Strategic Behavior under Uncertainty in Multiagent Settings

by

Xia Qu

(Under the direction of Prashant Doshi)

Abstract

Sequential decision making under uncertainty involves selecting a sequence of actions in the presence of noise to maximize an agent’s expected utility. In multiagent settings, agents are uncertain not only about their actions’ outcomes, their observations, and states, but also about actions of other agents sharing the environment. Therefore, an agent’s behavior must be strategic and consider these uncertainties. One recognized framework relevant for decision making in multiagent settings is the interactive partially observable Markov decision process (I-POMDP). This research focuses on strategic behavior of humans and normative agents in multiagent settings.

First, I study the behavior of humans in two classes of games and propose several new models of behavioral data collected when humans engaged in these games. The first class is a modified Centipede game for testing human recursive thinking. Recent experiments show that humans predominantly reason at lower levels; however, they display a higher level of reasoning if games are made simpler and more competitive. I model the data using the finitely-nested I-POMDP, appropriately simplified and augmented with models simulating human learning and choice. Results suggest that this process-oriented behavioral modeling provides a good fit of the data. My modeling further showed that humans attribute the
same errors that they themselves make to others. The second class pertains to sequential bargaining where humans are widely observed as deviating from game-theoretic predictions. I construct a suite of new and existing computational process models that integrate different choice models with utility functions. Fairness and limited backward induction, both of which may possibly explain the behavioral deviations, are incorporated. My comparative analyses reveal that limited backward induction plays a crucial role in longer-round games while in shorter-round games, fairness remains the key consideration.

Second, I present new methods for computing the strategic behavior of normative agents in the context of I-POMDPs. A new technique provides the first formalization of planning in finitely-nested I-POMDPs as a probabilistic inference problem. My comprehensive experimental results demonstrate that we may obtain solutions represented as compact finite state controllers whose quality is significantly better than previous policy iteration techniques though convergence may take more time.

**Index words:** Human behavior modeling, Behavioral game theory, Sequential decision making under uncertainty, EM algorithm
Strategic Behavior under Uncertainty in Multiagent Settings

by

Xia Qu

B.M.S., Beijing Forestry University, China, 2006
M.S., The University of Georgia, 2008

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2014
Strategic Behavior under Uncertainty in Multiagent Settings

by

Xia Qu

Approved:

Major Professors: Prashant Doshi

Committee: Walter D. Potter
Adam S. Goodie

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
August 2014
Strategic Behavior under Uncertainty in Multiagent Settings

Xia Qu

July 21, 2014
Acknowledgements

First and foremost, I would like to acknowledge the contributions of my advisor, Dr. Prashant Doshi. He is unquestionably a true innovator in the field of computer science and a great teacher and motivator. I feel that he has prepared me for success as a computer scientist and I have learned a great deal from him. Through his guidance and motivation, my graduate study here at the University of Georgia has been an enjoyable learning experience. Second, I would like to thank my committee members, Dr. Walter Potter for his valuable discussions and Dr. Adam Goodie for providing experimental data.

I would like to thank my labmates and closest friends I have come to meet over the past few years for helping me in my journey, including Ekhlas Sonu, Muthu Chandrasekaran, Roi Ceren, Yu Qiu, Yingke Chen, William Richardson, Kenneth Bogert, Fadel Adoe, Indrajit Das, Kedar Marathe and Maulesh Trivedi.
Preface

My dissertation focuses on strategic behaviors under uncertainty in multiagent settings. These behaviors include those from humans and normative agents. Specifically, for human behaviors, I focused on two classes of strategic games: modified Centipede games and sequential bargaining games, where humans tend to deviate from game theoretic predictions. Different models incorporating human behavioral factors were investigated in order to model data collected on these games. For normative behaviors, I focused on developing a new approximation algorithm in solving existing multiagent decision making framework, I-POMDP.

Together with my advisor Prof. Prashant Doshi, I have attempted to disseminate the research outcome via symposiums, conferences, journals, and posters submissions. The list of papers given below, along with this dissertation, forms an accurate description of the work that I have completed towards my dissertation.

Publication List:


Contents

Acknowledgements v

Preface vi

List of Figures x

List of Tables xv

1 Introduction 1

1.1 Motivation ................................................. 2
1.2 Behavioral Game Theory ................................. 3
1.3 Decision Making in Multiagent Settings ................. 10
1.4 Contributions ........................................... 12
1.5 Organization ............................................. 16

Part I Strategic Behavior of Human Agents in Games 17

2 Related Work 18

2.1 Reasoning Model ........................................... 18
2.2 Decision Model ........................................... 21
2.3 Learning Model .......................................... 22
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4</td>
<td>Social Model</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>Modeling Human Behavior in Modified Centipede Games</td>
<td>26</td>
</tr>
<tr>
<td>3.1</td>
<td>Studies and Data</td>
<td>27</td>
</tr>
<tr>
<td>3.2</td>
<td>Empirically Informed Interactive POMDP</td>
<td>40</td>
</tr>
<tr>
<td>3.3</td>
<td>Performance Analysis</td>
<td>44</td>
</tr>
<tr>
<td>3.4</td>
<td>Summary</td>
<td>65</td>
</tr>
<tr>
<td>4</td>
<td>Modeling Human Behavior in Sequential Bargaining Games</td>
<td>67</td>
</tr>
<tr>
<td>4.1</td>
<td>Sequential Bargaining Games</td>
<td>70</td>
</tr>
<tr>
<td>4.2</td>
<td>Background on Modeling Sequential Bargaining Games</td>
<td>73</td>
</tr>
<tr>
<td>4.3</td>
<td>Dynamic Level Based Choice Models</td>
<td>78</td>
</tr>
<tr>
<td>4.4</td>
<td>Performance Analysis</td>
<td>87</td>
</tr>
<tr>
<td>4.5</td>
<td>Summary</td>
<td>92</td>
</tr>
</tbody>
</table>

**Part II** Strategic Behavior of Normative Agents in Multiagent Settings

5 | Existing Solutions to Multiagent Decision Making | 98 |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>I-POMDP Solution</td>
<td>99</td>
</tr>
<tr>
<td>5.2</td>
<td>Existing Algorithms</td>
<td>101</td>
</tr>
</tbody>
</table>

6 | Multiagent Planning as Inference | 105 |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Background</td>
<td>107</td>
</tr>
<tr>
<td>6.2</td>
<td>Planning in I-POMDPs as Likelihood Maximization</td>
<td>110</td>
</tr>
<tr>
<td>6.3</td>
<td>Computational Complexity</td>
<td>120</td>
</tr>
<tr>
<td>6.4</td>
<td>Experiments</td>
<td>121</td>
</tr>
<tr>
<td>6.5</td>
<td>Summary</td>
<td>125</td>
</tr>
</tbody>
</table>
7 Future Work

7.1 Behavior of Humans ......................................................... 127
7.2 Behavior of Normative Agents .......................................... 128

Bibliography 130
List of Figures

1.1 Student thinks that his teacher thinks that he is taking notes on his laptop. . 7
1.2 An example of centipede game. . . . . . . . . . . . . . . . . . . . . . . . . . . 8
1.3 An example of ultimatum game where the proposer decides how to split $10. $x$ is the amount he offers. The responder then decides whether to accept or reject the offer. A rejection from the responder leaves both players nothing. . 8
1.4 An example of Investor-Trustee game where Investor original has $10, $x$ is the amount he invests which will be tripled. Trustee will decide how to share the $3x$ returns. The amount he chooses to repay to Investor is denoted as $y$. . . . 9
1.5 Mental model of an I-POMDP agent . . . . . . . . . . . . . . . . . . . . . . . . . . 11
3.1 (a) A game tree representation (extensive form) of three-stage game. States of the game are indicated by the letters, A, B, C and D. (b) Arrows denote the progression of play in the game. An action of move by each player causes a transition of the states in the game. (c) An example general-sum game used in the study. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 28
3.2 (a) A myopic player II decides on her action by comparing the payoff at state B with that at C. Here, \( B \prec C \) denotes a preference of C over B for the player whose turn it is to play and \( B : C \) denotes the rational choice by the appropriate player between states B and C. Thus, player I exhibits first-level reasoning. (b) If player II is predictive, she reasons about I’s actions. Player I then exhibits second-level reasoning in deciding her action at state A.

3.3 General-sum games in Study 1: (a) Mean achievement score of the participants for both opponent groups across test blocks. Notice that subjects generally expected their opponents to play at zero level far more than at first level. Some learning also occurred as is evident from the increasing achievement scores. (b) Mean prediction score of the participants for both groups across test blocks. Overall, prediction scores are low and similar for both groups initially. (c) Rationality error rates among participants measured as actions which are not rational given predictions.

3.4 An example of fixed-sum game used in Study 1. The payoffs are probabilities of success for player I. The complements are the probabilities of success for player II.

3.5 Fixed-sum games in Study 1: (a) Mean achievement score of the participants for both groups across test blocks. (b) Mean prediction score of the participants for both groups across test blocks. Subjects generally expected their opponents to play at first level rather than at zero level. (c) Rationality error rates among participants.

3.6 A four-stage fixed-sum sequential game in (a) matrix and (b) tree format.

3.7 Study 2: (a) mean achievement score and (b) mean prediction score of the participants for three groups across test blocks. (c) rationality error.
3.8 Study 1, General-Sum game, first model: Comparison of initial model predictions with actual data for both groups. Comparisons of the mean achievement score (top), and mean prediction score (bottom) are shown. Notice the differences in the first test block where subjects started off by making an unusually high or low number of rationality errors.

3.9 Study 1, Fixed-Sum game, first model: Comparison of initial model predictions with actual data for both groups. Mean achievement (top), and mean prediction scores (bottom) are shown.

3.10 Study 1, second model: Comparison of model predictions with actual data for both groups in the general-sum game (top), and fixed-sum game (bottom).

3.11 Study 1, third model: Comparison of model predictions with actual data for both groups in the general-sum game (top), and fixed-sum game (bottom).

3.12 Comparison of $I$-POMDP$^\gamma,\lambda_{i,3}$ based simulations with actual data in test folds: (a) Mean achievement scores and (b) Mean prediction scores.

3.13 Comparison of $I$-POMDP$^{\gamma_1,\lambda_1,\lambda_2}_{i,3}$ based simulations with actual data in test folds: (a) Mean achievement scores and (b) Mean prediction scores. Notice that the model-based scores exhibit similar values and trends as the experiment data, albeit the larger discrepancy in the first block. Additionally, the trend in the model-based scores appears more smooth compared to the actual data, possibly due to the large number of simulation runs.

3.14 Comparison of wFP$^{\phi,\lambda}_i$ based simulations with actual data in test folds: (a) Mean achievement scores and (b) Mean prediction scores. Note the poor simulation performance, more so in the context of the myopic opponent group, despite some evidence of initial learning as seen from its prediction scores.
4.1 An example of a 5-round SBG from Neelin et al. (1988) with a pie size of $15 and discount factors, (0.34, 0.34). The split, \((a, b)\), denotes that portion \(a\) is allocated to agent \(i\) and \(b\) is allocated to agent \(j\). The equilibrium offer of \(i\) is to propose slightly more than $3.75 to \(j\) at the opening round.

4.2 Level 5 belief in DLK for a 5-round SBG where \(\tau = 1.5\). The decimals on the edges are the probabilities of each level. Notice that each agent has two roles at levels other than 0 in the final round and 5 in the opening round: the responder in round \(t\) and the proposer in round \(t + 1\). An agent’s level \(l\) rule is a best response to her opponent’s level \(l - 1\) rule.

4.3 The optimal split at each level for the level-\(k\) model using NORM. The pair \((U_i, U_j)\) at each level represents the optimal split for agent \(i\) and \(j\), respectively.

4.4 An agent at each level ascribes decision rules of all lower levels to the other, and maintains a belief distribution over them.

4.5 An agent at each level ascribes decision rules of all lower levels to the other, and maintains a belief distribution over them.

4.6 Distribution of opening offer proportions for test data set, NSS\(^5,1\), by the different models and from observed data. Notice that RQR’s distribution is diffused and compares poorly with the observed distribution, whose support ranges from 0.1 to 0.6.

6.1 DBN for I-POMDP\(_{i,l}\) with \(i\)’s level-\(l\) policy represented as a standard FSC, with “node state” denoted by \(n_{i,l}\). The dotted line in right highlights a clique of the graphical model; the junctions are \((s^t, n_{i,l}^t, n_{j,l-1}^t)\) for \(t = 1, \ldots, T\), which is used in the E-step.
6.2 Interleaved improvement of the FSCs across the different levels demonstrates anytime behavior for $I$-POMDP$_{1,2}$ in the tiger (top) and the larger money laundering contexts (bottom). ................................. 122

6.3 BCD causes substantially larger improvements in the initial iterations until we are close to convergence. We use 3 blocks for both the tiger (top) and money laundering problems (bottom). ................................. 123

6.4 Comparison of EM with greedy and BCD with $I$-BPI for tiger (top) and money laundering (bottom). ................................. 124
List of Tables

3.1 Parameter values obtained by learning $\gamma$ from predictions and $\lambda$ from decisions in Study 1. 46

3.2 Study 1, first model: Goodness of the fit measured using the mean squared residual error (MSE). We include the random model for assessing significance. All MSEs have been rounded to five decimal places. 49

3.3 Study 1, second model: Parameter values obtained on the subjects’ decision data. Observe the difference between these and the parameter values in Table 3.1, especially of $\gamma$. 51

3.4 Study 1, second model: Goodness of the fit of the model with the study data compared with the random model for assessing significance. All MSEs have been rounded to five decimal places. 53

3.5 Study 1, third model: Parameter values obtained on the subjects’ decision data from the two studies. 54

3.6 Study 1, third model: Goodness of the fit of our model with the study data. MSE for the rationality errors is shown as well. MSE for the prediction score remains approximately the same as before. 54

3.7 A normal form representation of the game in Fig. 3.6. The payoffs are the probabilities of winning for player 1. Action, stay $s$, denotes that the first stay occurs at state $s$ for the corresponding player. 58
3.8 Average parameter values learned from the training folds of the experiment
data for the three candidate models. 

3.9 Log likelihood of the different models evaluated on the test folds using param-
eters as shown in Table 3.8.

3.10 MSE of the different models in comparison with the experiment data. Differ-
ence in MSE of the achievement score for the myopic group between the two
I-POMDP models is significant.

4.1 Game designs and collected data on SBG for our modeling. Values in the
column, Game repetitions, indicate the number of games played by each
subject in the pool. Values in the columns, Mean opening and Game theoretic prediction, are fractions of the opening pie size. The superscripts $n, m$
in NSS$^{n,m}$ or OR$^{n,m}$ represent the number of rounds and the game repetitions,
respectively. NSS data is from Neelin et al. (1988) and OR data is from Ochs
and Roth (1989).

4.2 Utilities of two offers for the proposer when using three different social utility
models: NORM, ERC and FSC. The total pie size is $15$.

4.3 Parameters for all combinations of the choice and utility models, obtained as
maximum likelihood point estimates. A space indicates that the parameter is
not applicable. Highest likelihood for each model is indicated in bold.

4.4 Likelihoods for the different choice models integrated with the three utility
models. We include the fit on the training data, NSS$^{5,4}$, for completeness.
Highest likelihood for each data set is indicated in bold. * annotates likeli-
hoods in a row whose difference is not significant.
4.5 Opening offers proportions (along with standard error) for the different experiments as observed from the data and obtained from the model predictions. We include the predictions by a random model as well for comparison.
Chapter 1

Introduction

Sequential decision making under uncertainty involves selecting a sequence of actions in the presence of noise to maximize an agent’s expected utility. Partially observable Markov decision processes (POMDPs) [88, 50, 56] offer a formal approach for accomplishing this in a single agent setting. In a multiagent setting, in addition to uncertainty over its own action and observation, as well as environment state, the agent must also consider uncertainty about actions of other agents. Agent’s behaviors therefore need to be strategic, taking all of these uncertainties into consideration. Relevant research in this setting has converged around two complementary, general frameworks, namely, the decentralized partially observable Markov decision process (DEC-POMDP) [7], which is restricted in contexts where the agents are cooperative, and the interactive partially observable Markov decision process (I-POMDP) [25, 41], which models decision making from the subjective view of an individual agent in both cooperative and noncooperative contexts [26].

These decision making frameworks provide optimal solutions to normative agents under the assumption that they are rational, or utility maximizing. Humans, on the other hand, are well known to be limited in reasoning power [51, 43]. It is widely observed in various studies that when playing strategic games humans do not always choose the rational action
predicted by theoretic models.

1.1 Motivation

Since the term artificial intelligence was coined in 1956, researchers have put enormous effort into this new field, though many have various definitions of it. As summarized by Russell and Norvig [83], eight definitions are laid out along two dimensions. Focusing on the dimension of thought processes and reasoning, as opposed to behavior, those definitions can be classified into two categories, thinking rationally and thinking like humans. However, the development of research along these two lines is not on the same pace. Most research has been done along the line of thinking rationally. For example, in the area of multiagent decision making, researchers have proposed two prominent frameworks, DEC-POMDP for teamwork and I-POMDP for cooperative and noncooperative settings. Each framework offers a principled and theoretically sound formalism for decision making under uncertainty with guarantees of optimality of the solution under different circumstances. The underlining assumption for these frameworks is that agents are rational. Therefore, these frameworks cannot be directly used for agents who are trying to mimic human reasoning, as humans are well-known to be limited in reasoning power [51, 43] and cannot always make optimal decisions.

Real-world applications of decision making often involve mixed settings that consist of humans and human-controlled agents. Examples of these applications include UAV reconnaissance in an urban operating theater and online negotiations involving humans. For these applications, in order to interact with humans, we need to have a framework that can reason like a human. The existing frameworks are largely purely rational and, thus, would lead to non-optimal solutions when humans are present in the environments, as humans often violate the underlining rationality assumption.

In this dissertation, I study the behavior of humans in two classes of strategic games.
These games are controlled and well-designed with simple settings such that human behavioral data that reflect how humans make strategic decisions can be easily collected. We focus on developing process-oriented and principled computational models with empirical support whose predictions are consistent with the observed data. We expect that these models can provide insight into human cognitive reasoning, which may help build computer agents emulating human decision making and design a general human-like agent.

As a comparison, behavior of normative agents which can be solved by frameworks such as DEC-POMDP and I-POMDP in multiagent settings has been the focus of research for a long time. Similar to any framework extended from POMDP, these frameworks also suffer from two curses [73]: the curse of dimensionality and the curse of history. Both curses limit the solutions for these frameworks from scaling to larger problems. Although different algorithms have been developed for both frameworks, we seek to find one which can scale up to larger problems and provide good quality solutions.

1.2 Behavioral Game Theory

Game theory is the mathematical study of interaction among self-interested agents [87] which usually happens in the context of strategic games. These games have been studied extensively by researchers from different areas, including economics, computer science, and AI. Game theory assumes that agents in these games are rational and it provides principled methods for identifying rational strategies for agents. One popular strategy profile in game theory is the Nash equilibrium, where no agent can benefit by changing her strategy while the other agents keep their strategies unchanged [69]. However, when it comes to human play, game theory provides poor predictions of the behavior of humans. For example, in a well-known game called the Prisoners’ Dilemma (PD), the Nash equilibrium strategy is to defect for both agents. Human subjects are observed to choose to cooperate about half the
1.2.1 Existing Models

Researchers have been aware of this disparity between theoretic predictions and human behaviors in games and have proposed different behavioral models for predicting human play. We review some of the models below.

1) **Quantal Response Equilibrium (QRE):** Humans do not always follow the rational, or best, strategies that maximize their utility rather they choose sub-rational ones with some probability. Hence, it is not a deterministic model, but instead specifies a probability distribution over all possible strategies. The probability of following any strategy is positively related to the utility of the strategy [61, 62].

QRE is based on the notion that “better strategies should be played more often than worse strategies, but the best strategies are not always played”. The most common specification for QRE is its logit equilibrium. It has been frequently used by researchers trying to explain noise and experience effects [61, 62, 45, 20, 23, 98, 29, 77] exhibited in behavior of humans in different settings.

2) **Level-k:** This model assumes that agents adopt strategies based on their reasoning levels. These levels are built iteratively from level 0, where an agent at level 0 simply chooses a strategy randomly. A level-1 agent believes her opponents are at level 0 and chooses a best response to uniform play. Similarly, an agent at any other level believes all other agents are at one level lower than herself and chooses the best response to that level strategy [90, 91].

The level-k model assumes that each agent follows a rule, or level, and the agent is always ‘smarter’ than her opponents by choosing strategies from a higher level. It is mostly used in normal-form games [90, 91, 21, 98]. Ho and Su [53] developed a dynamic
level-k model where an agent is not fixed with a specific level but forms a belief over her opponents’ levels and chooses a level strategy which best responds to her belief. In this model, depending on her belief, an agent might choose a strategy from a level which is lower than some of her opponents’ levels. Ho and Su use this model to explain human behavioral deviation from backward induction in sequential games.

3) **Cognitive Hierarchy (CH):** Similar to the level-k model, each agent is associated with a level, levels are built iteratively starting from level 0, and an agent at level 0 chooses a random play strategy. However, rules at all other levels are defined differently. For CH model, an agent at any level other than level 0 believes that all her opponents are at levels lower than hers. The agent therefore chooses a strategy which best responds to strategies at all lower levels assuming a distribution over these lower levels [18].

Camerer et al. [18] employ a Poisson distribution which contains a single parameter as the distribution of lower levels, and call this model Poisson cognitive hierarchy (PCH). They test the performance of the model on several normal-form games, and conclude an average of 1.5 level fits data best in most of their games [18].

4) **Quantal Level-k:** Quantal level-k is a hybrid model which combines the previously mentioned QRE model and level-k model together. An agent at level 0 still chooses a random play strategy while an agent at any other level now chooses a strategy that quantal responds to the strategy at one lower level than hers [90].

Wright and Leyton-Brown [98] compared the performance of these four models on a large collection of data in normal-form games. They conclude that the quantal level-k model has the best overall performance.

5) **Fictitious Play:** Fictitious play is a belief learning model where each agent maintains a frequency count on strategies adopted by her opponents at each round. This frequency
count can be seen as the agent’s belief. At end of each round game, her belief will be updated based on the observed strategies from her opponents. The agent chooses a strategy which is a best response to her belief [12, 79].

In fictitious play, each agent’s belief is a history of her opponents’ strategies. The agent assumes that her opponents in the next game will play according to their strategy histories. Weighted fictitious play is a modified version based on fictitious play where history strategies are discounted in the updated belief [19]. In this setting, strategies observed at \( t \) round back from now are weighted by \( \phi^{t-1} \) in the current belief, assuming \( \phi \) is the discounting factor. This changing is consistent with the observations that past strategies are eventually forgotten by humans. Weighted fictitious play reduces to Cournot play when \( \phi = 0 \), that is, the agent best responds to her opponents’ previous round strategies, forgetting all their previous history strategies; it merges to the original fictitious play when \( \phi = 1 \) with all history strategies equally unforgotten.

These models we reviewed are the basic ones that are used frequently in research on predicting how humans behave in strategic situations.

1.2.2 Recursive Reasoning

Strategic recursive reasoning of the form “what do I think that you think that I think \( \cdots \)” arises naturally in multiagent settings. Specifically, an agent’s rational action in a two-agent game often depends on the action of the other agent, which, if the other is also rational, depends on the action of the subject agent. In cognitive science, theory of mind (TOM) [31, 72] seeks to attribute beliefs, desires, and other mental states to interacting humans. A key aspect of this modeling is the depth of the recursive reasoning that is displayed by human agents. Initial investigations on the depth of strategic reasoning displayed by humans[51, 18] show that humans generally reason at only the first or second level. Typically, the first level,
which attributes immediate reward maximization but no recursive reasoning to others, is more prominent. Evidence of these low levels of reasoning is not surprising since humans are limited by bounded rationality. Recent experiments confirm that humans predominantly reason at lower levels; however, they display a higher level of reasoning if games are made simpler and more competitive [43].

In strategic games, assumptions of common knowledge of elements of the game tend to reduce the need for recursive reasoning by agents. However, not all elements may be common knowledge. For example, an agents belief is private particularly in a noncooperative setting. Multiple decision-making frameworks such as the recursive modeling method [42] and I-POMDP [41] model recursive beliefs as an integral aspect of agents decision making in multiagent settings.

1.2.3 Social Factors

Besides humans limited reasoning capacity, another reason behind human behavioral deviations from theoretic predictions is that humans’ utilities may be affected by social factors. We list some of the social factors along with different games in which these factors are believed to help explain human behaviors.

1) **Altruism:** In Centipede games where the rational strategy is to “take” immediately, a
high percentage of “pass” strategies was observed [60]. McKelvey and Palfrey explain the deviation by assuming that there are a small percentage of altruistic players who always take pass in the game. Other normal players, realizing the existence of these altruistic players, may also choose to pass in the first few stages to maximize their outcomes.  

![Figure 1.2: An example of centipede game.](image)

2) **Fairness:** In ultimatum bargaining games, a rational responder will accept any money offered; a rational proposer anticipates this, will offer the smallest possible amount. However, in experiments, researchers found that most subjects offer slightly less than half and lower offers are rejected half the time [81]. One possible explanation is fairness. Players may define a fair split and have a preference for being treated fairly. Similarly, fairness can also be used to explain alternating-offer games where the last round is an ultimatum game.

![Figure 1.3: An example of ultimatum game where the proposer decides how to split $10. x is the amount he offers. The responder then decides whether to accept or reject the offer. A rejection from the responder leaves both players nothing.](image)

---

1Trust can also be used to explain the deviation. Players may trust each other to choose “pass” strategy so that their total prize increase.
3) **Trust:** In an Investor-Trustee game designed by Berg, Dickhaut, and McCabe [6], Investor decides how much to invest and Trustee later decides how to share the returns, the rational strategy is to keep all returns for Trustee and to invest nothing for Investor. The experiments showed that on average Investor put in about half while repaid was slightly less than what was invested. Berg et al. explained this observation by trust, that is, the willingness that Investor has to believe her Trustee will return back.

