete choice: A posterior-based approach. <http://people.bu.edu/miaoj/Discrete15.pdf>.
- Nedic, A., & Ozdaglar, A. (2009). Sub-gradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142, 205–228.
- Ng, A., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In *International Conference on Machine Learning (ICML)*.
- Ng, A. Y., & Russell, S. (2000). Algorithms for inverse reinforcement learning. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, (pp. 663–670).
- Puterman, M. L. (2005). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.
- Ratliff, N., Bagnell, J. A., & Zinkevich, M. (2006). Maximum margin planning. In *International Conference on Machine Learning (ICML)*.
- Recht, B. (2019). A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2, 253–279.
- Rossi, A. G., & Utkus, S. P. (2019). *Who Benefits from Robo-advising? Evidence from Machine Learning*. <http://dx.doi.org/10.2139/ssrn.3552671>.
- Rust, J. (1994). Structural estimation of markov decision processes. *Handbook of Econometrics*, 4, 3082–3139.
- Schulman, J., Levine, S., Abbeel, P., Jordan, M. I., & Moritz, P. (2015). Trust region policy optimization. In *ICML'15: Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, (pp. 1889–1897). <https://dl.acm.org/doi/10.5555/3045118.3045319>.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

- Shannon, C. E. (1959). Coding theorems for a discrete source with a fidelity criterion in institute of radio engineers. In *Institute of Radio Engineers, International Convention Record*, 7.
- Shimosaka, M., Kaneko, T., & Nishi, K. (2016). Modeling risk anticipation and defensive driving on residential roads with inverse reinforcement learning. In *17th International IEEE Conference on Intelligent Transportation Systems*, (pp. 1694–1700).
- Steiner, J., Stewart, C., & Matejka, F. (2017). Rational inattention dynamics: Inertia and delay in decision-making. *Econometrica*, 85, 521–553.
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press Cambridge.
- Tanaka, T., Sandberg, H., & Skoglund, M. (2018). *Transfer-Entropy-Regularized Markov Decision Processes, working paper*. <https://arxiv.org/abs/1708.09096v3>.
- Toussaint, M. (2009). Robot trajectory optimization using approximate inference. In *Proceedings of the 26th Annual International Conference on Machine Learning*. <https://doi.org/10.1145/1553374.1553508>.
- Vo, H. T., & Maurer, R. (2013). Dynamic asset allocation under regime switching, predictability and parameter uncertainty. <http://dx.doi.org/10.2139/ssrn.2165029>.
- Wang, M., Lin, Y.-H., & Mikhelson, I. (2020). Regime-switching factor investing with hidden markov models. *Journal of Risk and Financial Management*, 13.
- Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Machine learning*, vol. 8, (pp. 229–256). <https://doi.org/10.1007/BF00992696>.
- Yu, L. J., & Zhao, H. (2019). Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. In *BMC Medical Informatics and Decision Making*. <https://doi.org/10.1186/s12911-019-0763-6>.
- Ziebart, B. D. (2010). Modeling purposeful adaptive behavior with the principle of maximum causal entropy. In *Doctoral dissertation —Carnegie Mellon University*. <https://www.cs.cmu.edu/~bziebart/publications/thesis-bziebart.pdf>.
- Ziebart, B. D., Maas, A. L., Bagnell, J. A., & Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*. <https://www.aaai.org/Papers/AAAI/2008/AAAI08-227.pdf>.