![Diagram of the Investor-Trustee game](image)

Figure 1.4: An example of Investor-Trustee game where Investor original has $10, $x$ is the amount he invests which will be tripled. Trustee will decide how to share the $3x$ returns. The amount he chooses to repay to Investor is denoted as $y$.

4) **Reciprocity:** This means rewarding friendly actions and punishing hostile actions [16]. The former one is referred to positive reciprocity while the latter one is negative reciprocity. Some of the games we mentioned above can also be explained using reciprocity. For example, in the ultimatum game, a responder rejecting lower offers is negative reciprocity; in the Investor-Trustee game, the repayment back from Trustee can be explained by positive reciprocity where Trustee rewards Investor’s investment.

The factors discussed above are not independent, under some circumstances, they may seem related or equivalent to each other. For example, fairness and reciprocity in ultimatum games [81], trust and reciprocity in Investor-Trustee games [6].
1.3 Decision Making in Multiagent Settings

Sequential decision making under uncertainty involves selecting a sequence of actions in the presence of noise to maximize an agent’s expected utility. In multiagent settings, agents are uncertain not only about the outcomes of their actions, their observations, and the environment state, but also about the actions of other agents sharing the environment. Therefore, an agent’s behavior must be strategic and consider these uncertainties. DEC-POMDP and I-POMDP are two recognized frameworks for multiagent decision making. Compared to DEC-POMDP, which is applicable in settings where agents are cooperative with each other, I-POMDP can be used in cooperative and noncooperative environment. We review I-POMDP framework in detail.

I-POMDPs [41] generalize POMDPs [88, 50, 56] to multiagent settings by including other agents models as part of the state space. For clarity, we focus on a setting shared by two agents, i and j. The framework can naturally extend to multiple agents. The I-POMDP for an agent i in a setting with one other agent j is mathematically defined as the tuple:

\[ \text{I-POMDP}_i = (IS_i, A, \Omega_i, T_i, O_i, R_i, OC_i) \]

where:

- \( IS_i \) denotes a set of interactive states defined as \( IS_i = S \times M_j \), where \( S \) is the set of physical states, \( M_j \) is the set of computable models of agent j which contains intentional models, denoted by \( \Theta_j \), and subintentional models. For each intentional models of agent j: \( \theta_j = (b_j, \hat{\theta}_j) \) where \( b_j \in \Delta(IS_j) \) is j’s belief and the frame \( \hat{\theta}_j = (A, \Omega_j, T_j, O_j, R_j, OC_j) \), where the parameters of agent j are defined analogously.
- \( A = A_i \times A_j \) is the set of joint actions of both agents.
- \( T_i \) is the probabilistic transitions between the physical states, \( T_i : S \times A \times S \rightarrow [0, 1] \), which describes the results of the agent’s actions on the physical states of the world.
In the above definition, beliefs are infinitely nested. This poses a challenge for making the framework operational. A natural way to break this infinity in beliefs is to truncate the beliefs to finite levels by defining level 0 beliefs. This leads to a finitely nested framework, I-POMDP$_{i,l}$, with $l$ denoting the level, which approximates the original I-POMDP.

Specifically, level 0 interactive states are just the physical states, $IS_{i,0} = S$, and level 0 beliefs are probability distributions over the level 0 states, $b_{j,0} \in \Delta(IS_{j,0})$. Subsequently,
level 0 models contain level 0 intentional models, each of which consists of a level 0 belief and the frame, $\Theta_{j,0} = \{(b_{j,0}, \hat{\theta}_j)\}$, and subintentional models, $M_{j,0} = \Theta_{j,0} \cup SM_j$. Level 1 interactive states are combinations of the physical states and level 0 models of the other agent, $IS_{i,1} = S \times M_{j,0}$. Level 1 beliefs are distributions over the level 1 interactive states, and level 1 models contain level 1 intentional models and level 0 models, $b_{j,1} \in \Delta(IS_{j,1})$. Higher levels are constructed analogously.

The level $l$ interactive state space contains models of all levels up to $l - 1$. A common simplifying approximation is to consider models of the previous level only.

### 1.4 Contributions

My research focuses on behavior of humans and normative agents in multiagent settings. For behavior of humans, I investigate two classes of strategic games and propose different models of behavioral data collected when humans engaged in these games. These models are constructed according to the characteristics of each class of games. My process-oriented models provide a good fit of the data. For the behavior of normative agents, I present new methods of solving multiagent decision making for normative agents in the context of I-POMDPs. My methods could generate solutions with better quality at the cost of longer running times. In this section, I will summarize the dissertation research that I have accomplished thus far, and give further details in subsequent chapters.

#### 1.4.1 Modeling Human Behavior on Modified Centipede Games

The first class of games I investigate is modified Centipede games, which can be used for testing human recursive thinking. Recent experiments by Goodie et al. [43] show that humans predominantly reason at lower levels; however, they display a higher level of reasoning if games are made simpler and more competitive.
To model the data, I use a finitely-nested I-POMDP as the point of departure for its explicit consideration of recursive beliefs and decision making based on such beliefs. The finitely-nested I-POMDP is then appropriately simplified and augmented with models simulating human learning and choice. Models are explored by supplying data collected at different game points. Results suggest that my process-oriented behavioral modeling provides a good fit of data by displaying lower mean squared error on model predictions than a random model. The simulating measurements using my modeling align with the study data in general. Model performance is improved by adopting a dynamic learning model.

I further test my modeling on data collected from games which can be used to test up to the third level of reasoning. Model performance improves when considering opponent choice error, suggesting humans could attribute this error to others.

The major contributions of this work are listed below:

**Contributions**

- Development of models that use the interactive partially observable Markov decision process as the point of departure, which was appropriately simplified and augmented with well-known models simulating human learning and decision.

- Demonstration of a good fit of the data by our process-oriented behavioral modeling.

- Demonstration of experimental results revealing that humans could be attributing the same error in choice that they themselves make to others.

These contributions are outlined in Chapter 3.

### 1.4.2 Modeling Human Behavior on Sequential Bargaining Games

The second class of games I investigate are sequential bargaining games, where humans are widely observed as deviating from game-theoretic predictions. Previous explanations
on human behavioral deviations have focused on considerations of fairness in the offers, and social utility functions have been formulated to model the data. However, a recent explanation by Ho and Su [53] for observed deviations from game-theoretic predictions in sequential games, such as the Centipede game, is that players engage in limited backward induction.

I present a suite of new and existing computational models that integrate different choice models with utility functions. These include DeBruyn and Bolton’s recursive quantal response with social utility functions, models based on Ho and Su’s dynamic level-$k$, and analogous extensions of the cognitive hierarchy with dynamic components. A Bayesian belief update is integrated to account for learning in repeatedly played games. In total, nine different models, three choice models cross combined with three utility functions, are considered. These models are evaluated on the sequential bargaining game data where games have different rounds. Experiments show that the two level-based models that capture violations of backward induction perform better in longer sequential bargaining games with more rounds. However, on shorter ones, recursive quantal response with social utility functions provides a better fit. These results reveal that both limited backward induction and fairness play important roles in how humans engage in sequential bargaining games; however, the importance of each factor varies with the length of the game.

The major contributions of this work are listed below:

**Contributions**

- Development of two novel constructions of dynamic level-based models that take limited backward induction into consideration for sequential bargaining games.

- Comprehensive and comparative analyses of nine different models, which encompass three different choice models cross combined with three different utility models, on behavioral data pertaining to sequential bargaining games.
• Demonstration of experimental results showing that limited backward induction plays a crucial role in longer games with more rounds, while fairness remains the key consideration for shorter rounds.

These contributions are outlined in Chapter 4.

1.4.3 Multiagent Planning as Inference

I-POMDP is a recognized framework for sequential decision making in multiagent settings. It is applicable to both cooperative and noncooperative contexts. While previous algorithms on optimally solving finitely-nested I-POMDPs tackle the optimization in planning, I present a new technique which formalizes the original planning in I-POMDP as a probabilistic inference problem.

I derive an expectation maximization (EM) solution for the transformed inference problem. An obvious way of applying EM to finitely-nested I-POMDPs is to solve lower levels until convergence before moving to the higher level. In contrast, I facilitate anytime behavior by interleaving the improvement of the solutions at different levels of nesting. One drawback of EM iterative method is its slow convergence rate. To speed up the convergence, block-coordinate descent, an approach which theoretically exhibits faster rates of convergence under considerably relaxed conditions, is explored. I test different EM variants on two problems. The experimental results show that we may obtain solutions represented as compact finite state controllers whose quality is significantly better than previous policy iteration techniques, though convergence may take more time.

The major contributions of this work are listed below:

Contributions

• Development of a new technique which provides the first formalization of planning in finitely-nested I-POMDP as a probabilistic inference problem.
• Development of an approach for speeding up the non-asymptotic rate of convergence of the iterative expectation maximization algorithm.

• Demonstration of comprehensive experimental results showing that we may obtain solutions whose quality is significantly better than previous policy iteration techniques.

These contributions are outlined in Chapter 6.

1.5 Organization

The rest of this dissertation is mainly divided into two parts. In Part I, I study human behavior in two classes of games of strategy and propose several new models of behavioral data collected when humans engaged in these games. Within this part, in Chapter 2, I briefly introduce related works in modeling human behavioral data. In Chapter 3, I introduce the first class of games, a modified Centipede game for testing human recursive thinking. I model the data using the finitely-nested I-POMDP, appropriately simplified and augmented with models simulating human learning and choice. Performance of the models is discussed. The second class of games, sequential bargaining games, is covered in Chapter 4. I construct a suite of new and existing computational process models that integrate different choice models with utility functions. Fairness and a new proposed theory of limited backward induction, both of which may possibly explain the behavioral deviations, are incorporated into these models. In Part II, I present new methods for computing the strategic behavior of normative agents in the context of I-POMDPs. In Chapter 5, I present several existing algorithms for solving I-POMDPs. A new technique transforming planning in finitely-nested I-POMDPs as a probabilistic inference problem is presented in Chapter 6. Finally, I close my dissertation with Chapter 7 where a brief summary of the accomplished work and outlines some avenues of future work are discussed.
Part I

Strategic Behavior of Human Agents in Games
Chapter 2

Related Work

In strategic games, theoretic models, which produce equilibrium, or rational strategies, assume that all players: 1) form beliefs based on what others might do (strategic thinking); 2) best respond to those beliefs (optimization); and 3) adjust best responses and beliefs until they are mutually consistent (equilibrium) [17, 14]. Human behavioral deviations can be explained by their limitations on these three processes. Based on each process limitation, different models have been proposed. One may categorize these models into three: reasoning models, decision models, and learning models, each of which models humans’ limitations on the corresponding process. A category on social models is needed because social factors influence human behavioral deviation in some games (Section 1.2.3).

Although models are classified into four categories, many modeling methods may deal with multiple factors and they integrate models from different categories.

2.1 Reasoning Model

Reasoning models focus on human reasoning, such as what others might do and what beliefs they have on others’ actions. Theory of mind (TOM) which attributes beliefs, desires, and
other mental states to interacting humans is one model on this effort. A key aspect of TOM is the depth of the recursive reasoning displayed by human agents.

Harsanyi [49] recognized that infinite recursive thinking arises naturally among rational players. This infinite recursion leads to difficulty in solving games. In order to avoid dealing with recursive reasoning, he proposed the notion of types and common knowledge of the joint belief over the player types. However, common knowledge is itself modeled using an infinite recursive system [33, 32].

Since Harsanyi’s introduction of abstract agent types, researchers have sought to mathematically define the type system. Mertens and Zamer [63] showed that a type could be defined as a hierarchical belief system with strong assumptions on the underlying probability space. Subsequent work [11, 52] gradually relaxed the assumptions required on the state space while simultaneously preserving the desired properties of the hierarchical belief systems. In a similar vein, Aumann defined recursive beliefs using both a formal grammar [3] and probabilities [4] to formalize interactive epistemology.

Ficici and Pfeffer [37] investigated whether human subjects displayed sophisticated strategic reasoning while playing 3-player, one-shot negotiation games. They proposed two models, reflexive and strategic ones, depending on whether other players are considered in the modeling. Their results showed that strategic models have better performance on fitting data, indicating that human also reason about the other players.

Stahl and Wilson [91] studied human levels of recursive thinking. They found that only 2 out of 48 (4%) of subjects attributed recursive reasoning to opponents while playing 12 symmetric 3×3 matrix games. However, 34% of the subjects ascribed zero-level reasoning to opponents. Corroborating this evidence, Hedden and Zhang [51] utilized a sequential, two-player, general-sum game, sometimes also called the Centipede game, in the literature [10], but the sum of payoffs does not necessarily grow as the game progresses. They found that subjects predominantly began with first-level reasoning. When playing against first-level co-
players, some subjects began to gradually use second-level reasoning, although the percentage of such players remained generally low.

Evidence of recursive reasoning in humans and investigations into the level of such reasoning is relevant to multiagent decision making in mixed settings. In particular, these results are directly applicable to computational frameworks such as RMM [42] and I-POMDP [41], both of which explicitly model recursive beliefs as integral to agents decision making.

Other decision-theoretic approaches have been proposed to model TOM. A Bayesian theory of mind model (BToM) [5, 44] expresses theory of mind as representational beliefs and desires, and captures these attributions as Bayesian inference in a POMDP that does rational planning and belief updating. BToM was employed in controlled scenarios to infer participants initial beliefs and goals from observed data. Pynadath and Marsella [76] developed a social simulation tool called PsychSim, which employs a formal decision-theoretic approach using recursive modeling to provide a computational model of how agents influence each others beliefs. PsychSim was used to analyze toy school bully scenarios. In these studies, beliefs play a key role in the modeling of TOM for human social interactions.

Level-based models, such as level-k and cognitive hierarchy, are built iteratively starting from level 0 where a random play strategy is usually adopted. These models can also be considered reasoning models. They are mostly applied in normal-form games [90, 91, 21, 98]. Wright and Leyton-Brown [98] investigated the performance of different models, including level-k and cognitive hierarchy, for predicting human play in one-shot, normal-form games. They concluded that quantal level-k, a hybrid model of level-k and quantal response model, is the best-fit model for predicting unseen human play in normal-form games. In their recent study [99], they redefine level 0 behavior as a meta-model where level-0 agents construct a probability distribution over actions. The predictive accuracies of iterative models with level-0 meta-model on normal-form games are observed to be largely improved in many cases.

Kawagoe and Takizawa [57] compared the performance of the level-k model, agent quantal
response equilibrium (AQRE) and AQRE with altruistic types on McKelvey and Palfreys data [60]. For majority games, level-k model predicted observed data better than AQRE with altruistic types. Consequently, behaviors in Centipede games may also be explained using self-interest and bounded rationality without resorting to social factors such as altruism.

Ho and Su developed a dynamic level-k model where an agent is not fixed with a specific level but forms a belief over her opponents’ levels and chooses a level strategy best responding to her belief [53]. In this model, depending on her belief, an agent might choose a strategy from levels lower than her opponents’ levels.

2.2 Decision Model

In theoretic models, given beliefs over others’ actions, rational agents should take actions which best responds to their beliefs. However, humans are well known for their bounded rationality and are widely observed to deviate from the optimal actions in this process. Models focus on this deviation are known as decision models.

One simple decision model is to always choose a random play without considering others’ actions. Although this simple model predicts human behaviors poorly, it provides a starting point for constructing more complex models. For example, in level-based models, level 0 rule is usually defined as a random play, as specified in level-k [90, 91] and cognitive hierarchy models [18].

The most popular decision model is the quantal response equilibrium (QRE). As mentioned in Section 1.2.1, instead of always taking the optimal action, agents can take any actions while the probability for taking an action is positively related to the utility. McKelvey and Palfrey [61] first introduced QRE and employed its logistic function to model human errors in normal-form games. They continued to test QRE on sequential games represented in extensive form [62]. Their results demonstrated that QRE is applicable in
different kinds of games.

QRE is frequently incorporated into reasoning models to account for human errors. Wright and Leyton-Brown [98] compared the performance of QRE, level-based models, and the hybrid model. They conclude that the hybrid model which combines QRE with level-based models is the best fit one for predicting human play. De Bruyn et al. analyzed the effect of the fairness factor in sequential bargaining games by exploring different utility functions in a recursive QRE model [23]. Ray et al. investigated the role of finite levels of iterated reasoning and non-selfish utility functions in a POMDP that incorporates QRE on Investor-Trustee games [78].

A simplification on QRE is to adopt the rational action with probability $1 - \epsilon$; while with probability $\epsilon$ choose any other action randomly. However, this model is not consistent with the notion that better strategies, if not rational, should be played more often than worse strategies.

### 2.3 Learning Model

Theoretic predictions on games, though not consistent with human behaviors, can eventually be adopted by humans through long-term learning [15]. In repeated played games, human subjects adjust their strategies after each game. This adjustment is characterized by learning models.

One frequently applied learning model is belief learning. It assumes players update beliefs about what others will do based on history. One such model is fictitious play (Section 1.2.1) where players keep track of the frequency of their opponents’ strategies. Cheung and Friedman [19] introduced weighted fictitious play which generalizes on fictitious play and Cournot play. This model was tested on a variety of laboratory experimental data, all of which are collected from normal-form games. Results indicated that players were quite heterogeneous on
learning rate, which usually decreased in more informative environments. Bayesian learning is another type of belief learning. Players in this model have a prior belief over opponents’ types and can learn the probable types of the opponents from actions of others [55]. Cox et al. [22] tested Bayesian learning on a two-player repeated played normal-form game. Their experimental data reveal that this model does reasonably well at predicting the equilibria that subjects eventually play, suggesting that Bayesian learning can provide an empirically effective solution to the equilibrium selection problem when the players have beliefs with finite support.

Reinforcement learning is another learning model which also learns from history. In contrast to belief learning which learns from others’ history, reinforcement learning learns from players’ previous payoffs. It assumes that strategies are reinforced by their previous payoffs. As pointed out in [15], reinforcement learning is suitable for players with very imperfect reasoning ability or for players who know nothing about unexplored or forgone strategies.

Camera et al. proposed a hybrid learning model, called experience-weighted attraction (EWA), which combines reinforcement learning and belief learning [13]. This hybrid model has two variables, attractions and an experience weight, both of which will be updated at each round of game. EWA can be reduced to weighted fictitious play or reinforcement learning when restricted differently. It has been tested on data collected from games including a series of constant-sum game, the median effort games, and p-beauty contest games. Results show that EWA fits the data better than reinforcement learning in all cases and better than belief learning model in most cases.

Other efforts have also been tried on learning. For example, De Bruyn et al. [23] steadily increased the value of the quantal response parameter at each round game. The increasing in the parameter value represents players’ experience.
2.4 Social Model

As mentioned in Section 1.2.3, different social factors may also be used to explain human behaviors in games.

On observing that people are motivated by both their pecuniary payoffs and their relative payoff standings, Bolton and Ockenfels [68] constructed a model of inequity-aversion, called a theory of Equity, Reciprocity and Competition (ERC). It computes utilities based on payoffs players received and their relative shares. Players in this model prefer a fair payoff. Bolton and Ockenfels provided proofs on explaining human behaviors using equity, reciprocity, or competition in bargaining and market games.

Fehr and Schmidt [34] proposed a similar model of inequity-aversion, or envy and guilt, where players not only care about their own payoffs but also others’ payoffs. This model redefines players’ utilities by incorporating envy and guilt measurements into payoffs. Envy occurs when a player has less payoff than her opponent while guilt occurs when a player has more. The predictions of the model were consistent with the empirical evidence on games they investigated, including ultimatum games, market games, and dictator games. This model was claimed to be applicable to any game. Ray et al. [78] integrated it into their POMDP framework to model human play in multiround Investor-Trustee games.

Both inequity-aversion models take other players’ payoffs into consideration when computing their utilities; however, they are defined differently based on interpretations for inequity. In ERC model, a player’s utility depends on whether her payoff is fair or not, comparing with a reference point; while in Fehr-Schmidt model, it depends on whether every player’s payoff is fair or not.

De Bruyn and Bolton [23] investigated the role of fairness by employing ERC model and Fehr-Schmidt model. They incorporated utility functions from both models into a quantal response equilibrium framework, which recursively computes the expected utilities at each
round. The reported out-of-sample fits and model predictions on multiple data sets are consistent: the two models involving social factors exhibit better performance than the normative one.
Chapter 3

Modeling Human Behavior in Modified Centipede Games

Recursive reasoning of the form “what do I think that you think that I think …” arises naturally in multiagent settings. Initial investigations on the depth of reasoning displayed by humans show that humans generally reason at only the first or second level [51, 18]. Typically, the first level, which attributes immediate reward maximization but no recursive reasoning to others, is more prominent. Evidence of these low levels of reasoning is not surprising since humans are limited by bounded rationality. Recent studies by Goodie et al. [43] confirm that humans predominantly reason at lower levels; however, they display a higher level of reasoning if games are made simpler and more competitive.

Given the availability of behavioral data from Goodie et al.’s studies, we seek to computationally model these data. My objective is to develop principled and empirically supported ways of learning and decision making in controlled interactions. We expect that this will provide general insights toward a process-oriented, computational modeling of human strategic behavior, which will help build agents that emulate human decision making and effectively interact with humans in mixed settings.
In this chapter, we first review the methodology in Goodie et al.’s studies and their behavioral data [43]. We then present models [29, 77, 30, 75] which use a multiagent decision making framework, I-POMDP, as the point of departure, appropriately simplified and augmented with human learning and decision models.

3.1 Studies and Data

Goodie et al. [43] investigated levels of recursive reasoning that subjects generally exhibit in particular interactions. We briefly review two of these studies below.

3.1.1 General- and Fixed-sum Games

Study 1 includes two experiments: general-sum games and fixed-sum games.

General-Sum Game

Goodie et al. select a two-player, three-stage, modified Centipede game with complete and perfect information. A tree representation of the game is shown in Figure 3.1. In this game, player I (the leader) may select to move or stay. The action stay is considered to be the default choice in this game. If player I selects to move, player II (the follower) faces the choice of moving or staying, as well. The game ends by an action of stay from any player or after three stage-transitions or two moves of player I. Actions of all players are perfectly observable to each other.

The outcomes are set as probabilities of gain for each player. Similar to magnitudes, rational choice involves selecting an action that maximizes the probability of success.

For a rational player I, in order to decide whether to move or stay at state A, she must reason about player II’s action at state B, whether player II will move or stay at B. A rational player II’s choice in turn depends on whether player I will move or stay at C. Thus, the game
lends itself naturally to recursive reasoning and the highest level of reasoning is governed by
the length of the game tree. In this game, for player I, she may exhibit up to two reasoning
levels.

In the example illustrated in Figure 3.1(c), a rational player I assuming that player II
is rational and that II believes that I is rational will choose to stay at A. This is because I
thinks that if she moves to B, a rational player II will stay at B to obtain a payoff of 0.6. A
move by player II to C is not rational because player I will then choose to move to D which
would leave only 0.4 for player II.

1) Methodology:

Opponent models: In order to test different levels of recursive reasoning, player II is designed
to be a computer program. This program plays a game in two ways: (i) If player I chooses to move, II decides on her action by simply choosing between the outcomes at states B and C rationally. In this case, player II is myopic, or a zero-level player (see Figure 3.2(a)). (ii) If player I chooses to move, II decides on her action by reasoning what player I will do rationally at C. Based on the action of I, player II will select an action that maximizes her outcomes. In this case, player II is predictive, or a first-level player (see Figure 3.2(b)).

![Game Diagram](image)

Figure 3.2: (a) A myopic player II decides on her action by comparing the payoff at state B with that at C. Here, \( B \prec C \) denotes a preference of C over B for the player whose turn it is to play and \( B : C \) denotes the rational choice by the appropriate player between states B and C. Thus, player I exhibits first-level reasoning. (b) If player II is predictive, she reasons about I’s actions. Player I then exhibits second-level reasoning in deciding her action at state A.

To illustrate, in the game of Figure 3.1(c) if player I decides to move, then she thinks that a myopic player II will move to obtain a payoff of 0.8, while a predictive II will choose to stay because she thinks that player I will choose to move from C to D, if she moved. By choosing to stay, II will obtain an outcome of 0.6 in comparison to 0.4 if she moves.

**Payoff structure:** The rational choice of players in the game in Figure 3.1 depends on the
preferential ordering of the states rather than specific probabilities. Let \( a \prec b \) indicate that the player whose turn it is to play prefers state \( b \) over \( a \). Games that exhibit a preference ordering of \( D \prec C \prec B \prec A \) and \( A \prec B \prec C \prec D \) for player I are trivial because I will always choose to stay in the former case and move in the latter case, regardless of how she thinks that II plays. Furthermore, consider the ordering \( C \prec A \prec B \prec D \) for player I and an ordering \( D \prec B \prec A \prec C \) for II. A myopic opponent will choose to move while a predictive opponent will stay. However, in both these cases player I will choose to move. Thus, such games are not \textit{diagnostic} regardless of whether player I thinks that the opponent is myopic or is predictive, I will select the same action precluding a diagnosis of I’s level of recursive reasoning. Of all the 576 distinct preferential orderings among states that are possible for both players, only 48 are diagnostic and not trivial, e.g., \( B \prec C \prec A \prec D \) for player I and \( A \prec D \prec B \prec C \) for II. For this ordering, player I will move if she thinks that the opponent is myopic, otherwise I will stay if the opponent is thought to be predictive. The game in Figure 3.1(c) follows this preference ordering.

\textbf{Design of task:} Each subject experienced an initial \textit{training} phase of at least 15 games that were trivial or those in which a myopic or predictive opponent behaved identically. These games served to acquaint participants with the rules and goal of the task without unduly biasing them about opponent type. In the \textit{test} phase, each subject experienced 32 games instantiated with outcome probabilities that exhibited the diagnostic preferential orderings. The 32 critical games were divided into 4 blocks of 8 games each. In order to avoid subjects developing a mental set, these games were interspersed with 8 that exhibited orderings such as, \( D \prec C \prec A \prec B \) and \( C \prec B \prec A \prec D \), resulting in a total of 40 games. These games not only serve to distract the participants but also function as \textit{catch} trials allowing the identification of participants who may not be attending to the games.

Approximately half of the participants played against myopic opponents, while the remaining played against predictive ones. In each group, all the participants were presented
with the tree and grid representation of the games with payoffs. All the participants also experienced a screen asking them what they thought the opponent would play and their confidence in the prediction, for each game. This data would be indicative of any learning across the repeated games that may occur.

2) Results: 114 participants completed the test phase. Of these participants, 58 experienced myopic opponents while the remaining 56 experienced predictive opponents.

For each participant, the experiment measured the fraction of games that the subject played the best response choice at A in each test block. The best response is the action choice which is conditionally rational given the type of opponent. For example, in the game in Figure 3.1(c), the best response for player I, if the opponent is myopic, is to move. However, if the opponent is predictive, the best response for I is to stay. Goodie et al. compiled an achievement score which is the proportion of games in which the subject played the best response choice given the opponent type. Subjects’ predictions of their opponent’s actions were also analyzed. Prediction scores were calculated, which reflect the proportion of predictions that are consistent with second-level reasoning. Low scores suggest that participants use less second-level reasoning in making their predictions about how their opponent reasons.

83 participants performed actions consistent with first-level reasoning in the first game while the remaining 31 acted consistently with the second level. Figure 3.3(a) shows the mean achievement scores across all participants in each of the 2 groups. Observe that the score is significantly higher when the opponent is myopic as compared to when she is predictive. Statistical tests \( F(1,112) = 34.31, p < 0.001 \) confirm that participants playing against myopic opponents have statistically significant higher achievement scores compared to predictive opponents across all test blocks. The higher scores imply that a larger proportion of subjects displayed first-level reasoning when acting initially, which increased as the games progressed. They did not expect the opponent to reason about their subsequent
Figure 3.3: General-sum games in Study 1: (a) Mean achievement score of the participants for both opponent groups across test blocks. Notice that subjects generally expected their opponents to play at zero level far more than at first level. Some learning also occurred as is evident from the increasing achievement scores. (b) Mean prediction score of the participants for both groups across test blocks. Overall, prediction scores are low and similar for both groups initially. (c) Rationality error rates among participants measured as actions which are not rational given predictions.

play and acted accordingly. Prediction data collected from the participants about the opponents’ possible action and their confidence in the prediction confirmed this expectation. Figure 3.3(b) shows the mean prediction scores across all participants in each of the two groups. Although the prediction score for the predictive group is higher than that for the myopic group, both are generally low. Participants predominantly started the experiment expecting the other to be myopic. Additionally, some learning took place that was responsive to the opponent with both prediction and achievement scores showing change across
Finally, mean *rationality errors* were computed, which reflect the proportion of times that participants’ choices were not rational given their predictions about the opponents’ decisions. As shown in Figure 3.3(c), the error rates remained generally low with a decrease across test blocks and across both groups.

**Fixed-Sum Game**

Outcomes from general-sum games are remarkably similar to previous results such as those of Hedden and Zhang [51]. They confirm the prominence of low levels of reasoning by humans engaged in general strategic games. The primary motivation behind the second study was to show demonstrations of higher levels of recursive thinking either by default or through learning. Goodie et al. sought to increase the level of reasoning by making the game more competitive and simpler.

The game in this experiment differed from the general-sum one in that the payoffs were probabilities of success for each player that summed to 1. An example game is shown in Figure 3.4.

![Fixed-Sum Game Example](image)

**Figure 3.4:** An example of fixed-sum game used in Study 1. The payoffs are probabilities of success for player I. The complements are the probabilities of success for player II.

1) **Methodology:**

Opponent models and payoff structure: Models of the opponent were set to be identical to those in general-sum games. Similarly, the rational choice of players in the example game
in Figure 3.4 depends on the preferential ordering of states of the game rather than actual probabilities. Of the 24 distinct preferential orderings of the states, only one ordering is diagnostic: \( C \prec B \prec A \prec D \). For this ordering, player I will move if she thinks that her opponent is myopic, otherwise I will stay if her opponent is thought to be predictive. Note that the game in Figure 3.4 follows this preference ordering. To maintain the attention of subjects, the actual probability values are changed while following the diagnostic ordering, and catch trials are included. Remaining orderings are either trivial for players I or II, or not diagnostic.

**Design of task:** Participants were run through a training phase similar to the one in the general-sum games. In the test phase, each subject experienced 40 games instantiated with outcome probabilities that exhibited the diagnostic preferential ordering. The 40 critical games were divided into 4 blocks of 10 games each. In order to avoid subjects developing a mental set, these games were interspersed with 40 catch trials that exhibited the orderings, \( C \prec A \prec B \prec D \) and \( D \prec B \prec A \prec C \). All participants were presented with the tree and grid representation of the games. All participants experienced the screen asking them what they thought the opponent would play and their confidence in the prediction for the games.

2) **Results:** Of the 118 participants who completed the test phase, 58 experienced a myopic opponent and 60 experienced a predictive opponent. The mean achievement score was computed for each test block. Notice that in the game in Figure 3.4 the best response for player I, if the opponent is myopic, is to move. However, if the opponent is predictive, the best response for I is to stay.

Because opponents’ types are fixed and participants experience 40 games, they have the opportunity to learn how their opponent might be playing the games. Consequently, participants may gradually make more best response choices over time. Goodie et al. postulated that participants were deemed to have learned the opponents model at the game after which
performance was always statistically significantly better than chance, as measured by a binomial test at the 0.05 level and one-tailed. This implies making no more than one inaccurate choice in any following window of ten games.

Among all participants, 36 acted consistently with first-level reasoning, while 82 were consistent with the second level in the first game. In Figure 3.5, it shows the mean achievement scores across all participants for the two groups, mean prediction scores, and the rationality error rates. Two group-level findings are evident from the results in Figure 3.5(a): First, the mean achievement score is significantly higher when the opponent is predictive as compared to when she is myopic ($F(1, 116) = 84.37, p < 0.001$).

![Figure 3.5](image-url)

Figure 3.5: Fixed-sum games in Study 1: (a) Mean achievement score of the participants for both groups across test blocks. (b) Mean prediction score of the participants for both groups across test blocks. Subjects generally expected their opponents to play at first level rather than at zero level. (c) Rationality error rates among participants.

The higher achievement score when the opponent is predictive and the low score when
the opponent is myopic implies that subjects predominantly displayed second-level reasoning when acting. Mean prediction scores from Figure 3.5(b) reveal that the participants generally expected their opponent to reason about their subsequent play (first-level reasoning). The fact that myopic opponents did not do this resulted in participants’ choices not being the best responses. The mean prediction score for the myopic group decreases while the achievement score improves across test blocks.

Second, notice from Figure 3.5(b) that the mean prediction score changes over successive test blocks in both groups. Multivariate t-tests of both the main effect of block position and the interaction between block position and opponent type reveal that the changes in both groups were significant ($p < 0.001$). Thus, learning took place that was responsive to the opponent, although the learning was slow and not all participants learned the opponent type. Participants learned to play best-response significantly faster against a predictive opponent compared to a myopic one.

### 3.1.2 Third-Level Reasoning

In order to test level 3 recursive reasoning, Goodie et al. extended the Modified Centipede game with fixed-sum from Study 1 to four stages with five states. An example game is shown in Figure 3.6.

1) **Methodology:**

Goodie et al. designed the computer opponent (player II), projected as a human player, to play a game in three ways if player I chooses to move and the game proceeds to state B: (i) Player II decides on her action by simply choosing rationally between the default outcomes of staying at states B and C in Figure 3.6(b); II is a zero-level player and is called *myopic*. (ii) II reasons that player I will choose rationally between the default outcomes of stay at C and stay at D, and based on this action, II selects an action that maximizes its outcomes; II is a first-level player who explicitly reasons about I’s subsequent choice and is called *predictive*. 

36
(iii) II reasons that player I is predictive and who will act rationally at C reasoning about II’s rational action at D; II is a second-level player who explicitly reasons about I’s subsequent choice which is decided by rationally thinking about II’s subsequent action at D. This type of II is called *super-predictive*. Therefore, if player I thinks that II is super-predictive, then I is reasoning deeply to three levels.

To illustrate, in the game in Figure 3.6(b), if player I chooses to move, she thinks that a myopic player II will stay to obtain a payoff of 0.6 at state B compared to move which will obtain 0.3 at state C. She thinks that a predictive II thinking that I being myopic will move to D thereby obtaining 0.8 instead of staying at C, will decide to move thinking that she can later move from D to E, which gives II an outcome of 0.8 at E. A super-predictive II knows
that I is predictive, knowing that if I moves from C to D then II will move to E which gives I only 0.2, hence I will choose to stay at C, therefore II will stay at B.

In contrast to three-stage sequential games, no single preference ordering in four-stage games can be used to distinguish between the actions of the three opponent reasoning types. Therefore, Goodie et al. employed two differently ordered games to diagnose each level of reasoning from the other two.

The game depicted in Figure 3.6(b), which has the preference ordering of \( E \prec B \prec A \prec CD \) for player I, is the only ordering that permits I’s second level reasoning to be distinguished behaviorally from her third and first levels of reasoning. As analyzed above, player I with the second level reasoning will choose to move while with the first and third levels, she will choose to stay. Preference orderings \( C \prec B \prec A \prec D \prec E \) and \( C \prec B \prec A \prec E \prec D \) are the two orderings which distinguish the first level reasoning from the second and third levels. In these games, player I with the first level will choose to move while with the second and third levels to stay.

2) Results:

Participants were assigned randomly to different groups that played against a myopic, predictive or super-predictive opponent. Each participant experienced 30 trials with each trial consisting of two games whose combined payoff orderings are diagnostic and a catch trial controlling for inattention. The 30 trials were grouped into 6 blocks of 5 trials each.

From the participants’ data in each of the three opponent groups, the measured the achievement score, the prediction score, and the rationality error are shown in Figure 3.7. A metric, \( L \), is defined as the trial after which performance over the most recent 10 trials never failed to achieve statistical significance (cumulative binomial probability < .05). This implies making no more than one incorrect choice in any window of 10 trials. For participants who never permanently achieved statistical significance, \( L \) was assigned a value of 30. Participants
in super-predictive opponent group had the highest overall achievement score with an average L score of 11.3, followed by those in myopic opponent group with L score, 27.2. These L scores are consistent with the observations that achievement scores in these two groups are increasing. Participants in predictive opponent group had achievement score close to zero and L score of 30, which means they never truly achieved a corresponding strategy. Similar to the mean achievement score, participants in the super-predictive opponent group exhibited the highest prediction scores while those in the predictive opponent group had the lowest scores (Figure 3.7(b)).

Figure 3.7: Study 2: (a) mean achievement score and (b) mean prediction score of the participants for three groups across test blocks. (c) rationality error
3.2 Empirically Informed Interactive POMDP

Data produced by studies reviewed in Section 3.1 are related to human recursive thinking. We seek process-oriented and principled computational models with empirical support whose predictions are consistent with the observed data. These models differ from statistical curve fitting, such as regression analysis and kernel-based density estimation on the data [70]. They may provide some insights into the judgement and decision-making processes that potentially led to the observed data. Consequently, these models tend to be more generally applicable than the former.

In order to computationally model the data, we look for a multiagent decision-making framework which is capable of modeling recursive reasoning in the decision-making process. Finitely-nested I-POMDPs [41] are a natural choice because of their explicit consideration of recursive beliefs and decision making based on such beliefs.

Although the Modified Centipede game is a sequential game which involves multiple decision points, decisions should be modeled recursively rather than sequentially. This is because decision points alternate and occur at distinct and fixed states for different players, and a player may not know her payoff until the other makes her move. We model the individual games from Study 1 and 2 using $\text{I-POMDP}_{i,2}$ and $\text{I-POMDP}_{i,3}$ respectively. The physical state space is perfectly observable (4 states in Study 1 and 5 in Study 2); i’s actions, $A_i = \{\text{Stay, Move}\}$ are deterministic and j has similar actions; i observes others actions, $\Omega_i = \{\text{Stay, Move}\}$; $O_i$ is not needed; and $R_i$ captures the diagnostic preferential ordering of the states depending on which game is being considered.

Because the opponent is conceived as human and guided by payoffs, we focus on intentional models only. Intuitively, the model set in Study 1, $\Theta_j = \{\theta_{j,1}, \theta_{j,0}\}$, where $\theta_{j,1}$ is the level 1 (predictive) model of the opponent and $\theta_{j,0}$ is the level 0 (myopic) model. In Study 2, the model set is extended to include $\theta_{j,2}$, the level 2 (super-predictive) model. Predic-
tions about the opponent’s action by the participants were consistent with opponent models being attributed. Parameters of these models are analogous to the I-POMDP for agent i, except for $R_j$ which reflects the preferential ordering of the states for the opponent. Note that the predictive model, $\theta_{j,1}$, includes the level 0 model of i, $\theta_{i,0}$, in her interactive state space. Agent i’s belief, $b_{i,2}$ in Study 1 and $b_{i,3}$ in Study 2, assigns a varied distribution to j’s possible models based on the game. This belief will reflect the general de facto thinking of the subjects about their opponent. It also assigns a marginal probability 1 to state A indicating that i decides at that state. In Study 1, both $b_{j,1}$ and $b_{j,0}$ that are part of j’s two models, respectively, assign a marginal probability 1 to B indicating that j acts at B. Belief $b_{i,0}$, that is part of $\theta_{i,0}$, assigns probability of 1 to C. Similarly marginal probabilities are assigned for beliefs in Study 2.

### 3.2.1 Learning and Decision Models

I-POMDP as a rational framework may not be directly applied to model data as humans are not always acting rationally. Thus, we augment I-POMDPs with elements that have support in behavioral and cognitive literature to model human decision making. Selection of these models and their parameters is informed by the data in Study 1 and 2: in Figures 3.3(a), 3.5(a), and 3.7(a), some subjects learn about the opponent model as they continue to play. However, the rate of learning varies across subjects, and, in general, the learning is slow and partial. This is indicative of the well-known cognitive phenomenon that subjects could be underweighting the observed evidence. In other words, observing that player II stayed in the game in Figure 3.1(c) may not fully convince the participant that II is predictive. We model this by making the observations slightly noisy and augmenting normative Bayesian learning
in the following way,

\[ b'_{i,l}(s, \theta_{j,l-1}|o_i; \gamma) = \alpha b_{i,l}(s, \theta_{j,l-1}) \times \left\{ \sum_{a_j} O_i(a_i|a_i, a_j, s') \times \Pr(a_j|\theta_{j,l-1}) \right\}^\gamma, \] (3.1)

where \( \alpha \) is the normalization factor, state \( s \) corresponds to A and \( s' \) to B, action \( a_i \) is to move, \( \Pr(a_j|\theta_{j,l-1}) \) is the probability of \( j \)'s action given the model \( \theta_{j,l-1} \), if \( \gamma < 1 \), then the evidence \( o_i \in \Omega_i \) is underweighted while updating the belief over \( j \)'s models.

In Figures 3.3(c), 3.5(c), and 3.7(c), there exist significant rationality errors. We utilize the quantal response model [60] to simulate human non-normative choice. The probability of choosing an action is represented as a sigmoid function of how close to optimal is the action. Mathematically,

\[ q(a^*_i \in A_i; \lambda) = \frac{e^{\lambda U(b_i, a^*_i)}}{\sum_{a_i} e^{\lambda U(b_i, a_i)}}, \] (3.2)

where \( q(a^*_i \in A_i; \lambda) \) is the probability assigned to action \( a^*_i \), given \( \lambda \), by the model; \( U(b_i, a_i) \) is the utility for \( i \) as computed by the I-POMDP performing the action, \( a_i \), given her belief, \( b_i \); and \( \lambda \) is the parameter that controls how responsive is the model to value differences. Within the I-POMDP, we may replace utility maximization with this model in a straightforward way.

In addition to the above models, we seek to ascertain the prior beliefs of agent \( i \) for different games. We utilize a distribution of heterogenous initial beliefs. This is informed by the predictions of the subjects about the opponent’s action for the first three games. For example, approximately just 2\% of the subjects experiencing the general-sum game in Study 1 thought that the opponent is predictive in the first three games consistently, while about 19\% of the participants thought the same while experiencing the fixed-sum game in Study 1. The corresponding beliefs were set to be highly informative.
3.2.2 Analysis on Social Factors

Research suggests that some social factors could play a role in human decision making in strategic games. We reviewed some of the work in Section 2.4. In those games, it is possible to end at a “win-win” situation thereby motivating the exploration of these factors. However, this is not always the case in the games that we model. The setting of fixed-sum with no increase in total payoffs makes the game competitive and precludes these models. In the general-sum game, players may reciprocate in the presence of a state where payoffs for both players are more desirable. Among the 40 games, only five catch games have outcome orderings that can be used to distinguish between a participant exhibiting behavior that possibly could be explained using reciprocity, or due to the models in $\theta_j$, by choosing different actions at state A. In all other games, behavioral outcomes from attributing different factors to the opponent would be equivalent to behavior predicted by one of the models in $\theta_j$. One such catch game has outcome probabilities of 0.6, 0.4, 0.8, and 0.2 for player I and 0.2, 0.6, 0.4, and 0.8 for II (preferences over states, I: $D \prec B \prec A \prec C$, II: $A \prec C \prec B \prec D$). Player I who thinks her opponent is either myopic or predictive would stay at A. If she expects reciprocity from the opponent, she would move from A given that her opponent would reciprocate by moving to C, which has better outcomes for both players compared to A. In these five game types, resulting in a total of 570 games played by all participants, the action move in conjunction with a prediction of move for the opponent is taken at a low rate of 8.98%. This indicates that the possible effect of reciprocity is not observed significantly in the games that we model.

In the context of inequity aversion, Goodie et al. report that under the myopic opponent condition, the percentage of time that the action move is chosen at A in games where outcomes at A are unequal between the two players is 53.4%, less than the move percentage of 57.6% across all games considered. Furthermore, we find that in the general-sum study, under myopic opponent group, there are 4 diagnostic games whose outcomes at state C are
closer for both players compared to those in state D. One such games outcomes in each state for player I are: 0.4, 0.2, 0.6, and 0.8, while for player II are: 0.4, 0.6, 0.8, and 0.2. In these games, staying at C is consistent with inequity aversion. However, participants stayed only 6.29% of the times at C. Therefore, there is insignificant evidence of the effect of inequity aversion in the data.

Other factors may also affect how the games are played. For example, subjects may simply choose to move because staying terminates the game right away, which is less “fun”. However, in the fixed-sum context, more than 90% of the participants choose to stay at A for the predictive opponent group (Figure 3.5(a)) and terminate the game immediately. This indicates that extending the game for “fun” was not a predominant factor in how participants played the game. Finally, Goodie et al. report another experiment on avoiding uncertainty brought about by moving and default staying and conclude that these factors do not motivate choice either.

3.3 Performance Analysis

Augmenting I-POMDP with models mentioned above requires us to find values for the parameters $\gamma$ (learning rate) and $\lambda$ (non-normative choice). In the experiments, the participants’ predictions used for computing the prediction score and their actions used for computing the achievement score, were collected. They provide two data sets from which parameters could be learned. We formulate the problem of finding these parameters as that of maximizing data likelihood. Optimization techniques such as the simplex method [65] may be used.

3.3.1 Model Performance in General- and Fixed-sum Games

Considering the existence of two data sets: predictions and actions, models may be supplied with different data.
1). Learning $\gamma$ Using Predictions and $\lambda$ Using Decisions

Participants’ predictions of the opponent’s action indicates their belief about the opponent’s type. Therefore, prediction data is particularly appropriate for learning parameter $\gamma$. We seek the value of $\gamma$ that maximizes the likelihood of the observed predictions as predicted by the belief updated using Equation 3.3. Formally, we want to maximize:

$$L = \prod_{i=1}^{\text{|Subj|}} \prod_{g=1}^{N} \prod_{l=0}^{l_j} \mathcal{I}(e_i^t, \theta_{j,l}) \ b_{i,2}^g(s, \theta_{j,l}; \omega_i; \gamma),$$

where $|\text{Subj}|$ and $N$ are the total number of subjects and games, respectively; $e_i$ is the prediction of the subject, $i$, in the $g^{th}$ game translated into the corresponding opponent type; $\mathcal{I}(e_i, \theta_{j,l})$ is an indicator function which is 1 if its arguments are equal, and 0 otherwise; and $b_{i,2}^g(\theta_{j,l}; \omega_i)$ is the probability, updated according to Equation 3.1, assigned by $i$ to $j$’s model, $\theta_{j,l}$, on observing $\omega_i$. Initial belief is set according to subject’s predictions of opponent types on the first three games.

In order to learn the rate of non-normative choice, $\lambda$, we seek the value that maximizes likelihood of the observed actions whose probabilities are computed by the augmented I-POMDP$_{i,2}$:

$$L = \prod_{i=1}^{\text{|Subj|}} \prod_{g=1}^{N} q(a_i^*|A_i)$$

$$= \prod_{i=1}^{\text{|Subj|}} \prod_{g=1}^{N} e^{\lambda U(b_{i,2}^g; a_i^*)} \sum_{a_i \in A_i} e^{\lambda U(b_{i,2}^g; a_i)} \quad \text{(from Equation 3.2)},$$

where $a_i^*$ is the action from $A_i$ selected by the subject $i$ in the $g^{th}$ game. Subject’s belief, $b_{i,2}^g$, is updated according to Equation 3.1 after each game instance, $g$. Notice that the ideal choice model, $q$, assigns a probability 1 to each of the actions played by each subject resulting in the maximum value of $X$ ($=0$).
The computation of the likelihood may be simplified by taking its log.

**Model Performance:**

We integrated models simulating human belief update (Equation 3.1) and choice (Equation 3.2) within the finitely-nested I-POMDP model. We utilize the converged values of the parameters within the augmented I-POMDP. We report on the performance of my initial model below.

1) *Parameters:* We randomly separated behavioral data from each experiment into training and test sets of approximately equal size. We used training set to learn the value of parameters, $\gamma$ and $\lambda$.

**General-sum game:** 57 of the 114 subjects participating in general-sum games were randomly selected, and their behavioral data used to learn the value of parameters. Of these subjects, 29 had experienced a myopic opponent while 28 had faced a predictive one. Because subjects facing myopic or predictive opponents displayed learning and acted significantly differently with different levels of rationality errors, we learn parameters separately on two groups and report parameter values in Table 3.1. Notice that some parameter values differ substantially between groups. The lower $\gamma$ for the myopic group results in weakly updated belief about the myopic model after an observation. This reflects the smaller rate of decrease in prediction scores observed across the test blocks in Figure 3.3, compared to the rate of increase in prediction score for the predictive group.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>General sum</th>
<th>Fixed sum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>myopic</td>
<td>predictive</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.0067</td>
<td>0.0546</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.82</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Table 3.1: Parameter values obtained by learning $\gamma$ from predictions and $\lambda$ from decisions in Study 1.

**Fixed-sum game:** Of the 118 subjects who participated in the critical games, 59 were selected
randomly and their data used to learn the value of parameters. We learn separately on the two groups of subjects facing different opponents (29 faced myopic while 30 faced a predictive opponent). The learned values of parameters are given in Table 3.1. Note that the relatively high value of $\lambda$ for the predictive group, which reflects the low rationality errors for subjects in that group. Additionally, the lower $\gamma$ for the myopic group is indicative of the poorer learning among subjects faced with a myopic opponent, as is evident from Figure 3.5.

2) Results: We utilized the learned values in Table 3.1 to parameterize the underweighting and the quantal response choice models within the $I$-POMDP$_{1,2}$. We cross-validated our model on the test set containing data of randomly picked subjects to evaluate its accuracy. Using the subjects’ predictions of the opponent type in the first three games, we set the distribution of prior beliefs. For example, consistent predictions of the opponent type (possibly incorrect) resulted in assigning a highly informative prior; otherwise, inconsistent predictions led to uninformative priors. We sampled the updated belief to obtain the model’s prediction of the opponent type and sampled the quantal distribution to obtain the model’s predicted action. We did this for the total number of games experienced by a subject and for the number of subjects in the test data set experiencing the two opponent types, in each study. General-sum game: We obtained model predictions that correspond to the 57 subjects in test data, of which 29 faced a myopic opponent and 28 faced a predictive opponent. First, we visually compare the mean achievement and prediction scores of the model predictions with those of the study data. As seen in Figure 3.8, model-based achievement and prediction scores align with the study data in general, but with some significant difference between the two on the final two test blocks. Each model data point is the average of ten runs.

We measure the goodness of the fit by computing mean squared error (MSE) of the predictions by the model ($I$-POMDP$_{1,2}$), and those of a random model (null hypothesis) for comparison. We show the MSE for both, the achievement and prediction scores, in Table 3.2. While the random model shows a low error in the achievement score for the
general-sum game, all differences in MSE between our model and the random model for both groups excepting the difference in achievement score for the predictive group between the two models, are significant ($p \leq 0.05$ on a paired Student’s t-test). In particular, level 1 predictions by our model when the opponent is myopic are consistent with the data thereby displaying a very small MSE; this was the prominent level of recursive reasoning in the general-sum game.

Figure 3.8: Study 1, General-Sum game, first model: Comparison of initial model predictions with actual data for both groups. Comparisons of the mean achievement score (top), and mean prediction score (bottom) are shown. Notice the differences in the first test block where subjects started off by making an unusually high or low number of rationality errors.
Fixed-sum game: Predictions from the model corresponded to the 40 game instances played by 29 subjects that faced a myopic opponent and by 30 that faced a predictive opponent, for the fixed-sum game. We compute the mean achievement and prediction scores of the model predictions and visually compare it with that of the actual data, in Figure 3.9. Notice the differences in the achievement scores for both the groups, which is, in part, due to the fact that subjects made more rationality errors initially compared to other test blocks and, in part, due to the low $\gamma$ that was learned. In Table 3.2, we show the MSE of the predictions by our model, and compare it with the MSE of the predictions by a random model (null hypothesis). Differences in the MSE between our model and the random model are statistically significant ($p < 0.05$ using a paired t-test) except for the difference in achievement score for the myopic group where $p < 0.1$.

<table>
<thead>
<tr>
<th>Game type</th>
<th>Opponent type</th>
<th>Mean Squared Error (MSE)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Achievement score</td>
<td>Prediction score</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Random</td>
<td>I-POMDP$_{1,2}$</td>
<td>Random</td>
</tr>
<tr>
<td>General sum</td>
<td>myopic</td>
<td>0.01936</td>
<td>0.00429</td>
<td>0.08588</td>
</tr>
<tr>
<td></td>
<td>predictive</td>
<td>0.00624</td>
<td>0.00621</td>
<td>0.02872</td>
</tr>
<tr>
<td>Fixed sum</td>
<td>myopic</td>
<td>0.02887</td>
<td>0.01231</td>
<td>0.09044</td>
</tr>
<tr>
<td></td>
<td>predictive</td>
<td>0.21269</td>
<td>0.05891</td>
<td>0.09033</td>
</tr>
</tbody>
</table>

Table 3.2: Study 1, first model: Goodness of the fit measured using the mean squared residual error (MSE). We include the random model for assessing significance. All MSEs have been rounded to five decimal places.

2). Learning Both $\gamma$ and $\lambda$ Using Decisions

Some of the first model predictions are not consistent with the study data. In particular, Figure 3.9 shows a large error between the predictions by the empirically-informed I-POMDP and the observed data in the context of predictive opponents. Surprisingly, one may discover that participants’ predictions of others’ actions did not accurately guide their final decisions.
Figure 3.9: Study 1, Fixed-Sum game, first model: Comparison of initial model predictions with actual data for both groups. Mean achievement (top), and mean prediction scores (bottom) are shown.

on how to act. Importantly, a significantly high percentage of subjects whose predictions were incorrect given their predictive opponents went on to perform the accurate choice. This is partly evident from the difference in the mean prediction and achievement scores for the predictive group in Figure 3.9. We think that this discrepancy is due to inattention by the participants to the screen collecting their predictions, and an increased focus on acting correctly that would result in their obtaining the monetary incentive.
Consequently, the possibly unreliable predictions of the subjects may not correctly reflect their actual learning rates (parameter $\gamma$). In order to improve the accuracy of the model predictions, we begin by learning $\gamma$ from the participants’ action data.

We utilize the likelihood as defined by Equation 3.4 in order to learn $\gamma$ from the subjects’ choices. The new parameter values learned from the training set are shown in Table 3.3. Overall, the parameter $\gamma$ is larger with a substantial increase for the predictive context of the fixed-sum game. These revised parameter values lead to revisions in $\lambda$ as well because $\gamma$ is utilized in the learning procedure of computing $\lambda$.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>General sum</th>
<th></th>
<th>Fixed sum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>myopic</td>
<td>predictive</td>
<td>myopic</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.075</td>
<td>0.035</td>
<td>0.04</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.74</td>
<td>0.69</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table 3.3: Study 1, second model: Parameter values obtained on the subjects’ decision data. Observe the difference between these and the parameter values in Table 3.1, especially of $\gamma$.

**General-sum game:** We show the improved achievement score of our model predictions for the general-sum game setting in Figure 3.10(a). In comparison to the previous performance (Figure 3.8(a)) the significantly higher $\gamma$ in the context of myopic opponents leads to model predictions on the test set that are more consistent with the observed data. This is reflected in the lower MSE (Table 3.4) of the model-based achievement score, which is roughly half of the previous MSE, and its larger difference from the random model is statistically significant ($p < 0.05$ on a paired Students t-test).

**Fixed-sum game:** The largest improvement in the model performance is in its second-level predictions in the fixed-sum game setting in the context of predictive opponents. Figure 3.10(b) shows the close fit of the model to the study data for the predictive group clearly. This is reflected in the residual MSE of the models predictions for the predictive group (Table 3.4), which is less by more than an order of magnitude from the previous MSE.
Figure 3.10: Study 1, second model: Comparison of model predictions with actual data for both groups in the general-sum game (top), and fixed-sum game (bottom).

The corresponding difference from the MSE of the random model is statistically significant ($p < 0.001$). On the other hand, the MSE of the prediction score is slightly higher indicating a somewhat poorer fit of the predictions data, which is to be expected.
<table>
<thead>
<tr>
<th>Game type</th>
<th>Opponent type</th>
<th>Mean Squared Error (MSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Achievement score</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Random</td>
</tr>
<tr>
<td>General</td>
<td>myopic</td>
<td>0.01936</td>
</tr>
<tr>
<td>sum</td>
<td>predictive</td>
<td>0.00624</td>
</tr>
<tr>
<td>Fixed</td>
<td>myopic</td>
<td>0.02887</td>
</tr>
<tr>
<td>sum</td>
<td>predictive</td>
<td>0.21269</td>
</tr>
</tbody>
</table>

Table 3.4: Study 1, second model: Goodness of the fit of the model with the study data compared with the random model for assessing significance. All MSEs have been rounded to five decimal places.

3). Dynamic $\lambda$

In addition to the discrepancy between the subjects’ predictions of opponents’ actions and their subsequent choice, it is observed that subjects’ rationality error rates decreased substantially as they experienced more games. Such decline has been observed previously in a wide range of experimental data in economics and cognitive science; for example, McKelvey and Palfrey report it [60]. This could be because subjects started paying more attention as the study progressed. This is clearly evident among the predictive group participants in Figure 3.3(c) and among the subjects in both groups who experienced the fixed-sum game (Figure 3.5(c)). By learning a fixed $\lambda$ across all games, our previous model did not capture this dynamic nature of the rationality error rate.

We model a linear change in the parameter, $\lambda$:

$$\lambda_k = \lambda_1 + (k - 1) \times c,$$  \hspace{1cm} (3.5)

where $\lambda_1$ is the parameter value for the first test block, $k$ indicates the current test block, $1 \leq k \leq 4$, and $c$ is the parameter by which $\lambda$ increases across test blocks. Recall that
an increase in the value of $\lambda$ likely leads to reduced non-normative choice from the quantal response model.

We learn the initial $\lambda_1$ by minimizing the error surface, $X$, represented by Equation 3.4 with the change that $\lambda$ is substituted by $\lambda_1$ and variable $g$ iterates over games in the first test block only. The learned value of $\lambda_1$ using training set is shown in Table 3.5.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>General sum</th>
<th>Fixed sum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>myopic</td>
<td>predictive</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.072</td>
<td>0.039</td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>0.56</td>
<td>0.44</td>
</tr>
<tr>
<td>$c$</td>
<td>0.15</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Table 3.5: Study 1, third model: Parameter values obtained on the subjects' decision data from the two studies.

<table>
<thead>
<tr>
<th>Game type</th>
<th>Opponent type</th>
<th>Mean Squared Error (MSE)</th>
<th>Achievement score</th>
<th>Rationality error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Random</td>
<td>I-POMDP$_{1,2}$</td>
</tr>
<tr>
<td>General</td>
<td>myopic</td>
<td>0.01936</td>
<td>0.00034</td>
<td></td>
</tr>
<tr>
<td></td>
<td>predictive</td>
<td>0.00624</td>
<td>0.00290</td>
<td></td>
</tr>
<tr>
<td>Fixed</td>
<td>myopic</td>
<td>0.02887</td>
<td>0.01079</td>
<td></td>
</tr>
<tr>
<td></td>
<td>predictive</td>
<td>0.21269</td>
<td>0.00218</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.6: Study 1, third model: Goodness of the fit of our model with the study data. MSE for the rationality errors is shown as well. MSE for the prediction score remains approximately the same as before.

We compare the predictions of the revised model with observed data from the two studies and show it visually in Figure 3.11. This model replaces the fixed $\lambda$ in the model of the previous subsection with a dynamic one. Notice from Figure 3.11(a) that the revised model fits the observed data for the myopic group of the general-sum game significantly better than a fixed $\lambda$. The MSE in Table 3.6 reveals error that is less by almost an order of
magnitude, and this difference is significant ($p < 0.001$). On the other hand, the revised model fit for the predictive group in the fixed-sum game is slightly worse than in the previous case. Nevertheless, the differences in the two model errors are not statistically significant ($p = 0.093$). Furthermore, MSEs for all other groups are less than the corresponding errors when a fixed $\lambda$ is utilized. In comparison with the random model, all fits continue to remain significantly better, except for the predictive group in the fixed-sum game whose
improvement narrowly missed statistical significance.

For the sake of completeness, we also show the MSE for the rationality error rate computed from the predictions of the revised model for all groups in both game contexts, in Table 3.6. Because λ increases across test blocks, the error rates reduce. The MSE continues to remain low and is significantly less than that of the random model.

### 3.3.2 Model Performance in Third-Level Reasoning

In Study 2, the number of opponent groups increases into 3, with a new model θ_{j,2} representing the level 2 (superpredictive) model of the opponent added. This makes the model set of player II increase, \( \Theta_j = \{\theta_{j,2}, \theta_{j,1}, \theta_{j,0}\} \). As the model proposed in Study 1 continue to apply to the Study 2 data, we extend the I-POMDP model to longer Centipede games and label it as I-POMDP\(^{\gamma, \lambda}_{i,3} \). We also present two other models.

1). With Ascribing Choice Model to Opponent, I-POMPD\(^{\gamma, \lambda_1, \lambda_2}_{i,3} \)

The methodology for the experiments reveals that the participants are deceived into thinking that the opponent is human. Therefore, participants may justify unexpected actions of the opponent as errors in their decision making rather than due to their level of reasoning. Hence, we generalize the previous model by attributing quantal response choice to opponent’s action selection as well. Let \( \lambda_1 \) be the quantal response parameter for the participant and \( \lambda_2 \) be the parameter for the opponent’s action. Then,

\[
Q(a_i^*; \gamma, \lambda_1, \lambda_2) = \frac{e^{\lambda_1 \cdot U(b'_{i,3}, a_i^*; \gamma, \lambda_2)}}{\sum_{a_i \in A_i} e^{\lambda_1 \cdot U(b'_{i,3}, a_i; \gamma, \lambda_2)}}. \tag{3.6}
\]

parameters, \( \lambda_1, \lambda_2 \in [-\infty, \infty] \); \( a_i^* \) is the participant’s action and \( Q(a_i^*) \) is the probability assigned by the model. \( U(b'_{i,3}, a_i; \gamma, \lambda_2) \) is the utility for \( i \) on performing action, \( a_i \), given its updated belief, \( b'_{i,3} \), with \( \lambda_2 \) parameterizing \( j \)’s action probabilities, \( \Pr(a_j | \theta_{j,1-1}) \), present in
Equation 3.1 and in computation of the utility. This model is labelled as IPOMDP\(^{\gamma, \lambda_1, \lambda_2}\).

In IPOMDP\(^{\gamma, \lambda_1, \lambda_2}\), three parameters are involved: \(\gamma\) representing participants’ learning rate, \(\lambda_1\) and \(\lambda_2\) representing non-normative actions of the participant and her opponent, respectively. The empirically informed IPOMDP model gives a likelihood of the experiment data given specific values of the three parameters.

We begin by learning \(\lambda_2\) first. Because this parameter characterizes expected opponent behavior, we utilize the participants’ expectations of their opponent’s action in each game, denoted as \(a_{ij}\), as the data. Denoting this set of expectations as \(P\), the likelihood of \(P\) is obtained by taking the product of \(Q(a_{ij}^*; \lambda_2)\) over \(G\) games and \(N\) participants as the probability is hypothesized to be conditionally independent between games given the model and is independent between participants.

\[
L(P; \lambda_2) = \prod_{i=1}^{N} \prod_{g=1}^{G} Q(a_{ij}^*; \lambda_2)
\]

Here, \(a_{ij}^*\) is the observed expectation by participant \(i\) of \(j\)’s action in game \(g\), and \(Q(a_{ij}^*; \lambda_2)\) is the probability assigned by the model to the action, whose computation is analogous to Equation 3.6 except that \(j\)’s lower-level belief replaces \(i\)’s belief and \(j\) does not ascribe non-normative choice to its opponent.

In order to learn the values of parameters, \(\gamma\) and \(\lambda_1\) (or \(\gamma\) and \(\lambda\) in IPOMDP\(^{\gamma, \lambda}\)), we utilize the participants’ actions at state A. Data consisting of these actions is denoted as \(D\). The likelihood of this data is given by the probability of the observed actions of participant \(i\) as assigned by our model over all games and participants.

\[
L(D; \gamma, \lambda_1, \lambda_2) = \prod_{i=1}^{N} \prod_{g=1}^{G} Q(a_{ij}^*; \gamma, \lambda_1, \lambda_2)
\]

\[
= \prod_{i=1}^{N} \prod_{g=1}^{G} \frac{e^{\lambda_1 U(a_{ij}^*; \gamma, \lambda_2)}}{\sum_{a_i \in A_i} e^{\lambda_1 U(a_{ij}^*; \gamma, \lambda_2)}} \quad \text{(from Eq. 3.6)}
\]
The computation of the likelihood may be simplified by taking its log.

2). Weighted Fictitious Play

A different reason for participant behavior that relies more heavily on past patterns of observed actions of the opponent, instead of ascertaining the mental models of the opponent as in the previous I-POMDP based models, is applicable. A well-known learning model in this regard is weighted fictitious play [19]. To apply this model, we first transform the game of Figure 3.6 into its equivalent normal form, which we show below in Table 3.7.

<table>
<thead>
<tr>
<th>$A_i$ (player I)</th>
<th>$A_j$ (player II)</th>
<th>stay</th>
<th>stay</th>
<th>move</th>
</tr>
</thead>
<tbody>
<tr>
<td>stay A</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>stay C</td>
<td>0.4</td>
<td>0.7</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>move</td>
<td>0.4</td>
<td>0.8</td>
<td>0.2</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.7: A normal form representation of the game in Fig. 3.6. The payoffs are the probabilities of winning for player $I$. Action, stay $s$, denotes that the first stay occurs at state $s$ for the corresponding player.

Let $E_i(a_j)$ be the observed frequency of the opponent’s action, $a_j \in A_j$. We update this as:

$$E_i^t(a_j; \phi) = I(a_j, o_i) + \phi E_i^{t-1}(a_j) \quad t = 1, 2, \ldots$$ (3.7)

where parameter, $\phi \in [0, 1]$, is the weight put on the past observations; $I(a_j, o_i)$ is an indicator function that is 1 when $j$’s action in consideration is identical to the currently observed $j$’s action, $o_i$, and 0 otherwise. $E_i^0(\cdot; \phi)$ may be initialized to 1 for all actions. The weighted frequency, when normalized, is deemed to be representative of agent $i$’s belief over what $j$ will do in the next game in the trials.

Due to the presence of rationality errors in the data, we combine the belief update of
Eq. 3.7 with quantal response:

\[ Q(a_i^*; \phi, \lambda) = \frac{\lambda \sum_{a_j} E_i(a_j; \phi) R_i(a_i^*, a_j)}{\sum_{a_i \in A_i} e^{\lambda \sum_{a_j} E_i(a_j; \phi) R_i(a_i, a_j)}}, \]  

(3.8)

Here, \( \bar{E}_i \) is the normalized frequency (belief) as obtained from Eq. 3.7 and \( \lambda \in [-\infty, +\infty] \).

This model is labelled as wFPi\( ^{\phi, \lambda} \).

Parameters for wFPi\( ^{\phi, \lambda} \) are learned by maximizing the log likelihood of the data, \( D \), in which the quantal response function is as shown in Equation 3.8. In this regard, note that the experiment data includes actions performed by the participants at states A and C, and programmed opponent actions at states B and D, if the game progressed to those states.

We evaluate the comparative fitness of the different generative models to the data and visually compare the experiment data with model simulations.

**Model Performance:**

1) **Parameters:** In order to learn \( \lambda_2 \), we use the expectations data of the catch games only. This is because no matter the type of the opponent, the rational action for the opponent in catch games is to move. Hence, expectations of stay by the participants in the catch trials would signal a non-normative action selection for the opponent. This also permits learning a single \( \lambda_2 \) value across the three groups. However, this is not the case for the other parameters. In Figure 3.7, observe that for different opponents, the learning rate, L, is different. Also, in Figure 3.7(c), the rationality errors differ considerably between the opponent groups. Therefore, we learn parameters \( \gamma \) and \( \lambda_1 \) given the value of \( \lambda_2 \) (and \( \lambda \) in I-POMDP\( \gamma, \lambda \), \( i, 3 \)), separately from each group’s diagnostic games. Analogously, we learn \( \phi \) and \( \lambda \) for wFPi\( ^{\phi, \lambda} \) from the diagnostic games as well. We report the learned parameters averaged over raining folds in Table 3.8.

In Table 3.8, see that \( \gamma \) for the predictive opponent group is close to zero. This is
I-POMDP $\gamma, \lambda$

<table>
<thead>
<tr>
<th>model</th>
<th>param.</th>
<th>myopic</th>
<th>pred</th>
<th>super-pred</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>0.232</td>
<td>0.079</td>
<td>0.357</td>
<td></td>
</tr>
<tr>
<td>$\lambda$</td>
<td>3.259</td>
<td>3.826</td>
<td>3.667</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.8: Average parameter values learned from the training folds of the experiment data for the three candidate models.

consistent with the observation that participants in this group did not make much progress in learning the opponent type. Consequently, we focus our analysis on the myopic and super-predictive opponent groups here onwards. Furthermore, note the dichotomy in the value of $\phi$ between the myopic and super-predictive opponent groups. A value of close to zero for the super-predictive group indicates that the previously observed action is mostly sufficient to predict the opponent’s action in the next game. However, $\phi$’s value close to 1 is indicative of the past pattern of observed actions not helping much in modeling the behavior of the other groups.

2) Results: We show the log likelihoods of the different models, including a random one that predicts other’s actions randomly and chooses its own actions randomly, in Table 3.9. The random model serves as our null hypothesis. note that $\text{I-POMDP}_{i,3}^{\gamma, \lambda_1, \lambda_2}$ has the highest likelihood in the myopic context, although the likelihood of the other $\text{I-POMDP}$ based model is slightly better for the super-predictive group. On the other hand, $\text{wFP}_{i}^{\phi, \lambda}$ exhibits a vast difference in the log likelihoods between groups, with the low likelihood for the myopic group, though still better than that for the random model, indicating a poor fit. As seen next, this
is due to the potentially poor descriptive prediction of the opponent’s actions in the games by relying solely on observed empirical play.

<table>
<thead>
<tr>
<th>model</th>
<th>log likelihood</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>myopic</td>
<td>super-predictive</td>
</tr>
<tr>
<td>Random</td>
<td>-1455.605</td>
<td>-1414.017</td>
</tr>
<tr>
<td>I-POMDP$^{\gamma,\lambda_{1,2}}_{i,3}$</td>
<td>-522.3421</td>
<td>-339.8796</td>
</tr>
<tr>
<td>I-POMDP$^{\gamma,\lambda}_{i,3}$</td>
<td>-548.0494</td>
<td>-337.4591</td>
</tr>
<tr>
<td>wFP$^{\phi,\lambda}_{i}$</td>
<td>-1288.06</td>
<td>-775.96</td>
</tr>
</tbody>
</table>

Table 3.9: Log likelihood of the different models evaluated on the test folds using parameters as shown in Table 3.8.

We utilize the learned values in Table 3.8 to parameterize the underweighting and quantal responses within the I-POMDP based models and fictitious play. We cross-validated the models on the test folds. Using a participant’s actions in the first 5 trials, we initialized the prior belief distribution over the opponent types.

We show the simulation performance of I-POMDP$^{\gamma,\lambda}_{i,3}$, I-POMDP$^{\gamma,\lambda_{1,2}}_{i,3}$ and wFP$^{\phi,\lambda}_{i}$ in Figures 3.12, 3.13 and 3.14, respectively. While the model-based scores and trends for I-POMDP$^{\gamma,\lambda}_{i,3}$ also broadly reflect those of the experiment data, the performance of wFP$^{\phi,\lambda}_{i}$ bears some remarks. Notice the high prediction scores for the super-predictive group, which is due to the model using just the previous observation for its prediction and not accounting for any human error, present in the data. Furthermore, as was evident from the likelihood, the model simulates the experiment data for the myopic group especially poorly. In particular, the achievement score is not only lower in comparison to the experiment data but does not improve across blocks either. This is somewhat surprising given the initial improvement in the model’s prediction scores across the blocks. This is, in part, due to the observed play not translating into the conditionally rational choices given the myopic opponent.

We also measure the goodness of the fit by computing the mean squared error (MSE)
Figure 3.12: Comparison of I-POMDP_{γ,λ} based simulations with actual data in test folds: (a) Mean achievement scores and (b) Mean prediction scores.

Notice from Table 3.10, that both I-POMDP based models have MSEs that are significantly lower than the random model. Recall that the log likelihood of I-POMDP_{γ,λ}^{i,3} is higher than I-POMDP_{γ,λ}^{i,3} for the myopic opponent (see Table 3.9). This reflects in the
Figure 3.13: Comparison of I-POMDP\textsuperscript{\gamma,\lambda_1,\lambda_2} based simulations with actual data in test folds: (a) Mean achievement scores and (b) Mean prediction scores. Notice that the model-based scores exhibit similar values and trends as the experiment data, albeit the larger discrepancy in the first block. Additionally, the trend in the model-based scores appears more smooth compared to the actual data, possibly due to the large number of simulation runs.

difference in MSE of the achievement score for the myopic group between the two that is significant (Student’s paired t-test: $p = .015$). However, other MSE differences between the two models are insignificant and do not distinguish one model over the other across scores and groups. Although attributing non-normative action selection to the opponent did not
Figure 3.14: Comparison of $\text{wFP}_i^{\phi,\lambda}$ based simulations with actual data in test folds: (a) Mean achievement scores and (b) Mean prediction scores. Note the poor simulation performance, more so in the context of the myopic opponent group, despite some evidence of initial learning as seen from its prediction scores.

result in significantly more accurate expectations for any group, we think that it allowed the model to generate actions for agent $i$ that fit the data better by supporting an additional account of $j$’s (surprising) myopic behavior. This single positive result should be placed in the context of increased expense of learning an additional parameter, $\lambda_2$. Large MSE of $\text{wFP}_i^{\phi,\lambda}$ reflects its weak simulation performance although it does improve on the par set by
Table 3.10: MSE of the different models in comparison with the experiment data. Difference in MSE of the achievement score for the myopic group between the two I-POMDP models is significant.

random model for the super-predictive group.

3.4 Summary

The games in Goodie et al. [43] are particularly well-suited to rigorously measuring recursive thinking. This is because at each state the corresponding player’s rational action depends on how the others will act, if given the chance, and not on others’ previous action(s) in that game. The games that we modeled from Study 1 and Study 2 tested recursive reasoning up to two and three levels, respectively. We proposed a process-oriented model using finitely-nested I-POMDPs as the point of departure due to their explicit consideration of recursive beliefs and decision making based on such beliefs. The original finitely-nested I-POMDP was simplified to fit Goodie et al.’s game settings. We further augment I-POMDPs with elements that have support in behavioral and cognitive literature to model human decision making. These include an underweighting learning model for slow learning and a quantal response choice model for human non-normative choices.

In the Study 1 data model, we investigated the performance of our models using different...
measurements from the data. Results show that the model fits data better when relying only on action data instead of action and prediction data. This is because subjects’ prediction data may not be as accurate as action data, since their payoffs depend on action only. We further improved the performance of our models by adopting a changing quantal response parameter. This is from the observation that the subjects’ rationality error rates decreased substantially as they experienced more games.

We applied the extended base model developed in Study 1 to model the Study 2 data. Since subjects were deceived into thinking that the opponent is human, they may justify unexpected actions of the opponent as errors in their decision making rather than due to their level of reasoning. Hence, we generalize the previous model by attributing the quantal response choice to opponent’s action selection as well. The improved model performance, though not significant, suggests that humans may attribute choice error to other participants.
Chapter 4

Modeling Human Behavior in Sequential Bargaining Games

Sequential bargaining games (SBG) [92, 82], also called alternating-offer games, typically consist of a finite number of rounds of bargaining over the partition of a pie between a proposer and a responder. If a partition from the proposer is accepted by the responder, the game ends and the pie is divided according to this partition; otherwise, both players exchange their roles in the next round and repeat the process until a partition is accepted or the predefined number of rounds elapse in which case both players get nothing. Discount factors are applied to the pie in subsequent rounds for both players in order to make a disagreement in bargaining costly. The game-theoretic solution is derived by applying backward induction [92, 82].

Researchers have conducted several experiments on how humans behave in a variety of SBGs [8, 64, 67, 81] and related negotiation games [39]. Results from these experiments show that subjects primarily do not take the rational actions prescribed by backward

---

1The sequential negotiation game studied by Gal and Pfeffer [39] involves a proposer who offers an exchange of colored chips to a responder who may accept or reject it. If the offer is rejected, the proposer and responder swap roles. While sharing important similarities with the SBG, it differs in a key aspect that the utility of the chips does not reduce over the finite rounds.
induction. For example, in one-round bargaining games, also called ultimatum games, proposers in the experiments predominantly offer some amount that is less than half the pie but more than the rational offer to responders, and lower offers were rejected frequently [81]. In two-round SBGs, most opening offers are between the equal split and subgame perfect equilibria [8], though different discount factors for the proposers and responders affect the results slightly [67]. In the three-round and five-round extended games, we seldom observe subgame perfect equilibria amounts in the opening offers, and a majority of these are close to the second round pie size [64].

One reason commonly explored to explain human behavioral deviation from rational play is that social factors such as fairness and reciprocity may affect a human player’s utility [54, 39]. Based on this hypothesis, De Bruyn and Bolton [23] investigate the role of fairness by employing two different utility functions and incorporating them into a quantal response equilibrium framework [61] for recursively computing the expected utility at each round. Here the quantal response model accounts for noise and experience. The two compared utility models include one with an equity-reciprocity competition [68] and the Fehr-Schmidt model [34] that generalizes the previous utility model. The reported out-of-sample fits and model predictions on multiple data sets are consistent: the two models involving social factors exhibit better performance than the normative model, though the difference in performance varies for different data sets. Consequently, fairness is an important factor that needs to be considered while modeling bargaining.

Another reason that may potentially explain the deviation from backward induction in SBG is that humans may be exhibiting limited backward induction [54, 53]. Backward induction based solution is achieved under the assumption that players are rational and they believe that their opponents are rational who in turn believe that others are rational. However, this may not be the case for human players as research suggests that humans have bounded recursive thinking power [51]. Limited backward induction causes the deviation
from rational play to grow as the number of stages in a game increases.

Ho and Su [53] introduce a model that attributes level-$k$ rules of varying levels to the other agent with a Bayesian update of the distribution over the levels, allowing it to account for limited backward induction. This dynamic level-$k$ model was fitted to the experimental data [60] on Centipede games [80] giving a better likelihood compared to the default static level-$k$ and backward induction. In a preliminary outline, Ho and Su proposed its applicability toward SBG as well but did not use it to fit any data. SBG’s differ from Centipede games in having a much larger number of action choices for the proposer and role swapping, which significantly complicates the construction of these models.

We construct a dynamic level-$k$ model for SBG and a new model that generalizes the Poisson cognitive hierarchy [18] with a Bayesian belief update to account for learning in repeated SBG. In contrast to the level-$k$ model, a level-$k$ player, $k > 0$, in the Poisson cognitive hierarchy chooses the action that optimizes its belief, which assumes a Poisson distribution over all of the lower levels. While the dynamic level-$k$ model generally follows Ho and Su’s paradigm, our construction differs in two ways: an agent at level 0 maximizes its expected utility instead of acting randomly and I integrate a quantal response function [61] to model decision errors.

Next, we perform a comprehensive and comparative analysis of three different models for SBG data in combination with three different utility models (two of which involve social factors). Given the new ways of thinking about how humans play sequential games, such an analyses is previously lacking and is needed crucially for SBG where data from multiple experiments is available. By combining the new level-based models with social utility models, we analyze the fit of two dominant theories on human behavior in two-player sequential games. For this analyses, we construct a large data set from 6 different experiments and perform out-of-sample testing that evaluates the models on previously unseen data sets for robustness. Interestingly, no one particular model performs the best on all the data sets.
Rather, the level-based models perform better on data from SBG with an extended number of bargaining rounds while the recursive quantal response with social utility better models the behavioral data from SBG with less rounds.

4.1 Sequential Bargaining Games

SBGs are widely studied because of their widespread use and economic impacts [92, 82, 8, 64, 67, 81, 54, 23]. We describe the preliminaries of SBG and an outline of the experimental data.

In a multi-round SBG, two agents, \( i \) and \( j \), bargain over how to split a pie of \( c \) units in \( n \) rounds. In odd rounds (initial round, \( t = 1 \)), agent \( i \) proposes a split and offers \( x \) of the pie to agent \( j \). If agent \( j \) accepts this proposal, the game ends and the pie is divided accordingly with \( i \) receiving \( c - x \) and \( j \) receiving \( x \). If agent \( j \) rejects the offer, the game proceeds to round \( t + 1 \) and the pie shrinks according to the predefined discount factor, \((\delta_i, \delta_j)\), where \( \delta_i \) is a real number between 0 and 1 representing agent \( i \)'s cost of delay, and analogously for \( j \). After each round, the two agents exchange their roles (agent \( j \) is the proposer in even rounds). If no agreement is achieved after \( n \) rounds, the game ends and both players get nothing.

**Example 1** (Five-round SBG). An example of a 5-round game used by Neelin et al. [64] is shown in Figure 4.1 with a pie size of \( c = $15 \), discount factors, \( \delta_i = 0.34 \), \( \delta_j = 0.34 \). The smallest possible offer is \$0.01. Illustrations in this chapter will use this game.

The subgame perfect equilibrium of an \( n \)-round SBG is computed by backward induction. To illustrate, in 1-round (ultimatum) games, player \( j \) prefers some amount than nothing and would accept any offer greater than zero; player \( i \) therefore should propose the smallest possible offer. In 2-round games, the proposer in the second round is \( j \) who would offer the smallest possible offer to player \( i \) and keep slightly less than \( c\delta_j \), which is the size of the pie.
Figure 4.1: An example of a 5-round SBG from Neelin et al. (1988) with a pie size of $15 and discount factors, (0.34, 0.34). The split, \((a, b)\), denotes that portion \(a\) is allocated to agent \(i\) and \(b\) is allocated to agent \(j\). The equilibrium offer of \(i\) is to propose slightly more than $3.75 to \(j\) at the opening round.

In that round. Therefore, in the first round, in order to make \(j\) accept the offer, proposer \(i\) needs to offer slightly larger than what \(j\) would get in round 2. Hence, \(i\) should offer \(c\delta_j\) to \(j\) in the first round. In this way, we can compute the subgame perfect equilibrium for games with any number of rounds.

**Example 2** (Backward Induction). For the 5-round SBG shown in Figure 4.1, in the fifth round, responder \(j\) would accept any non-zero offer and be indifferent to an offer of 0 since this offer and rejection both give her 0. Proposer \(i\) knows this and would offer the smallest amount of 0.01 to \(j\). Hence, the equilibrium offer is, \(x_5^* = 0.01\) and the equilibrium split is \((0.20, 0.01)\) in round five where the pie size is 0.21. In the fourth round, proposer \(j\) knows that responder \(i\) would get 0.20 in the fifth round if \(i\) rejects her offer. Therefore, \(j\) needs to propose \(x_4^* = 0.21\) to make sure \(i\) will accept. The equilibrium split in round four is
Similarly, in round three, \( x^*_3 = 0.40 \), and the equilibrium split is \((1.34, 0.40)\); in round two, \( x^*_2 = 1.35 \), and the equilibrium split is \((1.35, 3.75)\). At the opening round, \( x^*_1 = 3.76 \), and the equilibrium split is \((11.24, 3.76)\).

Extensive experimentation with human subjects playing \textbf{SBG} [8, 46, 64, 67, 9, 81, 47] has yielded much data that may inform the computational modeling of behavior in \textbf{SBG}. Among these experiments, we select data from those in which the \textbf{SBGs} exhibit more rounds. This allows the performance of level-based models, such as level-\(k\) and Poisson cognitive hierarchy, to be adequately tested and distinguished. Specifically, it facilitates attributing the other agent different levels and maintaining beliefs over these levels, which is otherwise precluded by 1- and 2-round games. Subsequently, we select data for \textbf{SBGs} with 3- and 5-rounds. Furthermore, experiments in which a small pool of subjects repeatedly play \textbf{SBGs} allow the applicability of the dynamic aspects of these models. This forms our second criteria for selecting the data sets.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Pie Size</th>
<th>Num. of Rounds</th>
<th>Discount factor</th>
<th>Num. of subjects</th>
<th>Game repetitions</th>
<th>Mean opening</th>
<th>Game theoretic prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSS(^{5,1})</td>
<td>$5$</td>
<td>5</td>
<td>(0.34, 0.34)</td>
<td>80</td>
<td>1</td>
<td>0.343</td>
<td>0.250</td>
</tr>
<tr>
<td>NSS(^{5,4})</td>
<td>$15$</td>
<td>5</td>
<td>(0.34, 0.34)</td>
<td>30</td>
<td>4</td>
<td>0.359</td>
<td>0.250</td>
</tr>
<tr>
<td>OR(^{3,10})</td>
<td>$30$</td>
<td>3</td>
<td>(0.40, 0.40)</td>
<td>20</td>
<td>10</td>
<td>0.433</td>
<td>0.240</td>
</tr>
<tr>
<td>OR(^{3,10})</td>
<td>$30$</td>
<td>3</td>
<td>(0.60, 0.40)</td>
<td>20</td>
<td>10</td>
<td>0.450</td>
<td>0.160</td>
</tr>
<tr>
<td>OR(^{4,10})</td>
<td>$30$</td>
<td>3</td>
<td>(0.60, 0.60)</td>
<td>18</td>
<td>10</td>
<td>0.451</td>
<td>0.235</td>
</tr>
<tr>
<td>OR(^{4,10})</td>
<td>$30$</td>
<td>3</td>
<td>(0.40, 0.60)</td>
<td>18</td>
<td>10</td>
<td>0.466</td>
<td>0.350</td>
</tr>
</tbody>
</table>

Table 4.1: Game designs and collected data on \textbf{SBG} for our modeling. Values in the column, \textbf{Game repetitions}, indicate the number of games played by each subject in the pool. Values in the columns, \textbf{Mean opening} and \textbf{Game theoretic prediction}, are fractions of the opening pie size. The superscripts \(n, m\) in NSS\(^{n,m}\) or OR\(^{n,m}\) represent the number of rounds and the game repetitions, respectively. NSS data is from Neelin et al. (1988) and OR data is from Ochs and Roth (1989).

Having searched a wide range of experiment data sets, we find six that satisfy our requirements: data from 2 experiments with 5-round games under different conditions by Neelin et
al. [64] and data from 4 experiments with 3-round games under different conditions by Ochs and Roth [67] \(^2\). We summarize the experimental settings, results and the corresponding equilibrium predictions in Table 4.1. The four data sets from OR exploit different configurations for discount factors of both players and use higher values than NSS data sets. Observe that the averaged opening offers deviate substantially from equilibrium-based game theoretic predictions. They are close to the second round pie size (mean opening proportion is close to discount factor for the second player) as mentioned in [64].

4.2 Background on Modeling Sequential Bargaining Games

In this section, we cover the existing social utility models and a quantal response based choice model. The social utility models are functions that give the utilities for players based on the offer split. The utilities differ from the monetary payoffs by considering social factors such as fairness or reciprocity. The choice model provides a way for players to select one or more actions based on their utilities. A choice model may be integrated with different social utility models by considering different utility functions.

4.2.1 Social Utility Models

A well-known reason for behavioral deviations from rational play is that subjects may be influenced by social factors. One factor especially relevant to SBG is fairness (or equitability). For example, fairness plays an important role in the ultimatum game where the equilibrium which is accepting for the responder and offering smallest amount for the proposer is seldom observed and offers are between equilibrium and equal split [81]. Human decision making in bargaining games and in negotiation [39] not only considers the absolute payoff that is

\(^2\)I sought to collect more data sets on SBG by contacting several researchers but no more data was accessible.
received from an offer but also the relative utility by comparing the received offer to that of the opponent’s.

We consider two prominent utility models that include fairness considerations: equity-reciprocity competition (ERC) [68] and Fehr-Schmidt model (FSC) [23, 34].

In ERC, the utility of an offer proportion $\sigma$ for the responder, $U_{rsp}(\sigma; b)$, is defined as:

$$
U_{rsp}(\sigma; b) = \begin{cases} 
  c \left( \sigma - \frac{b}{2} \left( \sigma - \frac{1}{2} \right)^2 \right) & \text{if } \sigma < 1/2, \\
  c \sigma & \text{if } \sigma \geq 1/2,
\end{cases}
$$

where $c$ is the pie size, $\sigma$ is the proportion of the offer that a proposer makes and a responder receives, and $b$ is a fairness factor measuring the importance of any inequitable allocation. The utility function defined in ERC when the proportion is less than half is nonlinear and consists of two components: the relative utility due to negative reciprocity, $c \frac{b}{2} \left( \sigma - \frac{1}{2} \right)^2$, reduced from the absolute payoff, $c \sigma$. Observe that the model is asymmetric because the utility is increasingly less when the agent is offered less than half the pie but it is the same as the payoff when the proportion is more than half the pie. This asymmetricity is reasonable as players in bargaining games have been found to reject unfair offers more than propose fair offers [23].

We may obtain the utility for the proposer, $U_{prp}(\sigma; b)$, by substituting $\sigma$ with $1 - \sigma$ in the piecewise functions (Note that doing the substitution above will also reverse the order of the piecewise functions):

$$
U_{prp}(\sigma; b) = \begin{cases} 
  c (1 - \sigma) & \text{if } \sigma < 1/2, \\
  c \left( 1 - \sigma - \frac{b}{2} \left( 1 - \sigma - \frac{1}{2} \right)^2 \right) & \text{if } \sigma \geq 1/2,
\end{cases}
$$

In the second model FSC, the utility of offer proportion $\sigma$ for the responder, $U_{rsp}(\sigma; \alpha, \beta)$,
is defined as:

$$U_{rsp}(\sigma; \alpha, \beta) = \begin{cases} 
    c\left(\sigma - \alpha \left(\frac{1}{2} - \sigma\right)\right) & \text{if } \sigma < \frac{1}{2}, \\
    c\left(\sigma - \beta \left(\sigma - \frac{1}{2}\right)\right) & \text{if } \sigma \geq \frac{1}{2}
\end{cases} \quad (4.3)$$

where $\sigma$ is as defined in Equation 4.1, $\alpha$ is the negative reciprocity parameter measuring the penalty for inequity where the agent receives less than her opponent while $\beta$ is the positive reciprocity parameter measuring the penalty for inequity where she receives more. Another equivalent form is,

$$U_{rsp}(\sigma; \alpha', \beta') = c\left(\sigma - \alpha' \cdot \max((1 - \sigma) - \sigma, 0) - \beta' \cdot \max(\sigma - (1 - \sigma), 0)\right)$$

The utility function in FSC is symmetric, a more wholesome measure of fairness and is a linear variant of the one used in ERC when a player receives less than half the pie. When the proportion is greater than half, relative utility due to positive reciprocity is reduced as well. This symmetricity makes FSC more expressive than ERC, although at the expense of an additional parameter.

Similarly, we may obtain the utility for the proposer, $U_{prp}(\sigma; \alpha, \beta)$, by substituting $\sigma$ with $1 - \sigma$ in the piecewise functions in Equation 4.3 and reverse the order.

For comparison, we also consider the normative utility model (NORM) in which the utilities are simply defined as the payoffs that each player receives.

$$U_{rsp}(\sigma) = c \cdot \sigma$$
$$U_{prp}(\sigma) = c - c \cdot \sigma$$ \quad (4.4)

Relative utilities due to negative and positive reciprocity may also be crudely characterized as envy and guilt [78].
Example 3 (Social Utility Models). Let us significantly penalize unfairness by setting \( b = 10 \) in ERC or setting FSC’s negative reciprocity parameter, \( \alpha = 2 \) and positive reciprocity parameter \( \beta = 0.1 \). \( \alpha \) is often greater than \( \beta \) because humans take greater cognizance when they are treated unfairly. For the opening round with pie size $15, suppose there are two offers: offer 1 is to propose $9 to the responder (\( \sigma = 0.6 \)); offer 2 is to propose $4.5 to the responder (\( \sigma = 0.3 \)). The utilities of the two offers for the proposer using the different social utility models are shown in Table 4.2 using Equations 4.2-4.4.

Notice that when a player receives more than the fair value, there is no penalty by ERC. FSC differs where parameter \( \beta \) represents positive reciprocity.

<table>
<thead>
<tr>
<th>Utility</th>
<th>NORM ERC ((b = 10))</th>
<th>FSC ((\alpha = 2, \beta = 0.1))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( U_{prp}(\sigma = 0.6) )</td>
<td>6</td>
<td>5.25</td>
</tr>
<tr>
<td>( U_{prp}(\sigma = 0.3) )</td>
<td>10.5</td>
<td>10.5</td>
</tr>
</tbody>
</table>

Table 4.2: Utilities of two offers for the proposer when using three different social utility models: NORM, ERC and FSC. The total pie size is $15.

These three social utility models, NORM, FSC, and ERC are incorporated into different choice models.

4.2.2 Quantal Response Based Choice Model

De Bruyn and Bolton [23] incorporate different social utility models into a recursive quantal response framework, denoted as RQR, which provides a distribution over the subject’s actions. The quantal response choice model assigns probabilities to actions proportionally to their utilities. The cumulative probability for the responder accepting an offer proportion, \( \sigma^t \in [0, 1] \), in round \( t \), is computed by the logit function:

\[
P_{rsp}^{\sigma^t}(acc; \lambda) = \frac{e^{\lambda U_{rsp}(\sigma^t)}}{e^{\lambda U_{rsp}(\sigma^t)} + e^{\lambda U_{rsp}(\sigma^t)}},
\]

(4.5)
where $U_{rsp}(\sigma^t)$ is the utility of the responder accepting the offer $\sigma^t$; $U_{rsp}(\emptyset)$ is the utility of the responder rejecting it: it is 0 when $t$ is the last round, otherwise, it is the expected utility that the responder can get in the next round $t+1$ as a proposer, $E(U_{prp}(\sigma^{t+1}))$. The quantal parameter $\lambda$ controls how rational the agent is. The probability for the responder rejecting an offer proportion $\sigma^t$, $P_{rsp}^{\sigma^t}(rej)$, is defined similarly by replacing $U_{rsp}(\sigma^t)$ in the numerator of Equation 4.5 with $U_{rsp}(\emptyset)$.

The probability of the proposer making an offer $\sigma^t$ at round $t$, $P_{prp}(\sigma^t)$, is defined as:

$$P_{prp}(\sigma^t) = \frac{e^{\lambda \cdot E(U_{prp}(\sigma^t))}}{\sum_{\sigma'} e^{\lambda \cdot E(U_{prp}(\sigma'))}},$$

(4.6)

where $E(U_{prp}(\sigma^t))$ is the proposer’s expected utility of offering proportion $\sigma^t$ to the responder. It takes the probability of responder accepting this offer, $P_{rsp}^{\sigma^t}(acc)$, and rejecting it, $P_{rsp}^{\sigma^t}(rej)$, into consideration, $E(U_{prp}(1 - \sigma^t)) = P_{rsp}^{\sigma^t}(acc) \cdot U_{prp}(\sigma^t) + P_{rsp}^{\sigma^t}(rej) \cdot U_{prp}(\emptyset)$.

**Example 4 (RQR Model).** Allowing for deviation from the maximum expected utility action, let $\lambda$ be 0.4 and use NORM as the utility function. RQR computes the probability of the opening offer based on Equation 4.6. This requires the expected utilities of each possible offer at the opening round, which are computed by recursively calling Equation 4.5 and Equation 4.6 for later rounds. This recursion stops at the final round, and we illustrate how the final round works.

In the final round (the fifth round), the pie size is $0.21$. For any offer proportion $\sigma^{t=5} \in [0, 1]$, responder $j$’s utility is $0.21 \times \sigma^{t=5}$ for accepting and 0 for rejecting. Suppose the offer is $0.1$ ($\sigma^{t=5} = 0.1/0.21 = 0.476$), from Equation 4.5, the probability of $j$ accepting is 0.51.
The expected utility of this offer for proposer $i$ is,

\[
E(U_{prp}(\sigma^t)) = P_{rsp}^{\sigma^t}(acc) \cdot U_{prp}(\sigma^t) + P_{rsp}^{\sigma^t}(rej) \cdot U_{prp}(\emptyset)
\]

\[
= 0.51 \cdot 0.11 + 0.49 \cdot 0
\]

\[
= 0.0561
\]

The expected utility of $i$ can be similarly computed for each possible offer. Equation 4.6 gives the probability for each possible offer – for example, it is 0.0455 for the offer $\$0.1$. The probability of an offer at previous rounds is computed analogously.

RQR adapts between games by incrementing the quantal parameter with $\lambda'$ to account for experience effects, $\lambda^g = \lambda + (g - 1) \cdot \lambda$, where $\lambda^g$ represents the quantal parameter used in the $g$th game.

Both social utility models, ERC and FSC, along with the normative model NORM, were integrated with RQR and compared by De Bruyn and Bolton [23]. The results showed that overall the social utility models, RQR+ERC and RQR+FSC, outperform the normative model RQR+NORM; between the two social utility models, RQR+FSC performs relatively better than RQR+ERC.

### 4.3 Dynamic Level Based Choice Models

Previous modeling of SBG data has explored the role of fairness with the conclusion that considering it provides an improved fit of the data compared to the normative model [23]. Recently, Ho and Su [53] illustrate that the constraint of limited backward induction helps explain behavioral data in multi-stage sequential games such as Centipede games, and possibly better compared to considerations of social factors such as positive and negative reciprocity as in Fehr and Schmidt [34].
Multi-stage games exhibit the property of limited induction when a larger deviation from the backward induction based predictions is observed in extended stage games compared to games with lesser stages. In other words, limited induction causes the deviation to grow as the number of stages increases.

A dynamic level-$k$ model [53] captures this systematic violation of backward induction. Johnson et al. [54] noted that subjects in a SBG study paid significantly less attention to the later rounds and that social factors do not explain the data, thereby providing preliminary evidence that subjects are engaging in limited induction.

Our first contribution is a construction of the dynamic level-$k$ and Poisson cognitive hierarchy [15] models for the multi-round SBG, in order to investigate whether the limited backward induction factor provides an improved explanation of the behavioral data. Note that the latter represents violations of backward induction as well. We generalize these models to include belief-based learning for repeated SBG. This allows an agent to adapt its behavior to previously observed strategies of others.

4.3.1 Level-$k$ Model With Belief-Based Learning

Our dynamic level-$k$ model (DLK) for predicting the actions of an agent, say $i$, in an $n$-round two-agent SBG consists of iterative decision rules and a belief distribution over these rules. Experiments [64, 54] show that humans may treat SBG as truncated games with less rounds, or less rounds look ahead (i.e. they sometimes even do not check pie size of later rounds before making decisions). The decision rules therefore model SBG as truncated games with different rounds.

For an $n$-round game, agent $i$ may treat it as a truncated game with different rounds varied from 1 to $n$ believing the other agent $j$ has different steps of look ahead from 0 to $n - 1$. Rules of levels ranging from 0 up to $n - 1$ are therefore correspondingly attributed to $j$ for predicting its possible action(s). A level $l$ decision rule, $\theta_{j,l}$, where $0 \leq l \leq n - 1$
and having $l$ steps look ahead, is obtained by applying backward induction to an $l$-round truncated subgame.

Proposer $i$ at level $n$ initially has a Poisson distribution, parameterized by $\tau$, over the different decision rules of agent $j$ (probabilities are normalized over all possible levels):

$$b_i(\theta_{j;l}; \tau) \propto \frac{\tau^l e^{-\tau}}{l!} \quad 0 \leq l \leq n - 1,$$

(4.7)

where $b_i(\theta_{j;l}; \tau)$ represents the probability that $i$ believes $j$ is at level $l$.

**Example 5** (Belief in DLK). An example of proposer $i$’s belief in a 5-round SBG is given in Figure 4.2 where $\tau = 1.5$. Agent $j$’s level $l$ decision rule, $\theta_{j,l}$, where $0 \leq l \leq 4$, allows a look ahead of $l$ more steps from the current stage (first round responder and second round proposer). For example, agent $j$ at level 1 has one more step to look ahead, and believes that the game ends after one more move of $i$ who is at level 0.

![Figure 4.2: Level 5 belief in DLK for a 5-round SBG where $\tau = 1.5$. The decimals on the edges are the probabilities of each level. Notice that each agent has two roles at levels other than 0 in the final round and 5 in the opening round: the responder in round $t$ and the proposer in round $t + 1$. An agent’s level $l$ rule is a best response to her opponent’s level $l - 1$ rule.](image-url)
In an $n$-round game, each agent is associated with one specific level in her turn to play in the role of the current round responder as well as the next round proposer (the first move agent is only taking the role of proposer while the last move agent only responder). Each level rule is specified by a rational offer (or a rational split, which specifies the rational portions for both agents) and agents will act based on this rational split: the responder will accept an offer if the utility of this offer is no less than that of her rational split proportion and reject otherwise; the proposer will propose the rational offer to her opponent. An agent at level 0 acts as a utility maximizer believing she is the only agent in the game. An agent at level $l$ where $0 < l \leq n - 1$ best responds to her level $l - 1$ opponent.

Specifically, a rational proposer at level 0 in round $t$ always proposes an offer, $\sigma^t$, that maximizes her utility, $U_{prp}(\sigma^t)$, for example, offers nothing and keeps the whole pie if the utility model is NORM. A responder at level 0 in the final round accepts an offer, $\sigma^n$, if $U_{rsp}(\sigma^n) > U_{rsp}(\emptyset)$; rejects it if $U_{rsp}(\sigma^n) < U_{rsp}(\emptyset)$; is indifferent otherwise. A responder at any higher level in the final round acts similarly.

A proposer at level $l > 0$ offers a proportion of the pie, $\sigma^t$, to the responder, such that this offer maximizes her utility and the responder’s utility, $U_{rsp}(\sigma^t)$, is greater than $U_{prp}(\sigma_{opt}^{t+1})$ (the utility of the responder in the next round $t + 1$ as a proposer) in order for the responder to accept this offer. A rational responder in an intermediate round $t$ ($t < n$) accepts an offer $\sigma^t$ whose utility is greater than the utility she would get in the next round, $U_{prp}(\sigma_{opt}^{t+1})$, as a proposer, and rejects it otherwise.

In DLK, the proposer at level $n$ in the opening round, $t = 1$, offers the proportion, $\sigma^1$, that both maximizes her utility and the responder’s utility $U_{rsp}(\sigma^1)$ is greater than the expectation over the utilities of the responder’s rules at levels 0 to $n - 1$ in the next round. These utilities would differ because the optimal proportion of the offer is predicated on the

\[^4\text{Consequently, rules of levels higher than } n - 1 \text{ attributed to the other agent produce the same behavior as the rule of level } n - 1. \text{ Therefore, } k \text{ need not be greater than } n.\]
level. The expectation uses the Poisson distribution over the lower levels.

**Example 6 (DLK Level Rules).** With NORM as the utility model for simplicity, we show how each level rule computes in Figure 4.3. Note that each level rule is built iteratively from level 0 by applying backward induction while rules at level 0 in different rounds have different rational splits.

Agent \( j \) at level 4 reasons all the way up to the final round and therefore has the same split at level 4 as the equilibrium split (see Example 2 and Figure 4.3). Agent \( j \) at level 3 has 3 more rounds to look ahead. Working backwards, agent \( i \) at the corresponding level 0 in the role of responder in round four and as proposer in round five believes that there is no other agent in the game; she will keep all of the pie, \$0.21. Therefore the rational split is, \((0.21, 0)\) for this level 0. Rational splits for increasing levels are obtained by applying backward induction from one lower level split.

Agent \( i \) at highest level (level 5) in the opening round is calculated by computing the expected utility for \( j \) (at levels ranging from 0 to 4) in round two, here \( E(U_j) = 3.9725 \). The rational offer should guarantee that \( j \) gets more than \( E(U_j) \) and maximizes \( i \)'s utility. Hence, in the opening round, the rational split is \((11.02, 3.98)\).

Additionally, analogous to RQR, we utilize a quantal response to model potential errors in the agents’ behaviors. Given an offer in the current round \( t \), \( \sigma^t \), a responder may accept or reject this offer, denoted by \text{acc} \ and \text{rej}, respectively. The probability accepting an offer in round \( t \), \( \sigma^t \), at level \( l \) is given as:

\[
P^{\sigma^t, l}_{\text{rsp}}(\text{acc}; \lambda) = \frac{e^{\lambda U_{\text{rsp}}(\sigma^t)}}{e^{\lambda U_{\text{rsp}}(\emptyset)} + e^{\lambda U_{\text{rsp}}(\sigma^t)}},
\]

(4.8)

where \( U_{\text{rsp}}(\sigma^t) \) is the utility of the responder accepting the offer; \( U_{\text{rsp}}(\emptyset) \) is the utility of rejecting the offer: it is 0 if \( t \) is the final round, otherwise, it is the utility of next round’s optimal proportion that the responder would get (as the proposer), \( U_{\text{prp}}(\sigma_{\text{opt}}^{t+1, l}) \).

82
Figure 4.3: The optimal split at each level for the level-$k$ model using NORM. The pair $(U_i, U_j)$ at each level represents the optimal split for agent $i$ and $j$, respectively.

The probability for the responder rejecting the offer, $P_{rsp}^t(rej)$, is one minus the probability in Equation 4.8.

In order to compute the probability of making an offer, $\sigma^t$, for a proposer at level $l$, where $1 \leq l \leq k - 1$, we first need to decide whether the offer would be accepted or not. As we mentioned previously, an offer is accepted if the utility to the responder, $U_{rsp}^t(\sigma^t)$, is greater than the utility of the optimal proportion that the responder would get in the next round, $U_{prp}^t(\sigma_{opt}^{t+1})$, as a proposer. In this case, the utility of the offer to the proposer is simply $U_{prp}^t(\sigma^t)$. On the other hand, if the offer is rejected, its utility to the proposer is the utility of the optimal proportion that she gets in the next round, $U_{rsp}^t(\sigma_{opt}^{t+1})$, as a responder. The probability of proposing an offer $\sigma^t$ is:

$$P_{prp}^t(\sigma^t) = \frac{e^{\lambda U_{prp}^t(\sigma^t)}}{\sum_{\sigma'} e^{\lambda U_{prp}^t(\sigma')}}.$$  \hspace{1cm} (4.9)
where $\hat{U}_{prp}(\sigma^t)$ is a piecewise function: it is $U_{prp}(\sigma^t)$ if the offer is accepted, or $U_{rsp}(\sigma_{opt}^{t+1})$, otherwise.

At level $n$, the computation of $U_{prp}(\sigma_{opt}^{t+1})$ is complicated due to agent $i$’s Poisson distribution, $b_i$, over the lower level decision rules of $j$. Specifically, the decision rules of different levels lead to differing optimal proportions as computed by the responder in the next round as the proposer. Consequently:

$$U_{prp}(\sigma_{opt}^{t+1}) = \sum_{l=0}^{n-1} b_i(\theta_j,l;\tau)U_{prp}(\sigma_{opt}^{t+1,l})$$

Given the above modification, Eq. 4.9 now applies to a proposer at level $n$, where $U_{rsp}(\sigma_{opt}^{t+1})$ is computed analogously.

In the DLK model, agent $i$ updates its distribution, $b_i$, between games on observing the action of the responder to an offer, $\sigma^t$, when it plays the SBG repeatedly. We may update this distribution using a simple Bayesian belief update given the observation of acc or rej, which is then used as the belief in the next game,

$$b_i'(\theta_j,l|acc) \propto P_{rsp}(\sigma^t,l|acc)b_i(\theta_j,l;\tau), \quad (4.10)$$

where $P_{rsp}(\sigma^t,l|acc)$ is as defined in Equation 4.8. Analogously, the belief update on observing a rejection involves, $P_{rsp}(\sigma^t,l|rej)$. Note that the posterior may not remain a Poisson.

The level-$k$ model described above differs from Ho and Su’s dynamic model [53] for SBG in two important ways: (i) While in our model, the proposer and responder at level 0 acts to maximize its own utility only, Ho and Su let the level 0 proposer and responder select a random threshold, which serves as the demand and the acceptance threshold. (ii) The initial belief over the rules of the responder assumes a Poisson distribution in our model. On the other hand, Ho and Su’s construction places a probability 1 on a particular rule for the
responder, which may vary for different participants.

We may integrate social utility models defined in the previous section by substituting the normative utility function $NORM, U(\cdot)$, with those of ERC and FSC.

4.3.2 Cognitive Hierarchy Model With Belief-Based Learning

Another behavioral model that could systematically capture violations of backward induction is the cognitive hierarchy model [15, 18], which while sharing aspects with the level-$k$ model also differs in key ways. In the past, this model has predominantly been utilized to model behavioral data on single-shot normal form games [15, 98]. We extend this model significantly to the context of repeated SBGs by including a belief update, and denote it as DCH.

While agent $i$ at level $k$ models the other agent, $j$, with decision rules at all levels from 0 up to $k - 1$ where $k = n$ for an $n$-round SBG, DCH differs from DLK in that the agent at any level, $l$, ascribes decision rules of levels ranging from 0 to $l - 1$ to the other. An agent starts with a Poisson belief distribution over the rules of all lower levels.

Example 7 (Belief in DCH). We illustrate the belief structure for the proposer $i$ in the opening round for a 5-round SBG in Figure 4.4 using a Poisson distribution parameter value, $\tau = 1.5$, at each level. Notice that the agent at all levels ascribes decision rules of all lower levels to her opponent. This is different from DLK where only the proposer at the highest level at the opening round has a probability distribution over all lower levels while agents at lower levels ascribe decision rules of one less level only (see Figure 4.2).

This difference from DLK in the structure of the model leads to changes in how the proposer computes the optimal proportion to offer. A proposer at level 1 in any round $t$ computes $\sigma_{opt}^t$ similarly as in DLK since there is only one level below. However, a proposer at level 2 computes the optimal proportion that maximizes her expected utility where the expectation is due to the belief over her opponent in round $t + 1$ being at level 1 or 0. We may
apply this reasoning to a proposer at any level, \(1 < l \leq n\). Formally, the optimal proportion to offer in round \(t\) for a proposer, say \(i\), of level, \(L \geq 1\), is, \(\sigma_{opt}^{t,L} = \arg \max_{\sigma^t} U_{prp}(\sigma^t)\), such that:

\[
U_{rsp}(\sigma^t) \geq \sum_{l=0}^{L-1} b_i(\theta_{j,l}; \tau) U_{prp}(\sigma_{opt}^{t+1,l}),
\]

where \(b_i(\theta_{j,l}; \tau)\) is the belief that \(i\) has over the decision rules of the responder \(j\) and \(U_{prp}(\sigma_{opt}^{t+1,l})\) is the utility of \(j\) as a proposer in round \(t+1\) at level \(l\).

Example 8 (DCH Level Rules). With NORM as the utility model, we illustrate how each level rule computes the split in Figure 4.5. For level \(l\) where \(l > 1\) – similar to the highest level at opening round in DLK – an agent has a belief over all possible lower levels and the split is computed based on the expected utility of the opponent. Take agent \(j\) at level 4 acting as a round 1 responder and round 2 proposer for example. Then, expected utility of agent \(i\) as the responder in round 2 is:

\[
U_{rsp}(\sigma^{t=2}) = \sum_{l=0}^{3} b_j(\theta_{i,l}; \tau = 1.5) U_{prp}(\sigma_{opt}^{t=3})
\]

\[
= 0.239 \times 1.74 + 0.358 \times 1.13 + 0.269 \times 1.27 + 0.134 \times 1.29 \quad \text{(from Figure 4.5)}
\]

\[
= 1.3349
\]

Therefore, \(\sigma_{opt} = 1.34\), and the optimal split is \((1.34, 3.76)\) as shown in Fig 4.3. Similarly, we can compute the splits at all levels. Working backwards, the optimal split of the pie for the highest level at the opening round is \((11.08, 3.92)\).

Given the above sophistication in computing the optimal proportion that a responder at any level greater than 1 would offer in the next round, the probability distributions for the responder’s and proposer’s actions are computed identically to Equation 4.8 and Equation 4.9, respectively. Furthermore, the proposer’s belief over the decision rules of different levels ascribed to the responder is also updated similarly to Equation 4.10, with
one change from DLK: beliefs of the proposer at all levels from 2 to $k$ are updated. Finally, we may integrate the social utility models as in DLK.

4.4 Performance Analysis

A second contribution is a comparative analysis of the different potential models on the SBG population data. We compare the previously prominent choice model – RQR in combination with the social utility models, NORM, ERC and FSC – with the new level-based choice models, DLK and DCH. The normative utility functions in the latter models may be substituted with those emphasizing fairness and reciprocity such as ERC and FSC. Because the level-based models are representative of a different behavioral explanation, we seek to answer the following important questions:

1. Are there types of SBG where the recent hypothesis of limited backward induction offers a better explanation of the observed behavior, and, if so, why?

2. Is a combination of being fair and limited backward induction influencing behavior in SBG thereby offering an improved explanation of SBG data in comparison to each individually?

We note that these are new questions whose answers could potentially reveal new insights into bargaining behavior.

Our methodology is to utilize one of the data sets in Table 4.1 for learning the parameters of the different models by maximizing the log likelihood of the models. We refer to the latter as the fit of the model and a greater log likelihood indicates a better fit. This is followed by an out-of-sample evaluation of the performance of the models. By evaluating on data sets some of which obtained from different participant pools and on games exhibiting different properties, out-of-sample testing facilitates a more robust evaluation of the generality of the
learned models. It has been previously utilized in behavioral game theory to predict the performance of models [23, 98].

We evaluate the model performances in multiple ways:

- comparative fit based on the log likelihood,
- mean error in the predicted opening offers, and
- a visual analysis of the distributions of opening offers.

We select the NSS$^{5.4}$ data set for learning parameters in which each game is composed of 5 rounds and a subject repeats it 4 times though not with the same opponent.

4.4.1 Learned Parameters

The choice model, $\text{RQR}$, is parameterized by the quantal response parameter, $\lambda$, and the update parameter, $\lambda'$, that represents learning across the repeated games. Choice models, $\text{DLK}$ and $\text{DCH}$, are both parameterized by the quantal response parameter, $\lambda$, and the Poisson distribution parameter, $\tau$. Additionally, the social utility models, $\text{NORM}$, $\text{ERC}$ and $\text{FSC}$, use zero, one and two parameters, respectively.

We learn the parameters of the different choice models in combination with the varying utility models as maximum likelihood point estimates on the NSS$^{5.4}$ data set. In Table 4.3, we report their values and the corresponding log likelihoods. For each choice model, we emphasize the highest log likelihood on using the different utility models, in bold. The overall best fit is also underlined.

Observe that the level-based choice models, $\text{DLK}$ and $\text{DCH}$, fit the data set better in comparison to $\text{RQR}$. Integrating the social utility models with the level-based choice models further improves the fit, with both $\text{DLK}$ and $\text{DCH}$ exhibiting log likelihood that is substantially greater than that of $\text{RQR}$. This indicates that in the longer 5-round SBG, consider-
Table 4.3: Parameters for all combinations of the choice and utility models, obtained as maximum likelihood point estimates. A space indicates that the parameter is not applicable. Highest likelihood for each model is indicated in bold.

<table>
<thead>
<tr>
<th>Choice model</th>
<th>Parameter</th>
<th>Social utility models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NORM</td>
</tr>
<tr>
<td>RQR</td>
<td>$\lambda$</td>
<td>0.396</td>
</tr>
<tr>
<td></td>
<td>$\lambda'$</td>
<td>-0.005</td>
</tr>
<tr>
<td></td>
<td>$b$</td>
<td>12.293</td>
</tr>
<tr>
<td></td>
<td>$\alpha$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td></td>
</tr>
<tr>
<td>log likelihood</td>
<td>-309.510</td>
<td><strong>288.873</strong></td>
</tr>
<tr>
<td>DLK</td>
<td>$\lambda$</td>
<td>0.601</td>
</tr>
<tr>
<td></td>
<td>$\tau$</td>
<td>0.313</td>
</tr>
<tr>
<td></td>
<td>$b$</td>
<td>7.671</td>
</tr>
<tr>
<td></td>
<td>$\alpha$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td></td>
</tr>
<tr>
<td>log likelihood</td>
<td>-263.358</td>
<td>-238.384</td>
</tr>
<tr>
<td>DCH</td>
<td>$\lambda$</td>
<td>0.579</td>
</tr>
<tr>
<td></td>
<td>$\tau$</td>
<td>0.301</td>
</tr>
<tr>
<td></td>
<td>$b$</td>
<td>7.170</td>
</tr>
<tr>
<td></td>
<td>$\alpha$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td></td>
</tr>
<tr>
<td>log likelihood</td>
<td>-263.492</td>
<td>-238.426</td>
</tr>
</tbody>
</table>

Notice that the value of $\tau$ increases from approximately 0.3 in the normative case to greater than 12 on integrating the social utility models. This indicates that the probability of modeling the other agent as a level 0 agent drops from 0.72 to about 0. The former is consistent with the observation that opening offers in the NSS data being between $5 and $6 about 73% of the times, where $5.1 is the second round’s pie size and the best offer the proposer would make if she assumes her opponent is on level 0. Interestingly, social considerations provide an alternative explanation for this offer – mild levels of fairness –
thereby allowing the weight on level 0 to reduce dramatically.

### 4.4.2 Model Predictions

We perform out-of-sample predictions using the parameters learned previously as shown in Table 4.3 on the remaining data sets of Table 4.1. These data include observations of subjects playing a 5-round SBG with no repetition and multiple 3-round SBGs with repetition. Out-of-sample testing differs from \( n \)-fold cross validation because the test data sets may represent distributions that are different from that of the training set. This could be due to subjects being drawn from different pools and games with much different parameters. Subsequently, such testing provides a robust performance evaluation and comparison.

The out-of-sample log likelihoods of the different choice models in combination with the social utility models are shown in Table 4.4. They show distinct trends in the comparative performance of the different models: (a) DLK and DCH provide the best fit for NSS\(^5\)\(^1\) performing significantly better than RQR (Student’s paired t-test, \( p \)-value = 0.002). Among the different utility models, ERC and FSC do not significantly improve on normative for both DLK and DCH on this data set. (b) RQR in combination with either ERC or FSC provides the best fit for all the OR data sets. Furthermore, the performance of the level-based choice models degrades significantly for some of the OR data sets.

Based on the learned parameters, we may predict the mean opening offer proportions by the proposer averaged across all participants and games. We report the models’ predictions in Table 4.5. Furthermore, taking the NSS\(^5\)\(^1\) test data set as an example, we show in Figure 4.6, the distributions of the opening offer proportions as predicted by the choice and utility models, in comparison with the observed distribution from the data.

The observed initial offers from the data in Table 4.5 are close to the size of the pie in the next round for the responder. From Table 4.5, we observe that surprisingly the mean offer proportion predicted by RQR is closer to the observed offer in comparison to the level-based
Table 4.4: Likelihoods for the different choice models integrated with the three utility models. We include the fit on the training data, NSS\(^5,4\), for completeness. Highest likelihood for each data set is indicated in bold. * annotates likelihoods in a row whose difference is not significant.

<table>
<thead>
<tr>
<th>Choice model</th>
<th>Dataset</th>
<th>Social utility models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NORM</td>
</tr>
<tr>
<td>RQR</td>
<td>NSS(^5,4)</td>
<td>-309.51</td>
</tr>
<tr>
<td></td>
<td>NSS(^5,1)</td>
<td>-241.85</td>
</tr>
<tr>
<td></td>
<td>OR(^1,10)</td>
<td>-387.84</td>
</tr>
<tr>
<td></td>
<td>OR(^2,10)</td>
<td>-442.92</td>
</tr>
<tr>
<td></td>
<td>OR(^3,10)</td>
<td>-346.37</td>
</tr>
<tr>
<td></td>
<td>OR(^4,10)</td>
<td>-313.80</td>
</tr>
<tr>
<td>DLK</td>
<td>NSS(^5,4)</td>
<td>-263.36</td>
</tr>
<tr>
<td></td>
<td>NSS(^5,1)</td>
<td><strong>-219.49</strong>*</td>
</tr>
<tr>
<td></td>
<td>OR(^1,10)</td>
<td>-390.05*</td>
</tr>
<tr>
<td></td>
<td>OR(^2,10)</td>
<td>-764.19*</td>
</tr>
<tr>
<td></td>
<td>OR(^3,10)</td>
<td>-717.59</td>
</tr>
<tr>
<td></td>
<td>OR(^4,10)</td>
<td>-750.41</td>
</tr>
<tr>
<td>DCH</td>
<td>NSS(^5,4)</td>
<td><strong>-263.49</strong></td>
</tr>
<tr>
<td></td>
<td>NSS(^5,1)</td>
<td><strong>-220.24</strong>*</td>
</tr>
<tr>
<td></td>
<td>OR(^1,10)</td>
<td><strong>-383.32</strong>*</td>
</tr>
<tr>
<td></td>
<td>OR(^2,10)</td>
<td><strong>-746.28</strong>*</td>
</tr>
<tr>
<td></td>
<td>OR(^3,10)</td>
<td><strong>-699.48</strong>*</td>
</tr>
<tr>
<td></td>
<td>OR(^4,10)</td>
<td><strong>-730.91</strong></td>
</tr>
</tbody>
</table>

choice models though the latter demonstrated better log likelihoods previously. However, a closer look reveals the higher standard error of RQR’s offer predictions in comparison to the standard errors of the level-based choice models for the data set. Additionally, the prediction of the opening offer by the random model that picks an offer at random is close to the observed data for the OR data sets, but removed from the opening offer for the NSS data sets. Because the pie size in the next round for the OR data sets is close to 50% of the first round’s pie size, the close predictions of the mean random offer are coincidental.

The distributions of the opening offers in Figure 4.6 show that with minimal difference
between them, the two level-based choice models better simulate the distribution of opening offers compared to RQR by displaying merging peaks. Simulation results were improved with the integration of social models. However, the difference between different social models is not big.

### 4.5 Summary

Several experiments on humans engaging in SBGs have generated a large amount of behavioral data that is available for computational modeling. While previous models have focused on the relevance of fairness considerations while playing the games, we additionally explored whether considerations of a recently proposed theory of limited backward induction could provide an improved model of the data. In this respect, we utilized two extended models that capture violations of backward induction for modeling SBG. While these models have traditionally been applied to single-shot normal-form games, their application to SBG is
Our comprehensive empirical analyses with 9 different models provide evidence that limited backward induction and fairness both play important roles in how humans engage in SBG. However, the importance of each factor varies with the length of the game. In longer SBG with more rounds, limited backward induction plays a crucial role. Also, between the two models that capture violations of backward induction, we did not observe a significant difference indicating that the simpler of the two models, DLK, is sufficient. While for shorter rounds of SBG, fairness of the offer remains the key consideration. This is a significant finding as it suggests a non-social factor involving violations of backward induction by humans playing longer SBGs in the context of previous knowledge about SBG that predominantly relies on social fairness.
Figure 4.4: An agent at each level ascribes decision rules of all lower levels to the other, and maintains a belief distribution over them.
Figure 4.5: An agent at each level ascribes decision rules of all lower levels to the other, and maintains a belief distribution over them.
Figure 4.6: Distribution of opening offer proportions for test data set, NSS$^{5,1}$, by the different models and from observed data. Notice that RQR’s distribution is diffused and compares poorly with the observed distribution, whose support ranges from 0.1 to 0.6.
Part II

Strategic Behavior of Normative Agents in Multiagent Settings
Chapter 5

Existing Solutions to Multiagent Decision Making

Decision making is a key feature for any intelligent agent. As I mentioned in Chapter 1, I-POMDP is a framework which generalizes on POMDP, a single agent decision making framework, to a multiagent setting. Different from DEC-POMDP, which also extends from POMDP, I-POMDP does not make any assumption on the relationship between considered agents in the environment. Recent applications of I-POMDPs in multiple domains testify to its significance. They are being used to explore strategies for countering money laundering [66] and enhanced to include trust levels for facilitating defense simulations [86]. They have been used to produce winning strategies for playing the lemonade stand game [101], explored for use in playing Kriegspiel [40]. In Chapter 3, I-POMDPs was modified to include empirical models for simulating human behavioral data pertaining to strategic thoughts and actions.

Though the I-POMDP framework provides a guarantee of solution optimality in multiagent settings, its high computational complexity has motivated much research into methods of solving it more tractably. In this chapter, we will briefly review the solution to I-POMDPs as well as different existing algorithms on solving I-POMDPs.
5.1 I-POMDP Solution

In sequential decision making, considering an infinite horizon criterion with discounted future rewards, the objective of agents is to maximize their expected future rewards,

\[ V = E\left\{ \sum_{t=0}^{\infty} \gamma^t r_t \right\}, \]

where \( \gamma \) is the discount factor on reward \( r \). I-POMDP solutions (also known as a policy) are mappings from belief over interactive states to actions. In I-POMDPs, belief is a sufficient statistic for optimal decision making that fully summarizes the agent’s observation history [41].

5.1.1 Belief Update

Beliefs are updated after the agent’s action and observation using Bayes rule. Two differences complicate the belief update in multiagent settings: first, since the state of the physical environment depends on the actions performed by both agents the prediction of how it changes has to be made based on the probabilities that the subject agent attributes to various actions of the other agent by solving its models; second, changes in the models of the other agent have to be included in the update. The changes reflect the other’s observations and, if it is modeled intentionally, the update of other agent’s beliefs. In this case, the agent has to update its beliefs about the other agent based on what it anticipates the other agent observes and how it updates. Agent \( i \)'s updated belief over the interactive state,
is \_i,l = \langle s', \langle b'_{j,l-1}, \hat{\theta}'_j \rangle \rangle$, may be formalized using the following state estimation function:

\[
b'_i,i,i,l(a_i, o_i, b_{i,l}) = \beta \sum_{is_i,l(\hat{\theta}_j)=\hat{\theta}'_j} b_{i,l}(is_i,l) \sum_{a_j \in A_j} Pr(a_j|\theta_{j,l-1}) T_i(s, a_i, a_j, s') O_i(s', a_i, a_j, o_i) \times \sum_{o_j \in \Omega_j} O_j(s', a_i, a_j, o_j) Pr(b'_{j,l-1}|b_{j,l-1}, a_j, o_j),
\]

where $\beta$ is the normalization constant, $b_{i,l}$ and $b'_i,l$ are the initial and updated level $l$ beliefs of agent $i$, respectively. Interactive state $is_{i,l}$ is defined as $\langle s, \langle b_{j,l-1}, \hat{\theta}_j \rangle \rangle$. $T_i$ and $O_i$ are the transition function and observation function of agent $i$ respectively and are obtained from the I-POMDP definition, and $O_j$ is the observation function of agent $j$ as defined in $\hat{\theta}_j$. $Pr(a_j|\theta_{j,l-1})$ is the probability that $a_j$ is rational for a Bayesian agent $j$ modeled using $\theta_{j,l-1}$, and $Pr(b'_{j,l-1}|b_{j,l-1}, a_j, o_j)$ is 1 if $b_{j,l-1}$ updated using action, $a_j$, and observation, $o_j$, equals $b'_{j,l-1}$, and 0 otherwise. \(^1\)

### 5.1.2 Value Function

A solution to an I-POMDP is a policy, $\pi_i : \Delta(IS_{i,l}) \rightarrow \Delta(A_i)$, which maps agent $i$’s belief to a distribution over its actions, analogous to that of a POMDP. Each belief state in an I-POMDP has a value which is the maximum sum of future discounted rewards the agent can expect starting from that belief state. The value associated with belief state, $b_{i,l}$ is composed of the best of the immediate rewards that can be obtained in $b_{i,l}$, together with the discounted expected sum of utilities associated with belief states following $b_{i,l}$. Using

\(^1\)Precluding considerations of computability, if the prior belief over $IS_{i,l}$ is a probability density function, then $\sum_{is_i,l(\hat{\theta}_j)=\hat{\theta}'_j}$ is replaced by an integral over the continuous space. In this case, $Pr(b'_{j,l-1}|b_{j,l-1}, a_j, o_j)$ is replaced with a Dirac-delta function, $\delta_D(SE_{\theta_j,l-1}(b_{j,l-1}, a_j, o_j) - b'_{j,l-1})$, where $SE_{\theta_j,l-1}()$ denotes state estimation involving the belief update of agent $j$. These substitutions also apply elsewhere as appropriate.
Bellman’s equation,

\[ V_i(b_{i,l}, \hat{\theta}_i) = \max_{a_i \in A_i} \left\{ \sum_{i' \in A_i} b_{i',l}(i') \cdot E R_i(i', a_i) + \gamma \sum_{o_i \in \Omega_i} \Pr(o_i | a_i, b_{i,l}) \cdot V_i(SE_{\theta_i}(b_{i,l}, a_i, o_i)) \right\} \]

(5.2)

where \( E R_i(i', a_i) = \sum_{a_j} R(s, a_i, a_j) \Pr(a_j | \theta_{j,l-1}) \), and \( SE(.) \) is the state estimator function which denotes the belief update of agent \( i \) given initial belief \( b_{i,l} \), action, \( a_i \), and its observation, \( o_i \), as shown in Equation 5.1.

### 5.2 Existing Algorithms

Approaches for tractably solving POMDP-family models suffer two curses [73]: the curse of dimensionality in which belief space scales exponentially with the number of states and the curse of history in which the number of distinct action-observation histories considered grows exponentially with the planning horizon. As an extension and generalization of POMDPs to multiagent settings, solutions to I-POMDPs also suffer from these two curses.

A possible solution is to transform the original I-POMDPs into POMDPs by collapsing models within it to noise, and utilize the existing algorithms in POMDPs. However, this transformation is neither straightforward nor optimal considering the update of other agents’ nested beliefs is now missing.

Interactive particle filtering technique was the first effort towards solving I-POMDPs. Later on, two other algorithms, each of which exploited the traditional methods of value iteration and policy iteration in MDP-family models, were proposed. I briefly review each of these algorithms below.
5.2.1 Interactive Particle Filtering

Interactive particle filtering (I-PF) [27] is the first applicable algorithm for computing approximately optimal policies for the finitely-nested I-POMDPs. It generalizes particle filtering to multiagent settings. This generalization is not trivial because of the interactive belief hierarchy in I-POMDPs.

I-PF was proposed on the observation that while an agent’s belief defined over other agents’ models may be a complex infinite space, sampling methods are able to approximate distributions over such large spaces to arbitrary accuracy. It sought to mitigate the adverse effect of the curse of dimensionality by forming a sampled, recursive representation of the agent’s nested belief as a set of nested particles which is then propagated over time. This propagation is complicated as it needs to be performed at each of the hierarchical levels of the beliefs. The total number of particles needed for an arbitrary accuracy and the associated complexity grow exponentially with the nesting level.

I-PF reduces the infinite belief space to a finite sampled set. Solutions can be obtained by applying value iteration on the belief set propagated from the initial sampled belief in the form of a look ahead reachability tree. To mitigate the curse of history, a complementary method based on sampling observations while building the look ahead reachability tree during value iteration was developed.

Even though I-PF attempts to restrict the dimensionality by sampling, its accuracy is still impacted by the number of models which increases the need for more samples. I-PF does not scale to longer time horizons and it is better suited for solving I-POMDPs with a given prior belief.
5.2.2 Interactive Point-based Value Iteration

Interactive point-based value iteration (I-PBVI) [28] generalizes point-based value iteration (PBVI) to multiagent settings. I-PBVI exploits the piecewise linear and convex (PWLC) nature of I-POMDP value function and computes general solutions independent of the agent’s initial belief in the form of value vectors corresponding to various optimal policies.

Analogous to PBVI, I-PBVI maintains a finite set of belief points ($\tilde{B}_{i,l}$) and uses them to prune the back propagated value vectors. Therefore, at any given time, there are at most $|\tilde{B}_{i,l}|$ optimal policies.

However, since the models of the other agent are distributions over its interactive states, the interactive state of the subject agent is continuous. This is not suitable for computational purpose. Therefore, I-PBVI makes a simplifying assumption by limiting the initial models of the other agent to a finite set, $\tilde{\Theta}_{j,l-1}$. It then computes a set of all models reachable from the initial set of models in a fixed number of time steps ($H$), denoted as $\text{Reach}(\tilde{\Theta}_{j,l-1}, H)$. This restricts the set of interactive states to $\tilde{IS}_{i,l} = S \times \text{Reach}(\tilde{\Theta}_{j,l-1}, H)$. $\tilde{B}_{i,l}$ is chosen from set of distributions over $\tilde{IS}_{i,l}$.

Although I-PBVI attempts to mitigate the curse of history by limiting the number of policies generated and has succeeded to a large extent, it still suffers from the curse of dimensionality. The size of $\text{Reach}(\tilde{\Theta}_{j,l-1}, H)$ increases exponentially with $H$ thereby aggravating the curse of dimensionality.

5.2.3 Generalized and Bounded Policy Iteration

Generalized and Bounded Policy Iteration (I-BPI) [89] attempts to tackle both the curses: curse of history and curse of dimensionality. It does so by proposing an extension of a scalable policy iteration algorithm called bounded policy iteration (BPI) to I-POMDP. Unlike I-PF and I-PBVI which focus on solving finite horizon I-POMDPs, I-BPI computes discounted
infinite horizon solutions for I-POMDP.

Generally, infinite horizon policies are represented as a finite state controllers for the agent. Each node in the controller represents a policy and is associated with a value vector. I-BPI begins by assigning arbitrary controllers for the agents at each level (similar to one described in chapter 6). It then iteratively evaluates and improves these controllers interleaving these operations at each level.

I-BPI reduces the interactive states of agent $i$ at level $l$ to a cross over the set of physical states and the set of converged policies of $j$ represented by the nodes of its respective controller at level $l - 1$. Similarly, for agents at all lower levels down to level 1. At level 0, the interactive states are the same as physical states.

The policies in I-BPI are stochastic in actions as well as in observation, i.e. each policy (node in the controller) maps to a distribution over actions and on receiving an observations, the agent may stochastically transition to another policy (node).

The policy is evaluated by solving a system of linear equations. Policy improvement is done using linear program to replace a node in the controller by a node of better quality. Hence, the size of the controller remains fixed at each level. Therefore, the curse dimensionality and curse of history have no effect on solution complexity between iterations.

Despite tackling both curses, I-BPI is prone to get stuck at locally optimal solution. Although an escape technique has been suggested, the solution may never converge to a global optima. Much better solutions are possible for controllers of same size using alternate methods. I explore one such method in chapter 6.
Chapter 6

Multiagent Planning as Inference

A number of algorithms for solving MDP-family frameworks are based on two categorical algorithms: value iteration and policy iteration [83]. Attias ([2]) showed a new approach to MDP by transforming the original planning problem as a likelihood maximization problem which computes the posterior distribution over actions conditioned on reaching the goal state within a specified number of steps. In the new formulation, the “data” is the initial state and the final goal state or the maximum total reward. Given the parameters of the MDP, we are interested in finding the most likely sequence of actions. Toussaint et al. ([95]) extend the planning as inference to infer policies for POMDPs when the planning horizon is finite and infinite. For the latter case, the policy may assume the form of a finite state controller [74]. They infer the most likely controller using expectation-maximization (EM) on a mixture of increasing horizon dynamic Bayesian networks (DBN). Experiments indicate that good quality controllers of small sizes are inferred in multiple small problem domains though run time is a concern.

Given the compelling potential of this approach in bringing advances in inferencing to bear on planning, we generalize it to planning in multiagent settings as formalized by the finitely-nested interactive POMDP (I-POMDP) [41], which provides a principled framework
for sequential planning in possibly noncooperative multiagent settings. Previously, Kumar and Zilberstein ([58]) extended it to cooperative planning in multiagent settings. Our generalization allows its use in noncooperation where we may not assume common knowledge of an initial belief or common rewards, due to which others’ beliefs, capabilities and preferences are modeled.

Analogously to POMDPs, we formulate a mixture of finite-horizon DBNs, which include nodes of the other agent’s controllers as well. Models ascribed to the other agent differing in beliefs and possibly other parameters translate to multiple finite state controllers because the initial belief is a model parameter in the inferencing. The $E$-step is formulated as synchronous iterations of the forward-backward algorithm while the $M$-step obtains the most likely parameters of the stochastic controller. In addition to performing the maximization exactly, we explore a fast version discussed by Toussaint et al. ([94]) that greedily assigns a high probability mass to one action and next node state for each controller node, which is equivalent to greedily performing multiple maximization steps given a single expectation step.

An obvious way of applying EM to finitely-nested I-POMDPs would be to improve the controllers of the lower level until convergence before moving to the higher level. In contrast, we facilitate anytime behavior by interleaving the improvement of the controllers at different levels of nesting.

A second contribution is an approach for speeding up the non-asymptotic rate of convergence of the iterative EM. The objective function that measures the quality of the solution is multidimensional. Instead of the traditional approach of modifying the values of a large number of variables in each iteration, we may decompose the problem into optimization subproblems in which the objective function is optimized with respect to a single or a small subset (block) of variables, while holding the other variables fixed. This approach of block-coordinate descent is theoretically shown to exhibit faster rates of convergence under consid-
erably relaxed conditions on the objective function [35, 96, 84]. While it forms a candidate tool for multidimensional optimization and has been applied in other contexts [1, 36], we presents its first application toward planning as inference.

6.1 Background

In this section, we briefly outline previous EM-based planning in the context of POMDPs. Then, we review the technique of block-coordinate descent.

6.1.1 Expectation-Maximization for POMDPs

Attias ([2]) casts planning under uncertainty in the context of POMDPs as an action sequence likelihood maximization problem. Given model parameters, $T_i(s^t, a^t_i, s^{t+1}), O_i(s^{t+1}, a^t_i, o^{t+1}_i)$, $\Pr(a^{t+1}_i|a^t_i)$, and $\Pr(r^t|a^t_i, s^t)$, where the latter is the probability of obtaining reward, $r^t$, given action and state at time step, $t$, the model may be viewed as a hidden Markov model. Here, all the variables are hidden except for the initial observation and the reward upper bound, and whose transition function is $\Pr(a^{t+1}_i|a^t_i, o^0_i, R_{max})$, which can be computed using forward-backward passes.

Solutions of POMDPs are policies, which go beyond a simple sequence of actions. We may represent the policy of agent $i$ for the infinite-horizon case as a stochastic finite state controller (FSC) [74]. An FSC, $\pi_i$, for agent $i$ is defined as:

$$\pi_i = (\mathcal{N}_i, \mathcal{T}_i, \mathcal{L}_i, \mathcal{V}_i),$$

where $\mathcal{N}_i$ is the set of nodes in the controller, $\mathcal{T}_i : \mathcal{N}_i \times A_i \times \Omega_i \times \mathcal{N}_i \rightarrow [0, 1]$ represents the node transition function; $\mathcal{L}_i : \mathcal{N}_i \times A_i \rightarrow [0, 1]$ denotes agent $i$’s action distribution at each node; and a distribution over the nodes is denoted by, $\mathcal{V}_i : \mathcal{N}_i \rightarrow [0, 1]$. For convenience of
presentation, we group together $\mathcal{V}_i$, $\mathcal{T}_i$ and $\mathcal{L}_i$ in $\hat{f}_i$.

To infer a FSC, Toussaint et al. ([94, 95]) formulate a series of DBNs unrolled over increasing time steps, $T = 0 \ldots$, each of which emits a single reward, $r_T = 1$. In order to infer the controller, nodes and parameters of the controller are included in the DBN. The likelihood maximization problem becomes that of finding $\pi_i$, which maximizes the likelihood:

$$\arg \max_{\pi_i} \sum_{T=0}^{\infty} \Pr(r_T = 1 | T; \pi_i) \Pr(T).$$

As the trajectory of states, observations, nodes and actions are hidden at planning time, the likelihood maximization is solved in the schema of expectation-maximization (EM) [24]. The $E$-step takes an expectation of the hidden sequences in the DBN:

$$Q(\pi_i' | \pi_i) = \sum_{T=0}^{\infty} \Pr(r_T = 1, z_i^{0:T}, T; \pi_i) \log \Pr(r_T = 1, z_i^{0:T}, T; \pi_i'),$$

where sequence, $z_i^{0:T} = \{s^t, o_i^t, n_i^t, a_i^t\}_{t=0}^T$. The full joint, $\Pr(r_T = 1, z_i^{0:T}, T; \pi_i)$, may be computed using a simultaneous forward and backward pass. The $M$-step derives the distributions, $\mathcal{V}_i$, $\mathcal{T}_i$, and $\mathcal{L}_i$, of the controller, $\pi_i'$, which maximize the expectation, $Q(\pi_i' | \pi_i)$.

Instead of revising the parameters of the previous iteration, a greedy variant of the $M$-step selects parameter values for the controller, $\mathcal{V}_i(\cdot)$, $\mathcal{T}_i(\cdot | n_i, a_i)$, and $\mathcal{L}_i(\cdot | n_i)$, for each $n_i$, $a_i$, and $o_i$, that maximize the function, which may be seen as the expectation divided by the previous parameters.

### 6.1.2 Block-Coordinate Descent

Block-coordinate descent (BCD) [35, 96, 84] is an established iterative technique to gain faster non-asymptotic rate of convergence in the context of large-scale $N$-dimensional optimization problems. In this technique, within each iteration, a set of variables referred to as coordinates are chosen and the objective function, $Q$, is optimized with respect to one of the coordinate blocks while the other coordinates are held fixed.
Let $\Psi$ denote a block of coordinates, which is a non-empty subset of $\{1, 2, \ldots, N\}$. We may define a set of such blocks as, $B = \{\Psi_0, \Psi_1, \ldots, \Psi_C\}$, which is a set of subsets each representing a coordinate block with the constraint that, $\Psi_0 \cup \Psi_1 \cup \ldots \cup \Psi_C = \{1, 2, \ldots, N\}$. Note that $B$ could be a partition of the coordinates, although this is not required and the blocks may intersect. We also define the complement of a coordinate block, $\Psi_c$, where $c \in \{0, 1, \ldots, C\}$, as, $\tilde{\Psi}_c = B - \Psi_c$. To illustrate, let the domain of a real-valued, continuously differentiable, multidimensional function, $Q$, with $N = 10$ be, $\{v_1, v_2, v_3, \ldots, v_{10}\}$, where each element is a variable. We may partition this set of coordinates into two blocks, so that, $B = \{\Psi_0, \Psi_1\}$. Let $\Psi_0 = \{v_2, v_5, v_8\}$, and therefore, $\Psi_1 = \{v_1, v_3, v_4, v_6, v_7, v_9, v_{10}\}$. Here, $\tilde{\Psi}_0$ denotes the block, $\Psi_1$.

BCD converges to a fixed point such as a local or a global optima of the objective function under relaxed conditions such as pseudoconvexity of the function and requires the function to have bounded level sets [96]. In the absence of pseudoconvexity, BCD may oscillate without approaching any fixed point of the function. Nevertheless, BCD still converges if the function has unique optima in each of the coordinate blocks. In order to use BCD, we must ensure that each coordinate is chosen sufficiently often [96]:

**Definition 1 (Cyclic rule).** There exists a constant, $T \leq N$, such that every block, $\Psi_c$, is chosen at least once between the $i^{th}$ iteration and the $(i + T - 1)^{th}$ iteration, for all $i$.

In the context of the cyclic rule, BCD does not mandate a specific partitioning or an ordering scheme for the blocks. A simple way to meet this rule is by sequentially iterating through each block although we must continue iterating until each block converges to the fixed point.
6.2 Planning in I-POMDPs as Likelihood Maximization

We generalize planning as inference of Section 6.1.1 to the framework of I-POMDP\(_i,l\). For presentation simplicity, we focus on a two-agent setting; the other agent could have differing frames, which need not be the same as that of agent \(i\). Our approach generalizes to multiple other agents as mentioned below.

6.2.1 Transforming the Interactive State Space

Define a controller at level \(l\) for agent \(i\) as, \(\pi_{i,l} = \langle \mathcal{N}_{i,l}, \hat{f}_{i,l} \rangle\), where \(\mathcal{N}_{i,l}\) is the set of nodes in the controller and \(\hat{f}_{i,l}\) groups the remaining parameters of the controller as mentioned in Section 6.1.1. Analogously, we may define \(j\)’s level \(l-1\) controller, \(\pi_{j,l-1}\). Next, we define a set, \(\mathcal{F}_{j,l-1}\), where \(f_{j,l-1} \in \mathcal{F}_{j,l-1}\) is, \(f_{j,l-1} = \langle b_{j,l-1}^0, \hat{f}_{j,l-1}, \hat{\theta}_j \rangle\).

As we reviewed in Section 1.3, an intentional model, \(\theta_{j,l-1} \in \Theta_{j,l-1}\), consists of \(j\)’s belief, \(b_{j,l-1}\), and the frame. Consequently, each model, \(\theta_{j,l-1}\), may be mapped to \(f_{j,l-1}\) in the following way: \(b_{j,l-1}^0\) in \(f_{j,l-1}\) is the belief in \(\theta_{j,l-1}\), \(\hat{f}_{j,l-1}\) are the parameters of \(j\)’s controller and \(\hat{\theta}_j\) is the model’s frame. We denote this mapping as, \(\mathcal{K}_{\pi_j} : \Theta_{j,l-1} \rightarrow \mathcal{F}_{j,l-1}\).

The interactive state space, \(IS_{i,l} = S \times \Theta_{j,l-1}\), then becomes, \(IS_{i,l} = S \times \mathcal{F}_{j,l-1}\) where \(f_{j,l-1} \in \mathcal{F}_{j,l-1}\) is obtained as the output of \(\mathcal{K}_{\pi_j}(\theta_{j,l-1})\), for some model, \(\theta_{j,l-1} \in \Theta_{j,l-1}\). The general undecidability of the problem of obtaining an exact infinite-horizon solution of a POMDP [71] and that not all POMDPs admit an optimal FSC [48], implies that, in practice, the transformation of the interactive state space may be with some loss in generality. Given \(\pi_{j,l-1}\), this transformation of \(IS_{i,l}\) involves simply replacing the models with \(\mathcal{F}_{j,l-1}\), and is a constant-time operation.

If there are \(|\hat{\Theta}_j| > 1\) frames with differing capabilities and preferences, \(f_{j,l-1} \in \mathcal{F}_{j,l-1}\) may additionally differ in \(\hat{\theta}_j\) and possibly in \(\hat{f}_{j,l-1}\) as well. \(\mathcal{F}_{j,l-1}\) continues to contain as
many controllers as there are models. If there are $X$ other agents, the interactive state space becomes, $I S_{i,l} = S \times_{x=1}^{X} F_{x,l-1}$, where $F_{x,l-1}$ represents the mapped set for each agent $x$. This is because the agents may differ in their frames and receive private observations. Consequently, controllers may evolve or transition differently from each other.

### 6.2.2 Problem Formulation and Mixture Models

Analogously to POMDPs, the planning problem in multiagent settings formalized by finitely-nested I-POMDPs is modeled as a mixture of DBNs of increasing time from $T = 0$ onwards where the time prior $\Pr(T) = \gamma^T (1 - \gamma)$ for DBN with time length $T$. We show the networks with $T = 0$ and $T$-time slices in Figure 6.1.

*Figure 6.1: DBN for I-POMDP$_{i,l}$ with $i$’s level-$l$ policy represented as a standard FSC, with “node state” denoted by $n_{i,l}$. The dotted line in right highlights a clique of the graphical model; the junctions are $(s^t, n_{i,l}^t, n_{j,l-1}^t)$ for $t = 1, \ldots, T$, which is used in the E-step.*
The transition and observation functions of $\text{i-POMDP}_{i,l}$ parameterize the network, $\hat{r}_i^T$ is a binary random variable whose value is a reward of 0 or 1 emitted at the end of $T$ time steps with probability proportional to the reward function, $R_i(s,a_i,a_j)$, specifically, $\Pr(\hat{r}_i^T|a_i^T,a_j^T,s^T)$ is proportional to $\frac{R_i(s^T,a_i^T,a_j^T)-R_{\text{max}}}{R_{\text{max}}-R_{\text{min}}}$.

The networks include nodes of agent $i$’s level-$l$ FSC. Therefore, functions in $\hat{f}_{i,l}$ parameterize the network as well, which are to be inferred. Additionally, because the transformed interactive state space embeds the other agent’s possible controllers as well, the network includes nodes from each of $k$ possible $j$’s level $l-1$ controllers. As the candidate models ascribed to $j$ are independent of each other, no edges exist between the $k$ nodes in each time slice.

We transform planning in multiagent settings formalized by finitely-nested $\text{i-POMDPs}$ as a likelihood maximization problem:

$$\pi_{i,l}^* = \arg\max_{\pi_{i,l}} (1 - \gamma) \sum_{T=0}^{\infty} \gamma^T \Pr(\hat{r}_i^T = 1|T;\pi_{i,l}), \quad (6.2)$$

where $\pi_{i,l}$ is a level-$l$ FSC of agent $i$. The hidden variables in the mixture model, denoted by $z$, are $z = \{s, n_{i,l}, n_{j,l-1}, a_i, a_j, o_i, o_j\}$.

**Proposition 1** (Correctness). The likelihood maximization problem as defined in (6.2) with the mixture models as given in Fig 6.1 is equivalent to the problem of solving the original $\text{i-POMDP}_{i,l}$ of discounted infinite horizon whose solution assumes the form of a finite state controller.

**Proof.** Each finite-time $\text{i-POMDP}$ defines a likelihood of observing the binary reward at the last time step $T$,

$$L_{T}^{\pi_{i,l}} = \Pr(\hat{r}_i^T = 1|T,\pi_{i,l}) = E\{\hat{r}_i^T = 1|T,\pi_{i,l}\},$$

112
and for the full mixture of I-POMDPs we have,

\[ \Lambda^{\pi_{i,l}} = \sum_T \Pr(T) L_T^{\pi_{i,l}} = \sum_T \Pr(T) E\{\tilde{r}^T = 1 | T, \pi_{i,l}\}. \]

We defined \( \Pr(\hat{r}^T | a_i, a_j, s) \) as proportional to the reward function in the original I-POMDP. Therefore, the expectation \( E\{\hat{r}^T = 1 | T, \pi_{i,l}\} \) is proportional to the reward expectation \( E\{r^T; \pi_{i,l}\}_{t=T} \) at time step \( t = T \) of the original I-POMDP. Hence,

\[ \Lambda^{\pi_{i,l}} \propto \sum_t P(t) E\{r_t; \pi_{i,l}\} = (1 - \gamma) V^{\pi_{i,l}}. \]

6.2.3 EM for I-POMDP_{i,l}

Given the mixture model of Section 6.2.2, our challenge is to generalize the EM-based iterative likelihood maximization for POMDPs to the framework of I-POMDPs.

We show the \( E- \) and \( M- \)steps in detail.

**E-step**

In a multiagent setting, the hidden variables additionally include what the other agent may observe and how it acts over time. Therefore, let \( z^{0:T}_i = \{s^t_i, n^t_{i,l}, n^t_{j,l-1}, a^t_i, a^t_j, o^t_i, o^t_j\}_{t=0}^T \), where the observations at \( t = 0 \) are null. The full log likelihood involves an expectation over the hidden variables:

\[ Q(\pi'_i | \pi_{i,l}) = \sum_{T=0}^{\infty} \sum_{z^{0:T}_i} \Pr(\tilde{r}^T_i = 1, z^{0:T}_i, T; \pi_{i,l}) \times \log \Pr(\tilde{r}^T_i = 1, z^{0:T}_i, T; \pi_{i,l}') \]  \hspace{1cm} (6.3)

For clarity, let agent \( i \) ascribe a single model to \( j \). In the above expected log likelihood,
from Figure 6.1, \( \Pr(\hat{r}_i^T = 1, z_i^{0:T}, T; \pi_{i,t}) \) may be written as:

\[
\Pr(\hat{r}_i^T = 1, z_i^{0:T}, T; \pi_{i,t}) = \Pr(T) \Pr(\hat{r}_i^T = 1, z_i^{0:T}|T; \pi_{i,t})
\]

\[
=(1 - \gamma)^T \Pr(\hat{r}_i^T = 1|a_i^T, a_j^T, s^T) \mathcal{V}_i(n_{i,l}^0) b_i^0(s^0, n_{i,l-1}^0) \mathcal{L}_i(n_{i,l}^T, a_i^T) \mathcal{L}_j(n_{j,l-1}^T, a_j^T)
\]

\[
\times \prod_{t=1}^{T} \mathcal{L}_i(n_{i,l}^{t-1}, a_i^{t-1}) \mathcal{L}_j(n_{j,l-1}^{t-1}, a_j^{t-1}) \mathcal{T}_i(n_{i,l}^{t-1}, a_i^{t-1}, o_i^t, n_{i,l}^t) \mathcal{T}_j(n_{j,l-1}^{t-1}, a_j^{t-1}, o_j^t, n_{j,l-1}^t)
\]

\[
\times O_i(s^t, a_i^{t-1}, o_i^t) O_j(s^t, a_j^{t-1}, o_j^t) T_i(s^{t-1}, a_i^{t-1}, a_j^{t-1}, s^t)
\]  

(6.4)

In addition to the parameters of I-POMDP, which are given, parameters of both agents’ controllers are present in Equation 6.4 as well. If we assume that agent \( j \)'s level \( l-1 \) controller is given, Equation 6.4 allows us to find the parameters of agent \( i \)'s level \( l \) controller that maximize the likelihood. However, agent \( j \)'s lower-level controller is not given but is inferred as well. Equation 6.4 easily generalizes to more models ascribed to \( j \).

In a DBN of \( T \) time steps, the observed evidence includes the reward, \( \hat{r}_i^T \), at the end and the initial belief, and we seek to obtain the likely distributions, \( \mathcal{V}_i, \mathcal{T}_i, \) and \( \mathcal{L}_i \), for each time slice. Consequently, we may realize the full joint in the expectation using a forward-backward algorithm on a hidden Markov model whose state is, \((s^t, n_{i,l}^t, n_{j,l-1}^t)\). The transition function of this Markov model is,

\[
\Pr(s^t, n_{i,l}^t, n_{j,l-1}^t|s^{t-1}, n_{i,l}^{t-1}, n_{j,l-1}^{t-1})
\]

\[
= \sum_{a_i^{t-1}, a_j^{t-1}, o_i^t, o_j^t} \mathcal{V}_i(n_{i,l}^{t-1}, a_i^{t-1}, o_i^t, n_{i,l}^t) \mathcal{T}_i(n_{i,l-1}^{t-1}, a_i^{t-1}, o_i^t, n_{i,l}^t) \mathcal{L}_i(n_{i,l}^t, a_i^t) \mathcal{L}_j(n_{j,l-1}^{t-1}, a_j^t)
\]

\[
\times O_i(s^t, a_i^{t-1}, o_i^t) O_j(s^t, a_j^{t-1}, o_j^t) T_i(s^{t-1}, a_i^{t-1}, a_j^{t-1}, s^t)
\]  

(6.5)

For multiple models, we include the nodes, \( n_{j,l-1} \), in each controller.

The forward message, \( \alpha^t = \Pr(s^t, n_{i,l}^t, n_{j,l-1}^t) \), represents the probability of being at some
state of the Markov model at time step $t$:

$$
\alpha^t(s^t, n_{i,t}, n_{j,t-1}^t) = \sum_{s^{t-1}, n_{i,t-1}, n_{j,t-1}^{t-1}} \Pr(s^t, n_{i,t}, n_{j,t-1}^t | s^{t-1}, n_{i,t-1}, n_{j,t-1}^{t-1}) \alpha^{t-1}(s^{t-1}, n_{i,t-1}, n_{j,t-1}^{t-1}),
$$

(6.6)

where $\alpha^0(s^0, n_{i,t}, n_{j,t-1}^0) = \mathcal{V}_i(n_{i,t}) \theta^0_i(s^0, n_{j,t-1}^0)$, and $\Pr(s^t, n_{i,t}, n_{j,t-1}^t | s^{t-1}, n_{i,t-1}, n_{j,t-1}^{t-1})$ as in Equation 6.5.

The backward message gives the probability of observing the reward in the final $T^{th}$ time step given some state of the Markov model, $\beta^t(s^t, n_{i,t}, n_{j,t-1}^t) = \Pr(\hat{r}_i^T = 1 | s^t, n_{i,t}, n_{j,t-1}^t)$:

$$
\beta^h(s^t, n_{i,t}, n_{j,t-1}^t) = \sum_{s^{t+1}, n_{i,t+1}, n_{j,t-1}^{t+1}} \Pr(s^{t+1}, n_{i,t+1}, n_{j,t-1}^{t+1} | s^t, n_{i,t}, n_{j,t-1}^t) \beta^{h-1}(s^{t+1}, n_{i,t+1}, n_{j,t-1}^{t+1}),
$$

(6.7)

where $\beta^T(s^T, n_{i,t}^T, n_{j,t-1}^T) = \sum_{a_i^T, a_j^T} \Pr(\hat{r}_i^T = 1 | s^T, a_i^T, a_j^T) \mathcal{L}_i(n_{i,t}^T, a_i^T) \mathcal{L}_j(n_{j,t-1}^T, a_j^T)$, and $0 \leq h \leq T$ is the horizon. Here, $\Pr(\hat{r}_i^T = 1 | s^T, a_i^T, a_j^T) \propto R_i(s^T, a_i^T, a_j^T)$.

Messages $\alpha^t$ and $\beta^h$ give the probability of a state of the hidden Markov model (some time slice in a Bayesian network). As we consider a mixture of Bayesian networks, we seek the probabilities for all states in the mixture Markov model. Subsequently, we may compute the forward and backward messages at all states for the entire mixture model in one synchronous sweep.

$$
\tilde{\alpha}(s, n_{i,t}, n_{j,t-1}) = \sum_{t=0}^{\infty} \Pr(T = t) \alpha^t(s, n_{i,t}, n_{j,t-1})
$$

(6.8)

$$
\tilde{\beta}(s, n_{i,t}, n_{j,t-1}) = \sum_{h=0}^{\infty} \Pr(T = h) \beta^h(s, n_{i,t}, n_{j,t-1})
$$

(6.9)
M-step

We update the parameters, $\mathcal{L}_i$, $\mathcal{T}_i$ and $\mathcal{V}_i$, of $\pi_{i,l}$ to obtain $\pi'_{i,l}$ based on the expectation in the E-step. Specifically, taking the log of the likelihood in Equation 6.4 with $\pi_{i,l}$ substituted with $\pi'_{i,l}$ and focusing on terms involving the parameters of $\pi'_{i,l}$ gives us:

$$\log \Pr(\hat{r}^T = 1, z^0:T; \pi'_{i,l}) = \langle \text{terms indep. of } \pi'_{i,l} \rangle + \sum_{t=0}^T \log \mathcal{L}_i'(n_{i,l}^t, a_i^t) + \sum_{t=0}^{T-1} \log \mathcal{T}_i'(n_{i,l}^t, a_i^t, o_{i,l}^{t+1}, n_{i,l}^{t+1}) + \log \mathcal{V}_i'(n_{i,l})$$

(6.10)

Update of action probabilities – $\mathcal{L}_i$:

In order to update, $\mathcal{L}_i'$, we partially differentiate the Q-function of Equation 6.3 with respect to $\mathcal{L}_i'$. To facilitate differentiation, we focus on the terms involving $\mathcal{L}_i$, as shown below.

$$Q(\pi'_{i,l} | \pi_{i,l}) = \langle \text{terms indep. of } \mathcal{L}_i' \rangle + \sum_{T=0}^{\infty} \Pr(T) \times \sum_{t=0}^T \Pr(\hat{r}_i^T = 1, z_i^0:T; \pi_{i,l}) \log \mathcal{L}_i'(n_{i,l}^t, a_i^t)$$

(6.11)

$\mathcal{L}_i'$ on maximizing the above equation is:

$$\mathcal{L}_i'(n_{i,l}^t, a_i^t) = \frac{\mathcal{L}_i(n_{i,l}^t, a_i^t)}{C(n_{i,l}^t)} \sum_{s^t,n_{i,l-1}^t,a_j^t} \mathcal{L}_j(n_{j,l-1}^t, a_j^t) \tilde{\alpha}(s^t, n_{i,l}^t, n_{j,l-1}^t) \left[ \Pr(\hat{r}_i^T = 1 | s^t, a_i^t, a_j^t) + \gamma \frac{1}{1-\gamma} \sum_{s^{t+1}, n_{i,l}^{t+1}, n_{j,l}^{t+1}, o_{i,l}^{t+1}} \tilde{\beta}(s^{t+1}, n_{i,l}^{t+1}, n_{j,l}^{t+1}) \mathcal{T}_i(s^t, a_i^t, a_j^t, s^{t+1}) \times \mathcal{T}_j(n_{i,l}^t, a_i^t, o_{i,l}^{t+1}, n_{i,l}^{t+1}) \mathcal{T}_j(n_{j,l-1}^t, a_j^t, o_{j,l}^{t+1}, n_{j,l-1}^t) \mathcal{O}_i(s^{t+1}, a_i^t, a_j^t, o_{i,l}^{t+1}) \times \mathcal{O}_j(s^{t+1}, a_i^t, a_j^t, o_{i,l}^{t+1}) \right],$$

(6.12)

where $C(n_{i,l})$ is a normalization constant. For convenience, let us rewrite the above as,
\[
\mathcal{L}_i'(n_{i,l}^t, a_i^t) \propto \mathcal{L}_i(n_{i,l}^t, a_i^t) \, f_L(n_{i,l}^t, a_i^t).
\]

**Update of node transition probabilities – \( T_i \)**

Partially differentiating the Q-function with respect to \( T_i' \) and setting it to 0 yields:

\[
T_i'(n_{i,l}^t, a_i^t, o_{i}^{t+1}, n_{i,l}^{t+1}) = \frac{T_i(n_{i,l}^t, a_i^t, o_{i}^{t+1}, n_{i,l}^{t+1})}{C(n_{i,l}^t, a_i^t, o_{i}^{t+1})} \sum_{s', n_{j,l-1}^{t+1}, n_{j,l-1}^{t+1}} \alpha(s', n_{i,l}^t, n_{j,l-1}^{t+1}) \, \beta(s^{t+1}, n_{i,l}^{t+1}, n_{j,l-1}^{t+1})
\]
\[
\times \sum_{a_j^t, o_{j}^{t+1}} T_i(s^t, a_i^t, a_j^t, s^{t+1}) \, T_j(n_{j,l-1}^{t+1}, a_j^t, o_{j}^{t+1}, n_{j,l-1}^{t+1}) \, O_i(s^{t+1}, a_i^t, o_i^{t+1}) \, O_j(s^{t+1}, a_j^t, o_j^{t+1}) \, \mathcal{L}_i(n_{i,l}^t, a_i^t) \, \mathcal{L}_j(n_{j,l-1}^{t+1}, a_j^t),
\]

(6.13)

where \( C(n_{i,l}, a_i, o_i) \) is a normalization constant. We may rewrite the above as, \( T_i'(n_{i,l}^t, a_i^t, o_{i}^{t+1}, n_{i,l}^{t+1}) \propto T_i(n_{i,l}^t, a_i^t, o_{i}^{t+1}, n_{i,l}^{t+1}) \, f_T(n_{i,l}^t, a_i^t, o_{i}^{t+1}, n_{i,l}^{t+1}).

**Update of node distribution – \( V_i \):**

Analogously to the previous parameters, we may obtain the revised initial node distribution as:

\[
V_i'(n_{i,l}^0) = \frac{V_i(n_{i,l}^0)}{C} \sum_{s^t, n_{j,l}^{t+1}, n_{j,l-1}^{t+1}} \beta(s^t, n_{i,l}^t, n_{j,l-1}^{t+1}) \, b_i^0(s, n_{j,l-1}^{t+1}),
\]

where \( C \) is a normalization constant. We may rewrite the above as, \( V_i'(n_{i,l}^0) \propto V_i(n_{i,l}^0) \, f_V(n_{i,l}^0). \)

**Greedy M-step**

Analogously to the greedy maximization in EM for POMDPs, we may obtain a greedy variant of the \( M \)-step [93]. It may be viewed as greedily performing an infinite number of
maximizations while keeping the expectation function fixed. Let,

\[ a_i^{t*} = \arg \max_{a_i^t} f_L(n_{i,l}^t, a_i^t), \]

where \( f_L(n_{i,l}^t, a_i^t) \) is as defined in the previous subsection. Next, we obtain the revised action probability function of the controller by assigning a high probability to \( L_i(n_{i,l}^t, a_i^{t*}) \), and identical non-zero probabilities to the remaining actions whose value is in part random to possibly avoid local optima. Similarly, we maximize \( f_T(n_{i,l}^t, a_i^t, o_{i+1}^t, n_{i,l}^{t+1}) \) and \( f_V(n_{i,l}^t) \) for obtaining a greedy \( T_i \) and \( V_i \), respectively.

### 6.2.4 Interleaved Iterations for Anytime Behavior

Observe that EM iterations are influenced by the initial belief. Consequently, candidate models of the other agent with the same frame but differing in initial beliefs may result in converged controllers that may differ. Subsequently, IS\(_{i,l}\) embeds multiple \( j \)'s level \( l-1 \) controllers, one for each model regardless of whether the frame differs. This embedding of controllers terminates at level 0 where the corresponding state space consists of physical states only.

The presence of controllers at different levels for the agents leads to novel challenges and alternative approaches. An obvious way of extending EM to the context of finitely-nested I-POMDPs would be to improve the controllers of the lower level until convergence before moving to the higher level. Consequently, we may formulate the I-POMDP at a particular level as a POMDP with a complex state space. However, the higher-level controller may not be improved for some time until the lower level converges. Consequently, we present a more sophisticated approach that interleaves improvements of the other agent's controllers with improvements of the subject agent's controller. This approach facilitates anytime behavior with the challenge that the interactive state space may change dynamically at every iteration.
6.2.5 Block-Coordinate Descent in EM

BCD may speed up the non-asymptotic rate of convergence of EM for both I-POMDPs and POMDPs. The challenge is to determine which of the many variables should be grouped into blocks and how.

We empirically show that grouping the number of time slices, \( t \), and horizon, \( h \), in Equations 6.8 and 6.9, respectively, into coordinate blocks of equal size is beneficial. In other words, we decompose the mixture model into blocks containing equal numbers of Bayesian networks. Formally, let, \( \Psi_t \) be a subset of \( \{ t = 0, t = 1, t = 2, \ldots, t = T_{\text{max}} \} \). Then, the set of blocks is, \( B^t = \{ \Psi_1, \Psi_2, \Psi_3, \ldots \} \). In practice, because both \( t \) and \( h \) are finite (say, \( T_{\text{max}} \)), the cardinality of \( B^t \) is bounded by some \( C \geq 1 \). Analogously, we define the set of blocks of \( h \), denoted by \( B^h \).

Obtaining \( \hat{\alpha} \) in the E-step now computes \( \alpha^t \) for the time steps in a single coordinate block, \( \Psi^t_{c} \), only, while using the values of \( \alpha^t \) for the complementary coordinate blocks, \( \Psi^t_{c} \), from the previous iteration. Analogously, we obtain \( \hat{\beta} \) in each iteration by computing \( \beta^h \) for the horizons in \( \Psi^h_{c} \) only, while using \( \beta \) values from the previous iteration for the remaining horizons. In order to meet the cyclic rule, we choose a block, \( \Psi^t_{c} \), at iterations, \( i = c + qC \) where \( q \in \{0, 1, 2, \ldots\} \).

For example, let us bound \( t \) and \( h \) to \( T_{\text{max}} = 100 \) and form 5 equal-sized coordinate blocks. Then, say, in the 7th iteration, we will compute \( \alpha^t \) and \( \beta^h \) for the time steps and horizons in the block, \( \Psi^t_{2} \) and \( \Psi^h_{2} \) respectively.

\[
\hat{\alpha}(s, n_{i, t}, n_{j, t-1}) = \sum_{t \in \Psi^t_{2}} \Pr(T = t) \alpha^t(s, n_{i, t}, n_{j, t-1}) + \sum_{t \notin \Psi^t_{2}} \Pr(T = t) \alpha^t_{-i}(s, n_{i, t}, n_{j, t-1})
\]

\[
\hat{\beta}(s, n_{i, t}, n_{j, t-1}) = \sum_{h \in \Psi^h_{2}} \Pr(T = h) \beta^h(s, n_{i, t}, n_{j, t-1}) + \sum_{h \notin \Psi^h_{2}} \Pr(T = h) \beta^h_{-i}(s, n_{i, t}, n_{j, t-1})
\]

Here, \( \alpha^t_{-i} \) and \( \beta^h_{-i} \) are obtained from previous iteration. Subsequently, we may perform
either the exact or greedy $M$-step. We may view this scheme as performing several quick “mini-iterations” of the $E$- and $M$-steps in place of a larger iteration that was performed in the original EM. As a result, $\hat{\alpha}$ and $\hat{\beta}$ are updated over all the time steps at the end of an entire cycle of several mini-iterations.

### 6.3 Computational Complexity

In the $E$-step, we compute $\hat{\alpha}$ and $\hat{\beta}$ (Equations 6.8 and 6.9), which are then used in the $M$-step. Each of these has complexity, $\mathcal{O}(T_{\text{max}}S^2|N_{i,l}|^2(|M||N_{j,l-1}|)^2)$, where $T_{\text{max}}$ is a bound on $T$ in practice and $|M|$ is the number of models of the other agent considered at each level. In order to compute $\hat{\alpha}$ and $\hat{\beta}$, we need the transition function of the Markov model for given current and next states in the $E$-step, which has complexity of $\mathcal{O}(S^2|N_{i,l}|^2(|M||N_{j,l-1}|)^2|A_i||A_j||\Omega_i||\Omega_j|)$. $E$-step’s net complexity is given by $\mathcal{O}(T_{\text{max}}S^2|N_{i,l}|^2(|M||N_{j,l-1}|)^2)$.

Complexity of the $M$-step is in updating the parameters, $\mathcal{L}_i$, $T_i$ and $\mathcal{V}_i$. The first two parameters each require in the order of $\mathcal{O}(S^2|N_{i,l}|^2(|M||N_{j,l-1}|)^2|A_i||A_j||\Omega_i||\Omega_j|)$. Computing the distribution, $\mathcal{V}_i$, has complexity, $\mathcal{O}(|N_{i,l}|S||M||N_{j,l-1}|)$. Therefore, the $M$-step’s complexity is dominated by that of computing the action (or transition) probability of the controller. These complexities are incurred as many times as the number of iterations.

If $|M|$ is the maximum number of models ascribed to the other agent at each level, then $\text{I-POMDP}_{i,l}$ contains $\mathcal{O}(|M|^l + 1)$ models. Because we associate each model with a controller that is improved, the overall computational complexity of the algorithm is that of the $E$- and $M$-steps run until convergence scaled by the exponential number of models. This high complexity limits the number of candidate models we may ascribe to others in this approach.
6.4 Experiments

We begin by demonstrating the *anytime* property of the interleaved EM inference-based planning. Four variants of our EM-based approach are used as appropriate: the exact EM inference-based planning (labeled as I-EM); replacing the exact M-step with its greedy variant (I-EM-Greedy); iterating EM based on coordinate blocks as described previously (I-EM-BCD), and lastly, replacing the exact maximization in I-EM-BCD with its greedy variant (I-EM-BCD-Greedy).

We use two problem domains – the noncooperative version of the *multiagent tiger problem* [27] ($|S| = 2$, $|A_i| = |A_j| = 3$, $|O_i| = |O_j| = 6$ for level $l > 0$ and $|O_i| = |O_j| = 3$ at level 0, and $\gamma = 0.9$). In addition to this toy problem, our testbed utilizes the larger noncooperative *money laundering problem* [66]. This is a game between law enforcement (blue team) and the money launderers (red team) who aim to move their assets from a ‘dirty’ pot to a ‘clean’ one through a series of financial transactions while evading capture by the blue team. This problem exhibits 99 physical states for the subject agent (blue team), 9 actions for the subject agent and 4 for the opponent, and 11 observations for the subject and 4 for the other.

As EM requires an initial seed controller and its convergence is often sensitive to the seed, we experiment with a total of 8 seeds generated randomly for both problems. Unlike I-BPI for which we may escape local optima by adding nodes, EM keeps the number of nodes in the controllers fixed. For the tiger problem, our seed controllers contain 5 nodes, which is sufficient to generate good quality behavior. Smaller controllers of size 3 are utilized for the other. We consider level 2 I-POMDPs for both domains with 5 models (or frames) for the tiger problem at each level (total of 25 models across levels) and 3 models for the laundering problem. All experiments were run on Linux with Intel Xeon 2.6GHz CPUs and 32GB RAM.
Figure 6.2: Interleaved improvement of the FSCs across the different levels demonstrates anytime behavior for I-POMDP_{i,2} in the tiger (top) and the larger money laundering contexts (bottom).

In Figure 6.2, we compare the convergence of the interleaved I-EM with the straightforward I-EM that proceeds bottom up across the levels, for one seed. We utilize the greedy variant for the larger money laundering problem. In both domains, interleaved improvement clearly demonstrates the anytime characteristic while the top-most level in the non-interleaved approach does not start improving until about 700 seconds elapse for the tiger
problem and 15,000 seconds for the laundering problem. The corresponding anytime property of the interleaved approach is evident for all other seeds. Observe that the bottom up I-EM-Greedy converges to a slightly better value in the laundering problem though this occurs for some seeds only.

Next, we explore any improvements to the convergence rate introduced by BCD for both
domains. In contrast to previous works [58, 95], we report the time elapsed instead of the number of iterations. This is because each iteration could consume excessive time and our use of BCD introduces quick “mini-iterations” several of which replace one original iteration. Clearly, Figure 6.3 shows that BCD in both I-EM and its greedy variant speeds up the non-asymptotic rate of convergence with the controllers improving comparatively quickly, for the same seed as in Figure 6.2. Notice that BCD may not modify the final fixed point.

Figure 6.4: Comparison of EM with greedy and BCD with I-BPI for tiger (top) and money laundering (bottom).
Finally, we compare the quickest of the EM variants, I-EM-BCD-Greedy, with a previous policy iteration based technique for I-POMDPs that operates on FSCs as well, I-BPI [89]. All methods are given the same set of seeds and we report on the best convergence value and times among the seeds, for each method. The results for a level 2 I-POMDP for the tiger and the money laundering context are shown in Fig 6.4. While the EM based approaches take more than an order of magnitude more time to converge, with the same number of nodes in the controller, they can achieve significantly better values. In particular, we obtain the controller with a value of 7.5 for the tiger problem using I-EM-BCD-Greedy compared to −10 using BPI. Figure 6.4 also gives an indication of the total time taken by EM to produce a converged controller: about 21 mins for the tiger problem and 7 hours for the larger money laundering.

A limitation of EM is its convergence to local optima leading to controllers whose quality is unpredictable. Overall, EM based iterative inference is considerably slower in converging compared to approaches such as bounded policy iteration. In particular, each EM iteration for the larger laundering problem consumes about 5 mins at level 2. The greedy M-step reduces the number of iterations of the original EM on average by a factor of 15 while BCD further reduces the number of iterations taken to reach the stable values by a factor of 2.5.

6.5 Summary

Previous algorithms on solving finitely-nested I-POMDP adopt the classic approach which focus on planning by using Bellman’s equitation. Attias [2] showed that planning under uncertainty can be reformulated as a likelihood maximization problem, which may be solved using probabilistic inferencing techniques. Toussaint et al. ([94, 95]) and Kumar et al. [59] have successfully applied EM on solving POMDPs and DEC-POMDPs. Inspired by their works, we provide the first formalization of planning in I-POMDPs as an inference problem.
and derive EM solution for the transformed problem.

An obvious way of applying EM to finitely-nested I-POMDPs is to solve lower level until convergence before moving to the higher level. In contrast, we facilitate anytime behavior by interleaving the improvement of the solutions at different levels of nesting.

One drawback of EM iterative method is its slow convergence. To speed up the rate of convergence, block-coordinate descent, a approach which theoretically exhibits faster rates of convergence under considerably relaxed conditions, is explored. Two problems with different size are tested. The experiments results show that we may obtain solutions represented as compact finite state controllers whose quality is significantly better than previous policy iteration techniques though convergence may take more time.
Chapter 7

Future Work

In this dissertation, we presented promising results on modeling human behavioral data in two classes of games: modified Centipede games and sequential bargaining games. We also computed normative behaviors in multiagent settings within the contexts of I-POMDPs. I outline some of open avenues for future improvement here.

7.1 Behavior of Humans

In each class of games investigated in this dissertation, different modeling methods were developed based on its characteristics. Specifically, in modified Centipede games, where recursive reasoning is the key, I-POMDP was used as the point of departure for its explicit consideration of recursive beliefs and decision making based on such beliefs. In the sequential bargaining games where limited backward induction and fairness are both possible explanations for human behavioral deviation from theoretic predictions, both elements were incorporated into the modeling.

In fact, most research in modeling human behavioral data focuses on developing models for one specific class of games. For example, Ficici et al. focused on Colored Trails (three-
player negotiation) problems [38, 37], Ray et al. focused on Investor-Trustee games [78]. De Bruyn et al. worked on a large collection of data obtained from games in different studies; however, all games are sequential bargaining ones [23]. Similarly as modeling in normal-form games, though data are collected from different studies, games still share the same structure [98].

Research in human behavioral modeling rarely investigates more than one type of games. Ho and Su developed a dynamic level-k model to explain human behavioral deviations from theoretic predictions in sequential games [53]. They tested the applicability of the model on the Centipede game and the sequential bargaining game. However, solid results are only provided for the first type of game.

As such, one possible line of future work is to investigate the applicability of our models on other strategic games which share the same characteristics as our games.

As mentioned in Section 1.3, the finitely-nested I-POMDPs break the infinitely nested beliefs by defining level 0 beliefs. The finitely-nested I-POMDPs can also be seen as a level based framework. Given the similarity between the finitely-nested I-POMDP and the level based models developed in Chapter 4, another line of future work is merging these two models together. The simplified finitely-nested I-POMDPs in Section 3.2 serves as a salient starting point. One challenge is how to modify the specification of the reward function defined in I-POMDP so that it is capable of incorporating different utility functions.

### 7.2 Behavior of Normative Agents

As demonstrated in the experiments from Section 6.4, we transform planning in finitely-nested I-POMDP into a probabilistic inference problem. This approach provided better solutions than previous policy iteration techniques. However, as EM algorithm takes significantly longer time to converge, an obvious line of future work is to investigate other faster
inferencing techniques.

Sticking with EM algorithm, the significant longer running time for solving I-POMDPs is due to three factors: the number of total EM runs, the convergence time (or number of iterations) within each EM run, and each iteration’s running time within EM. In Sections 6.2.3 and 6.2.5, two speed-up techniques have been explored for the second factor. Two other factors are left as future work.

The first factor, the number of total EM runs, grows exponentially with the number of models considered in the belief. Models sharing the same frame still need to go through different EM runs if they have different initial belief. One line of future work is to avoid multiple EM-runs for each model within the belief. A possible solution is to construct a series of full and complete DBN where all models are included. The challenge for this solution is how to construct the big DBN and how to compute forward and backward messages within the DBN.

The last factor, each iteration’s running time within EM, depends on the running time of the E-step and the M-step in each iteration. Another line of future work is to reduce running time on the E-step which is the bottleneck. As a possible solution, Monte Carlo EM [97, 100] employs sampling method to speed up inference in the E-step.
Bibliography


